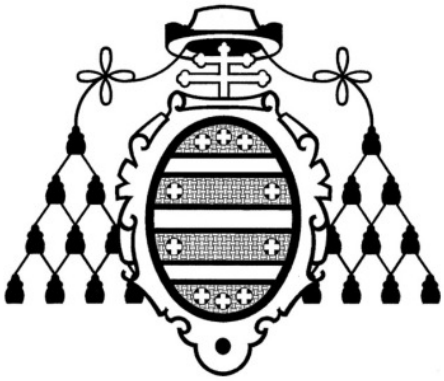
UNIVERSIDAD DE OVIEDO



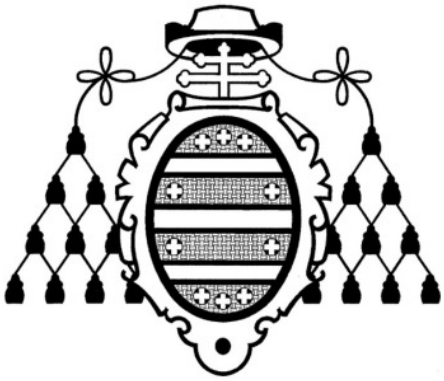
SISTEMAS Y SERVICIOS INFORMÁTICOS PARA INTERNET

TESIS DOCTORAL

MÉTODOS SEMÁNTICOS DE REUTILIZACIÓN DE DATOS ABIERTOS ENLAZADOS EN LAS LICITACIONES PÚBLICAS

Jose María Álvarez Rodríguez

UNIVERSIDAD DE OVIEDO



SISTEMAS Y SERVICIOS INFORMÁTICOS PARA INTERNET

TESIS DOCTORAL

MÉTODOS SEMÁNTICOS DE REUTILIZACIÓN DE DATOS ABIERTOS ENLAZADOS EN LAS LICITACIONES PÚBLICAS

Jose María Álvarez Rodríguez

Resumen

Las Administraciones Públicas son uno de los mayores compradores de la Unión Europea, ya que sus adquisiciones en conjunto representan un 17 % del Producto Interior Bruto. La tendencia está encaminada a la agilización de este gran mercado de oportunidades para facilitar el acceso a las empresas, especialmente a las pequeñas y medianas empresas (PYMEs), posibilitando el acceso a la información presente en los contratos públicos, permitiendo así la construcción de un entorno de mercado competitivo.

Desde otro punto de vista, el creciente uso de Internet durante los últimos años ha puesto de manifiesto un nuevo entorno de ejecución para las aplicaciones, utilizando como nueva plataforma la Web. Nuevas tecnologías y paradigmas están emergiendo para dar soporte al desarrollo y despliegue de aplicaciones y servicios, así como para la publicación de datos e información. En este sentido, iniciativas como la Web Semántica, en concreto el esfuerzo de *Linked Data* dentro de la nueva "Web de Datos", intentan elevar el significado de los elementos y recursos que están disponibles en la web con el objetivo de mejorar la integración e interoperabilidad entre aplicaciones y facilitar el acceso a la información y los datos a través de modelos y formatos de datos de conocimiento compartido unificados.

En el caso particular del proceso de contratación pública electrónica surge la necesidad de abordar problemas de gran calado tales como la dispersión de la información, la heterogeneidad de los formatos de los anuncios, la diversidad de formatos de explotación o el multilingüismo y multiculturalidad. Con el propósito de proporcionar una solución a estas necesidades mediante la aplicación de la iniciativa de Web Semántica y el esfuerzo propuesto por *Linked Data*, se realiza un estudio de una serie de métodos semánticos, materializados a través de un ciclo de vida de datos enlazados de carácter general, y su aplicación específica al dominio de la contratación pública electrónica.

De esta forma se mejora el acceso a la información y a los datos de los anuncios de licitación, aspecto clave para el incremento de las oportunidades de participación en los procesos de contratación pública, favoreciendo la publicidad de los anuncios de licitación e impulsando un entorno de datos abiertos, estratégico para las Administraciones Públicas tanto por su carácter económico como corporativo debido al movimiento *Open Data*.

Palabras clave

contratación pública electrónica, *e-Procurement*, *e-Government*, *linked data*, *open data*, *linking open data*, *linking open government data*, Web de Datos, web semántica

Abstract

Public administrations are one of the largest buyers of the European Union representing the 17 % of the total GDP. Current trends try to provide a new agile market of opportunities to ease the access to public procurement notices to companies, more specifically to Small and Medium Companies (SMEs), making the construction of a competitive pan-European market possible.

From other point of view, the emerging use of Internet in the last years has generate a new realm for the execution of applications and the deployment of new services using the Web as platform. New technologies and development models are arising to deal with the requirements of this new context in which new added-value services, publication of data and information are required. In this sense, initiatives such as Semantic Web, in particular through the *Linked Data* effort within the new Web of Data, seeks for raising the meaning of information resources on the Web with the major objective of improving the integration and interoperability among applications and easing the access to existing information and data by means of shared common models and formats.

In the e-Procurement context new relevant issues have to be resolved in order to fulfill the needs of existing problems with regards to the information dispersion, diverse publishing and exploitation formats or the multilingualism and multiculturalism. Taking into account the main goal of delivering a new solution to deal with these requirements applying the principles of the Semantic Web and *Linked Data* initiatives, an innovative and in-depth study to establish a set of semantic methods as part of a *Linked Data* life cycle and their application to the public contracts domain is carried out.

Thus the access to existing information and data extracted from the notices out is improved as key-enabler to increase the opportunity of joining in cross-border public procurement processes, boosting a new environment in which new advertising techniques for the notices based on open data are provided and aligning the outcomes of this work with the current corporate strategy of public administrations related to open data in one of the main economic domains.

Key-words

e-Procurement, *e-Government*, *linked data*, *open data*, *linking open data*, *linking open government data*, *Web of Data*, *semantic web*

Agradecimientos

Quiero que mis primeras palabras sean de agradecimiento a mi familia, mis padres, mi hermana y mi cuñado, por la confianza y el apoyo incondicional que me han brindado durante estos últimos años, por su espíritu de colaboración, su paciencia, su "saber hacer", sustento fundamental para conseguir la tranquilidad y seguridad necesarias para encarar la gran responsabilidad que supone afrontar la elaboración de un trabajo de investigación de este calado.

Agradecer también a todos mis compañeros de trabajo, por su apoyo y estímulo, tanto a aquellos con los que he compartido las fases iniciales de mi andadura profesional en el campo de la investigación convirtiéndose, sin duda, en "algo más", "Luis", "Milín", "Cabo Miguel", Diego, "Wikier", Iván Mínguez, Tejo, Alejandro, etc., como a mis actuales compañeros en el Grupo WESO y en el Departamento de Informática de la Universidad de Oviedo.

De igual forma mi reconocimiento y gratitud a todos mis amigos: "Ares", "Lea", Hugo, "Viñas", Héctor ("Ettore"), "Mele", "Basi", "Rake", Carla, Irene, "Berto", la peña "De la que vas plas", etc., por todos los buenos momentos que hemos pasado ajenos al entorno digital, por su lealtad y confianza tanto a nivel personal como en mi trabajo.

También agradecer al proyecto "10ders Information Services", y particularmente a la empresa Euroalert.net, la oportunidad que me han ofrecido para desarrollar mi trabajo y experimentación, posibilitando el acercamiento de la investigación a un entorno real.

Mención especial a mi director y mentor Dr. José Emilio Labra Gayo, por su incombustible aliento e impulso en toda mi trayectoria, guía en el camino de la investigación, por su empuje y por animarme a pensar yendo un paso más allá de lo evidente afrontando con éxito los retos que se presentan y a hacer las cosas "bien".

Historial de este documento

Fecha	Versión	Comentarios
Mayo/2011	0.1	Inicio del documento
Junio/2011	0.2	Estructura del documento
Septiembre/2011	0.3	Introducción y Estado del Arte
Octubre/2011	0.4	Contratación Pública y *e-Procurement*
Noviembre/2011	0.5	*Linked Data*, *Open Data* y Web Semántica
Diciembre/2011	0.6	Ciclo de Vida de Datos Enlazados
Enero/2012	0.7	Aplicación de MOLDEAS al *e-Procurement*
Febrero/2012	0.8	Validación y Experimentación
Marzo/2012	0.9	Versión alfa del documento
Abril/2012	1.0	Versión final del documento

Sobre este documento

Este documento recoge el estudio realizado sobre la aplicación de métodos semánticos para la producción, publicación y consumo de datos enlazados abiertos en el contexto de la administración pública electrónica y concretamente en el campo de la contratación pública electrónica o *e-Procurement*. Esta memoria de investigación titulada como "Métodos Semánticos de Reutilización de Datos Abiertos Enlazados en las Licitaciones Públicas" se ha elaborado para la obtención del título de Doctor por la Universidad de Oviedo.

El documento se encuentra organizado de la siguiente forma:

- "Introducción" (ver Capítulo 1): se realiza una motivación del problema de investigación a resolver identificando el plan y la metodología de trabajo seguida durante el proceso de elaboración de la tesis doctoral.
- "Contratación Pública y *e-Procurement*" (ver Capítulo 2): se presenta la casuística del dominio del problema a estudiar, así como las soluciones actuales más cercanas al enfoque desarrollado durante la investigación en el dominio de la administración electrónica centrando los esfuerzos en la valoración de los procesos de contratación pública electrónica.
- "Panorámica de uso de la Web Semántica y *Linked Data*" (ver Capítulo 3): se repasan desde un punto de vista general el estado del arte de estas iniciativas, así como su aplicación en los distintos dominios.
- "Definición de Métodos Semánticos" (ver Capítulo 4): se realiza la definición teórica de una serie de procesos, métodos y tareas a realizar para la promoción de datos siguiendo las directrices de *Linked Data* y atendiendo a los principios de *Open Data*.
- "Métodos Semánticos en el ámbito de las Licitaciones Públicas" (ver Capítulo 5): se materializan los procesos, métodos y tareas definidos en el capítulo anterior para su aplicación en el campo de los anuncios de licitación públicos.
- "Sistema MOLDEAS" (ver Capítulo 6): se describe la implementación de los componentes realizados para dar soporte a los procesos, métodos y tareas de carácter semántica en el dominio de las licitaciones públicas.
- "Experimentación y Validación" (ver Capítulo 7): se diseña el experimento y se evalúan y valoran los resultados obtenidos de la investigación realizada.
- "Conclusiones y Trabajo Futuro"(ver Capítulo 8): se comentan las conclusiones obtenidas tras la realización de esta memoria de investigación. También se marcan las líneas de evolución futura.
- "Impacto y Difusión" (ver Apéndice A): se detallan las actividades de difusión realizadas a través de los distintos canales durante la elaboración de la tesis doctoral valorando su impacto en la comunidad científica e industrial.

- “Trabajos publicados” (ver Apéndice B): se enumeran las publicaciones científicas realizadas que han surgido como parte del trabajo realizado durante la elaboración de la investigación.
- “Tablas de Validación” (ver Apéndice C): se presentan las tablas de validación pormenorizadas atendiendo a todos los criterios definidos en la experimentación de la investigación.

Sobre el autor

Jose María Alvarez Rodríguez es Ingeniero en Informática (2007) e Ingeniero Técnico en Informática de Sistemas (2005) por la Universidad de Oviedo. En junio de 2008 recibió el Premio al Mejor Proyecto Fin de Carrera de Ingeniería Informática por su proyecto "Activación de Conceptos en Ontologías mediante el algoritmo de Spreading Activation" otorgado por el Colegio Oficial del Principado de Asturias.

Desde abril del año 2005 a enero del año 2010 desempeña su actividad investigadora en el área de tecnologías semánticas del departamento de I+D+i de la Fundación CTIC de Asturias, especializándose en servicios web semánticos, sistemas basados en reglas, sistemas de búsqueda y datos enlazados. Durante esta etapa participa activamente en la redacción y ejecución de proyectos de investigación en distintos ámbitos destacando: regional (proyecto PRAVIA-PCTI Asturias cod. IE05-172), nacional (proyecto PRIMA-Plan Avanza-cod. TSI-020302-2008-32) y europeo (proyecto ONTORULE-FP7 cod. 231875). Como fruto de esta actividad es autor de diversas publicaciones en el área de semántica de carácter internacional en conferencias y *workshops* realizando su trabajo de investigación en 2009 bajo el título de "Interoperabilidad e Integración en Arquitecturas Orientadas a Servicios basadas en Semántica" para el Programa de Doctorado 34.1-"Sistemas y servicios informáticos para internet" del Departamento de Informática de la Universidad de Oviedo. Durante el curso 2008/2009 obtiene una plaza de profesor asociado (ref. Código: F036-75-DL0X041-AL6H. Bopa Nº 181, 4 de Agosto de 2008) impartiendo docencia en el área de Ciencias de la Computación e Inteligencia Artificial del Departamento de Informática de la Universidad de Oviedo.

Actualmente se encuentra contratado por el grupo WESO del Departamento de Informática de la Universidad de Oviedo dirigido por el profesor Dr. José Emilio Labra Gayo para la ejecución de distintos proyectos de investigación a nivel regional y nacional, destacando el proyecto *10ders Information Services* (Plan Avanza-cod. TSI-020100-2010-919) y *ROCAS-Reasoning On the Cloud Applying Semantics* (Ministerio de Ciencia e Innovación-TIN2011-27871) lo que le ha habilitado para la consecución de becas de carácter competitivo internacional, como la obtenida en marzo de 2012 a través de la red europea: *HPC-Europa2: Pan-European Research Infrastructure for High Performance Computing*. También ha compaginado su actividad investigadora con una plaza de profesor asociado en el área de Lenguajes y Sistemas Informáticos del Departamento de Informática de la Universidad de Oviedo durante el curso académico 2011/2012 participando en la co-dirección de proyectos fin de carrera. Finalmente y como parte esencial de la actividad investigadora es autor y revisor de varias conferencias y revistas de carácter internacional en el área de tecnologías semánticas.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

¿Por dónde empiezo, Majestad?
Empieza por el principio. Y
sigue hasta que llegues al final.
Entonces te detienes.

Alicia en el país de las Maravillas
LEWIS CARROL

Las Administraciones Públicas son uno de los mayores compradores de la Unión Europea (UE), ya que sus adquisiciones en conjunto representan un 17 % [12] del Producto Interior Bruto (PIB). La tendencia se dirige a intentar agilizar este gran mercado de oportunidades, que asciende a unos 2, 155 billones de euros, para facilitar el acceso a las empresas, especialmente a las pequeñas y medianas empresas (PYMEs), posibilitando el acceso a la información presente en los contratos públicos y permitiendo así la construcción de un entorno de mercado competitivo. Por otra parte, un punto clave reside en el impulso de la transparencia y la eficiencia en los propios procesos administrativos asociados a la contratación pública. Esta situación ha supuesto la germinación de una voluntad en las Administraciones Públicas para tratar de simplificar el proceso de contratación pública tanto desde un punto vista legal, como operativo. Las acciones de cambio vienen impulsadas desde la Unión Europea a través de distintas directivas que son transpuestas en los distintos Estados Miembros, dando soporte a estos procesos de adquisición de bienes y servicios. No obstante, las oportunidades comerciales que representan estos contratos públicos están fragmentadas en cientos de fuentes, en general de carácter local, representando una dilatada sucesión de anuncios de licitación que no llegan a tener una difusión adecuada, lo cuál es inaudito en la era de las comunicaciones instantáneas y globales.

Por otra parte, teniendo en cuenta que la información incluida en los anuncios de licitación tiene la consideración de Información del Sector Público (PSI), se deben proveer los mecanismos necesarios de acceso a esta información para facilitar la consulta de los datos contenidos en los anuncios de licitación. Actualmente la corriente *Open Data* dentro del seno de las organizaciones e individuos en general y de las entidades públicas en particular, ha generado un entorno de datos abiertos en el cual deben formar parte necesariamente los contratos públicos. Esta sensibilidad emergente respecto a la apertura de datos implica una necesaria relación con la contratación pública debido al valor de la información que contienen, tanto para crear nuevos servicios de valor añadido agregando datos de distintas fuentes, como medio de información para las empresas interesadas en atender a procesos de contratación pública.

Desde otro punto de vista, el creciente uso de Internet durante los últimos años ha puesto de manifiesto un nuevo entorno de ejecución para las aplicaciones, utilizando como nueva plataforma la Web, el gran sistema distribuido. Nuevas tecnologías y paradigmas están emergiendo para dar

soporte al desarrollo y despliegue de aplicaciones y servicios, así como para la publicación de datos e información. Los modelos de desarrollo están evolucionando hacia un estilo más colaborativo en el cual las empresas ofrecen su software como servicios (*Software as a Service*-SaaS), materializado a través del paradigma de *Cloud Computing*, implementado con tecnología de servicios con el objetivo de que terceros puedan utilizar estos nuevos servicios y datos para la construcción de aplicaciones agregadas con valor añadido.

En este sentido, iniciativas como la Web Semántica [31] dentro de la nueva "Web de Datos" o *Web of Data* a través de modelos y formatos de datos de conocimiento compartido unificados, intentan elevar el significado de los elementos y recursos que están disponibles en la web con el objetivo de mejorar la integración e interoperabilidad entre aplicaciones, impulsando la implantación de este enfoque. Dentro de la iniciativa de Web Semántica hay que destacar principalmente dos esfuerzos:

1. La iniciativa *Linked Data* [23], propone la publicación de datos enlazados siguiendo el modelo RDF [16] para facilitar la creación de una web de datos en la que éstos se puedan mostrar, intercambiar y conectar a través de URIs. La tendencia actual de publicación de datos enlazados está marcando una evolución en la interoperabilidad de aplicaciones con el consiguiente efecto que conlleva para las relaciones *Business to Business* (B2B), *Business to Client*(B2C) o *Administration to Administration* (A2A). Entre los casos de éxito se podrían destacar: administración electrónica (iniciativa de *Open Government Data* [184]-OGD), contratación pública electrónica de bienes y servicios (*e-Procurement*), oferta formacional, contextualización de aplicaciones, etc.

2. El desarrollo de lenguajes y formalismos lógicos para representar el conocimiento sobre un universo de discurso que permita la inferencia de nuevos datos a partir de los datos ya publicados. En este contexto se han reimpulsado el uso de técnicas de razonamiento y de sistemas basados en conocimiento, como pueden ser las ontologías y los sistemas basados en reglas lógicas (Ontobroker, XSB, etc.) o de producción (Drools, JRules, etc.). La aplicación de estos sistemas está ampliamente asentada en la resolución de diversos problemas (diagnóstico, planificación, reglas de negocio, etc.) pero siempre utilizando un enfoque para la representación del conocimiento y de los datos, en muchos casos específico y no estandarizado, pero que con la aplicación de los principios de la Web Semántica se ve favorecido por un nuevo contexto de estandarización.

Por lo tanto, teniendo en cuenta el contexto estratégico que supone la contratación pública para el mercado común y la sostenibilidad económica y considerando iniciativas como la Web Semántica, que animan y potencian la reutilización de datos, información y modelos de conocimiento compartido, queda patente que la inclusión de la tecnología semántica en el contexto de la administración electrónica y en concreto en los procesos de contratación pública electrónica, mejora e impulsa este dominio con un nuevo enfoque. El estudio e implantación de la semántica conlleva una mejora cualitativa en cuanto a la provisión de un modelo estandarizado, cooperativo y con nuevas capacidades, para mejorar e instrumentar este entorno desde un punto de vista tecnológico y realizar la visión estratégica que desde la política se propugna. Si se aplica el principio de evolución de las especies de *Charles Darwin*, se diría que la semántica aplicada a la contratación pública supone una evolución natural a la forma de entender este proceso administrativo, favoreciendo su supervivencia en el nuevo contexto de la *Web of Data*.

It is not the strongest of the species that survive, nor the most intelligent, but the ones most responsive to change.

1.1 Motivación

La aplicación de las tecnologías semánticas e iniciativas como *Open Data* y *Linked Data*, están de rigurosa actualidad, convirtiéndose en pieza clave para el desarrollo de nuevas aplicaciones y servicios de valor añadido, reutilizando datos provenientes de distintas fuentes: redes sociales, empresas, personas o instituciones públicas. Por otra parte, la investigación desarrollada en estas áreas se está materializando a través de múltiples publicaciones de alto valor científico, realización de conferencias internacionales, desarrollo de herramientas y transferencia tecnológica a empresas de distintos dominios, en las cuales se han de integrar fuentes de datos heterogéneas y generar servicios inteligentes contextualizados para los usuarios, así, se pueden encontrar soluciones en dominios tan dispares como salud, control de procesos industriales, administración electrónica, gestión de recursos humanos, etc.

En concreto, en el caso de la administración electrónica, uno de los procesos administrativos más influyentes por su impacto económico en la sociedad es la compra y adquisición de bienes y servicios, ya que impulsa la actividad económica de las empresas como proveedores de la Administración Pública y permite disponer nuevos servicios a los ciudadanos que, en último término, son los clientes y contribuyentes de la Administración. Es por ello, que la aplicación de las últimas tendencias de investigación y tecnológicas en este contexto se plantean como clave para mantener a la Administración Pública como paradigma de funcionamiento, organización y actualización.

1.1.1 Motivación Investigadora

La investigación en el campo de la Web Semántica se materializa a través de distintas líneas de estudio, que van desde la definición de modelos lógicos formales y su combinación con los ya existentes, hasta actividades más orientadas a la innovación para dar un nuevo enfoque a la resolución de problemas subyacentes a la informática, como es la integración de fuentes de datos heterogéneas o gestión de la información. Este amplio espectro de aplicación permite iniciar actividades de investigación en múltiples líneas y con distinto carácter: básica, aplicada, innovadora, etc. En el caso objeto de estudio, la investigación realizada se centra en aplicar los principios de la Web Semántica y *Linked Data* al dominio de la contratación pública electrónica, con el objetivo de facilitar el acceso a la información que estos documentos contienen. Por lo tanto, la actividad de investigación a realizar, centra su foco en cómo puede la semántica facilitar el acceso a la información de los anuncios de licitación pública.

En cualquier actividad de investigación a partir de una serie de hipótesis iniciales se intenta demostrar la validez de ciertos supuestos siguiendo una metodología que permita caracterizar y cuantificar la veracidad de los principios marcados. En este sentido, la aplicación de semántica a los contratos públicos provee un escenario extremadamente rico para realizar la experimentación necesaria que evalúe la mejora de la conjunción de la semántica como modelo para la gestión de la información y datos contenidos en los anuncios de licitación públicos, en contraposición con los actuales procesos tradicionales. En concreto, la investigación a realizar posee un objetivo muy específico pero de gran calado para la innovación en el ámbito de la administración pública electrónica y de los procesos de contratación pública.

Por otra parte y en cuanto a los resultados esperados, la investigación realizada debe demostrar las ventajas de la semántica en este dominio, ya probada en muchos casos para la resolución de problemas en otros contextos, para que de este forma se pueda promocionar el *know-how* tanto verticalmente, abordando otras situaciones sensibles en la contratación pública, como horizontalmente en otros procesos administrativos.

En resumen, la investigación realizada debe proveer los resultados y mecanismos adecuados para

crear concienciación, tanto en la comunidad científica, como en la industrial, de la bondad de la aplicación de la semántica al dominio de la contratación pública electrónica.

1.1.2 Motivación Técnica

Los procesos de contratación pública conllevan una serie de actividades que hacen uso de las más avanzadas técnicas para la gestión de la información y de datos, así como para el soporte a la decisión. La relevancia de la información que se maneja en estos procesos sumada a la trascendencia económica, implica que la tecnología que da soporte a todo el proceso debe asegurar la fiabilidad y corrección del mismo, impulsando y mejorando la cadena de valor tanto para la propia Administración como para los proveedores.

En general, la Administración Pública suele ser referente tecnológico por la inversión en infraestructuras que realiza. Es por ello, que la aplicación de la semántica en este dominio debe, no sólo ser capaz de proveer las mismas capacidades para el despliegue de servicios, sino mejorar la calidad de los mismos. Este reto tecnológico implica, en muchos casos, el desarrollo de herramientas que faciliten la adición de semántica aplicando un enfoque de ingeniería cuantificable. También, y desafortunadamente el uso de tecnología semántica implica, más veces de las deseadas, la depuración de la misma, por ello dentro de esta investigación la tecnología semántica juega un papel relevante tanto para probar su valía en este dominio, como para mejorar y ampliar sus capacidades.

1.1.3 Motivación Personal

La carrera en investigación propone desafíos para la resolución de problemas de diversos ámbitos aplicando diferentes técnicas. A lo largo de mi trayectoria profesional en el campo de la investigación, he tenido la oportunidad de participar en diversas actividades en este entorno, tales como proyectos, conferencias, redacción de artículos o formación, en los cuales he podido conocer y relacionarme con personas y organizaciones provenientes de distintos ámbitos así, sector privado, público y académico. Cada uno de ellos, con distintos objetivos y perfiles, han contribuido al enriquecimiento de mi propia experiencia, tanto a nivel personal, como evidentemente profesional, el cual queda materializado en la realización de esta tesis, fiel reflejo de la contribución de la misma, ya que afirmar que se trata de un trabajo personal aislado sería entender al individuo como un ente apartado del mundo, lo que implicaría negar el factor social inherente al mismo. Por ello, a nivel personal, la realización de la tesis constituye un punto de inflexión en mi carrera de investigación, supone no sólo un compendio de mi historia personal hasta el momento presente, sino también y con seguridad un elemento clave para mi devenir profesional y personal tanto a corto, como a medio y largo plazo.

Por otra parte, se plantean también dos retos: 1) investigador, que reside en el estudio, análisis, aplicación y prueba de la validez de uso de las tecnologías semánticas en el campo de la contratación pública electrónica y 2) tecnológico, al aplicar conceptos de ingeniería a la resolución de un problema, mediante una serie de métodos y herramientas cuantificables permitiendo así desarrollar tanto la motivación intrínseca (desafío de resolver un problema concreto) como extrínseca (difusión del conocimiento y solución generadas).

Finalmente, actúa como incentivo la obtención del propio título de Doctor, que supone la mayor distinción a nivel académico y formativo que se puede lograr en el actual sistema de educación, que se trasluce igualmente en gran satisfacción y orgullo no sólo a nivel personal, familiar y social sino también profesional.

1.1.4 El proyecto *10ders Information Services*

10ders Information Services [229] es un proyecto de investigación cofinanciado por el Ministerio de Industria, Turismo y Comercio dentro del plan Avanza 2 con código TSI-020100-2010-919, liderado por Gateway S.C.S. [133] (creadores de la plataforma de alertas de anuncios de licitación Euroalert.net [132]) y desarrollado en colaboración con la empresa Exis TI [236] y el equipo de investigación WESO de la Universidad de Oviedo.

El objetivo principal de este proyecto es el siguiente:

Construir una plataforma pan-europea y unificada que agregue todas las licitaciones públicas de la Unión Europea con el fin de diseñar productos y servicios de información baratos y accesibles que ayuden a las PYME a ser más competitivas en este mercado.

De la misma forma, este objetivo se desarrolla dentro de un proceso de ingeniería y tecnología cuya meta es:

Diseñar una arquitectura escalable capaz de localizar e interoperar con miles de fuentes de información públicas completamente heterogéneas para agregar la información de los cientos de miles de anuncios de licitación publicados diariamente en la Unión Europea en múltiples idiomas y culturas, con el fin de comprender y extraer conocimiento para el diseño de nuevos productos de información para PYMES.

Teniendo en cuenta este objetivo y el entorno tecnológico predefinido, los siguientes problemas que atañen a la contratación pública deben ser abordados y resueltos con la aplicación de nuevos métodos.

Dispersión de la información. Los anuncios de licitación se difunden en multitud de fuentes públicas de ámbito europeo, nacional, regional y local. Adicionalmente, las licitaciones de menor valor correspondientes a procedimientos con publicidad restringida y que suponen una larga serie de oportunidades comerciales, se publican en los perfiles del contratante de cada entidad, agencia o autoridad pública. Conseguir el diseño de un sistema eficiente de rastreo y monitorización para miles (cifra estimada) de fuentes de información, capaz de detectar las nuevas oportunidades publicadas diariamente, y que además pueda agregar nuevas fuentes a medida que sean descubiertas, constituye el primer problema tecnológico.

Mismo anuncio en más de una fuente. Dependiendo de la diferente normativa legal local, regional o nacional, el mismo anuncio de licitación se publica en más de una fuente al mismo tiempo. Como dificultad adicional, lo más habitual es que no se utilice ningún tipo de código o referencia que identifique el anuncio de forma coherente entre las diferentes fuentes. En ocasiones, entre los datos suministrados por cada una de las fuentes existirá más detalle en determinados aspectos del mismo anuncio de licitación; así por ejemplo, una fuente puede disponer de mejores datos sobre la autoridad contratante, mientras que otra puede que los suministre más completos sobre el objeto del contrato. En ese contexto, existe un gran interés en poder identificar de forma algorítmica qué datos provenientes de distintas fuentes se refieren al mismo anuncio, y al mismo tiempo ser capaces de utilizar los de mayor calidad en cada caso, no sólo para evitar tener información duplicada, sino para disponer de la calidad más elevada posible en los datos a explotar, agregando la mejor información existente en cada publicación.

Heterogeneidad de los formatos de los anuncios. Aunque los datos deben ser públicos, no existe un formato de anuncio de licitación unificado, ni tan siquiera un número limitado de posibilidades. Descargar y procesar la información de cada fuente para que sea utilizable de forma agregada, supone resolver diversos problemas complejos para interpretar la información estructurada o no, de los anuncios y extraer datos que permitan homogeneizar todas estas fuentes. Además los datos disponibles y su formato varían a lo largo del tiempo, ya que ningu-

na de las fuentes públicas ofrece una garantía de estabilidad o compatibilidad futura de los formatos utilizados para la publicación.

Almacenamiento. La cantidad de información que deberá procesar el sistema (solo el Diario Oficial de la Unión Europea-DOUE supone más de 20,000 documentos cada día) hace que sea necesario resolver la forma en que deba almacenarse para que pueda explotarse de forma eficiente. El sistema de almacenamiento diseñado, tiene el reto adicional de ser escalable para incluir nuevas fuentes de información.

Diversidad de formatos de explotación. Un almacén de información unificado con los anuncios de compras públicas ofrece múltiples posibilidades de explotación en diferentes productos, algunos de los cuales surgirán para responder a necesidades identificadas mucho después de cerrar el diseño del sistema. El último gran problema tecnológico es diseñar la arquitectura, en la que puedan convivir múltiples subconjuntos de información optimizados para requisitos de explotación diversos. En el supuesto más extremo, determinados casos de minería de datos estarán interesados en explotar aquellos datos que no tienen utilidad al margen de ese caso concreto, luego la arquitectura debe reducir la barrera del re-procesado en tiempo real de la información disponible, para extraer datos adiciones sin penalizar con ello otros servicios en explotación.

Multilingüismo y multiculturalidad. Como reto añadido se advierte que los documentos de licitación pueden publicarse en una o varias de las 23 lenguas oficiales de la Unión Europea. Adicionalmente cada país, en ocasiones utiliza otras lenguas también oficiales en sus respectivos territorios. Todos los algoritmos y sistemas que se incorporen a la arquitectura, deberán procesar, agregar e interpretar información que estará escrita en un número elevado de lenguas, que además estará expresada incorporando particularidades culturales que van desde la moneda, hasta el sistema impositivo pasando por la forma en que se codifican las direcciones postales. Suplementariamente, cuando el mismo anuncio se publica en más de un idioma, no siempre bastará con centrarse en uno de los idiomas, sino que en ocasiones habrá que analizar todos y cada uno de ellos utilizando las partes con datos de mayor calidad obtenidos entre las distintas versiones del mismo documento.

Una vez contextualizado el proyecto *10ders Information Services*, las actividades en las que participa la Universidad de Oviedo, se enclavan en la aplicación de las tecnologías semánticas e iniciativas como *Linked Data*, añadiendo una capa de conocimiento basada en semántica para proporcionar una visión de los datos que sea compatible con los principios y directrices de las actuales corrientes de *Open Data* y *Linked Data*. Con ello se pretende dar respuesta a los retos que presenta la contratación pública a nivel europeo, y como se ha reseñado en los puntos anteriores las tecnologías basadas en semántica, aportan una solución coherente y sostenible a la identificación y gestión de la información (**Dispersión de la información** y **Mismo anuncio en más de una fuente**) mediante modelos de datos y formatos compartidos y consensuados (**Heterogeneidad de los formatos de los anuncios** y **Diversidad de formatos de explotación**) en un entorno intrínsecamente para dar soporte a la internacionalización y acceso ubicuo a la información de forma contextualizada (**Multilingüismo y multiculturalidad**).

1.2 Planteamiento y Definición Inicial del Problema

Una vez que se ha repasado y presentado en las anteriores secciones una introducción a la casuística de la contratación pública y tecnología semántica, se han vislumbrado los distintos orígenes de la motivación y activado el contexto de ejecución dentro del proyecto *10ders Information Services*, es

el momento adecuado para plantear el problema a resolver y definir la hipótesis de partida objeto de estudio en esta tesis y de redacción en este documento.

Se parte de un dominio en el cual la publicación de la información de forma accesible, multiformato, multicanal, multilingüe, etc., resulta clave para facilitar las oportunidades de participación en los procesos de contratación pública y favorecer la publicidad de los anuncios de licitación mejorando la eficiencia de la propia Administración al generar un entorno competitivo. Por otra parte, la iniciativa de la Web Semántica y una parte de su realización como es el enfoque de *Linked Data*, proporciona la base para el modelado de datos e información de forma estándar y consensuada. Finalmente, el tercer componente a considerar reside en la tendencia actual de apertura de datos por parte de las instituciones públicas, gracias a la sensibilización generada por el movimiento de *Open Data*. A partir de la conjugación de estos tres puntos clave se formula la siguiente hipótesis, vertebradora de nuestra investigación.

Es posible mejorar el acceso a la información contenida en los anuncios de licitación de las distintas instituciones públicas europeas, tanto en términos cuantitativos como cualitativos, mediante métodos semánticos basados en aplicar y cumplir los principios de la iniciativa *Linked Data* y de la misma forma mantener y favorecer los principios de la corriente *Open Data*.

1.3 Objetivos Científico/Técnicos

La cooperación de las iniciativas basadas en semántica como *Linked Data*, unido al movimiento emergente de *Open Data* en el contexto de la licitación pública, presenta una serie de beneficios que permiten dar respuesta a problemas de gran calado que dificultan el despegue del mercado económico común. El estudio de la aplicación de la semántica en este contexto, junto con el despliegue tecnológico que soporte las capacidades necesarias que se han subrayado dentro del proyecto *10ders Information Services*, se materializa en los siguientes objetivos científico-técnicos:

1. Estudiar, analizar y valorar las capacidades actuales del dominio de la contratación pública.
2. Estudiar, analizar y valorar las corrientes de *Open Data* y Web Semántica, más en concreto *Linked Data*, para su aplicación en el dominio de la contratación pública.
3. Definir los métodos basados en semántica para producir, publicar, consumir y validar la información de los anuncios de licitación siguiendo las directrices de *Open Data* y *Linked Data*.
4. Definir los algoritmos y procesos para dar soporte a la aplicación de los métodos basados en semántica a la información de los anuncios de licitación.
5. Implementar y reutilizar los componentes software necesarios para dar soporte a los métodos semánticos.
6. Promover el uso de estándares y la reutilización de información y modelos de conocimiento compartido.
7. Aplicar los métodos semánticos definidos al contexto de los anuncios de licitación pública.
8. Establecer un conjunto de prueba y validación de los componentes implementados para verificar la corrección, validez y rendimiento de los métodos propuestos.
9. Difundir, formar y transferir la tecnología y conocimiento generado tanto a las comunidades científicas como industriales.

Cada uno de estos objetivos se lleva a cabo en el conjunto de tesis y se presentan a lo largo de los capítulos que se desarrollan en este documento, según se desglosa en la Tabla 1.1.

Objetivo(s)	**Capítulo(s)**
1	Estudio, análisis y valoración de la panorámica de la contratación pública en el Capítulo 2.
2	Estudio, análisis y valoración de la panorámica de la Web Semántica y *Linked Data* en el Capítulo 3.
3 y 4	Definición de los métodos semánticos necesarios para dar soporte a la contratación pública mediante *Linked Data* en los Capítulos 4 y 5.
5 y 6	Implementación y reutilización de los componentes software necesarios para aplicar los puntos anteriores en el Capítulo 6.
7 y 8	Experimentación y validación de los métodos semánticos aplicados al contexto de la licitación pública en los Capítulos 7 y 8.
9	Difusión y publicaciones realizadas tras el proceso de estudio, investigación y desarrollo en el Apéndice B.

Tabla 1.1: Alineación de Objetivos y Capítulos.

1.4 Metodología de la Investigación

La realización de la tesis debe plantearse como un proceso de trabajo, evolución y realimentación, en el cual a partir de una hipótesis inicial, unos objetivos concretos y un plan de trabajo se puedan alcanzar los objetivos planteados de una forma sistemática. Con el objetivo de facilitar este proceso y asegurar la calidad del trabajo, es conveniente plantear el siguiente esquema para identificar correctamente el "Problema a Resolver".

- ¿Por qué no se ha resuelto todavía?
- ¿Cómo se puede resolver?, ¿Resolución total, parcial, etc.?
- ¿Qué enfoques se pueden utilizar para el problema propuesto?, ¿Se puede mejorar alguno?
- ¿Cuáles son los factores críticos de éxito?, ¿Se pueden minimizar riesgos?, ¿Cuál sería el caso peor?
- ¿Es un problema actual?
- ¿Cuál sería el impacto de su resolución?
- ¿Cuáles son los escenarios relevantes para su aplicación posterior?

Teniendo presentes las posibles respuestas a estas preguntas cabe definir claramente el proyecto de tesis describiendo pormenorizadamente el problema o problemas a resolver, por ello es conveniente fijar un enfoque general, repasar el trabajo realizado y proponer una planificación para la consecución en tiempo y esfuerzo de los objetivos prefijados. En este sentido, existen métodos [162] para dar soporte a la formulación, evaluación y validación del problema.

Constructs. Proveen lenguaje, terminología y espacio, en la que un problema es definido y explicado.

Models. Cubren los hechos y conceptos en un dominio de interés o tipo de situaciones. Usan *Constructs* como lenguaje de descripción del espacio del problema.

Methods. Describen procesos y guían a los usuarios sobre cómo identificar soluciones aplicables a un problema investigación. Abarcan desde la parte teórica de algoritmos matemáticos a su realización. Desarrollo de una metodología, técnica o algoritmo.

Implementations. Implementan los anteriores con el objeto de demostrar su viabilidad. Implementaciones de referencia.

Como resultado de su aplicación se puede, por ejemplo, generar una nueva área de investigación, desarrollar un *framework* de trabajo, resolver problemas estancados, explorar minuciosamente un área, refutar los resultados actuales, crear nuevas metodologías o algoritmos más genéricos.

Esta metodología de trabajo sólo pretende ser una pequeña guía para centrar los esfuerzos y economizar el tiempo de desarrollo de la tesis. Las tareas aquí mencionadas serán responsabilidades tanto del Director de la tesis como del Doctorando. El objetivo final, será obtener un resultado relevante para la comunidad científica que pueda ser proyectado posteriormente en un entorno industrial.

1.4.1 Plan de Trabajo

El plan de trabajo que da soporte a la realización de la investigación, se divide en las siguientes fases:

1. Trabajo de Investigación, resumen y evaluación de la investigación realizada identificando problemas, métodos y propuesta de tema objeto de estudio. El resultado de este trabajo se materializa en:
   - Entrega del Trabajo de Investigación en el Departamento de Informática y obtención del Diploma de Estudios Avanzados.
   - Realización, al menos, de dos publicaciones en congresos. En este sentido, se debe adecuar el lugar de publicación al avance del trabajo realizado. Se fijan como lugares objetivos: *PhD Symposium*, *workshops* o conferencias internacionales de un nivel asequible. El objetivo principal será obtener realimentación de la línea de investigación marcada.
   - El tiempo para cubrir esta primera etapa será de doce meses.
2. Elaboración y evaluación de la solución propuesta. En esta etapa, se deben resolver las cuestiones principales de la tesis. Encontrar posibles soluciones, formulación, evaluación y validación. El resultado de este trabajo se materializa en:
   - Documento con la definición formal, evaluación y validación.
   - Realización, al menos, de dos publicaciones en congresos o revistas (JCR) de carácter y prestigio internacional. Retroalimentación de expertos: colaboración en listas de correos, proyectos, etc.
   - El tiempo para cubrir esta segunda etapa será de doce meses.
3. Escritura del documento de tesis. Finalización, redacción de toda la documentación necesaria. El resultado de este trabajo se materializa en:
   - Documento final de la tesis para la obtención del título de Doctor.

- Realización, al menos, de dos publicaciones en congresos de carácter y prestigio internacional y dos revistas (JCR).
- Realimentación de expertos: colaboración en listas de correos, proyectos, etc.
- El tiempo para cubrir esta última etapa será de doce meses.

Este proceso de elaboración de la tesis es ideal, y ha sido cubierto teniendo en cuenta las fluctuaciones habituales procedentes del ejercicio de otras tareas. Para minimizar los retrasos, se ha establecido un método de monitorización consistente en revisiones periódicas entre el Director y el Doctorando.

1.4.2 Metodología de Trabajo

En esta sección se presenta un resumen de la metodología de trabajo que se ha llevado a cabo durante la duración del trabajo en el proceso para la consecución de la tesis. Para alcanzar los objetivos científico-técnicos planteados, se han definido tareas de "investigación y prototipado" y otras de carácter transversal que cubren las necesidades generales de la realización de la tesis, como pueden ser la ejecución de pruebas o la redacción de documentación y publicaciones.

El estudio realizado en cuanto a contratación pública, como proceso administrativo y dentro del marco de la administración electrónica, así como la panorámica de uso de la Web Semántica, *Linked Data* y *Open Data* afectan de forma crucial en las tareas propias de "investigación y prototipado", ya que se produce un proceso de realimentación que permite acotar las referencias a consultar y focalizar el proceso de investigación. Esto conlleva que aquellas tareas propias de la ingeniería, sirvan como retroalimentación para el estudio realizado en las distintas líneas de investigación. De esta forma, la validación y las pruebas sirven para definir y realimentar los activos experimentales de los componentes de software.

La investigación sigue un modelo diferente a los clásicos modelos de ciclo de vida de ingeniería de software, más orientado al problema científico a resolver. Con ello, se realiza un enfoque de "investigación y experimentación concurrente", en el que las tareas de investigación y experimentación se realizan en paralelo obteniendo así un ciclo de vida en "espiral", que se adapta mejor al dinamismo de la investigación. De esta forma se sientan las bases para un modelo de estudio y desarrollo con realimentación continua de la experimentación, se orienta a la consecución de investigación y prototipos iterativos.

En contraste con las aproximaciones donde los proveedores de software tradicionales evalúan el resultado del proyecto durante la fase final del mismo, se realiza una aproximación de ingeniería concurrente dónde la investigación, el prototipado de software, la experimentación y la validación se realizan en paralelo. De esta manera, las actividades de investigación reciben realimentación de su relevancia en una fase temprana de la ejecución.

Finalmente, el *software* realizado durante la tesis se desarrolla mediante un modelo basado en comunidad de software libre, siendo el que mejor se adapta para obtener una colaboración y coordinación eficiente de las diferentes partes involucradas en la ejecución del proceso: validación por parte del Director, publicación de resultados, etc. Este modelo es fácil de gestionar y permite incorporar nuevos investigadores y contribuciones externas sin esfuerzo. Para dar soporte a este modelo se usa la infraestructura proporcionada por Google Code, adicionalmente, se emplean otras herramientas de software libre disponibles como:

- Herramientas de desarrollo software: Eclipse IDE y *plugins*, lenguajes de programación: Java, Python, etc.
- Herramientas de razonamiento y distribución: razonadores como Pellet, motores de inferencia como Drools, Prolog, Apache Hadoop, Apache Mahout, etc.

- Herramientas de integración de software: ANT, Maven, Continuum, etc.
- Herramientas de edición de documentos: Open-Office y LaTeX.
- Otras herramientas: Google Refine, Protégé, SNORQL, Pubby, etc.

1.5 El Camino hacia la Tesis

En el verano del año 2002 tras superar el primer curso en la Escuela de Ingeniería Técnica Informática de Oviedo y adquirir los primeros conocimientos de programación, algorítmica y matemáticas aplicados a la informática decidí, *motu proprio*, introducirme en el mundo de Internet. Recuerdo, el portal de *Geocities* en el cual se proporcionaba una herramienta tipo *Microsoft Front Page*, para la creación de sencillas páginas web, incluyendo formularios, encuestas, etc., en aquel momento, aunque no entendía el código HTML ni la forma de publicar páginas web, cree mi primera página, que incluía con una encuesta con una votación sobre los participantes de un programa de talentos, que en ese momento se emitía en televisión. Una vez concluida la edición de la página, cuando pulse el botón "Publicar" rápidamente avisé a mi hermana para mostrarle, orgulloso, que mis conocimientos estaban en aumento, sin embargo su reacción no fue ni mucho menos la esperada, con un gesto que denotaba incredulidad y asombro comentó: "¿qué es esto?". Esta anécdota que suele recordar cada cierto tiempo, o más bien cuando le muestro algo nuevo que he hecho, me recuerda la evolución que he experimentado en el transcurso de estos últimos años en el campo de la informática y de la tecnología web en concreto.

En el camino hacia la Tesis, existe otro punto de inflexión que tiene lugar en el invierno del año 2004 tras finalizar y obtener el título de Ingeniero Técnico en Informática de Sistemas, cuando surgió la posibilidad de formar parte de un nuevo equipo multidisciplinar, en el cual se iba a iniciar actividad de investigación en el área de Web Semántica. Este equipo comandado por expertos de la Universidad de Oviedo como José Emilio Labra, Enrique Del Teso (Filología), Guillermo Lorenzo (Filología) o Roger Bosch (Filosofía) y ubicado en la Fundación CTIC bajo la supervisión de Antonio Campos, comenzó su andadura con la tecnología semántica para desarrollar un sistema de búsqueda para el Boletín Oficial del Principado de Asturias (BOPA). Con aquel equipo, con el cual compartí casi 5 años y formado por Diego Berrueta, Luis Polo, Emilio Rubiera, Manuel Cañon e Iván Frade se desarrolló un buscador "semántico" [34] para las disposiciones que se publicaban en el BOPA. Se basada en la utilización en una técnica de "Concept Expansion", utilizando ontologías que después se transformaba a una consulta en *Apache Lucene*, los resultados no pudieron ser mejores y fue un gran éxito a nivel estratégico, de tal forma que actualmente todavía está en producción (la parte de búsqueda sintáctica) en la Administración del Principado de Asturias y, siendo seleccionado como un caso de uso [81] por el W3C.

Al presente valorando y analizando el desarrollo de aquel proyecto en el que se abordaron varios problemas, tales como:*screen scrapping* de los datos de las disposiciones, reconciliación de entidades, transformación de clasificaciones de productos o *mapeo* de palabras a conceptos, que atañen hoy al mundo de la Web Semántica y de la iniciativa *Linked Data* y también casi accidentalmente, se proporcionaron servicios como suscripción RSS a un consulta, navegación gráfica sobre las ontologías de dominio, búsqueda contextualizada por varios criterios y múltiples formatos para visualizar la información de las disposiciones (HTML y RDF). Hoy en día este proyecto estaría de actualidad en cualquier Administración, probablemente bajo un nombre tipo: "Portal multicanal de *Open Data* y *Linked Data* para el acceso a las disposiciones del boletín x", la cuestión reside en que de una manera intuitiva, se desarrolló investigación y tecnología a la cual en nuestros días se le designa con un nombre concreto y que además se puede comercializar. Este antecedente simplemente trata de revelar la evolución que en este área se ha producido.

Durante los siguientes años, seguí mi andadura en CTIC [80] participando en proyectos de investigación relacionados siempre con semántica y sobretodo con servicios web, en los cuales me gustaría destacar la actividad desarrollada en el proyecto PRAVIA [92] y su sucesor a nivel nacional el proyecto PRIMA [228], en ambos casos se trataba de utilizar servicios web semánticos para facilitar la interacción con los proveedores de servicios, en el sector asegurador, si bien se consiguieron realizar distintos prototipos, la gran lección aprendida fue la relativa a la notable complejidad de esta iniciativa, que por entonces, y probablemente de momento no se puede resolver de forma sencilla. Este trabajo me facilitó la posibilidad de participar en otros proyectos como EzWeb [227], en la parte de recomendación y descubrimiento de servicios, MyMobileWeb [226], adición de semántica para gestionar el contexto en aplicaciones móviles, SAITA [93], creación de una pasarela REST-SOAP, etc. Es en esta época cuando surge mi trabajo de investigación, cuyo tema principal era la interoperabilidad e integración en arquitecturas orientadas a servicios usando tecnologías semánticas. Finalmente, en mi última etapa en CTIC participé en el proyecto europeo ONTORULE [95], en el cual realicé actividades relacionadas con la gestión del proyecto y desde el punto de vista de la investigación, con la esfera de las reglas de negocio y ontologías, específicamente en tareas relacionadas con RIF [48](*Rule Interchange Format*).

1.5.1 Elaboración de la Tesis en el contexto del equipo de investigación WESO

Desde noviembre del año 2010 y como miembro del equipo de investigación de Web Semántica Oviedo dirigido por José Emilio Labra Gayo se disfruta de un entorno dinámico, colaborativo y flexible para la realización de trabajos de investigación e innovación mediante la realización de tareas y actividades como participación en proyectos, viajes, redacción de artículos, asistencia a conferencias, etc.

La metodología de trabajo utilizada en este equipo favorece la interacción entre las personas y el impulso de la investigación para conseguir resultados tanto desde punto de vista cualitativo como cuantitativo.

Mi llegada al equipo WESO ha supuesto tanto un nuevo reto como un impulso a mi carrera profesional e investigadora en la cual he podido perfeccionar y aumentar mi conocimiento aplicándolo a nuevos desafíos y trabajos en los cuales he tenido la suerte de participar con excelentes profesionales que de una forma u otra también forman parte de este trabajo.

1.6 Convención utilizada en este documento

A lo largo de este documento se utilizan distintos tipos de letra y puntuación para señalar partes destacadas del mismo y centrar la lectura del lector. La convención seguida en todo el documento es la siguiente:

Letra en *cursiva*. Se utiliza para destacar las palabras procedentes de otro idioma, por ejemplo en inglés *Linked Data*, o para señalar tecnicismos, por ejemplo *cloud computing*.

Letra en *cursiva* en un cuadro. Se utiliza para referenciar partes de texto que han sido tomadas de otros documentos en las que es importante mantener exactamente la definición que realizan. Por ejemplo, algunos textos tomados de directivas de la Unión Europea.

Letra en negrita. Se utiliza para destacar palabras o párrafos de alto valor para el correcto entendimiento del documento. Por ejemplo, en la formulación de la hipótesis.

Texto "entrecomillado". Se utiliza para destacar palabras o párrafos en los cuales se cita alguna frase u oración de un autor conocido o bien para simular la primera persona en la expresión escrita.

Figuras. Sirven para designar: diagramas, imágenes o trozos de ejemplo de código fuente. Por ejemplo, se utilizan para formalizar de forma gráfica el diseño del sistema MOLDEAS.

Tablas. Se utilizan tablas para indicar una serie de características y su evaluación. Por ejemplo, durante la evaluación, se usa este formato de presentación para condensar gran cantidad de información facilitando así su lectura.

El principal objetivo del empleo de esta notación, es amenizar y hacer más atractiva la lectura y comprensión del texto.

Igualmente, se utiliza indistintamente el concepto de licitaciones, anuncio de licitación o contrato público, para designar al documento que contiene la información y los datos de un contrato público. También se utilizan tecnicismos provenientes del idioma inglés, cuyo uso está perfectamente asentado e integrado en la prosa habitual en lo concerniente a temas como Web Semántica, *Linked Data*, *Open Data*, *Linking Open Data* o Web de Datos. Entre ellos, se puede destacar URI, IRI, recurso RDF, etc., para evitar la cacofonía y aunque la traducción al español implica el uso del género masculino, en muchos casos se opta por adaptar la expresión escrita a su uso habitual. De esta manera, se habla de "la URI de un recurso RDF", en lugar de "el URI de un recurso RDF" y en general, para ciertas traducciones se emplea el género femenino en lugar del masculino con el objetivo de facilitar la lectura y adaptar la forma escrita al modo habitual de expresión.

Capítulo 2

Contratación Pública y e-Procurement

> Cuando creíamos que teníamos
> todas las respuestas, de pronto,
> cambiaron todas las preguntas.
>
> *Citas Célebres*
> MARIO BENEDETTI

2.1 Introducción

La contratación pública es uno de los sectores clave de la economía europea, alcanzando un 17 % [12] del Producto Interior Bruto (PIB). La modernización y apertura de las fronteras de los Estados Miembros se plantea crucial para la mejora de la competitividad y la creación de nuevas oportunidades de negocio. En general, el proceso de contratación pública se considera largo y tedioso, con un gran consumo de recursos tanto para la propia Administración, como para los potenciales empresarios que acudan a los procesos de licitación. Por ello, desde la Unión Europea se elaboró el "Plan de Acción sobre Contratación Electrónica" [76] en el año 2004 con varios objetivos, entre los que destaca la creación de un marco jurídico común para el proceso de contratación pública y la elaboración de distintas acciones, proyectos y actividades que pudieran impulsar tecnológicamente este servicio de la Administración y mejorar la participación de las distintas empresas. Como resultado, en la Figura 2.1, se muestran los objetivos, resultados e impacto esperado que forman parte de los informes [77, 104] elaborados en 2010 por Siemens y la propia Comisión.

De igual forma, en este informe [104] se establecen una serie de métricas, ver Figura 2.2, resultados e impacto, que deben servir como guía para las distintas acciones realizadas a nivel europeo y que deben manifestarse en los distintos Estados Miembros.

A través del Plan de Acción de 2004 y de las posteriores evaluaciones continuas que se vienen realizando en la Unión Europea, queda patente la importancia que desde las instituciones públicas se aplica al proceso de contratación pública.

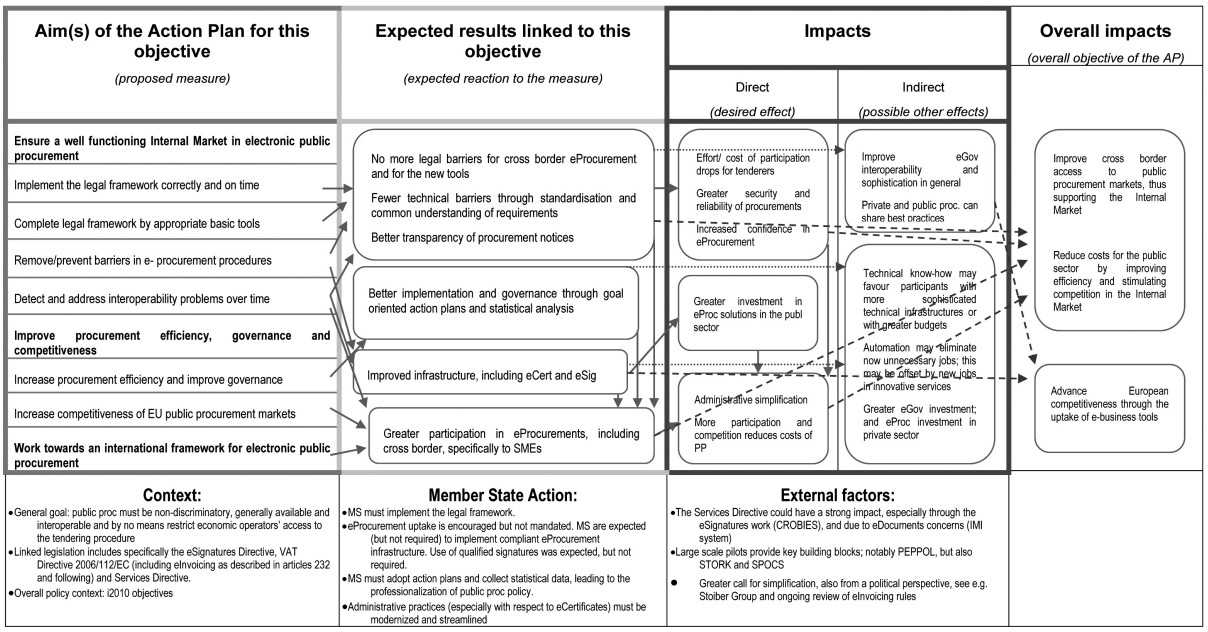


Figura 2.1: Estrategia en e-Procurement de la Unión Europea elaborado por Siemens.

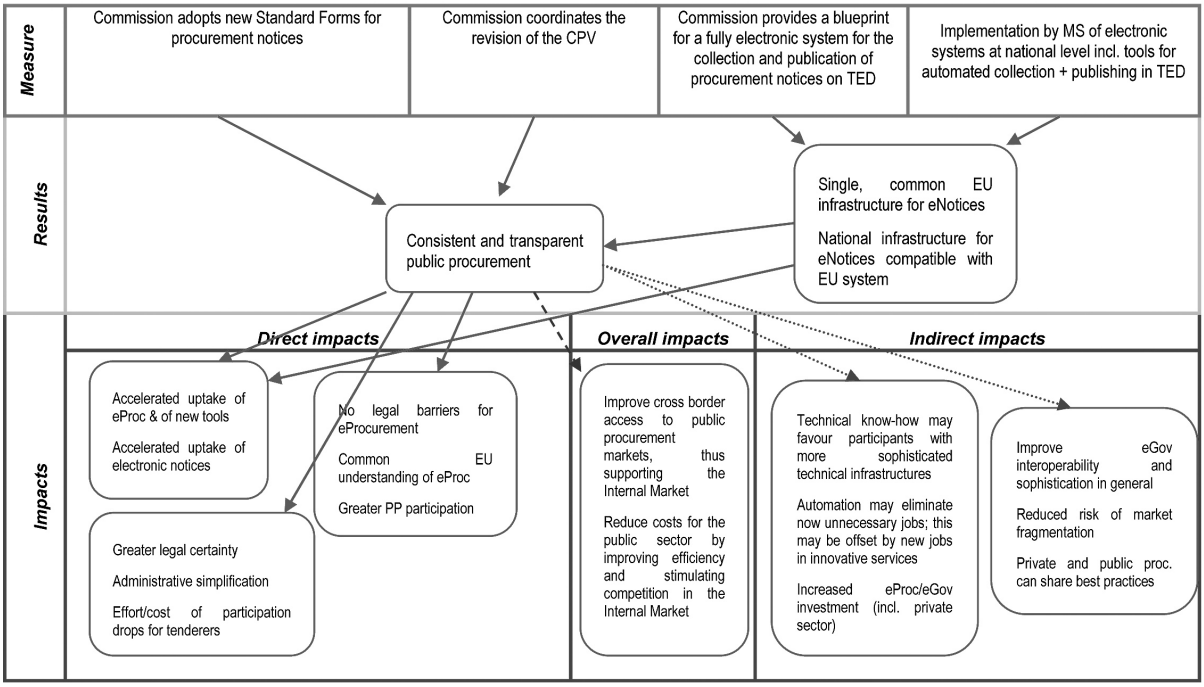


Figura 2.2: Métricas en e-Procurement de la Unión Europea elaborado por Siemens.

2.1.1 Terminología

Con el objetivo de favorecer la lectura del documento y disponer de una clara definición de cada uno de los términos habitualmente utilizados en los procesos de contratación pública, se dispone a continuación de una serie de definiciones (se incluyen las más importantes para el ámbito de este documento):

Adjudicatario (*Successful tenderer*). Empresa ganadora de un contrato público. Anteriormente en España el proceso de adjudicación se dividía en provisional y definitivo. La diferenciación del procesos en dos partes, se establecía con el objetivo de comprobar si el adjudicatario cumplía las garantías necesarias. Una vez que el proceso de selección previo se había superado, ya se establecía la formalización del contrato con un determinado órgano y siguiendo la lista de candidatos que la Mesa de Contratación había valorado.

Adjudicación (*Awarding*). Pasos que el órgano de contratación y los colectivos de soporte, comités de expertos o técnicos, siguen para decidir qué licitadores ofertan las mejores opciones para ganar el contrato. Habitualmente, se utilizan diversas técnicas para cada paso (técnica Delphi, entropía, etc.) aunque una de las máximas es obtener la mejor relación oferta/precio, bajo las condiciones más óptimas.

Agente externo. Actor dentro del proceso de licitación que interviene en alguna parte del proceso. Por ejemplo, en el ámbito español la Agencia Estatal de Administración Tributaria (AEAT) o la Tesorería General de la Seguridad Social (TGSS), disponen de información susceptible de ser solicitada por los sistemas de contratación electrónica de los órganos de contratación.

Anuncio previo (*Prior Information Notice*-PIN). Mensaje publicado por el órgano de contratación para informar sobre la intención de contratar en un determinado período. Suelen contener información sobre contratos de obras, suministros y servicios para los siguientes doce meses.

Anuncio de adjudicación (*Award notice*). Mensaje publicado por el órgano de contratación mediante los boletines oficiales como el "Diario Oficial de la Unión Europea" (DOUE) y el "Boletín Oficial del Estado" (BOE), en los cuales se informa de los datos pertinentes de una licitación y su adjudicatario. Suele corresponder a la información que se notifica como resolución al proceso.

Anuncio de licitación (*Contract notice*). Mensaje publicado por el órgano de contratación mediante el cual se anuncia una nueva licitación, describiendo los términos legales, administrativos y técnicos mínimos para que los empresarios expresen su interés de participación. La publicación se realiza de forma oficial a través del DOUE, BOE, etc., y también forma parte del perfil del comprador. En el ámbito español, existen otros tipos como el "Anuncio Simplificado de Licitación", utilizado en la fase de contratación específica de un Sistema Dinámico de Adquisición.

Contrato (*Contract*). Documento de acuerdo entre un órgano de contratación y un licitador, que establece las condiciones de contratación de acuerdo a los pliegos publicados por el primero y la oferta realizada por el segundo. Incluye firmas de las partes contratantes y toda la información pertinente. Este documento representa la culminación del acto de licitación.

Empresario (*Economic operator*). Es el "operador económico" que designa tanto a un "contratista" como a un "proveedor" o "prestador de servicios". Otros términos que se utilizan para especificar esta figura son "Licitador" o "Adjudicatario".

Espacio de trabajo de la licitación. Espacio digital en el cual se registra toda la documentación y aspectos relativos al proceso administrativo. Es el expediente electrónico del proceso, que puede ser implementado mediante carpetas virtuales, registros, repositorios, etc.

Licitar (*Call for tenders* o *tendering*). En su acepción más amplia: acto por el cual el Estado concede contratos para la ejecución de servicios, suministros y obras de interés público; la adjudicación definitiva de dichos contratos puede ir precedida de un concurso o subasta en la que varias personas, físicas o jurídicas, presentan sus cotizaciones o los precios que cobrarían por la ejecución del contrato. También se utiliza para denominar al subconjunto de tareas ejecutadas por el empresario, en interacción con el Órgano de contratación, destinadas a presentar una Oferta en respuesta a un Anuncio de Licitación o una invitación a oferta.

Oferta (*tender*). Documento que un empresario envía a un órgano de contratación para licitar.

Órgano de contratación (*Contracting authority*). Organizaciones, departamentos, secciones o colegiados con la capacidad legal y estatutaria para celebrar contratos en nombre de las entidades públicas.

Perfil del contratante (*Buyer profile*). Página web en la cual se publicita toda la información referente a los contratos de un determinado órgano de contratación.

Plataforma de contratación (*e-Procurement System*). Plataforma electrónica a través de la cual se gestiona el proceso administrativo de la contratación electrónica.

Otros. A continuación se disponen de algunos términos habitualmente utilizados en la bibliografía relativa a contratación pública y que se consideran de interés: acuerdo de exclusión, acuerdo marco (*framework agreement*), agencia de garantía (*guarrantee agency*), agencia registradora (*registration agency*), artefacto (*artifact*), agente externo (*external agent*), claúsulas administrativas, comité de expertos (*technical committee*), diálogo competitivo (*competitive dialogue*), expediente (*procurement File*), garantía definitiva (*tender guarantee*), invitación, mesa de contratación, oferta indicativa, oferta temeraria, pliegos (*tender information*), prescripciones técnicas, procedimiento (*procedure*): abierto, negociado, restringido, Registro Oficial de Licitadores y de Empresas Clasificadas (ROLEC en el ámbito español), sistema dinámico de adquisición (*dynamic purchasing system*), etc.

2.2 Contratación Pública

En general los participantes en un proceso de contratación pública, no difieren de la ejecución de cualquier contrato habitual salvo que uno de los actores, como parte contratante sea la Administración Pública, mientras que por la parte contratada pueden distinguirse tanto a personas físicas como jurídicas con capacidad de obrar y con solvencia acreditada tanto a nivel económico-financiero como técnico.

La Administración se estructura en diferentes órganos, compuestos por distintas unidades organizativas o agentes responsables de llevar a cabo las actividades de los procesos administrativos y entre ellos el proceso de contratación pública. Para estudiar los actores participantes en el proceso de contratación, se pueden seguir diferentes criterios pero por sencillez se suele utilizar una distribución funcional según las actividades a desarrollar por cada uno de ellos. A continuación, se detallan algunos de estos agentes con referencia a la situación actual en el ámbito de la Administración General del Estado en España.

Contratación. Son las entidades con capacidad para obrar un contrato. Normalmente, se agrupan bajo el término "Órganos de Contratación" y dependen de la propia estructura y organización de la Administración. Dependiendo de distintas variables como el tipo de contrato, el importe del mismo o el lugar de celebración, los órganos cambiarán de acuerdo a su potestad. Como ejemplo y siguiendo la estructura y organización de la Administración General del Estado, se

establecen distintas unidades organizativas que tendrán capacidad de actuar como Órganos de Contratación. En este caso y siguiendo la Ley 6/1997 [88], de 14 de abril de Organización y Funcionamiento de la Administración General del Estado, la cual en su exposición de motivos, refiere a uno de los preceptos fundamentales que la Constitución Española de 1978 dedica en su artículo 103 a la actividad de la Administración, así, objetividad, generalidad, eficacia, jerarquía, descentralización, desconcentración y coordinación se constituyen como principios básicos de funcionamiento de la Administración.

Esta misma ley, en su artículo 3, establece que la organización y actuación de la Administración General del Estado se realizará con respeto al principio de legalidad y de acuerdo a los siguientes principios (entre otros):

- De organización:
  - Jerarquía.
  - Descentralización funcional.
  - Desconcentración funcional y territorial
  - Economía, suficiencia y adecuación estricta de los medios a los fines institucionales.
  - Coordinación.
- De funcionamiento:
  - Eficacia en el cumplimiento de los objetivos fijados.
  - Eficiencia en la asignación y utilización de los recursos públicos.
  - Programación y desarrollo de objetivos y control de la gestión y de los resultados.
  - Responsabilidad por la gestión pública.
  - Racionalización y agilidad de los procedimientos administrativos y de las actividades de gestión.
  - Servicio efectivo a los ciudadanos.
  - Objetividad y transparencia de la actuación administrativa.

También y como ejemplo, en el Capítulo II de dicha ley se establece la "Organización Administrativa", tanto de la Administración General del Estado como de sus Organismos públicos. La Administración General del Estado responde a una organización funcional en Departamentos ministeriales y de gestión territorial integrada en Delegaciones de Gobierno en las propias Comunidades Autónomas. La organización de la Administración Central distingue entre órganos superiores (S) y órganos directivos (D):

- Ministros. (S)
- Secretarios de Estado. (S)
- Subsecretarios y Secretarios generales. (D)
- Secretarios generales técnicos y Directores generales. (D)
- Subdirectores generales. (D)

Esta estructuración administrativa resulta idónea para exponer una distribución modelo que dependiendo de cada caso puede variar: Comisión Europea, Universidad, etc.

Gestión. Contribuyen al proceso de contratación pública, dando el apoyo a las actividades de ejecución de las tareas económico-administrativas para tramitar el proceso. En general, suelen formar parte del propio Órgano de Contratación, aunque también pueden tener carácter general, como por ejemplo la Dirección General del Tesoro y Política Financiera siendo los encargados en última instancia de realizar el pago de los importes de los contratos.

Control. Se encargan de asegurar la legalidad de los procesos de contratación. En el ámbito de la Administración General del Estado, se podría destacar la Intervención General de la Administración del Estado y el Tribunal de Cuentas.

Asesoría. Especialistas colaboradores de la contratación para diferentes aspectos: jurídicos, técnicos, económicos u otros que pudieran requerirse. Siguiendo con el ejemplo anterior, y en general: la Junta Consultiva de Contratación Administrativa o los Servicios Jurídicos del Estado, por otra parte, en particular podríamos citar: las Oficinas Técnicas de Supervisión de Proyectos o la Comisión Interministerial de Adquisición de Bienes y Servicios Informáticos (CIABSI).

Publicidad. Canales de comunicación para hacer públicas las convocatorias y los procedimientos de contratación. En el ámbito de la Administración General del Estado podemos citar el BOE o la Plataforma de Contratación del Estado. En el ámbito europeo hay que mencionar el DOUE, *Tenders Electronic Daily* (TED) o el Sistema de Información para la contratación pública europea (SIMAP).

Otros. Funciones de carácter transversal, como las obligaciones tributarias o de seguridad social. Continuando con el ejemplo citado, la Agencia Tributaria o la Tesorería de la Seguridad Social serían agentes participantes que dan soporte a las actividades derivadas de la contratación pública.

2.2.1 Ejemplo: la Administración General del Estado en España

En esta sección se recoge parte de la casuística implicada en un proceso de contratación pública con la Administración General del Estado, con el objetivo de mostrar las actividades que se han de desarrollar y desvelando algunas de las interacciones que surgen entre los distintos agentes con las consiguientes necesidades de interoperabilidad e integración.

Actualmente cada Comunidad Autónoma tiene su propia plataforma de contratación y perfil del contratante, ver Tabla 2.1, en la cual se pueden realizar las operaciones del proceso de contratación pública electrónica. En muchos de los casos están enlazadas con la Plataforma de Contratación del Estado [102] y siguen sus estándares como "Componentes y Documentos Interoperables para la Contratación Electrónica" (CODICE) para la gestión de la información y documentación generada a lo largo del proceso. Evidentemente y dependiendo de la capacidad de cada órgano de contratación podrán desplegar su propia plataforma, la administración en España sigue un modelo descentralizado y coordinado desde la Administración General del Estado, o bien utilizar la infraestructura común.

Comunidad Autónoma	**Plataforma de Contratación**
Andalucía	http://www.juntadeandalucia.es/contratacion
Aragón	http://portal.aragon.es/portal/page/portal/CPUBLICA#
Castilla y León	http://www.jcyl.es/web/jcyl/Portada/es/Plantilla100/1246947706077/_/_/_
Castilla La Mancha	http://pagina.jccm.es/contratacion/
Cantabria	https://aplicaciones5.cantabria.es/PerfilContratante
Canarias	http://www.gobcan.es/perfildelcontratante/contenido
Cataluña	https://contractaciopublica.gencat.cat

Comunidad Autónoma	Plataforma de Contratación
Región de Murcia	http://www.carm.es/web/pagina?IDCONTENIDO=6427&IDTIPO=100
Comunidad de Madrid	http://www.madrid.org/cs/Satellite?cid=1203334374251&language=es&pagename=PortalContratacion/Page/PCON_contenidoFinal
Comunidad Valenciana	https://www.contratacion.gva.es/WebContrataP
Cdad. Foral de Navarra	http://www.navarra.es/home_es/Servicios/Portal+contrataciones/
Extremadura	https://contratacion.juntaextremadura.net
Galicia	http://www.contratosdegalicia.es/
Illes Balears	http://www.plataformadecontractacio.caib.es/
La Rioja	http://www.larioja.org/npRioja/default/defaultpage.jsp?idtab=26275
Principado de Asturias	https://sede.asturias.es
País Vasco	http://www.euskadi.net/r33-2288/es/contenidos/informacion/perfil_anuncios/

Tabla 2.1: Plataformas de Contratación en las Comunidades Autónomas.

En estas plataformas las operaciones a realizar, los idiomas disponibles, etc., son variados, si además se amplia el rango de órganos de contratación a Universidades, Ayuntamientos, Parroquias Rurales, Mancomunidades, etc., y ahora se proyecta a nivel europeo es factible advertir la diversificación y la replicación de esfuerzos que se están efectuando en este ámbito. Por ello, iniciativas como la Plataforma de Contratación del Estado o TED ayudan a unificar y facilitar el acceso a las distintas fases del proceso de contratación pública.

2.3 Contratación Pública Electrónica: *e-Procurement*

El término *e-Procurement* [251] se utiliza para designar el uso de medios y tecnologías electrónicos para llevar a cabo las operaciones y comunicaciones relacionadas con la contratación pública, y en general para todos aquellos procesos [4] en los que participen proveedores y consumidores de un cierto servicio o producto. En el caso particular de las Administraciones Públicas se utiliza esta terminología para definir las operaciones llevadas a cabo para adquirir bienes o servicios mediante distintos métodos: contratación directa, centrales de contratación, contratos de urgencia y emergencia, etc.

La importancia del proceso de contratación pública electrónica, va más allá de la simple transición del papel ya que puede aportar mejoras relevantes en eficiencia de las adquisiciones, gestión global del proceso de contratación pública y en general, al funcionamiento de los mercados de las administraciones.

La tendencia actual para acelerar la implantación de procesos de contratación, se alinea con las directrices de impulso de la administración electrónica como camino para la transformación del funcionamiento de la Administración Pública y la prestación de servicios. En este sentido, la Agenda

Digital de la Comisión Europea está trabajando en una guía o "libro blanco", para fijar las medidas que aseguren el despliegue de una infraestructura integradora de la contratación pública para atender las necesidades de los distintos países. Como primer paso para la consecución de este gran objetivo, se ha elaborado una declaración de intenciones a través de un "libro verde" en el cual se pone de manifiesto el ambicioso objetivo de uso de las TICs en este dominio de la Administración Pública.

El nacimiento de esta propuesta está fechada en el año 2005 cuando los Ministros de la Unión Europea manifestaron su deseo de conseguir "en el año 2010, que como mínimo el 50 % de toda la contratación pública que rebase el umbral de contratación pública de la Unión Europea se lleve a cabo por medios electrónicos". Teniendo presente esta intención se modificó la legislación pertinente y se puso en práctica el "Plan de Acción sobre Contratación Electrónica" [76] en el año 2004. Aunque desde un punto de vista teórico se han sentado las bases para la contratación pública electrónica, la realidad es que el grado de implantación efectiva es muy inferior al esperado, ya que la complejidad subyacente, ver Figura 2.3, abarca distintos niveles: técnico, logístico y administrativo. Siguiendo la evaluación realizada por la Comisión, se señala que al menos el 5 % del presupuesto total destinado a la contratación en los Estados Miembros que iniciaron esta línea de acción se adjudica por canales electrónicos.

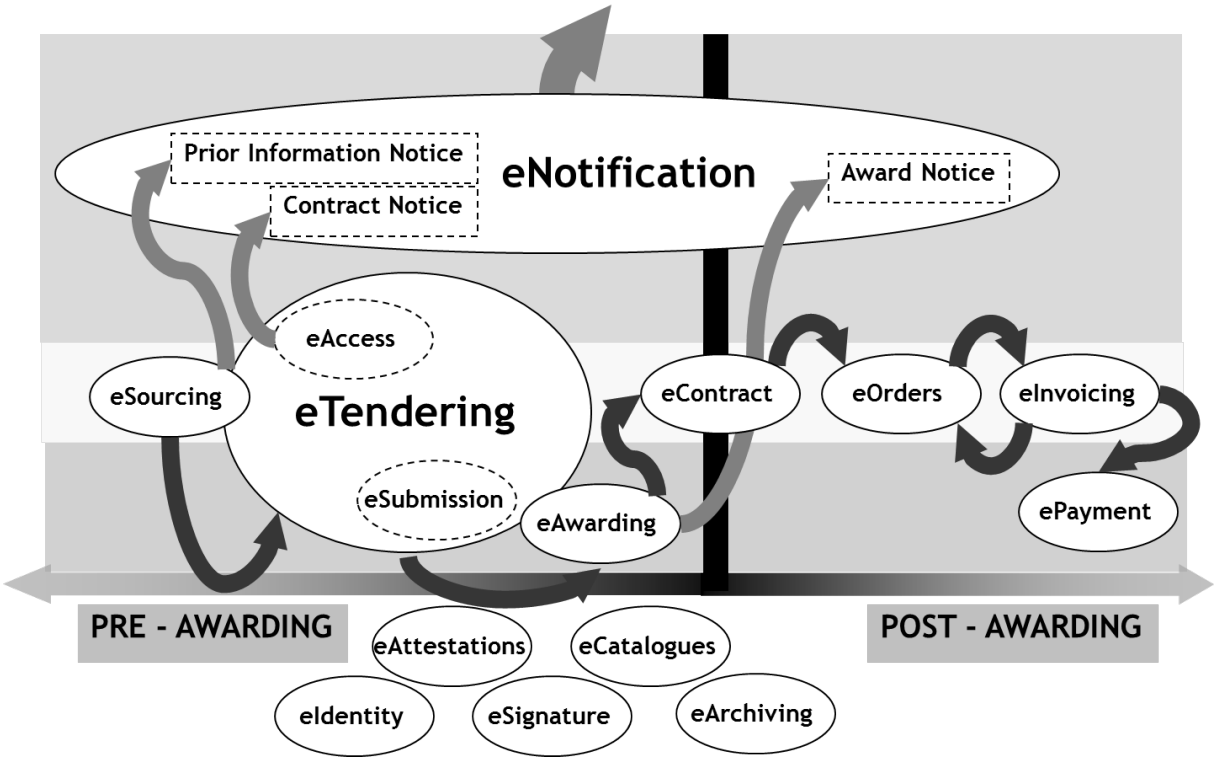


Figura 2.3: Diagrama de Complejidad y Fases de e-Procurement por la Unión Europea.

Ahora bien, el crecimiento de este porcentaje depende en buena medida de que la contratación electrónica sea implantada a todos los niveles administrativos (local, regional y estatal), para ello se debe contar con la tecnología necesaria, en muchos casos ya disponible, así como de plataformas de contratación que se puedan integrar a nivel operativo para que el tráfico a través de las mismas genere la masa crítica suficiente como para alcanzar los requisitos fijados por la Comisión para lograr la innovación en este proceso. Además de incrementar el uso de medios electrónicos en la contratación pública, surge la oportunidad de armonizar este proceso administrativo de forma transfronteriza de tal forma que se impulse la participación de las empresas en procesos de contratación pública

más allá de su región de origen, fomentando así la competitividad y mejorando tanto los productos como los servicios a contratar y por extensión potenciando tanto las posibilidades de las entidades dispuestas a licitar, como los servicios prestados por la propia administración.

2.3.1 Definición de Contratación Electrónica

Siguiendo la definición realizada por la Unión Europea [110], se puede definir contratación electrónica como:

La contratación electrónica es un término general utilizado para designar la sustitución de los procedimientos basados en soporte de papel por el tratamiento y la comunicación mediante TIC a lo largo de toda la cadena de contratación pública. Supone la introducción de procedimientos electrónicos para sustentar las distintas fases del proceso de contratación, es decir, publicación de los anuncios de licitación, suministro del pliego de condiciones, presentación de ofertas, evaluación, adjudicación, pedido, facturación y pago.

Los procedimientos vinculados a la facturación y el pago (posteriores a la adjudicación) no son específicos de la contratación; por tanto, es posible aplicar en el ámbito de la contratación pública electrónica soluciones desarrolladas para un mercado más amplio (de empresa a empresa) 6 . No obstante, en algunas fases (notificación, presentación de propuestas, valoración y pedido) se requieren soluciones específicas. Las fases de presentación, valoración y pedido son las más arduas ya que exigen la aplicación de una serie de protocolos y normas consensuados para organizar el intercambio de documentos complejos, así como la interacción entre el comprador público y los proveedores.

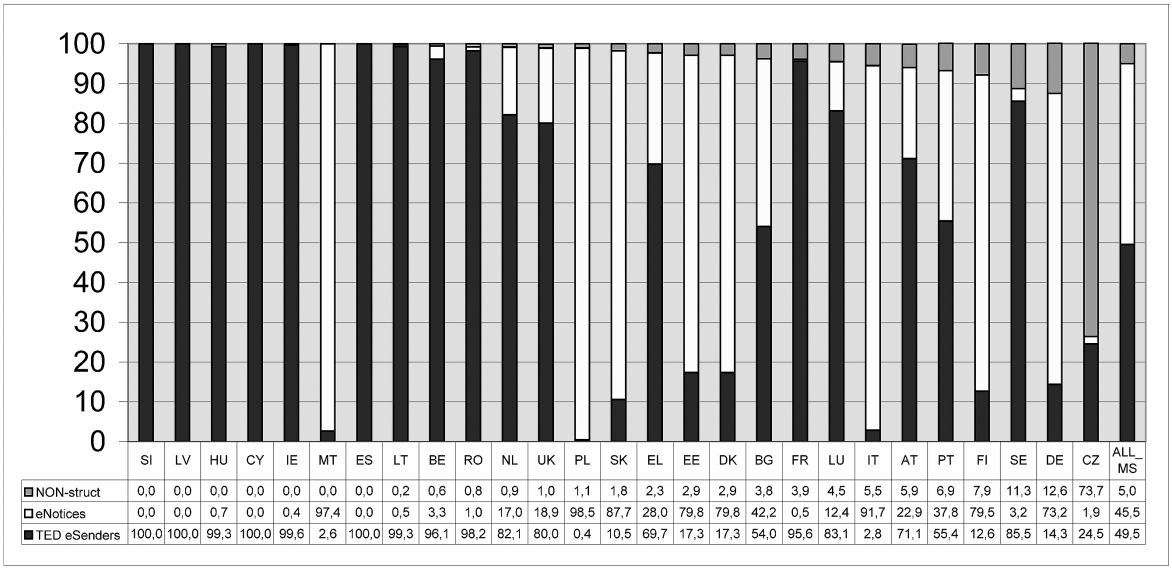
No obstante, consultando la actividad propuesta por la Unión Europea para la contratación electrónica se han fijado algunos aspectos que deberán todavía coexistir de forma no automatizada. Es el caso de la documentación de procesos de contratación complejos como pueden ser los planos y diseños de obras, ya que es necesario disponer de los originales de forma impresa y normalizados para poder ser consultados por los propios expertos de la administración y realizar las validaciones pertinentes.

Ahora bien para dar cabida a la contratación pública electrónica se ha necesitado un esfuerzo por parte de las diferentes Administraciones para la creación de portales en los cuales contener los anuncios de licitación y facilitar el acceso a la descripción de los mismos, y más en particular a los pliegos de condiciones técnicas y económicas. De esta forma, se da un paso hacia delante al disponer de un servicio público electrónico de "principio a fin". Este tipo de plataformas han sido denominadas como "plataformas de contratación electrónica", como ejemplo podemos citar la "Plataforma de Contratación del Estado" [102] entre otras. También y de forma obligatoria, aquellos entes que quieran hacer uso de la contratación pública electrónica deben publicar su "Perfil de Contratante" [96]. En el caso de España, existen otras características que impulsarán la contratación pública electrónica como la factura electrónica [94].

Por otra parte, el proceso de contratación pública y en concreto el que concierne a la contratación pública electrónica cuenta con distintas fases [104] con diferentes objetivos y que se deben abordar en un contexto común, definiendo para cada uno de ellos sus características.

eNotification. Esta fase cubre la publicación de anuncios de licitación. Se trata de un proceso unilateral que conlleva la comunicación entre órgano de contratación y el sistema de publicación. Es el primer paso para facilitar el acceso a la información contenida en los anuncios de licitación. El principal desafío de esta fase es realizar la publicación de la información de la forma más estandarizada posible. Actualmente, se utiliza XML como formato de publicación y mediante herramientas como eSenders es posible enviar la información a través de distintos canales como fax o correo electrónico. Los órganos de contratación se encargan de esta tarea buscando agilizar el proceso de publicación y asegurando que no existan errores. A nivel europeo, en el

año 2009, alrededor del 89 % de los anuncios de licitación utilizan un medio electrónico para ser publicados. En la Figura 2.4, se puede ver el uso de los medios para la publicación de anuncios realizados por los distintos Estados Miembros de forma agregada. Por ejemplo y comparando algunos países, en España sería un 88,8 % (ver Figura 2.5) y en Alemania un 82,9 % (ver Figura 2.6).



	SI	LV	HU	CY	IE	MT	ES	LT	BE	RO	NL	UK	PL	SK	EL	EE	DK	BG	FR	LU	IT	AT	PT	FI	SE	DE	CZ	ALL_MS
NON-struct	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,2	0,6	0,8	0,9	1,0	1,1	1,8	2,3	2,9	2,9	3,8	3,9	4,5	5,5	5,9	6,9	7,9	11,3	12,6	73,7	5,0
eNotices	0,0	0,0	0,7	0,0	0,4	97,4	0,0	0,5	3,3	1,0	17,0	18,9	98,5	87,7	28,0	79,8	79,8	42,2	0,5	12,4	91,7	22,9	37,8	79,5	3,2	73,2	1,9	45,5
TED eSenders	100,0	100,0	99,3	100,0	99,6	2,6	100,0	99,3	96,1	98,2	82,1	80,0	0,4	10,5	69,7	17,3	17,3	54,0	95,6	83,1	2,8	71,1	55,4	12,6	85,5	14,3	24,5	49,5

Figura 2.4: Porcentaje y Tipo de Publicación de Anuncios de Licitación en TED.

Media used by national awarding authorities to send notices to the Publications Office

Percentage of Notices / Month

media	Last 12 months	2010-10	2010-11	2010-12	2011-01	2011-02	2011-03	2011-04	2011-05	2011-06	2011-07	2011-09	2011-10	2011-11
TED eSenders / eNotices	**88,8**	89,6	91,2	85,8	88,0	89,8	89,2	82,0	91,2	87,2	89,4	92,7	100,0	100,0
paper / fax / email	11,2	10,4	8,8	14,2	12,0	10,2	10,8	18,0	8,8	12,8	10,6	7,3	0,0	0,0
	100,0	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**

Count of Notice_id / Publ_Month

media	Last 12 months	2010-10	2010-11	2010-12	2011-01	2011-02	2011-03	2011-04	2011-05	2011-06	2011-07	2011-09	2011-10	2011-11
TED eSenders	2.476	148	164	239	125	164	185	198	188	239	285	240	215	234
eNotices	15.567	1607	1751	1835	1539	1533	1732	1482	1515	1481	1432	1267	0	0
EMAIL	703	50	46	44	49	55	57	200	53	106	59	34	0	0
FAX	1.215	130	116	138	144	113	138	135	95	132	131	73	0	0
PAPER	369	24	23	160	34	24	38	35	16	15	13	11	0	0
Grand Total	**20330**	**1959**	**2100**	**2416**	**1891**	**1889**	**2150**	**2050**	**1867**	**1973**	**1920**	**1625**	**215**	**234**

Percentage of Notice_id / Publ_Month

media	Last 12 months	2010-10	2010-11	2010-12	2011-01	2011-02	2011-03	2011-04	2011-05	2011-06	2011-07	2011-09	2011-10	2011-11
TED eSenders	12,2	7,6	7,8	9,9	6,6	8,7	8,6	9,7	10,1	12,1	14,8	14,8	100,0	100,0
eNotices	76,6	82,0	83,4	76,0	81,4	81,2	80,6	72,3	81,1	75,1	74,6	78,0	0,0	0,0
EMAIL	3,5	2,6	2,2	1,8	2,6	2,9	2,7	9,8	2,8	5,4	3,1	2,1	0,0	0,0
FAX	6,0	6,6	5,5	5,7	7,6	6,0	6,4	6,6	5,1	6,7	6,8	4,5	0,0	0,0
PAPER	1,8	1,2	1,1	6,6	1,8	1,3	1,8	1,7	0,9	0,8	0,7	0,7	0,0	0,0
	100,0	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**

Figura 2.5: Porcentaje y Tipo de Publicación de Anuncios de Licitación en TED de España.

eAccess. Esta fase cubre la habilidad de obtener copias de cualquier documento del proceso de contratación electrónica, incluyendo los anuncios de licitación. Los medios habituales para posibilitar el acceso son la publicación a través de distintos sitios web o mediante correo electrónico. Se trata igualmente de un proceso unilateral, cuya relevancia radica en que las buenas prácticas y la estandarización deben estar presentes como puntos clave para facilitar el acceso a la información de los anuncios de licitación.

Media used by national awarding authorities to send notices to the Publications Office

Percentage of Notices / Month

media	Last 12 months	2010-10	2010-11	2010-12	2011-01	2011-02	2011-03	2011-04	2011-05	2011-06	2011-07	2011-09	2011-10	2011-11
TED eSenders / eNotices	**82,9**	82,2	82,0	81,9	84,0	81,2	84,2	79,7	83,2	83,0	83,0	82,1	82,9	87,5
paper / fax / email	17,1	17,8	18,0	18,1	16,0	18,8	15,8	20,3	16,8	17,0	17,0	17,9	17,1	12,5
	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0

Count of Notice_id / Publ_Month

media	Last 12 months	2010-10	2010-11	2010-12	2011-01	2011-02	2011-03	2011-04	2011-05	2011-06	2011-07	2011-09	2011-10	2011-11
TED eSenders	5.701	390	338	507	327	460	455	488	468	523	537	554	482	562
eNotices	33.230	2779	2555	2986	2542	2625	3193	2822	2726	2834	2923	2636	2509	2879
EMAIL	5.366	446	428	498	341	454	456	568	450	491	470	456	433	321
FAX	1.374	138	108	155	100	147	92	166	118	85	137	106	75	85
PAPER	1.298	104	98	120	106	113	135	109	78	110	103	132	107	87
Grand Total	**46969**	**3857**	**3527**	**4266**	**3416**	**3799**	**4331**	**4153**	**3840**	**4043**	**4170**	**3884**	**3606**	**3934**

Percentage of Notice_id / Publ_Month

media	Last 12 months	2010-10	2010-11	2010-12	2011-01	2011-02	2011-03	2011-04	2011-05	2011-06	2011-07	2011-09	2011-10	2011-11
TED eSenders	12,1	10,1	9,6	11,9	9,6	12,1	10,5	11,8	12,2	12,9	12,9	14,3	13,4	14,3
eNotices	70,7	72,1	72,4	70,0	74,4	69,1	73,7	68,0	71,0	70,1	70,1	67,9	69,6	73,2
EMAIL	11,4	11,6	12,1	11,7	10,0	12,0	10,5	13,7	11,7	12,1	11,3	11,7	12,0	8,2
FAX	2,9	3,6	3,1	3,6	2,9	3,9	2,1	4,0	3,1	2,1	3,3	2,7	2,1	2,2
PAPER	2,8	2,7	2,8	2,8	3,1	3,0	3,1	2,6	2,0	2,7	2,5	3,4	3,0	2,2
	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0

Figura 2.6: Porcentaje y Tipo de Publicación de Anuncios de Licitación en TED de Alemania.

eSubmission. Esta fase trata el envío de las ofertas realizadas por un agente económico al órgano de contratación. Se trata de un proceso bilateral en el cual se establece una comunicación continua entre estos dos agentes. Unos de los principales desafíos radica en el uso de sistemas de autenticación y autorización, que permitan la interoperabilidad pan-europea entre los distintos sistemas de autenticación de los Estados miembros.

eEvaluation/eAwarding. Estas dos fases desarrollan el proceso de toma de decisión sobre qué oferta de las presentadas se ajusta mejor al pliego de condiciones técnicas. El desafío clave en estas fases radica en el uso de técnicas totalmente automatizadas.

eOrdering. Esta fase conlleva la tarea de automatizar el proceso de publicar la información en línea, especialmente mediante el uso de catálogos (*eCatalogues*) que serán utilizados posteriormente por las herramientas que dan soporte a las distintas fases.

eInvoicing. Esta fase se encarga de la gestión de las facturas a través de medios electrónicos. Su principal desafío radica en el uso de medios electrónicos para el tradicional proceso de facturación requiriendo la comunicación e interoperabilidad entre los agentes participantes.

ePayment. Esta fase está relacionada con la gestión de los pagos y las transferencias de dinero entre los agentes participantes. Se trata por tanto de transacciones B2B, en las cuales la seguridad es un factor clave. Por todo ello, se ha definido el *Payment Services Directive* como resultado del *Single Euro Payments Area*, con el objetivo de eliminar posibles barreras en los sistemas electrónicos de pago.

Atendiendo a estas definiciones y según la exposición realizada en este documento y trabajo de investigación los resultados están orientados a mejorar tres fases de las identificadas por la Comisión: *eNotification*, *eAccess* y *eOrdering*. No obstante, tratándose de la aplicación de tecnología semántica con las características intrínsecas que conlleva, también se podrían aplicar a otras fases con el objetivo de impulsar la interoperabilidad entre los agentes implicados en el proceso de contratación electrónica.

2.4 Marco legal en Contratación Publica

El proceso de contratación pública electrónica se ha definido como clave en la Unión Europea, por ello se han redactado diferentes reglamentos que deben ser transpuestos a la legislación de cada país para unificar los criterios y establecer un marco común en todo el proceso. De esta manera, se establece una base legislativa que da soporte, para que la contratación pública electrónica en la Unión Europea sea transfronteriza. Es importante destacar que la legislación sobre contratos públicos se ve sometida a continuos cambios que permitan adaptar los procesos a las necesidades de la propia Administración Pública y de los licitadores, de esta manera las distintas actividades de publicación, adjudicación, caracterización de los tipos de contrato etc., se actualizan con el objetivo de adaptarse a la casuística que va surgiendo y cambiando a lo largo del tiempo. A continuación, se dispone de una descripción sintética de la lista de los Reglamentos, Leyes, Reales Decretos, etc., más destacados que se han formalizado, ordenados por ámbito cronológicamente:

- Reglamento (CE) Nº 1177/2009 [108] de la Comisión, de 30 de noviembre de 2009 por el que se modifican las Directivas 2004/17/CE, 2004/18/CE y 2009/81/CE del Parlamento Europeo y del Consejo en lo que concierne a sus umbrales de aplicación en materia de procedimientos de adjudicación de contratos.
- Reglamento (CE) Nº 1150/2009 [107] de la Comisión, de 10 de noviembre de 2009 por el que se modifica el Reglamento (CE) no 1564/2005 en lo que respecta a los formularios normalizados para la publicación de anuncios en el marco de los procedimientos de adjudicación de contratos públicos con arreglo a las Directivas 89/665/CEE y 92/13/CEE.
- Reglamento (CE) No 213/2008 [67] de la Comisión, de 28 de noviembre de 2007 que modifica el Reglamento (CE) no 2195/2002 del Parlamento Europeo y del Consejo, por el que se aprueba el Vocabulario común de contratos públicos (CPV), y las Directivas 2004/17/CE y 2004/18/CE del Parlamento Europeo y del Consejo sobre los procedimientos de los contratos públicos, en lo referente a la revisión del CPV.
- Directiva 2004/18/CE del Parlamento Europeo y del Consejo de 31 de marzo de 2004.
- Real Decreto Legislativo 3/2011 [221], de 14 de noviembre, por el que se aprueba el texto refundido de la Ley de Contratos del Sector Público.
- Ley 24/2011 [187], de 1 de agosto, de contratos del sector público en los ámbitos de la defensa y de la seguridad.
- Dictamen 31/11 de la Junta Consultiva de Contratación Administrativa sobre la obligatoriedad de integrar el perfil de contratante de los organismos y entidades del sector público estatal en la Plataforma de Contratación del Estado así como de las informaciones a publicar en la misma.
- Real Decreto 300/2011 [222], de 4 de marzo por el que se modifica el Real Decreto 817/2009, de 8 de mayo, por el que se desarrolla parcialmente la Ley 30/2007, de 30 de octubre, de contratos del sector público y se habilita al titular del Ministerio de Economía y Hacienda para modificar sus anexos.
- Ley 2/2011 [224], de 4 de marzo, de Economía Sostenible.
- Anteproyecto de Ley de Contratos del Sector Público en el ámbito de la Defensa y la Seguridad.
- Ley 30/2007 [188] incorporando las modificaciones de la Ley 34/2010 [189], y otras normas posteriores a su publicación.

- Acuerdo de la Junta Consultiva de Contratación Administrativa en relación con los supuestos de derecho transitorio que pueden derivar de la entrada en vigor de la Ley 34/2010, de 5 de agosto .
- Ley 34/2010 [189], de 5 de agosto, de modificación de las Leyes 30/2007, de 30 de octubre, de Contratos del Sector Público, 31/2007, de 30 de octubre, sobre procedimientos de contratación en los sectores del agua, la energía, los transportes y los servicios postales, y 29/1998, de 13 de julio, reguladora de la Jurisdicción Contencioso-Administrativa para adaptación a la normativa comunitaria de las dos primeras.
- Orden EHA/1490/2010 [219], de 28 de mayo.
- Resolución [225] de 3 de marzo de 2010, de la Dirección General del Patrimonio del Estado, por la que se publica la Recomendación de la Junta Consultiva de Contratación Administrativa sobre el envío de anuncios a la Comisión Europea.
- Orden EHA/3497/2009 [223], de 23 de diciembre por la que se hacen públicos los límites de los distintos tipos de contratos a efectos de la contratación administrativa a partir del 1 de enero de 2010.
- Real Decreto 817/2009 [220], de 8 de mayo, por el que se desarrolla parcialmente la Ley 30/2007, de 30 de octubre, de Contratos del Sector Público.
- Instrumentos para la aplicación de la Ley 30/2007 de 30 de Octubre de Contratos del Sector Público.
- Orden EHA/1220/2008 [218] de 30 de abril, por la que se aprueban las instrucciones para operar en la Plataforma de Contratación del Estado.
- ...

En conclusión, se puede observar como las directivas europeas se transponen a los Estados Miembros, en este caso España, con el objetivo de unificar en la medida de lo posible la gestión de la contratación pública mediante medios electrónicos y así poder facilitar el acceso a la misma a las distintas empresas y personas de todos los Estados Miembros en un mercado abierto común.

2.5 Necesidades de *e-Procurement*

Desde la Comisión Europea existe el compromiso firme de afianzar y desarrollar la contratación pública electrónica [69, 76, 110] con el objetivo de facilitar el acceso a los procesos administrativos a todas las empresas interesadas en el mercado común. Con ello se pretende conseguir una serie de ventajas que enlazan con las iniciativas de *Open Government* (OG), ver Sección 3.2.2:

- Reducción de costes tanto temporales como materiales. Habitualmente un proceso de contratación pública implica un gran despliegue de recursos humanos que deben dar soporte a las distintas actividades implicadas en el proceso administrativo y en las comunicaciones con los participantes. Esta situación se puede agilizar y optimizar para hacer un uso eficiente de los recursos propios de la administración y facilitar la interacción con las entidades implicadas.
- Incremento de la transparencia y accesibilidad. La automatización de los procesos de contratación pública centralizando los flujos de comunicación para concurrir a un concurso de licitación mejora la transparencia de la Administración Pública, suministrando información en un punto común y favoreciendo la accesibilidad a la misma, estableciendo el lugar en el cual encontrar la

información. En general, el coste de búsqueda de oportunidades se reduce, encontrándose ante un entorno más abierto, favoreciendo la divulgación a través de la red global de Internet. Esto trae consigo un incremento de la competencia ya que un mayor número de empresas pueden ser conscientes de las oportunidades de concurrir a los procesos, impulsando la apertura del mercado y obteniendo ofertas más competitivas.

- Mejora de la eficacia de la gestión administrativa. Dependiendo del tipo de contrato, por ejemplo las centrales de contratación, los procedimientos pueden ser centralizados y de esta manera ser racionalizados con una eficiente gestión de recursos.

- Incremento de la integración entre mercados de la Unión Europea. El proceso tradicional de contratación está basado en el uso de papel como modelo de intercambio de información, con los consiguientes problemas asociados: multilingüismo, localización, etc., que impiden que las empresas puedan concursar más allá de su lugar de ubicación. El uso de la contratación publica electrónica supera estas barreras tradicionales proveyendo un canal de comunicación común, en el cual cualquier entidad dispone de las mismas oportunidades independientemente de su país de procedencia. Con ello se consigue una participación transfronteriza de las distintas empresas.

En general, estas necesidades y ventajas asociadas al uso de las tecnologías de la información aplicadas al proceso de contratación pública deben contribuir a lograr una mayor eficacia en el uso de los recursos de la Administración Pública y de las entidades participantes en las distintas actividades. El contexto económico actual invita a reducir costes corrientes, entre los que se encuentra el exceso de burocracia presente tanto en el ámbito de los recursos humanos como materiales. Teniendo en cuenta el volumen de contratos públicos en el ámbito europeo queda justificado que el impulso de la administración electrónica en este campo debe ser prioritario. Por otro lado, la implantación de estos nuevos procesos, el coste del cambio, exige fuertes inversiones tanto técnicas como de formación y difusión para concienciar a los distintos agentes participantes del proceso de las necesidades y ventajas de este nuevo enfoque. Según informes [110] de la Unión Europea, el coste a escala nacional y regional de la implantación de portales electrónicos de contratación se situaría entre $0,5$ y 5 millones de euros, el coste de mantenimiento de los mismos se estima [69] entre miles y varios millones de euros, dependiendo de la escala del sistema y la sofisticación del mismo.

2.6 *e-Procurement* en la Unión Europea

El despliegue de un sistema de contratación pública a nivel europeo requiere una fuerte inversión que debe recaer a nivel nacional y regional, ya que las necesidades particulares y capacidades de cada entidad pública quedan de manifiesto en estos ámbitos. Por otra parte, la legislación de la Unión Europea en este campo permite a las entidades adjudicadores seleccionar el método y canal de comunicación más apropiado de acuerdo a su funcionamiento (electrónico o tradicional) si se ha sobrepasado un cierto umbral de contratación, es por ello que a nivel europeo la estrategia se sitúa en proveer el marco legal necesario para liberar el mercado de la contratación pública mediante una organización descentralizada y descoordinada con los siguientes objetivos:

- Permitir a las entidades adjudicadoras seleccionar el medio y canal para llevar a cabo el proceso de contratación pública.

- Asegurar que el proceso se desarrolla bajo las directrices de la Unión Europea y de acuerdo a los tratados que fijan los umbrales de contratación y sus condiciones.

- Fomentar el desarrollo de soluciones tecnológicas convergentes que permitan la viabilidad electrónica del proceso dando lugar a una serie de buenas prácticas que sean aplicables a otros servicios.
- Facilitar la participación de las distintos agentes en un ámbito global a nivel europeo eliminando en la medida de lo posible los costes de la contratación transfronteriza. Para ello las soluciones técnicas deben asegurar la eliminación de barreras intrínsecas a la diversidad europea, por ejemplo el idioma.

Enmarcando estos objetivos en una dimensión superior, la Unión Europea debe ejercer la función de coordinación e impulso general del sector de la contratación pública, para armonizar un sector altamente divergente pero de un gran calado para la economía de toda Europa. Hasta el momento las medidas tomadas por la Comisión se centran en lo ya comentado sobre el aseguramiento del marco legislativo, no obstante, se han promovido varias acciones como: 1) la modificación de ciertas Directivas de contratación pública que permitieran la introducción de técnicas electrónicas y sistemas automáticos para la adquisición de bienes y servicios; 2) plan de acción para que "...cualquier empresa europea que disponga de un ordenador y una conexión a Internet pueda participar en una adquisición pública llevada a cabo con medios electrónicos." y 3) cofinanciación de investigación para el desarrollo de tecnología que permitiese la contratación pública transfronteriza como las iniciativas PEPPOL [73] y la herramienta e-CERTIS [72]. Como ejemplo de la actividad de la Unión Europea en este área y del Plan de Acción, el uso de *Tenders Electronic Daily* (TED) en el año 2009 superó el 90 % para las licitaciones que superaban un cierto umbral, ver Tabla 2.2, generando 300 *billion* de euros en licitación. TED supone un gran avance para el campo de la contratación pública electrónica proporcionando formularios estándar [74] (19) para las distintas fases y herramientas que permiten la gestión de la documentación e información como *eNotices* [71]. En resumen, las acciones identificadas en el Plan de Acción de 2004 siguen vigentes hoy en día y el contexto de actuación sigue siendo clave para el impulso de la contratación pública electrónica y la participación transfronteriza en la contratación en línea.

Tipo de Contrato	**Umbral**
Public works	5,000,000 €
Service contracts	200,000 €
Supplies contracts	200,000 €
Supplies in the sectors of water, energy and transport	400,000 €
Supplies in the telecommunications sector	750,000 €
Contracts falling under the GATT agreements	130,000 €

Tabla 2.2: Tipos de Contratos y Umbrales para publicar en TED.

2.7 Desafíos europeos en *e-Procurement*

Las ventajas y necesidades revisadas anteriormente implican una serie de retos a distintos niveles que suponen una barrera de entrada a la adopción integral de las soluciones basadas en tecnologías de la información para los procesos de contratación pública. Por ello, se fijan distintos puntos de acción que deben ser abordados para conseguir la implantación exitosa de la administración electrónica en el campo de los contratos públicos:

- Uno de los principales problemas en relación a la implantación de un nuevo proceso consiste en la concienciación al cambio tanto a las entidades adjudicatarias como a los posibles provee-

dores. La inercia en la Administración Pública habitualmente conlleva una lenta adopción de los nuevos trámites, debido a cuestiones de reorganización interna y desconocimiento de las nuevas ventajas. Por otra parte, la reacción de los proveedores cuando interactúan con la Administración, en una situación de transición, suele venir marcada por cierto escepticismo debido a la falta de confianza en las nuevas tecnologías, la pérdida de la interacción personal, etc. Especialmente en este tipo de operaciones de centralización, las PYME suelen inquietarse por el temor a caer en un estado de obsolescencia en su relación con la Administración, es por ello que la capacidad de concienciación por parte de la Administración Pública al entorno afectado, se plantea como un reto crítico.

- Homogeneización de la normativa aplicada a la contratación pública electrónica. Como se ha comentado en anteriores secciones la Unión Europea a través de distintas directivas marca el paso que han de seguir los distintos Estados Miembros para transponer la normativa a sus entornos nacionales y regionales. Sin embargo, la aparición de numerosas plataformas y casos particulares de contratación conllevan una complicación superior a los proveedores candidatos a concursar. Por ejemplo en el caso de la fase de presentación de las ofertas, el aprendizaje necesario para el uso de estos nuevos portales opera como una importante barrera de entrada a los procesos. En este punto, la Unión Europea actúa con el objetivo de normalizar al máximo los procesos transfronterizos con la intención de unificar criterios y obtener sistemas interoperables que no confundan a sus usuarios. Algunas de estas líneas transversales se centran en la autenticación de usuarios, intercambio de documentación, facturas, catálogos, etc., de forma estándar.
- Enredo de requisitos técnicos que impiden la adopción ágil de los nuevos procesos. El típico caso se ve reflejado en los sistemas de autenticación e identificación que van desde la simplicidad de un par, usuario y contraseña, hasta sistemas más complejos basados en firma o certificados digitales.
- Adopción paulatina. Las capacidades y prioridades de cada Estado Miembro varían dependiendo de distintos parámetros y aunque el objetivo final es la implantación completa e integrada en cada uno de los países, cada uno de ellos avanza con su propio ritmo. No obstante, cabe resaltar que el desafío reside en que paulatinamente todos los instrumentos necesarios para la contratación transfronteriza se hagan realidad mediante este esfuerzo común.

La Unión Europea, como órgano integrador y vertebrador de los distintos Estados Miembros, tiene la obligación de aunar y coordinar los esfuerzos nacionales y regionales para la adopción definitiva de la contratación pública electrónica dando respuesta tanto a las necesidades como a los retos que esta iniciativa conlleve.

2.8 Iniciativas y proyectos en *e-Procurement*

Dentro del Plan de Acción realizado en 2004 para el impulso de la contratación pública electrónica, se habían seleccionado diferentes líneas de actuación que se han materializado en actividades a lo largo de estos años algunas de las cuales ya se han dado por finalizadas mientras que otras siguen su curso. Atendiendo al mapa de actividades [111] realizadas podemos destacar las siguientes:

1. **e-Certis**. Herramienta de información en línea y gratuita que se ha lanzado en las últimas fechas (2010) para aportar datos sobre los diferentes tipos de certificados y declaraciones que se suelen exigir en los procedimientos de contratación de los 27 Estados miembros, los países candidatos (Turquía y Croacia) y los países del EEE (Islandia, Liechtenstein y Noruega). La principal función de esta herramienta es ayudar a los agentes económicos y a las entidades adjudicadoras a

comprender qué tipo de información se solicita, así como facilitar la información aceptable. Esta acción sigue su curso, siendo el órgano responsable *The Internal Market and Services Directorate General* (DG-MARKT), centrando su aplicación en la fase de *eTendering* (*eAttestation*).

2. **Fiscalis 2013**. Esta iniciativa trata de luchar contra el fraude fiscal, mejorar las prácticas administrativas y asegurar el intercambio de información entre las Agencias Tributarias nacionales y las entidades adjudicatarias, a través de los sistemas transfronterizos de fiscalidad. Es una acción iniciada en el año 2008 y cuyo plazo de finalización está previsto para el año 2013. El órgano responsable es **Taxation and Customs Union** (DG-TAXUD) y su principal aplicación se centra en las fases de: *eInvoicing* y *ePayment*.

3. **ePRIOR**. Se trata de la implementación de servicios interoperables electrónicos a nivel europeo para la fase posterior a la adjudicación del contrato público. A través del programa *Interoperable Delivery of European eGovernment Services to public Administrations, Businesses and Citizens*(IDABC) y *eInvoicing* y *eOrdering*, iniciados en verano del año 2007 por el DG-MARKT y *Directorate-General for Informatics* (DIGIT) de la Comisión Europea, buscan contribuir a los objetivos de la iniciativa **i2010** mediante el establecimiento de un marco de políticas comunes para la sociedad de la información y medios digitales. Se ha realizado un prototipo en el cual se han incluido la metodología RUP@EC, se ha realizado un estudio sobre el uso de *eCatalogues* y se ha desarrollado un conector a la infraestructura del proyecto PEPPOL. Como resultado principal se ha obtenido el despliegue de la plataforma ePRIOR (*electronic PRocurement Invoicing and ORdering*) en octubre del año 2009. Los órganos responsables son el DG-MARKT y DIGIT cubriendo las fases o temas de *eInvoicing*, *eOrdering* y *eCatalogues*.

4. **Open e-PRIOR**. Es la versión de la iniciativa anteriormente mencionada, en la cual los formatos de fuente son abiertos, permitiendo el intercambio de información y documentos a través de la infraestructura PEPPOL. Tanto los órganos responsables como las fases cubiertas son las mismas que en el caso precedente.

5. **PEPPOL - Pan-European Public Procurement Online**. Es el gran proyecto de contratación pública electrónica transfronteriza gestionado por organismos del sector público de diferentes países y cofinanciado por la Unión Europea. Su principal objetivo es disponer de una infraestructura para facilitar el despliegue de servicios tecnológicos en un entorno homogéneo y de escala global para el desarrollo y la gestión de las operaciones implicadas en los procesos de contratación pública paneuropea. El núcleo de este proyecto lo constituye una red de transporte que permite a los socios comerciales conectar sus recursos tecnológicos para el intercambio de documentos de forma segura y fiable. También se considera su aplicación para los pedidos, la facturación electrónica, firma y validación, expedientes virtuales para la empresa, creación de catálogos electrónicos, etc. El órgano responsable es *The Information Society and Media Directorate General* (DG-INFSO) (CIP ICT/PSP programme) y las fases cubiertas por este proyecto corresponden a *eOrdering*, *eInvoicing*, *eCatalogue* y *eSignature*. Este proyecto, debido a su gran envergadura se encuentra dividido en diferentes paquetes de trabajo (8), cubriendo diferentes fases y colaborando con otras iniciativas tales como: *WP1- eSignature, WP2- Virtual Company Dossier, WP3 – eCatalogue, WP4 –eOrdering, WP5- eInvoicing, WP6-Project Management, WP7- Consensus and awareness building y WP8- Solutions architecture, design and validation*. De esta forma y a través de los subproyectos se cubren nuevas etapas de la contratación como *eNoticing, eTendering, eAwarding y eOrdering*.

6. **STORK - Secure idenTity acrOss euRope linKed**. El objetivo de esta iniciativa es establecer "...a European eID Interoperability Platform", que permita a los ciudadanos y a las empresas establecer relaciones virtuales transfronterizas a través del uso de los documentos de identificación nacionales. El órgano responsable de esta acción es el DG-INFSO y cubre las fases de *eSignature* y *eID*.

7. Iniciativas en el campo de la estandarización que se han realizado a lo largo de estos últimos años entre las que se pueden destacar:
   - *CEN BII - Workshop on 'Business Interoperability Interfaces on public procurement in Europe' Phase 2 (WS/BII 2)*.
   - *CEN eCAT - Multilingual eCataloguing and eClassification in eBusiness*.
   - *CEN WS/eBES - Workshop on 'e-business Board for European Standardization'*.
   - *CEN - Workshop on eInvoicing Phase 3 (CEN WS/INV3)*.
   - *OASIS TC/UBL - The Universal Business Language*.
   - *UN/CEFACT - United Nations Centre for Trade Facilitation and Electronic Business*. En los grupos de trabajo específicos de *International Trade and Business Group (TBG)* y en concreto, *TBG1 - Supply Chain*, *TBG6 – Architecture, Engineering and Construction* y *TBG19 - eGovernment*.
8. Participación en otros grupos de trabajo, comunidades y redes de trabajo como *eProcurement Forum*.

Finalmente y dentro de esta gran actividad de han completado proyectos de diversa índole como: *Evaluation of eProcurement uptake* (encuesta en línea) y *CROBIES - Cross-Border Interoperability of eSignatures*, participado en foros de discusión: *CEN eBIF - European e-Business Interoperability Forum* y *CEN BII - Workshop on "Business Interoperability Interfaces on public procurement in Europe" Phase 1*, redactado documentos estratégicos *"Evaluation of the 2004 Action Plan for Electronic Public Procurement"* (2010), *"Green Paper on expanding the use of eProcurement in the EU"* (2010), *"Summary of the responses to the Green Paper on expanding the use of eProcurement in the EU"* (2011), *"Reaping the benefits of electronic invoicing for Europe"*, *"e-Catalogues Gap Analysis between pre-awarding business requirements and the post-awarding implementation in e-PRIOR, Version 1.0"*, etc., mediante los cuales se ha intentado concienciar a la comunidad sobre las ventajas del impulso de la contratación pública electrónica, utilizando estándares y las nuevas tecnologías de la información. Como ya se ha comentado, siempre enclavadas en un marco legislativo referente *"New Directive on eInvoicing"* (2010) o *"Provisions relating to eProcurement introduced by the public procurement Directives 2004/17/EC13 and 2004/18/EC14"* (2004) y también otros como el *"New Common Procurement Vocabulary"* (2008).

En conclusión, se puede observar como la actividad de la Unión Europea en este campo y de sus distintos organismos, apuesta firmemente por el impulso de la contratación pública electrónica como medio para la mejora de la economía y de la competitividad en un ámbito paneuropeo, en el cual las tecnologías de la información y los estándares juegan un papel clave. La financiación de todas estas actividades requiere un esfuerzo común de las administraciones nacionales y regionales pero el retorno es evidente tanto para las empresas y ciudadanos, como para la propia Administración ya que se consigue una mejora en toda la cadena de valor en la que un contrato público actúa.

2.9 Modelo de Información para los Anuncios de Licitación

Una de las principales líneas de actuación para el impulso de la contratación pública electrónica es disponer de sistemas interoperables con la capacidad de comunicarse entre sí e intercambiar documentación, información y datos. Las propuestas e iniciativas relacionadas en la sección anterior se centran en estos puntos. De esta manera, en las etapas de contratación pública, ver Figura 2.3, se produce un flujo constante de información entre las entidades licitadoras y las partes licitantes (personas físicas o jurídicas). La flexibilidad y la capacidad de procesamiento automática de toda esta información es clave para el impulso de la contratación pública electrónica. Teniendo en cuenta que las barreras tecnológicas se pueden rebasar utilizando productos y herramientas en un estado de madurez alto, el principal problema se encuentra en la interoperabilidad.

2.9.1 Necesidad de Interoperabilidad

La interoperabilidad es una característica de los sistemas que se define como la capacidad para compartir y hacer uso de la información [196]. No debería considerarse desde un único punto de vista, sino que también tiene connotaciones en relación con la experiencia de usuario en el sentido de satisfacer sus expectativas en cuanto al intercambio y utilización de la información entre diferentes dispositivos y proveedores. Además, un entorno interoperable mejora las relaciones económicas entre las empresas ya que facilita su comunicación en entornos heterogéneos e impulsa las relaciones B2B.

The ability of two or more systems or components to exchange information and to use the information that has been exchanged.

En el actual entorno de ejecución como es la Web y con el despunte de las arquitecturas orientadas a servicios (SOA), la interoperabilidad es una de las claves para mejorar la comunicación entre los componentes. La agilidad proveniente de un entorno interoperable conlleva una mejora importante en los distintos procesos que engloba SOA: despliegue de nuevos servicios, evolución de la plataforma, etc. Las aplicaciones necesitan un entorno interoperable para mejorar su capacidad de evolución y no caer en la obsolescencia. Como ejemplo se puede pensar en una nueva aplicación que realiza un servicio totalmente innovador, suponemos igualmente que otra aplicación, estaría interesada en utilizar este servicio por la ventaja que podría obtener con respecto a sus competidores. Si no se dispone de un entorno interoperable para comunicar estas dos aplicaciones, la ventaja competitiva de las mismas decae y se necesitan aplicar recursos en tareas que podrían ser más o menos automatizadas en un entorno interoperable. Algunos de los puntos de la necesidad de la interoperabilidad en SOA son los siguientes:

- Mejorar la comunicación entre aplicaciones.
- Homogeneizar un entorno heterogéneo en cuanto a formato de datos y protocolos de comunicaciones.
- Aplicar estándares acordados de forma comunitaria.
- Agilizar los procesos de desarrollo y de mantenimiento.
- Facilitar los entornos de pruebas.
- Reutilizar lógica de negocio.

Desde un punto de vista técnico, podemos identificar dos tipos de interoperabilidad en el ámbito del software y las comunicaciones:

Sintáctica. Sucede cuando los sistemas son capaces de intercambiar información en un formato de datos determinado y bajo unos ciertos protocolos de comunicaciones. El caso más habitual es el uso de un vocabulario XML compartido para expresar los datos. Es la base de la interoperabilidad entre aplicaciones en los actuales entornos distribuidos.

Semántica. Más allá de la capacidad de intercambiar datos bajo ciertas condiciones existe la posibilidad de la interpretación automática de los mismos. Cuando las aplicaciones son capaces de compartir datos (sintaxis) e interpretarlos automáticamente (semántica) conseguimos este tipo de interoperabilidad. El valor añadido de la semántica a las aplicaciones se aplica a su autonomía. En un entorno como SOA, existen ciertas operaciones (descubrimiento o selección de servicios) que se pueden automatizar en un estado consciente, aumentando el grado de flexibilidad de la arquitectura. Por otro lado, la adición de semántica tiene que estar en buena relación con el determinismo o la precisión de las operaciones que la usan, es decir, añadir semántica en un entorno y trabajar con incertidumbre no tiene sentido para ciertas operaciones.

La importancia de la interoperabilidad queda por tanto probada en el sentido de fluidificar la comunicación entre distintas aplicaciones, es decir, que "hablen" el mismo lenguaje. Esta característica dentro de las arquitecturas globales es esencial, ya que junto a la integración se mejora por un lado la comunicación de aplicaciones y el formato de las mismas. En el momento en el que aparecen heterogeneidades se pueden utilizar adaptadores de un formato a otro, por lo que no representa en sí un gran problema una vez fijadas las características anteriores.

2.9.1.1 Cómo conseguir Interoperabilidad

Para conseguir un entorno interoperable se pueden seguir una serie de acciones que mejoran la comunicación de las aplicaciones, se pueden destacar las siguientes:

- Uso de estándares. Los estándares representan esfuerzos comunitarios en cierto sector de negocio para mejorar la comunicación entre los distintos sistemas. En este sentido, como paradigma se puede citar *eXtensible Business Reporting Language* (XBRL), el cual nace de la propuesta lanzada en 1998 por Charles Hoffman, experto contable y auditor, para simplificar la automatización del intercambio de información financiera mediante el uso del lenguaje XML. La idea subyacente en esta iniciativa no era otra que la de estandarizar el formato con el que la información financiera se distribuye entre los diferentes proveedores y consumidores.

- Tecnología común de comunicaciones. Para transportar la información utilizando cierto estándar establecido se debe utilizar un protocolo de comunicaciones común. Por ejemplo, en los servicios web se utiliza SOAP [50] (peticiones *POST* sobre HTTP) o REST [119].

- Implementaciones usando estándares. Las aplicaciones dispuestas al uso de un determinado estándar, por ejemplo servicios web WSDL [62]+SOAP, deben realizar un esfuerzo para saber trabajar con esos datos y adaptarlos a su lógica de negocio. En resumen, las aplicaciones deben ser capaces de importar y exportar datos utilizando los estándares propuestos.

2.9.1.2 Semántica e Interoperabilidad

La semántica representa un esfuerzo en cuanto a la consecución de la interoperabilidad, tanto desde el punto de vista de la obtención de modelos de datos comunes como de proporcionar los mecanismos adecuados para que la información pueda ser procesada de forma automática por las aplicaciones. La aplicación de semántica para mejorar la interoperabilidad se ve reflejada en diferentes ámbitos: modelar el conocimiento de un dominio de forma estándar, utilizar este modelo para describir los servicios del dominio o establecer un formato de datos estándar.

- Modelo de datos y formato estándar.

- El contenido es usable desde que está disponible.

- Es un formato procesable por las máquinas.

- La información existente es fácilmente representable.

- Flexibilidad en la representación de la información.

2.9.2 Propuestas y Modelos de Información actuales

Una vez revisado el concepto de interoperabilidad y su necesidad de aplicación en el contexto de la contratación pública electrónica, teniendo en cuenta la diversidad de plataformas para el proceso de contratación, modelos de información para definir las distintas fases de *e-Procurement* y las iniciativas en desarrollo que se están promoviendo desde la Unión Europea para ser transpuestas a nivel nacional y regional, es conveniente señalar que recomendaciones se siguen actualmente.

A nivel europeo el modelo desarrollado en TED es preceptivo para los anuncios de licitación que superen cierto umbral, ver Sección 2.6, en el desarrollo de esta iniciativa se ha conseguido interoperabilidad a nivel sintáctico, suministrando una serie de XML Schemas que permiten establecer un modelo común para la información contenida en los anuncios de licitación. Actualmente, estos XML Schemas no están disponibles de forma pública, salvo que se disponga de una suscripción a la herramienta *eSenders*, en su caso se pueden encontrar guías de implementación.

A nivel nacional, se puede utilizar como paradigma la Plataforma de Contratación del Estado que de nuevo han diseñado una serie de XML Schemas que permiten establecer un modelo formal para todas las fases de *e-Procurement*. En concreto, para las fases que son objeto de estudio en este documento: *eNotification* y *eAccess*, se puede encontrar el siguiente diagrama de componentes CODICE v1, ver Figura 2.8. De igual forma, existen modelos similares, hasta 15 (*Appeal Notification, Awarded Notification, Contract Award Notice, Contract Documents, Contract Documents Request, Contract Notice, Declaration, Guarantee Document, Invitation to Tender, Prior Information Notice, Qualification Result Notification, Rectification Request, Tender, Tender Reception Notification* y *Unawarded Notification*), para cada una de las fases implicadas en el proceso de contratación pública electrónica, ver Figura 2.7. En todos ellos se utilizan vocabularios XML estándar como *Universal Business Language* (UBL), códigos ISO para identificar monedas y países, etc., con el objetivo presente de la interoperabilidad como máxima para impulsar el uso de estas plataformas. A nivel regional, la casuística vuelve a ser diversa dependiendo de distintos modelos de información basados en diferentes especificaciones formales.

El principal problema de todos estos modelos reside en la sobre-especificación que se realiza de la información. La necesidad de que las distintas herramientas accedan a diferentes datos obliga a modelar con un grado de especificidad muy alto cada tipo de dato, proporcionando además una gran cantidad de metainformación. Todos estos esquemas siguen este modelo y realmente son extremadamente difíciles de generar y tratar, especialmente para terceras partes interesadas en acceder a esta información. Si bien supone un gran avance para la contratación pública electrónica, también en muchas ocasiones supone una considerable barrera de entrada ya que llega a comprender toda la información necesaria, etc., requiriendo un enorme esfuerzo integrador que conlleva que posibles agentes involucrados en el consumo de esta información no estén dispuestos a realizar.

Esta situación ha sido identificada dentro del proyecto *10ders Information Services* y por ello se ha desarrollado un nuevo vocabulario XML con el objetivo de aunar la información de los anuncios de licitación, contratos, etc., intentando unificar la versiones previas de TED, CODICE v1 y v2 y modelos regionales. Este trabajo se ha plasmado en la creación de *opXML* que, nuevamente, incluye una serie de modelos en XML Schema para expresar dicha información. El objetivo principal es servir como elemento vertebrador para expresar la información de forma común. Evidentemente, es similar a las especificaciones previas pero busca una definición simplificada que sea más manejable por terceros y que tenga transformación directa con las distintas plataformas de contratación. La principal ventaja es que surge por parte de una empresa experta en el dominio. Por otra parte, las desventajas de este enfoque residen en que reiteradamente se define un vocabulario XML cuando ya existen varios en TED, CODICE, etc., con este objetivo y no existe ningún consenso con otras entidades como puede ser la propia Administración o proveedores de servicios. A continuación y en contraste con los modelos anteriores de información se puede observar el enfoque seguido en *opXML*, ver Figura 2.9.

La principal conclusión que surge de esta análisis es que se han realizado numerosos y valiosos esfuerzos para modelar la información disponible en el proceso de contratación pública electrónica, pero que presentan desventajas en el sentido de una excesiva sobre-especificación, falta de consenso entre los agentes implicados, etc., que conlleva a una replicación de esfuerzos y a la especificación de las mismas partes repetidamente. No obstante, todo el conocimiento generado sobre este dominio debe a, largo plazo, convertirse en la semilla de algún tipo de estándar que pueda ser aplicado transversalmente en cualquier entidad con necesidades de contratación de productos y servicios.

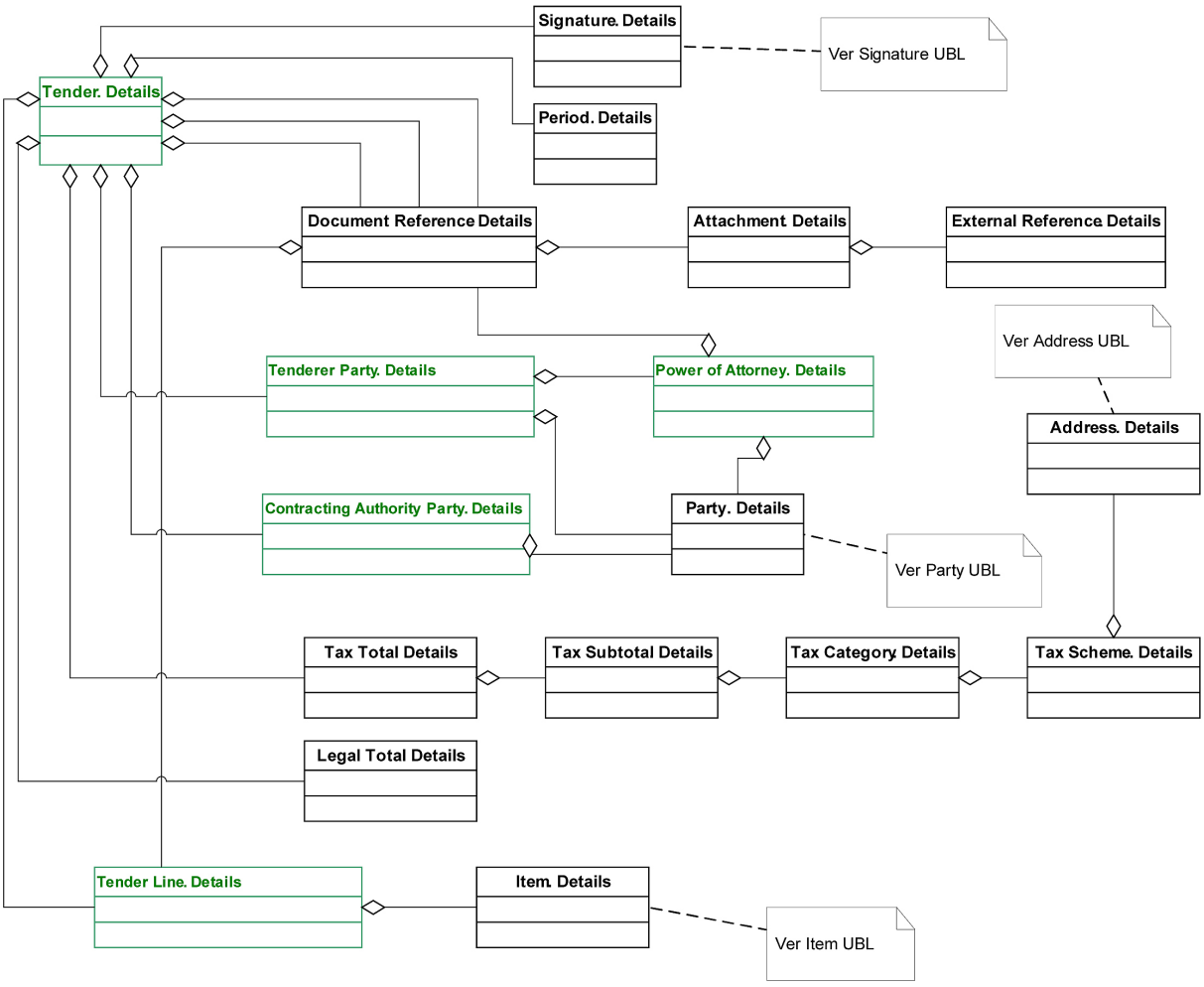


Figura 2.7: Modelo del Proceso de Contratación Pública Electrónica en CODICE v1.

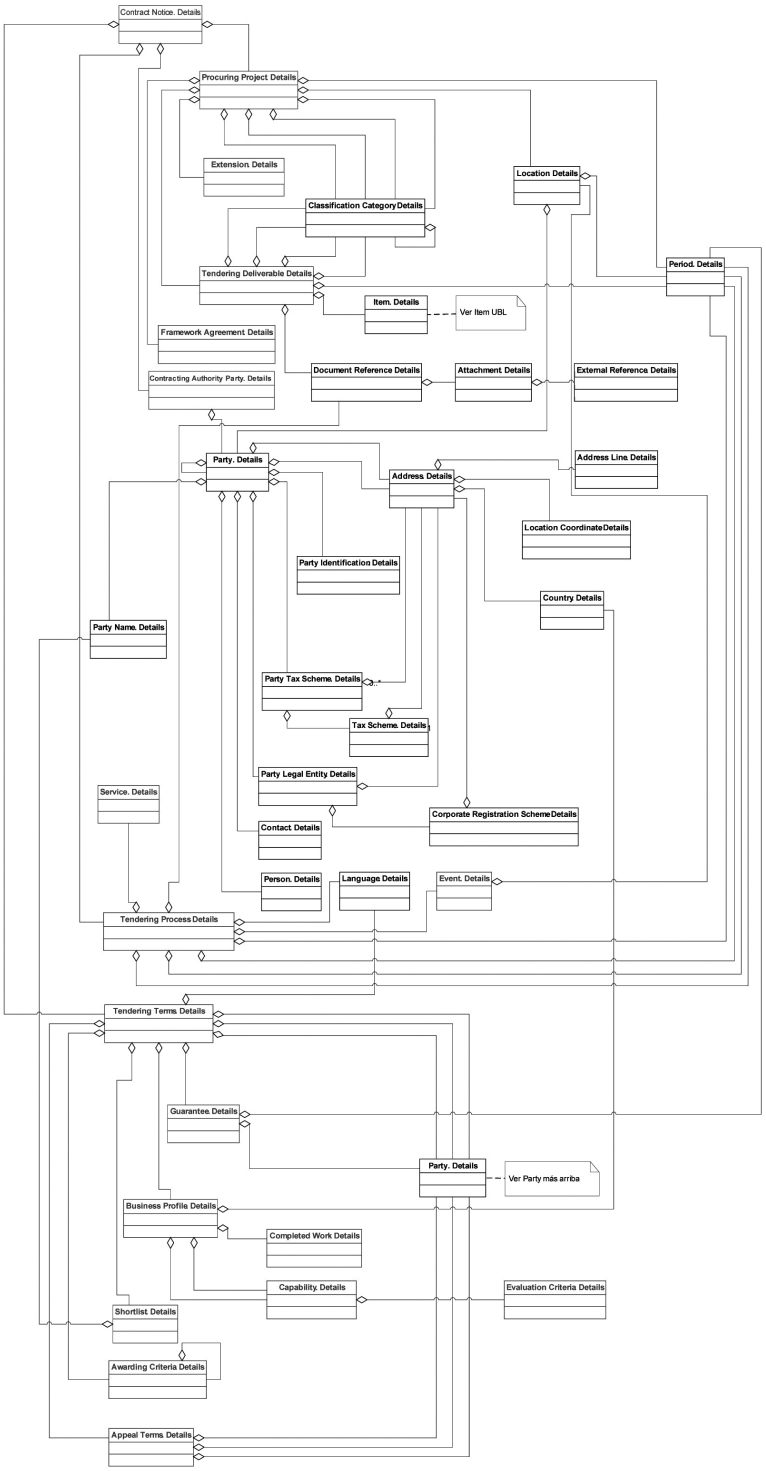


Figura 2.8: Modelo de Información de Anuncio de Licitación en CODICE v1.

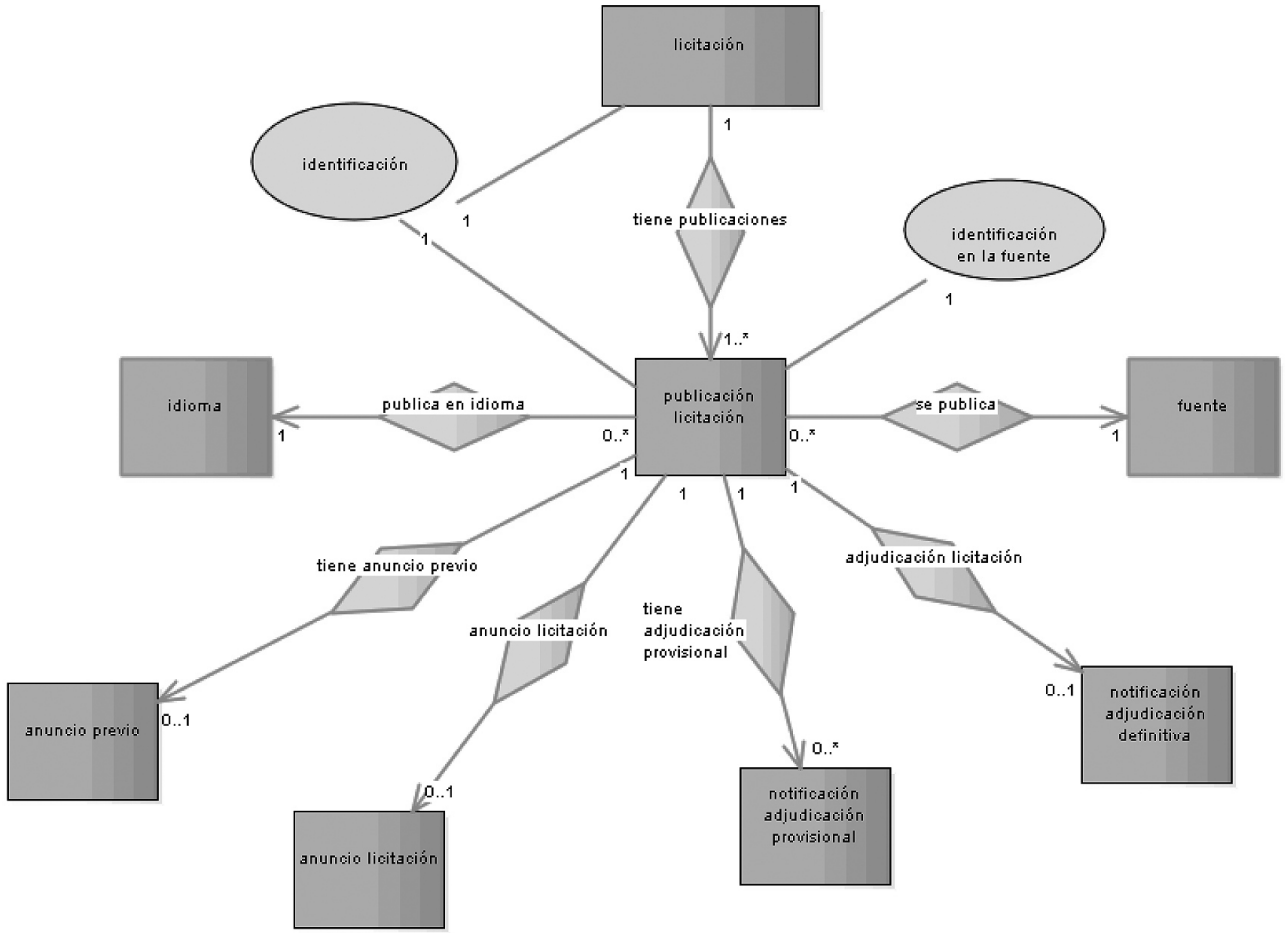


Figura 2.9: Modelo de Información de Anuncio de Licitación en *opXML*.

2.10 Clasificaciones Estándar de Productos

Los anuncios de licitación contienen información muy relevante para la búsqueda de oportunidades de negocio para las distintas empresas, relativa a la localización, la cuantía del contrato, la duración, etc., importante para concursar en la adjudicación del contrato, sin embargo, la información clave en un anuncio de licitación es el tipo de contrato que se va a adjudicar ya que con ello se crea el mayor filtro posible para las posibles empresas. En este sentido, han surgido muchas iniciativas [270] para aunar la adición de descriptores estándares [203] a los anuncios de licitación. A este tipo de información se le denomina, de forma particular, esquemas o catálogos estándar de productos y servicios, en general se pueden asimilar a un sistema de organización de conocimiento como tesauros, taxonomías o sistemas de clasificación jerárquicos, utilizados ampliamente para la organización [202] de grandes colecciones de información como documentos, textos, páginas web o recursos multimedia tanto en el ámbito público como privado [204]. Estos vocabularios permiten que los usuarios puedan anotar los objetos de información de una forma sencilla para que posteriormente su consulta y acceso se simplifique. El uso de estas técnicas de etiquetado con vocabularios controlados está ampliamente asentado [271] para describir el *topic* de distintos objetos de información, suponiendo un primer paso para el tratamiento automático de la información.

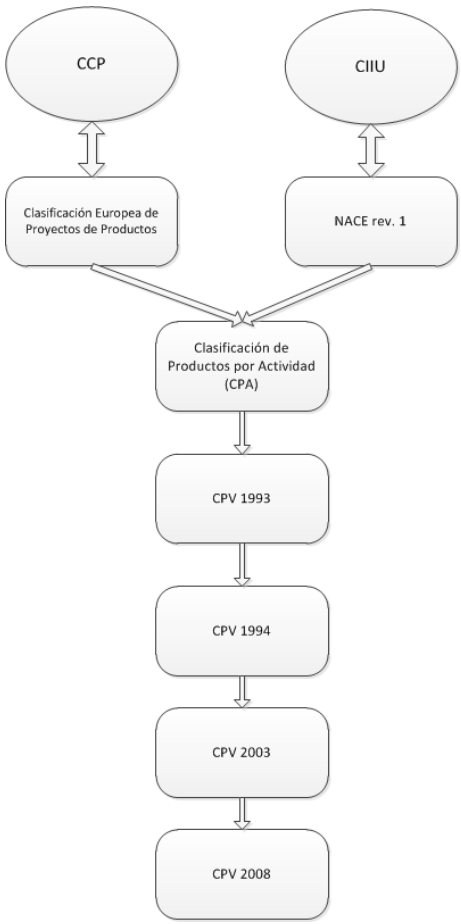


Figura 2.10: Evolución *Common Procurement Vocabulary*.

Por ejemplo en el contexto europeo han surgido muchas clasificaciones de este tipo para distintos dominios, disponibles en el servidor de metadatos RAMON. Entre ellos se puede destacar "*the European Schedule of Occupational Diseases*" en el campo de la salud, "*the European Education Thesaurus or the European Glossary on Education*" en materia de educación, "*the International Standard Classification of Occupations*" en el ámbito trabajo, etc. La estructura de estos sistemas de clasificación es similar:

jerarquía de entidades y multilingüismo, contienen sin embargo, factores heterogéneos que impiden su interoperabilidad y que deben ser abordados para conseguir un aprovechamiento eficiente del esfuerzo que supone realizar este tipo de clasificaciones.

En el campo de la contratación pública electrónica es evidente que este tipo de paradigmas de clasificación controlada de documentación como anuncios de licitación, lotes, contratos, etc., es clave para realizar una gestión eficiente de la gran cantidad de información y documentos que son generados constantemente. Como consecuencia, una de las primeras medidas impulsadas desde la Comisión Europea ha sido la realización del "*Common Procurement Vocabulary*" (CPV) [70], se trata de una clasificación para describir los objetos de un contrato permitiendo a los agentes implicados una forma sencilla de buscar los anuncios de licitación. De esta forma, se ha conseguido armonizar el sistema de codificación de los objetos de contrato. Además, el uso del CPV es obligatorio desde el 1 de febrero de 2006 según se establece en la Directiva Regulation (EC) No 2195/2002 [109] del Parlamento Europeo. En el ámbito europeo el CPV es, sin duda, el vocabulario controlado clave en la esfera de la contratación pública electrónica, que ha sido fruto del esfuerzo y trabajo desde otras clasificaciones, ver Figura 2.10, como la "Clasificación Central de Productos" (CPC), la "Clasificación Industrial Internacional Uniforme" (CIIU) y la "Clasificación de Productos por Actividades" (CPA)

Además del CPV existen otras muchas clasificaciones [262] interesantes, ver Tabla 2.3, en comercio electrónico o bien para la extracción de estadísticas, que han sido formuladas por diferentes organismos como la UNESCO, ONU o el Gobierno de Estados Unidos, que actualmente tienen en el mejor de los casos *mapeos* entre ellas, pero cuyo acceso suele ser complicado tanto por los formatos en los que se presentan, habitualmente MSExcel o PDF, como por las aplicaciones y sistemas de búsqueda muy rígidos.

Clasificación	**Acrónimo**	**Organismo**
Combined Nomenclature 2012 (desde 1995)	CN	Unión Europea
Central Product Classification, version 2 (2008)	CPC	Unión Europea
Clasificación de Productos por Actividad (2008)	CPA	Unión Europea
Integrated Tariff of the European Communities)	TARIC	Unión Europea
International Standard Industrial Classification of All Economic Activities, Rev.4	ISIC	*United Nations Statistics Division*
North American Product Classification System	NAPCS	Agencias estadísticas de Canadá, México y Estados Unidos
North American Industry Classification System 2007 y 2012	NAICS	Gobierno de Estados Unidos
PRODuction COMmunautaire	PRODCOM	Unión Europea
Standard International Trade Classification, Revision 4	SITC	*United Nations Statistics Division*
Nomenclature générale des activités économiques dans les Communautés européennes	NACE	Unión Europea
United Nations Standard Products and Services Code	UNSPSC	Consorcio entre ellos la ONU
...		

Tabla 2.3: Catálogo de Clasificaciones Estándar de Productos.

El uso de estas clasificaciones es clave para la trazabilidad de proveedores de productos y servicios, sistemas de anotación y etiquetado, extracción de información, etc. En el contexto de los anuncios de licitación representan el método de especificar el objeto del contrato pero no todo son ventajas, ya que al existir una amplia variedad, es difícil mantener la consistencia entre ellas y sus distintas versiones, forzando a que se dupliquen esfuerzos en los distintos organismos para extraer información basada en estos códigos.

2.11 Información sobre Organizaciones

An organization is considered to be a set of constraints on the activities performed by agents.

Este enfoque de organización fue presentando por Max Weber [299], posteriormente Mintzberg realizó un análisis de las estructuras organizacionales distinguiendo entre cinco tipos y configuraciones diferentes. Estos enfoques ponen de manifiesto los mecanismos que operan dentro de las organizaciones con el objetivo de coordinarse para conseguir un determinado objetivo materializado a través de diferentes metas, procesos y reglas de negocio, posición en el mercado y comunicación. Existen trabajos [125] en los cuales se intenta modelar la estructura de una organización, pero que no ha abordado problemas como: 1) falta de información para definir el *status* de una organización, 2) maraña de subdivisiones para especificar la estructura entre los elementos de la organización, 3) ausencia de capacidades léxicas para describir los conceptos y relaciones que se manejan en una organización, 4) distintos tipos de nombrado, etc.

La problemática para definir la estructura de una organización y su actividad no es sencilla, ya que puede ser muy diversa dependiendo de su organización, objetivos particulares, etc., es por ello que en el campo de la contratación pública electrónica resulta de gran valor la posibilidad de disponer de la información de los agentes implicados en el proceso administrativo, para así poder trazar su actividad a lo largo del tiempo. Desde un punto de vista del órgano contratante, su perfil debe estar disponible al público, el cual, además de los anuncios de licitación debe comprender la información de contacto, persona encargada, etc., toda esta información actualmente está accesible pero de una forma ligeramente rudimentaria. De igual forma, cuando los contratos son adjudicados, se dispone de la información relativa a la empresa adjudicataria pero de nuevo el acceso a esta información sigue un modelo tradicional. Un proceso lógico sería que una entidad especificara su perfil de contrante siguiendo un estándar en cuanto a su organización interna y que por otro lado las empresas pudieran definir un perfil en el cual indicar sus intereses y así facilitar el proceso de encaje entre tipos de contrato e intereses.

2.12 Evaluación del estado actual del mundo de *e-Procurement*

En este capítulo se han repasado los conceptos clave sobre contratación pública y en concreto la contratación pública electrónica. Desde la Unión Europea existe un claro compromiso y liderazgo por impulsar este proceso administrativo por la relevancia y el impacto que tiene en la sociedad, ya que sin lugar a dudas las cifras de inversión y gasto que se manejan resultan de gran interés para las empresas de todo tipo y sector. Sin embargo, aunque las distintas iniciativas han conseguido avanzar en la consecución de este gran reto, existen puntos clave en los cuales las soluciones existentes no han conseguido desarrollar toda su capacidad para atraer a las empresas y así conseguir un mercado competitivo a nivel paneuropeo. En este sentido, la disponibilidad de soluciones técnicas ya existentes así como las que se están desarrollando y la transferencia tecnológica desde los entornos académicos deben mejorar las fases de *e-Procurement*. Esta inversión en nuevas capacidades es clave para la adopción y uso efectivo de este proceso administrativo de forma electrónica. Por ello, son

esenciales tanto la publicación como la accesibilidad transfronteriza a la información y sistemas de contratación pública electrónica.

Los retos que se plantean son diversos y abarcan desde la generación de confianza suficiente entre entidades adjudicadoras y proveedores a través de un marco legislativo adecuado, hasta la generación de herramientas y medios que permitan la comunicación efectiva entre los distintos agentes participantes en el proceso. Muchas veces, los requisitos técnicos en lugar de facilitar la adopción de una nueva metodología de trabajo, suponen una gran barrera de entrada, un simple sistema de autenticación se puede convertir en un gran inconveniente para el uso de los medios electrónicos. No obstante, la transición hacia este nuevo mundo electrónico para la contratación se está realizando de forma sostenible y a distintas velocidades, dependiendo de las capacidades de las propias Administraciones. Las prioridades para realizar el despliegue definitivo se centran en el uso de medidas que actúan como incentivo o penalización para acelerar este nuevo método, impulsar la participación paneuropea en los distintos procesos de licitación pública, simplificando las condiciones a satisfacer para el uso de los nuevos sistemas y facilitando la autenticación entre los distintos agentes económicos. Finalmente, todos estos retos y prioridades se realizarán bajo una base normalizada, tanto en tecnología como en estándares, para la distinta documentación generada durante el proceso con el objetivo final de que paulatinamente todo tipo de empresas puedan competir por los contratos públicos bajo las mismas condiciones con las consiguientes ventajas para el mercado económico común.

Capítulo 3

Panorámica de uso de la Web Semántica y *Linked Data*

> ¿Por qué esta magnífica
> tecnología científica,
> que ahorra trabajo y nos hace la
> vida mas fácil,
> nos aporta tan poca felicidad?
> La repuesta es simplemente:
> porque aún no hemos
> aprendido a usarla con tino.
>
> *Citas Célebres*
> ALBERT EINSTEIN

3.1 Web Semántica

El término "Web Semántica" se puede interpretar como una evolución de la web actual. El uso del adjetivo "semántica" aporta un atractivo extra a la Web, dotando a la misma de características cercanas a lo que se entiende habitualmente por inteligencia artificial y que tan utilizadas son en películas de ciencia-ficción. No se debe confundir un avance tecnológico con una revolución en la concepción de la web actual y es preciso ser cautos con el campo de aplicación de la "semántica" en la red y seleccionar aquellos escenarios en los que realmente la semántica enriquezca el desarrollo actual.

Siguiendo las directrices que marcan las grandes empresas y que corresponden en muchos casos a la opinión de *Gartner* ("Semantic Web Technologies Take Middleware to Next Level" [186]), ya se vaticinaba que en el año 2005 la integración de aplicaciones empresariales estaría guiada por ontologías y por tanto impulsaba, la Web Semántica.

El despliegue de tecnología semántica, confiere una visión más realista, en la cual existen grandes discusiones abiertas sobre temas diversos, por ejemplo reglas, razonamiento distribuido, procesamiento de *Big Data*, lenguajes de formalización creados por distintas instituciones con distintos niveles de expresividad, herramientas en continuo desarrollo no estables, aplicaciones que se venden como Web Semántica cuando no lo son estrictamente o bien simplemente se han unido a la "iniciativa" de una tecnología, con el único objetivo de disponer de este sello en sus aplicaciones.

Las expectativas se presentan ambiciosas y tentadoras, pero no menos que en la década de los

años setenta con la eclosión de la inteligencia artificial. Por todo ello, es necesario un análisis riguroso y exhaustivo para dotar a este avance tecnológico de un planteamiento teórico y práctico adecuado donde realmente sea aplicable.

Esta introducción aunque un tanto pesimista, trata simplemente de otorgar un enfoque realista a esta tecnología de vanguardia, ciertamente aunque las barreras a superar son importantes, también hay que resaltar que el impulso realizado por todas las partes implicadas: empresa, universidad, administración, desarrolladores, etc., está siendo considerable y los resultados estables no se dilatarán excesivamente en el tiempo pero contando siempre con una base firme.

3.1.1 Definición

La Web Semántica siguiendo la definición propuesta por el W3C se presenta como:

> *Una web extendida, dotada de mayor significado, en la que cualquier usuario en Internet podrá encontrar respuestas a sus preguntas de forma más rápida y sencilla gracias a una información mejor definida.*

Surge como respuesta a las dificultades que aparecen al tratar de automatizar muchas tareas en la web actual. Hoy en día, los contenidos de la web son creados considerando que van a ser consumidos principalmente por personas, lo que hace difícil su interpretación por parte de agentes *software*. En consecuencia, algunas tareas comunes como la búsqueda de información en la web actual, son notoriamente mejorables; otras tareas aparentemente sencillas son casi imposibles de implementar. En el origen de este problema se encuentra la incapacidad de los agentes *software* (las máquinas) para encontrar, interpretar, extraer y combinar la información ya disponible en la web.

Para tratar de paliar esta situación, *Tim Berners-Lee*, reconocido como "padre de la web", impulsa a través del W3C el desarrollo de la Web Semántica [29,30]. A través de esta iniciativa, se desarrollan los lenguajes y formalismos que permiten extender la web actual, de tal manera que sus contenidos sean accesibles tanto a personas como a máquinas. Esta tecnología abre la puerta a una nueva generación de aplicaciones informáticas capaces de encontrar, seleccionar y combinar la información dispersa en la web para realizar tareas que actualmente se ejecutan de forma manual: seleccionar los resultados relevantes en una búsqueda, agregar información procedente de distintas fuentes, etc.

Una de las fortalezas de la web actual es la ingente cantidad de información que se encuentra publicada en ella y la infinidad de servicios a los que se puede acceder. Sin embargo, si la explotación de estos recursos requiriese necesariamente la intervención humana, como sucede ahora, su utilidad estaría limitada, de ahí que el W3C pretenda "guiar a la web hacia su máximo potencial" como herramienta universal y multipropósito. La Web Semántica proporciona una infraestructura para explotar eficientemente el potencial de la web [21, 101], se encuentra, sin embargo en continuo desarrollo, gestándose gradualmente su progresión.

En el contexto de la iniciativa de la Web Semántica, se han desarrollado y se continúan desarrollando estándares que son el soporte necesario para hacer realidad la misma. En la base se utilizan tecnologías estándar ya asentadas, como XML [51], HTTP [28] y URIs [27], que son también la base de la web actual. Pero además se han creado nuevos mecanismos para describir semántica y formalmente la información, se trata, entre otros, de RDF [195], RDF Schema [52] y OWL [164, 248]. Con estos lenguajes y modelos se puede describir la información de manera precisa y y carente de ambigüedad basándose en teorías lógicas como las *Description Logics [11] (DLs)*. En este sentido juegan un papel fundamental las taxonomías, los tesauros y las ontologías, como estructuras capaces de dotar de significado a los datos [19].

3.1.1.1 No es Web Semántica

En muchas ocasiones, no existe manera más conveniente para definir un concepto que encontrar un contraejemplo. Así, *Tim Berners-Lee* insiste en que la Web Semántica **No** es inteligencia artificial:

> *El concepto de documento entendible por una máquina no implica algún tipo de inteligencia artificial mágica que permita a las máquinas comprender el farfullar de los humanos. Sólo indica una habilidad de la máquina para resolver un problema bien definido a base de realizar operaciones bien definidas sobre unos datos bien definidos. En vez de pedir a las máquinas que entiendan nuestro lenguaje, se le pedirá a la gente que haga un esfuerzo extra. A roadmap to the Semantic Web. What the semantic Web isn't but can represent*. 1998.
>
> Fuente: http://www.w3.org/DesignIssues/RDFnot.html

Por tanto, desposeída de su aura mágica al no constar entre sus objetivos la conquista del lenguaje natural, la Web Semántica queda reducida a un intercambio de información eficiente entre los agentes.

3.1.2 Infraestructura para la Web Semántica

La iniciativa de la Web Semántica puede ser divergente en sus enfoques para los diferentes escenarios de aplicación, pero lo que está perfectamente definido es su apoyo en el uso de **ontologías** [276].

3.1.2.1 Ontologías en la Web Semántica

La evolución de la web a través de la Web Semántica, proporciona un nuevo espacio de publicación de recursos (documentos) que junto con las nuevas tecnologías de la información generan un entorno con unas condiciones inigualables para la divulgación de contenido.

Uno de los principales problemas de la web actual es la inmensa cantidad de información que se publica obviando cualquier procedimiento de control, creando grandes bases de datos, recursos y conocimiento, cuya explotación no es eficiente, ya que en muchos casos sólo está preparada para ser procesada o bien por agentes humanos o por agentes *software*.

Las ontologías se presentan como una forma para la organización de conocimiento y contenido heterogéneo en el ámbito de la Web Semántica, estableciendo un enlace entre el procesamiento humano de recursos y el automático, realizado entre agentes *software*. Paralelamente, las ontologías se utilizan como instrumentos para la desambiguación del lenguaje natural. Constituyendo en definitiva un mecanismo para la realización de la comunicación entre agentes de forma eficiente basándose en un conocimiento común y compartido. En la Sección 3.1.3 se describen más amplia y genéricamente los modelos de conocimiento basados en ontologías.

3.1.2.2 Arquitectura para la Web Semántica

En un primer momento, la arquitectura para la Web Semántica, más conocida como "tarta o pila de la Web Semántica", ver Figura 3.1, se diseñó con diferentes capas de abstracción, construyendo un *framework* semántico que recoge, supuestamente, todas las necesidades para la gestión del conocimiento. Cada una de las capas enriquece a la inmediatamente inferior proveyendo nuevos servicios o niveles de formalización superior del conocimiento.

A continuación se explica someramente la intención de cada una de las capas y el por qué de su presencia:

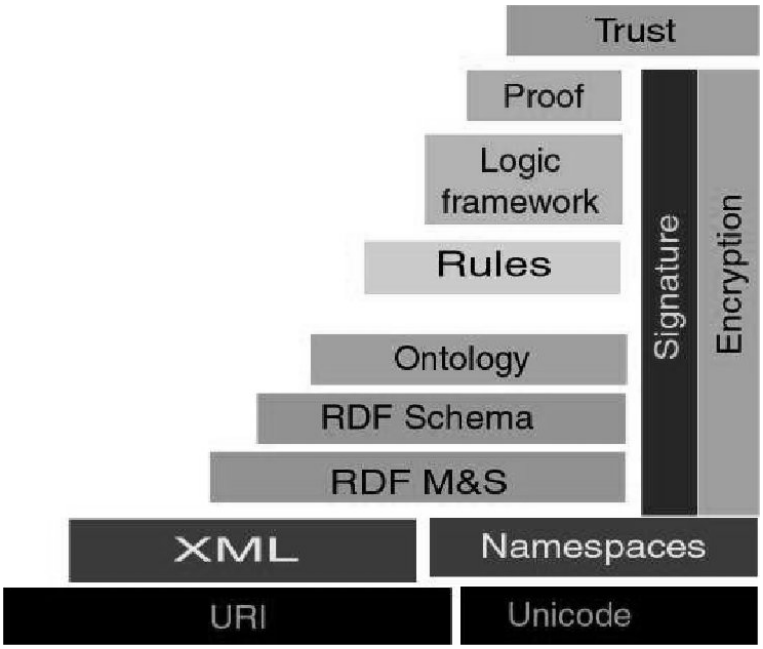


Figura 3.1: Arquitectura Web Semántica 2002.

URI/IRI-Unicode. Para dar un soporte estándar a una tecnología existen dos factores fundamentales: nombrado e identificación única de cada uno de los recursos y codificación de los mismos. Por esta doble razón aparecen los *Unified Resource Identifier* y la codificación estándar Unicode, implementada principalmente en UTF-8.

Ejemplo 3.1.1. *URI: protocolo:dirección:directorio:recurso,* `http://www.josemalvarez.es/foaf.rdf`

XML [51]. Lenguaje extensible de marcas (*eXtensible Markup Language*), es un formato estándar creado para la estructuración de datos, en la actualidad se encuentra regulado por el W3C que es el encargado de realizar las distintas especificaciones y versiones desde Febrero de 1998. XML está basado en SGML (*Standard Generalized Markup Language*, ISO 8879) que ya había sido establecido en 1986. El uso principal de XML es la estructuración de datos, pero en general es utilizado por todas aquellas aplicaciones informáticas en las que se pueda representar la información jerárquicamente. El lenguaje XML es en sí mismo un metalenguaje utilizado ampliamente para definir otros lenguajes. Consta entre otros, de elementos y atributos. Su orden y jerarquía son los encargados de formar el lenguaje definido con XML. No se dispone de etiquetas predefinidas, como podría ser HTML, siendo el usuario el responsable de definir un conjunto de elementos con sus etiquetas asociadas, la semántica de los documentos es proporcionada por la aplicación que use esos datos. Para definir la estructura de un documento XML, es decir, una gramática que indique el orden de los elementos y su jerarquía, se consigue de dos formas utilizando: 1) DTD o 2) XML Schema [113].

Es importante resaltar que un documento bien formado no es lo mismo que un documento válido, un documento estará bien formado cuando siga las reglas sintácticas de formato XML, mientras que un documento será válido si todos sus elementos están en orden correcto de anidación y está bien formado. La utilización de XML resulta de interés en base a alguna de las razones siguientes:

- Estándar para el intercambio de datos.
- Facilidad de uso.
- Legibilidad.

- Implantación.
- Extensibilidad.
- Separación entre formato y contenido.
- Tratamiento multiplataforma.
- Es libre, especificaciones disponibles.

```
<?xml version="1.0" encoding="utf-8"?>
<events xmlns="http://purl.org/weso/events" xmlns:xhtml="http://www.w3.org/1999/xhtml
    " xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http:
    //purl.org/weso/events events.xsd">
        <event type="family">
                <title>Bautizo</title>
                <date>28/10/2011</date>
                <start>12:00:00</start>
                <place>Tapia de Casariego</place>
                <description>
                        Bautizo de <xhtml:em>Pablo</xhtml:em> en
<xhtml:a href="http://maps.google.com/maps" title="Tapia de Casariego">Tapia de
    Casariego</xhtml:a>.
                </description>
        </event>
        <event type="work">
                <title>Seguimiento</title>
                <date>04/11/2010</date>
                <start>10:00:00</start>
                <end>13:00:00</end>
                <place>Facultad de Ciencias, Oviedo</place>
                <persons>
                        <person>Chema</person>
                </persons>
                <description>
                        Kick-off de 10ders.
                </description>
        </event>
</events>
```

Figura 3.2: Ejemplo de fichero XML.

Como aplicación práctica de XML, ver Figura 3.2, y debido a la necesidad de tratamiento automático para diferentes fines (formatos de presentación, transformación de un vocabulario a otro, etc.) hay que destacar XSL [65] para la generación del contenido a partir de un documento XML de una manera rápida, sencilla y eficaz, cuyo objetivo principal es presentar al usuario final un interfaz del documento más comprensible con el que pueda realizar un tratamiento de la información contenida en él de una forma más asequible, sin necesidad de conocer el formato XML.

Por ello desde el consorcio W3C se optó por realizar la especificación XSL con el objetivo de presentar un modelo para el procesamiento de la información almacenada en formato XML. La especificación redactada por el consorcio web para la transformación del contenido XML se denomina XSL, se encarga de definir una hoja de estilo con la cual se transforma el fichero XML en otro formato, teniendo en cuenta que la información en XML se almacena de forma jerarquizada, con la definición de XSL se pretende poder presentar al usuario dicha información en otros formatos estructurados como pueden ser: HTML, PDF, etc., XSL es una especificación, pero el lenguaje del que hace uso para la realización de la transformación es XSLT con sintaxis de XPath [66].

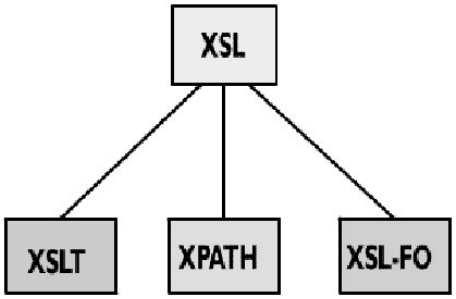


Figura 3.3: Componentes de XSL.

En XSLT se define cómo se ha de realizar la transformación del documento XML y no cuándo, así la transformación del documento se realiza en distintos pasos:

- Generación de un árbol a partir del fichero fuente de XML, esto se realiza mediante un procesador que analiza el documento, realizando al mismo tiempo una validación sintáctica del mismo.
- Procesamiento del árbol generado construyendo un nuevo árbol con la información procesada, el recorrido del árbol generado se realiza en preorden, con la posibilidad de variar el tipo de recorrido, es importante tener en cuenta el orden de evaluación de los nodos del árbol para así poder aplicar las plantillas correctamente.

El principal uso de XSLT como ha quedado patente en la introducción, es la transformación de documentos XML para generar contenido web, lógicamente esto no aportaría demasiada potencia a este lenguaje si solo sirviera para generar contenido web estático, en cambio usando este lenguaje se genera contenido dinámico aplicando distintas reglas, como ejemplo se podría obtener de una base de datos un fichero XML y con dicho fichero y una transformación apropiada generar una página dinámica con contenidos auto-actualizables cada cierto tiempo, otro ejemplo de aplicación podría ser la conversión de un programa escrito en un lenguaje a otro, definiendo las reglas correctas. XSLT utiliza también XPath, ver Figura 3.4, que es una especificación para el acceso a los valores de los distintos nodos del árbol generados a partir del procesamiento del fichero XML.

XML Schema [113]. En primer lugar es interesante diferenciar las dos tecnologías establecidas para la definición de la estructura de un documento XML: 1. *Document Type Definition* (DTD) es un vocabulario para definir las reglas de construcción de un documento XML. 2. XML Schema: Vocabulario XML utilizado para definir otros vocabularios XML.

El incremento en el uso de XML hace necesario la utilización de tecnología para poder expresar la estructura del documento y así proceder a su validación. Aunque en principio el objetivo de ambos es el mismo: definir vocabularios XML para poder validarlos sintácticamente e imponer distintos tipos de restricciones como de cardinalidad o integridad, cada uno presenta unas características diferentes lo que supone que escoger entre uno u otro implique realizar un estudio de lo que se pretende desarrollar. También disponemos de herramientas que nos permiten realizar la transformación de uno a otro pero sólo en el sentido de DTD a XML Schema. Se puede pensar que XML Schema es el sucesor de las DTD.

La utilización de XML Schema, ver Figura 3.5, se apoya en diferentes características que hacen que su uso sea interesante en el ámbito de la Web Semántica:

- Espacios de nombres, permite utilizar los mismos identificadores en el mismo documento, evitando así la ambigüedad.
- Esquema, es la estructura que va a presentar el vocabulario XML construido, con sus distintos elementos y atributos así como con las validaciones que sean necesarias.

```
<?xml version="1.0" encoding="iso-8859-1"?>
<xsl:stylesheet xmlns:evt="http://purl.org/weso/events"
        xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
                xmlns:xhtml="http://www.w3.org/1999/xhtml"
        exclude-result-prefixes="evt"
        version="1.0">
    <xsl:output method="xml" doctype-system="http://www.w3.org/TR/xhtml11/DTD/xhtml11
        .dtd" doctype-public="-//W3C//DTD XHTML 1.1//EN" indent="yes" encoding="iso
        -8859-1"></xsl:output>

    <xsl:template match="/">
        <html>
            <body>
                <xsl:apply-templates select="//evt:events"/>
            </body>
        </html>
    </xsl:template>

    <xsl:template match="evt:events">
        <div id="events">
            <ul>
            <xsl:apply-templates select="//evt:event">
                <xsl:sort select="concat(substring-after(substring-after(string(./
                    evt:date),'/'),'/'),substring-before(substring-after(string(./
                    evt:date),'/'),'/'),substring-before(string(./evt:date),'/'))"
                    data-type="number"/>
            </xsl:apply-templates>
            </ul>
        </div>
    </xsl:template>

    <xsl:template match="evt:event">
            <xsl:element name="li">
                <xsl:attribute name="class">
                    <xsl:value-of select="./@type"/>
                </xsl:attribute>
                <h2 class="title"><xsl:value-of select="./evt:title"/></h2>
                <h3 class="date"><xsl:value-of select="./evt:date"/></h3>
                <h4 class="start"><xsl:value-of select="./evt:start"/></h4>
                <xsl:if test="./evt:end">
                    <h4 class="end"><xsl:value-of select="./evt:end"/></h4>
                </xsl:if>
                <xsl:if test="./evt:place">
                    <strong class="place"><xsl:value-of select="./evt:place"/></
                        strong>
                </xsl:if>
                <xsl:apply-templates select="./evt:persons"></xsl:apply-templates>
                <xsl:apply-templates select="./evt:description"></xsl:apply-templates
                    >
            </xsl:element>
    </xsl:template>

  ...
</xsl:stylesheet>
```

Figura 3.4: Ejemplo de hoja de estilo XSL.

- Documento instancia, documento XML creado con una estructura definida en un XML Schema y contra el cual se "valida".

Las razones que pueden llevar a elegir XML Schema son, entre otras, las siguientes:

- Utiliza sintaxis XML.
- Perfectamente documentado.
- Define los elementos y atributos que pueden aparecer en un documento instancia.
- Establece la jerarquía y orden de los distintos elementos.
- Creación de distintos tipos genéricos.
- Extensible.
- Inclusión de documentos externos.
- Expresividad.
- Posibilidad de documentación.
- Tipos de datos simples, complejos, derivados.
- Expresión de restricciones de integridad y cardinalidad.
- Reutilización de tipos por extensión o restricción.
- ...

La utilización de XML Schema es crítica en los servicios web basados en WSDL [61] y SOAP [50] y está muy asentada en entornos de desarrollo basados en Java a través de herramientas como JAXB.

RDF [195]. *Resource Description Framework*, se encarga de describir los recursos, añadiéndoles la meta-información necesaria en un formato procesable. En la siguiente Sección 3.1.2.4 se aborda la descripción de RDF de manera más extensa.

Ontologías: Los documentos etiquetados constituyen una gran cantidad de información disponible para utilizar por la máquina. Están disponibles los datos, pero todavía no hay capacidad semántica, es necesario construir un modelo donde "encajar" esos datos. Para añadir la componente semántica mediante ontologías, ver Sección 3.1.3, se pueden utilizar diferentes lenguajes, cuyo estudio se abordará en la Sección 3.1.2.3.

Capas superiores: La arquitectura propone diferentes niveles en los que se colocan las reglas y la lógica, debido a su complejidad todavía es precipitado definirlas por completo y las discusiones se mantienen abiertas en grupos de trabajo del W3C como el RIF [48]. Los módulos transversales de "firma" [13] y "encriptación" [174] están definidos y se pueden encontrar como recomendaciones del W3C.

Aunque las ventajas de un diseño basado en capas está sobradamente demostrado en bibliografía de ingeniería del *software*, hay que resaltar que esta primera aproximación ha sido modificada con el objetivo de recoger la realidad de la arquitectura para la Web Semántica. Hay que tener en cuenta que la primera versión propuesta por *Tim Berners Lee* (web XML) ofrecía una visión ideal que contemplaba todas las partes implicadas, pero que no tenía el efecto experiencia de la construcción de aplicaciones y de los posibles problemas que se podrían encontrar. Por ello, esta arquitectura ha sido rediseñada planteando un modelo más cercano a la realidad, al menos de las aplicaciones y problemas, siendo también conocida como las "dos torres" [193] (web RDF), ver Figura 3.6.

```
<?xml version="1.0" encoding="iso-8859-1"?>
<xs:schema xmlns="http://purl.org/weso/events"
        xmlns:xs="http://www.w3.org/2001/XMLSchema"
        xmlns:xhtml="http://www.w3.org/1999/xhtml"
        targetNamespace="http://purl.org/weso/events"
        elementFormDefault="qualified" attributeFormDefault="unqualified">

<xs:import namespace="http://www.w3.org/1999/xhtml"
        schemaLocation="http://www.w3.org/2002/08/xhtml/xhtml1-strict.xsd"/>
        <xs:simpleType name="customString">
                <xs:restriction base="xs:string">
                        <xs:minLength value="1"/>
                        <xs:maxLength value="255"/>
                </xs:restriction>
        </xs:simpleType>

        <xs:simpleType name="customDate">
                <xs:restriction base="xs:string">
                        <xs:pattern value="([0-9]{2,2}/){2,2}([0-9]{4,4})"/>
                </xs:restriction>
        </xs:simpleType>

        <xs:simpleType name="categories">
                <xs:restriction base="xs:NMTOKEN">
                        <xs:enumeration value="family"/>
                        <xs:enumeration value="master"/>
                        <xs:enumeration value="work"/>
                        <xs:enumeration value="other"/>
                </xs:restriction>
        </xs:simpleType>

        <xs:complexType name="event">
                <xs:sequence>
                        <xs:element name="title" type="customString" minOccurs="1"
                            maxOccurs="1"/>
                        <xs:element name="date" type="customDate" minOccurs="1"
                            maxOccurs="1"/>
                        <xs:element name="start" type="xs:time" minOccurs="1"
                            maxOccurs="1"/>
                        <xs:element name="end" type="xs:time" minOccurs="0" maxOccurs
                            ="1"/>
                        <xs:element name="place" type="xs:string" minOccurs="0"
                            maxOccurs="1"/>
                        <xs:element name="persons" type="personsType" minOccurs="0"
                            maxOccurs="1"/>
                        <xs:element name="description" type="descriptionType"
                            minOccurs="1" maxOccurs="1"/>
                </xs:sequence>
                <xs:attribute name="type" type="categories" use="required"/>
        </xs:complexType>
        ...

        <xs:element name="events">
                <xs:complexType>
                        <xs:sequence>
                                <xs:element name="event" type="event" maxOccurs="
                                    unbounded"/>
                        </xs:sequence>
                </xs:complexType>
        </xs:element>
</xs:schema>
```

Figura 3.5: Ejemplo de XML Schema.

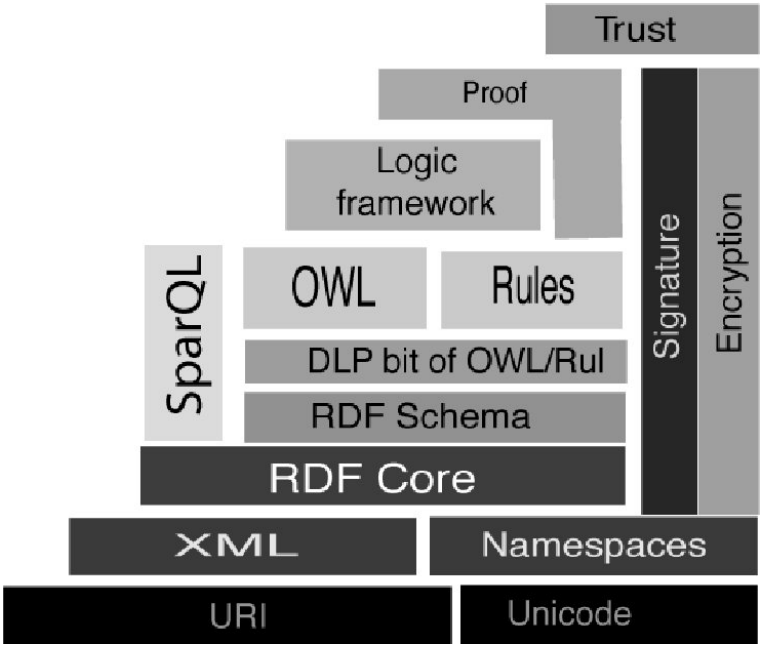


Figura 3.6: Arquitectura Web Semántica 2005.

La propuesta de cambio surge porque la Lógica Descriptiva, que utiliza OWL, no tiene un poder expresivo suficiente para solucionar el problema de la representación de conocimiento y razonamiento en la Web Semántica. OWL no puede tratar con información de una forma dinámica, no tiene predicados *n-arios*, no soporta símbolos de función, etc. De ahí que se valorara la posibilidad de extender las bases de conocimiento en OWL con reglas para aumentar la expresividad de los modelos y ganar en capacidad de inferencia.

Las reglas, en programación lógica como PROLOG, tienen una larga tradición en las ciencias de la computación y se llevan estudiando desde hace más de 20 años, sin embargo, la extensión de OWL con reglas no es una tarea sencilla. En general, las reglas están basadas también en un subconjunto de la lógica de primer orden, la lógica *Horn*, aunque su semántica formal es diferente, no está basada en teoría de modelos. La diferencia fundamental, aparte del tratamiento de la negación y cuestiones de expresividad, tiene que ver con OWL, ya que, debido a su semántica de primer orden, utiliza la hipótesis de mundo abierto, mientras que las reglas habituales de programación lógica utilizan mundo cerrado.

La extensión más conocida de OWL en este sentido es SWRL [169], que permite cláusulas Horn como axiomas en las bases de conocimiento en OWL. Este lenguaje mantiene el mundo abierto de OWL, pero penaliza en rendimiento debido a su alto coste computacional.

Debido a esta serie de problemas, se ha sugerido la opción de separar ontologías y reglas (lógica descriptiva y programación lógica) y utilizar cada tecnología en los escenarios pertinentes y apropiados. La combinación entre ambos, un objetivo verdaderamente ambicioso de la Web Semántica, ya no sería mediante la extensión de OWL con algún tipo de formalismo para la expresión de reglas, sino mediante la construcción de un interfaz lógico que permita desde las reglas utilizar información definida en las ontologías, y desde éstas, información inferida por el comportamiento de los sistemas de reglas.

Destacar la aparición del lenguaje de consulta para RDF, SPARQL [259] y su última versión SPARQL 1.1, creado por la necesidad de disponer de un lenguaje de acceso a los recursos definidos en forma de grafo y con una semántica definida [192]. El modelo semántico [156] proporcionado por RDF permite tratar cualquier recurso como una entidad con descripción asociada. No obstante, el uso de RDF no es suficiente para la elaboración de modelos de dominio más ricos y descriptivos, razón

por la cual surge un lenguaje como OWL que apoyado sobre el modelo de RDF permite realizar formalizaciones utilizando diferentes lógicas (principalmente *Description Logics*).

3.1.2.3 Lenguajes para la Web Semántica: Creando ontologías

La construcción de una base de conocimiento [145] utilizando ontologías puede realizarse con distintos lenguajes y diferentes grados de expresividad lógica [170, 194]. La selección de la lógica apropiada para la modelización de nuestra base de conocimiento no es una cuestión sencilla y deben contemplarse diferentes factores: grado de computabilidad, decidibilidad, soporte de los razonadores, etc. Todos ellos determinarán la lógica a utilizar ya que no se puede señalar arbitrariamente que se debe utilizar un nivel lógico cuando no es necesario para el modelo formal.

3.1.2.4 RDF

El primer lenguaje que se encuentra para la creación de ontologías es RDF (*Resource Description Framework*) como soporte básico a la par que potente, para añadir semántica a los recursos (documentos entre otros). La primera observación a realizar es la distinción entre: 1) modelo de datos, basado en tripletas Sujeto-Predicado-Objeto y 2) formato de datos, puede utilizar RDF/XML (normativo) como formato para la serialización del modelo.

El modelo de datos de RDF utiliza tripletas, ver Figura 3.7, encargadas de describir recursos:

Sujeto: recurso sobre el que vamos a realizar una afirmación. Están identificados de forma única a través de URIs, por ejemplo: Yo.

Predicado: es la afirmación sobre el sujeto, por ejemplo: "tengoNombre".

Objeto: valor del predicado para este sujeto, por ejemplo: "Jose María"@es.

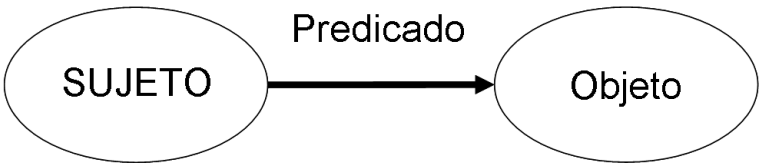


Figura 3.7: Modelo de tripletas RDF.

Usando RDF pueden realizarse afirmaciones simples sobre "cosas": "Este documento de memoria tiene de creador" a "Jose María Alvarez" puede expresarse en RDF usando la tripleta (memoria, "tieneCreador", "Jose María Alvarez"). A su vez se pueden generar nuevas tripletas (memoria, "tieneFecha", "Enero 2012") o (memoria, "tieneFormato", "PDF"), generando así un modelo semántico simple pero perfectamente válido. Si se le añade el uso de colecciones y la capacidad de "reificación" (afirmaciones sobre otras afirmaciones) se conseguirá una capacidad de expresión muy potente asentada en un sencillo lenguaje. En cuanto, a la serialización de RDF o su formato, existen diferentes estándares perfectamente válidos para su tratamiento automático por las máquinas y más o menos amigables para los humanos.

- RDF/XML [16] formato estándar y normativo por excelencia, ver Figura 3.9.
- Otros como: RDFa [2], Turtle [17], N3 [26] ,RDF/JSON [279] o RDF binario [117].

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:dc="http://purl.org/dc/elements/1.1/">
    <rdf:Description rdf:about="http://petra.euitio.uniovi.es/~i1637566/">
      <dc:creator>Jose M. Alvarez</dc:creator>
    </rdf:Description>
</rdf:RDF>
```

Figura 3.8: Ejemplo de tripletas de RDF en RDF/XML.

```
@prefix dc <http:// http://purl.org/dc/elements/1.1/>
<http://petra.euitio.uniovi.es/~i1637566/> dc:creator "Jose M. Alvarez"
```

Figura 3.9: Ejemplo de tripleta de RDF en N3.

En definitiva, RDF provee un mecanismo muy útil para describir recursos que no realiza presunciones sobre un dominio y capaz de representar a cualquiera de ellos. La declaración de propiedades (atributos) y su valor semántico se definen en el contexto de RDF con RDFS (*RDF Schema*). RDFS se construye por encima de RDF y sirve, no solo para definir las propiedades del recurso, sino también el tipo de recurso. Se pueden crear *clases de recursos*, restringir combinaciones de las clases, de las relaciones, además de ser el primer nivel de semántica que permite detectar estas aserciones. RDFS está basado en el metamodelado de objetos, el principal problema lo constituye la posibilidad de que una misma clase pueda desarrollar un doble rol de clase o de instancia, aunque se puede utilizar como un lenguaje de ontologías desde el punto de vista del manejo de clases, propiedades, rangos y dominios sobre propiedades. La posibilidad de generación de jerarquía de conceptos es un lenguaje muy limitado para la expresión de datos en detalles y no asegura la computabilidad. Es interesante conocer algunos de los vocabularios RDF, especialmente con el advenimiento de la iniciativa de *Linked Data*, que se han creado con distinto propósito, con muy buena acogida en la comunidad de Internet y cada día con un uso más extendido:

Dublin Core. Vocabulario RDF con las propiedades más comunes y semántica bien definida para el etiquetado de cualquier documento.

- Cada etiqueta es opcional y puede estar repetida. Un documento no necesita tener resumen y puede tener varios autores.
- La mayoría de etiquetas tienen cualificadores para refinar (nunca extender) su significado. Por ejemplo, la etiqueta "fecha" tendrá como cualificadores "de publicación", "de creación", etc.
- Principio del Uno-a-Uno. Los metadatos se refieren a un documento concreto, no a lo representado.
- Principio del *Dumb-down*. Si no se procesan las restricciones al significado de las propiedades, estas deben seguir proporcionando información útil. Se pierde nivel de detalle, pero la información sigue siendo válida.
- Las buenas prácticas de uso para una etiqueta determinada pueden variar por el contexto. El creador no puede dar por supuesto que serán interpretadas exclusivamente por una máquina, esto impondrá algunas restricciones a cómo se construyen los metadatos, pero no se debe eludir que el requisito fundamental de estos es su utilidad para descubrir información.

Usando estos principios, *Dublin Core* se impuso el objetivo de lograr un lenguaje de etiquetado: simple y fácil de mantener, semántica esencial y de significado común, alcance internacional y extensible. Se pueden establecer dos escenarios muy habituales en el uso de este vocabulario:

1. En los metaelementos presentes en HTML y XHTML, ver Figura 1.

```
  <head>
<title>Dublin Core Metadata Initiative (DCMI)</title>
<link rel="schema.DC" href="http://purl.org/dc/elements/1.1/" />
<meta name="DC.title" content="Dublin Core Metadata Initiative (DCMI) Home Page" />
<meta name="DC.description" content="The Dublin Core Metadata Initiative is an open
    forum . . . metadata standards and practices." />
<meta name="DC.date" content="2011-10-01" />
<meta name="DC.format" content="text/html" />
<meta name="DC.contributor" content="Dublin Core Metadata Initiative" />
<meta name="DC.language" content="en" />
</head>
```

Figura 3.10: Ejemplo de *Dublin Core* en HTML/XHTML.

2. En documentos RDF propiamente dichos como metadatos de los recursos. A continuación, se puede ver en la Figura 2 el uso de *Dublin Core* en combinación con RSS.

```
<rdf:RDF
 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns="http://purl.org/rss/1.0/"
 xmlns:dc="http://purl.org/dc/elements/1.1/"
 xmlns:syn="http://purl.org/rss/1.0/modules/syndication/"
 xmlns:taxo="http://purl.org/rss/1.0/modules/taxonomy/"
>
<channel rdf:about="http://dublincore.org">
<title>Dublin Core Metadata Initiative</title>
<link>http://dublincore.org/</link>
<description>Making it easier to find information.</description>
<dc:language>en-us</dc:language>
<dc:rights>1995-2007</dc:rights>
<dc:date>2007-10-01</dc:date>
<items>
<rdf:Seq>
  <rdf:li rdf:resource="http://dublincore.org/news/2007/#dcmi-news-20071001-01" />
  <rdf:li rdf:resource="http://dublincore.org/news/2007/#dcmi-news-20071001-02" />
  <rdf:li rdf:resource="http://dublincore.org/news/2007/#dcmi-news-20071001-03" />
  <rdf:li rdf:resource="http://dublincore.org/news/2007/#dcmi-news-20071001-04" />
</rdf:Seq>
</items>
</channel>

</rdf:RDF>
```

Figura 3.11: Ejemplo de *Dublin Core* con RDF.

SKOS-Core [217]. Vocabulario RDF utilizado para la representación de conocimiento de forma simple a través de conceptos, para ser procesado por máquinas de forma automática: vocabularios controlados, taxonomías, tesauros, esquemas conceptuales, glosarios o esquemas categoriza-

dos. Algunas de las propiedades de SKOS que confieren interés a este vocabulario son las siguientes:

- Identificación de conceptos a través de URIs.
- Etiquetado de conceptos: *prefLabel*, *altLabel*, *prefSymbol*, *altSymbol*, etc.
- Descripción y documentación (conceptos): *definition*, *example*, *scopeNote*, *version*, *changeNote*, etc.
- Relaciones entre conceptos: *broader*, *narrower*, *related*, etc.
- Indexación: *subject*.
- Soporte multilingüe.

En la siguiente Figura 3.12 se presenta un ejemplo más completo del uso de SKOS-Core para la descripción de conceptos:

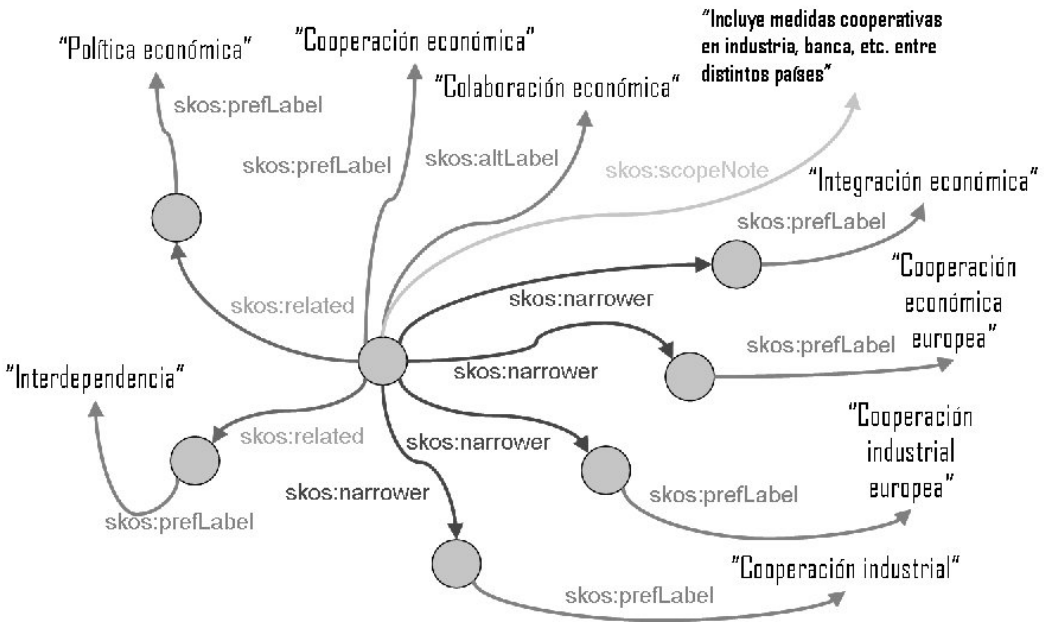


Figura 3.12: Concepto expresado en SKOS-Core.

Debe puntualizarse que el uso de SKOS-Core está siendo objeto de estudio para la representación de recursos lexicográficos [10] desde un punto de vista de la lingüística y como un avance tecnológico en este campo.

FOAF y DOAP. Vocabularios RDF para definir personas y amistades (*Friend Of A Friend*), ver Figura 3.13, o proyectos (*Description Of A Project*), ver Figura 3.14. Cada vez hay un mayor interés por este tipo de tecnologías y algunos proyectos de gran envergadura como Apache o RDF Ohloh [118], que incluyen muchos subproyectos, programas, etc., utilizan estas descripciones semánticas para cada módulo.

SIOC. (*Semantically-Interlinked Online Communities*). Es una ontología desarrollada por el equipo de Web Semántica de DERI Galway para describir semánticamente distintas comunidades *online*. SIOC integra, ver Figura 3.15, distintos vocabularios RDF con el objetivo de reutilizar las definiciones ya realizadas en FOAF, SKOS, RSS o *Dublin Core*. Se ha empleado con éxito en distintas aplicaciones para gestionar y describir toda la información disponible de las comunidades *online*: listas de correo (SWAML [206]), foros, etc.

```
<#me> a foaf:Person;
        foaf:family_name "Alvarez Rodr\u00EDguez";
        foaf:givenname "Jose Mar\u00EDa";
        foaf:homepage <http://josemalvarez.es>;
        foaf:knows _:bnode2016979200;
        foaf:mbox_sha1sum "0d1d9ad2de64fd900d03c18e3d2608171832d155";
        foaf:name "Jose Mar\u00EDa Alvarez Rodr\u00EDguez";
        foaf:nick "chema";
        foaf:phone <tel:+34-666-714-721>;
        foaf:schoolHomepage <http://www.uniovi.es/inicio/>;
        foaf:title "Sr.";
        foaf:workplaceHomepage <http://www.weso.es>.
_:bnode2016979200 a foaf:Person;
        rdfs:seeAlso <http://www.di.uniovi.es/~labra/labraFoaf.rdf>;
        foaf:mbox_sha1sum "5fa5d69bac0c1396825c475ec19325ec0ffd5569";
        foaf:name "Jose Emilio Labra".
```

Figura 3.13: Ejemplo parcial de documento FOAF en N3.

```
<http://rdfohloh.wikier.org/project/moldeas/rdf>
        dct:isFormatOf  <http://rdfohloh.wikier.org/project/moldeas>;
        a foaf:Document;
        rdfs:label "MOLDEAS's DOAP document serialized in RDF/XML";
        foaf:primaryTopic <http://rdfohloh.wikier.org/project/moldeas>.
        <http://rdfohloh.wikier.org/project/moldeas> dct:updated "2012-01-22T13:02:25
            Z";
        rdfohloh:ohloh-page <http://www.ohloh.net/projects/moldeas>;
        doap:created "2011-10-14T09:19:11Z";
        doap:description "This work aims to apply the semantic web and
            LOD approaches to public procurement notices...";
        doap:download-page <http://code.google.com/p/moldeas/downloads/list>;
        doap:homepage <http://purl.org/weso/moldeas/>;
        doap:name "MOLDEAS";
        doap:programming-language "JavaScript";
        a doap:Project;
        = <http://rdfohloh.wikier.org/project/586667>;
        skos:subject <http://dbpedia.org/resource/Java>,
        <http://dbpedia.org/resource/JavaScript>.
```

Figura 3.14: Ejemplo parcial de documento DOAP en N3.

```
<rdf:RDF
  xmlns:sioc='http://rdfs.org/sioc/ns#'
  xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
  xmlns:dc='http://purl.org/dc/elements/1.1/'
  xmlns:mvcb='http://webns.net/mvcb/'
>
  <sioc:Site rdf:about="http://groups.google.com/group/sioc-dev">
    <sioc:host_of>
      <sioc:Forum rdf:about="http://swaml.berlios.de/demos/sioc-dev/index.rdf#SIOC-
          Dev">
        <dc:description>SIOC development mailing list</dc:description>
        <mvcb:errorReportsTo rdf:resource="http://swaml.berlios.de/bugs"/>
        <sioc:container_of rdf:resource="http://swaml.berlios.de/demos/sioc-dev/2006-
            Dec/post-419.rdf"/>
        <dc:date>2007-06-18</dc:date>
        <dc:title>SIOC-Dev</dc:title>
        <mvcb:generatorAgent rdf:resource="http://swaml.berlios.de/doap.rdf"/>
        <sioc:has_host rdf:resource="http://groups.google.com/group/sioc-dev"/>
        <sioc:has_subscriber rdf:resource="http://swaml.berlios.de/demos/sioc-dev/
            subscribers.rdf#s26"/>
      </sioc:Forum>
    </sioc:host_of>
  </sioc:Site>
</rdf:RDF>
```

Figura 3.15: Ejemplo de descripción con SIOC.

RSS. Existen diferentes versiones y definiciones de este vocabulario RDF, ver Figura 3.1.2.4: *Rich Site Summary* (RSS 0.91), XML; *RDF Site Summary* (RSS 0.9 y 1.0), RDF; *Really Simple Syndication* (RSS 2.0). La definición realizada en el W3C es la siguiente:

> *Vocabulario RDF basado en XML que permite la catalogación de información (noticias y eventos) de los usuarios. Los archivos RSS contienen metadatos sobre fuentes de información especificadas por los usuarios, cuya función principal es notificar de forma automática cualquier cambio que se realice en esos recursos de interés.*

En resumen, se dispone de un lenguaje muy útil y sencillo para describir recursos (RDF) y un esfuerzo en forma de vocabulario para la definición de recursos más detallada (RDFS) pero incompleto. Por ello, a continuación, se expondrán algunos de los lenguajes más completos que han surgido para intentar mejorar los puntos débiles de esta primera aproximación para la expresión de modelos semánticos. En cuanto a vocabularios RDF existen una infinidad [258] de ellos para ser utilizados en distintos contextos que han sido enormemente impulsados por la corriente *Linked Data* y *Open Data*.

```
<?xml version="1.0" encoding="iso-8859-1"?>

<rdf:RDF
 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns="http://purl.org/rss/1.0/"
 xmlns:slash="http://purl.org/rss/1.0/modules/slash/"
 xmlns:taxo="http://purl.org/rss/1.0/modules/taxonomy/"
 xmlns:dc="http://purl.org/dc/elements/1.1/"
 xmlns:syn="http://purl.org/rss/1.0/modules/syndication/"
 xmlns:admin="http://webns.net/mvcb/"
>

<channel rdf:about="http://barrapunto.com/">
<title>Barrapunto</title>
<link>http://barrapunto.com/</link>
<description>La informaci&#243;n que te interesa</description>
<dc:language>EN</dc:language>
<dc:rights>Copyleft 1999-2006, Barrapunto</dc:rights>
<dc:date>2007-10-04T10:10:15+00:00</dc:date>
<dc:publisher>Barrapunto S.L.</dc:publisher>

<dc:creator>webmaster@barrapunto.com</dc:creator>
<dc:subject>Tecnolog&#237;a, Software libre</dc:subject>
<syn:updatePeriod>hourly</syn:updatePeriod>
<syn:updateFrequency>1</syn:updateFrequency>
<syn:updateBase>1970-01-01T00:00+00:00</syn:updateBase>
<items>
 <rdf:Seq>
  <rdf:li rdf:resource="http://ciencia.barrapunto.com/article.pl?sid
      =07/10/04/0958226&amp;from=rss" />
  <rdf:li rdf:resource="http://barrapunto.com/article.pl?sid=07/10/04/0754202&amp;
      from=rss" />
  <rdf:li rdf:resource="http://softlibre.barrapunto.com/article.pl?sid
      =07/10/04/023251&amp;from=rss" />
  <rdf:li rdf:resource="http://barrapunto.com/article.pl?sid=07/10/04/0152245&amp;
      from=rss" />
  <rdf:li rdf:resource="http://espana.barrapunto.com/article.pl?sid=07/10/03/1638217&
      amp;from=rss" />
  <rdf:li rdf:resource="http://barrapunto.com/article.pl?sid=07/10/03/1124232&amp;
      from=rss" />
  <rdf:li rdf:resource="http://entrevistas.barrapunto.com/article.pl?sid
      =07/10/03/1113255&amp;from=rss" />
  <rdf:li rdf:resource="http://barrapunto.com/article.pl?sid=07/10/03/1050215&amp;
      from=rss" />
  <rdf:li rdf:resource="http://barrapunto.com/article.pl?sid=07/10/03/1033250&amp;
      from=rss" />
  <rdf:li rdf:resource="http://barrapunto.com/article.pl?sid=07/10/03/0946239&amp;
      from=rss" />
 </rdf:Seq>

</items>
<image rdf:resource="http://barrapunto.com/images/topics/topicbarrapunto.png" />
<textinput rdf:resource="http://barrapunto.com/search.pl" />
</channel>

</rdf:RDF>
```

Figura 3.16: Ejemplo de canal RSS.

3.1.2.5 OIL

El *Ontology Inference Layer* [115] (OIL) desarrollado en el proyecto Europeo *OntoKnowledge* está construido por encima de RDF y RDFS utilizando muchos de sus constructores e intentando mantener la compatibilidad hacia atrás. OIL provee características de modelado basada en lógica de marcos [194] y *Description Logic* [11].

La importancia de OIL reside en la unificación de:

- *Description Logic*, heredando su semántica formal y la capacidad de razonamiento efectivo (FaCT++, Racer o Pellet).
- Sistemas basados en marcos, incorpora las características básicas de modelado basado en marcos (conceptos, superclases, atributos, etc.).
- Estándares web, construido sobre RDF y RDFS y utilizando como formato de intercambio de datos XML.

Las ontologías creadas con OIL distinguen distintos meta niveles, con el objetivo de proporcionar distintos niveles de servicio que pueden ser válidos para un gran número de ontologías y aplicaciones:

Primer meta nivel: también conocido como *definición de ontologías*, en la cual se proveen las definiciones de ontología, terminología que debería ser instanciada para definir un vocabulario estructurado con la semántica adecuada.

Segundo meta nivel: conocido como "meta meta" nivel o *contenedor de ontologías*, describe las características de la ontología tales como autor, nombre, ámbito, etc. Las anotaciones de este nivel se hacen mediante *Dublin Core*.

La capacidad de razonamiento es otra de las características importantes de OIL, que no estaba presente en RDF. Habitualmente esta capacidad se utiliza para realizar las operaciones de clasificación, validación de consistencia e inferencia de nuevo conocimiento de acuerdo a distintos niveles de expresividad:

Core OIL: prácticamente compatible con RDFS exceptuando la capacidad de "reificación".

Standard OIL: lenguaje que captura las primitivas y constructores necesarios para construir modelos semánticos con capacidad de inferencia.

Instance OIL: inclusión de "instancias".

Heavy OIL: capacidades extras de representación y razonamiento.

Cabe concluir, por tanto, que las principales ventajas de OIL se resumen en: 1) una aplicación no está obligada a trabajar con un lenguaje más expresivo de lo necesario; 2) aplicaciones que utilicen la expresividad más baja pueden añadir aspectos de otras ontologías y 3) aplicaciones con un alto nivel de complejidad pueden utilizar características de otras más simples. Finalmente, el trabajo realizado en OIL tiene su continuación en el siguiente lenguaje de modelado de ontologías DAML+OIL realizado por la cooperación de iniciativas europeas y americanas.

3.1.2.6 DAML+OIL

DAML+OIL [158] es un lenguaje de marcado semántico para recursos web creado por la cooperación de las iniciativas europeas (OIL) y americanas (DAML-ONT, DARPA *Agent Markup Language*) en el desarrollo de ontologías. El objetivo de desarrollo de este lenguaje está especialmente centrado para la Web Semántica, modelización de dominios concretos, utilizando los estándares (XML y RDF) y añadiendo una serie de primitivas de orientación a objetos (clases, propiedades, axiomas y aserciones), sistemas basados en marcos y parte de *Description Logic*. Además, DAML+OIL da soporte a los tipos de datos de XML Schema, separando las instancias de una clase y de las instancias de tipos de datos. El significado de DAML+OIL está definido por un modelo semántico estándar basado en las interpretaciones, consistiendo éstas en un dominio del discurso y una función de interpretación.

3.1.2.7 OWL

OWL, lenguaje de ontologías para la web, sucesor de DAML+OIL, actualmente está siendo desarrollado por el W3C, su primera versión estable es OWL 1.0, aunque ya se ha publicado una nueva versión OWL 1.1 [248] y OWL2 [164].

Como lenguaje para ontologías es una potente herramienta que dota de la expresividad necesaria para mejorar algunos de los servicios más utilizados en la red de Internet: búsqueda, manejo del conocimiento, interacción entre agentes automáticos etc. Desde un punto de vista más formal, OWL está formado por un conjunto de primitivas o constructores de metamodelado que son el punto de partida para operaciones más complejas, como las de razonamiento. Actualmente, aunque las características genéricas de OWL tienen su origen en OIL, se distinguen al menos 3 versiones de OWL (en su versión 1.0 y 1.1) dependiendo de su expresividad, complejidad y grado de computabilidad, cada versión superior de OWL contiene a la anterior ($Lite \subset DL \subset FULL$).

OWL-FULL. Se pueden utilizar todos los constructores y primitivas definidos en OWL y no restringe el uso de RDF. Esto implica que no se garantiza su decidibilidad pero en cambio, como familia de *Description Logics* posee un gran poder expresivo: tratar clases como instancias, definir propiedades sobre tipos de datos (*string*, *float*, etc.) .

OWL-DL. Subconjunto decidible de OWL-FULL, supone la máxima expresividad computable. Cada modelo realizado en OWL-DL genera directamente un modelo semántico en *Description Logics*.

OWL-Lite. Añade restricciones adicionales en el uso de los constructores de OWL, básicamente se pueden modelar jerarquías con restricciones sencillas.

Para advertir a que nivel de expresividad y complejidad se puede trabajar dependiendo del lenguaje utilizado se dispone la Figura 3.17, extraída del laboratorio OntoText.

Algunas de las características que hacen interesante el uso de OWL en el ámbito de la Web Semántica son:

- Las ontologías en OWL son una serie de axiomas y hechos que pueden reutilizar otras ontologías (importándolas). Además, como documento que son (serializado en RDF/XML), están identificadas por un URI y tienen asociada cierta metainformación (autor, dominio, fecha, etc.) que permite que sean referenciables como cualquier otro recurso en la red.
- Los axiomas son utilizados para asociar una serie de características (descripciones, restricciones, etc.) a las clases y propiedades de la ontología.
- Los hechos proporcionan información particular de una determinada instancia.

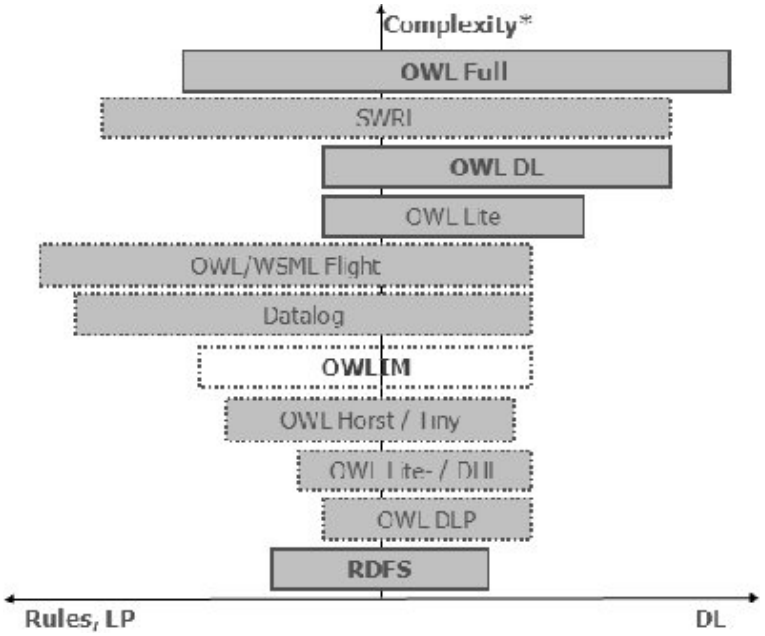


Figura 3.17: Algunos lenguajes para la Web Semántica por OntoText.

Como ejemplo de utilización de OWL, en su versión DL y con características de OWL2, se presenta una sencilla ontología, ver Figura 3.18, que modela un sistema de diagnóstico psicológico con capacidad de clasificar individuos de acuerdo a sus síntomas, ver Figura 3.19, con sintaxis Manchester.

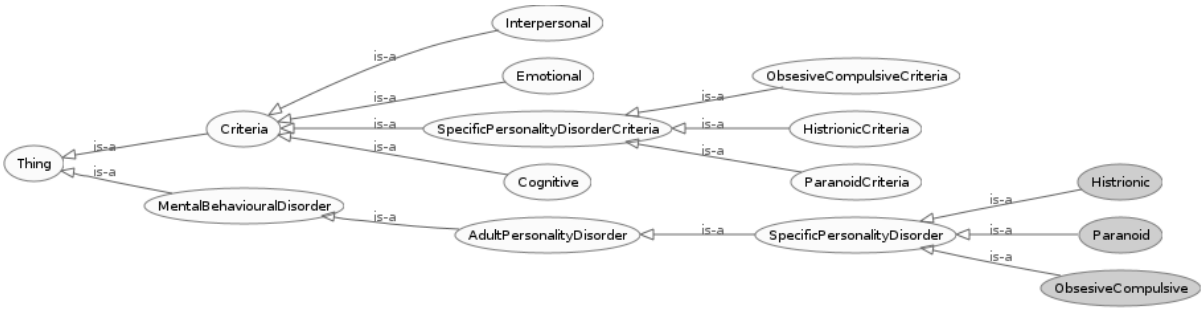


Figura 3.18: Ontología de ejemplo en OWL2 para diagnóstico psicológico.

3.1.2.8 WSML

WSML [91] es la propuesta de lenguajes formales realizado en WSMO [302] (relevante por su relación con los Servicios Web Semánticos) para la construcción de ontologías y que recoge distintas variantes de lógica. En general, teniendo en cuenta los distintos tipos de lógica de acuerdo a su expresividad es posible obtener diferentes serializaciones de los modelos realizados con distintos lenguajes OWL, WSML, F-Logic, etc. Esta característica es muy interesante para el uso de razonadores con distintos tipos de algoritmos y capacidades de razonamiento, independientemente del lenguaje en el que se haya modelado el dominio.

```
Class: <http://example.org/psydiag.owl/Histrionic>

    EquivalentTo:
        <http://example.org/psydiag.owl/has-criteria> min 3
              <http://example.org/psydiag.owl/HistrionicCriteria>,
         (not (<http://example.org/psydiag.owl/has-criteria> some
              <http://example.org/psydiag.owl/ObsesiveCompulsiveCriteria>))
          and (not (<http://example.org/psydiag.owl/has-criteria> some
              <http://example.org/psydiag.owl/ParanoidCriteria>))
          and (<http://example.org/psydiag.owl/has-criteria> only
              <http://example.org/psydiag.owl/HistrionicCriteria>)

    SubClassOf:
        <http://example.org/psydiag.owl/SpecificPersonalityDisorder>

DisjointClasses:
    <http://example.org/psydiag.owl/Histrionic>,
    <http://example.org/psydiag.owl/ObsesiveCompulsive>,
    <http://example.org/psydiag.owl/Paranoid>
```

Figura 3.19: Algunos axiomas de ejemplo en OWL2 para diagnóstico psicológico.

3.1.3 Ontologías

3.1.3.1 Antecedentes y Definición

El desarrollo de las ontologías se deriva directamente de la filosofía, Aristóteles acuña el término "Categoría" como palabra para describir las diferentes clases en las que se dividían las cosas del mundo. El término "ontología" es relativamente moderno (siglo XIX), proviene del griego *Ontos* (Ser) y *Logos* (Palabra), se empezó a utilizar para diferenciar el estudio de la Teoría de Categorías que se hacía en biología, de hecho, el trabajo de categorización surge en muchas áreas de la ciencia (filosofía, biología, medicina, lingüística, etc.).

El cuerpo de un esquema de conocimiento, *lo* que se puede representar, está basado en una conceptualización: objetos, conceptos y otras entidades, que se asume existen en un determinado dominio y que mantienen unas relaciones entre sí.

Una ontología es una especificación (explícita) de una conceptualización [145].

Las ontologías, por consiguiente, son sistemas basados en el conocimiento (*SBC*) que sirven como modelo unificado en la representación del mismo, adquiriendo incluso grado de ingeniería [18, 137]. En muchos casos, la Web Semántica y por extensión las ontologías se apoyan en la reutilización de conocimiento compartido. Las definiciones realizadas para las ontologías son variadas y todas dejan claro que sirven para modelar cierto dominio (conjunto de conceptos y relaciones):

- Una ontología define los términos básicos y las relaciones establecidas en cierta área.
- Especificación explícita de una conceptualización.
- Especificación formal y explícita de una conceptualización compartida.
- Teoría lógica que provee de forma explícita y parcialmente una conceptualización.
- ...

En el artículo [282] los autores agregan expresividad realizando la siguiente descripción completa de ontología:

- Conceptualización, modelo abstracto de algún fenómeno del mundo, proveniente de la identificación de los conceptos relevantes de dicho fenómeno.
- Explícita, conceptos y restricciones usados se definen explícitamente.
- Formal, capacidad de ser legible e interpretable por las máquinas.
- Compartida, captura conocimiento consensuado.

Es posible fusionar las definiciones anteriores en nueva descripción del término ontología:

Definición 3.1.2. *Modelo conceptual organizado mediante una taxonomía que permite definir relaciones entre conceptos, funciones, instancias (elementos) y axiomas en un determinado dominio.*

Para completar la definición de ontología hay que tener en cuenta la nomenclatura que habitualmente se utiliza para nombrar a las distintas entidades posibles y que en algunos casos puede dar lugar a errores:

- Clase, concepto, categoría o tipo.
- Instancia, individual.
- Entidad, objeto (clase o instancia).
- Propiedad, relación, slot, atributo, rol.

En la actualidad, la construcción de sistemas basados en conocimiento conlleva la creación partiendo de cero de nuevas bases de conocimiento. La aspiración debería ser utilizar componentes reutilizables [146], de esta manera los desarrolladores deberían crear sistemas con la agregación del conocimiento ya existente. El conocimiento definido, las técnicas de resolución de problemas y otros servicios, podrían ser compartidos entre varios sistemas, impulsando la creación de grandes sistemas con un bajo coste. Desde esta perspectiva, se pueden usar las ontologías como infraestructura para sistemas ubicuos.

3.1.3.2 Componentes

Una ontología consta de un conjunto no vacío de conceptos identificados como relevantes en el dominio a modelar, un conjunto de atributos para describir los conceptos que pueden proveer de distintas fuentes: propios, heredados, etc., un conjunto de funciones, un conjunto de axiomas que formalizan las condiciones que deben cumplir los distintos conceptos y un conjunto de instancias o realizaciones particulares de los conceptos.

Conceptos: cualquier entidad que se puede describir, tiene asociado un identificador único, puede poseer diferentes atributos y establecer relaciones con otros conceptos.

Relaciones: representan la interacción entre los conceptos de dominio. Formalmente, se definen como subconjuntos del producto cartesiano de n conjuntos $R : C_1 \times C_2 \times \ldots \times C_n$.

No todas las relaciones tienen el mismo significado, existen relaciones binarias de especialización como (*is-a*) o de composición (*part-whole*), que se modelan con las propiedades clásicas simétricas, reflexivas, etc.

Funciones: relaciones en las cuales el elemento *n-ésimo* es único para los $n-1$ anteriores. Formalmente, se definen como $F : C_1 \times C_2 \times \ldots \times C_n$. Por ejemplo una relación *serPadreDe* se puede modelar como una función ya que el atributo que evalúa es único para cada caso.

Axiomas: modelan "verdades" que siempre se cumplen en el modelo. Existen dos tipos de axiomas:

- Estructurales, condiciones relacionadas con la estructura jerárquica de la ontología.
- No estructurales, establecen relaciones entre atributos de un concepto y son específicos de cada dominio.

Instancias: representan realizaciones específicas del dominio de la ontología.

3.1.3.3 *DescriptionLogics* y ontologías

Description Logics [11] (DLs) son un conjunto de lógicas formales en el área de *Knowledge Representation*, utilizadas para la representación y razonamiento del conocimiento en un dominio de forma no ambigua. Las DLs se basan en una semántica perfectamente definida que provee un conjunto de constructores y primitivas con un significado lógico preciso.

Los pilares de estas lógicas son dos conjuntos: uno de ellos, *atomic concepts* o predicados unarios y otro, *atomic roles* o predicados binarios. Una DL provee además un conjunto de operadores, llamados *constructores*, que permiten crear conceptos y roles más complejos a partir de los más sencillos. Tanto los conceptos atómicos como los complejos, se denominan uniformemente *conceptos* y de igual forma, los roles atómicos y los complejos se denominan *roles*.

Los conceptos se utilizan para representar conjuntos de objetos y los roles sirven para establecer relaciones binarias entre objetos. Podemos definir el conjunto de los números reales como la disyunción de números racionales e irracionales

Ejemplo 3.1.3. $\mathbb{R} \equiv \mathbb{Q} \sqcup \mathbb{I}$

En general, una base de conocimiento basada en DL consiste en:

- *TBox*, contiene los axiomas de inclusión de conceptos $C_1 \sqsubseteq C_2$.
- *RBox*, contiene los axiomas de inclusión de roles $R_1 \sqsubseteq R_2$.
- *Abox*, contiene axiomas (aserciones sobre conceptos) $C(a)$, las aserciones sobre los roles $R(a, b)$, a y b son son nombre de objetos, R es un rol y C es un concepto.

Los constructores booleanos de conceptos son la $\sqcup$ (disyunción o unión), $\sqcap$ (conjunción o intersección) y $\neg$ (negación). Una DL que provee, implícita o explícitamente, todos los operadores booleanos se considera *cerrada* (proposicionalmente). Las DLs "cerradas" serán las que sean interesantes para su procesamiento en la Web Semántica. Aparte de los operadores booleanos, habitualmente las DLs proveen otros constructores para generar conceptos complejos a partir de roles. En este apartado, se encuentran los operadores existencial ($\exists$) y universal $\forall$.

Las DLs que proveen estos cinco operadores se denominan $\mathcal{ALC}$, pero esta lógica no permite axiomas de inclusión en roles y la componente *RBox* es vacía, supone que si bien se pueden realizar operaciones de razonamiento, la lógica a este nivel es poco expresiva. Añadiendo nuevos constructores a $\mathcal{ALC}$ se obtiene el conjunto de lógicas $\mathcal{ALC}_{\mathcal{HR}^+}$ (también conocida como $\mathcal{SH}$), que resultan de añadir la inclusión de axiomas (permitiendo diferentes tipos), sobre la *RBox*. Esta familia de lógicas, ver Tabla 3.1 extraída de [142], es muy interesante porque posee un gran poder expresivo y puede ser probada sobre razonadores DL, como FaCT++, Racer, Pellet o Hermit.

Nombre constructor	Sintaxis	Lógica
Concepto atómico	A	
Concepto universal	$(\top)$	
Rol atómico	R	
Conjunción de conceptos	$C \sqcap D$	
Disyunción de conceptos	$C \sqcup D$	
Negación de concepto	$\neg C$	
Restricción existencial	$\exists R.C$	
Restricción universal	$\forall R.C$	
Rol transitivo	$Trans(R)$	$\mathcal{S}$
Jerarquía de roles	$R_1 \sqsubseteq R_2$	$\mathcal{H}$
Inversión de roles	(R^-)	$\mathcal{I}$
Nominales (instancias)	$\{o\}$	$\mathcal{O}$
Restricciones funcionales de número	$\geq 2S(\geq 1S)$	$\mathcal{F}$
Restricciones no cualificadas de número	$\geq nS(\leq nS)$	$\mathcal{N}$
Restricciones cualificadas de número	$\geq nS.C(\leq nS.C)$	$\mathcal{Q}$

Tabla 3.1: Familia de lógicas $\mathcal{SH}$.

La lógica DL es importante para la construcción de ontologías ya que permite construir bases de conocimiento formales, computables y no ambiguas. Aunque no siempre será indispensable este nivel de lógica, tanto para la descripción de dominios como para la Web Semántica, es necesario presentar y dar a conocer este conjunto de lógicas, para así validar los modelos y mantener unos criterios formales en la construcción de ontologías. Las ontologías construidas con lógica DL proporcionan una base sólida para el desafío de la Web Semántica.

3.1.3.4 Ontología como *SBC*

Como sistema basado en el conocimiento, ver Figura 3.20, y teniendo en cuenta la importancia de la utilización de *Description Logics* como lógica para la creación de ontologías, se pueden distinguir los tres componentes heredados de la definición de DL. Pero desde el punto de vista tanto del razonamiento como la inferencia, operaciones importantes en cualquier SBC, resultan de interés los siguientes componentes: 1) *Tbox*, parte terminológica (organizado jerárquicamente) o conocimiento definido por intensión, consistente en conceptos, roles y construcciones más complejas por combinación de éstos y 2) *Abox* o parte extensional, es decir, las afirmaciones sobre individuos"concretos". Sobre estas dos componentes se podrá realizar razonamiento, operación especialmente relevante para la Web Semántica, de dos formas:

Razonamiento *Tbox*, intensional o estructural: permite consultar la estructura de conocimiento e inferir información a partir de ella. El mecanismo de razonamiento estructural por excelencia es la subsunción de conceptos, permite calcular todos los subconceptos a partir de un concepto dado o consultar si un concepto es subconcepto de otro. Los razonamientos usuales en la *Tbox* son:

- Consistencia, comprueba si el conocimiento tiene o no sentido.

- Subsunción, comprobación de sí todos los individuos que pertenecen a un concepto (el subsumido) también pertenecen a otro concepto (el que subsume).
- Equivalencia, comprueba si dos clases denotan el mismo conjunto de instancias.

Todos estos razonamientos son aplicables al problema de la satisfacibilidad de fórmulas lógicas siempre que se utilice un lenguaje de definición de conceptos que sea cerrado con respecto a la negación.

Razonamiento *Abox* o extensional: permite inferir nuevas instancias a partir de las definidas de forma explícita en la *Abox*. Los razonamientos usuales en la *Abox* son:

- Comprobación de instancias, verifica que un determinado individuo es una instancia de un concepto específico.
- Consistencia de la base de conocimiento, implica verificar que cada concepto que existe en la base de conocimiento admite, al menos, una instancia o individuo.
- Realización, encuentra el concepto más especifico del que un individuo es instancia.

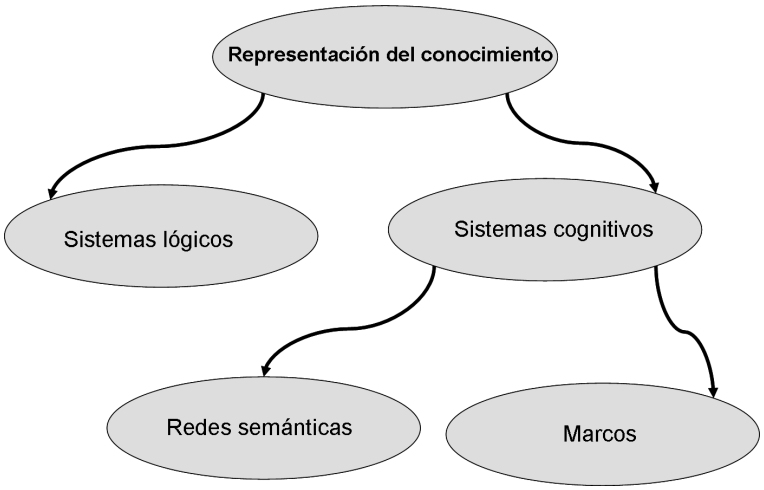


Figura 3.20: Sistemas basados en conocimiento.

3.1.3.5 Clasificación de ontologías

Las ontologías se pueden clasificar atendiendo a diferentes a criterios, a continuación se exponen algunos de ellos.

Grado de axiomatización. Atendiendo a Sowa [276] las ontologías se pueden clasificar en:

Terminológicas: define términos y sus relaciones en taxonomías que involucran tanto relaciones de subtipo y supertipo, como las que relaciona partes con un todo (*part-whole*), no incluyen axiomas y definiciones expresadas en lógica o lenguaje formal interpretable por una máquina. Por lo tanto, existe menos información sobre el dominio modelado pero la simplicidad de su especificación permite construir ontologías de gran tamaño.

Ejemplo 3.1.4. *WordNet base de conocimiento (y datos) léxica.*
Fuente: `http://wordnet.princeton.edu/`

Formales: consta de categorías restringidas por axiomas y definiciones expresadas en alguna lógica formal, con menos conceptos pero preparadas para soportar un procesamiento automático y servicios de razonamiento.

Una ontología terminológica podrá convertirse en formal a medida que se añadan axiomas, aunque esta evolución no es ni mucho menos trivial.

Según contexto. Dependiendo del grado de dependencia del contexto las ontologías pueden ser:

Dominio: modelan conceptualizaciones específicas, con restricciones de estructura y contenido del dominio. Se reaprovecharán sólo en aquellas aplicaciones que trabajen en ese dominio. Grado de dependencia alto.

Ejemplo 3.1.5. *Ontología de medicina Galen.*
Fuente: `http://www.opengalen.org/`.

Generales o de sentido común: vocabularios o estructuras taxonómicas genéricas. Alto grado de reaprovechamiento y bajo de dependencia del dominio.

Ejemplo 3.1.6. *Cyc y OpenCyc.*
Fuente: `http://www.cyc.com/`.

Metaontologías u ontologías genéricas: ontologías de dominio genérico, conceptos universales.

Ejemplo 3.1.7. *DOLCE: Descriptive Ontology for Linguistic and Cognitive Engineering.*
Fuente: `http://www.loa-cnr.it/DOLCE.html`.

Objeto de creación. Atendiendo al objetivo de creación de una ontología se establecen diferentes tipos:

- Ontologías para la representación de conocimiento, como: OWL, DAML+OIL, RDF, RDF(S), OKBC, etc.
- Ontologías *top level*, definen *frameworks*, conceptos universales como las ya comentadas *DOLCE* o *Cyc*.
- Ontologías para la definición de términos lingüisticos como *WordNet* o *EuroWordNet* [295].
- Ontologías de dominio, modelan conocimiento de cierto dominio: comercio electrónico o *e-commerce*, de medicina *Galen* o SNOMED, etc.

Las ontologías como sistemas basados en conocimiento, serán más productivos cuanto mejor "informados" estén, por ello, será interesante modelar utilizando el mayor de grado de particularidad posible en el dominio, pero teniendo presente, que también es importante mantener el objetivo de reutilizar conocimiento, podría ser interesante enclavar los nuevos conceptos de dominio definidos dentro de las categorías de alto nivel, para así alinear estas definiciones en un marco conceptual genérico.

3.1.3.6 Principios de diseño

Las ontologías deben o deberían cumplir una serie de principios:

Claridad: proporcionar el significado pretendido a los términos definidos. Definiciones tan objetivas como sea posible.

Completitud: las definiciones deberían recoger todas las condiciones necesarias y suficientes, además de poseer una buena documentación en lenguaje natural.

Coherencia: los términos definidos deben ser coherentes, concluyendo sólo aquellas inferencias consistentes con el modelo definido.

Extensibilidad: anticipando la necesidad de reutilización y facilitando la adición de nuevo conocimiento.

Mínimo compromiso ontológico: minimizar el número de afirmaciones a realizar sobre el dominio modelado, permitiendo así que otros agentes las puedan refinar.

Diversificación de jerarquías: haciendo uso de la herencia múltiple, usar tantos criterios de clasificación como sea posible. Así, añadir un nuevo concepto es sencillo utilizando los ya existentes y la capacidad de clasificación.

Minimizar la distancia semántica entre hermanos: conceptos similares deben agruparse.

Estandarización: independencia simbólica, utilización de nombrado estándar.

Granularidad: coherencia en el grado de particularidad, evitando el uso de términos ambigüos.

Finalmente, existen diversos procesos de desarrollo o metodologías que definen un procedimiento para la captura de conocimiento de un dominio y la pertinente construcción de ontologías. Por ejemplo: *Methontology Framework* o *Sensus method*.

3.1.3.7 Operaciones con ontologías

En esta sección se describen las operaciones [53] que se pueden realizar con las ontologías de acuerdo a su principio de reaprovechar el conocimiento ya definido. La reutilización del conocimiento consensuado es una de las máximas de la creación de ontologías, para ello se definen tres operaciones básicas (*mapping* o correspondencia entre ontologías, *merging* o unión de ontologías y *alignment* o descubrimiento de las correspondencias de *mapping*) que ayudan a la agregación de conocimiento basado en ontologías. La importancia de estas operaciones se pone de manifiesto en la mediación de datos entre fuentes heterogéneas, aspirando a resolver los conflictos que se producen entre sistemas basados en conocimiento, que deben interaccionar entre sí pero que han sido creados independientemente. El proceso de mediación, adquiere especial interés para la propuesta de servicios web semánticos, en los cuales esta operación es básica debido a la integración de ontologías provenientes de diferentes modelos de negocio.

***Mapping* o *mapeo* de ontologías:** especificación declarativa del solapamiento semántico entre dos ontologías. Las correspondencias entre entidades de ontologías diferentes son expresadas mediante axiomas en un determinado lenguaje de *mapeo*. Este proceso consta de tres fases: descubrimiento, representación y ejecución. Existen diferentes enfoques para llevar a cabo esta operación como MAFRA, RDFT o C-OWL.

***Alignment* o alineamiento de ontologías:** proceso mediante el cual se descubren las similitudes entre dos ontologías. El resultado es una especificación de los puntos en común, realizada a través del algoritmo *Match operator*. Existen diferentes implementaciones como Anchor-PROMPT, GLUE, *Semantic Matching* o QOM.

***Merging* o unión de ontologías:** creación de una ontología nueva tomando como fuente dos o más ontologías. La nueva ontología unifica y reemplaza las ontologías fuente. Se establecen dos enfoques para realizar esta operación: 1) entrada de *n* ontologías y salida de una sola ontología, unión y reemplazo de las demás (por ejemplo: algoritmo PROMPT [239]) y 2) entrada *n* ontologías que no son reemplazadas, sino que se genera una ontología *bridge* que importa a las ontologías originales y especifica las correspondencias mediante axiomas *bridge*, por ejemplo *OntoMerge*.

A la hora de afrontar la implementación de estas operaciones sobre ontologías se pueden generar dos tipos básicos de conflictos que impiden el éxito de la operación y requieren intervención humana para facilitar la realización automática de las operaciones:

1. Conflictos entre "conceptualizaciones" distintas del mismo dominio. A su vez, se distinguen dos categorías: 1) conflicto de ámbito, ocurre cuando dos clases tienen solapamiento en sus extensiones (el conjunto *s* de instancias), y no coincide exactamente y 2) conflicto en la cobertura del modelo y su granularidad, ocurre si dos ontologías cubren parte de cierto dominio (por ejemplo: empleados de universidad y estudiantes) o bien si una es más específica que otra (por ejemplo: una ontología define "persona" y otra define "persona joven").

2. Conflictos entre las especificaciones de los conceptos. También, se diferencian tres categorías: 1) conflicto en el estilo de modelado, cada ontología específica los conceptos de una manera determinada (por ejemplo: tratamiento de las unidades de tiempo) o la descripción de los conceptos difiere (por ejemplo: utilización de subclases vs atributos); 2) conflicto en la terminología, dos conceptos son equivalentes pero no utilizan el mismo nombre, problema de sinónimos, o viceversa, son diferentes y utilizan el mismo nombre, homónimos y 3) conflicto de codificación, no se utilizan las mismas nomenclaturas, unidades de medida, etc.

Todos estos conflictos, vienen en muchos casos provocados por no ajustarse a los principios de diseño establecidos en la Sección 3.1.3.6. No obstante, es habitual afrontar los problemas surgidos en la integración de ontologías ya que los modeladores provienen de distintas partes, con diferente formación y puede que sigan procedimientos particulares para la realización del modelado de las ontologías, facilitando las tareas a las aplicaciones que las consuman.

3.1.3.8 Aplicación de las ontologías

Las ontologías se convierten en la pieza fundamental de ciertas áreas así pueden señalarse las siguientes:

Ingeniería del conocimiento: las ontologías se pueden manifestar durante la ejecución de las siguientes tareas:

- Construcción del modelo conceptual, generando los términos del glosario y de las relaciones que se establecen entre ellos.
- Construcción de la base de conocimiento, utilizando la ontología de modelado conceptual se pueden crear bases de conocimiento con la aplicación de reglas, restricciones, etc.

Procesamiento del lenguaje natural: mantenimiento de la definición de términos gramaticales del lenguaje y las relaciones entre ellos.

Integración de sistemas heterogéneos: gestión de las diferencias existentes entre diversos sistemas de información con el objetivo de facilitar la comunicación entre los mismos.

Búsqueda semántica: utilizando conceptos y no términos para realizar las búsquedas.

Web Semántica: las ontologías son la base de la Web Semántica, por ello cualquier aplicación que tenga un carácter semántico se apoyará, muy probablemente, en ontologías.

Las ontologías se abren paso con fuerza sobre todo en el ámbito de las aplicaciones [249] de la Web Semántica, en particular en el escenario de la integración de aplicaciones (servicios web semánticos [**?**]) y contextualización del usuario. En la construcción de una ontología, hay que afrontar el modelado desde el punto de vista de la lógica (tipo y lenguaje de expresión) siguiendo los principios de diseño, ver Sección 3.1.3.6, y no de la orientación a objetos, muy frecuente en el ámbito de la ingeniería del *software*.

3.2 *Linked Data*

Actualmente se está ante un entorno de datos fluyendo constantemente, esperando a ser utilizados para generar información sobre un determinado acontecimiento. De esta manera, podemos ser conscientes de las últimas noticias, productos de una determinada marca o simplemente sobre información de nuestros contactos en las redes sociales. Esta nueva situación implica que existe un nuevo mercado para la construcción de aplicaciones y servicios [199] que exploten estos datos, generando nuevos negocios [212] y conectando a las comunidades científicas, y en general fomentando el progreso de una forma sostenible. Por ejemplo, algunos servicios de búsqueda como Google [244] utilizan los datos publicados en las tiendas de Internet como Amazon [175], para extender y procurar resultados de búsqueda más ajustados. En esta línea, prestando atención a los resultados de los principales buscadores, las primeras sugerencias corresponden a páginas de Wikipedia [297], Facebook [305], Linkedin [179], etc., por lo que efectivamente queda patente que los servicios de búsqueda [254], entre otros, están sacando partido de los datos estructurados publicados en distintos sitios web, para así identificar descripciones, organizaciones, personas, perfiles profesionales o productos y ofrecer resultados más acordes con la intención de los usuarios.

La evolución hacia esta nueva web, también de denominada Web 3,0, globalmente enlazada mediante datos supone una transición [7] de la web tradicional orientada a documentos (Web 1,0) y dirigida hacia las personas principalmente, incluyendo relaciones (Web 2,0), a una nueva *Web of Data*, en la cual los datos se publican para ser tratados tanto por los humanos como por los agentes automáticos, proporcionando descripciones más completas a todos los niveles. En este sentido, *Tim Berners-Lee* explica convenientemente los principios de esta nueva web realizando una analogía de una bolsa de patatas [24] con una página web, en la cual la parte delantera estaría orientada hacia las personas, mientas que la información (meta información) de la parte trasera serían datos dirigidos hacia otros agentes.

Actualmente la publicación de datos se haya envuelta en distintos dominios en los que previamente nunca se había pensado compartir datos, pero en los que se están generando nuevos servicios y oportunidades de negocio a partir de esta iniciativa que busca la compartición de datos. En este estado pueden surgir varias preguntas:

- ¿Cómo se deben publicar los datos para ser reutilizados?
- ¿Existen mecanismos para la reutilización automática?
- ¿Cuál es la forma más ágil para integrar fuentes de datos heterogéneas?
- ¿Se puede asegurar la calidad de los datos en términos de disponibilidad, autenticidad, evolución, etc.?

Estamos, en consecuencia, ante una revolución del uso de datos en la Web, entorno en el cual se pueden realizar diferentes operaciones sobre los mismos: descubrimiento, consulta, integración y explotación. La Web ofrece su infraestructura [185], protocolos y amplio asentamiento para reducir el impacto de esta evolución. No obstante, esta nuevo entorno trae consigo nuevos requisitos arquitectónicos, necesidad de entender el comportamiento de las URIs [27] y del protocolo HTTP [28] para reutilizar la infraestructura existente y concienciar a la comunidad del empleo de este entorno.

3.2.1 Definición y necesidad de *Linked Data*

Tim Berners-Lee acuñó esta iniciativa en el año 2006 [23] como parte de la Web Semántica [29–31]. En un principio la Web Semántica surgió con el objetivo de integrar grandes bases de datos a través

de modelos de conocimiento compartidos u ontologías, de esta manera cualquiera podría reutilizar el conocimiento ya consensuado y en consecuencia los datos asociados a esta información. Si bien la Web Semántica tuvo un gran éxito desde un punto de vista teórico, en la práctica escasas aplicaciones reales llegaron a tener un notable impacto en el gran público [296]. Es por ello, que tomando la visión de la Web Semántica en su parte de modelos y formatos de datos estandarizados formalizados a través de conocimiento compartido, fue tomada para definir lo que actualmente se conoce como *Linked Data* [183]. En realidad se trata de un enfoque práctico de la Web Semántica, en el cual se hace uso de los fundamentos de la semántica y de la web para proveer mecanismos para la publicación [35, 39] y consumo de datos masivos sin tener grandes barreras de entrada y claramente orientado al desarrollo de aplicaciones. Análogamente en el arranque de la web orientada a documentos se realizó un esfuerzo para identificar cómo publicar documentos, por ejemplo utilizando HTML, pero no se hizo hincapié en modelar la información que se publicaba en los mismos, esto ocasionó que la Web que hoy conocemos despegara exponencialmente. En el caso de la *Web of Data* se realiza un enfoque similar, es decir, se establece cómo publicar datos, sin embargo no presta especial énfasis en que estos datos estén formalizados estrictamente sobre un modelo conocimiento. De esta manera se facilita y favorece el crecimiento y la aparición de nuevas fuentes de datos en la Web. No obstante, esta segunda parte de modelado y estructuración de datos deviene fundamental para ciertas tareas automáticas.

El objetivo por lo tanto de la *Web of Data* no se centra tan sólo en publicar datos sino en hacer los datos accesibles, tanto a las personas como a las máquinas, como son actualmente los documentos disponibles en la Web. La diferencia sustancial reside en que actualmente la web está orientada al uso por personas y las máquinas necesitan procesos dotados de un cierto grado de complejidad para acceder a esta información. Por lo tanto, con esta nueva iniciativa se facilitan los enlaces entre los datos para que puedan ser explorados eficientemente por cualquier agente y realizar operaciones de enriquecimiento y enlazado de forma sencilla.

La *Web of Data* se construye sobre enlaces como la web tradicional en la que los documentos son enlazados entre sí. La diferencia reside en que los datos utilizan enlaces en RDF para describir recursos de distintos tipos identificados de forma unívoca mediante un URI y siguiendo unos principios que, aunque analizados en detalle en la Sección 3.2.4, se presenta a continuación una breve descripción de los mismos:

- *Use URIs as names for things*. Utilizar URIs como nombres para identificar y acceder a los recursos.
- *When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL [259])*. El acceso a los recursos a través de URIs debe estar basado en estándares como RDF o SPARQL suministrando igualmente información sobre el acceso.
- *Include links to other URIs. so that they can discover more things*. Enlazar datos no sólo implica la publicación masiva de datos sino que se deben establecer enlaces entre ellos con el objetivo de proveer un entorno navegable [25, 250].
- *Use HTTP URIs so that people can look up those names*. Utilizar URIs HTTP para reaprovechar la actual infraestructura de la web y que el público potencial pueda acceder a los datos de la misma forma que se realiza en el presente con los documentos.

En este contexto formalizado con reglas muy explícitas, queda patente que uno de los factores clave de esta iniciativa reside en poder de reutilizar información bien estructurada desde la propia fuente de datos, facilitando así la creación de aplicaciones fiables en el sentido del tratamiento de datos y evitando el uso de otras técnicas basadas en estadística y procesamiento de lenguaje natural, que se siguen aplicando ya que no siempre se cumple con todos los requisitos que establece la iniciativa de *Linked Data*.

En resumen, el objetivo de la iniciativa de *Linked Data* es enlazar datos de fuentes heterogéneas entre sí, con el objetivo de enriquecer la información y ofrecer a todos los agentes tanto personas como máquinas un espacio homogéneo basado en estándares para la realización de las operaciones citadas anteriormente, descubrimiento, acceso y explotación. Para ello se hace uso de la tecnología semántica RDF que provee una forma flexible y escalable para la realización de descripciones de las entidades del mundo: personas, organizaciones, localizaciones, etc. Las sentencias en RDF permiten establecer de una forma sencilla enlaces entre ellas creando conexiones desde un entorno local hacia un entorno global. También cabe destacar que el uso de RDF difiere de los actuales documentos presentes en la web en dos puntos importantes:

- El enlace se produce a nivel de entidades, no de documentos.
- Cada enlace tiene un tipo definido.

Linked Data permite así establecer conexiones entre diferentes fuentes de datos y crear un espacio de datos global, esto es posible gracias al uso de estándares y de un modelo de datos común que facilita la creación de aplicaciones de carácter general capaces de operar en este nuevo espacio de datos.

El gran valor de *Linked Data* se centra por tanto en brindar una oportunidad para utilizar los datos y explotar su valor mediante sistemas basados en conocimiento, con capacidad de procesar una cantidad masiva de datos obteniendo así soluciones más fiables y aproximadas a las necesidades y expectativas de los agentes. En síntesis, ver Tabla 3.2, se pueden establecer una serie de características y criterios a satisfacer.

ID	Directriz	Descripción
Principios Linked Data		
1.1	*Use URIs as names for things*	
1.2	*When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)*	
1.3	*Include links to other URIs*	
1.4	*Use HTTP URIs*	
Modelo ⋆		
2.1	⋆	*Available on the web (whatever format) but with an open licence, to be Open Data*
2.2	⋆⋆	*Available as machine-readable structured data (e.g. excel instead of image scan of a table)*
2.3	⋆⋆⋆	*as (2) plus non-proprietary format (e.g. CSV instead of excel*)
2.4	⋆⋆⋆⋆	*All the above plus, Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff*
2.5	⋆⋆⋆⋆⋆	*All the above, plus: Link your data to other people's data to provide context*

Tabla 3.2: Características a tener en cuenta sobre *Linked Data*.

3.2.2 *Open Data*

Este término se enclava dentro de las definiciones de conocimiento abierto realizadas por la organización *Open Knowledge Foundation* [205] (OKFN), en el cual también se incluyen:

- Contenidos como libros, películas, música y cualquier otro material con autoría definida.
- Datos, en general, pertenecientes a distintos contextos: científicos, históricos o geográficos.
- Información proveniente de las Administraciones Públicas.

Siguiendo la definición realizada por esta institución, cuando una obra o datos satisfagan las siguientes condiciones serán consideradas como abiertas:

Acceso. La obra debe estar disponible íntegramente y sólo a un coste de reproducción razonable, preferiblemente para su descarga de modo gratuito en Internet. La obra también debe estar disponible en una forma conveniente para ejecutar modificaciones sobre ella.

Redistribución. La licencia no debe contener restricciones sobre la posibilidad de venta o distribución de la obra en sí misma o formando parte de un paquete constituido por obras de fuentes diversas. La licencia no debe exigir un pago u otro tipo de cuota para esta venta o distribución.

Reutilización. La licencia debe permitir hacer modificaciones y obras derivadas facultando que éstas sean distribuidas en las mismas condiciones que la obra original. La licencia puede imponer algún tipo de requisito relativo al reconocimiento y a la integridad: veáse el principio 5 (Reconocimiento) y el principio 6 (Integridad) reseñados más adelante.

Ausencia de restricciones tecnológicas. Se debe proporcionar la obra de modo que no presente ningún obstáculo tecnológico para ejecutar los actos mencionados anteriormente. Esto se puede conseguir ofreciendo la obra en un formato de datos abierto, por ejemplo un formato cuya especificación esté disponible públicamente y de manera gratuita, de modo que para su uso no se imponga ninguna restricción de tipo monetario o análogas.

Reconocimiento. La licencia puede exigir como condición para la redistribución y la reutilización el reconocimiento de los contribuyentes y creadores de la obra. Si se impone esta condición, no debe ser de manera onerosa. Por ejemplo si se exige un reconocimiento, la obra debería ir acompañada de una lista de aquellos susceptibles del mismo.

Sin discriminación de personas o grupos. La licencia no debe discriminar a ninguna persona o grupo de personas.

Sin discriminación de ámbitos de trabajo. La licencia no debe contener restricciones sobre el uso en un ámbito de trabajo específico. Por ejemplo, no se puede limitar el uso de la obra en un negocio, o que ésta sea utilizada para investigación militar.

Distribución de la licencia. Los derechos adjuntos a la obra deben aplicarse también a todo áquel a quien le sea redistribuida, sin necesidad de que éste disponga una licencia adicional.

La licencia no debe ser específica de un paquete. Los derechos adjuntos a la obra no deben depender de que la obra forme parte de un paquete particular. Si la obra se extrae de ese paquete y se utiliza o se distribuye en las condiciones de la licencia de la obra, todos aquellos a quien les sea redistribuida deberán tener los mismos derechos que los concedidos conjuntamente con el paquete original.

La licencia no debe restringir la distribución de otras obras. La licencia no debe imponer restricciones en otras obras distribuidas conjuntamente con la obra objeto de la licencia. Por ejemplo, la licencia no debe imponer que todas las otras obras que se distribuyan por el mismo medio sean abiertas.

La corriente de datos abiertos [121] ha sido especialmente acogida en el ámbito de las Administraciones Públicas, ver Figura 3.21, adaptando las directrices aquí fijadas para desplegar el movimiento de *Open Government Data*.

Figura 3.21: Estado de *Open Data* en España por el Ministerio de Hacienda y Administraciones Públicas

Las instituciones públicas han establecido distintos canales de comunicación [1] con los ciudadanos, intentando siempre mejorar la interacción en los distintos procesos administrativos. Por ello y para estar presentes en las principales iniciativas de comunicación han procurado atender a los avances tecnológicos. Desde principios de la década de los años noventa han tenido presencia en la web proveyendo distintos servicios para que el ciudadano pueda agilizar sus trámites y así obtener una administración más eficiente, tanto en tiempo como en uso de recursos. Esta situación ha derivado a lo que se denomina como "administración electrónica" o *e-government*, en la cual la información de los servicios y datos que obran en poder de la Administración resultan de fácil acceso a los ciudadanos a través de los principales canales de comunicación, como la web, dispositivos móviles, etc., creando de esta forma una administración con una ventanilla única, fijando como objetivo principal la transparencia y la objetividad, materializadas a través de distintas políticas, concretamente en el caso europeo se dispone de un Plan de Acción [68] específico para el período 2011-2015.

No obstante, aunque las intenciones están acertadamente fijadas existen muchos desafíos que han estado y están impidiendo el despegue definitivo de la administración electrónica, desde la tecnología hasta el necesario cambio tanto para los integrantes de la propia administración como para los ciudadanos. Proveer servicios y datos a través de la red se ha convertido en un desafío en el cual se encuentran inmersas las distintas administraciones, esta situación implica ejecutar una gran

inversión para dar cabida a todas las necesidades desde un punto de vista tanto en materia de infraestructura como de concienciación social. De esta forma, un gobierno basado en administración electrónica debe ser capaz de fijar los requisitos de los servicios y datos a liberar, asegurar la autenticidad y definir la legislación pertinente que establezca un marco jurídico seguro para que los ciudadanos pueden utilizar toda la información disponible para consultar o construir nuevos servicios de valor añadido sobre ellos.

Entre las actividades que se presentan como clave para el desarrollo de la administración electrónica se pueden referir las siguientes:

- Uso y aplicación de estándares.
- Transparencia y participación.
- Concienciación de los avances en la gestión e integración de datos.
- Relaciones y colaboraciones.

En este contexto se determinan por una parte los objetivos de la administración respecto a disponer un entorno flexible y dinámico para sus ciudadanos, y por otra parte se encuentra el comportamiento de las propias personas y organizaciones. Actualmente se advierte un entorno definido por las siguientes características: global, conectado, cambiante y accesible en su mayor parte.

El término de *Open Government Data* ha sido designado en la administración de Estados Unidos para indicar aquellos datos o registros de información que obran en poder de la Administración Pública y que han sido recabados con distintos objetivos en los distintos procesos administrativos. Como ejemplos de reutilización en el sector público se encuentran los siguientes dominios: salud, cartografía, datos meteorológicos, educación, datos bibliográficos, contratación pública, legislación, etc. Como se ha comentado con carácter previo, las organizaciones públicas producen, almacenan y distribuyen distintos tipos de información: legal, financiera, geográfica, etc., en su actividad diaria. Esta información proveniente del sector público (PSI) está enclavada dentro de un marco jurídico que puede variar de un país a otro, por ejemplo en Europa la Directiva 2003/98/EC [75,246], en España la Ley 37/2007 [97] y su Real Decreto 1495/2011 [98] de desarrollo, Reino Unido [141], Francia [128] o la iniciativa de la Casa Blanca en Estados Unidos [171]. Habitualmente se utilizaban distintas técnicas y formatos para la distribución de la información del sector público por diferentes canales desde el papel y correo postal, hasta formatos electrónicos y las propias páginas web. La evolución en los últimos años ha sido de gran calado de modo que ha hecho necesario dictar una legislación común que debe ser transpuesta en los diferentes países, como ocurre en el caso de Europa.

La transición hacia este nuevo enfoque administrativo todavía se encuentra en una etapa temprana debido a las dificultades que supone toda una nueva forma de actuar. En general, se suele hablar de *Government Data* y *Public Sector Information* para referirse al mismo concepto. Se han realizado diferentes definiciones como las presentes en la iniciativa *Open Government Data*, pero no sólo se trata de exponer de forma pública las grandes bases de datos gubernamentales, sino que éstas deberían cumplir unos principios [184] que a continuación se listan:

1. Data Must Be *Complete*. *All public data are made available. Data are electronically stored information or recordings, including but not limited to documents, databases, transcripts, and audio/visual recordings. Public data are data that are not subject to valid privacy, security or privilege limitations, as governed by other statutes.*
2. Data Must Be *Primary*. *Data are published as collected at the source, with the finest possible level of granularity, not in aggregate or modified forms.*
3. Data Must Be *Timely*. *Data are made available as quickly as necessary to preserve the value of the data.*

4. Data Must Be *Accessible*. *Data are available to the widest range of users for the widest range of purposes.*
5. Data Must Be *Machine processable*. *Data are reasonably structured to allow automated processing of it.*
6. Access Must Be *Non-Discriminatory*. *Data are available to anyone, with no requirement of registration.*
7. Data Formats Must Be *Non-Proprietary*. *Data are available in a format over which no entity has exclusive control.*
8. Data Must Be *License-free*. *Data are not subject to any copyright, patent, trademark or trade secret regulation. Reasonable privacy, security and privilege restrictions may be allowed as governed by other statutes.*

Con todo ello, se provee un entorno en el cual "público" designa a las condiciones que deben cumplir los datos para considerarse "abiertos", no se entra a valorar los datos que se deben liberar ni tampoco cuestiones legales. En cuanto a los "datos", se toma este término para designar a los registros electrónicos, es importante destacar que los documentos físicos no se consideran como tal, que obran en poder de la administración y finalmente, con "revisable" se denomina a aquellos datos para los cuales existe una persona de contacto que actúa como enlace para aquellos que quieran usar los datos, para gestionar el uso de los mismos y para los que existe un órgano administrativo o judicial que posee la potestad para aplicar adecuadamente los principios citados.

Por otra parte, existen cuestiones [20] que deben tener respuesta en el contexto de *Open Government Data* desde un punto de vista estratégico hasta funcional, como qué tipo de interfaz de acceso (API) se debe proveer para el acceso a los datos, por ejemplo, en cuanto a los datos a liberar, la iniciativa de datos abiertos no especifica qué datos deberían liberarse. Corresponde a la estrategia del organismo en cuestión establecer las prioridades, coste y beneficios de la liberación de un determinado conjunto de datos. En determinados casos como el del Gobierno de Reino Unido [140] o el Ayuntamiento de Zaragoza [99] se utilizan encuestas a los propios usuarios para especificar o establecer un orden en los datos a liberar. El objetivo final será disponer de una política y guías de apertura de datos que sean capaces de dar soporte a los siguientes puntos:

Inclusión. El uso de estándares facilita la accesibilidad y la usabilidad de los datos. De esta forma, se facilita la participación de terceros y el desarrollo de aplicaciones de distinto carácter que beneficiarán transversalmente a toda la sociedad con nuevos servicios.

Transparencia. En los últimos años y debido a los múltiples casos de corrupción a nivel político, deficiente gestión de fondos públicos y en general, cierta falta de claridad por parte de las instituciones ha hecho que esta corriente sitúe como un eje el impulso de la transparencia en el conjunto de las Administraciones Públicas.

Responsabilidad. La eficiencia en el uso de recursos debe ser uno de los objetivos de la Administración Pública, por ello conseguir información sobre su funcionamiento puede mejorar el mismo a nivel interno.

Una vez que se dispone de datos abiertos cabe definir cómo se pueden obtener mejoras en ciertos procesos de uso. Entre ellos se puede destacar:

Reutilizar. La información abierta permite plantear nuevos modelos de interacción, tanto para otras administraciones públicas como para organizaciones o comunidades que se beneficien de estos datos para el despliegue de nuevos servicios [212]. Actualmente se realizan actividades para el impulso de la reutilización de datos, como los concursos en los que liberados datos, los interesados presentan aplicaciones haciendo uso de ellos.

Generar múltiples vistas de los datos [84]. Cuando se dispone de los datos de una forma más o menos estandarizada y con facilidad de acceso se pueden generar múltiples vistas, dependiendo de las necesidades del usuario y adaptándolos al contexto. Por ejemplo, generación de informes en PDF, visualización en mapas de la información geográfica, etc. Partiendo de los mismos datos y en algunos casos y tras la ejecución de un proceso de enriquecimiento se puede mejorar la información que provee. El coste de realización de estos procesos es alto si está centralizado en la administración, pero se puede diversificar y abaratar delegando en terceras partes.

Mejorar los actuales sistemas de búsqueda [166]. Un sistema de búsqueda dispondrá de mayor precisión cuanto más "informado" esté, por lo tanto, si se pueden reutilizar los datos ya disponibles para mejorar los resultados de las consultas de los usuarios, se facilitará el acceso a la información.

Integrar fuentes de datos [6, 149]. Habitualmente cada organismo goza de cierta independencia para gestionar sus propios datos. La realidad es que en muchos casos diferentes organismos de la misma entidad tienen la necesidad de cruzar datos para obtener diferente información agregada. Si se utilizan datos abiertos se favorece la integración y la reutilización de datos disminuyendo los costes de la obtención y gestión de los mismos al no estar multiplicada.

Una vez que se han definido algunos de los beneficios del uso de datos abiertos en el ámbito de las administraciones públicas, cabe ahora especificar cómo han de ser publicados para que puedan ser consumidos por terceros de forma ágil. Para ello se han establecido una serie de directrices o métodos:

- Usar anotaciones basadas en microformatos o RDFa [2] en XHTML o HTML5 [163].
- Facilitar la accesibilidad siguiendo las directrices WCAG [57].
- Desplegar APIs específicas para el consumo de datos.
- Sindicar de contenidos mediante RSS [278] o ATOM [238].
- Proveer interfaces REST [119] o SOAP [50] de servicios de acceso a datos.
- Aplicar tecnologías semánticas mediante RDF.

Finalmente se ha de definir la misión y estrategia de apertura de datos, así la confianza (*trust*), procedencia [58, 216] (*provenance*) y calidad de los mismos, la evolución en tiempo, las limitaciones tecnológicas y la capacidad para seguir creciendo. De acuerdo a lo expuesto en esta sección se pueden establecer una serie de factores a evaluar sobre la publicación de datos abiertos de acuerdo a distintos criterios y que servirán para comprobar si se cumplen las guías definidas.

3.2.3 *Linking Open Data*

La convergencia entre las iniciativas de *Linked Data* y *Open Data* ha conllevado la creación de un nuevo término para designar a aquellos datos que cumplen las guías de ambas iniciativas y que se denomina *Linking Open Data*. Bajo esta definición se agrupan todos los datos que han sido liberados aplicando los principios de *Linked Data* y que además cumplen significativamente las directrices de *Open Data*. Habitualmente este término se aplica en el contexto de las Administraciones Públicas, en las que el esfuerzo por la apertura de datos ha sido intenso y en el que han logrado grandes avances.

El alto valor de los datos enlazados abiertos permite disfrutar a terceros de la posibilidad real de creación de servicios de alto valor añadido. No obstante, surge la duda de si el proceso de enlazado

de datos y de explotación, corresponde a la propia Administración o si cabe delegar el mismo. En este sentido, existen una serie de cuestiones que se han de plantear:

- **¿Debe la Administración explotar los datos y construir aplicaciones para su consumo?**

  En un primer momento, las Administraciones Públicas cubrían toda la cadena valor de uso de los datos, desde su apertura hasta su explotación. En estos momentos, se está produciendo un cambio de dirección por cuestiones de sostenibilidad que implica que la creación de servicios y aplicaciones se delega en terceros. Si bien como punto de partida es conveniente que la Administración demuestre el valor del uso de los datos, una vez que se ha producido la sensibilización necesaria resulta evidente que deben ser terceras partes quienes reutilicen los datos.

- **¿Qué nivel de *Linked Data* debe proveer la Administración?**

  El caso ideal sería que los datos liberados se encontrarán en el estatus de 5 $\star$ pero la realidad es que tanto por parte de la comunidad de desarrolladores, habituados a tecnologías como RSS o APIs REST, como por el propio coste que genera llegar a este nivel, en muchos casos no es realmente útil realizar este esfuerzo, con la simple apertura de datos y un modelo de acceso convenientemente homogéneo es suficiente.

- **¿Sobre la calidad de los datos?**

  Como se ha comentado es fundamental para motivar la creación de aplicaciones la confianza en los datos que se están utilizando. En este sentido poder verificar su procedencia, evolución en el tiempo, autenticidad y en general, su nivel de confianza y calidad es vital para el triunfo de esta corriente en aplicaciones de alto valor añadido. El esfuerzo de la Administración debe centrarse por consiguiente en este punto, asegurando a la comunidad la confianza en los datos que van a usar.

- **¿Qué estructura se debe seguir en los organismos públicos para liberar los datos?**

  Existen principalmente dos enfoques para abordar la apertura de datos masiva: *top-down* y *bottom-up*. El enfoque *top-down* se está utilizando en el Gobierno de Reino Unido y consiste básicamente en fijar las directrices, modelos, etc., que se han de seguir para liberar datos desde un organismo central que propaga esta metodología a todos los sectores y organismos públicos que deseen abrir sus datos. Por el contrario, el enfoque *bottom-up* tiene su paradigma en la Administración española en el cual han surgido numerosos conjuntos de datos aplicando sus propias directrices. Determinar cuál es el mejor modelo depende del nivel de precisión que se establezca en cuanto a la publicación de datos, en contraposición con el coste del mismo y el tiempo empleado. El enfoque *top-down* si bien es más homogéneo, requiere un enérgico esfuerzo común inicial tanto en tiempo como en coste, pero en el largo plazo es más sostenible puesto que todos los organismos reaprovechan la experiencia. En cambio, el enfoque *bottom-up* permite un despliegue inicial más rápido multiplicando el esfuerzo y los costes en cada organismo candidato a la apertura de datos, pero con la posibilidad de que uno de los modelos particulares se convierta en referente.

- **¿Cómo se debe prevenir la privacidad de los datos?**

  Si bien por la propia definición de datos enlazados abiertos no se debiera tener en cuenta esta cuestión, si que surge la necesidad en contextos determinados de ofrecer mecanismos de control para prevenir el uso inadecuado de datos, por ejemplo los denominados "secretos estadísticos" y así evitar infringir otras leyes como las referidas a datos personales.

- **¿Puede la Administración establecer un modelo de negocio, ver Figura 3.22, sobre los datos?**

  Este punto genera cierta controversia, ya que el coste de posesión de los datos recae sobre los ciudadanos, actualmente cualquier servicio público o proceso administrativo tiene unos costes

que se financian a través de determinadas tasas. En el caso de la apertura de datos se considera que no debe suponer un coste adicional para los administrados pero la realidad es que el coste implícito está presente, es más, algunos organismos utilizan este mecanismo como medio de auto-financiación. Desde un punto de vista de los agentes implicados, los terceros que utilicen estos datos no estarán dispuestos a pagar un sobrecoste por su uso, la propia Administración no estaría incentivada a la aplicación de tasas sobre estos datos presumiendo que el retorno provisto de terceros sea capaz de financiar la apertura de los datos. Reiterando que por definición los datos deberían ser libres también es necesario la creación de un entorno sostenible y probablemente el enfoque óptimo en el estadio inicial sea de conformidad a un modelo libre.

- **Otras cuestiones.**

  Relacionadas con los datos a liberar, realimentación de los mismos (*crowdsourcing*) para corregir errores, reconciliación, etc., siguen todavía abiertas y obtendrán respuesta paulatinamente, en función la propia demanda.

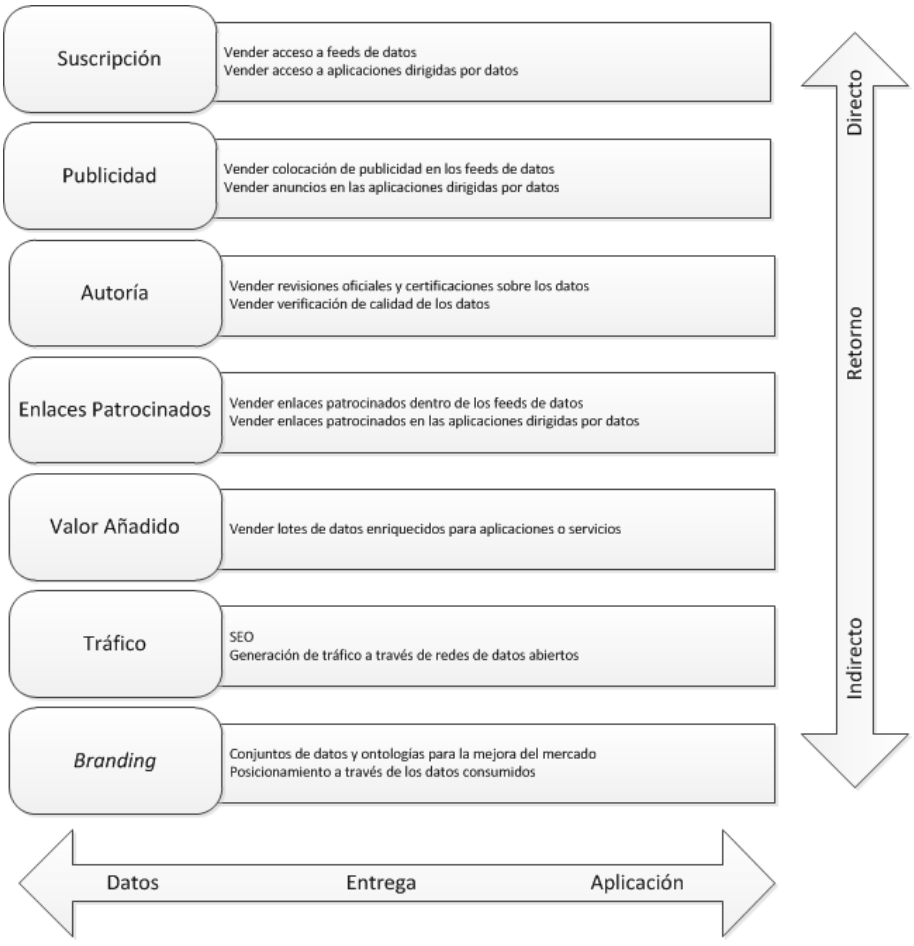


Figura 3.22: Modelos de negocio para datos.

En definitiva, la conjunción de los principios de *Linked Data* y *Open Data* ha conllevado una corriente de actuación en las Administraciones Públicas con amplios beneficios tanto para los ciudadanos como para las empresas. La experiencia inicial es altamente enriquecedora para las partes implicadas y su tendencia es seguir creciendo. No obstante, también se presentan varias cuestiones abiertas que se habrán de ir resolviendo a la vez que la tecnología y la experiencia se incrementan.

3.2.4 Principios de *Linked Data*

En las secciones anteriores se han repasado las líneas que marcan la definición de *Linked Data*, se pueden resumir en una corriente que aúna la tecnología semántica más madura con la apertura de datos, en un infraestructura perfectamente establecida como es la Web. Los principios de *Linked Data* [23] se reducen a los siguientes cuatro puntos:

- *Use URIs as names for things*. La utilización de URIs como nombres para identificar y acceder a los recursos. Este principio busca el uso de URIs como camino para referenciar de forma única a todos los recursos disponibles, no sólo documentos, así: contenidos digitales, vídeos, sensores [56, 190], organizaciones [284], personas [176], etc., con el objetivo de ampliar el contexto de un espacio virtual a uno real.
- *When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL [259])*. El acceso a los recursos a través de URIs debe estar basado en estándares como RDF o SPARQL proveyendo igualmente información sobre el mismo.
- *Include links to other URIs. so that they can discover more things*. El enlazado de datos no sólo implica la publicación masiva de los mismos, sino que se deben establecer enlaces entre ellos con el objetivo de proveer un entorno navegable [25, 250] y de consulta [150].
- *Use HTTP URIs so that people can look up those names*. La utilización de URIs HTTP para reaprovechar la actual infraestructura de la web y que el público potencial pueda acceder a los datos de la misma forma que se realiza actualmente con los documentos.

En general estos principios buscan una manera estándar de identificar, modelar, acceder y consumir recursos para que la creación de aplicaciones y servicios sea lo más sencilla y ágil posible. Por lo tanto, para atender a estos principios se deben facilitar una serie de mecanismos que permitan:

- Nombrar a todas las cosas, recursos, etc., mediante URIs.
- Referenciar y acceder a los contenidos mediante URIs. Uso de *hash*, *slash* URIs o código 303 de HTTP.
- Proveer información útil sobre los recursos de forma estándar mediante RDF. Uso de literales o de enlaces a otras descripciones RDF. Negociación de contenido.
- Incluir enlaces a otros datos. Podemos establecer tres tipos de enlaces principales: relación, identidad y vocabulario. Cada uno permite establecer relaciones, alinear unos recursos con otros y definir las descripciones respectivamente.

Los beneficios de la utilización de estos mecanismos que sustentan a los cuatro principios citados proporcionan las siguientes ventajas:

- Escala global. El uso de RDF permite unificar la información de los recursos, modelo, formato y acceso a los datos.
- Enlace entre recursos. RDF da soporte a la reutilización intrínseca de recursos ya definidos dando cabida a la unión de datos de diferentes fuentes.
- Expresividad. La información en RDF se puede representar combinando distintos vocabularios.
- Estructuración. Utilizando un modelo definido en RDF(S) u OWL se da soporte al modelo de los datos mediante mecanismos estándar, compartidos y reutilizables.

Además de estas ventajas, existen ciertas características que deberían evitarse, tales como los mecanismos de reificación de RDF, las colecciones o el uso de nodos en blanco con el objetivo de facilitar la consulta mediante el lenguaje SPARQL. En cuanto a los formatos en los que se pueden presentar los datos enlazados en RDF se encuentra: RDFa [2], RDF/XML [16], Turtle [17], N3 [26], RDF/JSON [117] o RDF binario [279].

En definitiva, se trata de proveer un entorno estándar y global de datos en el cual se utilice un modelo de datos estandarizado y único (RDF), con capacidad para enlazar a otros datos y que todos ellos se autodescriban en el sentido de facilitar a sus descripciones beneficiando el descubrimiento, acceso e integración de datos.

3.2.5 Construyendo una nueva Web de Datos

Este nuevo entorno de publicación de datos ha implicado que numerosas personas y organizaciones se hayan inclinado por la aplicación de los principios de los datos enlazados en la apertura de sus datos. Como resultado se puede establecer que el despegue de la Web de Datos es ya un hecho, creando un grafo global de billones de tripletas RDF de diferentes tipos y fuentes enlazadas [168] entre sí. En general, esta eclosión se debe a varias razones: genericidad, estandarización, dominio público, capacidad de representación de la información, expresividad, reutilización de tecnología estable, etc. Como origen de este enfoque se puede situar la iniciativa del *Linking Open Data* [144] (LOD) del W3C en el año 2007, en el cual se sentaron las bases de la actual Web de Datos, dando comienzo el desarrollo tecnológico así como la transformación de los primeros conjuntos de datos. Entre ellos podemos destacar el esfuerzo de la DBPedia [44] o herramientas como Pubby, D2RServer, etc. La evolución de la Web de Datos queda recogida en el diagrama de la nube de *datasets* (*LOD Cloud* [82]) que en su última versión recoge hasta 320 *datasets*, ver Figura 3.23, y continua en aumento si se consulta *The Data Hub* [123]. A modo informativo a fecha de septiembre de 2010 en el diagrama anteriormente citado estaban disponibles unos 203 *datasets* (en diciembre de 2011 llegan hasta 326), más de 25 billones de tripletas RDF y unos 395 millones de enlaces entre los diferentes datos.

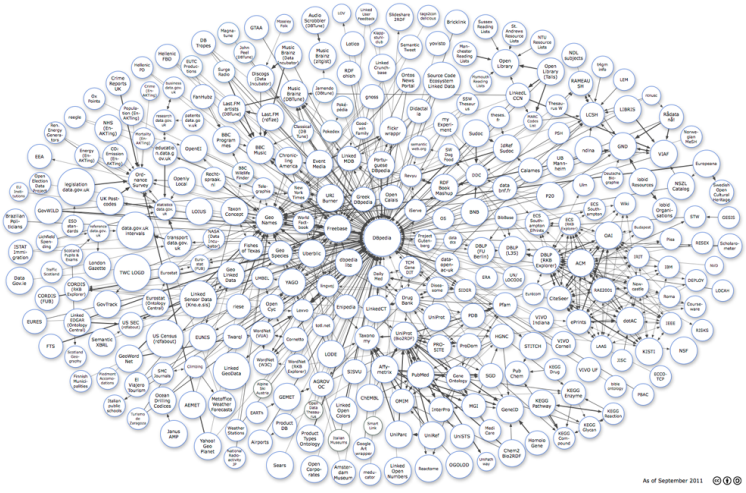


Figura 3.23: *Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch.*

Por otra parte, la información que se puede encontrar en la nueva Web de Datos es tan variada como en la actual web de documentos, a continuación algunos de los ejemplos de los datos que podemos encontrar, así:

- Genéricos o de carácter transversal, candidatos a ser reutilizados por cualquier conjunto de datos: DBPedia [43], Freebase [177], etc.

- Geográficos: Geonames [300], OpenStreetMap [243] o GeoLinkedData [281].
- "Government Data": Data Gov UK, Datos.es, etc.
- Educación, Multimedia, Bibliográficos, Científicos, Médicos, "Social Media", etc., que se pueden encontrar clasificados en la versión del diagrama de la nube de *Linking Open Data* bajo diferentes colores.

A la hora de unirse a la Web de Datos hay que tener en cuenta varias consideraciones, además de proveer datos enlazados se deben tener en cuenta unos principios relativos al diseño:

Diseño de las URIs. El objetivo es que las URIs referencien a objetos reales y que se puedan obtener descripciones a través de los protocolos propios de Internet, requisito alineado directamente con el segundo principio de datos enlazados y en este sentido hay que tener en cuenta varios aspectos:

- Uso de "minting HTTP URIs": son URIs que están bajo el control de áquel que publica datos, es decir, utiliza las URIs de su dominio para los datos y documentos.
- "Cool URIs" [22, 268]: las URIs deben permanecer en el tiempo, en todo caso los usuarios las pueden cambiar, pero los identificadores deberían permanecer inalterados. De la misma forma, se debe proveer un mecanismo para que a través de una misma URI se pueda acceder a distinto contenido mediante negociación de contenido, extensiones, etc.
- "Meaningful URIs" vs "ID based URIs" [79]: al igual que en el punto anterior se está ante la tesitura de diseñar identificadores que se puedan recordar y con significado respecto a su contenido o bien utilizar identificadores autogenerados. En cualquier caso, la decisión dependerá de la información a identificar, pero hay que tener presente este factor como clave en el diseño de la identificación de datos.

Descripción de recursos con RDF. En este sentido se han establecido una serie de buenas prácticas y patrones para describir los datos de los recursos susceptibles de publicación, ver Sección 3.2.6. Es importante seguir estas guías con el objeto de facilitar su consulta a través de lenguajes como SPARQL.

Metainformación sobre los datos publicados. No se trata sólo de publicar datos, ni de la apertura de las bases de datos a la web, sino también de suministrar información sobre qué, cómo y dónde se puede acceder a estos datos. Para ello, se debe atender a los siguientes criterios:

- Usar vocabularios para describir el *dataset* como voID [3] o "Semantic Sitemaps" [83].
- Añadir información sobre la procedencia [152, 230] de los datos (*provenance*).
- Añadir información sobre la licencia [197] de uso de los datos y los derechos de autoría, copia, distribución y modificación.
- Añadir información sobre la evolución temporal [143], se puede conseguir mediante diferentes técnicas, incluyendo la información en las URIs, en las descripciones RDF o mediante negociación de contenido.

Seleccionar vocabularios a utilizar. Existe un gran número de vocabularios [209] diseñados para describir diversos tipos de información. La tendencia es reutilizar los vocabularios existentes [256] en la medida de lo posible, evitando remodelar información ya descrita. Aunque existen catálogos de vocabularios, finalmente es la experiencia y la consulta de otros *datasets* la principal vía de información para seleccionar los vocabularios a reutilizar. En general se atenderá a los siguientes factores: uso actual, mantenimiento, cobertura y expresividad.

Enlazar con otros "*datasets*". Al igual que en el punto anterior no existe una forma estándar de realizar este proceso, puede ser manual o automático, pero si que es necesario fijar la atención en la semántica del enlace y seleccionar aquellos recursos que sean de mayor uso para asegurar de nuevo la coherencia de los datos y su difusión.

Publicar los datos. Una vez determinados los datos a publicar, los cuales se han transformado y enlazado (si resultara oportuno), es necesario proceder a su publicación de forma estandarizada mediante la aplicación de ciertos patrones definidos y teniendo en cuenta si los datos son o no dinámicos, la criticidad, el esfuerzo de duplicar bases de datos, los formatos que van a ser soportados, la exposición mediante servicios de consulta, el uso de un API, etc.

Consumir datos publicados. Este proceso conlleva la explotación de los datos disponibles tanto para generar nuevos servicios de negocio o simplemente para consultar la información que se haya publicada y que se puede enriquecer con el universo de los demás *datasets* publicados.

Difundir los datos publicados. Aunque no se trata de una acción o requisito funcional, el éxito de un conjunto de datos publicados dependerá en buena medida, además de su calidad [103] y facilidad de acceso, de la difusión que se realice sobre su uso y capacidad. Para ello, existen diversos catálogos [291] y herramientas de indexado, por ejemplo *thedatahub.org*, *prefix.cc*, etc., en los cuales se pueden incluir los datos publicados para que así, terceros los puedan consumir. La información de difusión deberá incluir aspectos relativos sobre cómo acceder a los datos, qué datos están publicados, el tipo de licencia, etc., para asegurar la confianza necesaria en los usuarios finales.

Como se ha reseñado en los puntos anteriores, pese a que existen patrones, recetas y métodos de producción, publicación y consumo de datos es decisión del responsable la selección de los mismos. Si bien la flexibilidad de actuación queda patente y beneficia la publicación masiva de datos, lleva consigo una desventaja, que reside en la posibilidad de errar en ciertos puntos provocando una baja calidad en los datos. Para ello, tal y como se expone en la siguiente sección existen diversos esfuerzos para poner en común estas guías bajo una metodología, especificando las actividades, participantes, herramientas y productos de entrada y salida en cada una de las partes del proceso. En el objeto de este estudio, esta situación es de especial relevancia para poder evaluar las ventajas de uso de datos enlazados, en contraposición con la actual situación de manejo de datos en el campo de la contratación pública electrónica.

3.2.6 Metodologías y Buenas Prácticas

En las anteriores secciones se han presentando las diferentes iniciativas en torno a datos enlazados en un contexto general, pero teniendo presentes las necesidades de la administración pública electrónica. Han quedado patentes las necesidades, ventajas y algunas desventajas del uso de esta iniciativa para los procesos de producción, publicación y consumo de datos. Si bien existen muchas guías y principios de diseño que se pueden extraer a través de la literatura, es conveniente repasar los enfoques que se están desarrollando para realizar una guía sobre cómo se deben aplicar los conceptos de *Linked Data*, *Open Data*, *Open Government Data*, *Linking Open Data*, etc., para ello existen diferentes guías e incluso plataformas que aglutinan una serie de características de interés que a continuación se listan, repasando las más destacadas.

- *Linked Data Design Considerations* del libro *Linked Data: Evolving the Web into a Global Data Space* [157].
- *Linked Data Patterns* [105].
- *Best Practices* [35,214] del grupo de trabajo del W3C *Government Linked Data First F2F June 29-30, 2011* [215].

- *Linked Data Cookbook* [173] del grupo de trabajo del W3C *Government Linked Data Working Group* [215].
- *Government Linked Data-Life Cycle* [155] del grupo de trabajo del W3C *Government Linked Data Working Group*.
- *Publishing Open Government Data* [20] borrador de trabajo del W3C.
- *LOD2 Stack* [285] del proyecto europeo *LOD2* [126].
- *Toward a Basic Profile for Linked Data* [234, 235] de IBM.
- Metodología [63, 64] propuesta por la Universidad de Oviedo en el ámbito de la Biblioteca del Congreso de Chile.
- *Talis Platform* [182].
- *Linked Open Data: The Essentials* [14].
- Documentación específica de cada país e incluso región para el despliegue de *Open Data*.

El objetivo de estas guías es definir unas directrices sobre cómo publicar los datos para obtener una Web de Datos de calidad. Además y en esta misma línea se están organizando continuamente conferencias, talleres, etc., específicos dentro de los grandes eventos científicos para divulgar las actividades relacionadas con *Linked Data*, como por ejemplo *Consuming Linking Open Data* [151] (COLD), *Linked Data On the Web* [42] (LDOW), *Social Data On the Web* [247] (SDOW) o *Linked Enterprise Data Patterns* [242] que sirven como concentrador de las actividades que se están realizando en este ámbito y que abordan los principales problemas que aparecen en este entorno emergente.

3.2.6.1 *Linked Data Design Considerations*

En el libro [157] realizado por Chris Bizer (investigador en la Freie Universitat de Berlin, Alemania y creador de la DBPedia) y por Tom Heath (uno de los investigadores principales de la plataforma Talis [182]) se ponen de manifiesto una serie de consideraciones, ver Tabla 3.3 de diseño a tener en cuenta cuando se pretende desplegar una plataforma de datos enlazados.

ID	Directriz	Descripción
1	***Using URIs as Names for Things***	
1.1	*Minting HTTP URIs*	Las URIs deben permitir acceder a los recursos que nombran. Uso del esquema HTTP.
1.2	*Use of Cool URIs*	Las URIs deben seguir unos criterios de diseño que permitan y animen a terceros su uso.
1.3	*Keep out of namespaces you do not control*	Las URIs de los recursos deben pertenecer a un dominio sobre el que se tenga control.
1.4	*Abstract away from implementation details*	Las URIs no incluyen detalles de implementación como el formato del recurso.
1.5	*Use Natural Keys within URIs*	Para asegurar que las URIs son únicas se utiliza una clave primaria o ID para identificar recursos.
1.6	*Hash URIs*	La concatenación de identificadores en el URI se realiza utilizando # como separador final del recurso.
1.7	*Slash URIs*	Igual que en el caso anterior pero utilizando /.

ID	Directriz	Descripción
2	***Describing Things with RDF***	
2.1	*RDF resources*	Cuando se accede a un recurso a través de un URI se debe proveer información útil.
2.2	*Literal Triples and Outgoing Links*	El sujeto de la información en RDF es siempre el URI del recurso que se está describiendo.
2.3	*Incoming Links*	Proveer enlaces a otros recursos para hacer los datos navegables.
2.4	*Triples that Describe Related Resources*	Describir parcialmente los recursos que son enlazados desde el actual.
2.5	*Triples that Describe the Description*	Añadir metainformación sobre los propios recursos como licencia.
3	***Publishing Data about Data***	
3.1	*Describing a Data Set*	Añadir metainformación sobre el conjunto de datos que se está publicando.
3.2	*Use of Semantic Sitemaps*	Utilizar la extensión de *Sitemap* para describir los datos proveyendo capacidades nuevas para los buscadores.
3.3	*Use of voiD Descriptions*	Utilizar el vocabulario voiD (*the Vocabulary of Interlinked Datasets*) para la metainformación de los datos. Es considerado el estándar de facto.
3.4	*Provenance Metadata*	Añadir información sobre la procedencia de los datos. Capacidades de monitorización y evolución en el tiempo.
3.5	*Licenses, Waivers and Norms for Data*	Incluir información sobre el uso posible de los datos a través de una licencia.
3.6	*Non-copyrightable Material*	$\approxeq$
4	***Choosing and Using Vocabularies to Describe Data***	
4.1	*SKOS, RDFS and OWL*	Modelar la información de los datos de acuerdo a un vocabulario estándar. Proveer un modelo formal para los datos.
4.2	*Annotations in RDFS*	Usar las anotaciones `rdfs:label` y `rdfs:comment` en las descripciones de RDF.
4.3	*Relating Classes and Properties*	Relacionar los recursos en RDF mediante propiedades de RDFS, SKOS, etc.: `rdfs:subClassOf`, `skos:related`, etc.
4.4	*Reusing Existing Terms*	Reutilizar las definiciones ya concebidas en los vocabularios existentes de los distintos dominios como: Dublin Core, FOAF, etc.
4.5	*Selecting Vocabularies*	Seleccionar los vocabularios a reutilizar teniendo en cuenta: *Usage and uptake*, *Maintenance and governance*, *Coverage* y *Expressivity*.
4.6	*Defining Terms*	Si los vocabularios existentes no cubren las necesidades de nuestros datos, definir nuevos términos.
5	***Making Links with RDF***	
5.1	*Publishing Incoming and Outgoing Links*	Publicar un volcado de los datos en RDF.

ID	Directriz	Descripción
5.2	*Making Links with External Data Sources*	Reutilizar y enriquecer las descripciones en RDF con otras ya existentes asegurando que existen.
5.3	*Choosing External Linking Targets*	Los enlaces a datos externos deben cumplir que sean referenciables mediante un URI y deberían ser reutilizados por otros.
5.4	*Choosing Predicates for Linking*	Similar al punto 22.
5.5	*Setting RDF Links Manually*	Incluir enlaces en las descripciones en RDF mediante edición manual (si de forma automática no es fiable).
5.6	*Auto-generating RDF Links*	Incluir enlaces las descripciones en RDF de forma automática. Uso de herramientas de reconciliación de entidades: SILK, LIMES, etc.
5.7	*Key-based Approaches*	Los enlaces se realizan mediante búsqueda de recursos con palabras claves.
5.8	*Similarity-based Approaches*	Los enlaces se realizan mediante búsqueda de recursos de acuerdo a una estructura.
6	***Recipes for Publishing Linked Data***	
6.1	*Linked Data Publishing Patterns*	Establecer el modelo para ofrecer los datos.
6.2	*From Queryable Structured Data to Linked Data*	Exportar una base de datos relacional mediante un mapeador a RDF.
6.3	*From Static Structured Data to Linked Data*	Transformar datos estáticos a RDF. Por ejemplo: CSV, MSExcel, etc.
6.4	*From Text Documents to Linked Data*	Extraer los datos RDF de documentos de texto.
6.5	*Data Volume: How much data needs to be served?*	Dependiendo de la cantidad de datos definir qué datos y cantidad se pueden exportar.
6.6	*Data Dynamism: How often does the data change?*	Dependiendo de los cambios en los datos definir qué patrón se adapta mejor a la exportación de los datos.
6.7	*Serving Linked Data as Static RDF/XML Files*	Publicar los ficheros en RDF en un servidor web. Forma más sencilla.
6.8	*Hosting and Naming Static RDF Files*	Utilizar una convención de nombrado y negociación de contenido para los ficheros en RDF. Reglas de reescritura en servidor web.
6.9	*Server-Side Configuration: MIME Types*	Servir el contenido apropiado de acuerdo a los tipos MIME.
6.10	*Making RDF Discoverable from HTML*	Permitir que desde HTML se puedan reutilizar las descripciones en RDF.
6.11	*Serving Linked Data as RDF Embedded in HTML Files*	Enriquecer HTML con las descripciones en RDF. Por ejemplo con RDFa.
6.12	*Serving RDF and HTML with Custom Server-Side Scripts*	Generación bajo demanda de RDF en el servidor.
6.13	*Serving Linked Data from Relational Databases*	Generación bajo demanda de RDF en el servidor *mapeando* a una base de datos.
6.14	*Serving Linked Data from RDF Triple Stores*	Utilizar las capacidades de un triple-store para ofrecer RDF.
6.15	*Serving Linked Data by Wrapping Existing Application or Web APIs*	Exportar los datos de las aplicaciones web existentes como RDF.

ID	Directriz	Descripción
6.16	*Testing and Debugging Linked Data*	Utilizar herramientas para verificar que se cumple con los principios de *Linked Data* y de la publicación de datos.
7	***Architecture of Linked Data Applications***	
7.1	*The Crawling Pattern*	Se navega por los recursos RDF con el objeto de extraer los datos para proveer servicios y datos a una capa superior.
7.2	*The On-The-Fly Dereferencing Pattern*	El patrón de los navegadores de *Linked Data* que extraen la información de los recursos en el momento en el que se necesita.
7.3	*The Query Federation Pattern*	Se crean consultas complejas ejecutadas sobre varias fuentes de datos que se presentan al usuario dentro de una aplicación.

Tabla 3.3: Consideraciones de Diseño *Linked Data*

Como se puede comprobar, en esta lista residen una serie de consideraciones a valorar en relación con la producción, publicación y consumo de datos enlazados. La implementación o el uso que se realice de cada uno de estos puntos depende de la estrategia seguida por la persona o institución implicada en el proceso de ofrecer sus datos bajo la iniciativa *Linked Data*. Finalmente y tras aplicar estos principios se puede utilizar el siguiente *checklist*, es similar al utilizado para añadir un conjunto de datos a la iniciativa LOD Cloud, para verificar, ver Tabla 3.4, si se han aplicado de forma correcta los principios de diseño propuestos por los autores y en consecuencia de *Linked Data*.

ID	Punto a cumplir	Descripción
1	*Does your data set links to other data sets?*	Existen enlaces externos desde las descripciones en RDF a otros externos que conformen un grafo global .
2	*Do you provide provenance metadata?*	Los datos disponen de información sobre su procedencia.
3	*Do you provide licensing metadata?*	Se establecen las restricciones necesarias sobre el uso de los datos.
4	*Do you use terms from widely deployed vocabularies?*	Las descripciones en RDF se basan en reutilizar clases y propiedades ya existentes y aplicadas ampliamente.
5	*Are the URIs of proprietary vocabulary terms dereferenceable?*	Las clases y propiedades definidas particularmente son referenciables.
6	*Do you map proprietary vocabulary terms to other vocabularies?*	Las clases y propiedades definidas se basan y relacionan con otras ya existentes y aplicadas ampliamente.
7	*Do you provide data set-level metadata?*	Se ofrece metainformación sobre el conjunto de datos en general
8	*Do you refer to additional access methods?*	Además del uso de URIs para referenciar los datos se proveen otros servicios como un *endpoint* de SPARQL.

Tabla 3.4: *Checklist Linked Data*

En conclusión, en este libro se ofrece una guía práctica y didáctica de los pasos a seguir para la aplicación de la iniciativa de *Linked Data* de forma correcta, asegurando las ventajas presentes en

los datos enlazados, minimizando las desventajas y proveyendo un entorno para la generación de servicios de alto valor y calidad. Se observa que tanto las directrices de la Tabla 3.3, como los puntos a cumplir de la Tabla 3.4, se pueden englobar en tres grandes procesos: producción, publicación y consumo de *Linked Data*.

3.2.6.2 *Linked Data Patterns*

Con el objetivo de proveer una forma estándar de transformar los datos siguiendo la iniciativa de *Linked Data* y disponer de unos criterios y soluciones estándar para modelar, publicar y consumir datos, se han publicado una serie de patrones [105] que resuelven los problemas más comunes que se pueden encontrar tales como el diseño de URIs, tipos de datos, etiquetado de recursos, etc. Al igual que en el caso anterior, en estos patrones se ofrecen una serie de guías para resolver problemas comunes que surgen en el momento de producir, publicar y consumir datos enlazados. Suponen una información muy valiosa, ya que permiten homogeneizar la creación de datos enlazados de tal forma que si una persona u organización asegura que ha seguido estos principios, se pueden construir aplicaciones que sepan qué tipo de datos se van encontrar y cómo, disminuyendo así el coste de reutilización de datos provenientes de terceros y asegurar la calidad de los mismos.

3.2.6.3 *Best Practices* del W3C

Se trata de una serie de documentos y buenas prácticas [214] dentro del grupo de trabajo del W3C *Government Linked Data Working Group* [215] cuya actividad se lanzó en el *Face 2 Face* de junio de 2011 y que consta de los siguientes objetivos:

- *The overarching objective is to provide best practices and guidance to create of high quality, re-usable Linked Open Data (LOD).*
- *Description of the full life cycle of a Government Linked Data project, starting with identification of suitable data sets, procurement, modeling, vocabulary selection, through publication and ongoing maintenance.*
- *Definition of known, proven steps to create and maintain government data sets using Linked Data principles.*
- *Guidance in explaining the value proposition for LOD to stakeholders, managers and executives.*
- *Assist the Working Group in later stages of the Standards Process, in order to solicit feedback, use cases, etc.*

Como grupo de trabajo su esfuerzo se centrará en cumplir los objetivos establecidos a través de la consecución de materiales como recomendaciones, notas, etc., en los siguientes ámbitos:

1. *Procurement*.
2. *Vocabulary Selection*.
3. *URI Construction*.
4. *Versioning* y *Stability*.
5. *Legacy Data*.
6. *Cookbook*.

Evidentemente y teniendo en cuenta los participantes en esta actividad, está claro que los puntos de actuación seleccionados son claramente estratégicos para la iniciativa de *Linked Data* y es por ello que aunque los resultados están en una etapa previa, es conveniente prestar suma atención a sus resultados con un doble objetivo, por una parte aplicarlos a futuros proyectos y por otra realimentar el esfuerzo realizado por sus participantes.

3.2.6.4 *Linked Data Cookbook* del W3C

Uno de los esfuerzos de la actividad comentada en la sección anterior, consiste en la elaboración de un libro de buenas prácticas [173] en cuanto a la producción, publicación y consumo de datos enlazados. Para ello se recopilarán las prácticas más comunes desde el punto de vista de la ingeniería que faciliten la adopción de *Linked Data* en diferentes entornos, asegurando una calidad y previniendo que el uso de datos enlazados no sea sinónimo de depuración.

Con carácter previo se han fijado una serie de pasos que se han de seguir para desplegar una infraestructura de datos enlazados acompañada de una serie de prácticas que aseguren la calidad y el proceso de adopción. Esta iniciativa es compatible con las buenas prácticas señaladas en las Secciones 3.2.6.1 y 3.2.6.2 y además ofrece una buena guía de todos los puntos que hay que tener en cuenta: diseño de URIs, vocabularios a reutilizar, herramientas, etc. Actualmente estas prácticas se están desarrollando y se dividen en 7 pasos que se presentan en la Tabla 3.5.

ID	Práctica	Descripción
1	*Model the Data*	*Identify, Model, Name and Test.*
2	*Name things with URIs*	*Following a name convention of current guides.*
3	*Re-use vocabularies whenever possible*	*Any given Linked Data set may include terms from an existing and widely used vocabulary.*
4	*Publish human and machine readable descriptions*	*Self-describing data suggests that "information about the encodings used for each representation is provided explicitly within the representation".*
5	*Convert data to RDF*	N/A
6	*Specify an appropriate license*	N/A
7	*Announce the new Linked Data Set(s)*	N/A

Tabla 3.5: *The 7 Best Practices for Producing Linked Data.*

3.2.6.5 *Government Linked Data-Life Cycle*

Siguiendo con las actividades que se desarrollan en este grupo de trabajo del W3C sobre datos enlazados en el entorno de la administración pública electrónica, se han identificado los siguientes ciclos de vida [155] propuestos por diversos autores:

- *Bernadette Hyland*, ver Figura 3.24.
- *Michael Hausenblas*, ver Figura 3.25.
- *Boris Villazón-Terrazas*, ver Figura 3.26
- *The DataLift Vision*, ver Figura 3.27

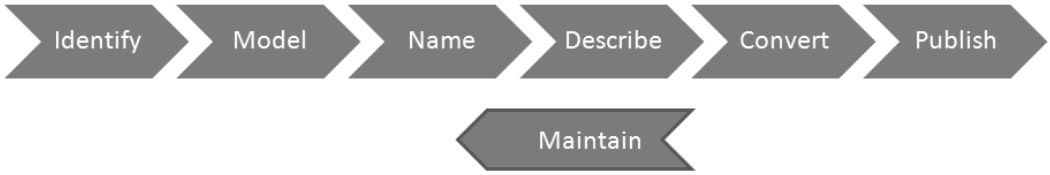


Figura 3.24: *Linked Data Lifecycle by B. Hyland*.

Figura 3.25: *Linked Data Lifecycle by M. Hausenblas*.

En todos ellos se recogen en grandes procesos los pasos a seguir para desplegar el modelo de *Linked Data* en una organización. Aunque existan cambios en la denominación, en el nivel de abstracción o en el orden de algunos procesos, el objetivo coincide en todos ellos. No obstante, estos modelos han surgido, probablemente, de la experiencia propia de los autores y aún estando en su etapa de desarrollo suponen un avance para afrontar la implantación de *Linked Data* como un proceso de ingeniería cuantificable.

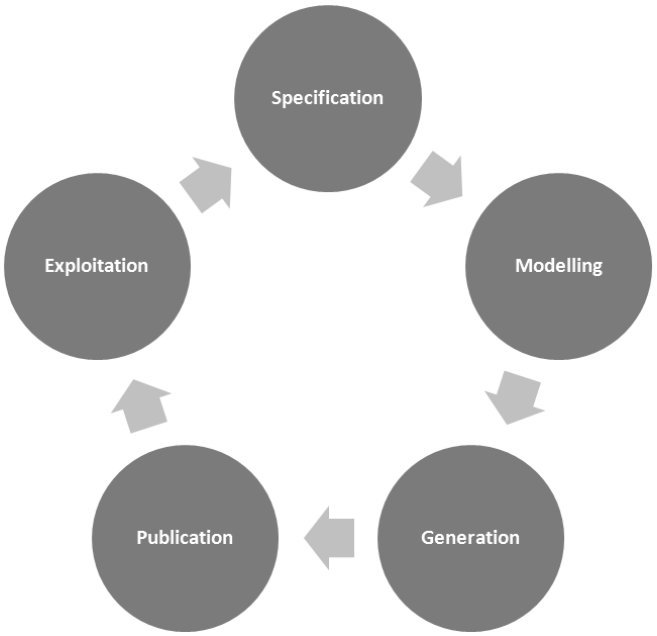


Figura 3.26: *Linked Data Lifecycle by B. Villazón-Terrazas.*

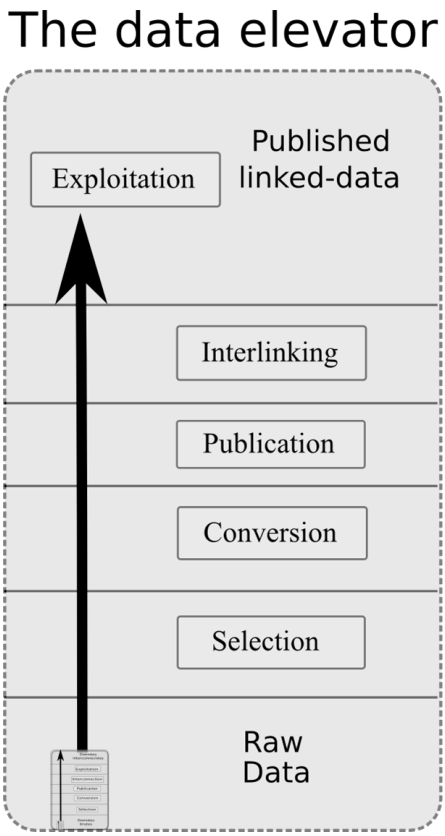


Figura 3.27: *DataLift Vision.*

3.2.6.6 *Publishing Open Government Data* del W3C

Se trata de un borrador de trabajo [20] (*W3C Working Draft 8 September 2009*) del grupo del W3C *eGovernment Interest Group* en el cual se hace una descripción somera sobre cómo publicar datos gubernamentales en la web. Reseñan fundamentalmente una serie de pasos para publicar datos sin grandes restricciones y de forma ágil, basándose en 3 pasos, ver Tabla 3.6, de carácter general pero que pueden servir especialmente a perfiles no técnicos presentes en las instituciones públicas.

ID	Paso	Descripción
1	*Step 1*	*The quickest and easiest way to make data available on the Internet is to publish the data in its raw form.*
2	*Step 2*	*Create an online catalog of the raw data (complete with documentation) so people can discover what has been posted.*
3	*Step 3*	*Make the data both human- and machine-readable.*

Tabla 3.6: *Straightforward Steps to Publish Government Data.*

De la misma forma, en el propio documento se exponen otros factores a valorar en el momento de publicación de los datos, en la línea de la iniciativa de *Linked Data*. Simplemente se trata de una guía básica prácticamente de carácter divulgativo.

3.2.6.7 *LOD2 Stack*

Dentro del proyecto europeo LOD2 [126] con nº de contrato: 257943, con duración desde septiembre de 2010 hasta agosto de 2014 y una financiación de más de 7 millones euros se está desarrollando una serie de documentos y tecnología asociada a la iniciativa de *Linking Open Data* con el objetivo de proveer buenas prácticas y herramientas que den soporte a toda la cadena de producción, publicación y consumo de datos enlazados en distintos dominios. Uno de los últimos resultados que han conseguido y que se encuentra en continua evolución es la denominada *LOD2 Stack* [285] que contiene una serie de herramientas para cada una de las etapas necesarias en la apertura de datos como *Linked Data*.

***OntoWiki* [287].** *OntoWiki is a tool providing support for agile, distributed knowledge engineering scenarios.*

***PoolParty* [253].** *PoolParty is a thesaurus management system and a SKOS editor for the Semantic Web including text mining and linked data capabilities.*

***Sig.ma* [290].** *Sig.ma is a tool to explore and leverage the Web of Data.*

***Comprehensive Knowledge Archive Network (CKAN) [122].** CKAN is a registry or catalogue system for datasets or other "knowledge" resources.*

***D2R Server* [37].** *D2R Server is a tool for publishing relational databases on the Semantic Web.*

***DBpedia Extraction* [43].** *DBpedia is a community effort to extract structured information from Wikipedia and to make this information available on the Web.*

***DL-Learner* [201].** *DL-Learner is a tool for supervised Machine Learning in OWL and Description Logics.*

***MonetDB* [49].** *MonetDB is an open-source high-performance database system that allows to store relational, XML and RDF data.*

SemMF. *SemMF is a flexible framework for calculating semantic similarity between objects that are represented as arbitrary RDF graphs.*

Silk Framework **[45].** *The Silk Linking Framework supports data publishers in setting explicit RDF links between data items within different data sources.*

Sindice **[291].** *Sindice is a state of the art infrastructure to process, consolidate and query the Web of Data.*

Sparallax **[277].** *Sparallax is a faceted browsing interface for SPARQL endpoints, based on Freebase Parallax.*

Triplify **[288].** *Triplify provides a building block for the "semantification" of Web applications.*

OpenLink Virtuoso **[180].** *Virtuoso is a knowledge store and virtualization platform that transparently integrates Data, Services, and Business Processes across the enterprise.*

WIQA **[40].** *The Web Information Quality Assessment Framework is a set of software components that empowers information consumers to employ a wide range of different information quality assessment policies to filter information from the Web.*

Como se puede observar este grupo de herramientas dan soporte a todo el ciclo de vida de los datos enlazados y proveen una plataforma genérica en la cual cualquier persona o entidad pueda integrar un nuevo conjunto de datos, desde recursos ya disponibles en RDF hasta documentos de texto. La realización de este proyecto es de una suma transcendencia para la comunidad, ya que en el despliegue de una infraestructura basada en datos enlazados se suele incurrir en las mismas dudas y problemas, por lo que disponer de una guía que parta desde un punto de vista teórico hasta su realización práctica, supone un gran avance respecto a las especificaciones teóricas, recetas y buenas prácticas que se han mencionado en las secciones anteriores.

Por otra parte, las herramientas desarrolladas se encuentran enclavadas dentro de un proceso o ciclo de vida de datos enlazados, ver Figura 3.28. Se trata de un modelo iterativo y realimentado, en el que se completan las distintas fases para cumplir con la producción, publicación y consumo de datos, utilizando las herramientas previamente comentadas y atacando las principales barreras que se suelen encontrar, así como aplicando los principios de diseño de *Linked Data*.

Para la aplicación de los principios de *Linked Data* es realmente conveniente prestar atención al trabajo desarrollado en este proyecto, no obstante, teniendo en cuenta que tan sólo se cuenta con un año de desarrollo es complicado aplicar su enfoque de forma integral.

3.2.6.8 *Toward a Basic Profile for Linked Data* de IBM

Dentro de la actividad de IBM en el campo de *Linked Data*, se ha identificado la necesidad de fijar una serie de reglas para aplicar esta iniciativa de forma cuantificable. El objetivo de esta guía es llegar a introducir dentro de una herramienta como *Rationale* (también perteneciente a IBM) el modelo arquitectónico propuesto por los datos enlazados, cumpliendo las directrices de la misma con la meta de obtener la tecnología necesaria que sirva para integrar datos de forma ágil en las aplicaciones.

El término utilizado por IBM es *Basic Profile Resources* [234] que se define como recursos *Linked Data* accesibles mediante HTTP que siguen una serie de patrones y convenciones comunes. En general, estos recursos dependen del dominio en el cual estén definidos y tan sólo se delimitan como comunes algunos considerados transversales a cualquier dominio. Estos recursos siguen una serie de reglas, ver Tabla 3.7, y reutilizan vocabularios existentes con el objetivo de cumplir con la iniciativa de *Linked Data*.

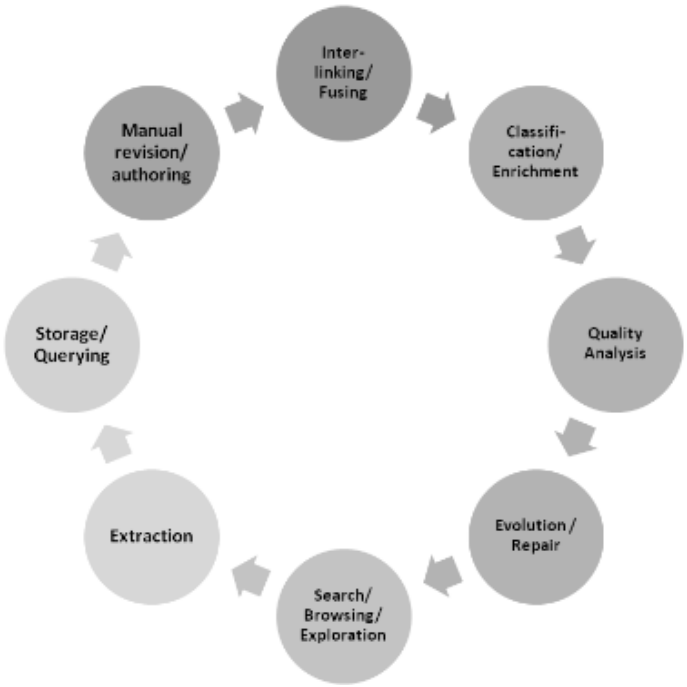


Figura 3.28: *Linked Data LifeCycle* (extraída de LOD2 Demo).

ID	Rule	Descripción
1	*Basic Profile Resources are HTTP resources*	*They can be created, modified, deleted and read using standard HTTP methods.*
2	*Basic Profile Resources use RDF to define their states*	*The state of a Basic Profile Resource (in the sense of state used in the REST architecture) is defined by a set of RDF triples.*
3	*You can request an RDF/XML representation of any Basic Profile Resource*	*The resource might have other representations.*
4	*Basic Profile clients use Optimistic Collision Detection during update*	*Because the update process involves getting a resource first, and then modifying it and later putting it back on the server, there is the possibility of a conflict (for example, another client might have updated the resource since the GET action). To mitigate this problem, Basic Profile implementations should use the HTTP If-Match header and HTTP ETags to detect collisions.*
5	*Basic Profile Resources use standard media types*	*Basic Profile does not require and does not encourage the definition of any new media types.*
6	*Basic Profile Resources use standard vocabularies*	*Basic Profile Resources use common vocabularies (classes, properties, and so forth) for common concepts.*
7	*Basic Profile Resources set* `rdf:type` *explicitly.*	*A resource's membership in a class extent can be derived implicitly or indicated explicitly by a triple in the resource representation.*
8	*Basic Profile Resources use a restricted number of standard data types*	*RDF does not define data types to be used for property values, so Basic Profile lists a set of standard datatypes to be used in Basic Profile.*

ID	Rule	Descripción
9	*Basic Profile clients expect to encounter unknown properties and content*	*Basic Profile provides mechanisms for clients to discover lists of expected properties for resources for particular purposes.*
10	*Basic Profile clients do not assume the type of a resource at the end of a link*	*Many specifications and most traditional applications have a çlosed model,"by which we mean that any reference from a resource in the specification or application necessarily identifies a resource in the same specification.*
11	*Basic Profile servers implement simple validations for Create and Update*	*Basic Profile servers should try to make it easy for programmatic clients to create and update resources.*
12	*Basic Profile Resources always use simple RDF predicates to represent links*	*Basic Profile makes it very simple to know how links will appear in representations and also makes it very simple to query them.*

Tabla 3.7: *Basic Profile Resources.*

Repasando esta lista de reglas se puede observar como algunas afectan a los principios 2º y 3º de *Linked Data* sin modificar su significado pero especificando de forma más concreta el funcionamiento esperado. La importancia de este reciente artículo reside en varios puntos: es realizado por IBM, es coherente con las reglas de *Linked Data* y ofrece un carácter práctico desde el punto de vista de la ingeniería para implementar, o en este caso, incluir la arquitectura y modelo de trabajo de datos enlazados en una plataforma existente como es *Rationale*.

3.2.6.9 Metodología y Proceso de Adopción de *Linked Data* en la Biblioteca del Congreso de Chile

En este trabajo se propone una metodología [63, 64] y proceso de adopción de la iniciativa de *Linked Data* en el contexto de las Administraciones Públicas y concretamente en la Biblioteca del Congreso de Chile para la publicación de la legislación actual e histórica. En general, se trata en realidad de la especificación de una infraestructura para dar soporte a los datos enlazados y un proceso de generación con diferentes etapas, que parten de un caso particular motivador pero que se podrían aplicar a un contexto genérico. Una de las diferencias respecto al proyecto LOD2 reside en las herramientas seleccionadas para llevar a cabo los distintos procesos de producción, publicación y consumo de datos enlazados y por otra parte la adición de servicios de consumo de datos especialmente dirigidos a los usuarios de la Administración, como es la visualización de las normas.

También, hay que destacar el proceso de adopción definido en este enfoque, ver Figura 3.29, en el cual quedan recogidos los pasos para promocionar las bases de datos legislativas utilizando datos enlazados. Este proceso es similar al propuesto por el W3C, ver Sección 3.2.6.4, y ciclo de vida definido en el proyecto LOD2, ver Figura 3.28, sin embargo difiere en el nombrado de los procesos y las herramientas a utilizar en cada una de las fases.

3.2.6.10 *Talis Platform*

En este caso la plataforma Talis [182] es un servicio o producto que sirve para la creación de una infraestructura de *Linked Data* en la nube. Ha tenido un gran éxito ya que sus creadores son grandes impulsores de la iniciativa (por ejemplo son los autores de *Linked Data Patterns*) y han trabajado en casos de enorme repercusión, como la apertura de datos públicos del Gobierno del Reino Unido. Ofrecen una *suite* de herramientas y servicios para cumplir las directrices de *Linked Data*, teniendo

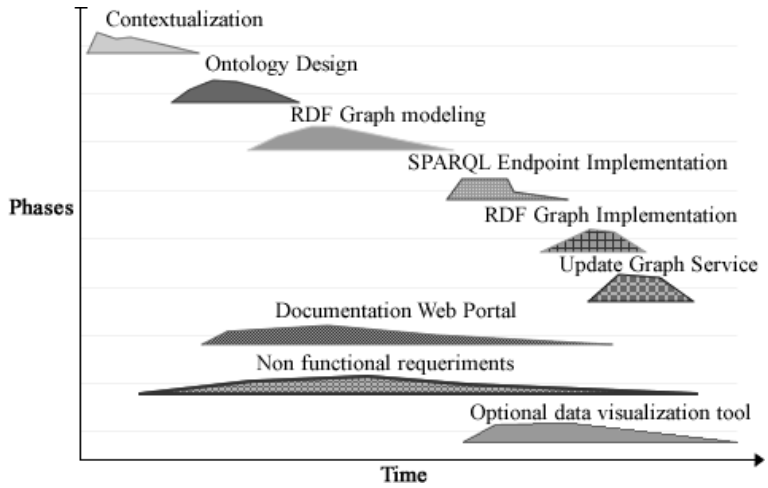


Figura 3.29: Proceso de implantación de *Linked Data* en la Biblioteca del Congreso de Chile.

en cuenta la mayor parte de la casuística presente en la producción, publicación y consumo de datos enlazados. En resumen, se provee una plataforma que cubre toda la cadena de valor de los datos enlazados y cumple todas las directrices necesarias.

Entre las características más llamativas que ofrecen dentro de esta plataforma se encuentran las siguientes (extraídas de la propia página web de Talis):

- *A simple, consistent web API for storing, managing and retrieving both structured and unstructured data.*
- *Flexible, schema-free metadata that allows applications to be easily evolved.*
- *A range of data access and query options enabling easy integration into both new and existing applications.*
- *Access control options to support hosting of both public and private data.*
- *A data hosting solution that is founded on open internet standards and web architectural best practices.*
- *Software as a Service, enabling rapid development with zero deployment costs.*
- *Low, even free, utility based pricing for services and hosting allowing costs to grow with usage.*
- *A highly available and scalable infrastructure to ensure that the repository grows in line with your applications needs.*

Evidentemente este tipo de soluciones son de extraordinario interés para las Administraciones Públicas, ya que consiguen un producto llave en mano con todas las capacidades necesarias para disponer de una infraestructura de datos enlazados de última generación. Es interesante destacar esta plataforma por dos motivos principales: sus creadores son relevantes desde un punto de vista científico y han realizado una gran transferencia tecnológica desde el campo de la investigación al industrial.

En la misma línea de la plataforma Talis se encuentran otras empresas que ofrecen productos y herramientas de alto valor como son Virtuoso de OpenLink o ToqQuadrant. Todos ellos son miembros activos en la comunidad de *Linked Data* y están presentes en los principales grupos de trabajo de esta iniciativa en el W3C.

3.2.6.11 *Linked Open Data: The Essentials*

Este libro [14] realizado en colaboración entre REEEP (*Renewable Energy and Energy Efficiency Partnership*) y la compañía *Semantic Web* es un manual que da respuesta a algunas de las preguntas comunes que surgen en el despliegue de una infraestructura de datos enlazados en el seno de una organización. En general, se trata de una recopilación de las buenas prácticas que se han revisado en los anteriores apartados y que focaliza en los beneficios de la publicación de datos enlazados para las organizaciones. La parte más destacada de este libro, ver Figura 3.30, se centra en la descripción de las tareas a realizar y de los posibles servicios haciendo hincapié en las tareas de publicación y consumo de datos enlazados.

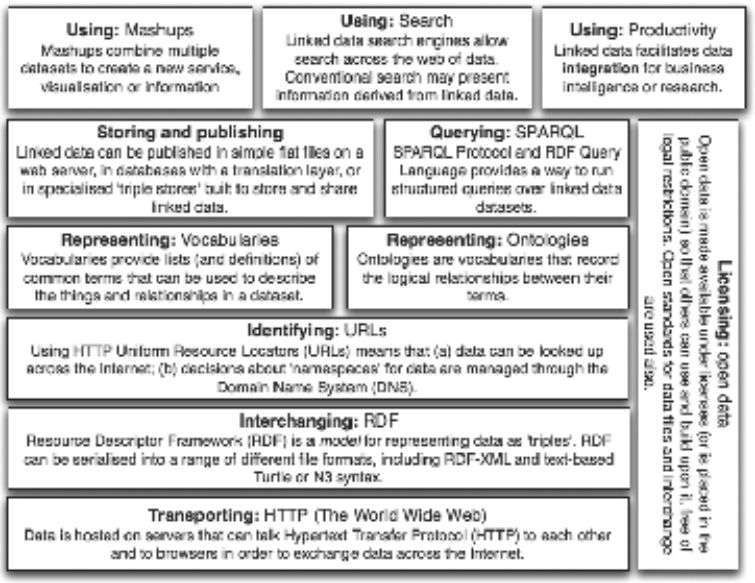


Figura 3.30: *Elementos of the Linked Open Data Puzzle*.

3.2.6.12 **Documentación específica**

Desde distintas organizaciones tales como Administraciones Públicas y Universidades, así como personalidades relevantes en esta materia se ha promovido la realización de proyectos, guías y artículos que tratan diversos aspectos relacionados con la iniciativa de *Linked Data*. Todos ellos tratan de aportar nuevos enfoques, herramientas y documentación a los distintos procesos implicados en el despliegue de esta iniciativa, entre los que pueden destacarse:

- "*A Proposal for Governmental Data URIs*" realizado por Gregory Todd Williams, Tim Lebo y Alvaro Graves.
- "*Designing URI Sets for the UK Public Sector*" [79] realizado por "The Cabinet Office" del Reino Unido.
- "*Federal Information Security Management Act (FISMA)*" proyecto realizado por el "National Institute of Standards and Technology" de Estados Unidos.
- La documentación disponible en la Biblioteca del Congreso de Estados Unidos.
- Los artículos de Jeni Tennison sobre distintos aspectos relativos al diseño de URIs, etc.
- El *framework* SERIMI [8] de reconciliación de entidades.

- Los documentos y software elaborados por la empresa Epimorphics.
- Las guías de datos abiertos y enlazados de las distintas Administraciones Públicas, tanto a nivel estatal: España, Reino Unido, Francia, Alemania, etc., como a nivel regional: Asturias, País Vasco o Cataluña o incluso local como en el Ayuntamiento de Zaragoza.

En general se trata de documentos o herramientas que nacen de la necesidad y experiencia en la resolución de ciertos problemas relacionados con la apertura de datos y su enlazado, que se reproducen en cada escenario en el cual se aplica esta iniciativa. Todo este trabajo y esfuerzo supone un gran valor para la comunidad de desarrolladores y de personas implicadas en la aplicación de estos principios.

3.2.7 Escenarios y Casos de Uso de Éxito

La irrupción de la corriente de *Linked Data* ha conllevado la realización práctica y real de una parte de la denostada Web Semántica. Muchas son las instituciones, administraciones públicas, empresas privadas, universidades, hospitales, etc., que bajo esta corriente están liberando sus datos en el entorno web siguiendo las directrices marcadas por *Tim Berners-Lee*. El principal objetivo consiste en la apertura de datos, para que una vez publicados sirvan como fuente para la creación de aplicaciones agregando distintos recursos (*mashups*), facilitando la transparencia de comportamiento en ciertos organismos, suministrando servicios de valor añadido basados en el contexto del usuario o simplemente desde un punto de vista de tendencia, para obtener presencia en la nueva Web de Datos.

No obstante, este nuevo enfoque conlleva varios desafíos a nivel científico que están siendo abordados actualmente, tales como el procesamiento de grandes cantidades de datos de forma eficiente, gestión-formalización-explotación del conocimiento subyacente mediante reglas, razonamiento distribuido, *stream-reasoning*, *real time linked data*, *complex event processing*, reconciliación de entidades, visualización etc., los casos de uso de aplicación de estas investigaciones representan un abanico muy amplio: recomendación de recursos (películas, libros, etc.), análisis de sentimientos y opiniones, domótica, *smart-cities*, computación ubicua, *green computing*, sistemas de soporte a la decisión en campos como *e-Health*, etc.

Hasta el momento esta iniciativa ha estado muy ligada al mundo técnico, es decir, tanto la publicación como el consumo de los datos estaba muy orientado a un perfil fundamentalmente técnico. No obstante, al igual que con la web que hoy conocemos, la Web de Datos todavía no trascendido al gran público, es por ello que surgen varios desafíos [84] para conseguir el despegue definitivo de esta corriente:

1. Necesidad de una capa de presentación. Si una persona buscara cierto recurso que tiene definido según unas necesidades, la expresión de esta consulta implicaría características de sistemas de *Information Retrieval* o tareas de análisis de sentimientos. Concretamente: a) descubrimiento de qué fuentes pueden tener disponible esa información; b) validación y cooperación entre las distintas fuentes de datos disponibles; c) consulta a las fuentes de datos seleccionadas y d) análisis, presentación y manejo de las vistas de los resultados. En este sentido, la tendencia actual reside en "ocultar" la existencia de varios datasets detrás de los sistemas de búsqueda, así como la consulta y manejo de los resultados, de tal manera que el usuario no es consciente de las oportunidades de explotación de datos de las que podría disponer. Es necesario aclarar que el conocimiento de la tecnología no debe ser la clave para la explotación de la misma por el usuario final, pero si que es necesario proveer los métodos y herramientas adecuadas para que el usuario obtenga el máximo partido de la información subyacente. De esta manera si la tendencia de *Linked Data* supone una revolución respecto a la actual web de documentos, se

deberían proveer nuevos mecanismos de acceso a la información y no restringirse a los tradicionales modelos de interacción con el usuario (por ejemplo un formulario de búsqueda).

2. Combinación efectiva de los distintos *datasets*. Teniendo en cuenta que existe un catálogo de datos y vocabularios en constante crecimiento, es necesario proveer los mecanismos necesarios (alineación, *mapeo* y fusión) para que la combinación de los mismos sea eficiente de modo que la identificación de recursos similares, el encaje mediante propiedades o la mezcla de recursos sea intuitiva, tanto desde un punto de vista técnico como del usuario final. Las técnicas que permiten llevar a cabo este desafío no son deterministas y en muchos casos se basan en algoritmos que chequean la estructura de los recursos, o bien realizan comparaciones basadas en procesamiento del lenguaje natural entre las descripciones de los recursos.

En el contexto de *Linked Data* se manifiestan dos principales vías de actuación sobre los datos: publicación y consumo, concretamente en el campo de consumo de datos surgen varios interrogantes como:

- Gestión de datos a nivel web.
- Procesamiento de consultas federadas [54].
- Búsqueda.
- Descubrimiento de *datasets* y recursos: URIs, datos adicionales o datasets relevantes para una consulta.
- *Datasets* evolutivos: cómo afectan los cambios en los datasets a las consultas, monitorización.
- Razonamiento e inferencia de conocimiento en la Web de Datos.
- Calidad de los datos: evaluación de la calidad de la información, confianza y *provenance*.
- Interfaz de Usuario para la interacción con la Web de Datos: interacción y usabilidad, visualización, interfaces basadas en lenguaje natural.

Por otra parte, tanto en las Administraciones Públicas como en otros dominios se está trabajando exhaustivamente en el campo de la reutilización de datos y servicios basados en semántica, en la actualidad se pueden señalar algunos ejemplos:

- Gestión de bibliotecas digitales [130], la Biblioteca del Congreso de Chile o del Congreso de Estados Unidos.
- Sistemas de búsqueda y recomendación basados en perfiles de usuario enlazados [200].
- Sistemas para el soporte a la decisión [131], especialmente en el dominio de *e-Health*.
- Gestión, publicación y servicios sobre información financiera [106, 136].
- Otros múltiples campos de actuación: GIS+*Linked Data*, turismo, tráfico, educación, domótica, *semantic sensors*, etc.

En último término todas las iniciativas basadas en *Linked Data* convergen al objetivo principal de la misma:

> *Linked Data is about using the Web to connect related data that wasn't previously linked, or using the Web to lower the barriers to linking data currently linked using other methods.*

Teniendo en cuenta el valor añadido de la visualización de datos para Linked Data, se pueden establecer una serie de requisitos [85] a cumplir, que actualmente están parcialmente cubiertos por algunas herramientas, no obstante, se puede establecer una guía de características necesarias para el manejo de datos enlazados y su visualización, siendo uno de los grandes elementos de investigación actual.

- Requisitos de alto nivel: a) capacidad para generar vistas agregadas de los datos; b) soporte para la creación de filtros y c) soporte para la visualización detallada de los recursos.
- Requisitos de consumo de datos: a) manejo de datos multidimensionales; b) manejo de datos jerárquicos;c) generación de estructuras de representación: grafos, árboles, etc.; d) identificación de las relaciones significativas en los datos; e) extracción de datos y nuevas relaciones y f) contextualización de usuario y multidispositivo.
- Requisitos directamente relacionados con *Linked Data*: a) navegación entre los distintos datasets; b) exploración de los datos para obtener consciencia de la estructura y de las relaciones; c) exploración que permita análisis con el objetivo de identificar incoherencias, etc.; d) consulta transversal a través de los distintos datasets mediante un lenguaje formal y e) capacidad para extraer partes de los datos para su posterior reutilización.
- Requisitos relacionados con la navegación semántica: a) navegación facetada y contextualizada del usuario; b) exploración del conocimiento, inferencia, etc.; c) consulta avanzada con filtros de lenguaje natural; d) análisis detallado y explicación de las inferencias y e) presentación de resultados de análisis.

3.3 Tendencias actuales en Semántica

El creciente uso de Internet durante los últimos años ha puesto de manifiesto un nuevo entorno de ejecución para las aplicaciones, utilizando como nueva plataforma la web, el gran sistema distribuido. Nuevas tecnologías y paradigmas están emergiendo para dar soporte al desarrollo y despliegue de aplicaciones y servicios, así como para la publicación de datos e información. Los modelos de desarrollo están cambiando a un estilo más colaborativo en el cual las empresas ofrecen su software como servicios (*Software as a Service*-SaaS) materializado a través del paradigma de *cloud computing* [9], implementado con tecnología de servicios con el objetivo de que terceros puedan utilizar estos servicios para la construcción de aplicaciones agregadas con valor añadido.

En este sentido, tal como se ha señalado, iniciativas como la Web Semántica, que a través de modelos y formatos de datos de conocimiento compartido unificados, intentan elevar el significado de los elementos y recursos que están disponibles en la web, con el objetivo de mejorar la integración e interoperabilidad entre aplicaciones, están impulsando la implantación de este enfoque. Dentro de la iniciativa de Web Semántica hay que destacar dos esfuerzos:

1. La iniciativa *Linked Data* que tal y como se ha descrito propone la publicación de datos enlazados siguiendo el modelo RDF para facilitar la creación de una web de datos en la que éstos se puedan mostrar, intercambiar y conectar a través de URIs. La tendencia actual de publicación de datos enlazados está marcando una evolución en la interoperabilidad de aplicaciones, con el consiguiente efecto que conlleva para las relaciones B2B, B2C o A2A. Entre los casos de éxito podrían destacarse: administración electrónica (iniciativa de *Open Government Data*, contratación pública de bienes y servicios (*e-Procurement*), oferta formacional , contextualización de aplicaciones, *mashups*, etc.

2. El desarrollo de lenguajes y formalismos lógicos para representar el conocimiento sobre un universo de discurso, permitiendo la inferencia de nuevos datos a partir de los datos ya publicados. En este contexto se ha impulsado nuevamente el uso de las técnicas de razonamiento y de sistemas basados en conocimiento, como las ontologías y los sistemas basados en reglas (Ontobroker, XSB, etc.) o de producción (Drools , JRules , etc.). La aplicación de estos sistemas está ampliamente asentada en la resolución de diversos problemas (diagnóstico, planificación, reglas de negocio, etc.) pero siempre utilizando un enfoque para la representación del conocimiento y de los datos, en muchos casos específico y no estandarizado. Por tanto, debe tenerse en cuenta la flexibilidad, tanto para la integración como para la interoperabilidad de las aplicaciones. Una arquitectura orientada a servicios sobre la plataforma web, utilizando los protocolos actuales y que se beneficie de las iniciativas relativas a la Web Semántica puede dar respuesta a este punto clave. La implantación de un sistema basado en conocimiento, concretamente de sistemas basados en reglas y de razonamiento, que hagan uso de una infraestructura estandarizada y cooperativa redunda en gran beneficio para la resolución de problemas basados en conocimiento declarativo y compartido, mejorando tanto la independencia tecnológica de las aplicaciones como la experiencia de usuario.

Al amparo de la visión de la Web Semántica, el W3C ha promovido la creación de varias recomendaciones que intentan ofrecer soluciones para las diferentes capas de la arquitectura. En concreto, RDF es el lenguaje de representación básico que permite representar tripletas de la forma sujeto-predicado-objeto. Dichas tripletas forman un grafo dirigido que puede integrarse automáticamente con otros grafos obtenidos de otros servidores. Otra de las tecnologías propuestas por el W3C es RDFS, que permite la definición de clases, propiedades e individuos y ofrece unos mecanismos básicos de inferencia mediante reglas. RDFS facilita la creación e integración de vocabularios pero carece de expresividad suficiente para describir relaciones avanzadas como complementos de conjuntos o cardinalidades. Las limitaciones expresivas de RDFS propiciaron la definición de OWL, un lenguaje de definición de ontologías basado en lógica descriptiva. La versión 2 de OWL llegó a status de recomendación del W3C en octubre del año 2009. Una de las principales mejoras de esta versión, estaba encaminada a la resolución del compromiso entre la expresividad y la complejidad de los razonamientos, mediante la definición de perfiles o fragmentos computacionales. Los tres perfiles de OWL2: EL, QL y RL, son sublenguajes del mismo que permiten alcanzar una complejidad polinómica para tareas de razonamiento estándar limitando la expresividad. La combinación de OWL con lenguajes basados en reglas que incluyen la negación, como los *dl-programs* [232] han sido desarrollados en las redes de excelencia *REWERSE* y *Knowledge web* y definen la interoperabilidad entre las reglas y las ontologías.

El problema del intercambio de reglas entre los distintos sistemas de razonamiento e inferencia también ha sido abordado recientemente, así encontramos la recomendación del W3C RIF de 22 de junio del año 2010. RIF es, de nuevo, una familia de lenguajes (*Core,Production Rule Dialect, BLD Basic Logic Dialect, Datatypes and Built-Ins, Framework for Logic Dialects*, etc.) con diferente expresividad, cuyo objetivo es convertirse en *lingua franca* para el intercambio de conocimiento basado en reglas en la web. El formato utilizado por RIF es XML y su combinación con ontologías permite que las reglas y el modelo de datos sobre los que se van a aplicar las reglas se pueden intercambiar entre distintos actores. Para interpretar RIF es necesario realizar una traducción desde este vocabulario al motor de inferencia deseado. En el ámbito de intercambio de reglas hay que resaltar el antecesor de RIF, RuleML que es una iniciativa internacional sin ánimo de lucro, que cubre los aspectos del intercambio y la interoperabilidad de reglas. Esta iniciativa mantiene una estrecha relación con los grupos de OASIS en reglas, así como con ISO Common Logic (estándar en el año 2007). RuleML, como grupo, también contribuye en OMG a SBVR, específicamente en el apartado de *Production Rule Representation* (PRR) cuya última versión data de diciembre del año 2009.

Para la consulta de datos RDF se ha desarrollado SPARQL, un lenguaje de consulta y un protocolo de acceso que permiten definir un terminal (o *endpoint*) en el que se publican conjuntos de

datos (o *datasets*) RDF y que están disponibles como servicios en la web. Actualmente se está trabajando en vocabularios para definir datasets y poder enlazarlos entre sí de forma sencilla. Con el uso de SPARQL han aparecido propuestas para la definición de reglas de producción con este lenguaje (extensión *SPARQL with Updates-SPARUL)* de modo que se ejecuten directamente sobre una base de datos en RDF, como puede ser SPARQL-Rules [252], también existen enfoques para la consulta de ontologías OWL con SPARQL [273] y en la especificación que se elabora actualmente, SPARQL 1.1, se dispone de un vocabulario para definir los servicios disponibles. Finalmente y con el objetivo de personalizar la vista de las aplicaciones por el usuario y su contextualización se ha aplicado el uso de reglas en formato JSON [135].

La evolución de los formalismos para definir ontologías lleva emparejado el desarrollo de razonadores que puedan llevar a cabo las inferencias necesarias. Desde el pionero KL-ONE hasta este momento, se han implementado múltiples razonadores basados en lógica descriptiva siguiendo diferentes técnicas. En la actualidad, se pueden destacar Fact++ [289], Pellet [274] y RacerPro [148] que se basan en la técnica conocida como *semantic tableaux*. A pesar de las altas complejidades, en el caso peor los algoritmos de razonamiento utilizados en estos sistemas son capaces de resolver muchas tareas prácticas gracias al uso de diversas optimizaciones. Para resolver dichas limitaciones, especialmente, al tratar con grandes cantidades de datos, se han buscado técnicas alternativas como el algoritmo de resolución utilizado en KAON2 [231], el sistema *hipertableau* empleado en HermiT [233], las técnicas de eliminación de tipos [267] o las recientes técnicas basadas en eliminación de consecuentes [272].

La utilización de reglas se ha propuesto como una alternativa para la implementación de razonadores. En este caso, las definiciones de la ontología son compiladas a un conjunto de reglas, que se aplica al conjunto de datos para obtener las inferencias correspondientes. La principal ventaja de estos razonadores es que se basan en técnicas ya conocidas en el ámbito de la programación lógica con diversas implementaciones disponibles, constituye sin embargo una desventaja, que generalmente es necesario utilizar subconjuntos de OWL, como el conocido OWL Horst [286]. Uno de los mayores retos de la Web Semántica es la búsqueda de técnicas que mejoren la escalabilidad de los razonadores, en esta línea, han surgido trabajos que proponen la utilización de tecnologías distribuidas para afrontar dicha complejidad. Por ejemplo, en [275] se propone la implementación de un sistema de inferencia paralelizable sobre OWL Horst, mediante un particionado de las reglas. Con el objetivo de mejorar la escalabilidad, en [298] se propone un algoritmo para realizar inferencias sobre RDFS que se aplica a un *benchmark* de 10,000 tripletas, proponiendo como trabajo futuro la posible aplicación de MapReduce [100, 293]. En [294] se describe una implementación basada en el algoritmo MapReduce logrando realizar el cierre parcial de 864 millones de tripletas en RDFS, en una hora utilizando 32 procesadores. Recientemente, los mismos autores desarrollaron el sistema WebPIE [292], el cual ha sido capaz de realizar inferencias sobre 1 billón y medio de tripletas, en 6,1 horas utilizando 32 nodos, mediante reglas de OWL Horst. El sistema SAOR se ha implementado para dar soporte al buscador semántico SWSE, incorporando un razonador sobre OWL Horst [165], dicho sistema ha diseñado un nuevo algoritmo distribuido, mejorando la escalabilidad de la implementación [167]. Finalmente, también se está trabajando [154] en la conjunción de *Linked Data* y *cloud computing* para el procesamiento de grandes cantidades de datos.

3.4 *e-Procurement* y Semántica

La construcción de un modelo semántico sobre un dominio de negocio concreto consiste en el desarrollo de un sistema de conocimiento que describa las entidades y propiedades, así como las relaciones lógicas que existen entre las mismas. Un aspecto, muchas veces minusvalorado, es que en estos modelos de dominio es necesario describir también los procesos, las restricciones y en general, los aspectos dinámicos del dominio. Este tipo de modelos formalizados mediante un lenguaje lógico, como ya se ha descrito, se denominan "ontologías". La selección de la lógica apropiada para la

modelización de la base de conocimiento no es una cuestión sencilla y se deberán contemplar diferentes factores tales como grado de computabilidad, decidibilidad, soporte razonadores, etc. Existen diferentes vocabularios y lenguajes que permiten modelar un cierto dominio, algunos de los cuales ya se han repasado (RDF, RDFS, OWL y WSML) en la Sección 3.1.2.2, no obstante, existen otros estrechamente ligados a entornos de negocio, que a continuación se presentan.

RIF (*Rule Interchange Format*) [48]. RIF constituye una familia de lenguajes con diferente expresividad, cuyo objetivo es convertirse en *lingua franca* para el intercambio de conocimiento basado en reglas en la Web. Define la compatibilidad con documentos RDF y OWL, además está especialmente diseñado para integrarse con sistemas de inferencia basados tanto en reglas de producción, como de Programación Lógica. La ventaja de RIF respecto de OWL es una mayor expresividad, que permite expresar conocimiento causal y procedimental.

SCOR Model (*Supply-Chain Operations Reference-Model*). Es un modelo conceptual de referencia para la especificación y formalización de los procesos de negocio asociados a las cadenas de suministro. Este modelo define los procesos de planificación, fabricación, entrega, evolución e inventarios de productos en una cadena logística.

SBVR (*Semantics of Business Vocabulary and Business Rules*). Es un estándar (ISO 704/1087) desarrollado por el *Object Management Group* (OMG). Define un metamodelo para el desarrollo de modelos semánticos de vocabularios y reglas de negocio. La idea es que a partir del lenguaje natural se puedan expresar vocabularios y reglas de negocio en un dominio concreto. La construcción de un modelo de negocio mediante el SBVR recoge el vocabulario de conceptos del dominio y la definición de lógica formal asociada.

Clasificaciones de productos de comercio electrónico [202, 203, 271]. Para facilitar el intercambio automático de información y la anotación de los diferentes productos y documentos que forman parte del comercio electrónico, distintos sectores de actividad económica han desarrollado sus propias terminologías o vocabularios [262] controlados, para armonizar mediante códigos unívocos la identificación de los objetos de negocio. Ejemplos de estos vocabularios son: el estándar eCl@ss o la clasificación UNSPSC entre otros, ver Sección 2.10.

ebXML (*Electronic Business using eXtensible Markup Language*). Es un lenguaje XML elaborado por OASIS y UN/CEFACT, también consta de una infraestructura que permite la comunicación entre entidades participantes en transacciones de negocio electrónico, favoreciendo la interoperabilidad y la seguridad de una forma homogénea entre las distintas partes. La propuesta original cubría cinco capas: *Business processes*; *Collaboration protocol agreements*; *Core data components*; *Messaging* y *Registries and repositories*.

eBXML ha sido aprobado por ISO en un conjunto de especificaciones, ISO 15000: ISO 15000-1: ebXML *Collaborative Partner Profile Agreement*; ISO 15000-2: ebXML *Messaging Service Specification*; ISO 15000-3: *ebXML Registry Information Model*; ISO 15000-4: ebXML *Registry Services Specification*; y ISO 15000-5: ebXML *Core Components Technical Specification*, Version 2.01.

XBRL (*extensible Business Reporting Language*). Es una norma elaborada en el año 1998 por Charles Hoffman, contable y auditor, para simplificar la automatización del intercambio de información financiera mediante el uso del lenguaje XML. La familia de lenguajes XBRL se ha realizado para satisfacer las exigencias principalmente de información financiera y empresarial, permite aplicar etiquetas identificativas multilenguaje con significado, por ejemplo indicando si es un valor monetario, también es posible mostrar la relación que guardan los elementos entre sí, así se podría saber cómo se calculan. Una característica muy importante es la capacidad de extensión, de esta forma es capaz de adaptarse para casos particulares de empresas.

3.4.1 Actividades de aplicación de Semántica en *e-Procurement*

Desde la creación del movimiento de la Web Semántica, en concreto de la realización práctica mediante *Linked Data*, se han aplicado los principios de estas iniciativas a múltiples dominios como ya se ha reseñado en las secciones anteriores. Evidentemente en un sector como la Administración Pública, caracterizado por su amplitud, casuística y carácter estratégico para todos los ciudadanos es por lo que la semántica y sobre todo las corrientes de *Open Data* y *Linked Data* han penetrado con mucha fuerza en los últimos años. En un principio los esfuerzos como en otros dominios, se centraban en el modelado de la administración como organización y de sus procesos, con el objetivo de mejorar la interoperabilidad e integración entre las aplicaciones y facilitar los trámites burocráticos. En muchos casos se han desplegado soluciones flexibles basadas en la reutilización de información, procesos y conocimiento a través de grandes bases de datos compartidas, servicios web y sistemas basados en reglas, no obstante la inmersión de la semántica propiamente dicha se centraba en la realización de modelos para formalizar cuestiones relativas a documentación y a procesos administrativos. Por ello, la irrupción de *Open Data* y *Linked Data* ha reorientado el esfuerzo de las Administraciones Públicas para aprovechar la semántica en su beneficio.

Por otra parte, también han surgido iniciativas [4] relacionadas con *e-Procurement*, pero desde un punto de vista de las cadenas de suministro, en las cuales se modelan de forma completa un entorno de proveedores y suministros mediante la coordinación de los recursos software, hardware y humanos propios de un entorno de este tipo, como son ERPs, robots o las propias personas.

3.4.1.1 Ontología de Contratos Públicos del proyecto LOTED

El diseño de la ontología en el proyecto LOTED [257], es de suma importancia debido a su consideración como primer gran esfuerzo por aunar las iniciativas de *Linked Data* y *Open Data* en el campo de la contratación pública electrónica. En este proyecto la propuesta principal consiste en la transformación directa de la información proveniente del RSS de TED a RDF para su posterior indexado en un repositorio y acceso mediante un *endpoint* de SPARQL. Como sucede un muchos casos relacionados con la iniciativa de *Linked Data* la ontología realizada, ver Figura 3.31, está orientada a la representación de los datos en sí, estableciendo un modelo formal en el cual enclavar los recursos RDF generados, pero si bien este enfoque es válido, tan sólo representa una parte de la casuística presente en el dominio de la contratación pública electrónica. El intento realizado en TED se centra en el diseño de las entidades presentes en la información del RSS para así suministrar un modelo formal a los recursos RDF. Por todo ello, esta ontología sirve como una gran fuente de información para comprobar que tipo de entidades se recogen o publican en el RSS de TED, pero en un espectro más amplio se considera insuficiente para la representación de información teniendo en cuenta los modelos de datos disponibles en las plataformas de contratación o el propio de opXML realizado en el proyecto "10ders Information Services", como ha quedado señalado en la Sección 2.9.2.

Por otra lado y desde el punto de vista del acceso a los datos, se ofrece la posibilidad de la realización de consultas en SPARQL seleccionando una serie de códigos y acotando los resultados por fechas, sin embargo este sistema parece desactualizado y no se han realizado cambios en los últimos dos años.

3.4.1.2 Ontología de Contratos Públicos de la República Checa

Esta ontología está siendo desarrollada por el grupo *Knowledge Engineering Group* de la *Charles University* de Praga en la República Checa. Parte del esfuerzo está siendo cubierto parcialmente dentro del proyecto europeo LOD2 en su paquete de trabajo *WP9A – LOD2 for a Distributed Marketplace for Public Sector Contracts*, cuya descripción es la siguiente:

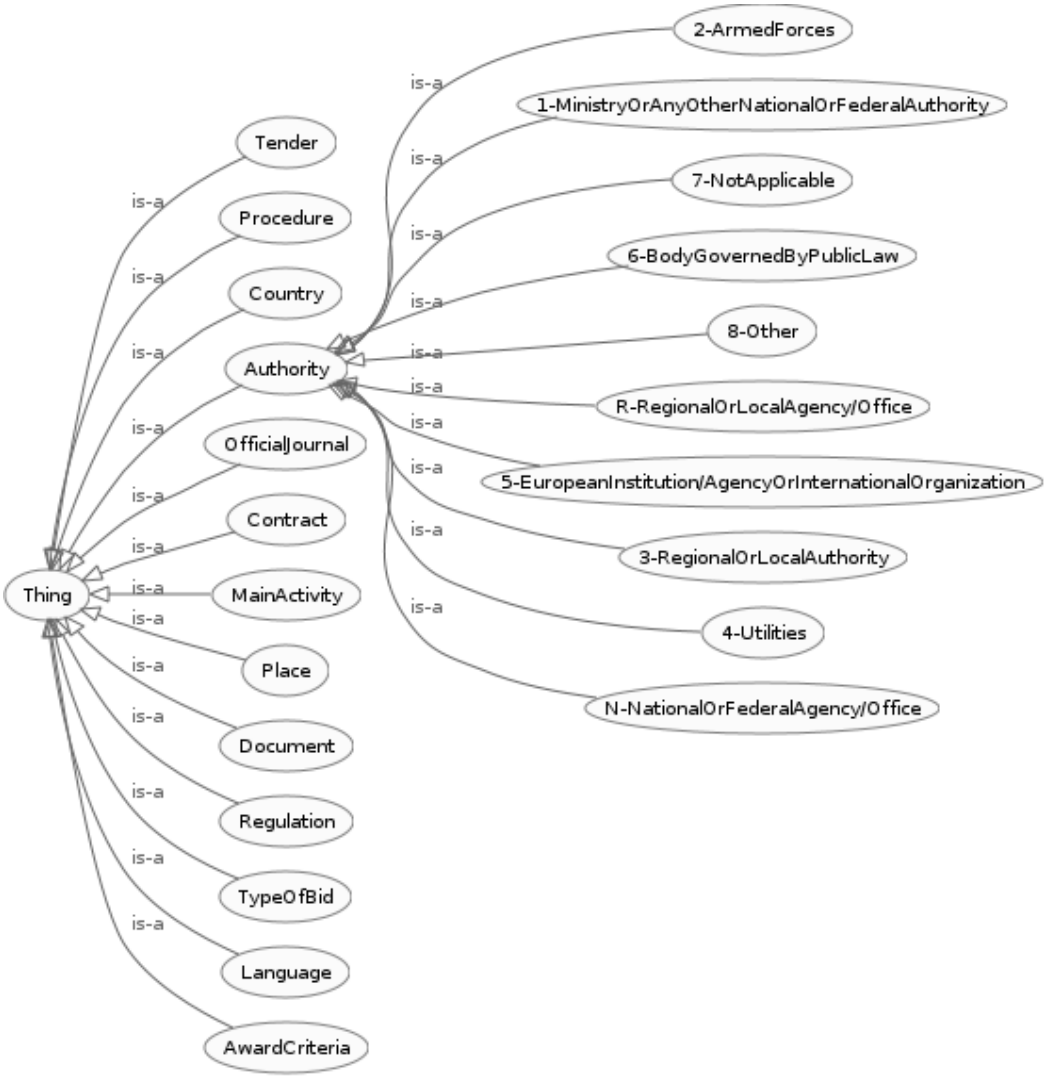


Figura 3.31: Ontología de Contratos Públicos del proyecto LOTED.

> *The objective of this use case is to explore and demonstrate the application of linked data principles for procuring contracts in the public sector...*

Este trabajo enlaza perfectamente con el propósito de este documento, esto es cubrir el sector de los contratos públicos con semántica y concretamente con la iniciativa *Linked Data*. Es por ello que se ha establecido contacto con los integrantes del grupo de investigación de esta Universidad para aprovechar y realimentar esfuerzos, como fruto de esta colaboración han empezado a reutilizar los códigos CPV, resultado de este trabajo y del proyecto "10ders Information Services".

La ontología que se ha desarrollado en este grupo, ver Figura 3.32, tiene como intención recoger la información y datos de los contratos públicos de forma estructurada, para que pueda ser consumida automáticamente tanto por personas como por máquinas, desarrollándose también dentro del ámbito de la iniciativa de *Open Data* de la República Checa. Desde un punto de vista del diseño reutiliza varios vocabularios y ontologías ya disponibles, como *Payments Ontology* del Reino Unido, lo que confiere a este modelo un carácter integrador y reutilizable. No obstante, abordar la descripción de toda la casuística del proceso de contratación pública electrónica parece muy ambicioso y podría presentar problemas de interoperabilidad, integración y reutilización ya que puede estar muy orientada a la problemática de un entorno particular. Otro de los puntos importantes que se deben abordar, y parcialmente recogidos en esta ontología y que deben servir de guía, están referidos a la adición de

metainformación para *provenance*, licencia, etc., que si bien en algunos *datasets* es importante, en la información de carácter público es fundamental y debe ser un requisito para las propias entidades públicas.

Figura 3.32: *Public Contracts Ontology from Czech Republic.*

3.4.1.3 Clasificaciones Estándar de Productos

Las clasificaciones de productos, son instrumentos claves de estandarización [203] que nacen con el fin de conseguir una clasificación común de productos y servicios. En general, son variadas [271] y obedecen a intereses particulares dependiendo del sector como e@Class o RossetaNET, o bien de carácter global como UNSPSC u otras destinadas a un dominio particular como el CPV en la Administración Pública, ver Sección 2.10. Las diferencias estriban en su alcance y cobertura sectorial, pero también en el grado de especificidad o nivel de profundidad que existe para describir los productos o servicios.

Como señala Hepp [159–161] todos estos estándares reflejan una combinación de componentes variables, que pueden ser utilizados para la construcción de una ontología derivada a partir de la

clasificación. Sin embargo, se puede identificar una estructura común subyacente a todas ellas y que es fundamental señalar para proporcionar un modelo de datos semántico universal para este tipo de clasificaciones, consistente en que todas estas clasificaciones se ordenan jerárquicamente.

Categorías de productos. Las clasificaciones se dividen en categorías o clases de productos, agrupando los distintos elementos del catálogo.

- Las categorías de la clasificación se organizan jerárquicamente.
- Cada elemento de la clasificación pertenece a una categoría de productos.
- Cada elemento de la clasificación pertenece sólo una categoría de productos, es decir, las categorías son disjuntas.

Estructura taxonómica. Además de la división en niveles de jerarquía de los elementos de la clasificación, su objetivo es organizar y agrupar los productos en sectores verticales mediante algún tipo de criterio establecido por la comunidad que desarrolla el estándar.

Estas son características genéricas de las clasificaciones de productos. Sin embargo, otras clasificaciones más sofisticadas incluyen un diccionario de propiedades estándar que se puede utilizar para describir productos con más detalle. Normalmente, estos diccionarios de propiedades también incluyen los tipos de datos que pueden ser valor de las mismas, así como su referencia con respecto a estándares internacionales para establecer las unidades de medida, este es el caso de la clasificación de productos de e@Class. En otras ocasiones, se construyen clasificaciones multilingües para la expresión de los descriptores de cada elemento.

La irrupción de la tecnología semántica y, sobre todo, la aparición de lenguajes web para la representación de conocimiento y gestión de metadatos, ha propiciado un creciente interés en el uso de las clasificaciones estándar de productos para mejorar tanto el intercambio de información, como su capacidad para estructurar información. La construcción de ontologías de productos con alto nivel de detalle implican un coste que es muy difícil de asumir en muchos casos. Los autores [114,241,304] comparten la opinión de Corcho [78] de que una ontología de productos sería muy útil para la organización conceptual del mercado, se entiende que estas ontologías tienen más carácter privado, para la organización de la producción, los departamentos de ventas y comerciales y, en general, para cualquier área de una empresa o institución que deba tratar con la gestión de productos.

El desafío básico más importante que hay que afrontar cuando se deriva una ontología de una clasificación de productos, se refiere a cómo interpretar la semántica original de la taxonomía. No existe una definición formal de las relaciones taxonómicas que construyen cada categoría de la clasificación y es tentador utilizar la propiedad de un vocabulario de ontologías, como *rdfs:subClassOf*, para intentar representar estas relaciones semánticas, sin embargo, esta suposición es errónea. Como señala [160], esta relación de jerarquía entre los elementos no se puede considerar equivalente a una relación de subclase o de herencia. En primer lugar, tomando el siguiente ejemplo: el elemento "Partes y accesorios de bicicleta" del CPV 2008 (34432000-4) tiene como antecesor a "Bicicletas" (34440000-0), donde la relación semántica entre los dos elementos no es herencia, es decir, no se puede expresar que una **parte de una bicicleta** sea una **bicicleta**. Para ser más precisos, debería modelarse como una relación de composición o agregación. En este sentido, ocurre igual con la relación taxonómica entre "Tubos de resina Epoxi"(19522110-5) y "Resina Epoxi" (19522100-2), en el que es difícil justificar de nuevo una relación de herencia clásica entre el primero y el segundo elemento del CPV, ya que en ningún caso se puede considerar que el continente y el contenido de un objeto complejo tenga el mismo estatus en una ontología de dominio, es decir, que un **tubo** no es un tipo de **resina**.

Pero no sólo es complicado interpretar correctamente las relaciones semánticas que codifica la taxonomía de una clasificación de productos, desde el punto de vista de las ontologías como modelos de conocimiento de dominio, muchos elementos de una clasificación de productos son difícilmente

interpretables como conceptos de dominio, en este sentido, un elemento como "Barras, varillas, perfiles y alambre de estaño" (27623100-9) del CPV 2003, que parece más una colección artificial de productos que una clase estructural, en la que sus instancias comparten algún tipo de propiedad común, estrictamente para interpretar correctamente el elemento "Barras, varillas, perfiles y alambre de estaño" como una clase, debería definirse como la unión de varias clases: por ejemplo, "Barras", Varillas" o "Alambre de estaño".

La problemática que presenta el modelado semántico de las clasificaciones de productos conlleva dificultades intrínsecas que no se encuentran en otros dominios. En este sentido han surgido vocabularios como *GoodRelations* y *ProductOntology*, que facilitan estas tareas de modelado y reutilización de descripciones de productos. *GoodRelations* es un vocabulario estándar (esquema, diccionario u ontología) para productos y datos empresariales que pueden ser introducidos en páginas web, ver Figura 3.33 de las tripletas extraídas utilizando el servicio *Any23*, tanto estáticas como dinámicas, para permitir de esta forma el procesamiento automático por las máquinas. La principal ganancia reside en el aumento de la visibilidad, por motores de búsqueda etc., de los productos y servicios etiquetados de esta manera, actualmente algunas de las empresas que utilizan este vocabulario son: Google, Yahoo, Best Buy, O'Reilly, Volkswagen UK, Renault UK, etc., en general para etiquetar sus productos y para que la información pueda ser procesada automáticamente.

Por otra parte, *ProductOntology* se utiliza para enlazar cualquier producto con una descripción que está disponible en la Wikipedia, de esta manera las instancias de la ontología obtienen una definición dinámica que se puede reutilizar en cualquier contexto. Finalmente, cabe destacar la iniciativa *schema.org* desarrollada por los grandes proveedores de servicios de búsqueda, cuyo objetivo es encajar descripciones en las propias páginas web para que el descubrimiento de la información por parte de los *crawlers* sea más sencillo.

En cuanto a los escenarios o casos de uso en los que las catálogos de clasificaciones de productos han sido utilizados son variados, pero podrían destacarse los siguientes: en servicios web semánticos para el proceso de descubrimiento, para el intercambio y actualización automática de catálogos de productos entre distintas aplicaciones, para el etiquetado de recursos mediante vocabularios controlados, etc.

Como conclusión, se puede observar que las clasificaciones de productos y servicios son sumamente interesantes para la mejora de la interoperabilidad e integración de las aplicaciones, que ha sido ampliamente impulsada por la irrupción de la corriente de la Web Semántica y los datos enlazados.

3.4.1.4 Información sobre Organizaciones

En el caso de las organizaciones la tarea de búsqueda de propuestas relacionadas con semántica se complica debido a que existen numerosas ontologías que tienen definido el concepto de "Organización". Por ejemplo FOAF hace uso de esta entidad para referirse a la compañía a la que pertenece una persona, *Inference Web* que está estrechamente relacionada con procesos de inferencia en la web, en el ámbito de *trust* y *provenance* [216], utiliza las organizaciones para establecer la relación de confianza que existe entre las diferentes entidades. En todos los casos la representación y cobertura del concepto "Organización" es mínimo y adaptado para el caso particular.

Según la evaluación realizada por Dave Reynolds (Epimorphics Ltd) en un informe web [87], convertido en los últimos tiempos en borrador [263] del W3C, se han desarrollado muchos enfoques los cuales estaban dirigidos por distintos objetivos, algunos centrados en la noción de organización como ente superior, así aparece en *upper-ontologies* como Proton, Sumo o SmartWeb. Estos modelos están construidos con múltiples objetivos, por lo que en principio no se adaptan fácilmente a la estructura de modelos pequeños y reutilizables para la descripción de las organizaciones, aunque evidentemente deben ser revisados. Por otra parte, si se utilizan los motores de búsqueda para obte-

```
<http://www.renault.co.uk/ownerservices/shop/item/renaulttoys/pedalcar/eco2pedalcar/
    default.aspx>
  dcterms:title "ECO2 Pedal Car - Renault Shop - Owner Services - Renault UK" .


<http://www.renault.co.uk/ownerservices/shop/item/renaulttoys/pedalcar/eco2pedalcar/
    default.aspx#offering> a
  <http://purl.org/goodrelations/v1#Offering> .

<http://www.renault.co.uk/ownerservices/shop/item/renaulttoys/pedalcar/eco2pedalcar/
    default.aspx#offering>
  gr:eligibleRegions  "GB"^^<http://www.w3.org/2001/XMLSchema#string> .


<http://www.renault.co.uk/ownerservices/shop/item/renaulttoys/pedalcar/eco2pedalcar/
    default.aspx#offering>
        foaf:page <http://www.renault.co.uk/ownerservices/shop/item/RenaultToys/
            PedalCar/ECO2PedalCar/default.aspx> ;
        gr:availableDeliveryMethods <http://www.renault.co.uk/ownerservices/shop/
            deliverydetails.aspx#delivery> ;
        gr:hasPriceSpecification <http://www.renault.co.uk/ownerservices/shop/
            deliverydetails.aspx#deliverycharges> ;
        gr:name "ECO2 Pedal Car" ;
        gr:hasPriceSpecification _:node16kpidu7qx455 .

_:node16kpidu7qx455 a gr:UnitPriceSpecification ;
        gr:hasCurrency "GBP"^^<http://www.w3.org/2001/XMLSchema#string> ;
        gr:hasCurrencyValue "260"^^<http://www.w3.org/2001/XMLSchema#float> ;
        gr:valueAddedTaxIncluded "true"^^<http://www.w3.org/2001/XMLSchema#boolean> ;
        gr:validThrough "2012-02-03T17:22:43Z"^^<http://www.w3.org/2001/XMLSchema#
            datetime> ;
        gr:hasUnitOfMeasurement "C62"^^<http://www.w3.org/2001/XMLSchema#string> .

<http://www.renault.co.uk/ownerservices/shop/item/renaulttoys/pedalcar/eco2pedalcar/
    default.aspx#offering>
      gr:hasBusinessFunction gr:Sell ;
        gr:hasInventoryLevel _:node16kpidu7qx456 .

_:node16kpidu7qx456 a gr:QuantitativeValue ;
        gr:hasMinValue "1"^^<http://www.w3.org/2001/XMLSchema#float> .

<http://www.renault.co.uk/ownerservices/shop/item/renaulttoys/pedalcar/eco2pedalcar/
    default.aspx#product> a gr:SomeItems ;
        gr:category "Pedal Car" ;
        gr:name "ECO2 Pedal Car" ;
        gr:description "Dimensions: 114 x 69 x 62cm. Weight: 10kg. Age 3 to 7 years."
            ;
        foaf:page <http://www.renault.co.uk/ownerservices/shop/item/RenaultToys/
            PedalCar/ECO2PedalCar/default.aspx> .
```

Figura 3.33: Ejemplo de tripletas de RDF en N3 extraídas de un producto de Renault utilizando *GoodRelations*.

ner los trabajos relacionados con las organizaciones, Swoogle obtiene alrededor de 3,900 resultados, Falcons 15, 881 del concepto "Organization" estando presente en 15 vocabularios diferentes y Google ofrece: 1) la "Organization Ontology 1.0" desarrollada en SHOE la cual proporciona una idea de la jerarquía básica de una organización, industrias y roles posibles de los empleados; 2) una "Organization Ontology for Enterprise Modelling", más orientada a entornos de cadenas de suministros y 3) "Enterprise Ontology", que es una ontología para representar la actividad de negocio de las empresas en Ontolingua. También *Jeni Tennison*, a través de su blog, ha apuntado a la ontología desarrollada por TSO para "London Gazette RDFa markup", en la cual se incluyen los conceptos de "Gazzete Organization" y "Gazette Person".

En general, se pueden sintetizar que los enfoques llevados a cabo en AKT Portal Ontology, Proton, *GoodRelations*, FOAF, SIOC, *Enterprise Modelling Ontology*, *Enterprise Ontology*, *Gazzete*, *Provenance Vocabulary Ontology* y otros orientados al ambiente académico como ECS (Universidad de Southampton) o "Academic Institution Internal Structure Ontology" (AIISO), deben ser tenidos en cuenta para obtener una enseñanza común y un vocabulario que pueda describir de forma genérica y extensible la casuística que surge para modelar la información empresarial. En conclusión, existe un amplio conjunto de vocabularios destinados a describir entidades similares pero con distintos objetivos y que se han reutilizado en la ontología de Organizaciones, ver Figura 3.34, desarrollada por Dave Reynolds. Este trabajo es un excelente punto de partida para el modelado de organizaciones en la *Web of Data* ya que ha condensado el conocimiento previo.

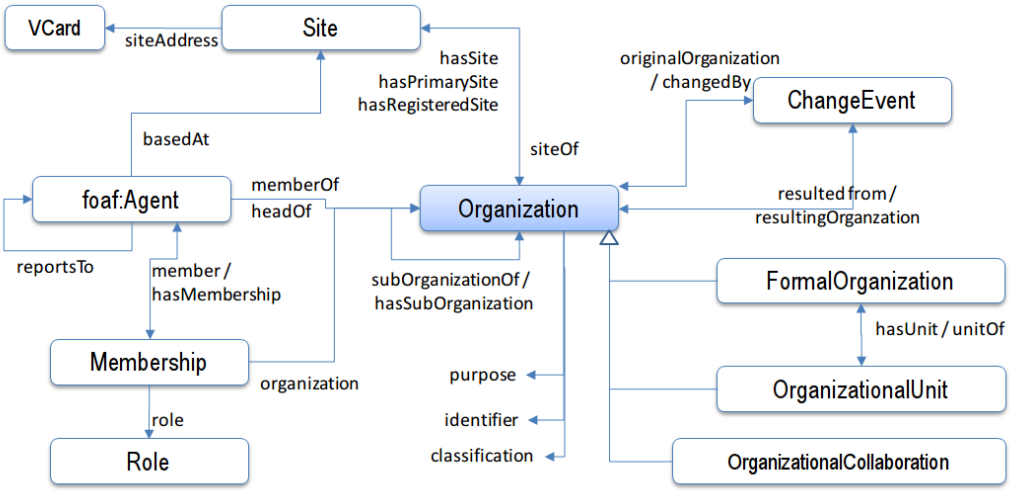


Figura 3.34: *Organizations Ontology. Overview.*

Desde otro punto de vista no tan centrado en la corriente de Web Semántica y ontologías, hay que referenciar la propuesta de datos abiertos realizada por *OpenCorporates* a través "*The Open Database Of The Corporate World*", han utilizado técnicas de *screen scrapping* y *crawling* para extraer información de más de 12 millones de compañías. Esta información, ver Figura 3.35, tiene un alto valor para reutilizar y trazar la actividad de una empresa ya que disponen en la mayoría de los casos del identificador de la empresa. La información de esta base de datos sigue un enfoque mixto entre *Open Data* y *Linked Data*, pero el gran problema reside en la ausencia de un modelo formal para la descripción de los datos, es por ello que disponer de un modelo formal que integre esta información es clave para la posible reutilización de la información y explotación de la misma de una forma estándar.

3.4.1.5 Proyectos de Investigación

Los proyectos de investigación de los principales programas competitivos también se han visto involucrados en el despliegue de sistemas de contratación pública electrónica utilizando semántica

```
<http://opencorporates.com/companies/nl/37136346.rdf?id=5828504>
  a <http://purl.org/dc/dcmitype/Text>,
            foaf:Document;
    dct:format "application/rdf+xml";
    dct:isFormatOf
     <http://opencorporates.com/companies/nl/37136346?id=5828504>;
    dct:title "Linked Data in RDF format for Benuma";
    foaf:primaryTopic
     <http://opencorporates.com/id/companies/nl/37136346> .
     ...
   <http://opencorporates.com/id/companies/nl/37136346>
   a <http://s.opencalais.com/1/type/er/Company>;
    :label "Benuma"
```

Figura 3.35: Información (parcial) sobre una Organización de "Open Corporates" en N3.

y datos enlazados. Entre ellos se puede destacar los siguientes, relacionados por su impacto, nivel europeo y por las entidades e investigadores participantes.

***LOTED Project* [257]** *Linked Open Tenders Electronic Daily* realizado por el grupo KMI de la *Open University* del Reino Unido permite la consulta en SPARQL de anuncios de licitación publicados a través de los RSS de TED. Se trata del tradicional enfoque de *Rdfizar*, información ya publicada y hacerla disponible a través de un *endpoint* de SPARQL. Sin duda se trata de una importante iniciativa tanto por su carácter innovador, como por tratarse de la primera apuesta real de utilización de semántica en los anuncios de licitación, sin embargo, tan sólo llega a los anuncios de licitación publicados en TED y también parece que el demostrador oficial ha dejado de ser mantenido por los autores. No obstante, es necesario considerar esta propuesta para aprovechar el efecto experiencia de la misma.

***LOD2 Project* [126].** Este proyecto europeo al que se ha referenciado en la Sección 3.2.6.7 por su esfuerzo en la iniciativa *Linked Data* desde un punto de vista genérico, también ha seleccionado la contratación pública electrónica como un caso de uso estratégico, es por ello que han aumentado los paquetes de trabajo para incluir el esfuerzo de la *Charles University* de la República Checa y su investigación sobre la aplicación de *Linked Data* y semántica en el campo del *e-Procurement*. Esto se ha manifestado en la constitución del paquete de trabajo *WP9A – LOD2 for a Distributed Marketplace for Public Sector Contracts*. La importancia del seguimiento de este proyecto reside tanto en las personas como en las instituciones implicadas, ya que generan una gran cantidad de tecnología y *know-how* especialmente relevante para el estudio objeto de este documento y para las iniciativas de *Linked Data* y *Open Data* en general.

***LATC Project* [255].** *Linked open data around-the-clock* es un proyecto europeo, una *Specific Support Action* en el contexto del 7º Programa Marco-ICT formado por más de 58 instituciones en los que se realizan y coordinan proyectos, personas. etc. El objetivo de este proyecto es gestionar la información y datos generados a través de distintas fuentes proveyendo la infraestructura y documentación necesaria para desplegar arquitecturas que den soporte a los datos enlazados. Prueba de su ingente capacidad es la cobertura y asesoramiento a proyectos y a investigadores de Europa (74 %) y de Estados Unidos (25 %). El consorcio está formado por instituciones y empresas tan relevantes en este campo como: *Digital Enterprise Research Institute* (DERI), *NUI Galway*, Irlanda; *Vrije Universiteit Amsterdam* (VUA), Países Bajos; *Freie Universität Berlin* (FUB), Alemania; *Institute for Applied Informatics* e.V. (InfAI), Alemania y *Talis Information Ltd.*, Reino Unido.

***PlanetData Project* [127].** Se trata de una Red de Excelencia en el ámbito europeo (FP7-257641) con un presupuesto total de 3,72 millones de euros y en la que participan los principales organismos de investigación con el objetivo de sumar los esfuerzos de la comunidad de investigadores para ofrecer a las organizaciones interesadas en la iniciativa de *Linked Data* soporte con la publicación de sus datos. Evidentemente cuentan con un estrecho lazo con los proyectos anteriores y su actividad cubre las cuestiones relacionadas con datos enlazados en diferentes ámbitos: *data streams, (micro) blog posts, digital archives, eScience resources, public sector data sets, and the Linked Open Data Cloud*.

***WebDataCommons.org* [41].** Es una iniciativa conjunta realizada por el grupo de investigación *Web-based Systems Group* en la *Freie Universität* de Berlín y el *Institute AIFB* perteneciente al *Karlsruhe Institute of Technology*, en el cual se han extraído y generado tripletas RDF de más de 65 millones de sitios web, contando un número de 3,2 *billion RDF quads* a fecha de febrero del año 2012, cubriendo información sobre productos, personas, organizaciones, lugares, eventos y recetas de cocina entre otros. Para la realización de este experimento se han utilizando técnicas de extracción de información de HTML que utilizan microformatos, RDFa, etc. Toda esta información y datos está disponible para su descarga y uso por terceros en RDF. La importancia de este proyecto reside tanto en las personas y organizaciones implicadas como en la magnitud de los datos que han conseguido extraer.

Otros proyectos. Estas iniciativas trabajan con tecnología semántica y de datos enlazados en diferentes contextos suministrando servicios a instituciones públicas o bien realizando investigación e innovación. Entre otros, se pueden citar: FP7 project SEALS (*Semantic Evaluation at Large Scale*), FP7 project SpitFIRE (*Semantic-Service Provisioning for the Internet of Things using Future Internet Research by Experimentation*), French DataLift project y Semic.EU.

Igualmente a nivel nacional se han impulsado diferentes iniciativas y proyectos, como los de la República Checa y "10ders Information Services", que intentan elevar el significado de la información de los anuncios de licitación utilizando tecnologías semánticas. También hay que considerar que la reutilización de vocabularios ya establecidos e impulsados por las distintas Administración Públicas, por ejemplo la iniciativa Joinup [191] de la Unión Europea, es de suma relevancia para proporcionar información utilizando datos enlazados, en este sentido todas las iniciativas y proyectos reseñados en las secciones anteriores ayudan a disponer de los *building blocks* necesarios para afrontar la aplicación de semántica a los procesos de contratación pública electrónica.

Capítulo 4

Definición de Métodos Semánticos

> No hemos sido los primeros,
> pero seremos los mejores.
>
> *Citas Célebres*
> STEVE JOBS

En el capítulo anterior se ha realizado una introducción y un repaso a los conceptos de Web Semántica, *Open Data*, *Linked Data* y *Linking Open Data*, en los cuales se ha podido comprobar la diversidad de iniciativas, enfoques y estrategias que se están desarrollando desde distintos ámbitos con el objetivo de hacer realidad la Web de Datos. En el contexto de las licitaciones públicas la aplicación de estos principios resulta clave para impulsar la competitividad y la transparencia en los procesos administrativos que forman parte de la contratación pública, no obstante una de las carencias que se puede vislumbrar es la falta de estructuración de los procesos, métodos y tareas a definir y desarrollar, para llevar a cabo una correcta realización práctica de estos conceptos cumpliendo de esta forma con los principios que se propugnan. Es por ello que a lo largo de este capítulo se realiza una clasificación de los procesos, métodos y tareas a efectuar, así como una definición teórica de los mismos. De esta manera, una vez que se contextualicen los métodos definidos en el ámbito de la contratación pública, se podrá verificar su eficiencia respecto a otros enfoques, en algunos casos, demasiado abiertos, que se alejan de la resolución del proceso encaminado a la apertura de datos enlazados complicando la iniciativa, provocando una sensación de inquietud y duda, tanto por parte de los encargados de liberar los datos como por los que deben consumirlos.

Una primera clasificación de estos métodos, realizada recientemente [151], los enclava en dos grandes grupos:

1. Publicación. Orientados a seleccionar los datos a liberar y a llevar a cabo todas las actividades relacionadas con los mismos: modelado, diseño de URIs, formatos de salida, etc.

2. Consumo. Orientados a establecer los mecanismos de reutilización de datos enlazados en distintas aplicaciones o bien en otros *datasets*. Entre las actividades que lleva aparejadas este gran grupo estarían: la consulta, enriquecimiento, calidad, establecimiento de cachés, etc.

Sin embargo, estos dos grandes grupos pueden ser divididos a su vez, en distintos tipos de métodos, que si bien pueden encajar en distintos momentos de la generación de *Linked Data*, la definición de los mismos debe ser única en cada caso. Existen métodos y, en consecuencia, actividades que se han de reutilizar en distintos procesos con diferente denominación, de esta manera los métodos relacionados con la producción de datos enlazados estarán en estrecha relación con aquellos que se encarguen de actualizar o consolidar los datos, por ejemplo un método que lleve a cabo la tarea de

descubrimiento automático de *datasets* será de gran utilidad tanto en producción como en consumo o actualización. Por todo ello, además de proveer una clasificación de métodos semánticos, ver Figuras 4.1 y 4.2, y distintas actividades relacionadas, se deben alinear con los ciclos de vida que se repasan en la Sección 3.2.6.5.

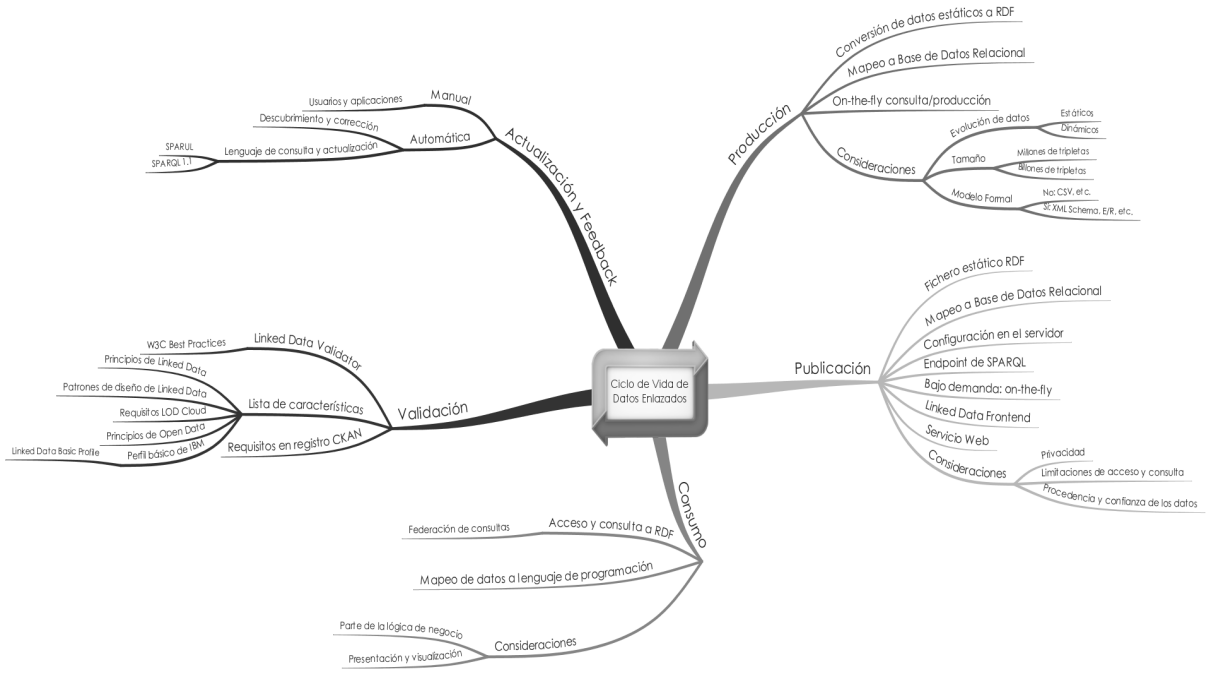


Figura 4.1: *Clasificación de Métodos Semánticos.*

Para llevar a cabo una nueva definición se necesita precisar qué procesos se deben contemplar en la apertura y enlazado de datos, una vez definidos se identifican los métodos que se pueden aplicar y dentro de los mismos las tareas a ejecutar. Por ejemplo, en el proceso de "Producción" se deberá identificar los datos a transformar, seleccionar un esquema de URIs y validar la transformación. Por lo tanto, se cuenta con una descripción de tres niveles: proceso, método semántico y tarea, que intenta dar respuesta a las siguientes preguntas, ver Tabla 4.1. La selección de estos procesos, métodos y tareas se ha llevado a cabo tratando de factorizar las actividades presentes en los distintos ciclos de vida de los datos enlazados.

Proceso	**Pregunta**
Producción	¿Qué datos y cómo los transformo?
Publicación	¿Cómo publico los datos transformados?
Consumo	¿Cómo reutilizo los datos enlazados disponibles?
Realimentación	¿Cómo actualizo y mejoro mis datos enlazados?
Validación	¿Cómo verifico que los datos enlazados en los distintos procesos son correctos?

Tabla 4.1: Procesos y Preguntas en *Linked Data*.

El proceso a seguir para desarrollar este capítulo es el siguiente:

1. Definir los métodos semánticos de forma general en el marco de un proceso.
2. Explicar los distintos enfoques que dan respuesta a ese método.

3. Comentar las tareas que implica llevar a cabo el enfoque anterior.
4. Ejemplificar el método con un ejemplo transversal.

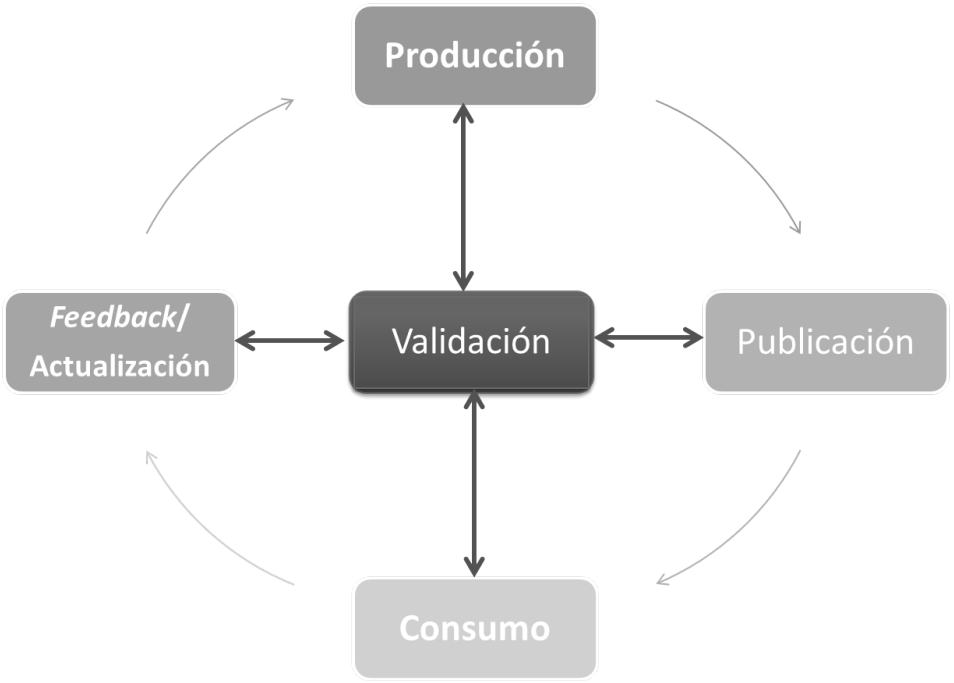


Figura 4.2: Procesos en *Linked Data*.

4.0.2 Ejemplo transversal

Para ejemplificar cada una de las definiciones realizadas a lo largo de este capítulo se utilizará un ejemplo de un conjunto de datos real. Se ha seleccionado la información disponible en el Nomenclátor de Asturias 2010 realizado por la Sociedad Asturiana de Estudios Económicos e Industriales, ya que contiene información de distintos ámbitos (nombres, estadísticas, etc.) e idiomas y facilita la comprensión y lectura del documento.

4.1 Definiciones Previas

Durante este capítulo se utilizarán algunas de las definiciones ya realizadas en las distintas especificaciones [52, 156, 164, 192, 195, 260] sobre el uso de Web Semántica y algunos de los conceptos clave pertenecientes a la iniciativa de *Linked Data*.

Tupla. Siguiendo las definiciones utilizadas en el modelo relacional y en matemáticas, una tupla es una función que *mapea* nombres con valores. En general, se trata de un conjunto de valores $a_1, a_2 ... a_n$ que guardan una relación de orden entre sí, pueden contener valores repetidos y otros objetos dentro de sí.

Dataset. Se puede definir como un conjunto de tuplas. Si se tiene en cuenta que en la mayoría de los casos al producir *Linked Data* los datos se obtienen de una base de datos relacional, XML, texto

separado por comas, etc., cada uno de los almacenes especifican una forma de guardar tuplas para generar un *dataset*.

***Internationalized Resource Identifiers* (IRIs).** Habitualmente se utiliza la denominación de IRI (RFC-3987) para referirse a una generalización de las URIs y URLs (RFC-3986), siendo totalmente compatibles con las definiciones de las anteriores. Normalmente se presentan de la siguiente forma $< IRI >$ y pueden ser relativas respecto a un IRI base o completas.

En las definiciones formales de RDF y SPARQL se utiliza el término IRI pero por su definición puede ser utilizado también el término URI. En el ámbito de *Linked Data* se puede afinar aún más y utilizar la siguiente nomenclatura: $IRI \longrightarrow URI \longrightarrow HTTP\ URI$.

***Uniform Resource Identifier*.** El concepto de URI surge como una cadena de caracteres para identificar de manera única a un elemento, en el contexto de la Web Semántica se ha utilizado este término para identificar recursos. Dentro de la definición de URI encaja *Uniform Resource Locator* (URL), que además de proveer información de nombrado sobre el elemento especifica cómo se ha de acceder a él. En el documento [195] se hace referencia al término HTTP URIs para identificar a aquellas URIs que utilizan el esquema de nombrado HTTP para identificar a los recursos dentro de la iniciativa de la Web Semántica.

Modelo RDF. Aunque ya se ha comentado en la Sección 3.1 el modelo RDF, cabe recordar su definición, RDF sigue y cumple con los principios y características de interoperabilidad, extensibilidad, capacidad de evolución y descentralización elaborados por el W3C. En particular, el modelo RDF se diseñó para tener un modelo de datos simple con una semántica formal y capacidades de inferencia basado en un vocabulario que hace uso de URIs para nombrar a los elementos. En el universo RDF, los elementos a modelar son un conjunto de recursos, en esencia todo aquello que sea susceptible de tener un URI. El lenguaje utilizado para describir estos recursos se compone de un conjunto de predicados (binarios). Las descripciones de recursos en RDF se basan en la estructura $< s, p, o >$ (sujeto, predicado, objeto), en los cuales los predicados y los objetos son también recursos o literales (en el caso de los objetos). Por otra parte, sujetos y objetos pueden ser anónimos, así tienen URI pero sólo de forma local, formando lo que se denominan *blank nodes* cuyo uso se ha evaluado y juzgado [210] recientemente.

En conclusión el modelo RDF consta de unos principios para construir un vocabulario bajo una determinada semántica, en muchos casos RDFS [52] u OWL [164], permitiendo así la herencia de clases y propiedades, definición de tipos y otras características. Este modelo es especificado en una serie de documentos que cubren su semántica, sintaxis, serilización, etc.

Tripleta RDF. Es una tupla $(s, p, o) \in (I \cup B) \times I \times (I \cup L \cup B)$, en la que I es el universo de todos los posibles recursos que se pueden identificar por un IRI, B es un conjunto infinito de recursos no nombrados y L es el conjunto de todos los literales RDF.

Grafo RDF. Un conjunto de tripletas RDF se interpreta estructuralmente como un grafo RDF $\mathcal{G} = (V, E)$, donde V es el conjunto de vértices, $V \subseteq (I \cup B \cup L)$, y E es el conjunto de ejes, $E \subseteq V \times I \times V$. Para cada tripleta (s, p, o) del modelo RDF existe una arista dirigida, etiquetada con el predicado p, entre los vértices que representan el sujeto s y el objeto o. Si este grafo no tiene nodos en blanco se denomina *ground*.

Recurso RDF. Es un conjunto de sentencias o tripletas (s, p, o) RDF, en las que el sujeto s es constante, tiene la misma IRI.

Grafo RDF nombrado. Es un grafo RDF identificado por un IRI.

***Dataset* RDF.** En la especificación de SPARQL [259] y su semántica [260] se habla de *dataset* RDF y se define como un conjunto $\mathcal{D} = \{\mathcal{G}, (< I_1 >, \mathcal{G}_1), (< I_2 >, \mathcal{G}_2) ... (< I_n >, \mathcal{G}_n)\}$, donde $\mathcal{G}$ y cada $\mathcal{G}_i$ son grafos RDF identificados a través de un IRI I_i y $n \geq 0$. Cada par $(< I_n >, \mathcal{G}_n)$ es un grafo

nombrado, en el cual I_n es el identificador del grafo $\mathcal{G}_n$; $\mathcal{G}$ es el grafo por omisión (*default graph*) y contiene todas las tripletas.

Esta definición sobre un *Dataset* RDF se puede extender para definir un modelo formal que sirva para integrar información proveniente de distintas fuentes de datos dentro de un repositorio RDF u otro sistema de almacenamiento. Un *dataset* integra varios grafos RDF $\mathcal{G}$, de forma que cada una de las tripletas se puede identificar, gestionar y referenciar de forma separada. En el modelo de datos de RDF, utilizando grafos nombrados las sentencias forman parte del conjunto de datos o de un grafo nombrado $\mathcal{G}_n$ o bien del grafo por omisión $\mathcal{G}$.

Por lo tanto, un *dataset* RDF puede representar información de distintos grafos RDF, si se añade una nueva dimensión a la definición de tripleta RDF $(s, p, o, \mathcal{G})$ se obtienen "sentencias contextualizadas", en las que los tres primeros elementos (s, p, o) indican una sentencia RDF y el cuarto elemento $\mathcal{G}$ representa el grafo en el que están definidas. De esta forma un *Dataset* RDF consta de un conjunto infinito de cuádruplas $(s, p, o, \mathcal{G})$. Esta definición es ampliamente utilizada en los repositorios RDF para identificar los distintos recursos y así poder acceder mediante SPARQL a los recursos definidos en determinados grafos.

Ontología. Se define una ontología $\mathcal{O}$ como una tupla $\mathcal{C}, \mathcal{R}, \mathcal{I}, \mathcal{A}$, donde $\mathcal{C}$ es un conjunto de conceptos, $\mathcal{R}$ es un conjunto de relaciones, $\mathcal{I}$ es un conjunto de instancias y $\mathcal{A}$ es un conjunto de axiomas. Todos los conceptos, relaciones, instancias y axiomas se expresan a través de un lenguaje como OWL o F-Logic que permite expresar un determinado formalismo lógico. Esta visión de una ontología encaja con la definición realizada en OKBC, los conceptos se corresponderían con clases en OKBC, relaciones con "slots", "facets" con tipos de axiomas y los "individuals" corresponderían con instancias.

La necesidad de definir el concepto de ontología reside en que las descripciones de recurso en RDF deberán haber sido modeladas previamente de acuerdo a un modelo formal. Teniendo en cuenta que en el contexto objeto de estudio se utilizarán ontologías, un recurso RDF se presenta como una instancia de un concepto de una ontología $\mathcal{O}$ que especifique el conocimiento de este dominio.

***Mapeo* entre instancias.** Los *mapeos* entre conceptos e instancias de una ontología $\mathcal{O}$ y en consecuencia de recursos RDF han sido ampliamente estudiados [239] en el campo de los servicios web semánticos para dar respuesta a los procesos de mediación. En [90] se especifican tres operaciones: *mapping*, *alignment* y *merging*, como tareas necesarias para realizar el proceso de mediación entre ontologías. En el caso que nos ocupa es necesario realizar operaciones de *alignment* para identificar recursos RDF similares a uno dado y poder establecer relaciones de equivalencia, igualdad y *mapeo* en general, por ejemplo, utilizando propiedades como `owl:sameAs`, `skos:exactMatch`, etc. En el ámbito de *Linking Open Data* se engloban estas técnicas en un problema denominado "Reconciliación de Entidades" y además de los enfoques basados en alineación de instancias de ontologías, se han realizado otros basados en procesamiento de lenguaje natural de descripciones textuales [8] o basados en la comparación de URIs [207]. En cualquier caso, es conveniente definir que se entiende por *alignment* en Web Semántica.

Ontology alignment es el proceso de descubrir similaridades entre dos ontologías, el resultado de la operación es una especificación de similaridades entre las dos ontologías seleccionadas, generalmente este proceso se basa en la aplicación del algoritmo *Match operator*.

4.2 Definición Genérica de Método Semántico

En primer lugar, debe definirse qué se entiende por método semántico en general y concretamente en el contexto de este trabajo.

Definición 4.2.1 (Proceso p)**.** *Se define un proceso p, como la aplicación de uno o varios métodos semánticos.*

Definición 4.2.2 (Método Semántico sm)**.** *Se define un método semántico sm, como la consecución de n tareas para llevar a cabo una operación sobre un dataset.*

Este *dataset* puede ser un conjunto de valores y datos $\mathcal{G}$, o bien un *dataset* RDF $\mathcal{D}$, dependiendo del proceso concreto.

Definición 4.2.3 (Tarea t)**.** *Es cada uno de los pasos que se han de llevar a cabo para realizar un método semántico*

En definitiva, un proceso p dentro de la iniciativa de *Linked Data* se puede realizar a través de distintos métodos semánticos sm, que conlleva a su vez la ejecución de n tareas t, implementadas mediante diferentes herramientas. De esta forma se puede responder a las preguntas formuladas en la Tabla 4.1 y que a continuación se ejemplifican.

> En el proceso p de "Publicación" se puede utilizar un *endpoint de SPARQL* (sm) para publicar los datos y para ello se necesita: t_1-disponer de un *dataset* RDF $\mathcal{D}$, t_2-*desplegar un endpoint de SPARLQ* y t_3-insertar el *dataset* RDF $\mathcal{D}$ en el *endpoint de SPARQL*.

En este caso, para una mayor precisión se podría utilizar por ejemplo un interfaz web para insertar los datos o bien una consola, un *script*, etc., pero no es necesario especificar hasta ese nivel de detalle ya que dependería de las herramientas utilizadas para cada caso.

4.3 Relación con Modelos de Ciclo de Vida

Los modelos de ciclo de vida definen las distintas fases o pasos a realizar para la apertura y gestión de los datos enlazados. En general, se trata de procesos de alto nivel que sirven como guía tanto desde un punto de vista estratégico como técnico y que crean concienciación de los puntos clave para la apertura y enlazado de datos. En las siguientes tablas se realiza la alineación de los métodos que se han identificado inicialmente.

Etapa/Fase/Paso	**Proceso**
Identify	Producción
Model	Producción
Name	Producción
Describe	Producción
Convert	Producción, Validación
Publish	Publicación
Manteinance	Realimentación, Producción, Publicación, Consumo

Tabla 4.2: Alineación Métodos Semánticos y Ciclo de Vida de Bernadette Hyland

En este primer ciclo de vida, ver Tabla 4.2, existe un proceso de sumo interés como es el de mantenimiento de los datos generados. Por otra parte, cada uno de los pasos que se describen forman parte de un modo más específico de tareas a realizar para la producción y publicación de datos. En el contexto que nos ocupa, se entiende un método semántico como una función que realiza una transformación de datos, con un objetivo dentro de un ámbito de un proceso de mayor orden como producción, publicación o consumo.

Por ello, este primer ciclo de vida describe las tareas a realizar y no los métodos semánticos, ni los procesos de alto nivel.

Etapa/Fase/Paso	**Proceso**
Data Awareness	Producción
Modelling	Producción y Validación
Publishing	Publicación
Discovery	Producción, Publicación, Consumo, Realimentación
Integration	Producción
Use Cases	Consumo

Tabla 4.3: Alineación Métodos Semánticos y Ciclo de Vida de Michael Hausenblas

Al igual que en el modelo de ciclo de vida anterior, la versión propuesta por Michael Hausenblas, ver Tabla 4.3, identifica actividades a llevar a cabo que no están enclavadas en un proceso de mayor calado y que pueden ser comunes a varias de las fases. En este ciclo de vida cabe resaltar dos fases como la de *Integration* y *Use Cases* ya que esto significa que proporciona un enfoque totalmente orientado a la explotación, en realidad se puede considerar como parte del proceso de consumo de datos.

Etapa/Fase/Paso	**Proceso**
Specification	Producción
Modelling	Producción y Realimentación
Generation	Producción y Validación
Publication	Publicación
Exploitation	Consumo y Realimentación

Tabla 4.4: Alineación Métodos Semánticos y Ciclo de Vida de Boris Villazón-Terrazas

Las fases establecidas por Boris Villazón, ver Tabla 4.4, constituyen una buena guía con procesos de un mayor grado de abstracción que no indican exactamente los pasos a realizar para su consecución, nuevamente se hace hincapié en la explotación de los datos enlazados como parte esencial de esta iniciativa.

Etapa/Fase/Paso	**Proceso**
Selection	Producción
Conversion	Producción y Validación
Publication	Publicación
Interlinking	Producción
Exploitation	Consumo y Realimentación

Tabla 4.5: Alineación Métodos Semánticos y *DataLift Vision*

La visión del ciclo de vida realizada por *DataLift*, ver Tabla 4.5, utiliza un enfoque similar a las definiciones provistas en este documento sobre procesos, pero no tiene en cuenta directamente ni la realimentación, ni el control de la calidad como parte necesaria del proceso de promoción de datos a la iniciativa de *Linked Data*.

Etapa/Fase/Paso	**Proceso**
Interlinking/Fusing	Producción
Classification/Enrichment	Producción
Quality Analysis	Producción
Evaluation/Repair	Producción y Validación
Search/Browsing/Exploration	Publicación
Extraction	Consumo
Storage/Querying	Consumo y Realimentación
Manual revision/Authoring	Validación y Realimentación

Tabla 4.6: Alineación Métodos Semánticos y Ciclo de Vida de LOD2 Project

El trabajo desarrollado en el proyecto LOD2, ver Tabla 4.6, establece además de procesos, ciertas tareas clave para el éxito del consumo de los datos publicados. También es conveniente resaltar que por primera vez se establece la necesidad de validación manual en determinadas tareas.

Etapa/Fase/Paso	**Proceso**
Contextualization	Producción y Realimentación
Ontology Design	Producción y Realimentación
RDF Graph Modelling	Producción, Realimentación
SPARQL endpoint implementation	Publicación
RDF Graph implementation	Publicación
Update Graph Service	Realimentación
Documentation	Consumo
Data Visualization Tool	Consumo

Tabla 4.7: Alineación Métodos Semánticos y Metodología BCN y Universidad de Oviedo

La metodología y proceso de adopción, ver Tabla 4.7, desarrollada por la Universidad de Oviedo en conjunción con la Biblioteca del Congreso de Chile fija las tareas a desarrollar circunscribiéndose a un contexto concreto, la documentación oficial que emana de la actividad propia del Congreso de Chile, estos destacan especialmente por dos etapas: el servicio de actualización de datos y la documentación.

En general, la experiencia de estos ciclos de vida y metodologías de adopción de *Linked Data* y *Linking Open Data* se centran en las tareas a desarrollar y suelen coincidir en su orientación, aunque no es un nombrado. Sin embargo, no se contemplan tareas como la revisión de la calidad de los datos o la privacidad (esto es cuestionable refiriéndose a *Linking Open Data*), es por ello que la factorización realizada en distintos procesos de mayor nivel de abstracción, permite aislar las operaciones y realizar una separación de responsabilidades. Sin constituir un objetivo la obtención de un nuevo YALDC (*Yet Another Linked Data Life Cycle*) evidentemente a efectos de organización de las tareas realizadas en la aplicación de *Linked Data* en el ámbito de las licitaciones públicas, se ha decido utilizar el enfoque que se describe en este capítulo. Además, la novedad de estos ciclos de vida, surgidos desde casos prácticos, fundamenta la reutilización tanto de la experiencia de sus autores como la propia, para realizar el enfoque desde la teoría a la práctica.

4.4 Tareas Comunes en los Procesos de *Linked Data*

Independientemente del proceso, fase o etapa en la que se realicen operaciones con datos enlazados, se presentan repetitivamente situaciones susceptibles de subsanación, por ejemplo la identificación de vocabularios a utilizar en el proceso de producción o de realimentación o cómo diseñar las URIs de los recursos objeto de publicación. Para ello, en los ciclos de vida que se han repasado en la sección anterior, en los libros, especificaciones y buenas prácticas de la iniciativa de *Linked Data* se encuentran soluciones a estos problemas comunes. El objetivo de esta sección es presentar algunas de estas tareas para a partir de ellas, construir el método semántico que implementa un proceso dentro de esta propuesta, para la realización de estas tareas se pueden establecer una serie de responsables, prácticas y resultados de salida esperados. Tomando como base las metodologías tradicionales de software como Métrica V3 [89], se definen estos conceptos de forma sencilla y abierta para su posible ampliación.

Participante/Rol	**Responsabilidad**
Propietario de datos	Es el encargado de establecer la estrategia de apertura de datos. Tiene dos perfiles: técnico y de gestión.
Experto en el dominio	Es el conjunto de personas que se encargan habitualmente de los datos.
Desarrollador	Es el encargado de llevar a la práctica la iniciativa de *Linked Data* sobre los datos escogidos por el **Propietario** y modelados por el **Experto en el domino**.
Usuario final	Beneficiario de la apertura de datos como *Linked Data*. Puede ser una persona, entidad, etc., de perfil técnico o simplemente un usuario de Internet.

Tabla 4.8: Participantes/Roles en *Linked Data*.

Esta lista de participantes y responsabilidades, ver Tabla 4.8, establece una clasificación sencilla para identificar a los agentes implicados en llevar a cabo la apertura de datos, puede ser más extensa incluyendo consultores, analistas, etc., pero el objetivo es determinar de forma intensiva una serie de roles, no realizar una descripción extensiva de cada uno de los posibles participantes.

ID	**Tarea**	**Responsables**	**Prácticas**	**Resultado**
t_1	Análisis del *dataset* a transformar	Desarrollador, Propietario de datos y Experto en el dominio	Documentación previa, Reunión y Esquemas	Documentación de especificación inicial
t_2	Limpieza de datos	Desarrollador y Propietario de datos	Documentación previa, Reunión y Catálogos	Conjunto de datos "limpios"
t_3	Selección de Vocabularios	Desarrollador y Experto en el dominio	Documentación previa, Reunión, Esquemas y Catálogos	Catálogo de vocabularios candidatos

ID	Tarea	Responsables	Prácticas	Resultado
t_4	Selección de otros *datasets* RDF	Desarrollador y Experto en el dominio	Documentación previa, Reunión, Esquemas y Catálogos	Catálogo de *datasets* RDF
t_5	Modelado de datos en RDF	Experto en el dominio	Documentación previa, Reunión, Esquemas y Catálogos	Ontología de dominio $\mathcal{O}$ y *dataset* RDF $\mathcal{D}$
t_6	Diseño de un Esquema de URIs	Desarrollador, Propietario de datos y Experto en el dominio	Reunión y Esquemas	Catálogo de URIs y *dataset* RDF $\mathcal{D}$
t_7	Diseño Plantilla Objetivo del Recurso RDF	Desarrollador y Experto en el dominio	Reunión y Esquemas	Plantilla recurso RDF
t_8	Enriquecimiento de los datos en RDF	Desarrollador y Propietario de datos	Reunión y Esquemas	*datasets* RDF enriquecidos
t_9	Transformación de los datos a RDF	Desarrollador	Herramienta de generación de datos en RDF (preferible con validación)	RDF *dataset* $\mathcal{D}$
t_{10}	Reconciliación de Entidades	Desarrollador y Propietario de datos	Reunión y Esquemas	Conjunto EM de tuplas de recursos RDF ponderados
t_{11}	Ponderación de Recursos RDF	Usuario final	Programas de consumo de datos RDF	Conjunto de tuplas M de tuplas $< r_{RDF}, k >$
t_{12}	Validación de Recursos RDF	Desarrollador, Experto en el dominio y Propietario de datos	Tablas	Tabla de grado de cumplimiento de características y metainformación
t_{13}	Consolidación de datos RDF	Desarrollador, Experto en el dominio y Propietario de datos	Reunión y Esquemas	*Dataset* RDF consolidado
t_{14}	Infraestructura para *Linked Data*	Desarrollador y Propietario de datos	Reunión y Esquemas	Especificación de componentes de infraestructura
t_{15}	Acceso y formato en datos RDF	Desarrollador y Propietario de datos	Reunión y Esquemas	Especificación de despliegue de datos
t_{16}	Añadir metainformación a los recursos RDF	Desarrollador y Propietario de datos		Dataset RDF $\mathcal{D}$ enriquecido con metainformación de *provenance* y *trust*

ID	Tarea	Responsables	Prácticas	Resultado
t_{17}	Documentación extra	Desarrollador, Experto del dominio y Propietario de datos	Reunión, Tablas y Esquemas	Documentación a los procesos realizados

Tabla 4.9: Resumen de especificación de tareas.

La ejecución de estas tareas se enclavan dentro de los distintos procesos y métodos del ciclo de vida generando un flujo de trabajo, ver Figura 4.3, en el cual la generación de datos enlazados se convierte en un proceso de ingeniería cuantificable delimitado en el cual se pueden establecer métricas de tiempo y esfuerzo para la optimización de recursos.

4.4.1 Tarea t_1-Análisis del *dataset* a transformar

Esta tarea conlleva estudiar los datos que se van a transformar, identificando los tipos de datos a modelar, los conceptos y relaciones que se establecen entre los mismos. También hay que tener en cuenta el grado de evolución de los datos, si son dinámicos, es decir si varían a lo largo del tiempo y tienen su origen en una base datos, o bien si son estáticos procedentes de un fichero XML, CSV, etc., y no cambian en un plazo corto de tiempo. El resultado de esta tarea será una primera especificación de los datos a modelar, probablemente en lenguaje natural y a la cual se le dará soporte formal en la tarea específica de modelado de datos en RDF.

La responsabilidad de esta tarea deberá ser la conjunción del esfuerzo entre el *Desarrollador* y el *Propietario de datos*, tanto desde un punto de vista técnico como de dominio. Aunque el desarrollador sea capaz de entender los datos que está tratando es necesario conocer el por qué de esos datos, su significado y la forma de conseguirlos, para ello se pueden utilizar informes previos, entrevistas, etc.

En el ejemplo citado con anterioridad, el Nomenclátor de Asturias 2010 dispone de la siguiente información:

- Códigos: 2 cifras para indicar el Concejo (CC), 2 cifras para indicar la Parroquia (PP) y 2 cifras para indicar la entidad de población (EE). La concatenación de estas 6 cifras da lugar a un identificador único de la entidad de población.
- Nombre: en español según el Instituto Nacional de Estadística y en asturiano o nombre tradicional.
- Categoría: existen varias categorías en jerarquía según el tipo de la entidad de población (Concejo, Parroquia, Lugar, Ciudad, Villa, etc.)
- Datos estadísticos: datos físicos para la altitud, distancia y superficie, número de hombres y mujeres y número de viviendas principales y no principales.

Estos datos se encuentran disponibles en el formato del programa MSExcel y se pueden descargar a través de la página web del SADEI. Una entidad de población de ejemplo podría ser la siguiente, ver Tabla 4.10, también existen algunas peculiaridades dependiendo del tipo de población, por ejemplo los concejos no tienen altura, por lo que habría que tomar la máxima o la mínima de sus entidades de población o bien descartar el uso de este valor en las entidades que no lo posean.

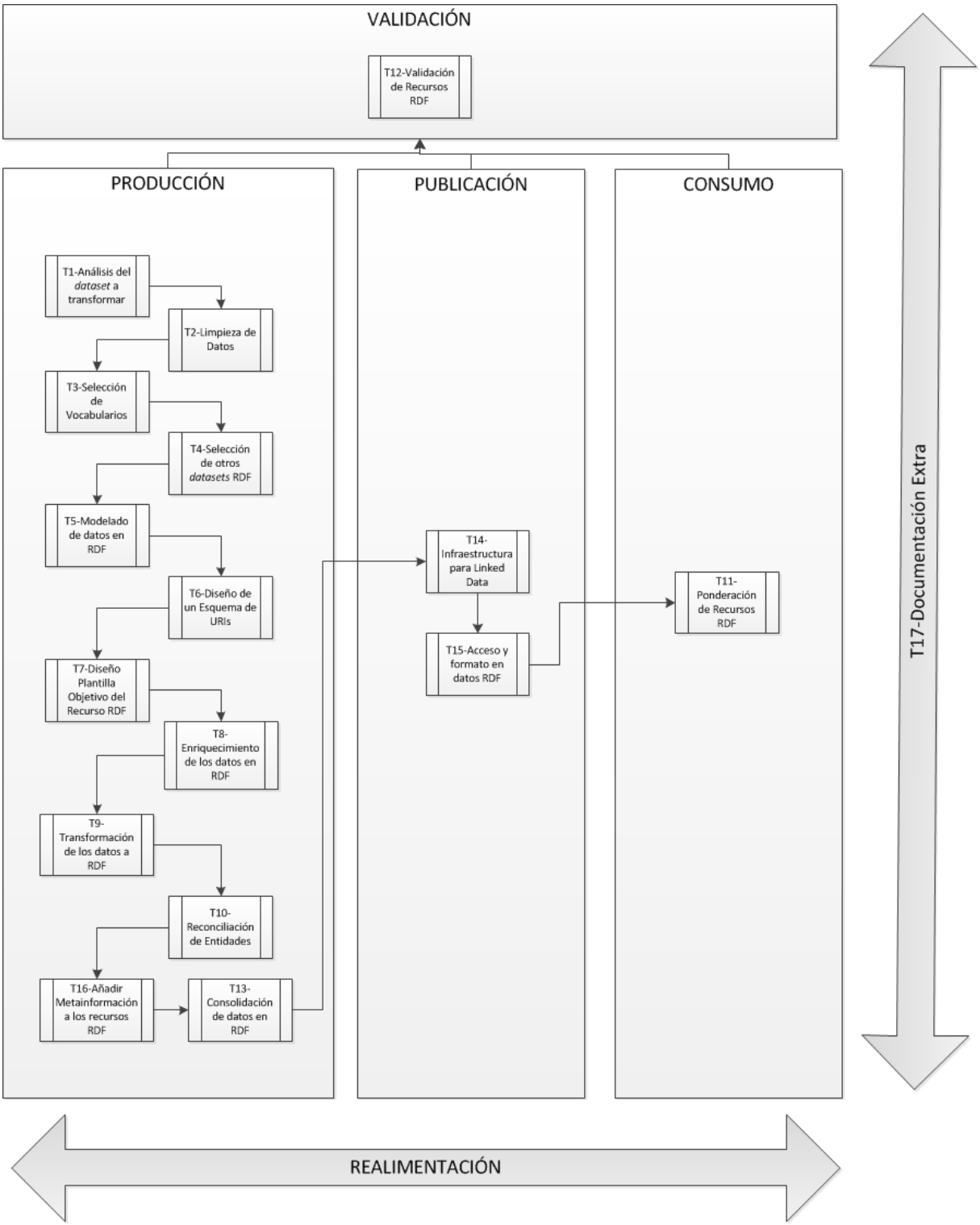


Figura 4.3: Flujo de Tareas en los distintos Procesos del Ciclo de Vida de Datos Enlazados.

Dato	Valor
Código Concejo	53
Código Parroquia	08
Código Entidad	02
Nombre	Llanuces
Nombre Tradicional	Chanuces
Tipo de Entidad	Lugar
Superficie km^2	7
Altitud m	870
Habitantes	28
Hombres	17
Mujeres	11
Viviendas	59
Viviendas Principales	15
Viviendas No Principales	44

Tabla 4.10: Ejemplo de entidad de población en el Nomenclátor de Asturias 2010.

```
"53","08","02","Llanuces","Chanuces","Lugar",,"7,00",870,28,17,11,59,15,44
```

Figura 4.4: Ejemplo de entidad de población en el Nomenclátor de Asturias 2010 en formato CSV.

4.4.2 Tarea t_2-Limpieza de Datos

La necesidad de esta tarea surge por la posibilidad de que las fuentes de datos a promocionar a la iniciativa de *Linked Data* dispongan de valores extraños o corruptos que no deberían formar parte del *dataset* final. Esta tarea denominada *Data Cleansing* es de vital importancia para asegurar que los datos generados cuentan con la calidad suficiente para ser reutilizados por terceros y no arrastrar valores incorrectos desde las fuentes originales. En realidad, adicionalmente cuando se realiza esta tarea se hace un proceso de depuración de datos que en muchos casos sirve para mejorar no sólo los datos publicados sino los datos procedentes de la fuente original.

El responsable de realización de esta tarea será el *Desarrollador* y el *Propietario de datos*, como resultado se obtendrá un conjunto de datos inicial con valores correctos.

En el caso del Nomenclátor de Asturias se producía la situación de que en cada tupla existía un carácter blanco extra en cada uno de los códigos, que no tiene ningún sentido para la generación final de datos, además otras cuestiones referidas al uso de minúsculas o mayúsculas no estaba unificado, por lo que la limpieza de estos datos para que sigan un convención como "*Camel Case*" o similar facilita posteriormente su transformación de forma correcta.

4.4.3 Tarea t_3-Selección de Vocabularios

Una vez identificadas las entidades y las relaciones de un dominio, es conveniente reutilizar los vocabularios que dan soporte para realizar estas definiciones y descripciones. El abanico de las posibilidades es muy amplio y existen diversas fuentes de consulta que mantienen una lista de los vocabularios más utilizados según el dominio: bibliografía, salud, etc. Debe atenderse a la existencia

de estas especificaciones e intentar impulsar la reutilización de conocimiento e información en su máxima expresión.

La responsabilidad de esta tarea deberá ser un esfuerzo conjunto entre el *Desarrollador* y un *Experto en el dominio* a modelar. El resultado de esta tarea debe ser una especificación de los vocabularios candidatos a ser reutilizados incluyendo la posibilidad de extensión de los mismos.

Siguiendo con el ejemplo citado, se identifica la necesidad de modelar las relaciones entre las entidades y por consiguiente, los vocabularios apropiados para ello podrían ser los siguientes:

- Información multilingüe, geográfica y estadística. La información multilingüe esta implícita en RDF, en cuanto a la geográfica, expresamente no se dispone de datos, pero podría obtenerse la georreferenciación de los lugares y emplear el vocabulario básico del W3C para albergar los datos geográficos, también existen otros vocabularios como el provisto por la iniciativa *GeoLinkedData*. Finalmente cabe citar que el vocabulario*RDF Data Cube* se ha desarrollado para la expresión de datos estadísticos.
- Jerarquía de entidades. Se puede valorar el uso de SKOS y SKOS-XL para la definición de una taxonomía.
- Tipos de conceptos: altitud, metros, superficie, kilómetros cuadrados, personas, hombres, mujeres, viviendas, etc. Se pueden reutilizar las definiciones realizadas en la DBPedia.

4.4.4 Tarea t_4-Selección de otros *datasets* RDF

Al igual que en la tarea anterior, una vez identificados los datos a transformar es necesario conocer con qué otros *datasets* se pueden enlazar los datos generados. Asumiendo que el objetivo final es la obtención de un *dataset* RDF 5 $\star$, esta tarea deviene fundamental.

La responsabilidad de esta tarea recae de nuevo en el *Desarrollador*, que verificará cómo se consumen esos datos, con qué datos de los que posee se pueden enlazar los recursos externos y finalmente el *Experto en el dominio* deberá validar que los *datasets* relacionados son adecuados en cuanto a su semántica y forma.

En el ejemplo citado se identifican los siguientes *datasets* a reutilizar:

- DBPedia para los conceptos identificados.
- GeoLinkedData y NUTS para la información geográfica.

Para buscar estos *datasets* candidatos se dispone de herramientas como *Datacatalogs.org*, *Freebase*, *Sindice*, *The Data Hub*, etc., pero en muchos casos por la propia experiencia o consultando *datasets* de temática similar se consigue averiguar cuáles reutilizar. También, consultando la documentación [209] de los grupos de trabajo del W3C se puede obtener una buena guía para la selección de los conjuntos de datos a reutilizar.

4.4.5 Tarea t_5-Modelado de datos en RDF

La iniciativa de *Linked Data* y *Linking Open Data* no sólo trata de publicar y consumir datos masivamente, en este caso habilitando el acceso a la base de datos corporativa, ficheros, etc., sería suficiente. Cada agente interesado en el consumo de datos debería investigar qué modelo siguen los mismos, un esquema relacional, XML Schema o si simplemente no siguen ningún modelo o, al menos, no es

posible acceder a él. Es por ello que surge la necesidad de establecer un modelo formal para las entidades, relaciones y datos. Dentro de la Web Semántica el uso de ontologías está ampliamente asentado y aceptado, por lo que habitualmente en el momento de la publicación de los datos se incluye una definición formal de los recursos RDF. Esta formalización puede ser implícita, por la reutilización de vocabularios y datos preexistentes, o bien explícita porque se haya creado un modelo particular para ese conjunto de datos.

La responsabilidad para la realización de este modelado recae sobre un *Experto en el dominio* de los datos a promocionar mediante *Linked Data*. La salida de esta tarea será una ontología de domino $\mathcal{O}$ para esos datos y parcialmente el conjunto $\mathcal{D}$ que define el *dataset* RDF.

Con el objetivo de ejemplificar esta tarea, se define una jerarquía en SKOS para modelar los tipos de entidades de población, ver Figura 4.5, y las estadísticas, ver Figura 4.6. Se trata de una versión parcial ya que este modelado tiene mayor calado, especialmente en la parte referida a estadísticas.

```
<http://purl.org/weso/nomenclator/ontology/Concejo>
        skosxl:prefLabel "Concejo"@es ;
        rdfs:subClassOf skos:Concept ;
        owl:sameAs <http://dbpedia.org/resource/Municipalities_of_Spain>;
        skos:example <http://purl.org/weso/nomenclator/asturias/2010/resource
            /01/00/00> ;
        rdfs:label "Concejo"@es .

<http://purl.org/weso/nomenclator/ontology/Parroquia>
        skosxl:prefLabel "Parroquia"@es ;
        rdfs:subClassOf skos:Concept ;
        skos:broaderTransitive <http://purl.org/weso/nomenclator/ontology/Concejo>;
        skos:example <http://purl.org/weso/nomenclator/asturias/2010/resource
            /01/01/00> ;
        rdfs:label "Parroquia"@es .

<http://purl.org/weso/nomenclator/ontology/Lugar>
        skosxl:prefLabel "Lugar"@es ;
        rdfs:subClassOf skos:Concept ;
        owl:sameAs <http://dbpedia.org/resource/Lugar>;
        skos:broaderTransitive <http://purl.org/weso/nomenclator/ontology/Parroquia>;
        skos:example <http://purl.org/weso/nomenclator/asturias/2010/resource
            /01/02/05> ;
        rdfs:label "Lugar"@es .
...
```

Figura 4.5: Modelo parcial de tipos de entidad con SKOS del Nomenclátor de Asturias 2010.

Atendiendo a las definiciones formales de una ontología se obtendría la siguiente descripción formal:

$\mathcal{O}_{nomenclator}$, donde $\mathcal{C} = \{Concejo, Parroquia...Lugar\}$, $\mathcal{R} = \{skosxl : prefLabel, rdfs : label...skos : broaderTransitive\}$, e $\mathcal{I}$ es el conjunto de todas las entidades de población.

$\mathcal{O}_{nomenclator\_stats}$, donde $\mathcal{C} = \{Region...Sexo\}$, $\mathcal{R} = \{area...refArea\}$, e $\mathcal{I}$ es el conjunto de todas observaciones estadísticas para cada entidad de población.

```
<http://purl.org/weso/nomenclator/stats/ontology/refArea> a rdf:Property, qb:
    DimensionProperty;
  rdfs:label "Region"@en;
  rdfs:subPropertyOf sdmx-dimension:refArea;
  rdfs:range skos:Concept;
  qb:concept sdmx-concept:refArea


<http://purl.org/weso/nomenclator/stats/ontology/physicaldata/area> a rdf:Property,
    qb:MeasureProperty
  rdfs:label "Area"@en;
  rdfs:subPropertyOf sdmx-measure:obsValue;
  rdfs:range xsd:decimal .
...
```

Figura 4.6: Modelo parcial de datos estadísticos del Nomenclátor de Asturias 2010.

4.4.6 Tarea t_6-Diseño de un Esquema de URIs

Se trata sin duda de una tarea clave para la correcta publicación y consumo de datos, el espectro de posibilidades para realizar un esquema de URIs contempla muchos factores y deberá dar respuesta, al menos, a las siguientes preguntas:

- ¿Qué tipo de URIs utilizar?, ¿"Slash" vs "Hash"?.
- ¿"Meaningful URIs" vs "ID based URIs"?.
- ¿La negociación de contenido forma parte del URI?.
- Factores de éxito del uso de "Cool URIs".
- Control de URIs-"minting HTTP URIs", ¿Referenciables?.
- Uso de fechas en URIs.
- URI para los recursos, URI para las definiciones, URI para la descripción del *dataset*, etc.
- URI para recursos y definiciones. URI base.
- ...

La responsabilidad de este diseño recae sobre el *Desarrollador*, el *Experto en el dominio* y el *Propietario de datos* ya que desde un punto de vista estratégico un correcto esquema de URIs sirve como documentación e incita a la reutilización de los datos enlazados. El resultado de esta tarea será una especificación de las URIs a utilizar y en consecuencia, el conjunto $\mathcal{D}$ que define un *dataset* RDF deberá quedar perfectamente establecido

En las Figuras 4.5 y 4.6 ya se había adelantado el diseño de URIs, no obstante, en la Tabla 4.11, se especifican con mayor detalle.

Característica	Decisión	Ejemplo
Tipo de separador en URI	*Slash*	
Tipo de URI	ID URI (promocionar nombres en una URI puede perjudicar su accesibilidad y usabilidad)	<base_uri>/CC/PP/EE
Negociación de Contenido en URI	No	Uso de cabeceras HTTP
Uso de "Cool URIs"	Si	
URIs referenciables	Si, el dominio está bajo nuestro control.	`http://purl.org/weso/nomenclator/asturias/2010/resource/53/08/02`
Uso de fechas	Si, el año 2010	$\equiv$
URIs para recursos	Si, sufijo *resource*	`http://purl.org/weso/nomenclator/asturias/2010/resource/{CC}/{PP}/{EE}`
URIs para definiciones	Si, sufijo *ontology*	
Base URI para recursos	Si	`http://purl.org/weso/nomenclator/asturias/2010/resource`
Base URI para definiciones	Si	`http://purl.org/weso/nomenclator/asturias/2010/ontology`
URI descripción del *dataset*	Si	`http://purl.org/weso/nomenclator/asturias/2010/resource/ds`

Tabla 4.11: Diseño de un esquema de URIs para el Nomenclátor de Asturias 2010.

Este esquema de URIs se prolongaría para el uso de estadísticas ya que se modela por un lado, los datos propios de la entidad de población, y, por otro, las estadísticas. Realmente con este diseño y modelado se estaría definiendo un *dataset* RDF con 4 grafos nombrados:

$\mathcal{D}_{nomenclator} = \{\mathcal{G},$
$(< http://purl.org/weso/nomenclator/asturias/2010 >, \mathcal{G}_1),$
$(< http://purl.org/weso/nomenclator/asturias/2010/ontology >, \mathcal{G}_2),$
$(< http://purl.org/weso/nomenclator/asturias/2010/stats >, \mathcal{G}_3),$
$(< http://purl.org/weso/nomenclator/asturias/2010/stats/ontology >, \mathcal{G}_4)\}$

4.4.7 Tarea t_7-Diseño Plantilla Objetivo del Recurso RDF

La experiencia y la coherencia en la promoción de datos a la iniciativa de *Linked Data* conduce habitualmente a la creación de un *recurso objetivo* que representa una plantilla o esqueleto, que sirve de guía para la transformación de los datos. Una vez que se ha realizado el proceso de producción,

este *recurso objetivo* suele desecharse ya que se dispone de miles de ejemplos en los datos ya transformados. Sin embargo, se puede dar el caso en el cual no todos los recursos RDF generados tengan el mismo número de tripletas y en consecuencia, resulte difícil validar si los datos generados son correctos respecto a la intención inicial. Es por ello, que la creación de un esqueleto de los recursos RDF durante toda la vida de los datos puede facilitar las tareas de validación en las cuales se chequean las relaciones que deben tener cada uno los recursos o el tipo de dato en los literales. Al igual que en la descripción de un *dataset*, se pueden agregar ejemplos de recursos, manteniendo esta información mediante una plantilla que pudiera ser objeto de consulta bajo una determinada URI. Por ejemplo, en el caso de la descripción de *datasets* utilizando voID se recomienda utilizar una convención de nombrado para ello, no obligatorio obviamente, válido para tareas como el descubrimiento automático. En este caso, se trataría de utilizar un enfoque similar para "guardar" la especificación de nuestros recursos RDF y así poder reutilizarla en posteriores procesos.

La responsabilidad de esta tarea recae en el *Desarrollador* y en el *Experto en el dominio*. Como resultado se obtiene un recurso RDF "ideal" que es el superconjunto de las propiedades utilizadas en nuestros recursos. Se podrían establecer varias plantillas por cada *dataset* dependiendo del modelado que se haya realizado, para así procurar soporte a distintas plantillas según una jerarquía.

```
<http://purl.org/weso/nomenclator/asturias/2010/resource/template>
      <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
            <http://purl.org/weso/nomenclator/ontology/Lugar> ;
      rdfs:label "Chanuces"@ast , "Llanuces"@es ;
      <http://purl.org/dc/elements/1.1/identifier>
            "53_08_02" ;
      <http://www.w3.org/2004/02/skos/core#broaderTransitive>
            <http://purl.org/weso/nomenclator/asturias/2010/resource/53/08/00> ;
      <http://www.w3.org/2008/05/skos-xl#prefLabel>
            "Chanuces"@ast , "Llanuces"@es .
```

Figura 4.7: Plantilla objetivo de un recurso RDF en el Nomenclátor de Asturias 2010.

No obstante y utilizando las actuales recomendaciones y con el objetivo de evitar un exceso de sobre-especificación, en el vocabulario de descripción de *datasets* voID se dispone de una propiedad: `void:exampleResource` que proporciona soporte a este enfoque.

4.4.8 Tarea t_8-Enriquecimiento de los datos en RDF

Esta tarea es posiblemente una de las más relevantes dentro de la iniciativa de *Linked Data* ya que su objetivo es generar los enlaces a otros *datasets* existentes, con la consiguiente reutilización de información y datos. Para la planificación de esta tarea se debe tener en cuenta qué tipo de datos se pretenden enriquecer y de acuerdo al catálogo de *datasets* preestablecido, llevar a cabo la ejecución del proceso de enlazado de datos. Los enfoques para realizar esta tarea son principalmente los dos siguientes: 1) manual y 2) automático. En ambos casos, la validación final manual es prácticamente inevitable ya que no se trata de un proceso determinista y puede verse sometido a cambios durante el ciclo de vida de los datos. Por ejemplo, se supone que las URIs de los recursos con los que se enlazan los datos perdurarán en el tiempo, pero pudiera ocurrir que se produjera algún cambio en las mismas. Un proceso automático sería recomendable, sin embargo en algunos casos la actuación manual es mucho más eficiente.

La responsabilidad de esta tarea recae sobre el *Desarrollador* en primer lugar y el encargado de mantenimiento por delegación del *Propietario de datos* en segunda instancia. El resultado de esta tarea será el enriquecimiento del *dataset* RDF.

Siguiendo con el ejemplo citado y una vez fijados los *datasets* a reutilizar se efectúa el enlazado de datos a través de relaciones definidas con este propósito, tales como `owl:sameAs`, `skos:relatedMatch`, etc.

```
<http://purl.org/weso/nomenclator/asturias/2010/resource/44/00/00>
      <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
            <http://purl.org/weso/nomenclator/ontology/Concejo> ;
      rdfs:label "Oviedo"@es , "Uvieu"@ast ;
      <http://purl.org/dc/elements/1.1/identifier>
            "44_00_00" ;
      <http://www.w3.org/2002/07/owl#sameAs>
            <http://geo.linkeddata.es/resource/Municipio/Oviedo> ,
            <http://dbpedia.org/resource/Oviedo> ;
      <http://www.w3.org/2004/02/skos/core#broaderTransitive>
            <http://nuts.psi.enakting.org/id/ES12> ;
      <http://www.w3.org/2008/05/skos-xl#prefLabel>
            "Oviedo"@es, "Uvieu"@ast .
```

Figura 4.8: Ejemplo de entidad RDF enriquecida del Nomenclátor de Asturias 2010.

4.4.9 Tarea t_9-Transformación de datos a RDF

De acuerdo a la ejecución de las tareas anteriores sobre el análisis del *dataset* a transformar, la selección del vocabularios, el diseño de URIs para los recursos, etc., cabe la realización de la transformación para la obtención del *dataset* $\mathcal{D}$ en RDF. Uno de los puntos clave para la ejecución de esta tarea consiste en la selección de la herramienta a utilizar, para ello se debe dar respuesta a la siguiente pregunta:

¿Se utilizará una herramienta de propósito general (ETL) como Google Refine o se implementará un programa *ad-hoc*?

Dependiendo de las capacidades de la herramienta y la necesidad de realizar otras tareas como la reconciliación de entidades pueden convertirse en buenas pistas para la selección de la misma. Igualmente, es conveniente que la expresión de las reglas de *mapeo* para cada una de las tuplas de entrada y su conversión en RDF sea lo más sencilla posible asegurando que los recursos generados son sintácticamente válidos. La implementación de un programa personalizado siempre es una posibilidad correcta en el caso de que la casuística de los datos de entrada sea muy diversa y no se puedan expresar todas las condiciones en las herramientas de propósito general. Finalmente, otro enfoque se centra en aunar las ventajas de cada una de las opciones y realizar la transformación inicial con una herramienta de propósito general y el enriquecimiento y validación de los datos en RDF con un programa personal.

El resultado de esta tarea es el *dataset* RDF $\mathcal{D}$ conteniendo todos los recursos de información de la entrada y para los cuales existe una regla de transformación. La responsabilidad, por lo tanto, recae sobre el *Desarrollador* en primer lugar y el encargado de mantenimiento por delegación del *Propietario de datos*, en segunda instancia.

4.4.10 Tarea t_{10}-Reconciliación de Entidades

El problema sobre el establecimiento de enlaces entre dos recursos RDF está siendo abordado actualmente desde diferentes puntos de vista con el objetivo de facilitar el enriquecimiento de los datos

de forma automática y determinista. Los enfoques actuales se centran en el procesamiento del lenguaje natural, utilizando las descripciones de los recursos en propiedades como `rdfs:label`, similitud en URIs, etc. Las herramientas que se pueden utilizar para este propósito abarcan desde sistemas de búsqueda como *Sindice*, a *frameworks* completos como *SERIMI*, utilizando para ello grandes bases de datos como *Freebase*.

Por tratarse de una subtarea del enriquecimiento de datos, la responsabilidad recae sobre el *Desarrollador* en primer lugar y el *Encargado de Mantenimiento* por delegación del *Propietario de datos*, en segunda instancia.

El resultado de esta tarea será un conjunto EM de tuplas $< r_{source}, r_{target}, k >$, donde r_{source} es el recurso RDF de entrada para el cual se buscan candidatos y r_{target} es un recurso RDF candidato valorado con un valor k.

También, adicionalmente se puede utilizar este tipo de tarea para descubrir entidades similares y realizar procesos de descubrimiento y análisis de información y datos de forma automática

4.4.11 Tarea t_{11}-Ponderación de Recursos RDF

El consumo de datos RDF puede requerir el establecimiento de un mecanismo para la obtención de un valor acerca del nivel de representatividad de un recurso en un determinado contexto. Básicamente el enfoque puede determinarse de acuerdo a un conjunto de relaciones y valores de entrada, otorgando una ponderación a cada recurso para que pueda ser utilizado en otras tareas.

El responsable de realizar esta tarea será un *Usuario final* como consumidor de los datos. Se puede aplicar también a la reconciliación de entidades, y en consecuencia al enriquecimiento de los datos en RDF. El resultado de esta tarea será un conjunto M de tuplas $< r_{RDF}, k >$, en el cual para cada recurso r_{RDF} perteneciente a un *dataset* RDF $\mathcal{D}$ se establece un valor k de relevancia.

Esta tarea resulta relevante igualmente para realizar búsquedas sobre un *dataset* RDF, ya que los actuales lenguajes de consulta como SPARQL no permiten establecer una prioridad para el encaje o "matching" de tripletas que impide una ordenación de los resultados por relevancia de acuerdo a ciertos criterios. Algunas herramientas como Virtuoso de OpenLink han implementado este tipo de características, pero carecen de especificación formal en las recomendaciones actuales del W3C.

4.4.12 Tarea t_{12}-Validación de Recursos RDF

El objetivo final de la exposición de datos con RDF debe ser la reutilización de los mismos para la creación de otras aplicaciones, para ello se deben proveer datos que cumplan ciertas características de acuerdo al modelo establecido. Es por ello que la validación de los recursos RDF generados debe ser prioritaria para asegurar la calidad de los datos que se han obtenido, así se puede atender a diferentes criterios:

- Deben ser datos RDF correctos en cuanto al estricto cumplimiento de la sintaxis normativa.
- Deben ser correctos en cuanto a rango y dominio de los literales utilizados y previamente definidos en el modelo. Se deben crear instancias del conjunto $\mathcal{I}$ de la ontología $\mathcal{O}$.
- Deben proveer la metainformación adecuada para cumplir con los requisitos de confianza (*trust*) y poder conocer su procedencia (*provenance*). En algunos casos, esta información debe estar disponible en cada recurso individualmente o a nivel de *dataset* RDF, dependiendo bien del contexto en el que se vayan a utilizar o bien con respecto a las características del propio *dataset*. Por ejemplo, tratándose de leyes que pueden tener un sólo identificador y que evolucionan en

el tiempo es conveniente que estas características se fijen a nivel de recurso. En cambio, si todos los datos del *dataset* varían en grupo será más conveniente añadir esta metainformación a nivel del propio *dataset* y no multiplicar esta información en cada uno de los recursos.

- Dependiendo de las relaciones y propiedades que se hayan establecido se debe asegurar que están presentes en todos los recursos.
- En relación con otras características igualmente recomendables como la licencia de uso, autoría, autenticidad, autorización de uso, el no repudio, etc., es necesaria su inclusión dentro del conjunto de metainformación provista con el objetivo de asegurar su consumo en condiciones de fiabilidad.

El responsable de la realización de esta tarea será el *Desarrollador*, el *Experto en el dominio* y el *Propietario de datos*. El resultado será un conjunto de directrices en el cual se especifique el grado de cumplimiento de cada una de ellas y cómo se verifique el mismo.

En el ejemplo que se está desarrollando y acoplando a lo largo de este capítulo se asegura que:

- Los datos RDF son válidos porque se han generado con herramientas que producen un RDF válido.
- Se sigue el modelo realizado para especificar cada uno de los recursos.
- Se añade metainformación, ver Figura 4.9, sobre el *dataset* a nivel del mismo ya que la evolución y el cambio en los datos se produce a nivel de conjunto.

4.4.13 Tarea t_{13}-Consolidación de datos en RDF

Esta tarea se encarga de proveer nuevos datos al *dataset* RDF agregando parte de los datos ya disponibles. El caso paradigmático ocurre en materia de estadísticas en el que además de las observaciones particulares se pueden diseñar recursos RDF integrando cierta información de alto valor para que no sea necesaria su réplica por cada uno de los consumidores. Por lo tanto, esta tarea consiste en la agregación de los datos RDF ya disponibles.

El responsable de esta tarea será el *Desarrollador*, el *Experto en el dominio* y el *Propietario de datos* que seleccionarán aquellos datos que deben ser agregados con el objetivo de proveer un mejor acceso a los mismos dotando así al *dataset* RDF de mejores características de accesibilidad y usabilidad de los datos, resultando un *dataset* RDF extendido.

Dentro del Nomenclátor de Asturias 2010 consta consta una sección dedicada a observaciones estadísticas que siguiendo las especificaciones para la publicación de datos estadísticos en RDF se pueden agrupar mediante "*Slices*", se trataría de un caso de consolidación de datos. Por ejemplo, se podrían proveer datos que dieran respuesta directamente a preguntas como: "Número de personas de género masculino en Llanuces en el año 2010". En el vocabulario *RDF Data Cube* este tipo de información se modelaría a través de un "Slice", ver Figura 4.10, en el cual se tendrían en cuenta 3 dimensiones (género, región y fecha) y una medida (personas).

```
<http://purl.org/weso/nomenclator/asturias/2010/resource/ds> a void:Dataset ;
              rdf:type skos:ConceptScheme ;
              rdfs:label "Nomenclator Asturias 2010"@en ;
              void:dataDump <http://purl.org/weso/nomenclator/asturias/2010/
                  nomenclator-asturias-2010.ttl> ;
              dcterms:title "Nomenclator Asturias 2010" ;
              dcterms:description "RDF data extracted from INE and Sadei" ;
              dcterms:source <http://www.sadei.com/> ;
              dcterms:license <http://opendatacommons.org/licenses/by/1.0/> ;
              dcterms:publisher <http://www.josemalvarez.es/foaf.rdf#me> ;
              dcterms:author <http://www.josemalvarez.es/foaf.rdf#me> ;
              dcterms:author <http://www.di.uniovi.es/~labra/labraFoaf.rdf#me> ;
              dcterms:contributor <http://purl.org/weso/nomenclator/asturias/2010/
                  resource/contributor/Weso> ;
              dcterms:modified "2011-11-10"^^xsd:date ;
              void:vocabulary skosxl: , skos: , geo: ;
              void:target <http://purl.org/weso/nomenclator/asturias/2010/stats/
                  resource/ds> ;
              foaf:homepage <http://purl.org/weso/nomenclator/> ;
              void:exampleResource <http://purl.org/weso/nomenclator/asturias/2010/
                  resource/01/00/00> ;
              void:exampleResource <http://purl.org/weso/nomenclator/asturias/2010/
                  resource/44/00/00> ;
              void:exampleResource <http://purl.org/weso/nomenclator/asturias/2010/
                  resource/53/00/00> ;
              skos:hasTopConcept   <http://purl.org/weso/nomenclator/asturias/2010/
                  resource/01/00/00> ,
               ...
               .
<http://purl.org/weso/nomenclator/asturias/2010/stats/resource/ds> a void:Dataset ;
              rdfs:label "Statistics Nomenclator Asturias 2010"@en ;
              void:dataDump <http://purl.org/weso/nomenclator/asturias/2010/stats/
                  nomenclator-asturias-2010-stats.ttl> ;
              dcterms:title "Statistics Nomenclator Asturias 2010" ;
              dcterms:description "RDF data extracted from INE and Sadei" ;
              dcterms:source <http://www.sadei.com/> ;
              dcterms:license <http://opendatacommons.org/licenses/by/1.0/> ;
              dcterms:publisher <http://www.josemalvarez.es/foaf.rdf#me> ;
              dcterms:author <http://www.josemalvarez.es/foaf.rdf#me> ;
              dcterms:author <http://www.di.uniovi.es/~labra/labraFoaf.rdf#me> ;
              dcterms:contributor <http://purl.org/weso/nomenclator/asturias/2010/
                  resource/contributor/Weso> ;
              dcterms:modified "2011-11-10"^^xsd:date ;
              void:vocabulary qb: , sdmx-code: , sdmx-dimension: ;
              foaf:homepage <http://purl.org/weso/nomenclator/> ;
              void:exampleResource <http://purl.org/weso/nomenclator/asturias/2010/
                  stats/resource/physicaldata/altitude/01/00/00> ;
              void:exampleResource <http://purl.org/weso/nomenclator/asturias/2010/
                  stats/resource/sex/m/01/00/00> ;
              void:exampleResource <http://purl.org/weso/nomenclator/asturias/2010/
                  stats/resource/sex/f/01/00/00> ;
```

Figura 4.9: Metainformación del *dataset* RDF del Nomenclátor de Asturias 2010.

```
nomen-stats: sliceByRegionSex a qb:SliceKey;
  rdfs:label "Slice by region"@en;
  rdfs:comment "Fixed year, region and sex change"@en;
  qb:componentProperty
  nomen-stats:refPeriod;
 .

nomen-stats: spopulation a
  qb:DataStructureDefinition;
  qb:component
    [qb:dimension nomen-stats:refPeriod; ],
    [qb:dimension nomen-stats:refArea; ],
    [qb:dimension sdmx-dimension:sex; ],
    [qb:measure
    nomen-stats:population; ];
qb:sliceKey nomen-stats: sliceByRegionSex .


nomen-stats:region/sex a qb:Slice;
  qb:sliceStructure
  nomen-stats: sliceByRegionSex;
  nomen-stats:refPeriod
  <http://reference.data.gov.uk/id/gregorian-interval/2010-01-01T00:00:00/P1Y> ;
  qb:observation
  nomen-obs:region/sex/m/53/08/02, ...
.

nomen-obs:region/sex/m/53/08/02 a qb:Observation;
  qb:dataSet nomen-stats:nomenclator2010;
  nomen-stats:refArea <http://purl.org/weso/nomenclator/asturias/2010/resource
      /53/08/02> ;
  nomen-stats:refPeriod <http://reference.data.gov.uk/doc/gregorian-interval
      /2010-01-01T00:00:00/P1Y> ;
  sdmx-dimension:sex sdmx-code:sex-M ;
  sdmx-attribute:unitMeasure <http://dbpedia.org/resource/Person>
  nomen-stats:population 17 ;
.
```

Figura 4.10: Ejemplo de *Slice* en el *dataset* RDF del Nomenclátor de Asturias 2010.

4.4.14 Tarea t_{14}-Infraestructura para *Linked Data*

La consecución de todos los procesos, métodos y tareas puede reflejarse en una infraestructura de herramientas que proporcionen soporte desde dos puntos de vista:

Completa. Cubre todos los procesos, métodos y tareas suministrando las herramientas adecuadas para cada uno de ellos. Los criterios de selección pueden variar dependiendo del estado de madurez, de los datos a tratar, de las tareas a realizar, etc. En este sentido la infraestructura definida en *LOD2 project*, ver Sección 3.2.6.7, se puede considerar como paradigmática.

Parcial. Cubre alguno de los procesos, métodos o tareas. Por ejemplo, para el proceso de publicación, se puede establecer una infraestructura especial que no tenga en cuenta ningún otro proceso y de por hecho la existencia de datos. A continuación y siguiendo el ejemplo que se desarrolla a lo largo de este capítulo se advierte un modelo para para la publicación de datos, ver Figura 4.11, que supone la existencia de datos en un cierto repositorio y que no provee más operaciones que la de acceso a datos sin realimentación.

El responsable de esta tarea será el *Desarrollador* y el *Propietario de datos* que deberán resolver, de acuerdo a unas determinadas especificaciones, el conjunto de componentes que den soporte a los datos. La especificación de este despliegue puede ser una simple lista de componentes o el uso de diagramas UML particulares para este tipo de procesos, como el diagrama de componentes o de despliegue.

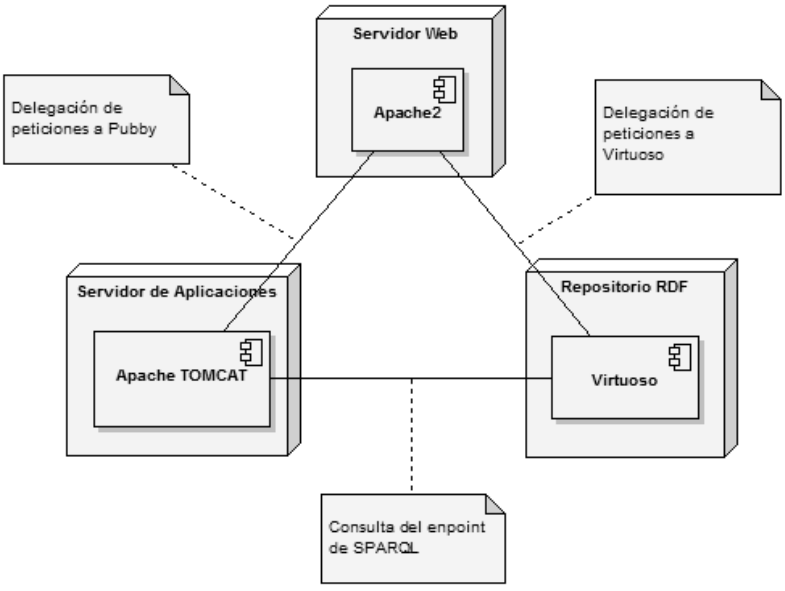


Figura 4.11: Infraestructura Parcial para el Nomenclátor de Asturias 2010.

4.4.15 Tarea t_{15}-Acceso y Formato de datos en RDF

Seguidamente al proceso de generación de datos en RDF como *Linked Data* es necesario establecer los mecanismos de acceso a la información, en esta tarea se debe dar respuesta a una serie de cuestiones:

- ¿Cómo se va a desplegar la información?, ¿Fichero, repositorio, generación *on-the-fly*, etc.?

- ¿En qué formatos se va a ofrecer esta información?
- ¿Qué tipo de acceso se va a permitir?, ¿Consulta directa a un *endpoint* de SPARQL?, ¿Un servicio web: REST, SOAP, etc.?

La responsabilidad de esta tarea recae en el *Desarrollador* y el *Propietario de datos* que han de decidir cómo se ha de hacer pública esta información de acuerdo a la estrategia de la entidad, de la infraestructura tecnológica y de los datos en sí mismos. El resultado de esta tarea debe proveer una especificación sobre cómo acceder a los datos en RDF y en qué formatos se van a facilitar.

En el ejemplo que se está desarrollando, se ha decidido la siguiente estrategia:

- Despliegue mediante un repositorio semántico con un *endpoint* de SPARQL habilitado y ficheros con los volcados completos de los datos.
- Los siguientes formatos deben ser prioritarios: HTML, RDF/XML y N3.
- Acceso directo a consultar sobre el *endpoint* de SPARQL y un servicio de negociación de contenido sobre los datos.

4.4.16 Tarea t_{16}-Añadir metainformación a los recursos RDF

El objetivo de esta tarea es dotar a los recursos de información de la metainformación necesaria para suministrar soporte a la procedencia de los datos y garantizar un cierto grado de confianza en su uso. En esta tarea, existen dos enfoques posibles:

1. Incluir la metainformación a nivel de *dataset*. Esta opción se utilizará cuando los recursos tengan un carácter estático y evolucionen poco en el tiempo y en conjunto, de esta forma, tiene sentido incluir sólo metainformación a nivel del conjunto completo de datos para no generar información superflua.
2. Incluir la metainformación a nivel de cada recurso. Este enfoque se utilizará cuando los recursos tengan un carácter dinámico y evolucionen en el tiempo independientemente de otros recursos en el conjunto de datos completo.
3. Enfoque híbrido. Evidentemente, si sólo algunos recursos del *dataset* son candidatos a incluir metainformación a nivel de cada recurso es preferible optar por este enfoque facilitando la información donde sea necesaria. La principal desventaja de esta opción reside en que la transformación de los datos no es homogénea generando incertidumbre en los posibles clientes.

La responsabilidad de esta tarea recae en el *Desarrollador* y el *Propietario de datos* que han de decidir qué metainformación se debe incluir y qué enfoque aplicar. El resultado de esta tarea es el propio *dataset* RDF, pero incluyendo nuevas tripletas para representar la información necesaria como metadatos.

4.4.17 Tarea t_{17}-Documentación extra

Uno de los puntos clave para la reutilización de datos enlazados reside en la provisión de una documentación adecuada en relación con el tipo de recursos de información que se han transformado y qué datos contienen. Desde un punto de vista estratégico las decisiones que se hayan tomado en cuanto a selección de vocabularios y otros *datasets* a reutilizar, suponen una buena guía de documentación facilitando a los clientes datos perfectamente conocidos, ahora bien, desde una perspectiva operativa es conveniente documentar las decisiones tomadas como el diseño de URIs, el tipo de

datos utilizado, el acceso a la información, las formas de consulta, posibles restricciones, etc., para que la selección del *dataset* se convierta en una decisión sencilla. Un ejemplo claro de documentación correcta es la provista por *GoodRelations* y *ProductOntology* en las cuales además de los datos en sí mismos, se ofrecen ejemplos de uso, características, casos de éxito en su reutilización, etc., favoreciendo su difusión entre la comunidad.

La responsabilidad de esta tarea recae en el *Desarrollador* y el *Propietario de datos* que han de establecer qué partes se pueden documentar públicamente y la forma de difusión la misma, mediante metadatos, etc. El resultado de esta tarea será, en general, una especificación de cómo se deben reutilizar los datos. En este sentido, en las últimas fechas ha habido una extensa discusión en la lista de correo de *Linking Open Data* en el W3C sobre el uso de `HttpRange-14` para consultar cómo se debería incluir información de unos recursos en otros.

El resultado de esta tarea será una serie de documentos de especificación y ejemplos, así como nuevas tripletas RDF en el *dataset* resultado para enlazar esta nueva metainformación.

4.5 Proceso de Producción

Definición 4.5.1 (Producción)**.** *Proceso que conlleva todas las tareas que implican la transformación de datos a un modelo RDF obteniendo como resultado un dataset RDF $\mathcal{D}$.*

4.5.1 Método Semántico de Producción de *Linked Data*

Definición 4.5.2 (Método Semántico de Producción)**.** *SPM es una función que para un conjunto de tuplas de entrada o dataset $\mathcal{G}$ y un conjunto de mapeos $\mathcal{M}$ genera un dataset RDF $\mathcal{D}$.*

$$SPM : \mathcal{G} \times \mathcal{M} \longrightarrow \mathcal{D}$$

Las características que tiene esta definición de un SPM son las siguientes:

- Para todo valor f_i del conjunto de entrada $\mathcal{G}$, existe al menos un *mapeo* m_i en el conjunto $\mathcal{M}$.
- Existen valores f_i del conjunto de entrada $\mathcal{G}$, para los que existen varios *mapeos* m_i formando un conjunto $\mathcal{M}_i$ subconjunto de $\mathcal{M}$.

Esta definición implica que toda la información de entrada permanece en el conjunto de salida. Si bien se podría pensar que existen valores de la entrada, por ejemplo un identificador generado automáticamente desde una base de datos que no sería candidato a persistir en el conjunto de salida, es conveniente mantener toda la información por la posibilidad de realizar realimentación y monitorización transversal. Si se apreciara la existencia de algún inconveniente con ocasión del consumo o de la publicación se deben proveer otros mecanismos para hacer transparente esta información.

4.5.2 SPM_1-Transformación de datos a RDF

Este método de producción de un *dataset* RDF, SPM_1, asume que los datos disponibles se encuentran en algún formato procesable automáticamente y que en cierto modo son estáticos o bien es una imagen estática de los mismos. Se trata en concreto del supuesto en el que se dispone de datos en formatos como CSV o MSExcel o bien se hace un volcado de una base de datos para generar posteriormente a partir de estos valores de entrada el *dataset* RDF.

Por lo tanto, este enfoque se aplicará ante la siguiente situación:

- Datos estáticos. Inicialmente sin disponibilidad de acceso mediante un lenguaje de consulta tipo XPath o SQL.
- Formato de datos preestablecido tipo CSV, MSExcel o XML.
- El tamaño del *dataset* de entrada no es excesivamente grande. No existe una forma de evaluar el tamaño de un *dataset*, pero se podría fijar que el resultado final fuera en orden de millones de tripletas RDF.

4.5.3 SPM_2-*Mapeo* con Base de Datos

Se trata de un método híbrido de producción y publicación de datos enlazados, SPM_2, que asume la existencia de una base de datos, en la mayoría de los casos relacional, con datos preestablecidos a los cuales se puede acceder a través de un lenguaje de consulta tipo XPath, XQuery o SQL. Este procedimiento se fundamenta en la realización de un *mapeo* con los datos que se pretenden exportar como RDF y la estructura de los mismos. Se pueden mezclar datos de diferentes tablas y utilizar los lenguajes de consulta para generar datos agregados. Este enfoque se suele utilizar cuando existe una gran base de datos de complejo volcado a ficheros y cuya transformación y posterior publicación sería un proyecto de gran magnitud en sí mismo. Por lo tanto, se opta por una solución en la cual se utiliza una capa intermedia que se encarga de acceder a los datos ya existentes y producir RDF bajo demanda. De esta forma, se pueden reutilizar las bases de datos existentes, se considera un método no intrusivo ya que sólo requiere los datos, pero puede provocar una sobrecarga en la actual base de datos, dada esta situación se suele optar por realizar el envoltorio semántico sobre una imagen de la misma, así se protege la operatividad ordinaria mientras que se confiere el acceso mediante datos enlazados. La solución que propone este método debería utilizarse cuando el tamaño de los datos es desmesurado y está en continua evolución, de tal manera que no es rentable realizar SPM_1 en cada actualización de datos. La ventaja implícita reside precisamente en este último punto ya que se ofrece soporte directo a la actualización automática de la vista en RDF.

Por lo tanto, este enfoque se aplicará ante la siguiente situación:

- Datos dinámicos. Con posibilidad de acceso mediante un lenguaje de consulta tipo XPath o SQL.
- Modelo y formato de datos formal, por ejemplo basado en el Modelo Entidad/Relación.
- El tamaño del *dataset* de entrada suele ser grande. El resultado final estaría en el orden de billones de tripletas RDF.

4.5.4 SPM_3-Consulta y transformación a RDF

El método SPM_3 se utiliza cuando se cuenta con un sistema de gestión de datos al que se puede acceder a través de un lenguaje de consulta y que además de almacenar datos estrictamente, también contiene otro tipo de información como por ejemplo texto. Este método se suele utilizar en bases de datos documentales en las que la unidad de información es un documento con metadatos. Los datos promocionados a *Linked Data* suelen incluir la metainformación de los documentos esquivando la transformación de toda la información con el objetivo de facilitar el acceso a los mismos. En resumen se provee acceso a la metainformación que está estructurada mientras que los textos completos y libres, o bien no se promocionan o bien se ofrecen de forma completa sin análisis previo a través de otros vínculos.

Por lo tanto, este enfoque se aplicará ante la siguiente situación:

- Datos dinámicos, sin posibilidad de acceso mediante un lenguaje de consulta tipo XPATH o SQL.
- Formato de datos preestablecido, sin embargo no tiene porque existir un modelo formal para los mismos.
- El tamaño del *dataset* de entrada no debería ser relevante, simplemente se generan los datos en RDF bajo demanda.

4.5.5 Tabla de Decisión del Proceso de Producción

En resumen y considerando las tres variables siguientes: 1) evolución y actualización de los datos; 2) la existencia de un modelo formal o acceso mediante lenguaje de consulta y 3) el tamaño del *dataset* resultado, es factible establecer la siguiente tabla de decisión, ver Tabla 4.12.

Método	**Tipo**		**Formato**		**Modelo Formal**		**Tamaño Final**	
	Estático	Dinámico	Estructura	Consulta	Si	No	Millones	Billones
SPM_1	$\star$		$\star$			$\star$	$\star$	
SPM_2		$\star$		$\star$	$\star$			$\star$
SPM_3		$\star$	$\star$	$\star$	$\star$		$\star$	$\star$

Tabla 4.12: Tabla de Decisión del Proceso de Producción de *Linked Data*.

4.6 Proceso de Publicación

Definición 4.6.1 (Publicación)**.** *Proceso que conlleva todas las tareas que implican el despliegue público de un dataset RDF $\mathcal{D}$.*

4.6.1 Método Semántico de Publicación de *Linked Data*

Definición 4.6.2 (Método Semántico de Publicación)**.** *$\mathcal{SPM}$ es una función que para un dataset RDF $\mathcal{D}$ y un conjunto de características de publicación $\mathcal{P}$, obtiene como resultado un dataset RDF $\mathcal{D}_{pub}$ publicado cumpliendo $\mathcal{P}$.*

$$\mathcal{SPM} : \mathcal{D} \times \mathcal{P} \longrightarrow \mathcal{D}_{pub}$$

Las características contenidas en esta definición de un $\mathcal{SPM}$ son las siguientes:

- $\mathcal{D} \approxeq \mathcal{D}_{pub}$, ya que se pueden utilizar técnicas para filtrar ciertos datos presentes en $\mathcal{D}$.
- El conjunto de características de publicación $\mathcal{P}$ se utilizan para la validación de $\mathcal{D}_{pub}$.

Esta definición de método semántico de publicación implica que para la existencia de datos en RDF se establecen un conjunto de características que se deben cumplir en el momento de la publicación como puede ser la negociación de contenido, URIs referenciables, acceso mediante protocolos estándar como HTTP, etc. El objetivo es proveer un *dataset* RDF que ofrezca las características necesarias para ser considerado como *Linked Data*, en este sentido se intenta que los datos publicados alcancen el nivel de 5 $\star$.

Adicionalmente, dependiendo del método seleccionado de publicación se debe proveer una infraestructura de almacenamiento de los datos que puede ser un fichero, un repositorio RDF o una base de datos, de acuerdo al tipo seleccionado se podrán publicar los datos de una u otra forma. Si bien el proceso de almacenamiento de los datos generados puede estimarse como una tarea de producción o publicación, realmente se trata de un proceso implícito, de ahí que el proceso de publicación sea dependiente del tipo de almacén utilizado. No obstante, existen soluciones que se combinan, por ejemplo se pueden publicar los datos de un fichero RDF mediante un *endpoint* de SPARQL sin necesidad de disponer de un repositorio RDF nativo, o bien se pueden utilizar extensiones de algunos sistemas de gestión de base de datos [181], siendo capaces de manejar recursos RDF. La decisión sobre la solución que se deba adoptar, se tomará con ocasión del diseño de la infraestructura de *Linked Data* que condicionará estratégicamente los procesos posteriores a la generación de los datos.

4.6.2 $\mathcal{SP}M_1$-Fichero estático en RDF

Este método de publicación, $\mathcal{SP}M_1$, simplemente ofrece el conjunto de datos a través de un fichero alojado en un servidor web, siendo accesible mediante un protocolo estándar. Es la forma de publicación que facilita en menor grado el acceso y reutilización de información, ya que para conseguir la descripción de cualquier recurso se debe obtener el *dataset* completo con la consiguiente sobrecarga para su consumo posterior. Además es necesario configurar correctamente los tipos MIME en el servidor web para que se ofrezca el contenido de modo adecuado.

La publicación de ficheros es una de las formas más utilizadas en multitud de circunstancias debido a su sencillez, no requiere ningún tipo de infraestructura adicional, tanto desde el punto de vista de *Open Data* como de *Linked Data*, ya que sólo requiere hacer pública la información sin valorar otro tipo de consideraciones.

Por ejemplo, en el supuesto referido sobre el Nomenclátor de Asturias 2010 es posible habilitar en el servidor web los volcados de los datos RDF (serializados en N3) en las siguientes URIs:

- <base_uri>= `http://purl.org/weso/datasets/nomenclator/asturias/2010`
- <base_uri>/nomenclator-asturias-2010.ttl
- <base_uri>/nomenclator-asturias-2010-ontology.ttl
- <base_uri>/stats/nomenclator-asturias-2010-stats.ttl
- <base_uri>/stats/nomenclator-asturias-2010-stats-ontology.ttl

Además, se debe configurar en el servidor web, en este caso Apache2 fichero `/etc/mime.types`, cómo servir el tipo MIME con extensión "ttl" mediante la adición de la línea `text/turtle ttl`. Como se puede observar este método de publicación es el más básico y requiere tanto la realización de configuración manual como la necesidad de verificar el correcto funcionamiento siguiendo las prácticas de publicación del W3C.

4.6.3 $\mathcal{SP}M_2$-*Mapeo* con Base de Datos

Como se ha expuesto en la Sección 4.5.3 se trata de un método híbrido de producción y publicación, $\mathcal{SP}M_2$, de datos enlazados. En este caso, los datos se generan bajo demanda mediante una función de *mapeo* entre la descripción RDF que se debe producir y los datos almacenados en la base de datos. En cuanto a la publicación, la descripción RDF se genera bajo demanda atendiendo a las características prefijadas y es similar al caso que se presentará en la Sección 4.6.6 sobre la publicación

con un *Linked Data Frontend*, con la diferencia de que los datos no se hayan previamente en RDF, sino que se generan dinámicamente.

4.6.4 $\mathcal{SPM}_3$-*Endpoint* de SPARQL

El uso de un *endpoint* de SPARQL está ampliamente asentado en la comunidad de *Linked Data* como método de publicación estándar, $\mathcal{SPM}_3$, mediante el cual además de publicar la información de los recursos RDF, se ofrece un servicio de consulta mediante el lenguaje SPARQL. Este método de publicación requiere el uso de un repositorio RDF en el que se almacenan los datos RDF previamente transformados por lo que es incompatible con los métodos de generación de RDF bajo demanda.

La aplicación de este método de publicación es la más interesante desde el punto de vista de la reutilización ya que permite la creación y ejecución de consultas sobre los datos RDF por lo que se pueden consolidar aquellos previamente no generados. Desde un punto de vista tecnológico requiere de infraestructura adicional ya que es necesario disponer de un repositorio RDF dotado de un servicio SPARQL.

4.6.5 $\mathcal{SPM}_4$-*On-the-fly*

Al igual que en el método de publicación desde una base de datos, se trata de un método híbrido de generación y publicación de datos enlazados bajo demanda, $\mathcal{SPM}_4$. Las descripciones en RDF de los recursos no existen físicamente sino que se generan bajo petición. Resulta interesante cuando el publicador de datos es diferente del propietario de datos y también para ofrecer dinámicamente una capa de datos enlazados sobre servicios preexistentes, pese a que la generación y publicación se realiza bajo demanda, en ningún caso se obvian las tareas implicadas en estos procesos.

Este método es el utilizado en servicios como RDFohloh [118] para generar y publicar la información sobre los proyectos disponibles en el portal Ohloh [46]. Por ejemplo, ver Figura 4.12, se puede obtener la descripción del proyecto MOLDEAS en las siguientes URIs y bajo las mismas características que se definirían en otros métodos como la negociación de contenido, etc.:

- `http://rdfohloh.wikier.org/project/moldeas/html`
- `http://rdfohloh.wikier.org/project/moldeas/rdf`
- `http://rdfohloh.wikier.org/project/moldeas/n3`
- ...

4.6.6 $\mathcal{SPM}_5$-*Linked Data Frontend*

La evolución en los métodos de publicación derivada de la necesidad de cumplir con los criterios de *Linked Data* ha motivado la creación de herramientas que proporcionen soporte a los mismos. De esta manera, ha surgido tecnología tal como los *Linked Data Frontend*, $\mathcal{SPM}_5$, que simplifican el método de publicación realizando de forma automática muchas de las tareas que antes se realizaban manualmente y que sirven para cumplir con las características de publicación $\mathcal{P}$, como la negociación de contenido, el uso de HTTP URIs, etc. Se puede considerar la evolución natural de la reescritura de URLs, que se utilizaba previamente, la cual requería un profundo conocimiento de los módulos de reescritura de los servidores web.

No obstante, este método de publicación se basa en la existencia de un *endpoint* en SPARQL, sobre el que se puedan ejecutar consultas *DESCRIBE*. El funcionamiento del mismo consiste en la

```
<http://rdfohloh.wikier.org/project/moldeas/rdf> dct:isFormatOf
        <http://rdfohloh.wikier.org/project/moldeas>;
        a foaf:Document;
        rdfs:label "MOLDEAS's DOAP document serialized in RDF/XML";
        foaf:primaryTopic <http://rdfohloh.wikier.org/project/moldeas>.
        <http://rdfohloh.wikier.org/project/moldeas> dct:updated "2011-12-25T10:04:51
            Z";
        rdfohloh:ohloh-page <http://www.ohloh.net/projects/moldeas>;
        doap:created "2011-10-14T09:19:11Z";
        doap:description "This work aims to apply the semantic web and LOD approaches
             to public procurement notices...";
        doap:download-page <http://code.google.com/p/moldeas/downloads/list>;
        doap:homepage <http://purl.org/weso/moldeas/>;
        doap:name "MOLDEAS";
        doap:programming-language "JavaScript";
        a doap:Project;
        = <http://rdfohloh.wikier.org/project/586667>;
        skos:subject <http://dbpedia.org/resource/JavaScript>.
```

Figura 4.12: Ejemplo de generación *On-the-fly* de *Linked Data*.

extracción e identificación de la URI del recurso y teniendo en cuenta $\mathcal{P}$, generar una representación del mismo preguntando al repositorio RDF.

La principal ventaja de este método reside en la configuración del mismo ya que se puede realizar para cada uno de los *dataset* RDF disponibles en un repositorio, obteniendo de forma sencilla el acceso a los recursos RDF. Además, este método no resulta intrusivo con las herramientas de almacenaje de RDF, ya que son aplicaciones externas añadidas como una capa nueva de acceso para facilitar la publicación y el consumo.

Por ejemplo, utilizando Pubby y una configuración, ver Figura 4.13, se puede publicar la información del Nomenclátor de Asturias 2010 que previamente ha sido almacenada en un repositorio RDF.

4.6.7 $\mathcal{SPM}_6$-Servicio Web

Este método de publicación, $\mathcal{SPM}_6$, se puede considerar un híbrido entre los citados anteriormente, ya que en realidad un *Linked Data Frontend* es en esencia un servicio web dotado de ciertas características que permiten el acceso mediante URIs definidas a los contenidos que se generan y publican en un repositorio RDF. También un servicio web puede generar las descripciones RDF bajo demanda, o bien constituir un envoltorio sobre una petición a una base de datos obteniendo la representación de los datos en RDF. En general, este es el método de publicación más abierto, ya que la fuente de los datos en RDF puede ser variada: fichero estático, repositorio, bajo demanda, etc., y la publicación se puede basar en: generación bajo demanda, consultas, etc. Habitualmente se utilizarán servicios bajo el modelo REST por su complicidad respecto a la iniciativa *Linked Data*, también sería posible invocar servicios bajo el protocolo SOAP incluyendo las representaciones RDF en el cuerpo de la respuesta del servicio.

```
<> a conf:Configuration;
    conf:projectName "Nomen 2010 Asturias";
    conf:projectHomepage <http://purl.org/weso/nomenclator/>;
    conf:webBase <http://purl.org/weso/nomenclator/>;
    conf:usePrefixesFrom <>;
    conf:defaultLanguage "en";
    conf:indexResource
      <http://purl.org/weso/nomenclator/asturias/2010/resource/ds>;

conf:dataset [
        conf:sparqlEndpoint <http://156.35.31.156/sparql>;
        conf:sparqlDefaultGraph <http://purl.org/weso/nomenclator/asturias/2010>;
        conf:datasetBase <http://purl.org/weso/nomenclator/>;
        conf:datasetURIPattern "asturias/2010/resource/.*";
        conf:webResourcePrefix "";
        conf:fixUnescapedCharacters "(),'!$&*+;=@";
        conf:metadataTemplate "metadata.ttl";
        meta:pubbyUser <http://purl.org/weso>;
        meta:pubbyOperator <http://purl.org/weso>;
];
conf:dataset [
        conf:sparqlEndpoint <http://156.35.31.156/sparql>;
        conf:sparqlDefaultGraph <http://purl.org/weso/nomenclator/ontology>;
        conf:datasetBase <http://purl.org/weso/nomenclator/>;
        conf:datasetURIPattern "ontology/.*";
        conf:webResourcePrefix "";
        conf:fixUnescapedCharacters "(),'!$&*+;=@";
        ...
    ];

conf:dataset [
       conf:sparqlEndpoint <http://156.35.31.156/sparql>;
       conf:sparqlDefaultGraph <http://purl.org/weso/nomenclator/asturias/2010/stats
           >;
       conf:datasetBase <http://purl.org/weso/nomenclator/>;
       conf:datasetURIPattern "asturias/2010/stats/resource/.*";
       conf:webResourcePrefix "";
       conf:fixUnescapedCharacters "(),'!$&*+;=@";
        ....
 ];
.
```

Figura 4.13: Ejemplo de configuración de Pubby para el Nomenclátor de Asturias 2010.

4.6.8 Privacidad en la Publicación de *Linked Data*

El concepto de privacidad en los datos enlazados se pone de manifiesto en el momento en el que existe información y datos que dependiendo de ciertos factores deben ser filtrados, como por ejemplo el tipo de usuario. Desde una perspectiva conceptual la publicación de datos enlazados puede verse sometida a cuestiones de privacidad, pero no así refiriéndose a *Open Data* o *Linking Open Data*, en los cuales no se deberían someter los datos a ningún tipo de filtro con el objetivo de facilitar la transparencia. No obstante, es conveniente sopesar como parte de la publicación de datos la posibilidad de establecer vistas o filtros que puedan agregar la información, bien para su consumo o por cuestiones legales, actualmente, la forma de obtención de esta característica desde la iniciativa de *Linked Data* se puede realizar mediante la generación de vistas en RDF:

- Usar grafos nombrados, en los cuales sólo estén disponibles los datos que se deseen públicos, evidentemente, si todos los datos están publicados en un *endpoint* de SPARQL, sortear esta restricción no supone un gran inconveniente.
- Filtrar ciertos valores o relaciones en las consultas sobre los datos publicados.
- Definir usuarios, roles, listas de control de acceso sobre los datos publicados y los grafos.

Sin duda la privacidad no es concepto nuevo, ya que las bases de datos tradicionales incluyen el concepto de *vista* para ofrecer soporte a este tipo de característica. En cualquier caso, esta cuestión no ha sido abordada completamente desde el punto de vista de la ingeniería y como ya se ha comentado se emplean soluciones parciales reutilizando la tecnología semántica. La forma más coherente y notable de adoptar la privacidad como característica de la publicación de datos sería seguir un enfoque parecido a los *Linked Data Frontends*, sin embargo aún no se ha contemplado como un propósito estratégico.

4.7 Proceso de Consumo

Definición 4.7.1 (Consumo)**.** *Proceso que conlleva todas las tareas que implican el acceso a los recursos de un dataset RDF $\mathcal{D}_{pub}$.*

4.7.1 Método Semántico de Consumo de *Linked Data*

Definición 4.7.2 (Método Semántico de Consumo)**.** *SCM es una función que para un dataset RDF $\mathcal{D}_{pub}$, con unas características de publicación $\mathcal{P}$ y un conjunto de mapeos $\mathcal{M}^1$, obtiene como resultado la representación de los recursos $r_k \in \mathcal{D}_{pub}$ en otro modelo formal.*

$$SCM : \mathcal{D}_{pub} \times \mathcal{M}^1 \longrightarrow \mathcal{D}_{consum}$$

Las características que tiene esta definición de un *SCM* son las siguientes:

- $\mathcal{D}_{pub}$ no es modificado por los métodos de consumo.
- $\mathcal{P}$ indica como se accede al *dataset* RDF.
- $\mathcal{M}^1$ indica como transformar el *dataset* $\mathcal{D}_{pub}$ a la representación objetivo.

Esta definición de consumo implica por un lado, que los datos se pueden consumir directamente para realizar operaciones como navegación y representación gráfica o que se pueden utilizar para cargar valores dentro de objetos de un lenguaje de programación formando parte de la lógica de negocio de una aplicación.

4.7.2 SCM_1-Directo de datos en RDF

Este método de consumo, SCM_1 está ampliamente asentado para la creación de aplicaciones que simplemente realicen una visualización, navegación o creación de *mashups* con los datos publicados. Es la forma más sencilla de consumo y encaja impecablemente con el concepto de Web 2.0, en el cual se consumen distintas fuentes de datos, servicios o APIs de distintos proveedores para generar un servicio agregado en el cual poder manejar los datos. La lógica de negocio de estas aplicaciones suele ser relativamente simple ya que las operaciones a realizar con los datos no trascienden de la mera agregación e integración de varias fuentes de datos.

Actualmente es considerable la relevancia que están adquiriendo este tipo de aplicaciones, de ahí que la representación de RDF en JSON haya logrado una gran trascendencia con el objetivo de facilitar a los desarrolladores, procedentes desde ámbitos ajenos a la Web Semántica, la reutilización de datos enlazados.

Por otra parte, los sistemas de visualización de datos están ampliamente asentados y existen múltiples herramientas que permiten la presentación de los mismos. La tendencia actual en este tipo de consumo se centra en el enriquecimiento automático y bajo demanda de los datos disponibles: descubriendo nuevos *datasets*, enriqueciendo la información textual con elementos multimedia o georreferenciación, etc.

4.7.3 SCM_2-*Mapeo* a Lenguaje de Programación

Tradicionalmente el acceso a datos ha sido una de las partes clave en el desarrollo de aplicaciones con el objetivo de asegurar la persistencia de los objetos de negocio y recuperarlos de forma sencilla. El uso de sistemas de bases de datos relacionales requería una transformación del Modelo Entidad/Relación a un modelo orientado a objetos, estructurado, funcional o basado en prototipos dependiendo del lenguaje de programación y del paradigma utilizado, en el que mediante una serie de *mapeos* permitían "cargar" los objetos de negocio con valores procedentes de la base de datos. La diversidad de datos y la heterogeneidad en los modelos de representación ha comportado la proliferación de herramientas que facilitan el *mapeo* entre distintas entidades, por ejemplo el uso de ORMs, como Hibernate o Linq, en los lenguajes orientados a objetos está ampliamente asentado. En el caso que nos ocupa, SCM_2, surge una situación similar en la que se dispone de un modelo fuente para la representación de los datos, RDF, y un modelo objetivo que puede depender del paradigma en el que se esté programando. En este escenario se manifiesta la necesidad de disponer de un tipo de ORM para combinar RDF con los objetos de negocio en un lenguaje de programación, pero la realidad es que todavía no se ofrecen herramientas que permitan realizar este promoción de los datos enlazados a objetos de nuestra lógica de negocio de una forma sencilla. Los enfoques que se suelen seguir son los siguientes:

1. Programar manualmente los accesos a datos mediante consultas SPARQL o un API RDF para crear nuestros objetos de negocio con los datos previos en RDF.

2. *Mapeo* semi-automático indicando a las propiedades de nuestro modelo objetivo qué propiedades del *dataset* RDF se deben cargar.

3. Transformación automática de ontologías a un modelo de un lenguaje de programación y en consecuencia a partir de los recursos RDF que sean instancias de esa ontología crear instancias del modelo objetivo.

Este tipo de enfoques para el consumo de datos han sido estudiados en el proyecto TRIOO [116], sin embargo lo que ocurre verdaderamente es que se sigue abordando el problema desde un enfoque en gran medida manual, con el consecuente perjuicio para las aplicaciones que consuman datos enlazados.

4.8 Proceso de Validación

Definición 4.8.1 (Método Semántico de Validación)**.** *SVM es una función que para un dataset RDF $\mathcal{D}$ y un conjunto de características a validar $\mathcal{V}$, obtiene como resultado un conjunto de aserciones $\mathcal{A}$.*

$$SVM : \mathcal{D} \times \mathcal{V} \longrightarrow \mathcal{A}$$

Las características que tiene esta definición de un SPM son las siguientes:

- El *dataset* RDF $\mathcal{D}$ no es modificado durante el proceso de validación.
- El conjunto de características a validar $\mathcal{V}$ depende del proceso a contrastar.
- El conjunto de aserciones obtenidas tras el proceso de validación $\mathcal{A}$, indica el grado de cumplimiento de $\mathcal{D}$ de acuerdo a $\mathcal{V}$.

La validación de los datos enlazados es aplicable a todos los procesos del ciclo de vida, de esta manera, se aseguran tanto características de calidad como principios de *Linked Data*, *Open Data* y *Linking Open Data*, validación de enlaces, etc.

El proceso de validación se puede realizar de dos formas:

Manual. Mediante la revisión personal de las características deseables para cada uno de los recursos del *dataset* RDF $\mathcal{D}$.

Automática. Mediante el uso de herramientas existentes que permiten verificar ciertas características de los datos enlazados, por ejemplo el validador Vapour [33] o el presente para la publicación de *datasets* en la nube de *Linked Data*.

El uso de recursos "plantilla" o ejemplos en la metainformación de un *dataset* RDF ayuda a la realización de los procesos de validación ya que permite extraer ciertas características que han de cumplir todos los recursos pertenecientes a un *dataset* determinado. No obstante, la validación de datos enlazados todavía no ha tenido un impulso enérgico por parte de la comunidad, por lo que el método más utilizado es la validación manual. Esta situación tiene como principal ventaja que se promueve la publicación masiva de datos sin poseer grandes barreras de entrada, sin embargo presenta como inconveniente la obtención de un entorno de datos en el que la incertidumbre sobre la calidad de los mismos o su consumo carece de respaldo. La tendencia en este tipo de proceso lleva a la convergencia hacia iniciativas ya asentadas, como las de accesibilidad web en la cual se tienen una serie de perfiles con un grado de cumplimiento de reglas que permiten asegurar que las páginas web cumplen ciertos requisitos.

4.9 Proceso de Realimentación

Definición 4.9.1 (Realimentación)**.** *Proceso que conlleva todas las tareas que implican la actualización de datasets RDF preexistentes, se trata de una agregación de los procesos anteriores.*

Definición 4.9.2 (Método Semántico de Realimentación)**.** *SFM es una función que para un dataset RDF $\mathcal{D}$ publicado como $\mathcal{D}_{pub}$ y un conjunto de relaciones y valores a modificar $\mathcal{RV}$, obtiene como resultado un nuevo dataset RDF $\mathcal{D}'$ que se publica como $\mathcal{D}'_{pub}$.*

$$SFM : \mathcal{D}_{pub} \times \mathcal{RV} \longrightarrow \mathcal{D}'$$

Las características que tiene esta definición de un SFM son las siguientes:

- El *dataset* RDF publicado $\mathcal{D}'$, se considera un nuevo conjunto de datos diferente a $\mathcal{D}_{pub}$.
- El *dataset* RDF $\mathcal{D}$, sólo se modifica en la parte de datos que se haya publicado como $\mathcal{D}_{pub}$. En la mayoría de los casos $\mathcal{D}_{pub} \cong \mathcal{D}$, por lo que la definición se puede extender fácilmente a $\mathcal{D}$.
- El conjunto de relaciones y valores a modificar $\mathcal{RV}$ es un conjunto de tripletas de la forma $< r_k, p, v >$, donde r_k es un recurso del *dataset* $\mathcal{D}_{pub}$, p es una relación presente en el recurso r_k y v es el valor de la relación p en r_k. Además de actualizar relaciones y valores existentes también puede implicar la adición o borrado de una tripleta $< r_k, p, v >$.
- El cambio en los valores de las tripletas de un recurso puede implicar cambios en el modelo formal $\mathcal{O}$ de los recursos.
- Esta definición puede exigir cambios en el proceso de generación y publicación ya que estas actualizaciones deberán ser almacenadas en función del método utilizado.

El método de realimentación se interpreta como una composición de los métodos producción, consumo y publicación, ya que tanto si se modifican tripletas existentes (en cuanto a valores) como si se añaden o borran tripletas de los recursos, puede ser necesario ejecutar los métodos anteriores de forma completa, se asume que:

- $\mathcal{M}$ es el conjunto de *mapeos* para la producción de un *dataset*.
- $\mathcal{P}$ es el conjunto de características de publicación.
- $\mathcal{M}^1$ es el conjunto de *mapeos* para el consumo de datos.
- $\mathcal{M}^1 = \mathcal{RV}$.
- $\mathcal{D}'_{pub}$ es el *dataset* RDF $\mathcal{D}'$ publicado, tras aplicar el método de realimentación.

Por tanto, la composición de los métodos ya definidos de consumo (SCM), producción (SPM) y publicación ($\mathcal{SPM}$) deben generar un nuevo *dataset* RDF publicado como $\mathcal{D}'_{pub}$.

$$\mathcal{SPM} \circ SPM \circ SCM(\mathcal{D}_{pub}, \mathcal{RV}) \equiv \mathcal{D}'_{pub} \tag{4.1}$$

$$\mathcal{SPM} \circ SPM \circ SCM(\mathcal{D}_{pub}, \mathcal{M}^1) \equiv \mathcal{D}'_{pub} \tag{4.2}$$

$$\mathcal{SPM} \circ SPM(\mathcal{D}_{consum}, \mathcal{M}) = \mathcal{D}'_{pub} \tag{4.3}$$

$$\mathcal{SPM}(\mathcal{D}', \mathcal{P}) = \mathcal{D}'_{pub} \tag{4.4}$$

$$\mathcal{D}'_{pub} = \mathcal{D}'_{pub} \tag{4.5}$$

Así, atendiendo a las definiciones de los métodos semánticos realizadas, el proceso de realimentación se puede formular como una composición de los procesos de consumo, producción y publicación.

$$SPM \circ SPM \circ SCM(\mathcal{D}_{pub}, \mathcal{RV}) \equiv SFM(\mathcal{D}_{pub}, \mathcal{RV}):$$

4.9.1 Lenguaje de Actualización

Si el método de publicación utilizado es un *endpoint* de SPARQL sobre un repositorio RDF con capacidades para SPARUL o SPARQL 1.1, entonces se pueden actualizar tripletas y recursos RDF mediante el propio lenguaje de consulta SPARQL con las nuevas características de actualización. Este método resulta sin duda el más apropiado, ya que es declarativo y se puede utilizar toda la potencia del lenguaje de consulta para seleccionar y actualizar aquellos recursos que se obtengan como respuesta.

Por ejemplo utilizando la consola de Virtuoso de OpenLink se pueden ejecutar este tipo de sentencias, ver Figura 4.14, (borrar e insertar una tripleta RDF en un grafo nombrado):

```
SPARQL DELETE FROM GRAPH <http://purl.org/weso/pscs/isic/v4> { ?s void:dataDump  ?o }
  FROM <http://purl.org/weso/pscs/isic/v4> WHERE {
      ?s  void:dataDump  ?o
    };

SPARQL INSERT INTO GRAPH <http://purl.org/weso/pscs/isic/v4> {
  <http://purl.org/weso/pscs/isic/v4/resource/ds> void:dataDump
    <http://purl.org/weso/datasets/pscs/isic/v4/isic-v4.ttl>
  };
```

Figura 4.14: Ejemplo de uso de un lenguaje de actualización .

4.9.2 Descubrimiento Automático

Un método de actualización se puede implementar a través de un proceso continuo en el que una vez actualizados los datos se ejecute la combinación de los métodos de consumo, producción y publicación. En este caso, la actualización es un *workflow* automático que puede ser inducido por cambios en una base de datos o por manifestarse nuevas relaciones en *datasets* que se hayan publicado posteriormente. En el primero de los casos, la estructura y definición de los recursos no debería variar, ya que se trata de una nueva remesa de datos pero con la misma estructura, rangos y dominios para las propiedades y valores. Por ejemplo en el caso de información de sensores se actualiza constantemente, pero los valores tomados y su paso por el proceso de producción y publicación es siempre el mismo. En el segundo caso tras estimar o descubrir la posibilidad de alterar el *dataset* publicado se añaden nuevas propiedades o valores, lo que puede implicar la necesidad de cambiar el modelo formal para que sea consecuente con los valores representados. Un ejemplo de este segundo caso, lo constituye la adición de referencias a otros *datasets* de forma automática utilizando las propiedades léxicas de los recursos previos.

4.9.3 Actualización Ocasional

En este caso se corrigen valores del *dataset* RDF publicado en recursos concretos y que no se hayan detectado durante las fases de validación, es un escenario de refinamiento por el uso de los datos con el objetivo de verificar su calidad y validez. Por ejemplo tratando con valores a nivel internacional para publicar datos económicos, puede ser que el separador utilizado para los "miles" no esté unificado en todos los casos, el hallazgo de un recurso con un valor "extraño" da lugar a su cambio sin alterar otros datos y recursos presentes en el *dataset*.

4.9.4 Actualización Incremental

Este método de actualización surge por la necesidad de un refinamiento progresivo al manifestarse nuevas necesidades en la descripción de los recursos, para que tengan en cuenta más información. La forma de proceder puede ser automática, ver Sección 4.9.2, o bien manual añadiendo más datos en un determinado tipo de recurso.

4.9.5 Usuarios y Aplicaciones

En la mayoría de los casos la actualización y cambios en los datos enlazados son controlados por el *Desarrollador* o el *Propietario de datos*. Dependiendo de la naturaleza, licencia de uso y nivel crítico de los mismos, se puede habilitar la edición para terceros como usuarios y aplicaciones. Actualmente este enfoque no se ha extendido ya que en definitiva requiere la validación por parte del *Propietario de datos*, pero contribuiría enormemente a realimentar los conjuntos de datos publicados. Un enfoque similar consiste en consultar a los usuarios finales qué datos les gustaría tener disponibles, pero en ningún caso se permite una especie de *crowdsourcing* de datos enlazados. También en aquellos enfoques en los que se genera RDF bajo demanda transformando datos de un sitio o servicio web, se podrían actualizar los datos de la fuente, por ejemplo en OhLoh disponiendo de los permisos de usuario adecuados se podría modificar la información que figura posteriormente en la versión RDF.

4.10 Tablas de Validación

El objetivo de esta sección es definir una serie de características en el ámbito de *Linked Data*, *Open Data* y *Linking Open Data* que sirvan para verificar que la aplicación de los métodos semánticos es correcta y con ello obtener datos enlazados que aporten las ventajas conocidas sobre el uso de esta iniciativa. De esta manera, si se consigue asegurar que los datos producidos mediante estos métodos semánticos concuerdan con las directrices marcadas, se puede confirmar la calidad en los datos, impulsar la reutilización de los mismos y mejorar, consecuentemente, el acceso a la información que conllevan. Una vez que se apliquen estos métodos a un determinado dominio, como el de las licitaciones públicas, se disponen de estas tablas como método de validación.

La elaboración de estas tablas surge de la agregación de varias fuentes:

1. Principios de *Linked Data* [23].
2. Principios de *Open Data* [205] .
3. Especificaciones y documentos del W3C como [20, 35, 173].
4. *Linked Data Design Considerations* del libro *Linked Data: Evolving the Web into a Global Data Space* [157].

5. *Linked Data Patterns* [105].
6. "*Basic Profile Resources*" [234] de IBM.
7. Principios para incluir *datasets* en la nube de datos enlazados y añadir al registro de CKAN [86].
8. Documentación específica, ver Sección 3.2.6.12.
9. Experiencia adquirida.

La forma de modelado de estas tablas es la de un cuestionario, de manera que la evaluación de una característica puede tener las siguientes respuestas:

Positiva $\oplus$**.** La característica es aplicable y se ha realizado en el actual *dataset*.

Negativa $\ominus$**.** La característica es aplicable y no se ha realizado en el actual *dataset*.

No aplicable $\odot$**.** La característica no es aplicable y no se ha realizado en el actual *dataset*.

Las tablas que se listan a continuación y cuya descripción completa se haya en el Apéndice C, permiten verificar los procesos, métodos semánticos, principios, patrones de diseño y características para difundir los *datasets* RDF:

1. La Tabla C.1, con identificador T^1, realiza una validación de las características, en total 69, propias de los procesos de producción y publicación. La validación del proceso de consumo vendrá determinada por la viabilidad de los dos anteriores. De la misma forma el proceso de realimentación como composición de los anteriores se puede verificar mediante las cuestiones planteadas en esta tabla.
2. La Tabla C.2, con identificador T^2, señala los patrones de diseño aplicados, en total 44, utilizados para el modelado de datos.
3. Las Tablas C.3 y C.4, con identificadores T^3 y T^3_1, indican el grado de cumplimiento de los principios de *Linked Data* y del Modelo de 5 $\star$, en total $4 + 5$.
4. Las Tabla C.5 y C.6, con identificadores T^4 y T^4_1, indican el grado de cumplimiento de los principios de *Open Data*, en total $8 + 14$.
5. La Tabla C.7, con identificador T^5, indica el grado de cumplimiento de los principios, en total 5, para formar parte del diagrama "Linking Open Data Cloud".
6. La Tabla C.8, con identificador T^6, indica el grado de cumplimiento de los principios, en total 47, para formar parte de un registro de *datasets* basado en CKAN.

Obviamente, el cumplimiento de ciertas preguntas, principios o directrices conlleva el de otros presentes en las diferentes tablas, no obstante, es conveniente separar estas características para efectuar correctamente la evaluación del nivel en el que se encuentra un *dataset*.

Capítulo 5

Métodos Semánticos en el ámbito de las Licitaciones Públicas

> Aunque nuestra visión hacia adelante es muy corta, podemos darnos cuenta de que hay mucho por hacer.
>
> *Citas Célebres*
> ALAN TURING

5.1 Anuncios de Licitación

El dominio de la contratación pública electrónica comprende un conjunto de fases bien diferenciadas como ha quedado patente en el repaso realizado durante el Capítulo 2, de todas las fases identificadas y teniendo en cuenta las necesidades de interoperabilidad e integración y en general de comunicación entre los distintos agentes implicados en este proceso administrativo, cabe destacar los procesos particulares de *eAccess* y *eNotification* en los cuales resulta especialmente relevante la publicación de información de forma estandarizada para impulsar la participación de cualquier tipo de organización, creando un mercado competitivo de oportunidades a lo largo de toda una comunidad como es la Unión Europea. Si bien el caso de estudio de este documento se centra en los anuncios de licitación públicos a nivel europeo, el modelo propuesto es aplicable a cualquier otro entorno en el que las necesidades de compartir información sean un factor clave para el impulso de un determinado sector de negocio.

Por otra parte, tal y como se ha señalado en la Sección 1.1.4 existen una serie de problemas inherentes al proceso de contratación pública que no han sido abordados completamente o se han intensificado con el traspaso de la actividad a un entorno electrónico. En este sentido, la dispersión de la información, múltiples fuentes de datos, la heterogeneidad de los formatos tanto en publicación como en explotación y el multilingüismo unido a la multiculturalidad, generan un entorno en el cual, evidentemente, el uso de semántica y la aplicación de los principios de *Linked Data* pueden mejorar ostensiblemente su comportamiento, confiriendo el carácter transversal necesario a la información y datos que son utilizados y publicados bajo este proceso administrativo.

El esfuerzo desarrollado desde la gran mayoría de las instituciones públicas con intereses en impulsar la contratación pública por medios electrónicos ha generado la aparición de múltiples proyectos, plataformas, especificaciones, etc., que si bien han abordado las necesidades de este entorno,

también han generado un maremágnum de modelos, subprocesos, pasos y etapas, etc., perjudicando en cierta medida las intenciones iniciales y absorbiendo gran parte del esfuerzo y del tiempo en la realización de acciones integradoras en lugar de generar servicios de valor añadido para los agentes implicados. De forma sintética esta introducción guía el proceso de aplicación del ciclo de vida de datos enlazados a los anuncios de licitación públicos.

A lo largo de este capítulo se utilizará un ejemplo real (no se incluyen todos los datos por su extensión) de generación de datos enlazados de un contrato público cuyas características son las siguientes:

- Anuncio de Gestión de Infraestructuras Sanitarias del Principado de Asturias S.A.U. (GISPASA) para la licitación del suministro, e instalación de camas y cunas y mesillas para el nuevo Hospital Universitario Central de Asturias.
- Objeto del contrato:
  1. Tipo: Suministro.
  2. Descripción: Adquisición de camas, cunas y mesillas, sus condiciones de suministro, e instalación de dicho mobiliario para el nuevo Hospital Universitario Central de Asturias.
  3. División por lotes y número de lotes/Número de unidades: Sí, 4 lotes. Cada uno de los lotes se ejecuta como un contrato separado en el cual las empresas pueden concursar, siendo el objeto del contrato, la cuantía y los plazos diferentes.
  4. CPV (Referencia de Nomenclatura): 33192000.- mobiliario para uso médico; 33192100.- camas para uso médico; 3914316.- cunas; 39143310.- mesitas bajas.
  5. Valor estimado del contrato: 3371282, 99 €.
  6. Lugar de ejecución/entrega: Carretera del Rubín, s/n. Finca "La Cadellada", Oviedo, Asturias – 33011.
  7. ...
- Fecha de publicación: 21 de enero de 2011 en el DOUE y 28 de enero de 2011 en el BOE.

5.1.1 Proceso de Producción de *Linked Data* de Anuncios de Licitación

Atendiendo a la definición realizada en la Sección 4.5, este proceso implica aquellas tareas que a partir de un *dataset* de entrada $\mathcal{G}$ y un conjunto de reglas de *mapeo* $\mathcal{M}$ obtiene un *dataset* RDF $\mathcal{D}$ como resultado. Para la realización del proceso de producción y como quedará patente en las siguientes secciones se ha utilizado un programa Java desarrollado en particular para dar cobertura a esta transformación debido al tamaño y extensión de los datos disponibles, en total de un millón de anuncios de licitación.

5.1.1.1 Tarea t_1-Análisis del *dataset* a transformar

La casuística del proceso de contratación pública electrónica ha conllevado la realización de múltiples modelos de datos capaces de recoger y representar la información implicada en este proceso administrativo. De acuerdo a las especificaciones desarrolladas en CODICE o en opXML y según los modelos desarrollados en las ontologías propuestas por el proyecto LOTED y por la *Charles University* de la República Checa, queda reflejado que la cobertura total de este proceso es verdaderamente complicada, por lo que es preferible la construcción de modelos de un tamaño sostenible y que dispongan de una gran capacidad de extensión e integración con otros ya existentes. Esta tendencia en el modelado de ontologías se ha considerado debido a que los intentos de realizar grandes

especificaciones formales de distintos dominios no ha obtenido los beneficios deseados, bien por su incapacidad para realmente representar todo el conocimiento o bien por el coste de realización de ciertos procesos automáticos como el razonamiento. Es por ello, que una buena práctica reside en la partición de conocimiento, ofreciendo así modelos sencillos, representativos, flexibles e intrínsecamente extensibles.

En cuanto a las reglas de negocio o conocimiento que se encuentran en un modelo o especificación de la información de los datos de los anuncios de licitación, cabe destacar los siguientes:

- En un proceso de licitación se puede publicar un lote de anuncios de licitación de distinto tipo.
- Un anuncio de licitación está identificado de forma única y se publica en, al menos, una fuente.
- La publicación de un anuncio de licitación se realiza en, al menos, un idioma.
- Un anuncio de licitación puede tener un anuncio previo, una adjudicación provisional o definitiva.
- Un anuncio previo se identifica a través de un código y una fuente, además dispone de una fecha prevista de licitación, información sobre los lotes, información sobre el tipo de proceso (restringido, emergencia, central de contratación, etc.) para el cual se adjuntan una serie de documentos o pliegos.
- El anuncio de licitación está compuesto por, al menos, un lote que dispone de la información de compra, es decir la adquisición que se realiza en el mismo. De nuevo tiene asociado un proceso de licitación, al que se adjuntan una serie de documentos o pliegos.
- La notificación provisional de adjudicación es relativa a un lote y en consecuencia, a una adquisición de compra dentro de un determinado proceso para el que existen documentos o pliegos. Evidentemente, consta de una serie de resoluciones, entre ellas la provisional y la definitiva, aparece además la figura de las organizaciones implicadas.
- La notificación definitiva de adjudicación posee los mismos elementos que la anterior pero realizada de forma firme.
- El proceso de contratación pública puede constar de un anuncio previo, un anuncio de licitación y las notificaciones provisional (opcional) y definitiva, para cada uno de ellos existen una serie de documentos asociados. Como metainformación se extrae: la descripción, fechas límites y eventos (localización, fecha inicio, fecha fin y persona de contacto).
- Un lote dentro de anuncio de licitación tiene una adquisición de compra de un producto o servicio y una serie de resoluciones.
- La adquisición está relacionada con un lote, sus notificaciones y anuncios, además de suministrar información sobre el tipo, título y descripción de compra, lista de códigos de adquisición, lista de items de adquisición (tipo y cantidad), importe y localización.
- La resolución de un lote tiene asociada una descripción, resultado y fecha, además de datos de ofertas (ítems y cantidades) con un importe determinado en un moneda específica.
- El importe de una adquisición tiene una cantidad total con impuestos y otra incluyendo impuestos. Además, la cuantía se expresa en una moneda determinada.
- La localización de cualquier anuncio, lote, etc., consta de una dirección postal, una localidad, un país y un idioma. El idioma puede estar asociado al país o a la región y descendiendo hasta el nivel de Comunidad de Autónoma (en el caso español).

- Una organización implicada en un proceso de licitación debe proveer información postal y una persona de contacto, especificando entre otros cargo, teléfono, correo postal, nombre, etc.
- Cualquier documento en todo el proceso de licitación tiene un título, una fecha y una descripción en un determinado idioma, estando asociado a un pliego oficial.
- Una fuente de publicación es aquella que se identifica por un código, un nombre, una descripción y una URL, además pertenece a una organización que realiza publicaciones en un determinado idioma.

El universo de discurso expuesto en la lista anterior guía el modelado posterior de la base de conocimiento y de los datos, ya que debe cubrir las entidades identificadas para que toda la meta-información presente en los anuncios de licitación se refleje en las entidades del dominio. De esta manera, se ha diseñado un *framework* general en el cual además de representar las propiedades intensionales de los documentos relativos al proceso de licitación también se indican otras entidades para su posterior ampliación. Con el objetivo de destacar las principales entidades y propiedades que guiarán el ciclo de vida de datos enlazados para los anuncios de licitación, se presentan las siguientes figuras:

- Figura 5.1, en la cual se realiza una descripción de alto nivel de las entidades participantes en un proceso de contratación pública.
- Figura 5.2, en la cual se detallan las entidades y algunas propiedades importantes de acuerdo al anterior diagrama y que se consideran relevantes desde el punto de vista de los anuncios de licitación.
- Figura 5.3, en la cual se detallan las entidades y las relaciones más importantes en un contrato público, teniendo en cuenta los distintos lotes.
- Figura 5.5, en la cual se presenta parcialmente la jerarquía de los órganos de contratación.

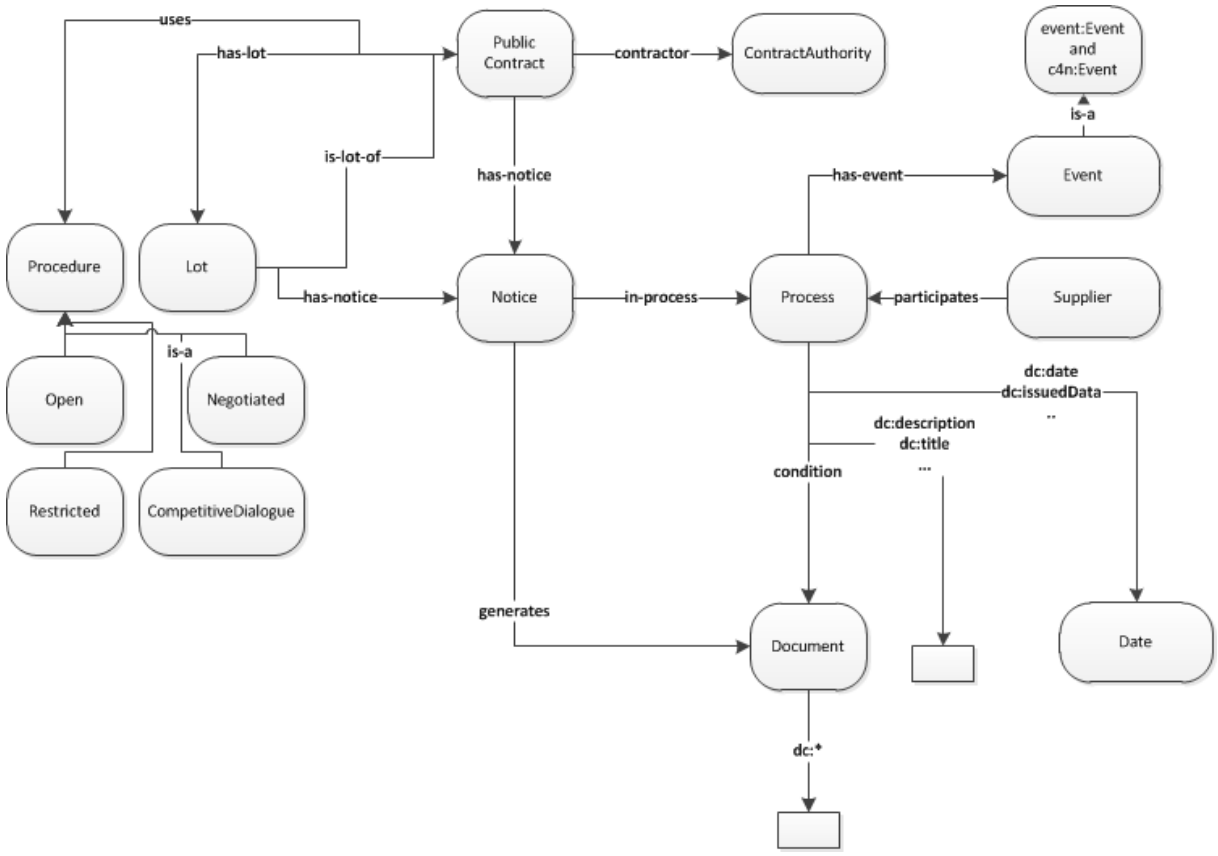


Figura 5.1: Modelo General de Entidades para el proceso de contratación pública electrónica (I).

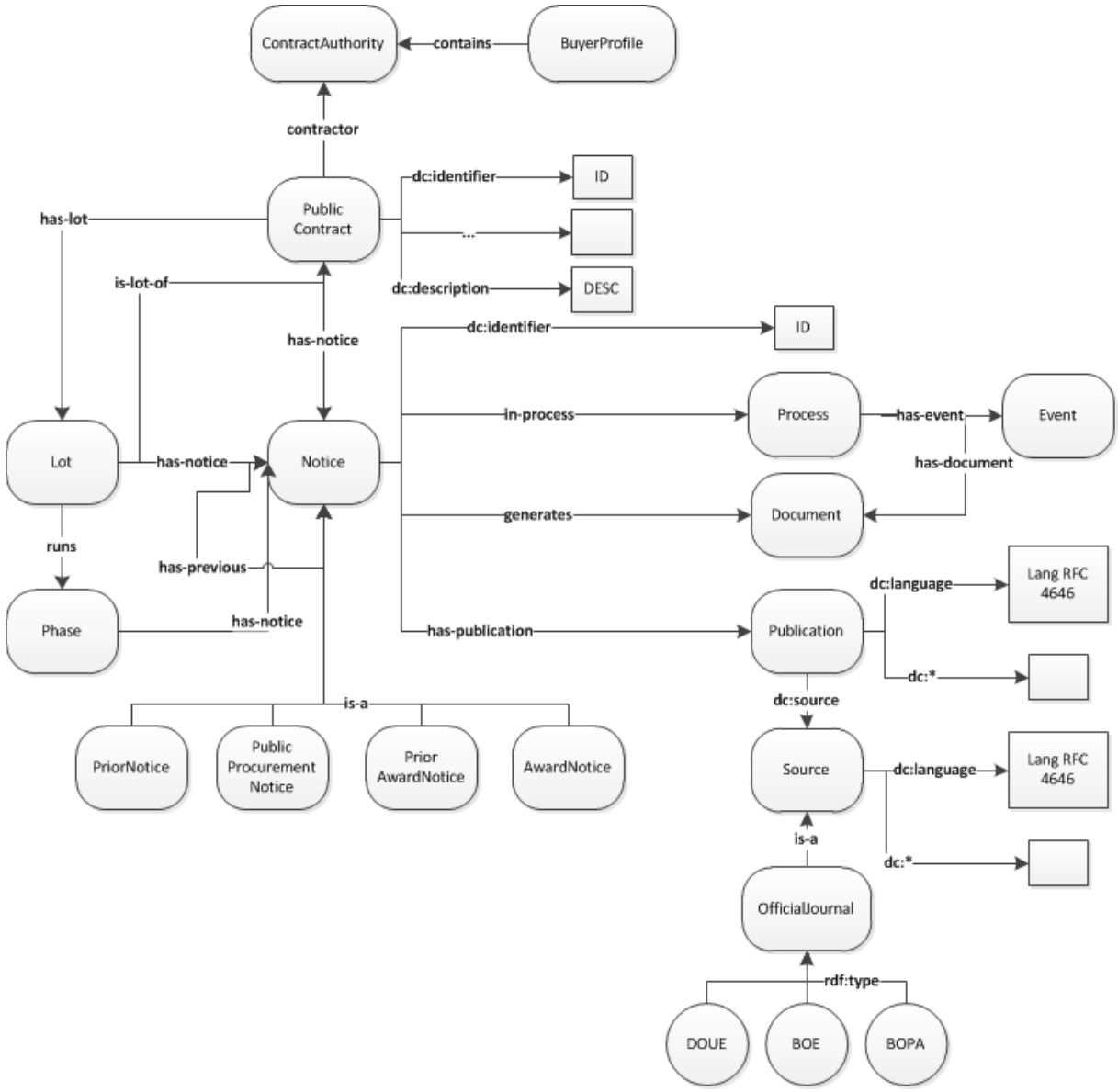


Figura 5.2: Modelo General de Entidades para el proceso de contratación pública electrónica (II).

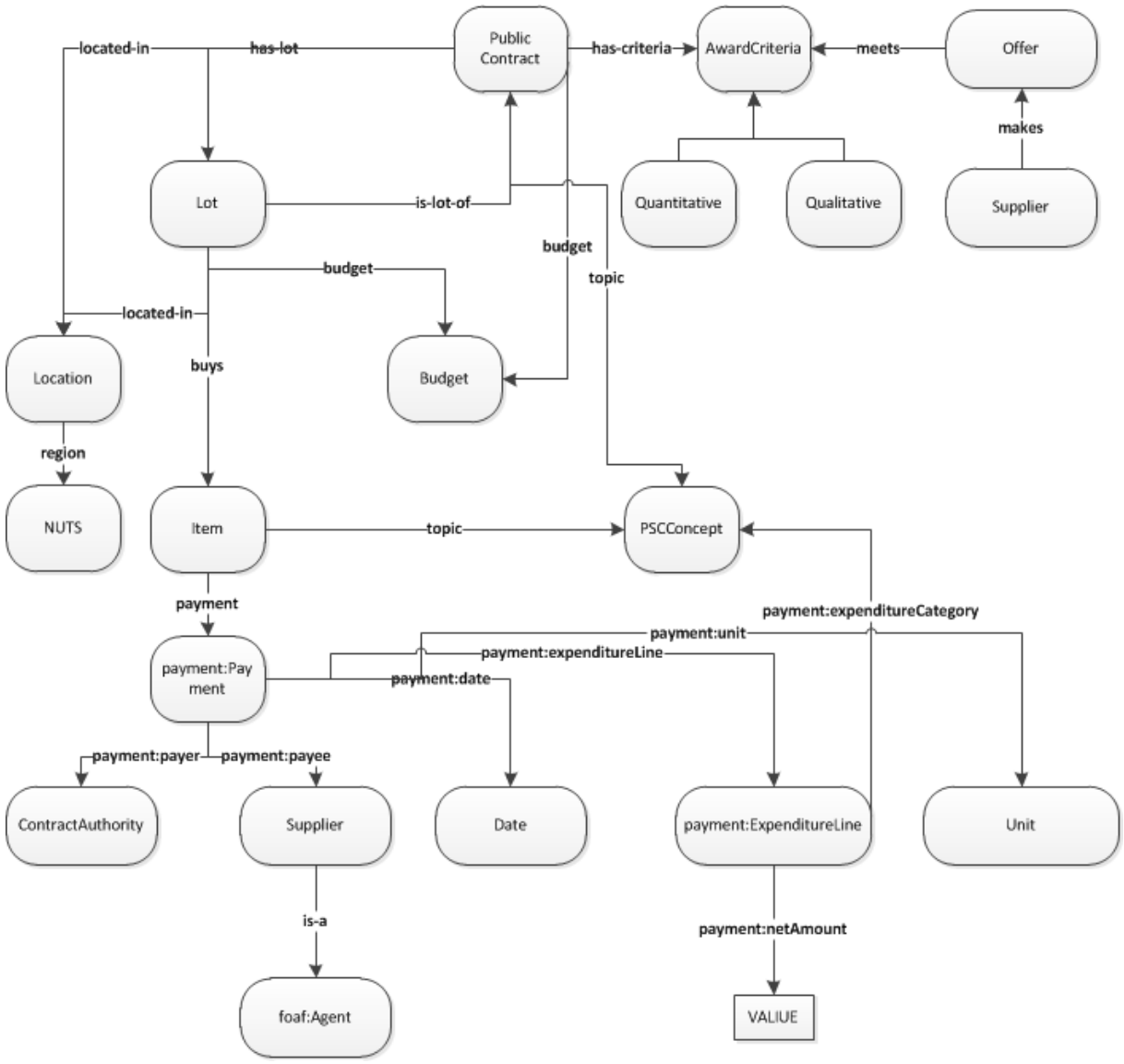


Figura 5.3: Relación entre los lotes en un contrato público.

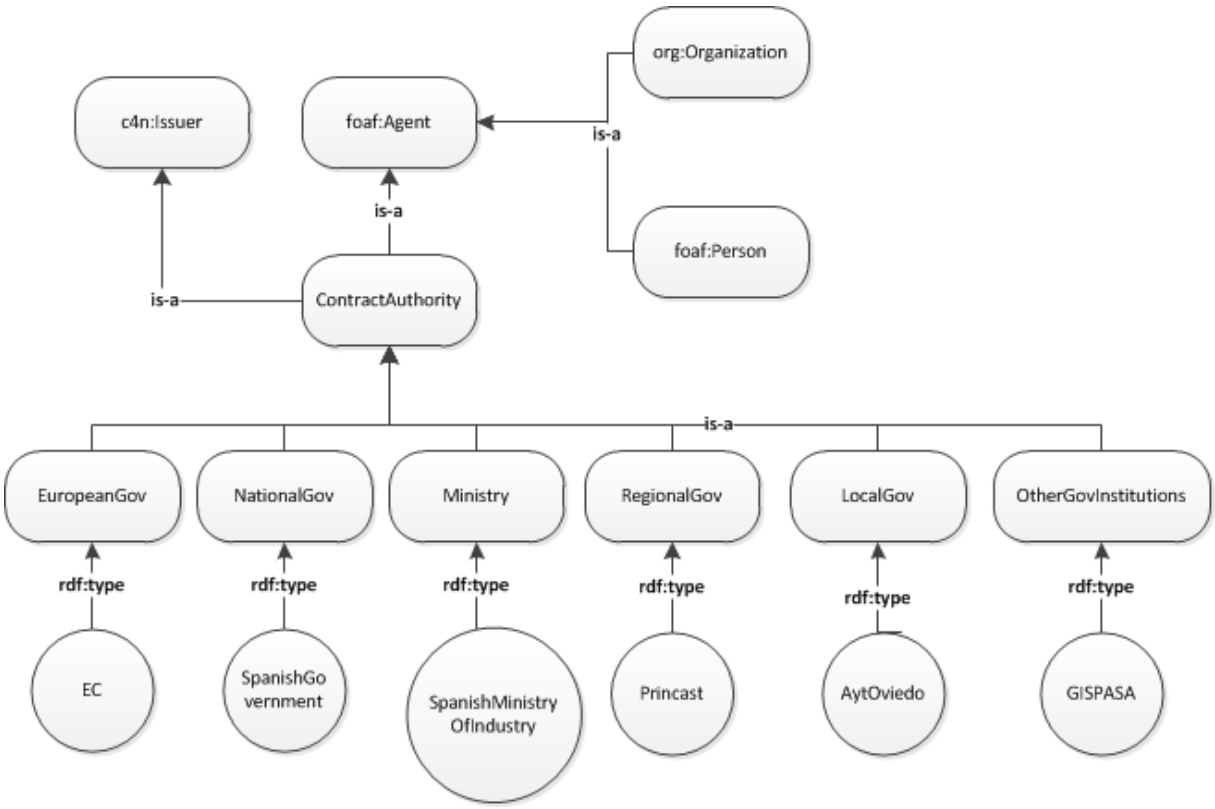


Figura 5.4: Modelo de Órganos de Contratación.

5.1.1.2 Tarea t_2-Limpieza de datos

Las fuentes de datos de anuncios de licitación son variadas y están disponibles en distintos formatos, en este trabajo y debido a la disponibilidad de datos de anuncios de licitación ya filtrados, facilitados por la empresa Euroalert.net, se ha decido utilizar este *dataset* de entrada, en formato CSV, en el cual la información ya ha sido recogida e integrada, evitando así la necesidad de recuperar información de distintas fuentes y de ejecutar tareas propias de la limpieza de datos. Aunque tratándose de fuentes públicas de datos se pudiera pensar que la información oficial ya debería estar disponible de forma homogénea, la realidad es que la diversidad de formatos provoca cierto desorden, en el que el esfuerzo para limpiar los datos consume gran parte del tiempo. No obstante, el sistema MOLDEAS se ha diseñado mediante un conjunto de adaptadores de tal forma que la fuente de datos es independiente pudiendo generar representaciones semánticas para la mayoría de las fuentes de datos habituales a nivel nacional y europeo.

5.1.1.3 Tarea t_3-Selección de Vocabularios

Los vocabularios seleccionados para modelar los anuncios de licitación teniendo en cuenta el análisis realizado en la tarea t_1 se presentan en la Tabla 5.1. En general, se trata de vocabularios que atienden a los siguientes criterios:

1. Formalización de una estructura taxonómica, como RDFS, SKOS u OWL.
2. Realización de *mapeos* entre conceptos, como SKOS y OWL.
3. Representación de tipos de datos, como XML Schema.
4. Gestión de información multilingüe, como SKOS-XL y RDFS.
5. Adición de metadatos y *provenance*, como Dublin Core, voID y *Provenance Ontology*.
6. Representación del tiempo e intervalos, como `Time Ontology` del W3C.
7. Representación de pagos, *Payments Ontology* del Gobierno del Reino Unido.
8. Llamadas a participación, *Call For Anything Ontology* de DERI, y eventos, *LODE Ontology*.

Prefijo	**Vocabulario**	**Fuente**	**Uso**
c4n	`http://vocab.deri.ie/c4n#`	*Michael Hausenblas*	Especificación de llamadas para participar en eventos.
dbpedia	`http://dbpedia.org/ontology/`	Comunidad *Linked Data*.	Reutilización de definiciones.
dc	`http://purl.org/dc/elements/1.1/`	*Dublin Core Metadata Initiative*	Creación de metadatos para los documentos.
dct	`http://dublincore.org/documents/dcmi-terms/`	≡	≡
event	`http://linkedevents.org/ontology/`	*Ryan Shaw*	*An ontology for Linking Open Descriptions of Event*.
foaf	`http://xmlns.com/foaf/0.1/`	Comunidad de Web Semántica.	Especificación de relaciones entre personas.

Prefijo	Vocabulario	Fuente	Uso
gr	http://purl.org/goodrelations/v1#	*Martin Heep*	Reutilización de definiciones para describir productos y servicios.
geo	http://www.w3.org/2003/01/geo/wgs84_pos#	W3C	Reutilización de elementos geográficos.
org	http://www.w3.org/ns/org#	Epimorphics Ltd.	Descripción de organizaciones.
owl	http://www.w3.org/2002/07/owl#	W3C	Realización de definiciones en el dominio.
po	http://www.productontology.org/	Martin Heep	Reutilización de datos provenientes de Productontology.
prov	http://purl.org/twc/ontology/w3c/prov#	W3C	Especificación de metadatos de procedencia.
payment	http://reference.data.gov.uk/def/payment#	Gobierno de Reino Unido	Especificación de pagos.
pscs	http://purl.org/weso/pscs/ontology/	Parte del estudio de esta tesis.	Ontología para las definiciones de las PSCs.
nuts	http://nuts.psi.enakting.org/def/	Universidad de Southampton	Especificación de las regiones europeas.
skos	http://www.w3.org/2004/02/skos/core#	W3C	Especificación de taxonomías.
skosxl	http://www.w3.org/2008/05/skos-xl#>	W3C	Representación de información lingüística.
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#	W3C	Descripción de recursos.
rdfs	http://www.w3.org/2000/01/rdf-schema#	W3C	Descripción de recursos con relaciones lógicas.
time	http://www.w3.org/2006/time#	W3C	Especificación de intervalos de tiempo y fechas.
time-entry	http://www.w3.org/2006/time-entry#	W3C	≡
intervals	http://reference.data.gov.uk/def/intervals/	Gobierno de Reino Unido	≡
vcard	http://www.w3.org/2006/vcard/ns#	W3C	Representación de información de contacto.
void	http://rdfs.org/ns/void#	Deri y W3C	Descripción de metadatos de un *dataset*.
xml	http://www.w3.org/XML/1998/namespace	W3C	Reutilización de definiciones.

Prefijo	Vocabulario	Fuente	Uso
xsd	http://www.w3.org/2001/XMLSchema#	W3C	Especificación de tipos de datos.

Tabla 5.1: Selección de Vocabularios para los Anuncios de Licitación.

5.1.1.4 Tarea t_4-Selección de otros *datasets* RDF

En este caso el principal *dataset* RDF a reutilizar es el relativo a la información geográfica de países europeos denominado NUTS y cuyo uso se detalla en la Sección 5.3. En general, en el proceso de producción de datos enlazados de los anuncios de licitación, clasificaciones de productos y organizaciones se reutilizan vocabularios y *datasets* similares, entre los cuales cabe destacar los presentes en la Figura 5.5.

5.1.1.5 Tarea t_5-Modelado de datos en RDF

El resultado de esta tarea debe ser una ontología de dominio $\mathcal{O}$, que modele la información de los anuncios de licitación, r_{ppn}, pertenecientes a un *dataset* RDF $\mathcal{D}$. Por lo tanto, se debe suministrar una ontología $\mathcal{O}$ de acuerdo al análisis realizado en la Sección 5.1.1.1, facilitando así la descripción formal de los datos de los recursos RDF:

- $\mathcal{C} = \{$*PublicContract, ContractAuthority, BuyerProfile, Lot, Item, Budget, Notice, PriorNotice, PublicProcurementNotice, PriorAwardNotice, AwardNotice, Publication, Source, Procedure, Open, Negotiated, Restricted, CompetitiveDialogue, Process, Event, Supplier,..., Process, Document, AwardCriteria, Offer, Location, Phase* $\}$
- $\mathcal{R} = \{$*rdf:type, rdfs:subClassOf, rdfs:label, rdfs:comment, dc:identifier, dc:language, dc:description, dc:source, dc:subject, dc:date, dc:*, foaf:topic, ..., contractor, has-lot, is-lot-of, runs, has-notice, has-publication, has-previous, in-process, generates, has-event, has-document, located-in, budget, buys, payment, topic, meets, has-criteria, makes*$\}$
- $\mathcal{I} = \{r_{ppn}\}$
- $\mathcal{A} = \{\star\}$, debido a su extensión se ha considerado la opción de realizar su representación gráfica en el diagrama de la ontología y a través de las reglas de negocio definidas previamente.

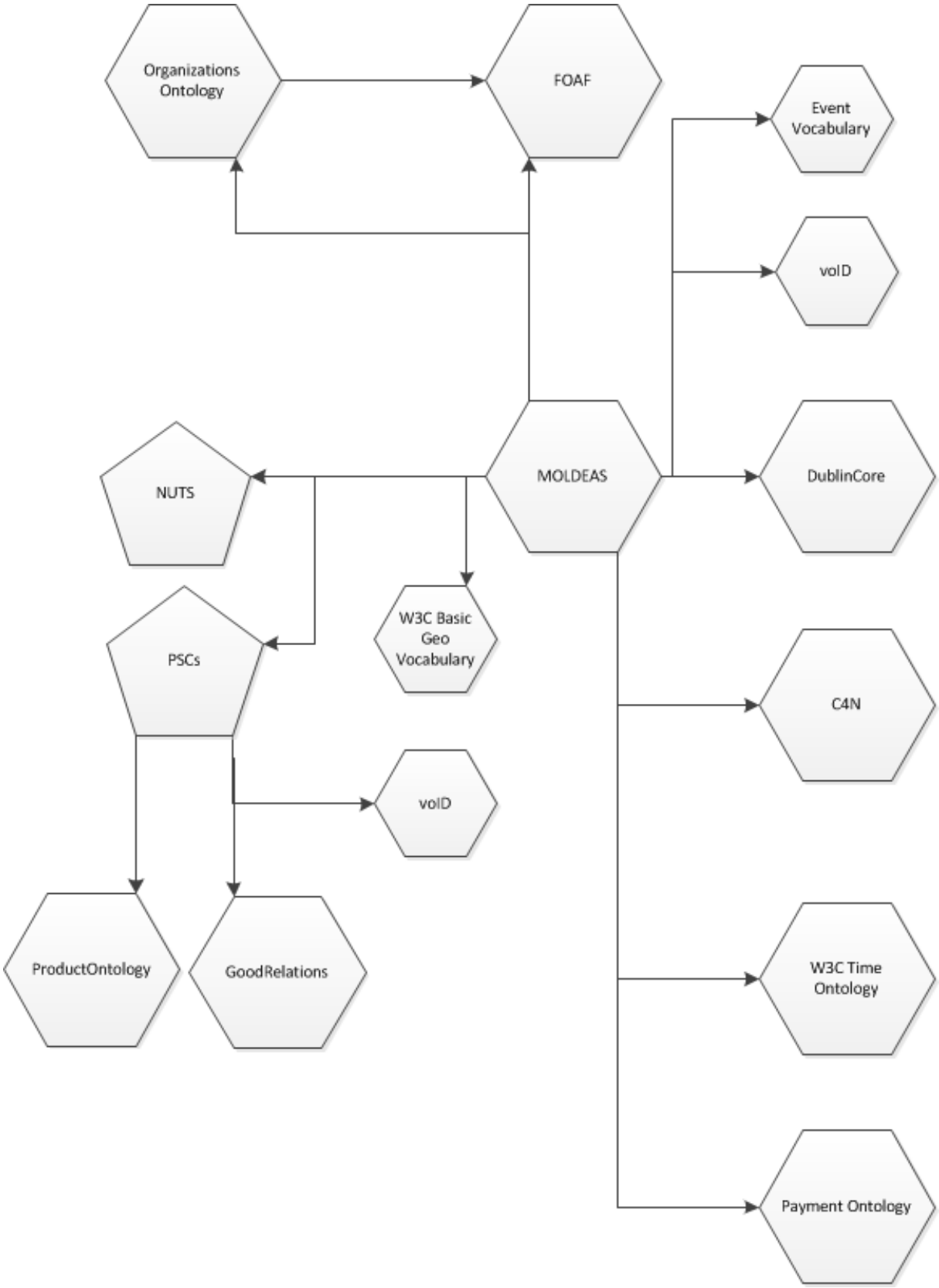


Figura 5.5: Principales Vocabularios (hexágonos) y *Datasets* (pentágonos) utilizados.

La estrategia utilizada para el diseño de este modelo se ha centrado en la reutilización de vocabularios y propiedades existentes, con el objetivo de facilitar su integración posterior en otros modelos y su extensión por terceros. Como segundo paso se diseñan las propiedades que deben tener los recursos (se adjuntan las más representativas del ejemplo en curso), teniendo en cuenta su ulterior aplicación ver Tabla 5.2, pertenecientes a ese *dataset* y que son comunes para todos los anuncios de licitación.

Propiedad	**Descripción**	**Ejemplo**
`rdf:type`	Especificación del tipo de un recurso del *dataset* RDF	a `PublicContract,...,Lot`
`rdfs:label`	Etiqueta o título del anuncio de licitación	"Suministro e instalación de camas nuevas..."^^xsd:string
`rdfs:comment`	Notas adicionales del anuncio de licitación	"Camas de hospitalización general"^^xsd:string
`dc:identifier`	Identificador utilizado en la URI del recurso	"010357"^^xsd:string
`dc:description`	Descripción concreta del recurso RDF	"Camas de hospitalización general"^^xsd:string
`dc:source`	Fuente de publicación de los anuncios de licitación	`moldeas-ppn:DOUE`
`dc:language`	Lenguaje utilizado en un documento	"ES"^^xsd:string
`dc:subject`	Tema u objeto del contracto	`cpv2008:33192000`
`dc:date`	Fecha de publicación de un recurso, pago, etc.	`dc:date` "21/01/2011"^^xsd:date
`dc:publisher`	Entidad contratante	`moldeas-org:GIPSA`
`dc:title`	Etiqueta o título del anuncio de licitación	"Suministro e instalación de camas nuevas..."^^xsd:string
`topic`	Tipo de licitación (códigos CPV) (similar a `dc:source`)	`cpv2008:33192000`
`contractor`	Entidad contratante	`moldeas-org:GIPSA`
`has-lot`	Parte de un lote de un contrato público	`moldeas-ppn:010357/lot/1`
`is-lot-of`	Contrato al que pertenece un lote	`moldeas-pc:010357`
`has-notice`	Anuncio de licitación de un contrato público	`moldeas-ppn:010357`
`has-publication`	Publicación de un contrato público o lote	`moldeas-ppn:010357/publication/1`
`has-previous`	Publicación previa de un contrato público o lote	`moldeas-ppn:010357/publication/1`
`in-process`	Tipo de procedimiento del contrato	`process` a `Open`
`located-in`	Localización del anuncio de licitación	`nuts:ES120`
`budget`	Presupuesto de un lote de un contrato público	`moldeas-ppn:010357/lot/1/budget`

Propiedad	Descripción	Ejemplo
buys	Adquisición realizada en un lote de un contrato público	moldeas-ppn:010357/lot/1/item/1
payment	Pago realizado sobre un ítem de un lote de un contrato público	moldeas-ppn:010357/lot/1/item/1/payment
...		

Tabla 5.2: Diseño de propiedades para los Anuncios de Licitación.

Una vez establecido el conjunto de propiedades de cada recurso RDF representando un anuncio de licitación, es necesario definir el conjunto de grafos en los cuales se encuadrarán los recursos, es decir, el *dataset* RDF $\mathcal{D}$. Para ello, en la Tabla 5.3 se indican las tuplas $(\mathcal{G}_k, I_k)$ correspondientes a cada uno de los grafos $\mathcal{G}_k$ identificados a través de la URI I_k.

$\mathcal{G}_k$	I_k
$\mathcal{G}$	http://purl.org/weso/ppn
$\mathcal{G}_1$	http://purl.org/weso/ppn/2008
$\mathcal{G}_2$	http://purl.org/weso/ppn/2009
$\mathcal{G}_3$	http://purl.org/weso/ppn/2010
$\mathcal{G}_4$	http://purl.org/weso/ppn/2011
$\mathcal{G}_5$	http://purl.org/weso/ppn/2012
$\mathcal{G}_6$	http://purl.org/weso/pc/ontology
$\mathcal{G}_7$	http://purl.org/weso/pc

Tabla 5.3: *Dataset* RDF $\mathcal{D}$ para Anuncios de Licitación.

5.1.1.6 Tarea t_6-Diseño de un Esquema de URIs

Esta tarea tiene como objetivo establecer la forma y estructura de las URIs tanto para las definiciones realizadas en la ontología $\mathcal{O}$, como para todos los recursos presentes en el *dataset* RDF $\mathcal{D}$ que se genera a partir de la transformación de los datos a RDF. Es una de las actividades clave ya que guiará tanto el método final de transformación como el proceso posterior de publicación. La Tabla 5.4 contiene la descripción de la estructura de URIs que se utilizarán en la generación de los anuncios de licitación. Se ha optado por un diseño en el cual las URIs refieren a la información de mayor utilidad y única, abordando la descripción pormenorizada de cada uno de los recursos a través de las propiedades, evitando así la generación de URIs excesivamente extensas con mucha metainformación.

URI	Descripción	Ejemplo
http://purl.org/weso/ppn/	URI base para los anuncios de licitación: <base_uri>	NA
http://purl.org/weso/pc/	URI base para los contratos públicos: <base_uri_pc>	NA
<base_uri_pc>/ontology	Definiciones comunes a todos los contratos públicos	<base_uri_pc>/ontology/PublicContract

URI	Descripción	Ejemplo
<base_uri_pc>/resource/ds	Descripción del catálogo de contratos públicos	<base_uri_pc>/resource/ds
<base_uri_pc>/resource/{id}	URI para un recurso representando un contrato público	<base_uri_pc>/resource/010357
<base_uri>/ontology	Definiciones comunes a todos los anuncios de licitación	<base_uri>/ontology/Item
<base_uri>/resource/ds	Descripción del catálogo de los anuncios de licitación	<base_uri>/resource/ds
<base_uri>/{year}	Espacio de nombres para una determinada PSC	<base_uri>/ppn/2012
<base_uri>/resource/{year}/{id}	URI para un recurso representando un anuncio de licitación	<base_uri>/resource/2012/010357
<base_uri>/resource/{year}/{id}/publication/{id}	URI para un recurso representando una publicación de un anuncio de licitación	<base_uri>/resource/2012/010357/publication/1
<base_uri>/resource/{year}/{id}/lot/{id}	URI para un recurso representando un lote de un contrato público	<base_uri>/resource/2012/010357/lot/1
<base_uri>/resource/{year}/{id}/lot/{id}/item/{id}	URI para un recurso representando un ítem de un lote de un contrato público	<base_uri>/resource/2012/010357/lot/1/item/1
<base_uri>/resource/{year}/{id}/lot/{id}/budget	URI para un recurso representando un presupuesto de un lote de un contrato público	<base_uri>/resource/2012/010357/lot/1/budget
<base_uri>/resource/{year}/{id}/lot/{id}/payment	URI para un recurso representando los pagos de un lote de un contrato público	<base_uri>/resource/2012/010357/lot/1/item/1/payment

Tabla 5.4: Diseño de URIs para los Anuncios de Licitación.

5.1.1.7 Tarea t_7-Diseño Plantilla Objetivo del Recurso RDF

El objetivo de esta tarea es establecer una plantilla de cada uno de los recursos RDF que están presentes en el *dataset* $\mathcal{D}$ para que sirvan como guía en los siguientes momentos: 1) en la ejecución propiamente dicha de la transformación de los datos originales a RDF y 2) en la validación de los recursos RDF generados, de esta manera, tratándose de *datasets* con una gran cantidad de recursos, se pueden identificar fácilmente recursos que no sean compatibles con este esquema favoreciendo la depuración de los recursos generados. Adicionalmente, un esquema de recurso sirve como documentación extra para el proceso de consumo. De acuerdo al recurso plantilla, ver Figura 5.6, y a las definiciones realizadas en la ontología que modela estos datos es posible realizar una validación en cuanto a los tipos de datos, cardinalidad de las relaciones, tipo de objetos, etc., que resulta de sumo interés para asegurar la calidad de los datos producidos.

```
<<base_uri>/resource/ppn/{year}/{id}>
      a        Notice;
     (rdfs:label ""@lang ;)+
     (rdfs:description ""@lang ;)+
     dc:identifier ""^^xsd:string ;
     dc:subject ""^^xsd:string ;
     dc:title ""^^xsd:string ;
     dc:date ""^^xsd:date ;
     (topic <uri>;)*
     (located-in <uri>;)*
      ...
      .
```

Figura 5.6: Plantilla Objetivo de un Recurso de los Anuncios de Licitación.

5.1.1.8 Tarea t_8-Enriquecimiento de los datos en RDF

En el caso de los anuncios de licitación la tarea de enriquecimiento se ha realizado en las propias clasificaciones de productos y en las organizaciones (países y localización). Directamente, se ha optado por evitar el enriquecimiento de los anuncios de licitación manteniendo así la información original de la forma más clara y sencilla posible. Evidentemente, la Tarea t_{10} sobre "Reconciliación de Entidades" se ha delegado, ya que además de proveer enlaces a los documentos oficiales en las fuentes no existen fuentes de datos semánticas en las que se disponga de información extra sobre una determinada licitación.

5.1.1.9 Tarea t_9-Transformación de los datos a RDF

Una vez realizadas las tareas anteriores se está en disposición de realizar la transformación de los datos de entrada a RDF. En esta tarea el punto clave de decisión reside en seleccionar o bien una herramienta ya disponible o bien implementar un programa que ejecute las reglas de transformación tomando como entrada los datos de las clasificaciones. En este caso, se ha optado por la implementación de un proceso Java particular debido a la casuística y al tamaño del *dataset* de entrada.

5.1.1.10 Tarea t_{16}-Añadir metainformación a los recursos RDF

En relación a la metainformación en los anuncios de licitación y teniendo en cuenta la interpretación posible realizada en la Sección 4.4.16 se ha optado por realizar un enfoque mixto en el cual se provee la descripción de los propios *datasets* RDF, ver Figuras 5.7 y 5.8 además de indicar en cada recurso su información particular de procedencia. Es relevante destacar la definición de la licencia de los datos con el objetivo de facilitar su posterior reutilización, en este caso se ha seleccionado una licencia de "*Open Data*" basada en las directrices fijadas en la guía provista en [198].

```
<http://purl.org/weso/ppn/data/resource/ds?output=ttl>
      rdfs:label "RDF description of Public Procurement Notices" ;
      foaf:primaryTopic <http://purl.org/weso/ppn/resource/ds> .

<http://purl.org/weso/ppn/resource/ds>
      a    <http://rdfs.org/ns/void#Linkset> ;
      rdfs:label "Public Procurement Notices"@en ;
      dcterms:author
            <http://www.di.uniovi.es/~labra/labraFoaf.rdf#me> ,
            <http://www.josemalvarez.es/foaf.rdf#me> ;
      dcterms:contributor
            <http://purl.org/weso/pscs/resource/10ders> ,
            <http://rdfohloh.wikier.org/project/moldeas/rdf> ;
      dcterms:description
            "Some Public Procurement Notices available in RDF" ;
      dcterms:license
            <http://opendatacommons.org/licenses/by/1.0/> ;
      dcterms:modified
            "2011-11-10"^^<http://www.w3.org/2001/XMLSchema#date> ;
      dcterms:publisher
            <http://www.josemalvarez.es/foaf.rdf#me> ;
       dcterms:title
            "Public Procurement Notices" ;
      void:target
            <http://purl.org/weso/ppn/2008/resource/ds> ,
            <http://purl.org/weso/ppn/2009/resource/ds> ,
            <http://purl.org/weso/ppn/2010/resource/ds> ,
            <http://purl.org/weso/ppn/2011/resource/ds> ;
            <http://purl.org/weso/ppn/2012/resource/ds> ;
      foaf:homepage <http://purl.org/weso> .
```

Figura 5.7: Descripción del *Linkset* de los Anuncios de Licitación.

```
<http://purl.org/weso/pscs/ppn/2012/resource/ds>
      a        void:Dataset , skos:ConceptScheme ;
      rdfs:label "Public Procurement Notices 2012"@en ;
      dcterms:author
            <http://www.di.uniovi.es/~labra/labraFoaf.rdf#me> ,
            <http://www.josemalvarez.es/foaf.rdf#me> ;
      dcterms:contributor
            <http://purl.org/weso/pscs/resource/10ders> ,
            <http://rdfohloh.wikier.org/project/moldeas/rdf> ;
      dcterms:description "Common Procurement Vocabulary" ;
      dcterms:license <http://opendatacommons.org/licenses/by/1.0/> ;
      dcterms:modified "2011-06-06"^^xsd:date ;
      dcterms:publisher <http://www.josemalvarez.es/foaf.rdf#me> ;
      dcterms:source
        <http://europa.eu/legislation_summaries/internal_market/businesses/
            public_procurement/l22008_en.htm> ;
      dcterms:title "PPN 2012" ;
      void:dataDump <http://purl.org/weso/pscs/ppn/2012/ppn-2012.ttl> ;
      void:exampleResource
        <http://purl.org/weso/ppn/2012/resource/010357>;
      void:uriRegexPattern
        "http://purl.org/weso/ppn/2012/resource/.+" ;
      void:vocabulary skosxl: , skos: , rdfs: ;
      foaf:homepage <http://purl.org/weso> .
```

Figura 5.8: Descripción del *dataset* de Anuncios de Licitación 2008.

5.1.1.11 Resultado Final y Ejemplos

El resultado final del proceso de producción de *Linked Data* tras el análisis y ejecución de las tareas identificadas y del método de producción seleccionado, genera como resultado un conjunto de datos en RDF mediante datos enlazados en los cuales se pueden extraer las siguientes estadísticas de producción de datos, ver Tabla 5.5, consultas, ver Figura 5.9 y un ejemplo conteniendo la información básica de los recursos generados de forma gráfica, ver Figuras 5.10 y 5.11.

"Dame 100 anuncios de licitación en el campo de maquinaria industrial que se han solicitado en Francia e Inglaterra durante los años 2008 y 2009"

Anuncios de Licitación	**Nº de Elementos**	**Tripletas**
PPN 2008	112843	677058
PPN 2009	399766	2398601
PPN 2009	431813	2590880
PPN 2011	67044	402264
Catálogo de Anuncios de Licitación (total)		
PPNs	1011466	6068803

Tabla 5.5: Estadísticas y Ejemplos de los Anuncios de Licitación.

```
SELECT DISTINCT * WHERE{
  ?ppn rdf:type moldeas-pc:Notice.
  ?ppn moldeas-ppn:located-in ?nutsCode.
  FILTER(?nutsCode = <http://nuts.psi.enakting.org/id/UK> or
    ?nutsCode = <http://nuts.psi.enakting.org/id/FR> ) .
  ?ppn moldeas-ppn:topic ?cpvCode.
  FILTER(?cpvCode = cpv2008:42000000 ) .
  ?ppn dc:identifier ?id.
  ?ppn dc:date ?date .
  FILTER ( (xsd:long(?date) >= xsd:long(2012))
    && (xsd:long(?date) <= xsd:long(2012)) )
} LIMIT 100
```

Figura 5.9: Ejemplo de consulta en SPARQL sobre los Anuncios de Licitación.

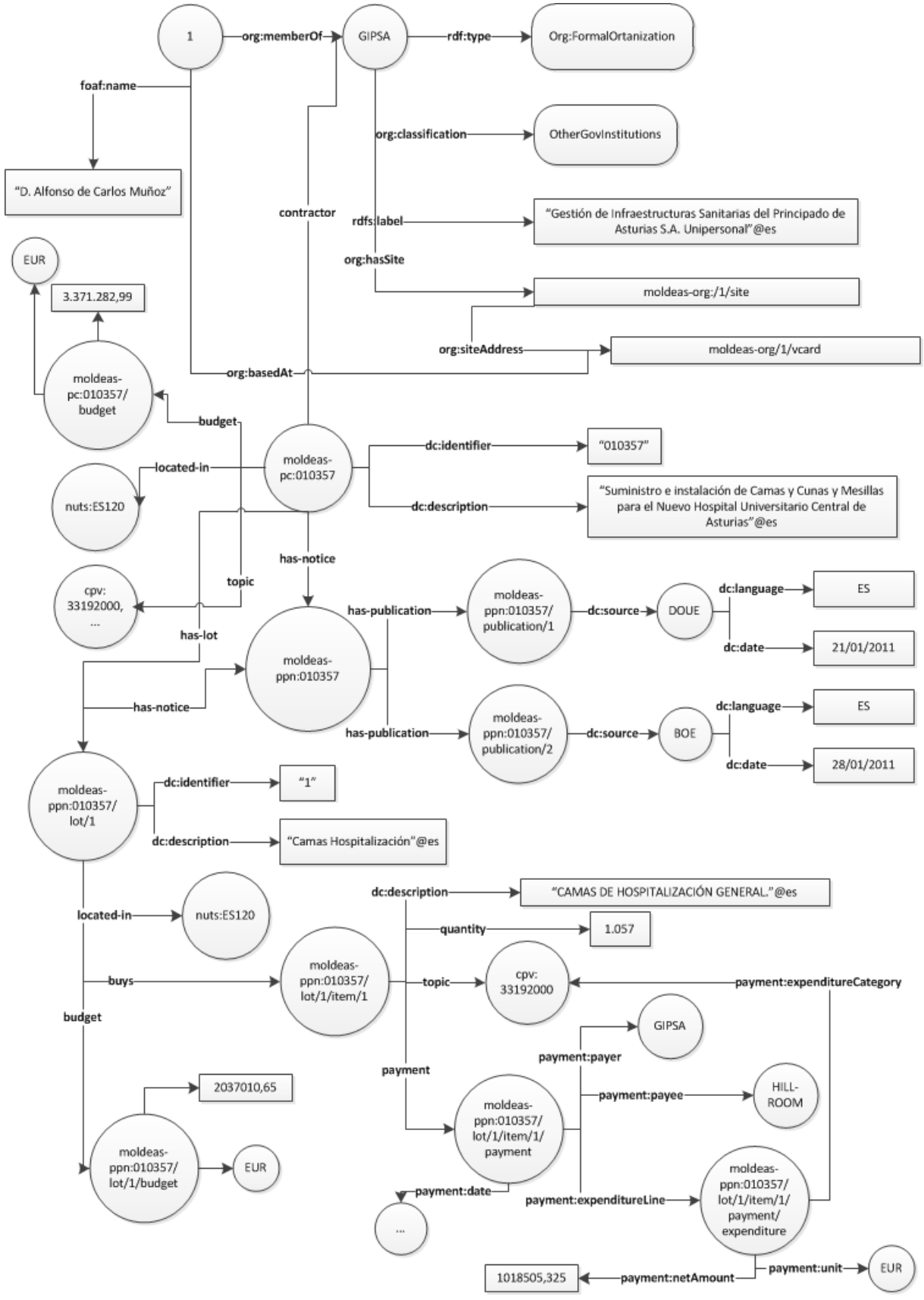


Figura 5.10: Ejemplo de modelado para la licitación pública de ejemplo (I).

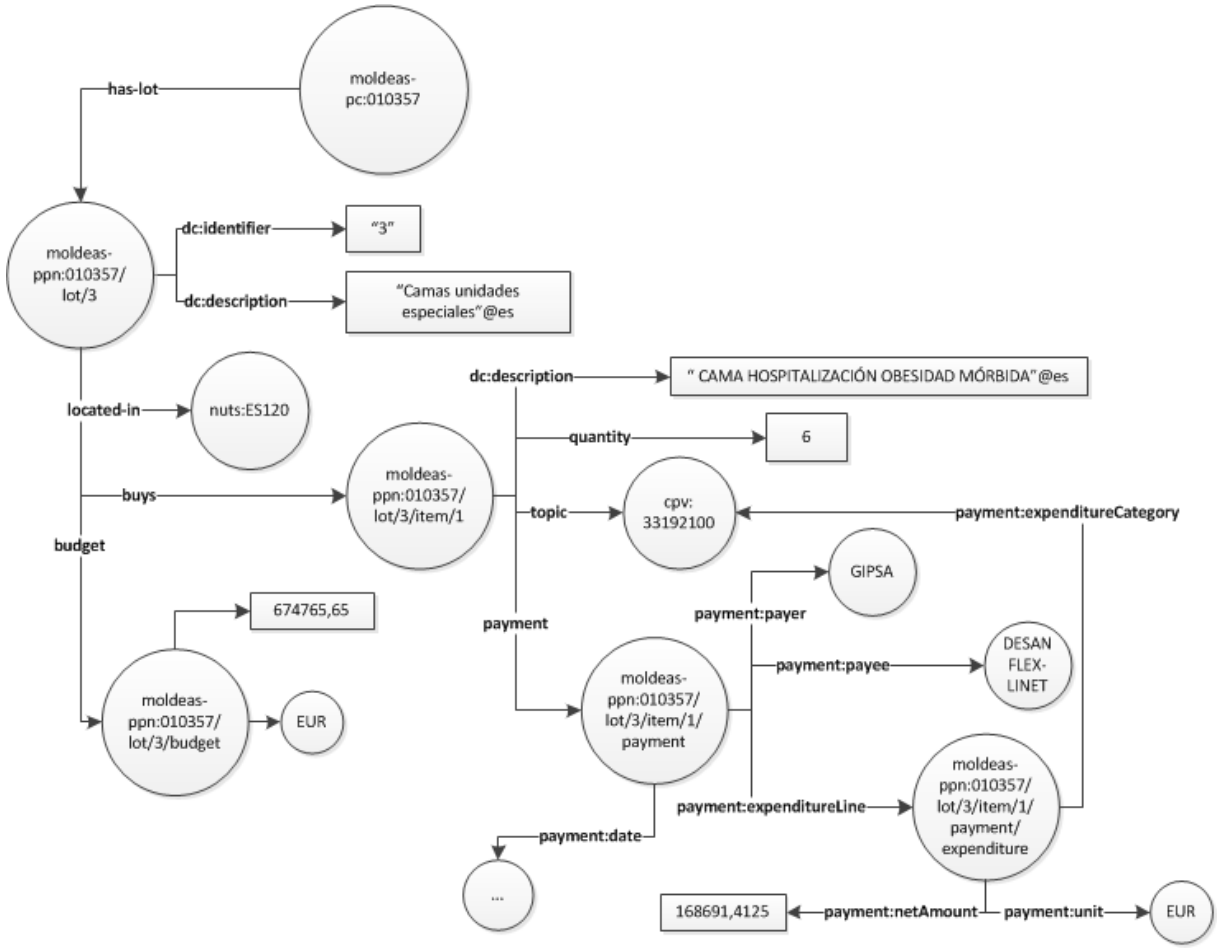


Figura 5.11: Ejemplo de modelado para la licitación pública de ejemplo (II).

5.1.1.12 Método de Producción de *Linked Data* de Anuncios de Licitación

De acuerdo al análisis y diseño de datos enlazados realizado para los anuncios de licitación a lo largo de las anteriores secciones y la tabla de decisión 4.12, el método semántico seleccionado para realizar la producción de datos enlazados es el SPM_1-"Transformación de datos a RDF", ver Sección 4.5.2, en el se transforman un conjunto de datos de entrada $\mathcal{G}$ a un *dataset* RDF $\mathcal{D}$. Según la definición de método semántico de producción, realizada en la Sección 4.5.1, y el estudio de los anuncios de licitación se pueden establecer los siguientes conjuntos:

- $\mathcal{G}$ es el *dataset* de entrada, conjunto de tuplas, conteniendo los datos de cada uno de los anuncios de licitación.
- $\mathcal{M}$ es el conjunto de *mapeos*, ver Tabla 5.6, extraídos según el análisis y diseño realizado en las secciones anteriores. Estos *mapeos* son directamente expresables en la herramienta de transformación y toman como parámetros el valor de una de las tuplas de entrada (posición X) y la propiedad a generar.
- *Dataset* RDF $\mathcal{D}$ es el *dataset* resultado, siguiendo el análisis y diseño realizado en las secciones anteriores y tras la ejecución de la tarea propia de transformación de datos.

$\mathcal{M}$	**Propiedad**	**Valor**
m_1	`rdf:type`	URI
m_2	`rdfs:label`	`xsd:string@lang`
m_3	`rdfs:comment`	`xsd:string@lang`
m_4	`dc:identifier`	`xsd:string`
m_5	`dc:description`	`xsd:string`
m_6	`dc:source`	URI
m_7	`dc:language`	`xsd:string`
m_8	`dc:date`	`xsd:string` o URI
m_9	`dc:publisher`	`xsd:string` o URI
m_{10}	`dc:title`	`xsd:string`
m_{11}	`topic`	URI
m_{12}	`contractor`	URI
m_{13}	`has-lot`	URI
m_{14}	`is-lot-of`	URI
m_{15}	`has-notice`	URI
m_{16}	`has-publication`	URI
m_{17}	`has-previous`	URI
m_{18}	`in-process`	URI
m_{19}	`located-in`	URI
m_{20}	`budget`	URI
m_{21}	`buys`	URI
m_{22}	`payment`	URI
...		

Tabla 5.6: Conjunto de *mapeos* $\mathcal{M}$ para los Anuncios de Licitación.

5.1.2 Proceso de Publicación de *Linked Data* de Anuncios de Licitación

Según la definición realizada en la Sección 4.6, este proceso conlleva todas las tareas que implican la publicación de un *dataset* RDF $\mathcal{D}$, a través de un conjunto de características $\mathcal{P}$, para la obtención de un *dataset* RDF $\mathcal{D}_{pub}$, método semántico de publicación. El método seleccionado para la publicación de datos es una conjunción de los detallados en las Secciones 4.6.2 (Fichero estático RDF), 4.6.4 (*Endpoint* de SPARQL) y 4.6.6 (*Linked Data Frontend*) con el objetivo de proporcionar cobertura a las necesidades de la mayor parte de los potenciales clientes. De esta forma, se suministra un entorno en el cual se pueden: 1) descargar todos los datos de una única consulta y un sólo formato de datos; 2) realizar consultas personalizadas con distintos formatos de salida a través del protocolo y lenguaje propuesto en SPARQL y 3) navegar a través de los datos mediante un interfaz con provisión para la negociación de contenido. Debido al ingente conjunto de características del conjunto $\mathcal{P}$ para la publicación a través de la web, incluyendo las propias de HTTP, se indican aquellas más importantes y su cumplimiento ($\oplus$, $\ominus$) de acuerdo a los métodos de publicación seleccionados, ver Tabla 5.7.

$\mathcal{P}$	**Característica**	**Valor**
p_1	Acceso HTTP	$\oplus$
p_2	Negociación de contenido	$\oplus$
p_3	Consultas en SPARQL	$\oplus$
p_4	Actualización en SPARQL	$\oplus$
p_5	Volcado de datos en RDF	$\oplus$
p_6	Licencia de los datos enlazados	$\oplus$

$\mathcal{P}$	**Característica**	**Valor**
p_7	Documentación extra	$\oplus$
	...	

Tabla 5.7: Conjunto genérico de características $\mathcal{P}$ de publicación.

5.1.2.1 Tarea t_{14}-Infraestructura para *Linked Data*

El objetivo de esta tarea es el diseño de la infraestructura necesaria para albergar los datos enlazados, permitiendo así el acceso a los mismos y su consiguiente reutilización siguiendo los métodos de publicación seleccionados. Dependiendo de los servicios a ofrecer se deberán contemplar distintos componentes software, aunque un ejemplo característico del posible diagrama de despliegue será el que se puede observar en la Figura 5.12 y que se utilizará como representativo para el contexto de los anuncios de licitación, incluyendo las clasificaciones de productos y las organizaciones.

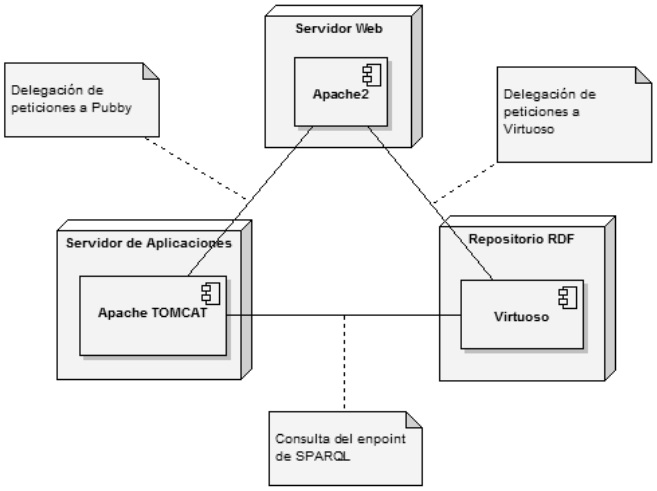


Figura 5.12: Infraestructura Objetivo para *Linked Data*.

Cada uno de estos componentes cumple una función específica que se detalla a continuación:

Servidor web. Teniendo presentes dos de los grandes objetivos de la publicación de datos enlazados como son el uso de HTTP URIs y que éstas sean en el mayor grado posible *cool uris*, la presencia de un servidor web se justifica como punto de entrada a la consulta de los datos enlazados evitando la presencia de números de puerto, etc., suministrando un acceso homogéneo a los datos. Para el cumplimiento de estos objetivos se ha seleccionado el servidor web Apache2. Por otra parte, con una configuración adecuada del servidor se da soporte al método de publicación basado en un "*Fichero estático RDF*".

Servidor de aplicaciones. Este componente es el encargado de albergar las aplicaciones que proporcionan servicios de valor añadido de acceso a los datos desde diferentes puntos de vista. También se puede desplegar un contenedor de servlets que en muchos casos es suficiente. En este caso, seleccionando Pubby como elemento de software para dar soporte al método de publicación mediante un "*Linked Data Frontend*".

Repositorio RDF. Con el objetivo de albergar los datos transformados en RDF, es decir el *dataset* RDF $\mathcal{D}$, es necesario el despliegue de un servidor RDF, para esta tarea existen diferentes op-

ciones [55] pero Virtuoso de OpenLink se considera una buena opción por su amplia aceptación, capacidad de extensión y soporte a las especificaciones.

***Endpoint* de SPARQL.** Del mismo modo que los datos transformados se han de almacenar, también es conveniente proveer un interfaz de acceso a los mismos mediante el protocolo y lenguaje de consulta SPARQL. Igualmente existen diferentes opciones, sin embargo formando parte de Virtuoso existe un módulo que se instala y está disponible automáticamente con el despliegue inicial del repositorio. El comportamiento de este componente se asemeja también a un servicio web REST con peticiones GET, facilitando de forma parcial otro método de publicación como el señalado en la Sección 4.6.7.

5.1.2.2 Tarea t_{15}-Acceso y formato en datos RDF

Las necesidades de las aplicaciones para el posible consumo de datos varían ostensiblemente dependiendo de su objetivo. En el caso objeto de estudio de este documento y con la infraestructura incluida en la anterior sección se suministra soporte a los siguientes formatos de datos:

Acceso	**Formato**	**Provisto por**
Petición GET	N3/Turtle	Apache2
Consulta SPARQL	*Spreadsheet*	*Endpoint* de SPARQL
≡	XML	≡
≡	JSON	≡
≡	Javascript	≡
Consulta SPARQL y petición GET	N3/Turtle	*Endpoint* de SPARQL y *Linked Data Frontend*
≡	RDF/XML	≡
≡	NTriples	≡
≡	HTML	≡

Tabla 5.8: Acceso y Formato de datos de los Anuncios de Licitación.

5.1.3 Proceso de Consumo de Anuncios de Licitación

El proceso de consumo de datos enlazados, según la definición realizada en la Sección 4.7.1, consiste en la reutilización de los mismos para ser aplicados en la construcción de una nueva aplicación o servicio de valor añadido. En general, la reutilización más sencilla consiste en la representación gráfica de los recursos o la simple consulta con selección de formato de datos de acuerdo a las características de publicación utilizadas. En el caso que nos ocupa y teniendo en cuenta el objetivo de realización de un prototipo experimental de extracción de anuncios de licitación como demostrador del consumo de datos enlazados, se ha escogido el método semántico de consumo SCM_2-"*Mapeo* a Lenguaje de Programación", cuya descripción está disponible en la Sección 4.7.3, orientado a obtener una representación de los recursos RDF en un lenguaje de programación (en este caso Java) como objetos de negocio. De acuerdo a este objetivo y a la definición del propio método es necesario precisar que:

- El *dataset* RDF $\mathcal{D}_{pub}$, es el conjunto de datos disponible tras aplicar el método de publicación.
- El conjunto $\mathcal{M}^1$, ver Tabla 5.9, indica como transformar el *dataset* anterior a la representación objetivo: objetos del lenguaje Java.

De esta manera, se obtienen una serie de objetos, $\mathcal{D}_{consum}$, con la información y datos necesarios, no se transforman necesariamente todos los datos disponibles en los recursos pertenecientes a $\mathcal{D}_{pub}$, para ser reutilizados como objetos de negocio en un lenguaje de programación. Es conveniente señalar que el acceso a los datos se realiza a través de la consulta al *endpoint* de SPARQL ejecutando consultas *SELECT* y *DESCRIBE*.

$\mathcal{M}^1$	**Propiedad**	**Tipo en Java**
m_1^1	URI recurso	`java.lang.String`
m_2^1	`rdf:type`	`org.weso.moldeas.to.PPNTO`
m_3^1	`dc:identifier`	`java.lang.String`
m_4^1	`dc:subject`	`java.lang.String`
m_5^1	`dc:date`	`java.util.Date`
m_6^1	`dc:publisher`	`org.weso...OrganizationTO`
m_7^1	`topic`	`java.util.List<PSCTO>`
m_8^1	`located-in`	`java.util.List<NUTSTO>`
m_9^1	`rdfs:label`	Map<String,String>(lang, value) para cada propiedad
m_{10}^1	`rdfs:description`	Map<String,String>(lang, value) para cada propiedad
...		

Tabla 5.9: Conjunto de *mapeos* $\mathcal{M}^1$ de consumo para los Anuncios de Licitación.

5.1.4 Proceso de Validación de Anuncios de Licitación

La validación como proceso transversal en cualquier etapa dentro del ciclo de vida de datos enlazados debe realizarse con el objetivo de asegurar la calidad de los datos. De acuerdo a la definición realizada en la Sección 4.8, este proceso consiste en la comprobación de que los recursos de un *dataset* RDF cumplen ciertas características, esta validación puede ser realizada manual o automáticamente dependiendo del caso, por ejemplo se dispone de herramientas adecuadas para las características de negociación de contenido o para la inclusión en la nube de datos enlazados, pero en cambio para comprobaciones relativas a los dominios y rangos de las propiedades, etc., no existe una herramienta completa. Por todo ello, se ha seguido un enfoque híbrido basado en la utilización de herramientas y validación manual. La descripción completa de la validación de acuerdo a todas las características se reseña en las Tablas de Validación disponibles en el Apéndice C.

5.1.4.1 Tarea t_{12}-Validación de Recursos RDF

Siguiendo con la definición realizada de esta tarea en la Sección 4.4.12, se puede asegurar que la transformación realizada de los anuncios de licitación a la iniciativa *Linked Data* cumple estrictamente los siguientes puntos:

- Los datos RDF son correctos, ya que se han utilizado el API de Jena que asegura la generación correcta de RDF.
- El dominio y rango en las propiedades es correcta, ya que se realiza la validación contra el modelo definido.
- Se ha establecido metainformación sobre la procedencia a nivel de *dataset*.

- Todos los recursos transformados siguen la plantilla objetivo RDF.

En conclusión, las tareas, métodos y el proceso de validación tienen una repercusión verdaderamente trascendente en el ciclo de vida de datos enlazados, es por ello que en este estudio se ha tenido especial interés en la consecución correcta de la validación de los datos enlazados.

5.1.5 Proceso de Realimentación de Anuncios de Licitación

Este proceso, según la definición realizada en la Sección 4.9, busca la mejora y perfeccionamiento de los datos promocionados a RDF. Esta situación emerge en el momento en el que los datos comienzan a ser reutilizados tanto por aplicaciones o servicios como por individuos. En el caso particular de los anuncios de licitación no han llegado a ser reutilizados por terceras partes, por lo que la realimentación ha quedado restringida a la captura de fallos por la propia aplicación MOLDEAS, tratándose en este caso de una forma de realimentación basada en *Usuarios y Aplicaciones* y de carácter *Actualización Ocasional*.

5.2 Clasificaciones Estándar de Productos

Los sistemas de organización del conocimiento, *Knowledge Organization Systems* (KOS), como tesauros, taxonomías o sistemas de clasificación, se han desarrollado en el seno de distintas comunidades e instituciones, principalmente con el objetivo de organizar grandes bases de datos que contienen recursos de diferente tipo tales como documentos, páginas web o elementos multimedia. Estos vocabularios permiten a los usuarios la anotación de los objetos de información de los recursos para simplificar su extracción y consulta, habitualmente se utilizan técnicas de indexado por un determinado tema con el objeto de facilitar el acceso también para las máquinas, suministrando los metadatos necesarios en la descripción de los objetos de información.

En el dominio de la contratación pública electrónica estas clasificaciones resultan de gran interés para la especificación de los objetos de contrato, así como para la extracción de estadísticas a posteriori. Como se ha reseñado en la Sección 2.10, una de las principales clasificaciones en este sentido es el "*Common Procurement Vocabulary*"(CPV) [70], pero también existen otros esquemas de clasificación de gran interés en la esfera del comercio electrónico, así la "*Combined Nomenclature*" o el "*North American Product Classification System*". En el ámbito de esta tesis se ha optado por promocionar las clasificaciones más importantes en este contexto para facilitar el acceso a los anuncios de licitación independientemente de la clasificación utilizada, para ello, se ha realizado la transformación siguiendo los principios de *Linked Data* y los métodos definidos en el capítulo anterior de las siguientes clasificaciones estándar de productos (PSCs) (un total de 9), ver Tabla 5.10.

Clasificación	**Acrónimo**	**Organismo**
Common Procurement Vocabulary, (2003 y 2008)	CPV	Unión Europea
Combined Nomenclature 2012 (desde 1995)	CN	Unión Europea
Central Product Classification, version 2 (2008)	CPC	Unión Europea
Clasificación de Productos por Actividad (2008)	CPA	Unión Europea
International Standard Industrial Classification of All Economic Activities, Rev.4	ISIC	*United Nations Statistics Division*

Clasificación	Acrónimo	Organismo
North American Industry Classification System 2007 y 2012	NAICS	Gobierno de Estados Unidos
Standard International Trade Classification, Revision 4	SITC	*United Nations Statistics Division*

Tabla 5.10: Catálogo de Clasificaciones Estándar de Productos seleccionadas.

5.2.1 Proceso de Producción de *Linked Data* de Clasificaciones Estándar de Productos

Siguiendo la definición realizada en la Sección 4.5, este proceso conlleva todas las tareas que implican la transformación de un *dataset* de entrada $\mathcal{G}$, mediante unas reglas de *mapeo* $\mathcal{M}$, para la obtención de un *dataset* RDF $\mathcal{D}$, método semántico de producción.

5.2.1.1 Tarea t_1-Análisis del *dataset* a transformar

Las PSCs como instrumentos claves de estandarización nacen con el fin de conseguir una clasificación común de productos y servicios. Las diferencias entre las clasificaciones no sólo se limitan a cuestiones de alcance y cobertura sectorial de producto, sino también al grado de especificidad que difiere de unas a otras.

Como ya se ha reseñado en la Sección 3.4.1.3, Hepp [159–161] apunta a estos estándares como una combinación de componentes variables que pueden ser utilizados para la construcción de ontologías derivadas, sin embargo, se puede identificar una estructura común subyacente a todas las PSCs y que debe considerarse fundamental para proporcionar un modelo de datos semántico universal para las PSCs, para ello, se utilizarán los conceptos los conceptos de *árbol* y *bosque* provenientes de la teoría de grafos con el objetivo de representar la estructura común de las PSCs.

Categorías de productos. Las clasificaciones se dividen en categorías o clases de productos. Estas categorías agrupan los distintos elementos de la PSC, Cat_{psc}, en distintos niveles de especialización semántica o niveles de jerarquía: $Cat_{psc} = \bigcup_{n=0}^{k} (Cat^n_{psc})$, desde términos genéricos, como el caso del elemento del CPV-"Servicios de reparación y mantenimiento" (código 50000000), hasta productos altamente específicos y directamente identificables, como en un elemento de la misma jerarquía anterior pero con un mayor nivel de especificidad como "Servicios de reparación y mantenimiento de instalaciones contra incendios" (código 50413200). Una PSC cumple las siguientes características:

- Las categorías de la PSC se organizan jerárquicamente: $Cat^0_{psc} \succ Cat^1_{psc} \succ ... \succ Cat^n_{psc}$.
- Cada elemento de la PSC, t^x_{psc}, pertenece a una categoría de productos.
- Cada elemento de la PSC, t^x_{psc}, pertenece sólo a una categoría de productos. Es decir, las categorías son disjuntas: $\bigcap_{n=0}^{k} (Cat^n_{psc}) = \emptyset$.

Estructura taxonómica. Además de la división en niveles de jerarquía de los elementos de una PSC, su objetivo es organizar y agrupar los productos en sectores verticales mediante algún tipo de criterio establecido por la comunidad que desarrolla el estándar. Formalmente, esta estructura

taxonómica de cada sector de productos tiene forma de árbol, T_{psc}: todos los elementos t^{n}_{psc} tienen un elemento de nivel superior t^{n-1}_{psc}. Además el conjunto de sectores de productos de una clasificación constituye la propia PSC, que puede definirse como un **bosque** ($\mathbb{F}_{psc}$) de árboles (T_{psc}), en el que se cumple que:

- $\mathbb{F}_{psc} = \bigcup_{m=0}^{k} (T^{m}_{psc})$
- Cada elemento t^{0}_{psc} es la **raíz** de una agrupación de productos estructurada jerárquicamente en forma de árbol, T^{x}_{psc}.
- Cada elemento t_{psc} pertenece a uno de estos árboles de productos T^{x}_{psc}.
- Por la propia definición de bosque, cada elemento t_{psc} pertenece sólo a un T^{x}_{psc}. Es decir, cada sector de productos es disjunto: $\bigcap_{m=0}^{k} (T^{m}_{psc}) = \emptyset$.

Estas son características genéricas de las clasificaciones de productos, sin embargo, otras PSCs más sofisticadas incluyen un diccionario de propiedades estándar que se puede utilizar para describir productos con más detalle. Normalmente, estos diccionarios de propiedades también incluyen los tipos de datos que pueden ser valor de las mismas, así como su referencia con respecto a estándares internacionales para establecer las unidades de medida, tal es el caso de la clasificación de productos de e@Class. En otras ocasiones, se construyen clasificaciones multilingües para la expresión de los descriptores de cada elemento de la PSC, el caso extremo es el CPV, donde se contempla un total de hasta 23 lenguas.

El desafío básico más importante que hay que afrontar cuando se deriva una ontología de una PSC, reside en cómo interpretar la semántica original de la taxonomía. No existe una definición formal de las relaciones taxonómicas que construyen cada T_{psc} de la clasificación y es tentador utilizar la propiedad de un vocabulario de ontologías, como *rdfs:subClassOf*, para intentar representar estas relaciones semánticas. La postura del enfoque seguido es radicalmente distinta, ver Sección 3.4.1.3, desde este punto de vista, las PSCs fueron construidas para solucionar problemas de comunicación, proporcionar una forma de organizar tipos de productos y agruparlos de acuerdo a unos conceptos y definiciones que funcionasen *de facto* como un estándar en determinados entornos de actividad comercial e industrial. Las PSCs no fueron diseñadas como modelos conceptuales de dominio, en el sentido actual que tiene el término "ontología", sino como una forma de estructurar la terminología y la forma de nombrar los productos, de ahí que se interpreten las PSCs como simples esquemas conceptuales en el que la relación taxonómica que jerarquiza los distintos elementos de cada T_{psc} no se interpreta como una relación de herencia o subtipo, sino como una relación de mayor o menor especificidad de los elementos. Resumiendo, se consideran las PSCs como simples vocabularios controlados y se utilizará una ontología RDF/OWL, SKOS Core, como modelo de datos común.

La adopción de una ontología como SKOS Core para modelar las PSCs está completamente justificada desde este enfoque, se trata de una ontología RDF/OWL que gira en torno a dos clases principales *skos:Concept* y *skos:ConceptScheme*, así todos los elementos de una PSC son considerados instancias de *skos:Concept*, de manera que se entiende que todos ellos son conceptualizaciones de recursos. Asumiendo este modelo, la interpretación de cada elemento t_{psc} de una PSC es natural, directa y no provoca incoherencias en su tratamiento, tal como se ha reseñado en los ejemplos anteriores de inconsistencia. El interés de este enfoque, radica en la posibilidad de poder replicar completamente la semántica original de las PSCs.

Para agrupar todos los elementos de una PSC bajo un paraguas común, se construye una clase

contenedora de todos los t_{psc}: $PSCConcept$, declarándola como subclase del concepto $skos : Concept$:

$$PSCConcept \sqsubseteq skos : Concept \tag{5.1}$$

Esta clase a su vez se puede dividir en distintas categorías que constituyen los niveles de jerarquía del PSC (si los hubiera). El atractivo en la utilización de SKOS Core reside en que es factible replicar completamente el modelo de datos común de las PSCs, de esta forma, por cada Cat^k_{psc}, se construye un concepto subclase de *PSCConcept*, en el caso particular del CPV:

$$Division \sqsubseteq PCSConcept; Grupo... \tag{5.2}$$

Y se mantiene la semántica original de la clasificación añadiendo los siguientes axiomas, perfectamente expresables en OWL-DL para el caso del CPV. En otras clasificaciones, no es posible establecer este tipo de axiomas debido a la falta de designación de categorías ya que la relación de jerarquía se establece simplemente mediante un código numérico.

$$PSCConcept \equiv Division \sqcup Grupo \sqcup Clase \sqcup Categoria \tag{5.3}$$

$$Division \sqcap Grupo \sqcap Clase \sqcap Categoria = \bot \tag{5.4}$$

Por otro lado, el segundo factor importante a destacar reside en la estructura taxonómica de cada T_{psc}, que como se ha referido, tenía una difícil interpretación como conjunto de relaciones de herencia clásica, sin embargo, bajo el paradigma SKOS Core se contemplan relaciones semánticas tales como: `skos:inScheme`, `skos:related`, `skos:broader`, `skos:narrower` y sus versiones transitivas, que constituyen el marco conceptual adecuado para el establecimiento de las relaciones entre los elementos de la PSC.

Finalmente, la clase *skos:ConceptScheme* deviene necesario para completar la interpretación de una PSC y su modelo, ver Figura 5.13, se reifica una PSC como una instancia de esta clase para también así contextualizar los términos y distinguirlos e identificarlos en lo referido a un entorno web.

$$PSCConcept \equiv\ni inScheme\ PSCScheme \tag{5.5}$$

5.2.1.2 Tarea t_2-Limpieza de datos

Los datos provenientes de las clasificaciones de productos se encuentran disponibles, en la mayoría de los casos, en hojas de cálculo MSExcel o bien directamente en formato CSV en los cuales ya se ha realizado un esfuerzo por los organismos creadores para evitar caracteres extraños o incorrectos. Habitualmente, todos estos datos provienen de fuentes oficiales por lo que este proceso ya ha sido realizado en la labor de creación de los mismos, es por lo que en la promoción de estos datos, esta tarea ya ha sido realizada y no es necesario aplicar ningún filtro especial a los datos ya disponibles. Esta circunstancia supone un gran avance para la transformación de datos, ya que se dispone de aquellos clasificados con 2$\star$ según los principios de *Linked Data* y no es necesario realizar procesos de *screen-scrapping* con la consiguiente ganancia en la calidad de los datos.

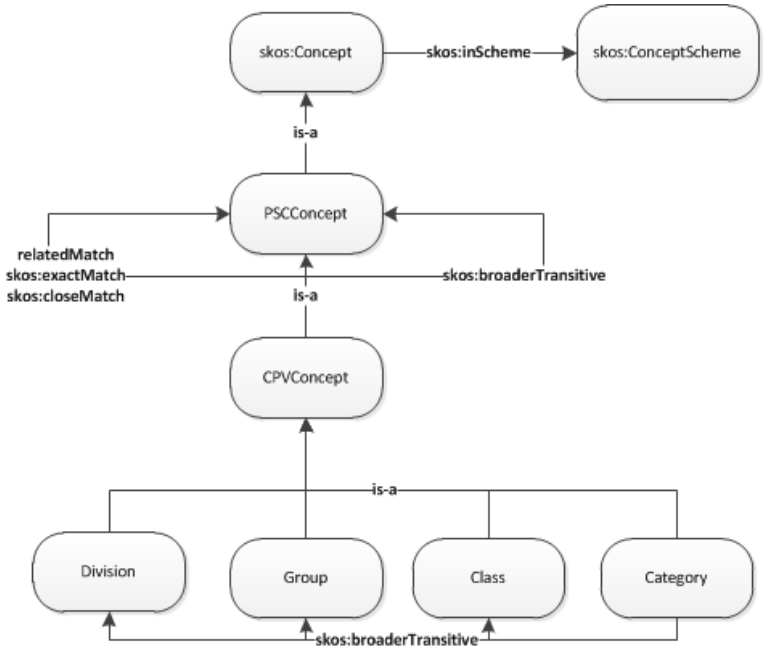


Figura 5.13: Modelo gráfico para las Clasificaciones Estándar de Productos.

5.2.1.3 Tarea t_3-Selección de Vocabularios

Los vocabularios seleccionados para modelar las clasificaciones de productos, teniendo en cuenta el análisis realizado en la tarea t_1 se presentan en la Tabla 5.11. En general, se trata de vocabularios que atienden a los siguientes criterios:

1. Formalización de una estructura taxonómica, como RDFS, SKOS u OWL.
2. Realización de *mapeos* entre conceptos, como SKOS y OWL.
3. Representación de tipos de datos, como XML Schema.
4. Gestión de información multilingüe, como SKOS-XL y RDFS.
5. Representación de información de negocio, como *GoodRelations* y *ProductOntology*.
6. Adición de metadatos y *provenance*, como *Dublin Core Terms*, voID y *Provenance Ontology*.
7. Representación del tiempo e intervalos, como *Time Ontology* del W3C.

Prefijo	Vocabulario	Fuente	Uso
dbpedia	http://dbpedia.org/ontology/	Comunidad *Linked Data*.	Reutilización de definiciones.
dc	http://purl.org/dc/elements/1.1/	*Dublin Core Metadata Initiative*	Creación de metadatos para los documentos.
dct	http://dublincore.org/documents/dcmi-terms/	≡	≡
foaf	http://xmlns.com/foaf/0.1/	Comunidad de Web Semántica.	Especificación de relaciones entre personas.

Prefijo	Vocabulario	Fuente	Uso
gr	http://purl.org/goodrelations/v1#	Martin Heep	Reutilización de definiciones para describir productos y servicios.
owl	http://www.w3.org/2002/07/owl#	W3C	Realización de definiciones en el dominio.
po	http://www.productontology.org/	*Martin Heep*	Reutilización de datos provenientes de Productontology.
skos	http://www.w3.org/2004/02/skos/core#	W3C	Especificación de taxonomías.
skosxl	http://www.w3.org/2008/05/skos-xl#>	W3C	Representación de información lingüística.
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#	W3C	Descripción de recursos.
rdfs	http://www.w3.org/2000/01/rdf-schema#	W3C	Descripción de recursos con relaciones lógicas.
void	http://rdfs.org/ns/void#	Deri y W3C	Descripción de metadatos de un *dataset*.
xml	http://www.w3.org/XML/1998/namespace	W3C	Reutilización de definiciones.
xsd	http://www.w3.org/2001/XMLSchema#	W3C	Especificación de tipos de datos.

Tabla 5.11: Selección de Vocabularios para las Clasificaciones Estándar de Productos.

5.2.1.4 Tarea t_4-Selección de otros *datasets* RDF

Al igual que en la sección anterior, los *datasets* a reutilizar se centran en vocabularios de negocio ya existentes y por consiguiente en los datos disponibles en los mismos, de esta forma, se reutilizan y enlazan datos provenientes de *GoodRelations* y *ProductOntology*, ver Tabla 5.12.

Prefijo	*Dataset*	Fuente	Uso
dbpedia-res	http://dbpedia.org/	Comunidad *Linked Data*.	Reutilización de datos provenientes de la DBPedia.
gr	http://purl.org/goodrelations/v1#	*Martin Heep*	Reutilización de definiciones para describir productos y servicios.
po	http://www.productontology.org/	*Martin Heep*	Reutilización de datos provenientes de ProductOntology.

Tabla 5.12: Selección de otros *datasets* para las Clasificaciones Estándar de Productos.

5.2.1.5 Tarea t_5-Modelado de datos en RDF

El resultado de esta tarea debe ser una ontología de dominio $\mathcal{O}$. que modele los recursos t_{psc}, pertenecientes a un *dataset* RDF $\mathcal{D}$. Inicialmente, se debe proveer la especificación del modelo formal u ontología $\mathcal{O}$ de acuerdo al análisis realizado en la Sección 5.2.1.1, dando soporte a la descripción de los datos de los recursos RDF:

- $\mathcal{C} = \{PSCConcept, CPVConcept, Division, Group, Class, Category\}$
- $\mathcal{R} = \{rdf:type, skos:inScheme, dc:identifier, dc:subject, level, pscs:relatedMatch, pscs:level, skos:exactMatch, skos:broaderTransitive, skosxl:prefLabel, rdfs:label, gr:description\}$
- $\mathcal{I} = \{t_{psc}\}$
- $\mathcal{A} = \{\star\}$, son los axiomas propios de una ontología en SKOS.

Como segundo paso se diseñan las propiedades que deben tener los recursos, teniendo en cuenta su ulterior aplicación, ver Tabla 5.13, pertenecientes a ese *dataset*, comunes para todas las PSCs.

Propiedad	Descripción	Ejemplo
`rdf:type`	Especificación del tipo de un recurso t_{psc}	a `gr:ProductOrServiceModel`, `PSCConcept`
`skos:inScheme`	URI al esquema de la PSC T_{psc} a la que pertenece t_{psc}	`skos:inScheme` cpv2008:ds
`dc:identifier`	Identificador utilizado en la URI del recurso t_{psc}	`dc:identifier` "55900000"^^xsd:string
`dc:subject`	Identificador original proveniente de la fuente de datos	`dc:subject` "55900000-9"^^xsd:string
`pscs:relatedMatch`	URI a un recurso t'_{psc} que encaja parcialmente con t_{psc}	`po:retail`
`skos:exactMatch`	URI a un recurso t'_{psc} que encaja totalmente con t_{psc}	`cpv2008:52900000`
`skos:closeMatch`	URI a un recurso t'_{psc} del CPV 2008 que encaja parcialmente con t_{psc}	`cpv2008:52900000`
`pscs:level`	Indica el grado de especificidad de un t_{psc} cuando no se puede inferir directamente su antecesor	`pscs:level` "5"
`skos:broaderTransitive`	URI a un recurso t'_{psc} antecesor de t_{psc} que indica la categoría Cat_{psc}	`skos:broaderTransitive` cpv2008:55000000
`skosxl:prefLabel`	Etiquetas y descripciones multingües	Retail trade services"@EN
`rdfs:label`	≡	≡
`gr:description`	≡	≡

Tabla 5.13: Diseño de propiedades para los elementos de las Clasificaciones Estándar de Productos.

Una vez especificado el conjunto de propiedades de cada elemento t_{psc} es necesario definir el conjunto de grafos en los cuales se encuadrarán los recursos, es decir, el *dataset* RDF $\mathcal{D}$. Para ello, en la Tabla 5.14 se indican las tuplas $(\mathcal{G}_k, I_k)$ correspondientes a cada uno de los grafos $\mathcal{G}_k$ identificados a través de la URI I_k.

$\mathcal{G}_k$	I_k
$\mathcal{G}$	http://purl.org/weso/pscs
$\mathcal{G}_1$	http://purl.org/weso/pscs/cpv/2008
$\mathcal{G}_2$	http://purl.org/weso/pscs/cpv/2003
$\mathcal{G}_3$	http://purl.org/weso/pscs/cn/2012
$\mathcal{G}_4$	http://purl.org/weso/pscs/cpc/2008
$\mathcal{G}_5$	http://purl.org/weso/pscs/cpa/2008
$\mathcal{G}_6$	http://purl.org/weso/pscs/isic/v4
$\mathcal{G}_7$	http://purl.org/weso/pscs/naics/2007
$\mathcal{G}_8$	http://purl.org/weso/pscs/naics/2012
$\mathcal{G}_9$	http://purl.org/weso/pscs/sitc/v4
$\mathcal{G}_{10}$	http://purl.org/weso/pscs/ontology

Tabla 5.14: *Dataset* RDF $\mathcal{D}$ para Clasificaciones Estándar de Productos.

5.2.1.6 Tarea t_6-Diseño de un Esquema de URIs

Esta tarea tiene como objetivo establecer la forma y estructura de las URIs tanto para las definiciones realizadas en la ontología $\mathcal{O}$ como para todos los recursos presentes en el *dataset* RDF $\mathcal{D}$ que se genera a partir de la transformación de los datos a RDF. Es una de las actividades clave ya que guiará tanto el método final de transformación como el proceso posterior de publicación. En la Tabla 5.15 se hace la descripción de la estructura de URIs que se seguirán para las PSCs.

URI	Descripción	Ejemplo
http://purl.org/weso/pscs/	URI base: <base_uri>	NA
<base_uri>/ontology	Definiciones comunes a todas las PSC	<base_uri>/ontology/PSCConcept
<base_uri>/resource/ds	Descripción del catálogo de las PSCs	<base_uri>/resource/ds
<base_uri>/{psc}/{version|year}	Espacio de nombres para una determinada PSC	<base_uri>/cpv/2008
<base_uri>/{psc}/{version|year}/ontology	Definiciones particulares de una PSC	<base_uri>/cpv/2008/ontology
<base_uri>/resource/{psc}/{version|year}/{id}	URI para un recurso de una PSC	<base_uri>/cpv/2008/resource/55900000
<base_uri>/resource/{psc}/{version|year}/ds	URI la descripción del *dataset* de una PSC	<base_uri>/cpv/2008/resource/ds

Tabla 5.15: Diseño de URIs para las Clasificaciones Estándar de Productos.

5.2.1.7 Tarea t_7-Diseño Plantilla Objetivo del Recurso RDF

El objetivo de esta tarea es establecer una plantilla de cada uno de los recursos RDF que están presentes en el *dataset* RDF $\mathcal{D}$ para que sirvan como guía en los siguientes momentos: 1) en la ejecución propiamente dicha de la transformación de los datos originales a RDF y 2) en la validación de los recursos RDF generados. De esta manera, tratándose de *datasets* con una gran cantidad de recursos se pueden identificar fácilmente aquellos que no sean compatibles con este esquema favoreciendo la depuración de los recursos generados. Adicionalmente, un esquema de recurso sirve como documentación extra para el proceso de consumo. En el caso de las clasificaciones de productos es necesario establecer un recurso plantilla, ver Figura 5.14, para los propios elementos de la PSC, t_{psc}, así como para la propia descripción de los *datasets*. No obstante, esta descripción se delega a la tarea t_{16}-"Añadir metainformación a los recursos RDF", ver Sección 5.2.1.11.

De acuerdo al recurso plantilla y a las definiciones realizadas en la ontología que modela estos datos es posible realizar una validación en cuanto a los tipos de datos, cardinalidad de las relaciones, tipo de objetos, etc., que resulta de sumo interés para asegurar la calidad de los datos producidos.

```
<<base_uri>/resource/{psc}/{version|year}/{id}>
      a        gr:ProductOrServiceModel , pscs:PSCConcept;
     (skosxl:prefLabel ""@lang;)+
     (gr:description ""@lang;)+
     (rdfs:label ""@lang ;)+
     dc:identifier ""^^xsd:string ;
     dc:subject ""^^xsd:string ;
     (pscs:relatedMatch <uri>;)*
     (pscs:level <uri>;)*
     (skos:broaderTransitive <uri>;)[0,1]
     (skos:exactMatch <uri>; )*
     skos:inScheme <<base_uri>/resource/{psc}/{version|year}/ds> .
```

Figura 5.14: Plantilla Objetivo de un Recurso de las Clasificaciones Estándar de Productos.

5.2.1.8 Tarea t_8-Enriquecimiento de los datos en RDF

El objetivo de esta tarea es enlazar los recursos generados del *dataset* RDF $\mathcal{D}$ con otros ya existentes. En el caso particular de las clasificaciones de productos y de acuerdo a los *datasets* identificados en la Tarea t_4, el enriquecimiento de los elementos se realizará con *ProductOntology* que a su vez dispone de enlaces a la DBPedia, con lo que el enlazado de los datos queda perfectamente justificado.

Para llevar a cabo el enriquecimiento es necesario realizar reconciliación de entidades entre las descripciones de elementos t_{psc} y los recursos objetivos. Análogamente a enfoques como los de *Silk Server* o SERIMI y teniendo en cuenta las particularidades de las descripciones de los productos se ha decido implementar un componente, como parte de `moldeas-transformer`, para realizar este enlazado, tomando como entrada la descripción del producto, en inglés, y tras la realización del procesamiento de lenguaje natural mediante las herramientas Apache Lucene [153] y Solr [120], obtiene una lista de recursos candidatos a ser enlazados. Dado que el *matching* de los recursos no se puede asegurar al 100 %, requeriría una validación manual, por lo que se ha creado una propiedad particular *pscs:relatedMatch* para indicar la relación que se establece entre el recurso actual t_{psc} y el recurso obtenido. Adicionalmente, en una segunda versión se ha especializado esta propiedad para incluir un valor de fiabilidad (*threshold*) que indica el grado de similitud entre los recursos, de esta manera se pueden obtener los recursos similares a uno dado por encima de un valor umbral.

Por otra parte, en el ámbito de los anuncios de licitación la clasificación de productos más rele-

vante es el CPV 2008 ya que es el sistema de clasificación vigente y normativo en la Unión Europea. Teniendo en cuenta que los anuncios de licitación objeto de estudio en este documento pertenecen a la Unión Europea y que por lo tanto están clasificados de acuerdo a esta taxonomía, es conveniente proveer los enlaces adecuados entre los distintos elementos de todas las PSCs para que puedan ser traducidos al CPV 2008 facilitando la capacidad expresiva para la realización de consultas y en consecuencia favoreciendo la accesibilidad a los anuncios de licitación. El soporte a este enfoque sigue un modelo similar al presentado anteriormente, con la diferencia de que los recursos origen son aquellos generados en la transformación de las PSCs y el *dataset* objetivo es el correspondiente al CPV 2008. De esta manera, se establecen dos tipos de enlaces:

1. *Mapping* exacto. Este enlace entre elementos de las PSCs se realiza cuando en las propias clasificaciones existen *mapeos* entre los elementos porque un organismo oficial se ha encargado de realizar esta tarea por cuestiones estratégicas. En este caso, el enlace entre los conceptos se modela mediante la propiedad `skos:exactMatch` cuya semántica dentro de la especificación de SKOS indica que su uso estará destinado a realizar enlaces entre conceptos *skos:Concept*, en los cuales se puede asegurar que representan al mismo recurso.
2. *Mapping* parcial. Al igual que en el caso anterior el proceso consiste en identificar los recursos del CPV 2008 que encajan parcialmente con un recurso origen en otra PSC. En este caso, la propiedad utilizada es `skos:closeMatch` que de acuerdo a su semántica permite establecer enlaces entre conceptos *skos:Concept*. Igualmente, se ha especializado este caso para contener un valor umbral del enlace.

En conclusión, la ejecución de esta tarea permite por un lado crear enlaces entre todas las PSCs y un *dataset* externo y por otra parte, crear enlaces entre todas las PSCs y el CPV 2008. El enriquecimiento realizado en las PSCs conduce finalmente a la siguiente estructura, ver Figura 5.15, de enlaces entre las distintas PSCs y los *datasets* externos.

5.2.1.9 Tarea t_9-Transformación de los datos a RDF

Una vez realizadas las tareas anteriores se está en disposición de realizar la transformación de los datos de entrada a RDF. En esta tarea el punto clave de decisión reside en seleccionar bien una herramienta ya disponible o bien implementar un programa que ejecute las reglas de transformación tomando como entrada los datos de las clasificaciones. En el caso objeto de estudio se ha optado por un enfoque híbrido realizando la transformación inicial mediante la herramienta Google Refine [178] y su extensión [208] para trabajar con RDF y las posteriores etapas de enriquecimiento a través de una implementación particular que tenga en cuenta la casuística específica de las clasificaciones de productos. En cualquier caso, en la ejecución de esta tarea se debe asegurar que el *dataset* RDF $\mathcal{D}$ generado es válido en cuanto a sintaxis y al modelo definido. Esta tarea se ha ejecutado secuencialmente para cada unas de las clasificaciones de productos.

5.2.1.10 Tarea t_{10}-Reconciliación de Entidades

El diseño y ejecución de esta tarea ha sido descrito en la Sección 5.2.1.8 ya que están estrechamente ligadas.

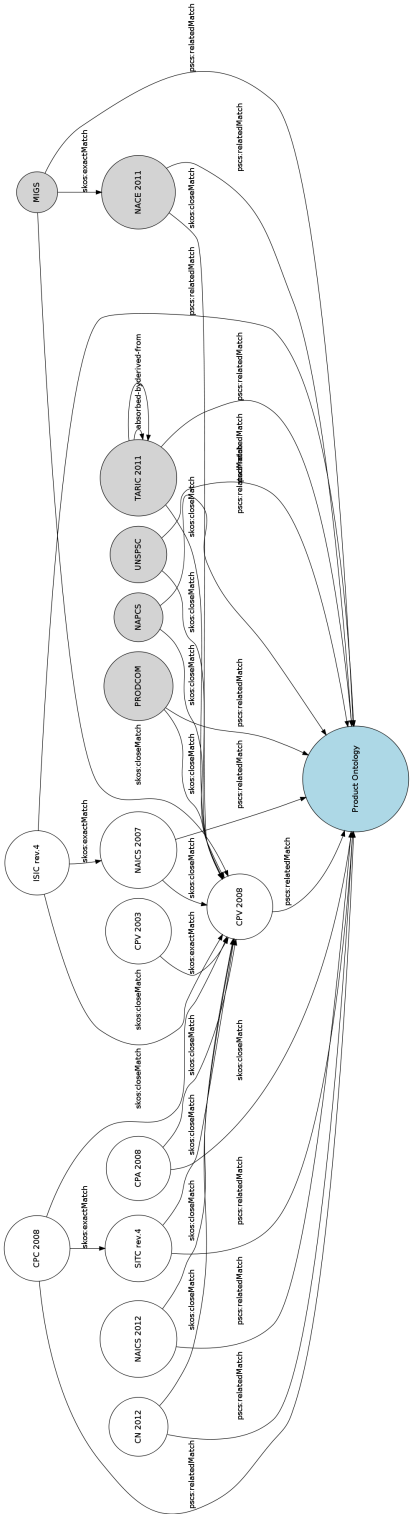


Figura 5.15: Enlaces entre las distintas Clasificaciones de Productos.

5.2.1.11 Tarea t_{16}-Añadir metainformación a los recursos RDF

En el caso de las clasificaciones de productos la metainformación en los recursos se puede interpretar en los dos sentidos que se han descrito en la Sección 4.4.16, bien en cada uno de los recursos generados o en el *dataset* completo. Considerando el carácter estático de las clasificaciones de productos en cuanto a su creación, depende de los organismos oficiales que se encargan de su mantenimiento, y dado que para los elementos que varían en el tiempo éstos generan un *mapeo* directo entre las versiones de los mismos, como en el caso del CPV 2003 y 2008, se considera una elección correcta situar la metainformación a nivel de *dataset* para cada unas de las PSCs. Para ello y de acuerdo a los vocabularios seleccionados se utiliza *voID* como especificación para indicar la metainformación de un *dataset* concreto. El enfoque seguido genera metainformación a nivel del catálogo de las clasificaciones de producto, ver Figura 5.16, y para cada una de las clasificaciones transformadas, ver Figura 5.17 como ejemplo de la descripción del *dataset* del CPV 2008. Es relevante destacar la definición de la licencia de los datos con el objetivo de facilitar su posterior reutilización, en este caso se ha optado por una licencia de "*Open Data*" basada en las directrices fijadas en la guía provista en [198].

```
<http://purl.org/weso/pscs/data/resource/ds?output=ttl>
      rdfs:label "RDF description of Product Scheme Classifications" ;
      foaf:primaryTopic <http://purl.org/weso/pscs/resource/ds> .

<http://purl.org/weso/pscs/resource/ds>
      a    <http://rdfs.org/ns/void#Linkset> ;
      rdfs:label "Product Scheme Classifications"@en ;
      dcterms:author
            <http://www.di.uniovi.es/~labra/labraFoaf.rdf#me> ,
            <http://www.josemalvarez.es/foaf.rdf#me> ;
      dcterms:contributor
            <http://purl.org/weso/pscs/resource/10ders> ,
            <http://rdfohloh.wikier.org/project/moldeas/rdf> ;
      dcterms:description
            "Some Product Scheme Classifications available in RDF" ;
      dcterms:license
            <http://opendatacommons.org/licenses/by/1.0/> ;
      dcterms:modified
            "2011-11-10"^^<http://www.w3.org/2001/XMLSchema#date> ;
      dcterms:publisher
            <http://www.josemalvarez.es/foaf.rdf#me> ;
       dcterms:title
            "Product Scheme Classifications" ;
      void:target
            <http://purl.org/weso/pscs/cpc/2008/resource/ds> ,
            <http://purl.org/weso/pscs/naics/2007/resource/ds> ,
            <http://purl.org/weso/pscs/sitc/v4/resource/ds> ,
            <http://purl.org/weso/pscs/cpv/2003/resource/ds> ,
            <http://purl.org/weso/pscs/naics/2012/resource/ds> ,
            <http://purl.org/weso/pscs/cpv/2008/resource/ds> ,
            <http://purl.org/weso/pscs/cpa/2008/resource/ds> ,
            <http://purl.org/weso/pscs/cn/2012/resource/ds> ,
            <http://purl.org/weso/pscs/isic/v4/resource/ds> ;
      foaf:homepage <http://purl.org/weso> .
```

Figura 5.16: Descripción del *Linkset* de las Clasificaciones Estándar de Productos.

```
<http://purl.org/weso/pscs/cpv/2008/resource/ds>
      a        void:Dataset , skos:ConceptScheme ;
      rdfs:label "CPV 2008"@en ;
      dcterms:author
            <http://www.di.uniovi.es/~labra/labraFoaf.rdf#me> ,
            <http://www.josemalvarez.es/foaf.rdf#me> ;
      dcterms:contributor
            <http://purl.org/weso/pscs/resource/10ders> ,
            <http://rdfohloh.wikier.org/project/moldeas/rdf> ;
      dcterms:description "Common Procurement Vocabulary" ;
      dcterms:license <http://opendatacommons.org/licenses/by/1.0/> ;
      dcterms:modified "2011-11-10"^^xsd:date ;
      dcterms:publisher <http://www.josemalvarez.es/foaf.rdf#me> ;
      dcterms:source
        <http://europa.eu/legislation_summaries/internal_market/businesses/
            public_procurement/l22008_en.htm> ;
      dcterms:title "CPV 2008" ;
      void:dataDump <http://purl.org/weso/pscs/cpv/2008/cpv-2008.ttl> ;
      void:exampleResource
        <http://purl.org/weso/pscs/cpv/2008/resource/18000000> ,
        <http://purl.org/weso/pscs/cpv/2008/resource/45000000> ,
        <http://purl.org/weso/pscs/cpv/2008/resource/33000000> ;
      void:uriRegexPattern
        "http://purl.org/weso/pscs/cpv/2008/resource/.+" ;
      void:vocabulary skosxl: , skos: , gr: ;
      skos:hasTopConcept
        <http://purl.org/weso/pscs/cpv/2008/resource/63000000> ,
        <http://purl.org/weso/pscs/cpv/2008/resource/76000000> ,
        <http://purl.org/weso/pscs/cpv/2008/resource/19000000> ,
        <http://purl.org/weso/pscs/cpv/2008/resource/79000000> ,
        ...
        <http://purl.org/weso/pscs/cpv/2008/resource/16000000> ;
      foaf:homepage <http://purl.org/weso> .
```

Figura 5.17: Descripción del *dataset* CPV 2008.

5.2.1.12 Resultado Final y Ejemplos

El resultado final del proceso de producción de *Linked Data*, tras el análisis y ejecución de las tareas identificadas y del método de producción seleccionado, genera como resultado un catálogo de clasificaciones de productos mediante datos enlazados, en los cuales se pueden extraer las siguientes estadísticas de producción de datos así como ejemplos de los recursos generados, ver Tabla 5.16. Por otra parte, el aumento de la expresividad en el momento de realizar consultas se puede observar en la Figura 5.18 en la que se expresa la siguiente consulta:

> *"Dame 100 productos o servicios relacionados, descripción en inglés, con el término construcción que aparezca en cualquiera de los catálogos disponibles y que tengan enlaces con productos o servicios disponibles en la CPV 2008."*

Clasificación	**Nº de Elementos**	**Ejemplo**	**Tripletas**	**Enlaces externos**	**Enlaces CPV 2008**
CPV 2003	8323	Figura 5.19	546135	8322	462 (del CPV 2008 al 2003)
CPV 2008	10357	Figura 5.20	803311	10355	N/A
CN 2012	14552	Figura 5.21	137484	2590	2390
CPC 2008	4408	Figura 5.22	100819	4408	4375 y 1503 (exactos)
CPA 2008	5429	Figura 5.23	92749	5429	5399
ISIC v4	766	Figura 5.24	18986	766	765
NAICS 2007	2328	Figura 5.25	36292	2328	2300
NAICS 2012	2212	Figura 5.26	35390	2212	2186
SITC v4	4017	Figura 5.27	70887	3941	3811
Catálogo de Clasificaciones Estándar de Productos (total)					
PSCs	52392	N/A	1842053	40351	23191

Tabla 5.16: Estadísticas y Ejemplos del Catálogo de Clasificaciones Estándar de Productos seleccionadas.

```
SELECT DISTINCT * WHERE{
  ?product pscs:relatedMatch <http://www.productontology.org/id/construction> .
  ?product skos:closeMatch ?cpv.
  ?product skosxl:prefLabel ?productLabel.
  ?cpv skosxl:prefLabel ?cpvLabel.
  ?product skos:inScheme ?scheme.
  FILTER (?scheme != <http://purl.org/weso/pscs/cpv/2008/resource/ds>).
  FILTER (lang(?cpvLabel)="en")
} LIMIT 100
```

Figura 5.18: Ejemplo de consulta en SPARQL sobre el Catálogo de Clasificaciones de Productos.

```
<http://purl.org/weso/pscs/cpv/2003/resource/52900000>
     a        gr:ProductOrServiceModel , cpv-def:Group ;
     skosxl:prefLabel
     gr:description
     rdfs:label
       "Detaljhandel med medicinska produkter"@SV ,
       "Detailhandel med medicinalvarer"@DA ,
       "Verkoopdiensten op de detailhandelniveau van medische producten"@NL ,
       "Storitve trgovine na drobno z medicinskimi izdelki"@SL ,
       "Servizi di vendita al dettaglio di prodotti medici"@IT ,
       "Retail trade services of medical products"@EN ,
       "Einzelhandel mit medizinischen Erzeugnissen"@DE ,
       ...
     dc:identifier "52900000" ;
     dc:subject "52900000-8" ;
     <http://purl.org/weso/pscs/ontology/relatedMatch>
       <http://www.productontology.org/id/medical> ,
       <http://www.productontology.org/id/retail> ,
       <http://www.productontology.org/id/product> ,
       <http://www.productontology.org/id/service> ,
       <http://www.productontology.org/id/trade> ;
     skos:broaderTransitive
            <http://purl.org/weso/pscs/cpv/2003/resource/52000000> ;
     skos:inScheme "http://purl.org/weso/pscs/cpv/2003/resource/ds" .
```

Figura 5.19: Ejemplo final de un Recurso del CPV 2003.

```
<http://purl.org/weso/pscs/cpv/2008/resource/55900000>
      a       gr:ProductOrServiceModel , cpv-def:Group ;
     skosxl:prefLabel
     gr:description
     rdfs:label
        "Servizi di vendita al dettaglio"@IT ,
        "Detailhandel"@DA ,
        "Detailhandelsdiensten"@NL ,
        "Servicios comerciales al por menor"@ES ,
        "Maloobchod"@CS ,
        "Storitve trgovine na drobno"@SL ,
        "Retail trade services"@GA ,
        "Retail trade services"@EN ;
        ...
      dc:identifier "55900000"^^xsd:string ;
      dc:subject "55900000-9"^^xsd:string ;
      <http://purl.org/weso/pscs/ontology/relatedMatch>
        <http://www.productontology.org/id/retail> ,
        <http://www.productontology.org/id/service> ,
        <http://www.productontology.org/id/trade> ;
      skos:broaderTransitive
         <http://purl.org/weso/pscs/cpv/2008/resource/55000000> ;
      skos:exactMatch
        <http://purl.org/weso/pscs/cpv/2003/resource/52900000> ,
        <http://purl.org/weso/pscs/cpv/2003/resource/52800000> ,
        <http://purl.org/weso/pscs/cpv/2003/resource/52400000> ,
        <http://purl.org/weso/pscs/cpv/2003/resource/52600000> ,
        <http://purl.org/weso/pscs/cpv/2003/resource/52700000> ,
        <http://purl.org/weso/pscs/cpv/2003/resource/52300000> ,
        <http://purl.org/weso/pscs/cpv/2003/resource/52500000> ,
        <http://purl.org/weso/pscs/cpv/2003/resource/52200000> ,
        <http://purl.org/weso/pscs/cpv/2003/resource/52100000> ;
      skos:inScheme <http://purl.org/weso/pscs/cpv/2008/resource/ds>  .
```

Figura 5.20: Ejemplo final de un Recurso del CPV 2008.

```
<http://purl.org/weso/pscs/cn/2012/resource/846721100080>
      a       <http://purl.org/weso/pscs/ontology/PSCConcept> ,
      gr:ProductOrServiceModel ;
      skosxl:prefLabel
      gr:description
      rdfs:label "Nombre de pi\'{e}ces"@fr , " Anzahl Stuck"@de , " Number of items"
          @en ;
      dc:identifier "846721100080" ;
      dc:subject "846721100080" ;
      <http://purl.org/weso/pscs/definitions/level>
              "6"^^xsd:int ;
      <http://purl.org/weso/pscs/ontology/relatedMatch>
        <http://www.productontology.org/id/number> ,
        <http://www.productontology.org/id/item> ;
      skos:closeMatch
        <http://purl.org/weso/pscs/cpv/2008/resource/GC25> ,
        <http://purl.org/weso/pscs/cpv/2008/resource/GC27> ,
        <http://purl.org/weso/pscs/cpv/2008/resource/GC31> ;
      skos:inScheme <http://purl.org/weso/pscs/cn/2012/resource/ds> ;
      skos:notation "p/st"@en ;
      skos:note "8467 21 10" .
```

Figura 5.21: Ejemplo final de un Recurso de CN 2012.

```
<http://purl.org/weso/pscs/cpc/2008/resource/62294>
      a       <http://purl.org/weso/pscs/ontology/PSCConcept> , gr:
          ProductOrServiceModel ;
      skosxl:prefLabel
      gr:description
      rdfs:label "Specialized store retail trade services, of paper and paperboard"
          @en ;
      dc:identifier "62294" ;
      dc:subject "62294" ;
      <http://purl.org/weso/pscs/ontology/level>
              "5" ;
      <http://purl.org/weso/pscs/ontology/relatedMatch>
        <http://www.productontology.org/id/retail> ,
        <http://www.productontology.org/id/paperboard> ,
        <http://www.productontology.org/id/trade> ,
        <http://www.productontology.org/id/service> ,
        <http://www.productontology.org/id/specialized> ,
        <http://www.productontology.org/id/store> ,
        <http://www.productontology.org/id/paper> ;
      skos:broaderTransitive
              <http://purl.org/weso/pscs/cpc/2008/resource/6229> ;
      skos:closeMatch
        <http://purl.org/weso/pscs/cpv/2008/resource/55000000> ,
        <http://purl.org/weso/pscs/cpv/2008/resource/55900000> ,
        <http://purl.org/weso/pscs/cpv/2008/resource/51544000> ;
      skos:inScheme <http://purl.org/weso/pscs/cpc/2008/resource/ds> ;
```

Figura 5.22: Ejemplo final de un Recurso de CPC 2008.

```
<http://purl.org/weso/pscs/cpa/2008/resource/470054>
      a  <http://purl.org/weso/pscs/ontology/PSCConcept> ,
        gr:ProductOrServiceModel ;
      dc:identifier "470054" ;
      dc:subject "47.00.54" ;
      skosxl:prefLabel
      gr:description
      rdfs:label "Retail trade services of electrical household appliances"@en ;
      <http://purl.org/weso/pscs/ontology/relatedMatch>
         <http://www.productontology.org/id/retail> ,
         <http://www.productontology.org/id/household> ,
         <http://www.productontology.org/id/service> ,
         <http://www.productontology.org/id/trade> ,
         <http://www.productontology.org/id/appliance> ,
         <http://www.productontology.org/id/electrical> ;
      skos:broaderTransitive
        <http://purl.org/weso/pscs/cpa/2008/resource/47005> ;
      skos:closeMatch
        <http://purl.org/weso/pscs/cpv/2008/resource/39711000> ,
        <http://purl.org/weso/pscs/cpv/2008/resource/55000000> ,
        <http://purl.org/weso/pscs/cpv/2008/resource/55900000> ;
      skos:inScheme <http://purl.org/weso/pscs/cpa/2008/resource/ds> .
```

Figura 5.23: Ejemplo final de un Recurso de CPA 2008.

```
<http://purl.org/weso/pscs/isic/v4/resource/479>
      a        <http://purl.org/weso/pscs/ontology/PSCConcept> ,
        gr:ProductOrServiceModel ;
      skosxl:prefLabel
      gr:description
      rdfs:label "Retail trade not in stores, stalls or markets"@en ,
        "Venta al por menor no realizada en comercios, puestos de venta o mercados"
            @es ;
      dc:identifier "479" ;
      dc:subject "479" ;
       "Retail trade not in stores, stalls or markets"@en , "Venta al por menor no
           realizada en comercios, puestos de venta o mercados"@es ;
      <http://purl.org/weso/pscs/ontology/relatedMatch>
        <http://www.productontology.org/id/retail> ,
        <http://www.productontology.org/id/trade> ,
        <http://www.productontology.org/id/market> ,
        <http://www.productontology.org/id/stall> ,
        <http://www.productontology.org/id/store> ;
      skos:broaderTransitive
        <http://purl.org/weso/pscs/isic/v4/resource/47> ;
      skos:closeMatch <http://purl.org/weso/pscs/cpv/2008/resource/55000000> ,
        <http://purl.org/weso/pscs/cpv/2008/resource/55900000> ,
        <http://purl.org/weso/pscs/cpv/2008/resource/45213260> ;
      skos:inScheme <http://purl.org/weso/pscs/isic/v4/resource/ds> .
```

Figura 5.24: Ejemplo final de un Recurso de ISIC rev4.

```
<http://purl.org/weso/pscs/naics/2007/resource/311811>
      a <http://purl.org/weso/pscs/ontology/PSCConcept> ,
        gr:ProductOrServiceModel ;
      gr:description
      skosxl:prefLabel
      rdfs:label "Retail Bakeries"@en ;
      dc:identifier "311811" ;
      dc:subject "311811" ;
      <http://purl.org/weso/pscs/ontology/relatedMatch>
        <http://www.productontology.org/id/retail> ,
        <http://www.productontology.org/id/bakery> ;
      skos:broaderTransitive
        <http://purl.org/weso/pscs/naics/2007/resource/31181> ;
      skos:closeMatch
        <http://purl.org/weso/pscs/cpv/2008/resource/55000000> ,
        <http://purl.org/weso/pscs/cpv/2008/resource/55900000> ,
        <http://purl.org/weso/pscs/cpv/2008/resource/15612500> ;
      skos:inScheme <http://purl.org/weso/pscs/naics/2007/resource/ds> .
```

Figura 5.25: Ejemplo final de un Recurso de NAICS 2007.

```
<http://purl.org/weso/pscs/naics/2012/resource/311811>
      a  <http://purl.org/weso/pscs/ontology/PSCConcept> ,
        gr:ProductOrServiceModel ;
      skosxl:prefLabel
      gr:description
      rdfs:label "Retail Bakeries"@en ;
      dc:identifier "311811" ;
      dc:subject "311811" ;
      <http://purl.org/weso/pscs/ontology/relatedMatch>
        <http://www.productontology.org/id/retail> ,
        <http://www.productontology.org/id/bakery> ;
      skos:broaderTransitive
         <http://purl.org/weso/pscs/naics/2012/resource/31181> ;
      skos:closeMatch
        <http://purl.org/weso/pscs/cpv/2008/resource/55000000> ,
        <http://purl.org/weso/pscs/cpv/2008/resource/55900000> ,
        <http://purl.org/weso/pscs/cpv/2008/resource/15612500> ;
      skos:inScheme <http://purl.org/weso/pscs/naics/2012/resource/ds>.
```

Figura 5.26: Ejemplo final de un Recurso de NAICS 2012.

```
<http://purl.org/weso/pscs/sitc/v4/resource/08195>
      a       <http://purl.org/weso/pscs/ontology/PSCConcept> ,
        gr:ProductOrServiceModel;
      skosxl:prefLabel
      gr:description
      rdfs:label "Dog or cat food, put up for retail sale"@en ;
      dc:identifier "08195" ;
      dc:subject "081.95" ;
      gr:description "Dog or cat food, put up for retail sale"@en ;
      <http://purl.org/weso/pscs/ontology/relatedMatch>
        <http://www.productontology.org/id/retail> ,
        <http://www.productontology.org/id/sale> ,
        <http://www.productontology.org/id/cat> ,
        <http://www.productontology.org/id/dog> ,
        <http://www.productontology.org/id/food> ;
      skos:broaderTransitive
        <http://purl.org/weso/pscs/sitc/v4/resource/0819> ;
      skos:closeMatch <http://purl.org/weso/pscs/cpv/2008/resource/33115200> ,
        <http://purl.org/weso/pscs/cpv/2008/resource/55900000> ,
        <http://purl.org/weso/pscs/cpv/2008/resource/QB48> ;
      skos:inScheme <http://purl.org/weso/pscs/sitc/v4/resource/ds> .
```

Figura 5.27: Ejemplo final de un Recurso de SITC v4.

5.2.1.13 Método de Producción de *Linked Data* de Clasificaciones Estándar de Productos

De acuerdo al análisis y diseño de datos enlazados realizado para las clasificaciones de productos a lo largo de las anteriores anteriores y la tabla de decisión 4.12, el método semántico seleccionado para realizar la producción de datos enlazados es el SPM_1-"Transformación de datos a RDF", ver Sección 4.5.2, en el se transforman un conjunto de datos de entrada $\mathcal{G}$ a un *dataset* RDF $\mathcal{D}$. Según la definición de método semántico de producción, realizada en la Sección 4.5.1, y el estudio de las clasificaciones de productos se pueden establecer los siguientes conjuntos:

- $\mathcal{G}$ es el *dataset* de entrada, conjunto de tuplas, conteniendo los datos de cada una de las clasificaciones de productos.
- $\mathcal{M}$ es el conjunto de *mapeos*, ver Tabla 5.17, extraídos según el análisis y diseño realizado en las secciones anteriores. Estos *mapeos* son directamente expresables en la herramienta de transformación y toman como parámetros el valor de una de las tuplas de entrada (posición X) y la propiedad a generar.
- *Dataset* RDF $\mathcal{D}$ es el *dataset* resultado, siguiendo el análisis y diseño realizado en las secciones anteriores y tras la ejecución de la tarea propia de transformación de datos.

$\mathcal{M}$	**Propiedad**	**Valor**
m_1	`rdf:type`	URI
m_2	`skos:inScheme`	URI
m_3	`dc:identifier`	`xsd:string`
m_4	`dc:subject`	`xsd:string`
m_5	`pscs:relatedMatch`	URI
m_6	`skos:exactMatch`	URI
m_7	`skos:closeMatch`	URI
m_8	`pscs:level`	`xds:int`
m_9	`skos:broaderTransitive`	URI
m_{10}	`skosxl:prefLabel`, `rdfs:label`, `gr:description`	`xsd:string@lang`

Tabla 5.17: Conjunto de *mapeos* $\mathcal{M}$ para las Clasificaciones Estándar de Productos.

5.2.2 Proceso de Publicación de *Linked Data* de Clasificaciones Estándar de Productos

Considerando que la estrategia definida para todos los *datasets* implicados en el proceso de contratación pública electrónica es común, el proceso de publicación es homogéneo siguiendo la estructura definida en la Sección 5.2.2.

5.2.2.1 Tarea t_{14}-Infraestructura para *Linked Data*

Nuevamente, la estrategia definida para todos los *datasets* implicados en el proceso de contratación pública electrónica es común, por ello la infraestructura utilizada es la misma que la definida en la Sección 5.1.2.1.

5.2.2.2 Tarea t_{15}-Acceso y formato en datos RDF

De la misma forma que en el apartado anterior, el acceso y formato de datos RDF para los datos contenidos y relativos a las organizaciones siguen el esquema proporcionado en la Sección 5.1.2.2.

5.2.3 Proceso de Consumo de Clasificaciones Estándar de Productos

El proceso de consumo de datos enlazados, según la definición realizada en la Sección 4.7.1, consiste en la reutilización de los datos enlazados para ser aplicados en la construcción de una nueva aplicación o servicio de valor añadido. En general, la reutilización más sencilla consiste en la representación gráfica de los recursos o la simple consulta con selección de formato de datos de acuerdo a las características de publicación utilizadas. En el caso que nos ocupa y teniendo en cuenta el objetivo de realización de un prototipo experimental de extracción de anuncios de licitación como demostrador del consumo de datos enlazados, se ha escogido el método semántico de consumo SCM_2-"*Mapeo* a Lenguaje de Programación", cuya descripción está disponible en la Sección 4.7.3, orientado a obtener una representación de los recursos RDF en un lenguaje de programación (en este caso Java) como objetos de negocio. De acuerdo a este objetivo y la definición del propio método es necesario definir:

- El *dataset* RDF $\mathcal{D}_{pub}$, es el conjunto de datos disponible tras aplicar el método de publicación.
- El conjunto $\mathcal{M}^1$, ver Tabla 5.18, indica como transformar el *dataset* anterior a la representación objetivo, objetos del lenguaje Java.

De esta manera, se obtiene una serie de objetos, $\mathcal{D}_{consum}$, con la información y datos necesarios, no se transforman necesariamente todos los datos disponibles en los recursos pertenecientes a $\mathcal{D}_{pub}$, para ser reutilizados como objetos de negocio en un lenguaje de programación. Es conveniente señalar que el acceso a los datos se realiza a través de la consulta al *endpoint* de SPARQL ejecutando consultas *SELECT* y *DESCRIBE*.

$\mathcal{M}^1$	**Propiedad**	**Tipo en Java**
m_1^1	URI recurso	`java.lang.String`
m_2^1	`rdf:type`	`org.weso.moldeas.to.PSCTO`
m_3^1	`skos:inScheme`	`java.lang.String`
m_4^1	`dc:identifier`	`java.lang.String`
m_5^1	`dc:subject`	`java.lang.String`
m_6^1	`pscs:relatedMatch`	`java.util.List<PSCTO>`
m_7^1	`skos:exactMatch`	`java.util.List<PSCTO>`
m_8^1	`skos:closeMatch`	`java.util.List<PSCTO>`
m_9^1	`pscs:level`	int
m_{10}^1	`skos:broaderTransitive`	java.util.List<PSCTO>
m_{11}^1	`skosxl:prefLabel`, `rdfs:label`, `gr:description`	Map<String,String>(lang, value) para cada propiedad

Tabla 5.18: Conjunto de *mapeos* $\mathcal{M}^1$ de consumo para las Clasificaciones Estándar de Productos.

5.2.4 Proceso de Validación de Clasificaciones Estándar de Productos

La validación como proceso transversal en cualquier etapa dentro del ciclo de vida de datos enlazados debe realizarse con el objetivo de asegurar la calidad de los datos. De acuerdo a la definición realizada en la Sección 4.8, este proceso consiste en la comprobación de que los recursos de un *dataset* RDF cumplen ciertas características.

La realización de esta validación puede ser realizada manual o automáticamente en función del caso concreto, por ejemplo para la característica de negociación de contenido o para la inclusión en

la nube de datos enlazados se dispone de herramientas adecuadas, pero en cambio para comprobaciones relativas a los dominios y rangos de las propiedades, etc., no existe una herramienta completa. Por todo ello, se ha seguido un enfoque híbrido basado en la utilización de herramientas y validación manual. La descripción completa de la validación de acuerdo a todas las características se reseña en las Tablas de Validación disponibles en el Apéndice C.

5.2.4.1 Tarea t_{12}-Validación de Recursos RDF

Siguiendo con la definición realizada de esta tarea en la Sección 4.4.12, se puede asegurar que la transformación realizada de los catálogos de clasificaciones de productos a la iniciativa *Linked Data* cumple estrictamente los siguientes puntos:

- Los datos RDF son correctos ya que se han utilizado herramientas y APIs (Google Refine y Jena) que aseguran la generación correcta de RDF.
- El dominio y rango en las propiedades es correcta, ya que se realiza la validación contra el modelo definido.
- Se ha establecido metainformación sobre la procedencia a nivel de *dataset*.
- Todos los recursos transformados siguen la plantilla objetivo RDF.

5.2.5 Proceso de Realimentación de Clasificaciones Estándar de Productos

Este proceso según la definición realizada en la Sección 4.9, busca la mejora y perfeccionamiento de los datos promocionados a RDF. Esta situación emerge en el momento en el cual los datos comienzan a ser reutilizados, tanto por aplicaciones o servicios como por individuos. En el caso particular del catálogo de las clasificaciones de productos y debido a su reutilización por la *Charles University* de la República Checa se han detectado errores que han sido corregidos, se trata de casos específicos no detectados por el proceso de validación, debido al procesamiento masivo de datos. En este sentido el proceso de realimentación se basa en un enfoque en el cual *Usuarios y Aplicaciones* realizan el descubrimiento de datos no correctos y cuyo cambio se centra en una *Actualización Ocasional*. Aunque este enfoque es completamente válido, es conveniente señalar que este proceso todavía no se ha definido de forma completa dentro de la iniciativa de *Linked Data* por lo que su ejecución automática constituye un objetivo a medio plazo en el que terceros puedan mejorar los datos previa validación del *Propietario de datos*.

5.3 Organizaciones

La información de las organizaciones se considera un punto clave para la realización de muchos servicios de valor añadido tanto para la Administración Pública como para otras entidades. El conocimiento del estado financiero de una entidad, la actividad que ha desarrollado a través de proyectos, los productos y servicios ofertados, el listado de clientes y socios, etc., se considera una información de alto interés para la monitorización y la selección de posibles entidades para el establecimiento de relaciones de negocios. Desde el punto de vista interno de la propia organización, su estructura y personas implicadas en el desarrollo de la actividad de negocio así como la explotación del conocimiento organizacional son factores clave que determinan el camino a seguir por una entidad y su valor como tal. Como se ha señalado en la Sección 3.4.1.4, existen varios enfoques para

la representación de la información de las organizaciones, en el caso objeto de estudio de este documento se centrará principalmente en los siguientes puntos: 1) información de contacto de la organización respecto a su localización y 2) persona de referencia para el establecimiento de relaciones. Este planteamiento es válido tanto para la representación de la información de los organismos que, cuentan con capacidad para la realización de contratos públicos, como para aquellos que deseen concursar a través de ofertas y cuya información debe perdurar a lo largo del tiempo hasta el fin del contrato. De acuerdo a la información disponible en las distintas fuentes de información de los anuncios de licitación las organizaciones o entidades modeladas cumplirán, al menos, con estos dos puntos clave que suponen un gran avance respecto a la información actual.

5.3.1 Proceso de Producción de *Linked Data* de Organizaciones

Al igual que en el caso de las clasificaciones de productos y según la definición realizada de este proceso en la Sección 4.5, este proceso implica todas las tareas que se han de realizar para la transformación de un *dataset* de entrada $\mathcal{G}$, mediante unas reglas de *mapeo* $\mathcal{M}$, con el objetivo de conseguir un *dataset* RDF $\mathcal{D}$, de esta forma se define el método semántico de producción.

5.3.1.1 Tarea t_1-Análisis del *dataset* a transformar

Las organizaciones, tanto órganos contrantes como ofertantes, durante el proceso administrativo de contratación pública generan información y datos abarcando un amplio espectro, sin embargo para los fines de apertura y enlazado de datos mediante un modelo formal se considera suficiente la representación de la información de contacto de una entidad, incluyendo tanto a la persona física o jurídica de la misma con especial atención a su localización. La realidad es que la obtención de la información y datos pormenorizados de una organización se haya disponible públicamente, pero a través de fuentes información de difícil consulta que carecen de relevancia crítica para el estudio realizado en este trabajo, por ello se ha decidido realizar un primer acercamiento a la información empresarial focalizando el esfuerzo en la selección de un modelo formal que suministre el marco de trabajo adecuado para la representación de información de las organizaciones de una forma sencilla y operativa en el ámbito de la contratación pública electrónica. En este sentido, emergen tres entidades: organizaciones, personas y países, en las que hay que representar distintos datos, principalmente en las dos primeras información relativa a los datos de contacto y en la última la geolocalización. Una organización puede tener diferentes identificadores, asumiendo que cada uno de sus departamentos o subdivisiones y también las personas varían de acuerdo a la unidad mínima de organización seleccionada y en consecuencia su localización. La invariante que se produce reside en que para un proceso de contratación pública, la tupla una organización (entidad, departamento, división, etc.), una persona de contacto, un país y un propósito determinado es única.

En este sentido, surge la necesidad de especificar las organizaciones, personas, direcciones y países en un modelo formal. Para la utilización de un modelo formal, extensible que suministre el soporte necesario para el almacenamiento de la información y datos de las entidades con carácter multilingüe, se ha seleccionado la ontología "*Organizations*" realizada por *Dave Reynolds* (Epimorphics Ltd) ya que conjuga los esfuerzos de anteriores propuestas facilitando la representación de cualquier tipo de entidad desde un punto de vista estructural, información de contacto, actividad desarrollada, etc., convirtiéndose de esta manera en el marco de trabajo apropiado para el tratamiento de la información de licitantes y licitadores. La información sobre las personas se representa con la ontología FOAF, mientras que finalmente los países y su información, además de utilizar el *dataset* de NUTS (regiones de Europa), reutilizan las definiciones realizadas en la DBPedia.

La información que se tomará como paradigmática para este enfoque surge de las hojas de cálculo disponibles en el servidor FTP de TED en la Unión Europea, que consta de los siguientes campos:

organisation official name, postal address (street), town, postalcode, countrycode, contact: attention, tel, fax, email, url, postal-code for the country (1 to 3 characters), contact point(s): telephone number, fax number, e-mail address, URL, un ejemplo completo se ha señalado en la Figura 5.28 extraído de la documentación oficial de TED. No obstante y de acuerdo a la ejecución del proyecto "10ders Information Services" se ha seleccionado el *dataset* provisto por la empresa Gateway S.C.S., reseñado en la Figura 5.38, y consistente en 50,000 entidades. La actualización de este *dataset* es incremental a lo largo del tiempo y crece generando una gran masa de información de datos, con el objetivo de realizar la prueba de concepto de representación de la información empresarial, se ha decidido acotar a las necesidades del proyecto.

```
<addr>
  <organisation>Westminster City Council</organisation>
  <address>City Hall, 64 Victoria Street</address>
  <contact>Westminster Accord</contact>
  <attention>Linda Pinder</attention>
  <countrycode>UK</countrycode>
  <town>London</town>
  <postalcode>SW1E 6QP</postalcode>
  <tel>020 7641 3173</tel>
  <email><txemail>lpinder@westminster.gov.uk</txemail></email>.
  <fax>020 7641 3017</fax>
</addr>
```

Figura 5.28: Ejemplo de datos sobre una Organización en TED.

```
ID contrato, ISO code, campos internos de la BBDD, tipo de entidad, nombre entidad en
     cada idioma, dir. postal, campos de la dir. postal
...
000170-2011,da,false,7,5,1,ContractingAuthority,Europa-Parlamentet,,,"plateau de
    Kirchberg, 2929, Luxembourg",plateau de Kirchberg,2929,Luxembourg,,"",,,
000170-2011,lv,false,7,5,1,ContractingAuthority,Eiropas Parlaments,,,"plateau de
    Kirchberg, 2929, Luksemburga",plateau de Kirchberg,2929,Luksemburga,,"",,,
000170-2011,nl,false,7,5,1,ContractingAuthority,Europees Parlement,,,"plateau de
    Kirchberg, 2929, Luxemburg",plateau de Kirchberg,2929,Luxemburg,,"",,,
000170-2011,pl,false,7,5,1,ContractingAuthority,Parlament Europejski,,,"plateau de
    Kirchberg, 2929, Luksemburg",plateau de Kirchberg,2929,Luksemburg,,"",,,
000170-2011,pt,false,7,5,1,ContractingAuthority,Parlamento Europeu,,,"plateau de
    Kirchberg, 2929, Luxemburgo",plateau de Kirchberg,2929,Luxemburgo,,"",,,
000170-2011,es,false,7,5,1,ContractingAuthority,Parlamento Europeo,,,"plateau de
    Kirchberg, 2929, Luxemburgo",plateau de Kirchberg,2929,Luxemburgo,,"",,,
...
```

Figura 5.29: Ejemplo de datos sobre Organizaciones.

5.3.1.2 Tarea t_2-Limpieza de datos

La oportunidad de reutilizar la información y datos suministrados por una empresa experta en el dominio de la contratación pública y con la capacidad para filtrar aquella información y datos relevantes resulta de gran valor para la ejecución de la transformación de estos datos. Es por ello que partiendo de este *dataset* original, esta tarea ya ha sido realizada y los datos a transformar se encuentran en condiciones óptimas de calidad, tanto en contenido como en formato (CSV).

5.3.1.3 Tarea t_3-Selección de Vocabularios

Los vocabularios seleccionados para modelar la información de las organizaciones teniendo en cuenta el análisis realizado en la tarea t_1 y la metodología seguida en los anteriores apartados se presentan en la Tabla 5.19. En general, se trata de vocabularios que atienden a los siguientes criterios:

1. Formalización de una estructura taxonómica, como RDFS, SKOS u OWL.
2. Realización de *mapeos* entre conceptos, como SKOS y OWL.
3. Representación de tipos de datos, como XML Schema.
4. Gestión de información multilingüe, como SKOS-XL y RDFS.
5. Representación de información de negocio (entidades y personas), como "*Organizations Ontology*", "FOAF" y "*DBPedia Ontology*".
6. Adición de metadatos y provenance, como *Dublin Core Terms*, voID y *Provenance Ontology*.
7. Representación de información de contacto como vCard en RDF.

Prefijo	Vocabulario	Fuente	Uso
dbpedia	http://dbpedia.org/ontology/	Comunidad *Linked Data*.	Reutilización de definiciones.
dc	http://purl.org/dc/elements/1.1/	*Dublin Core Metadata Initiative*	Creación de metadatos para los documentos.
dct	http://dublincore.org/documents/dcmi-terms/	≡	≡
foaf	http://xmlns.com/foaf/0.1/	Comunidad de Web Semántica.	Especificación de relaciones entre personas.
geo	http://www.w3.org/2003/01/geo/wgs84_pos#	W3C	Reutilización de elementos geográficos.
lgd	http://linkedgeodata.org/ontology/	*Linked Geodata Initiative*	≡
org	http://www.w3.org/ns/org#	W3C y Epimorphics Ltd.	Descripción de organizaciones.
owl	http://www.w3.org/2002/07/owl#	W3C	Realización de definiciones en el dominio.
prov	http://purl.org/twc/ontology/w3c/prov#	W3C	Especificación de metadatos de procedencia.
nuts	http://nuts.psi.enakting.org/def/	Universidad de Southampton	Especificación de las regiones europeas.
skos	http://www.w3.org/2004/02/skos/core#	W3C	Especificación de taxonomías.
skosxl	http://www.w3.org/2008/05/skos-xl#>	W3C	Representación de información lingüística.
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#	W3C	Descripción de recursos.

Prefijo	Vocabulario	Fuente	Uso
rdfs	http://www.w3.org/2000/01/rdf-schema#	W3C	Descripción de recursos con relaciones lógicas.
vcard	http://www.w3.org/2006/vcard/ns#	W3C	Representación de información de contacto.
void	http://rdfs.org/ns/void#	Deri y W3C	Descripción de metadatos de un *dataset*.
xml	http://www.w3.org/XML/1998/namespace	W3C	Reutilización de definiciones.
xsd	http://www.w3.org/2001/XMLSchema#	W3C	Especificación de tipos de datos.

Tabla 5.19: Selección de Vocabularios para las Organizaciones.

5.3.1.4 Tarea t_4-Selección de otros *datasets* RDF

En este caso los *datasets* a reutilizar se centran en la información sobre personas y organizaciones, otorgando igualmente valor a la información sobre países, específicamente a las regiones europeas. Los conjuntos de datos seleccionados se señalan en la Tabla 5.20.

Prefijo	*Dataset*	Fuente	Uso
dbpedia-res	http://dbpedia.org/	Comunidad *Linked Data*.	Reutilización de datos provenientes de la DBPedia.
nuts	http://nuts.psi.enakting.org/	University of Southampton	Especificación de las regiones europeas.

Tabla 5.20: Selección de otros *datasets* para las Organizaciones.

5.3.1.5 Tarea t_5-Modelado de datos en RDF

El resultado de esta tarea debe ser una ontología de dominio $\mathcal{O}$ que modele los recursos, organizaciones, personas y países, pertenecientes a un *dataset* RDF $\mathcal{D}$. Inicialmente, se debe proveer la especificación del modelo formal u ontología $\mathcal{O}$ de acuerdo al análisis realizado en la tarea t_1, dando soporte a la descripción de los datos de los recursos RDF:

- $\mathcal{C} = \{Organization, Country, Person, \ldots, Site\}$
- $\mathcal{R} = \{rdf{:}type, dc{:}identifier, org{:}purpose, org{:}classification, \ldots, org{:}memberOf, rdfs{:}label\}$
- $\mathcal{I} = \{r_{org}, r_{country}\}$
- $\mathcal{A} = \{\star\}$, son los axiomas propios de la ontología de organizaciones y los vocabularios FOAF y vCard.

Como segundo paso se diseñan las propiedades que deben tener los recursos, teniendo en cuenta su ulterior aplicación ver Tabla 5.21, pertenecientes a ese *dataset* (organizaciones, personas y países).

Propiedad	Descripción	Ejemplo
rdf:type	Especificación del tipo de un recurso	org:FormalOrganization
rdfs:label	Etiqueta para el recurso organización, persona o país	rdfs:label "Dutch Company Inc."@en
org:purpose	Perfil de características de una organización	N/A
org:classification	Tipo de organización de acuerdo a una taxonomía en SKOS	org:classification <SME>
org:hasSite	Localización física de una organización	org:hasSite <http://purl.org/weso/eprocurement/organization/resource/{id}/site>
org:siteAddress	Recurso vcard que especifica los datos de una localización física de una organización o persona	org:siteAddress <http://purl.org/weso/eprocurement/organization/resource/{id}/vcard>
org:memberOf	Pertenencia de un "agente", en este caso personas a una organización	
org:basedAt	Localización física de una persona	org:basedAt <http://purl.org/weso/eprocurement/organization/resource/{id}/vcard>
eproc:refCountry	Enlace a un país	eproc:refCountry <http://purl.org/weso/eprocurement/country/resource/ES>
eproc:refDbpedia	Enlace al recurso de un país en la DBPedia	eproc:refDbpedia <http://dbpedia.org/resource/ISO_3166-1:ES>
eproc:refNuts	Enlace al recurso de una región en NUTS	eproc:refNuts <http://nuts.psi.enakting.org/id/ES>
eproc:shortLabel	Código ISO 3166 de un país	eproc:shortLabel "ES"
geo:lat	Latitud de un recurso	geo:lat "40.46366700000001"
geo:long	Longitud de un recurso	geo:long "-3.74922"
vcard:*	Establecimiento de la información de contacto para un recurso organización o persona	N/A

Tabla 5.21: Diseño de propiedades para los elementos de las Organizaciones.

Una vez especificado el conjunto de propiedades de cada recurso es necesario definir el conjunto de grafos en los cuales se encuadrarán los recursos, es decir, el *dataset* RDF $\mathcal{D}$. Para ello, en la Tabla 5.22 se indican las tuplas $(\mathcal{G}_k, I_k)$ correspondientes a cada uno de los grafos $\mathcal{G}_k$ identificados a

través de la URI I_k realizando especial énfasis en la separación entre los recursos con datos en sí mismos, de las definiciones de aquellos. Hay que destacar que los recursos de la información asociada a personas no tienen entidad por sí mismos sino que se relacionan con una organización, es decir, no existen personas no relacionadas con organizaciones.

$\mathcal{G}_k$	I_k
$\mathcal{G}$	http://purl.org/weso/eprocurement/
$\mathcal{G}_1$	http://purl.org/weso/eprocurement/organization
$\mathcal{G}_2$	http://purl.org/weso/eprocurement/organization/person/
$\mathcal{G}_3$	http://purl.org/weso/eprocurement/country
$\mathcal{G}_4$	http://purl.org/weso/eprocurement/organization/ontology
$\mathcal{G}_5$	http://purl.org/weso/eprocurement/country/ontology

Tabla 5.22: *Dataset* RDF $\mathcal{D}$ para Organizaciones.

5.3.1.6 Tarea t_6-Diseño de un Esquema de URIs

Esta tarea tiene como objetivo establecer la forma y estructura de las URIs tanto para las definiciones realizadas en la ontología $\mathcal{O}$ como para todos los recursos presentes en el *dataset* RDF $\mathcal{D}$ que se genera a partir de la transformación de los datos a RDF. Es una de las actividades clave ya que guiará tanto el método final de transformación como el proceso posterior de publicación. En la Tabla 5.23 se hace la descripción de la estructura de URIs que se siguen para las entidades reconocidas de organizaciones, personas y países.

URI	Descripción	Ejemplo
http://purl.org/weso/eprocurement	URI base: <base_uri>	NA
<base_uri>/organization/ontology	Definiciones comunes para las organizaciones y personas	<base_uri>/ontology/Organization
<base_uri>/organization/resource/ds	Descripción del *dataset* de organizaciones	<base_uri>/resource/ds
<base_uri>/organization/resource/{id}	Descripción de un recurso organización identificado por id	<base_uri>/organization/resource/1
<base_uri>/organization/person/resource/ds	Descripción del *dataset* de personas dentro de las organizaciones	<base_uri>/person/resource/ds
<base_uri>/organization/person/resource/{id}	Descripción de un recurso persona identificado por id	<base_uri>/organization/person/resource/1
<base_uri>/country/ontology	Definiciones comunes para los países	<base_uri>/country/ontology/Country
<base_uri>/country/resource/ds	Descripción del *dataset* de países	<base_uri>/country/resource/ds

URI	Descripción	Ejemplo
<base_uri>/country/resource/{id}	Descripción de un recurso país identificado por id	<base_uri>/country/resource/ES

Tabla 5.23: Diseño de URIs para las Organizaciones.

5.3.1.7 Tarea t_7-Diseño Plantilla Objetivo del Recurso RDF

El objetivo de esta tarea es establecer una plantilla de cada uno de los recursos RDF que están presentes en el *dataset* RDF $\mathcal{D}$ para que sirvan como guía en los siguientes momentos: 1) en la ejecución propiamente dicha de la transformación de los datos originales a RDF y 2) en la validación de los recursos RDF generados. De esta manera, tratándose de *datasets* con una gran cantidad de recursos se pueden identificar fácilmente recursos que no sean compatibles con este esquema favoreciendo la depuración de los recursos generados, adicionalmente, un esquema de recurso sirve como documentación extra para el proceso de consumo. En este caso es necesario definir una plantilla para las organizaciones (ver Figura 5.30), las personas (ver Figura 5.31) y los países (ver Figura 5.32) con la información y datos que contendrán en cada caso.

De acuerdo al recurso plantilla y a las definiciones realizadas en la ontología que modela estos datos es posible realizar una validación en cuanto a los tipos de datos, cardinalidad de las relaciones, tipo de objetos, etc., que resulta de sumo interés para asegurar la calidad de los datos producidos.

```
<<base_uri>/organization/resource/{id}/vcard>  a v:VCard ;
  v:fn "Dutch Company Inc." ;
  v:org [ v:organisation-name "Dutch Company Inc." ;
          v:organisation-unit "Corporate Division" ] ;
  v:adr [ rdf:type v:Work ;
          v:country-name "Netherlands" ;
          v:locality "Amsterdam" ;
          v:postal-code "1016 XJ" ;
          v:street-address "Lijnbaansgracht 215" ] ;
  v:geo [ v:latitude "52.36764" ;
          v:longitude "4.87934" ] ;
  v:tel [ rdf:type v:Fax, v:Work ;
          rdf:value " +31 (10) 400 48 00"] ;
  v:email <mailto:company@mydutchcompany> ;
  v:logo <http://mydutchcompany.com/logo.png> .

<<base_uri>/organization/resource/{id}/site> a org:Site;
 rdfs:label "Site of Dutch Company Inc."@en ;
 org:siteAddress <<base_uri>/organization/resource/{id}/vcard> .

<<base_uri>/organization/resource/{id}> a
  org:FormalOrganization ;
  org:purpose <URI_purpose>;
  org:classification <URI_type>;
  org:hasSite <<base_uri>/organization/resource/{id}/site> ;
  eproc:refCountry <http://dbpedia.org/resource/ISO_3166-1:ES> ;
  (eproc:refNuts <http://nuts.psi.enakting.org/id/ES> ;) *
  rdfs:label "Dutch Company Inc."@en .
```

Figura 5.30: Plantilla Objetivo de una Organización.

```
<<base_uri>/organization/person/resource/{id}/vcard>  a v:VCard ;
  v:fn "Person name." ;
  v:org [ v:organisation-name "Dutch Company Inc." ;
          v:organisation-unit "Corporate Division" ] ;
  v:adr [ rdf:type v:Work ;
          v:country-name "Netherlands" ;
          v:locality "Amsterdam" ;
          v:postal-code "1016 XJ" ;
          v:street-address "Lijnbaansgracht 215" ] ;
  v:geo [ v:latitude "52.36764" ;
          v:longitude "4.87934" ] ;
  v:tel [ rdf:type v:Fax, v:Work ;
          rdf:value " +31 (10) 400 48 00"] ;
  v:email <mailto:company@mydutchcompany> ;
  v:logo  <http://mydutchcompany.com/logo.png> .

<<base_uri>/organization/person/resource/{id}> a
  foaf:Person ;
  org:memberOf <<base_uri>/organization/resource/{id}> ;
  org:basedAt <<base_uri>/organization/person/resource/{id}/vcard> ;
  .
```

Figura 5.31: Plantilla Objetivo de una Persona.

```
<<base_uri>/country/resource/{id}> a
  eproc:Country ;
  eproc:refDbpedia <http://dbpedia.org/resource/ISO_3166-1:ES> ;
  (eproc:refNuts <http://nuts.psi.enakting.org/id/ES> ;) [0,1]
  geo:lat "40.46366700000001" ;
  geo:long "-3.74922" ;
  eproc:shortLabel "ES" ;
  ...
  rdfs:label "Spain"@en .
```

Figura 5.32: Plantilla Objetivo de un País.

5.3.1.8 Tarea t_8-Enriquecimiento de los datos en RDF

El objetivo de esta tarea es enlazar los recursos generados del *dataset* RDF $\mathcal{D}$ con otros ya existentes. En el caso particular de las organizaciones, personas y países el esfuerzo se centra en realimentar el *dataset* de países según los conjuntos de datos candidatos seleccionados en la Tarea t_4.

Para realizar este enriquecimiento es necesario realizar reconciliación de entidades entre las descripciones de los recursos de los países y los recursos objetivos. Para la realización de este proceso se han utilizado los siguientes enfoques:

1. Reconciliación y enlazado con los países disponibles en la DBPedia, utilizando la herramienta de Google Refine y sus capacidades para la búsqueda de recursos en las grandes bases de datos ya publicadas como datos enlazados.
2. Enlazado con los países disponibles en la versión *Linked Data* de NUTS, realizando el enlazado manual (23 países).
3. Adición de los datos geográficos a cada país, mediante el servicio web de consulta de Google dentro de la herramienta de Google Refine.

Dado que la reconciliación de organizaciones es una tarea con un alto porcentaje de error debido a la casuística de nombrado y teniendo en cuenta que con el propósito de facilitar y favorecer el acceso a los anuncios de licitación y dado que no se trata de información crítica (disponer de una mayor información de una empresa no implica mejorar el acceso a los anuncios), se ha optado por posponer esta etapa a futuros trabajos específicos centrados en la monitorización de la actividad de las empresas, desarrollándose en la actualidad mediante la realización de un proyecto fin de carrera. De igual forma, la información relativa a las personas de contacto no se considera trascendente para la mejora en el acceso a los anuncios de licitación, si bien es relevante contemplar el modelado de estos datos como parte del proceso administrativo no se considera información crítica para el objeto de estudio de este documento, por lo que al igual que con las organizaciones, la identificación y enlazado de personas se ha pospuesto para posteriores mejoras.

5.3.1.9 Tarea t_9-Transformación de los datos a RDF

Una vez definidas las tareas anteriores es posible abordar la transformación de los datos de entrada a RDF, al igual que en las anteriores transformaciones se ha optado por un enfoque híbrido entre la herramienta Google Refine con su extensión para RDF y el uso de las utilidades del componente `moldeas-transformer`. De nuevo, la ejecución de esta tarea debe asegurar que el *dataset* RDF $\mathcal{D}$ generado es válido en cuanto a sintaxis.

5.3.1.10 Tarea t_{10}-Reconciliación de Entidades

El diseño y ejecución de esta tarea ha sido descrito en la Sección 5.3.1.8 ya que están estrechamente ligadas.

5.3.1.11 Tarea t_{16}-Añadir metainformación a los recursos RDF

Al igual que con las clasificaciones de productos, la metainformación cobra sentido a nivel de *dataset*, ver Figuras 5.33, 5.34 y 5.35, ya que no se consideran cambios probables en un corto espacio de tiempo en cada una de las entidades gestionadas. En cualquier caso, tratándose de las organizaciones la ontología seleccionada para representar el modelo formal de los recursos ya recoge la

posibilidad de añadir metainformación relativa a posibles eventos de cambios. Teniendo en cuenta que la información de las personas está asociada indisolublemente a las organizaciones, tampoco se considera necesario incluir un exceso de metainformación en cada uno de estos recursos. Finalmente y en el caso particular de los países, los cambios que se puedan producir constituirían una evolución real del *dataset* para verificar que la información de todos los países sigue siendo congruente. Por ello, se ha seguido el modelo de incluir la metainformación a nivel de *dataset* y no de recurso en particular. Continuando con los vocabularios seleccionados en los anteriores apartados, la especificación realizada por voID cubre las necesidades de metainformación para representar datos de procedencia, licencia, etc.

```
<http://purl.org/weso/eprocurement/organizations/resource/ds>
      a         void:Dataset ;
      rdfs:label "Organizations dataset 2011"@en ;
      dcterms:author
            <http://www.di.uniovi.es/~labra/labraFoaf.rdf#me> ,
            <http://www.josemalvarez.es/foaf.rdf#me> ;
      dcterms:contributor
            <http://purl.org/weso/pscs/resource/10ders> ,
            <http://rdfohloh.wikier.org/project/moldeas/rdf> ;
      dcterms:description "Dataset of organizations provided in 10ders project" ;
      dcterms:license <http://opendatacommons.org/licenses/by/1.0/> ;
      dcterms:modified "2011-12-10"^^xsd:date ;
      dcterms:publisher <http://www.josemalvarez.es/foaf.rdf#me> ;
      dcterms:source <http://purl.org/weso/pscs/resource/10ders> ;
      dcterms:title "Organizations dataset 2011" ;
      void:dataDump <http://purl.org/weso/eprocurement/organization/organization-10
          ders.ttl> ;
      void:exampleResource
        <http://purl.org/weso/eprocurement/organization/resource/1> ,
        <http://purl.org/weso/eprocurement/organization/resource/2> ,
        <http://purl.org/weso/eprocurement/organization/resource/3> ;
      void:uriRegexPattern
        "http://purl.org/weso/eprocurement/organization/resource/.+" ;
      void:vocabulary org: , rdfs: , vcard: ;
      foaf:homepage <http://purl.org/weso> .
```

Figura 5.33: Descripción del *dataset* de Organizaciones.

```
<http://purl.org/weso/eprocurement/organizations/person/resource/ds>
      a        void:Dataset ;
      rdfs:label "People of the Organization"@en ;
      dcterms:author
            <http://www.di.uniovi.es/~labra/labraFoaf.rdf#me> ,
            <http://www.josemalvarez.es/foaf.rdf#me> ;
      dcterms:contributor
            <http://purl.org/weso/pscs/resource/10ders> ,
            <http://rdfohloh.wikier.org/project/moldeas/rdf> ;
      dcterms:description "Dataset of people working on organizations provided in 10
          ders project" ;
      dcterms:license <http://opendatacommons.org/licenses/by/1.0/> ;
      dcterms:modified "2011-12-10"^^xsd:date ;
      dcterms:publisher <http://www.josemalvarez.es/foaf.rdf#me> ;
      dcterms:source <http://purl.org/weso/pscs/resource/10ders> ;
      dcterms:title "People of the Organization" ;
      void:dataDump <http://purl.org/weso/eprocurement/organization/person/person-
          organization-10ders.ttl> ;
      void:exampleResource
        <http://purl.org/weso/eprocurement/organization/person/resource/1> ,
        <http://purl.org/weso/eprocurement/organization/person/resource/2> ,
        <http://purl.org/weso/eprocurement/organization/person/resource/3> ;
      void:uriRegexPattern
        "http://purl.org/weso/eprocurement/organization/person/resource/.+" ;
      void:vocabulary foaf: , org: , rdfs: , vcard: ;
      foaf:homepage <http://purl.org/weso> .
```

Figura 5.34: Descripción del *dataset* de Personas.

```
<http://purl.org/weso/eprocurement/country/resource/ds>
      a        void:Dataset ;
      rdfs:label "ISO 3166 Countries"@en ;
      dcterms:author
            <http://www.di.uniovi.es/~labra/labraFoaf.rdf#me> ,
            <http://www.josemalvarez.es/foaf.rdf#me> ;
      dcterms:contributor
            <http://purl.org/weso/pscs/resource/10ders> ,
            <http://rdfohloh.wikier.org/project/moldeas/rdf> ;
      dcterms:description "Dataset of ISO 3166 Countries" ;
      dcterms:license <http://opendatacommons.org/licenses/by/1.0/> ;
      dcterms:modified "2011-12-12"^^xsd:date ;
      dcterms:publisher <http://www.josemalvarez.es/foaf.rdf#me> ;
      dcterms:source <http://purl.org/weso/pscs/resource/10ders> ;
      dcterms:title "ISO 3166 Countries1" ;
      void:dataDump <http://purl.org/weso/eprocurement/country/countries.ttl> ;
      void:exampleResource
        <http://purl.org/weso/eprocurement/country/resource/ES> ,
        <http://purl.org/weso/eprocurement/country/resource/BE> ,
        <http://purl.org/weso/eprocurement/country/resource/IT> ;
      void:uriRegexPattern
        "http://purl.org/weso/eprocurement/country/resource/.+" ;
      void:vocabulary geo: , rdfs: , dbpedia: ;
      foaf:homepage <http://purl.org/weso> .
```

Figura 5.35: Descripción del *dataset* de Países.

5.3.1.12 Resultado Final y Ejemplos

El resultado final del proceso de producción de *Linked Data*, tras el análisis y ejecución de las tareas identificadas y del método de producción seleccionado, genera como resultado tres *datasets* correspondientes a las entidades organización, persona y país mediante datos enlazados, en los cuales se pueden extraer las siguientes estadísticas de producción de datos, así como ejemplos de los recursos generados, ver Tabla 5.24. Por otra parte, el aumento de la expresividad en el momento de realizar consultas se puede observar en la Figura 5.36 en la cual se expresa la siguiente consulta:

> *"Dame la forma de contacto de todas las organizaciones que han publicado anuncios de licitación en un determinado país."*

Conjunto de datos	**Nº de Elementos**	**Ejemplo**	**Tripletas**	**Enlaces externos**
Organizaciones	50000	Figura 5.38	1150020	50000 (países)
Personas	50000	Figura 5.39	1100000	50000 (países)
Países	246	Figura 5.37	1756	1779
Organizaciones Personas y Países (total)				
Agregado	100246	N/A	2251776	101779

Tabla 5.24: Estadísticas y Ejemplos de las Organizaciones.

```
SELECT DISTINCT * WHERE{
  ?org rdf:type org:FormalOrganization .
  ?org eproc:refCountry ?country.
  ?country rdfs:label ?labelCountry.
  FILTER (?country != <http://nuts.psi.enakting.org/id/ES>).
  ?org org:hasSite ?site .
  ?org org:siteAddress ?vcard.
} LIMIT 100
```

Figura 5.36: Ejemplo de consulta en SPARQL sobre los datos de Organizaciones.

```
<http://purl.org/weso/webfoundation/webindex/countries/resource/ES>
        a eproc:Country ;
        eproc:refDbpedia <http://dbpedia.org/resource/ISO_3166-1:ES> ;
        eproc:refNuts <http://nuts.psi.enakting.org/id/ES>
        geo:lat "40.46366700000001" ;
        geo:long "-3.74922" ;
        rdfs:label "Spain"@en ;
        wi:shortLabel "ES" .
```

Figura 5.37: Ejemplo final de un País.

```
<http://purl.org/weso/eprocurement/organization/resource/100/vcard>  a v:VCard ;
  v:fn "University of Oviedo" ;
  v:org [ v:organisation-name "University of Oviedo" ;
          v:organisation-unit "Corporate Division" ] ;
  v:adr [ rdf:type v:Work ;
          v:country-name "Spain" ;
          v:locality "Oviedo" ;
          v:postal-code "33003" ;
          v:street-address "Plaza del Riego, S/N" ] ;
  v:geo [ v:latitude "43,375" ;
          v:longitude "-5,844" ] ;
  v:tel [ rdf:type v:Fax, v:Work ;
          rdf:value " +34  985 10 30 00"] ;
  v:email <mailto:gerencia@uniovi.es> ;
  v:logo <http://www.unioviedo.es/mia09/logos/logo_uniovi.jpg> .

<http://purl.org/weso/eprocurement/organization/resource/100/site> a org:Site;
 rdfs:label "Site of the University of Oviedo"@en ;
 org:siteAddress <http://purl.org/weso/eprocurement/organization/resource/100/vcard>
     .

<http://purl.org/weso/eprocurement/organization/resource/100> a
  org:FormalOrganization ;
  org:classification <http://purl.org/weso/eprocurement/organization/ontology/
      University>;
  org:hasSite <http://purl.org/weso/eprocurement/organization/resource/100/site> ;
  eproc:refCountry <http://dbpedia.org/resource/ISO_3166-1:ES> ;
  eproc:refNuts <http://nuts.psi.enakting.org/id/ES12> ;
  rdfs:label "University of Oviedo"@en .
```

Figura 5.38: Ejemplo final de una Organización.

```
<http://purl.org/weso/eprocurement/organization/person/resource/100/vcard>  a v:VCard
     ;
  v:fn "Vicente Gotor" ;
  v:org [ v:organisation-name "University of Oviedo" ;
          v:organisation-unit "Rector Division" ] ;
  v:adr [ rdf:type v:Work ;
          v:country-name "Spain" ;
          v:locality "Oviedo" ;
          v:postal-code "33003" ;
          v:street-address "Plaza del Riego, S/N" ] ;
  v:geo [ v:latitude "43,375" ;
          v:longitude "-5,844" ] ;
  v:tel [ rdf:type v:Fax, v:Work ;
          rdf:value " +34  985 10 30 00"] ;
  v:email <mailto:rector@uniovi.es> ;
  v:logo <http://www.unioviedo.es/mia09/logos/logo_uniovi.jpg>  .

<http://purl.org/weso/eprocurement/organization/person/resource/100> a
  foaf:Person ;
  org:memberOf
    <http://purl.org/weso/eprocurement/organization/resource/100> ;
  org:basedAt
    <http://purl.org/weso/eprocurement/organization/person/resource/100/vcard> ;
```

Figura 5.39: Ejemplo final de una Persona.

5.3.1.13 Método de Producción de *Linked Data* de Organizaciones

Siguiendo las directrices señaladas en el análisis y diseño de entidades participantes en el modelado de organizaciones y la tabla de decisión 4.12, el método semántico seleccionado para realizar la producción de datos enlazados es el SPM_1-"Transformación de datos a RDF", ver Sección 4.5.2, en el que se transforman un conjunto de datos de entrada $\mathcal{G}$ a un *dataset* RDF $\mathcal{D}$. Según la definición de método semántico de producción, realizada en la Sección 4.5.1, y el estudio de las organizaciones se pueden establecer los siguientes conjuntos:

- $\mathcal{G}$ es el *dataset* de entrada, conjunto de tuplas, conteniendo los datos de cada una de las organizaciones, las personas y los países.
- $\mathcal{M}$ es el conjunto de *mapeos*, ver Tabla 5.25, extraídos según el análisis y diseño realizado en las secciones anteriores. Estos *mapeos* son directamente expresables en la herramienta de transformación y toman como parámetros el valor de una de las tuplas de entrada (posición X) y la propiedad a generar.
- *Dataset* RDF $\mathcal{D}$ es el *dataset* resultado, siguiendo el análisis y diseño realizado en las secciones anteriores y tras la ejecución de la tarea propia de transformación de datos.

$\mathcal{M}$	**Propiedad**	**Valor**
m_1	`rdf:type`	URI
m_2	`rdfs:label`	`xsd:string@lang`
m_3	`org:purpose`	URI
m_4	`org:hasSite`	URI
m_4	`org:siteAddress`	URI
m_5	`org:memberOf`	URI
m_6	`org:basedAt`	URI
m_7	`eproc:refCountry`	URI
m_8	`eproc:refDbpedia`	URI
m_9	`eproc:refNuts`	URI
m_{10}	`eproc:shortLabel`	`xsd:string`
m_{11}	`geo:lat`	`xsd:double`
m_{12}	`geo:long`	`xsd:double`
m_{13}	`vcard:*`	URI (incluyendo más *mapeos* dependiendo de la información)

Tabla 5.25: Conjunto de *mapeos* $\mathcal{M}$ para las Organizaciones.

5.3.2 Proceso de Publicación de *Linked Data* de Organizaciones

Evidentemente, la estrategia definida para todos los *datasets* implicados en el proceso de contratación pública electrónica es común, por ello el proceso de publicación es homogéneo siguiendo la estructura definida en la Sección 5.2.2.

5.3.2.1 Tarea t_{14}-Infraestructura para *Linked Data*

Nuevamente, la estrategia definida para todos los *datasets* implicados en el proceso de contratación pública electrónica es común, por ello la infraestructura utilizada es la misma que la definida en la Sección 5.1.2.1.

5.3.2.2 Tarea t_{15}-Acceso y formato en datos RDF

De la misma forma que en el apartado anterior, el acceso y formato de datos RDF para los datos contenidos y relativos a las organizaciones siguen el esquema proporcionado en la Sección 5.1.2.2.

5.3.3 Proceso de Consumo de Organizaciones

El proceso de consumo de datos enlazados, según la definición realizada en la Sección 4.7.1, consiste en la reutilización de los datos enlazados para ser aplicados en la construcción de una nueva aplicación o servicio de valor añadido. En general, la reutilización más sencilla consiste en la representación gráfica de los recursos o la simple consulta con selección de formato de datos de acuerdo a las características de publicación utilizadas. En el caso que nos ocupa y teniendo en cuenta el objetivo de realización de un prototipo experimental de extracción de anuncios de licitación como demostrador del consumo de datos enlazados, se ha escogido el método semántico de consumo SCM_2-"*Mapeo* a Lenguaje de Programación", cuya descripción está disponible en la Sección 4.7.3, orientado a obtener una representación de los recursos RDF en un lenguaje de programación (en este caso Java) como objetos de negocio. De acuerdo a este objetivo y a la definición del propio método es necesario definir:

- El *dataset* RDF $\mathcal{D}_{pub}$, es el conjunto de datos disponible tras aplicar el método de publicación.
- El conjunto $\mathcal{M}^1$, ver Tabla 5.26, indica como transformar el *dataset* anterior a la representación objetivo, objetos del lenguaje Java.

De esta manera, se obtienen una serie de objetos, $\mathcal{D}_{consum}$, con la información y datos necesarios, no se transforman necesariamente todos los datos disponibles en los recursos pertenecientes a $\mathcal{D}_{pub}$, para ser reutilizados como objetos de negocio en un lenguaje de programación. Es conveniente señalar que el acceso a los datos se realiza a través de la consulta al *endpoint* de SPARQL ejecutando consultas *SELECT* y *DESCRIBE*.

$\mathcal{M}^1$	**Propiedad**	**Tipo en Java**
m_1^1	URI recurso	`java.lang.String`
m_2^1	`rdf:type`	`org.weso.moldeas.to.OrganizationTO, PersonTO, CountryTO`
m_3^1	`rdfs:label`	`Map<String,String>` (lang, value)
m_4^1	`org:purpose`	`List<org.weso.moldeas.to.PurposeTO>`
m_5^1	`org:hasSite`	`org.weso.moldeas.to.SiteTO`
m_6^1	`org:siteAddres`	`org.weso.moldeas.to.SiteAddressTO`
m_7^1	`org:memberOf`	`org.weso.moldeas.to.OrganizationTO`
m_8^1	`org:basedAt`	`org.weso.moldeas.to.SiteTO`
m_9^1	`eproc:refCountry`	`org.weso.moldeas.to.CountryTO`
m_{10}^1	`eproc:refDbpedia`	`org.weso.moldeas.to.CountryTO`
m_{11}^1	`eproc:refNuts`	`List<org.weso.moldeas.to.NUTSTO>`
m_{12}^1	`eproc:shortLabel`	`java.lang.String`

$\mathcal{M}^1$	**Propiedad**	**Tipo en Java**
m^1_{13}	`geo:lat` and `geo:long`	`com.javadocmd.simplelatlng.LatLng`
m^1_{14}	`vcard:*`	`org.weso.moldeas.to.VcardTO`

Tabla 5.26: Conjunto de *mapeos* $\mathcal{M}^1$ de consumo para las Organizaciones.

5.3.4 Proceso de Validación de Organizaciones

La validación como proceso transversal en cualquier etapa dentro del ciclo de vida de datos enlazados debe realizarse con el objetivo de asegurar la calidad de los datos. De acuerdo a la definición realizada en la Sección 4.8, este proceso consiste en la comprobación de que los recursos de un *dataset* RDF cumplen ciertas características. Siguiendo con la estrategia marcada para todos los datos implicados en el proceso de contratación pública electrónica, la descripción completa de la validación de acuerdo a todas las características, se reseña en las Tablas de Validación disponibles en el Apéndice C.

5.3.4.1 Tarea t_{12}-Validación de Recursos RDF

Al igual que en apartados anteriores, de acuerdo a la estrategia definida y con la definición realizada de esta tarea en la Sección 4.4.12, se puede asegurar que la transformación de la información correspondiente de las organizaciones a la iniciativa *Linked Data* cumple estrictamente los siguientes puntos:

- Los datos RDF son correctos ya que se han utilizado herramientas y APIs (Google Refine y Jena) que aseguran la generación correcta de RDF.
- El dominio y rango en las propiedades es correcta, ya que se realiza la validación contra el modelo definido.
- Se ha establecido metainformación sobre la procedencia a nivel de *dataset*.
- Todos los recursos transformados siguen la plantilla objetivo RDF.

5.3.5 Proceso de Realimentación de Organizaciones

Este proceso, según la definición realizada en la Sección 4.9, busca la mejora y perfeccionamiento de los datos promocionados a RDF. Esta situación emerge en el momento en el cual los datos comienzan a ser reutilizados tanto por aplicaciones o servicios como por individuos. En el caso particular de las organizaciones y sus entidades relacionadas (personas y países) no se han llegado a reutilizar por terceras partes, por lo que la realimentación ha quedado restringida a la captura de fallos por la propia aplicación MOLDEAS, tratándose en este caso de una forma de realimentación basada en *Usuarios y Aplicaciones* y de carácter *Actualización Ocasional*.

Capítulo 6

Sistema MOLDEAS

> La mayoría de la gente no sabe lo que quiere hasta que se lo enseñas.
>
> *Citas Célebres*
> STEVE JOBS

6.1 Introducción

En el caso objeto de estudio de este documento se ha planteado la aplicación de los principios de *Linked Data*, para el modelado y explotación de datos e información provenientes de los anuncios de licitación públicos. Para ello, tal y como se ha presentado en los anteriores capítulos, se ha definido un ciclo de vida de datos enlazados que a través de procesos, métodos y tareas suministra una metodología de actuación genérica para proceder en este sentido. Como ejemplo de su validez y aplicación se han utilizado los datos de los contratos públicos para ejemplificar los procesos de producción, publicación, consumo, validación y actualización, ver Figura 6.1. Aunque ciertas tareas se apoyan en el uso de aplicaciones de terceros como Google Refine o bien en la parametrización de bibliotecas ya existentes como Apache Lucene, es necesario suministrar un entorno en el cual los resultados de aplicación de las tareas puedan ser procesados para implementar algunas de las ya especificadas y así ejemplificar transversalmente el uso de datos enlazados en un determinado dominio. Por ello y de acuerdo al análisis y especificación realizado se plantea la necesidad de diseñar los componentes del sistema MOLDEAS como paso final para la cobertura en el uso de datos enlazados en el campo de la contratación pública electrónica y teniendo presentes los siguientes objetivos:

- Facilitar y dar soporte a las tareas del ciclo de vida que no puedan ser desarrolladas completamente por herramientas externas.
- Validar los datos generados por otras herramientas.
- Enriquecer con procesos *ad-hoc* la información y datos de los anuncios de licitación según el modelo y especificación fijado en el Capítulo 5.
- Implementar un demostrador público de consumo de datos enlazados.
- Proveer un sistema de búsqueda/recomendación de anuncios de licitación de acuerdo a criterios predefinidos por el cliente.

- Establecer un conjunto de prueba que realice la validación parcial de ciertos procesos apoyándose en tecnología pre-existente.
- Diseñar un sistema extensible y escalable para su futura ampliación.

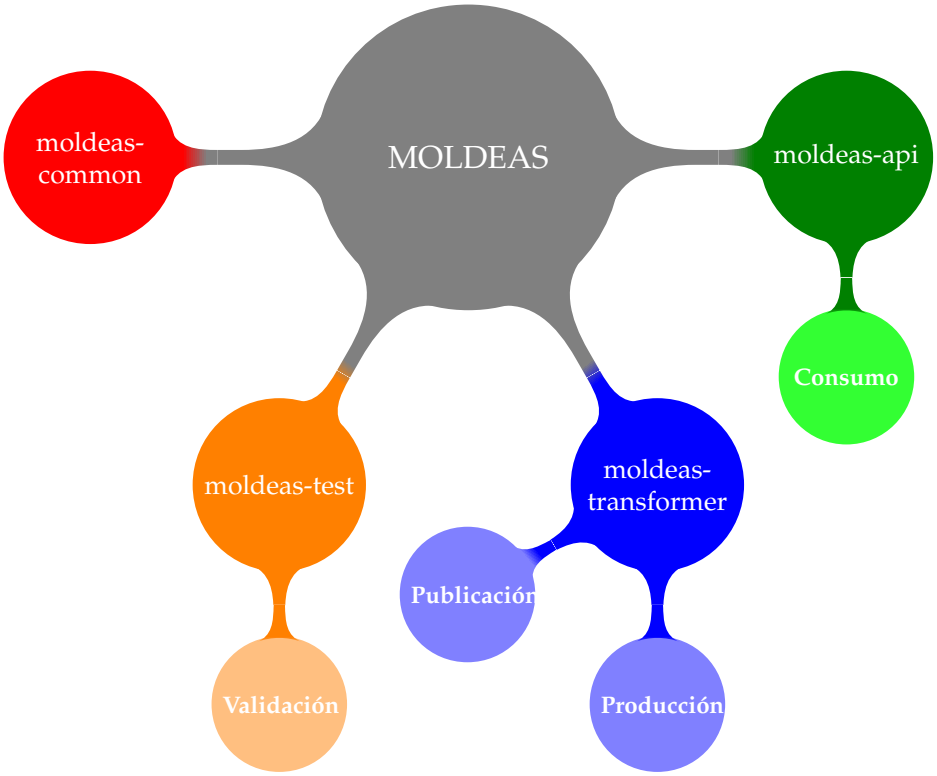


Figura 6.1: Alineación inicial de componentes de MOLDEAS y procesos del Ciclo de Vida de *Linked Data*.

Ante estos objetivos de gran envergadura, teniendo en cuenta el ciclo de vida de datos enlazados definido y las tareas especificadas para la información y datos presentes en las licitaciones públicas, cabe especificar el diseño de los componentes del sistema MOLDEAS como elemento vertebrador tanto de los procesos como de la información. De esta forma, a lo largo de este capítulo se realiza una descripción del diseño e implementación realizada en el sistema MOLDEAS haciendo especial hincapié en los detalles más relevantes del mismo.

6.2 Descripción del Sistema MOLDEAS

La tarea de análisis de un sistema conlleva la especificación de una serie de requisitos que guíen el posterior diseño e implementación del sistema MOLDEAS, de esta forma, se pueden extraer los siguientes objetivos:

- Analizar e identificar el trabajo relacionado.
- Identificar y definir los requisitos asociados a las actividades de investigación, innovación y desarrollo a realizar.
- Realimentar los puntos anteriores con los resultados obtenidos.

Estos objetivos ya han sido parcialmente cubiertos en los anteriores capítulos, en los cuales se ha repasado intensivamente la mayor parte de los trabajos más relevantes relacionados con el dominio de la contratación pública electrónica, así como puesto de manifiesto los procesos, métodos y tareas a desarrollar dentro del ciclo de vida de datos enlazados. Es por ello que este capítulo se centrará en presentar las partes más prominentes y destacadas del sistema MOLDEAS y su aplicación en las distintas tareas del ciclo de vida, así como en el desarrollo del sistema de búsqueda de anuncios de licitación. En primer lugar, cabe mostrar un esbozo del sistema con una vista funcional del mismo, ver Figura 6.2, en el cual se enclavan los distintos procesos del ciclo de vida y los objetivos generales del sistema: producción de datos enlazados de los anuncios de licitación para su posterior almacenamiento y publicación a través de un *endpoint* de SPARQL y un *Linked Data frontend* y consumo de los datos almacenados para la construcción de servicios de valor añadido como el sistema de búsqueda/recomendación de anuncios de licitación. Se ha optado por un enfoque totalmente práctico en este capítulo de ingeniería con el objetivo de resaltar y documentar la innovación del sistema MOLDEAS sin focalizar en metodologías intensivas de ingeniería del *software*.

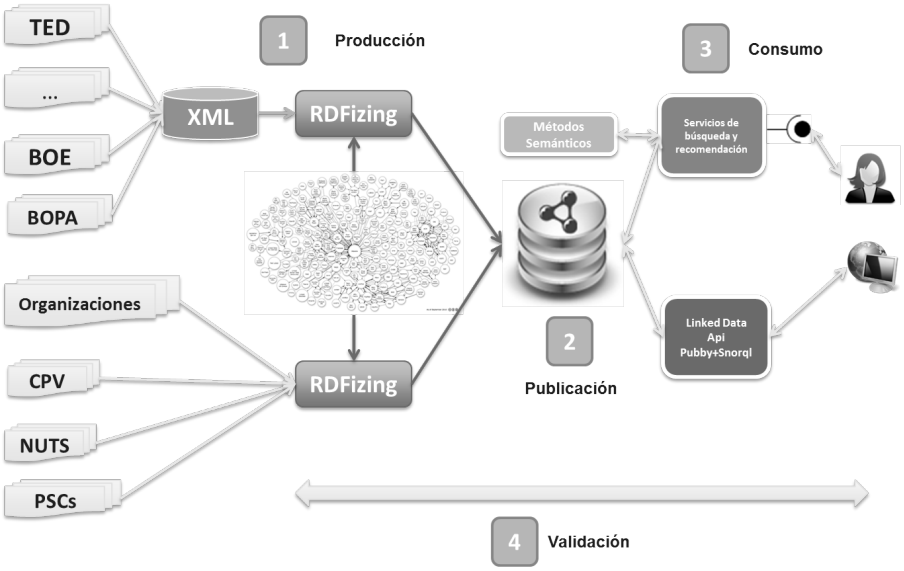


Figura 6.2: Arquitectura funcional del sistema MOLDEAS.

A la vista de la arquitectura funcional propuesta, se ha realizado una aproximación en distintos componentes, de acuerdo a sus responsabilidades e interacciones, ver Figura 6.3, y con las siguientes definiciones:

moldeas-common. Alberga utilidades necesarias a lo largo de todos los procesos del ciclo de vida de datos enlazados.

moldeas-transformer. Se encarga de dar soporte a la producción de datos enlazados cubriendo las tareas de transformación, enriquecimiento, reconciliación de entidades, etc.

moldeas-api. Se encarga del consumo de los datos enlazados publicados bajo unas ciertas características en un *endpoint* de SPARQL y que con la aplicación de varios métodos de expansión de consultas es capaz de generar consultas SPARQL cercanas al lenguaje natural para la recuperación de los anuncios de licitación.

moldeas-test. Se encarga de la validación de los datos enlazados y de los métodos de expansión definidos en `moldeas-api`.

moldeas-web. Se encarga de la presentación y consumo de datos enlazados suministrando un interfaz gráfico para `moldeas-api` en el cual el usuario puede seleccionar las características de los

anuncios de licitación para su posterior búsqueda y presentación con diferentes vistas (tabla, mapa, etc.). Además, suministra un interfaz de servicios REST para el acceso a los métodos disponibles en *moldeas-api*, por lo que cumple una doble función: 1) servir como demostrador público para el usuario y 2) ejemplificar las llamadas a `moldeas-api` desde un punto de vista del desarrollador.

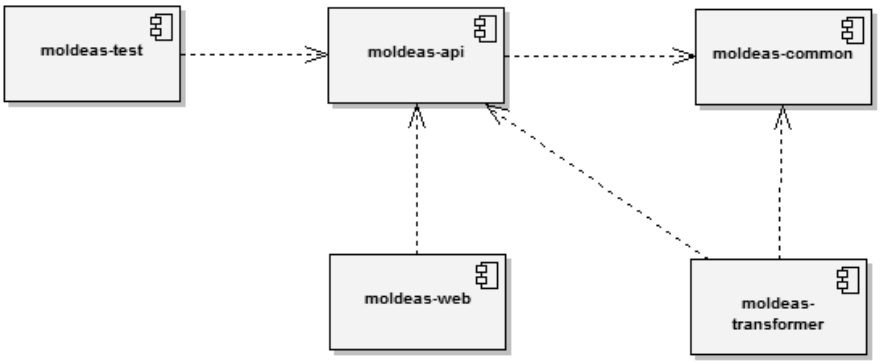


Figura 6.3: Componentes del sistema MOLDEAS.

6.2.1 Arquitectura de alto nivel

El despliegue de una infraestructura de datos enlazados requiere la cooperación de diferentes elementos *hardware* y *software*. Según el análisis realizado de cada uno de los componentes un diagrama de despliegue de la arquitectura propuesta en MOLDEAS, se presenta en la Figura 6.4, la descripción de cada uno de estos nodos y componentes es la siguiente:

- Nodo web. En el cual se encuentran disponibles un servidor web HTTP, Apache2 HTTP Server, este elemento *software* sirve como punto de entrada a los elementos del sistema, tanto para el consumo de datos enlazados directamente por otras máquinas como para las peticiones relativas al componente *moldeas-web*. También, se utiliza este servicio para albergar una aplicación de ejecución de consultas *on-line* en SPARQL como SNORQL, sin necesidad de utilizar directamente el interfaz propuesto por el *endpoint* de SPARQL.

- Nodo de aplicaciones web. En el cual se encuentra instalado un contenedor de aplicaciones web J2EE como Apache Tomcat, con el objetivo de albergar las aplicaciones relativas al *Linked Data frontend* y a la aplicación *moldeas-web*.

- Nodo repositorio RDF. En el cual se instala el repositorio RDF, en el cual se almacenan y publican los datos enlazados provenientes de los anuncios de licitación. La publicación de datos enlazados se realiza a través de su almacenamiento en un repositorio RDF nativo como Virtuoso de OpenLink, suministrando adicionalmente un *endpoint* SPARQL para que los datos puedan ser reutilizados y consumidos tanto por la propia aplicación de *moldeas-api* como por clientes de forma externa, en este caso por el *Linked Data frontend* y SNORQL.

Físicamente estos nodos y componentes se han diseñado de forma separada ya que la comunicación entre los mismos se realiza mediante delegación de consultas y comunicación mediante HTTP. De esta manera, se permite que el sistema sea escalable y flexible, pudiendo sustituir el entorno tecnológico seleccionado por otros proveedores.

Los componentes especificados en MOLDEAS tienen un doble carácter ya que algunos son utilizados de forma *off-line* como es el caso de `moldeas-transformer` y `moldeas-test`, específicamente para los procesos de producción y validación de datos enlazados, mientras que otros como *moldeas-web* sirven como cliente de los servicios proporcionados en `moldeas-api` de forma `on-line`. Transversalmente *moldeas-common* se utiliza a lo largo de cualquier ejecución ya que contiene las clases *Helper* necesarias para la ejecución de tareas comunes.

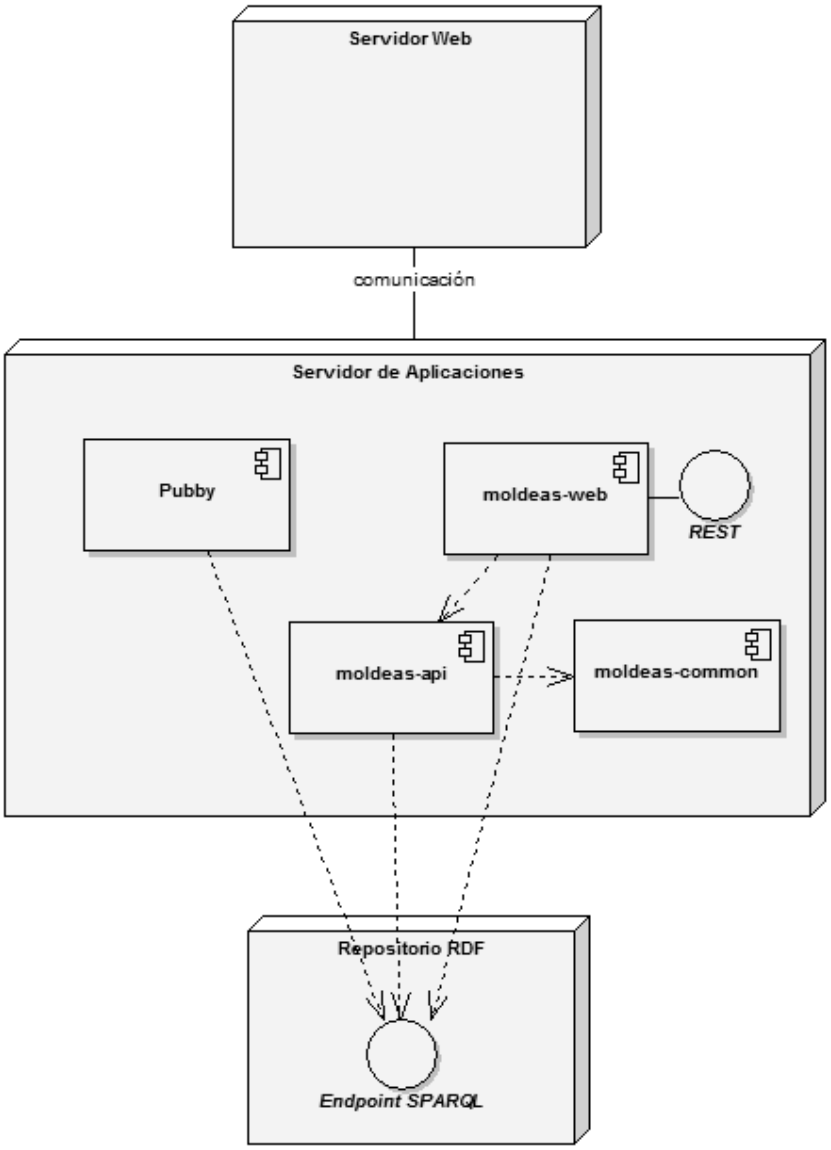


Figura 6.4: Diagrama de Despliegue de MOLDEAS.

6.2.2 Entorno Tecnológico

En el momento de analizar y diseñar un conjunto de componentes *software* es necesario seleccionar aquellas bibliotecas y herramientas que suministren determinadas funcionalidades de base. La selección estratégica de esta tecnología y herramientas se describe someramente a continuación:

- Lenguaje de programación Java 1.6. En el campo de la Web Semántica y en particular de la iniciativa de datos enlazados, la mayoría de las bibliotecas y funcionalidades externas se en-

cuentran programadas en este lenguaje, este hecho unido a la experiencia propia del autor, justifica perfectamente la decisión de uso de este lenguaje.

- Apache Maven2. El desarrollo de aplicaciones debe realizarse de forma sostenible, es por ello que esta herramienta da soporte a todo el proceso de construcción de *software*: compilación, pruebas, empaquetado, despliegue, documentación, ejecución, etc. Además de proveer los mecanismos apropiados para la gestión de dependencias de forma declarativa.
- Eclipse IDE. La edición del código fuente de las clases Java se ha realizado a través de este entorno de desarrollo, ampliamente asentado en la comunidad Java y con una excelente comunidad que proporciona herramientas extra a través de distintos *plugins*, proporcionando un entorno productivo y con altas capacidades para el desarrollador. Adicionalmente, Maven es capaz de generar a través de su fichero de configuración la estructura de un proyecto de Eclipse de forma automática.
- Repositorio de código fuente. Se ha seleccionado la forja de proyectos de Google Code para albergar los cambios y las actualizaciones maduras a través del sistema de control de versiones Mercurial.
- Jena 2.6.4. Biblioteca Java de base tecnológica para la Web Semántica que incluye las principales funciones para el tratamiento de RDF, OWL, etc., facilitando la ejecución de consultas SPARQL tanto de forma local (en memoria) como distribuida (en un *endpoint*).
- Log4j 1.2.14. Biblioteca Java para la gestión del sistema de registro de una aplicación mediante la cual se puede especificar los distintos niveles de registro así como la serialización de los mismos.
- Junit 4.0. *Framework* Java para la ejecución de pruebas unitarias con un amplio abanico de configuraciones y extras para el diseño de pruebas de integración, regresión, etc.
- Apache Lucene 2.9.0. Motor de búsqueda sintáctica en Java con amplias posibilidades tanto para el indexado de documentos, procesamiento de lenguaje natural y ejecución de consultas.
- Apache Solr 1.4.1. Plataforma empresarial de búsqueda en Java basada en Apache Lucene que añade nuevas funcionales como nuevos filtros para el procesamiento de lenguaje natural.
- Apache Mahout 0.4. Biblioteca Java con un amplio espectro de algoritmos de *data mining* y aprendizaje automático, con capacidades para integrarse con otras herramientas como Apache Hadoop y que propone un framework extensible para la creación de nuevos algoritmos.
- Spring 2.5. *Framework* Java para la creación de aplicaciones empresariales basado en la técnica inversión de control y el uso de *Plain Old Java Object* (POJOs) para el diseño flexible y el desarrollo ágil de *software*.
- Jersey REST 0.8. Biblioteca Java para la creación de servicios web REST mediante anotaciones.
- Jquery 1.4.1. *Framework* para el desarrollo de interfaces de usuario enriquecidos basados en HTML+CSS+Javascript.
- Exhibit 2.2.0. Biblioteca basada en Javascript para el desarrollo de interfaces con múltiples vistas para la presentación de datos en general.
- Pubby 0.3.3. *Linked Data frontend* para la negociación de contenido y presentación de datos enlazados disponibles a través de un *endpoint* de SPARQL.
- SNORQL. Aplicación basada en HTML+Javascript para la realización de consultas *on-line* sobre *endpoints* de SPARQL.

6.3 Consideraciones Generales de Diseño

En un sentido amplio, dado un problema y un dispositivo en el cual resolverlo, es necesario suministrar un método preciso de respuesta a la casuística planteada en el problema, de acuerdo al dispositivo objetivo, tiene la denominación de *algoritmo*. El diseño de algoritmos [147] tiene dos componentes importantes:

1. El primero se refiere a la búsqueda de métodos o procedimientos, secuencias finitas de instrucciones adecuadas al dispositivo establecido, que permita resolver el problema.
2. El segundo permite medir de alguna forma el coste (en tiempo y recursos) de consumo de un algoritmo con el fin de encontrar la solución, ofreciendo la posibilidad de comparar distintos algoritmos que resuelven un mismo problema.

Una vez que se dispone de un algoritmo que funciona correctamente, es necesario definir criterios con el objetivo de medir su rendimiento o comportamiento. Estos criterios se centran principalmente en su simplicidad y en el uso eficiente de los recursos. La sencillez es una característica sensiblemente interesante en el diseño de algoritmos, facilitando su verificación, el estudio de su eficiencia y el mantenimiento. De ahí que muchas ocasiones se priorice la simplicidad y legibilidad del código frente a alternativas más crípticas y eficientes del algoritmo. Por otra parte, el uso eficiente de los recursos suele medirse en función de dos parámetros: el *espacio*, es decir, memoria utilizada, y el *tiempo*, unidades de tiempo de ejecución. En ambos casos, se hace referencia a los costes que supone la búsqueda de la solución al problema planteado mediante un algoritmo, además estos dos parámetros son utilizados para una posible comparación ulterior de los algoritmos entre sí.

El tiempo de ejecución de un algoritmo dependerá de diversos factores como los datos de entrada suministrados, la calidad del código generado por el compilador para crear el programa objeto, la naturaleza y rapidez de las instrucciones máquina del procesador concreto que ejecute el programa, y la complejidad intrínseca del algoritmo. Existen dos tipos posibles de estudio sobre el tiempo:

1. Áquel que proporciona una medida teórica (a priori), consistiendo en la obtención de una función que acote (inferior o superiormente) el tiempo de ejecución del algoritmo para unos valores de entrada determinados.
2. Áquel que ofrece una medida real (a posteriori), consistiendo la medición del tiempo de ejecución del algoritmo para unos valores de entrada determinados y un entorno de ejecución particular.

Ambas medidas son importantes puesto que si bien la primera ofrece estimaciones del comportamiento de los algoritmos de forma independiente del ordenador en donde serán implementados y sin necesidad de ejecutarlos, la segunda representa las medidas reales del comportamiento del algoritmo.

6.3.1 Consideraciones sobre Diseño de Programas

El objetivo de implementación del sistema MOLDEAS recae en suministrar parcialmente soporte a los distintos procesos implicados en el ciclo de vida de datos enlazados, si bien algunas tareas se realizan mediante el uso de ciertas herramientas, existen otras que deben ser parametrizadas e implementadas para ofrecer un entorno homogéneo para el manejo de la información y datos provenientes de los anuncios de licitación. La separación de responsabilidades entre los distintos componentes se

realiza de acuerdo a su funcionalidad, de esta manera es posible efectuar cambios transparentes en los componentes sin que los otros sean involucrados en el proceso de cambio: implementación, prueba y empaquetamiento. Por ello, se han definido los interfaces de comunicación entre los mismos como un API para que cualquier futuro desarrollo se apoye en este *framework* para añadir nuevas funcionalidades. Las claves para un diseño abierto de un API coinciden en muchos sentidos con los de un lenguaje de programación [134]:

Concisión notacional: el API deberá proporcionar un entorno con un nivel de detalle adecuado: interfaces claras, simples, unificadas etc. Las posibles ampliaciones sobre el *framework* de *MOLDEAS* deben resultar sencillas y no presentar inconvenientes para que el programador comprenda y extienda su diseño.

Ortogonalidad: la funcionalidad del API debe suministrar el mecanismo adecuado para combinar nuevos componentes y rechazar algunos de los ya presentes. Por ejemplo la adición de nuevas restricciones no debe suponer una recodificación del código del algoritmo.

Abstracción: el diseño del API debe basarse en el uso de técnicas como los patrones de diseño e interfaces, favoreciendo la abstracción de algoritmos.

Seguridad: los algoritmos implementados deben tener puntos obligatorios de restricciones para verificar por ejemplo su terminación aunque se añadan nuevas restricciones.

Expresividad: el API debe ser lo suficientemente "rica" como para que nuevas ampliaciones puedan ser formuladas de forma sencilla de acuerdo a la información y datos presentes en los anuncios de licitación y en su modelo formal.

Extensiblidad: el API debe basarse en técnicas de programación que favorezcan la adición de nuevas características y su adaptación para nuevas configuraciones del algoritmo.

Portabilidad: el lenguaje seleccionado para proporcionar estas técnicas debe poseer esta característica.

Eficiencia: en cualquier implementación de un algoritmo o conjunto de los mismos esta propiedad es fundamental y aunque el API diseñado, atendiendo a las características anteriores pueda sobrecargar la ejecución básica de los procesos, su penalización en tiempo no es tan alta como para descartar los principios de diseño anteriores.

Librerías e interacción con el exterior: la ejecución de las funcionalidades provistas deberá ser independiente del lenguaje de ontologías utilizado, así se aisla el conjunto de técnicas de la fuente de conocimiento favoreciendo el uso del cualquier lenguaje de formalización.

Entorno: para facilitar la depuración de los algoritmos realizados se proveerá un entorno gráfico en el cual poder visualizar la ejecución.

Además de estas principios generales para el diseño del API, hay que tener en cuenta:

- El entorno de ejecución puede ser una aplicación web, por lo que se deberá tener en cuenta en el diseño para que se pueda integrar fácilmente con *frameworks* como Spring.
- La mayoría de las aplicaciones utilizando Web Semántica utilizan el lenguaje Java, por lo tanto, este será el lenguaje seleccionado en su última versión.

6.3.2 Patrones de Diseño

Con el objetivo de cumplir los criterios mencionados anteriormente, el sistema MOLDEAS se basará en la aplicación intensiva de patrones de diseño [5,129] como buena práctica de programación. Normalmente estos patrones indican la interacción que ha de realizarse entre los distintos elementos que participan en el problema, así para cada uno se cuenta con un conjunto de objetos, cada uno de los cuales realizará una función proveyendo servicios a los demás objetos implicados. Los patrones de diseño proponen soluciones con distintas características: elegantes, modulares, escalables y flexibles. Una posible definición de los mismos se presenta a continuación:

> *Los patrones de diseño representan soluciones para problemas recurrentes en la ingeniería del software.*

En general, se trata de soluciones estándar para un problema habitual en programación que utiliza distintas técnicas para la flexibilización del código, tratando al mismo tiempo de satisfacer ciertos criterios no funcionales. Se suelen asimilar a una estructura determinada de implementación que cumple una finalidad concreta y permite describir ciertos aspectos de un programa.

No obstante, pese a que los patrones ofrecen buenas soluciones puede que en algún caso resulte contraproducente el uso de los mismos, el empleo de estas técnicas debe realizarse, por tanto, en el momento justo. La problemática reside en establecer cuándo es adecuada su aplicación, así se pueden citar varios criterios:

- El código del programa ha crecido exponencialmente.
- Las clases de un programa aglutinan código semánticamente que no corresponde con su funcionalidad.
- Se deben realizar pruebas unitarias de clases.
- El diseño del programa es altamente complejo y las relaciones entre las distintas clases es opaca.
- La comunicación con otros desarrolladores ha decaído debido a la complejidad del código.

El libro *The Ganf of Four (Gof)* [129] fue pionero en introducir estas técnicas de programación caracterizando las mismas en tres niveles:

Patrones creacionales: se utilizan para la creación e inicialización, se pueden destacar: *Abstract Factory*, *Builder*, *Factory Method*, *Prototype* o *Singleton*.

Patrones estructurales: su objetivo es separar la interfaz de operaciones de la implementación. Tratan de organizar cómo las clases y objetos se agrupan para generar estructuras y organizaciones más grandes. Por ejemplo: *Adapter*, *Bridge*, *Composite*, *Decorator* o *Facade*.

Patrones de comportamiento: describen la comunicación entre los objetos. *Chain of Responsability*, *Command*, *State*, *Strategy* o *Visitor* son ejemplos de este conjunto.

En combinación con los patrones de diseño, se encuentra la técnica de *refactoring* [124], utilizada para reestructurar y refinar el código fuente (en muchos casos aplicando un determinado patrón) sin alterar la funcionalidad o comportamiento externo del mismo.

6.4 Diseño de Componentes del Sistema MOLDEAS

6.4.1 Diseño de `moldeas-common`

El objetivo de este módulo es aunar aquella funcionalidad y lógica de negocio común a todos los procesos del ciclo de vida. De esta forma y atendiendo al modelo de construcción de aplicaciones Java propuesto por Maven se pueden separar responsabilidades entre los distintos componentes, compilar y empaquetar (como fichero JAR) por separado las distintas clases de Java de manera transparente.

Entre las funcionalidades entregadas en este paquete hay que destacar las siguientes:

- Clases para la carga de ontologías y de documentos RDF desde distintas fuentes de datos como ficheros o un *endpoint* de SPARQL.

  El doble objetivo de diseño de estos paquetes es: 1) independizar la ejecución de los procesos de la base de conocimiento y 2) abstraer la localización (local, remoto, base de datos, etc.) y contenido del recurso.

  Para diseñar estos dos objetivos se han seguido los siguientes criterios:

  1. El patrón *DAO* que sirve como guía para este primer objetivo, abstrayendo las operaciones necesarias sobre la base de conocimiento a un interfaz que pueda ser implementado por distintos proveedores.
  2. El segundo objetivo, se obtiene factorizando la información invariante de los recursos: el contenido del recurso puede ser siempre un *InputStream* y todo recurso podrá ser localizado por un identificador único (nombre fichero, URL o un id. de una base datos), no obstante para facilitar la abstracción de la clave de un recurso se ha diseñado como un objeto (*KnowledgeResourcePK*) anticipando la necesidad de claves más complejas que una simple cadena identificativa.

  En la Figura 6.5, se realiza un diagrama de clases del módulo de acceso a la base de conocimiento.

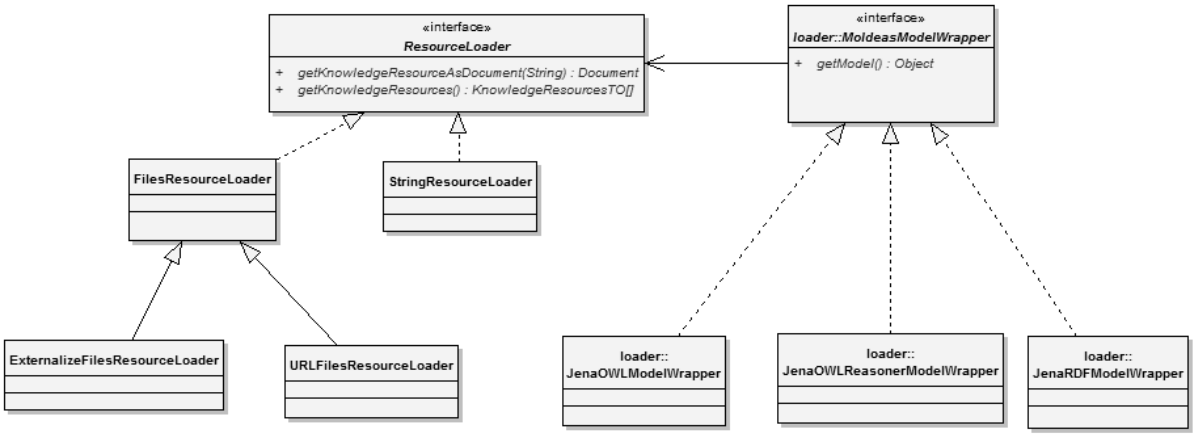


Figura 6.5: Diagrama de Clases del acceso a datos (ontologías y RDF) en MOLDEAS.

Uno de los puntos clave para facilitar un soporte escalable a los distintos lenguajes, consiste en la selección del API para trabajar con ontologías y recursos RDF, para ello y teniendo en cuenta el uso del lenguaje de ontologías OWL y su representación bajo el modelo de datos RDF en sus distintos formatos, se debe realizar una abstracción que permita procesar la información y los datos con comodidad, ocultando los detalles de representación sintáctica y ofreciendo las

abstracciones básicas de OWL y RDF: sentencias, sujeto, recurso, clases, propiedades, etc. En este sentido, las alternativas disponibles se centran en:

Jena: biblioteca Java de referencia para el trabajo en el campo de la Web Semántica conteniendo múltiples funcionalidades para el desarrollo de aplicaciones basadas en RDF y OWL. Se trata de *software libre* inicialmente desarrollado por *HPLabs* y actualmente bajo la fundación Apache.

OWL-API: es una biblioteca para Java específicamente diseñada para tratar ontologías expresadas en OWL. Se distribuye bajo licencia de *software* libre y ha sido creada como parte del proyecto europeo *WonderWeb*, en la actualidad es gestionada en la Universidad de Manchester.

Protégé-OWL-API: utilizada en el IDE Protégé para el desarrollo de ontologías en OWL. Combina aspectos tanto de OWL-API como de Jena, no obstante, su ámbito está más orientado hacia a este editor que como biblioteca para un usuario programador.

Una vez valoradas las posibilidades de estas herramientas, se ha seleccionado Jena ya que proporciona soporte para el tratamiento de varios lenguajes y formatos de modelado, permite la integración con herramientas y procesos externos como los razonadores, y además su comunidad ha crecido exponencialmente en los últimos tiempos gracias a su adhesión al proyecto Apache, lo que le confiere un grado de madurez y confianza extra que asegura un correcto funcionamiento de la misma.

- Clases para el almacenamiento de constantes a lo largo del ciclo de vida de los datos manejados por el sistema MOLDEAS. En este sentido, es conveniente parametrizar los valores de las URIs base, los grafos RDF, así como los prefijos y espacios de nombres de todos los vocabularios y *datasets* reutilizados. Esta práctica evita los errores ortográficos en la codificación de URIs tanto en la producción de datos como en su consumo a través de consultas SPARQL o mediante el propio API definido por Jena.

- Clases para la gestión de errores y excepciones que se puedan producir a lo largo de cada uno de los procesos del ciclo de vida. El diseño de la gestión de errores es uno de los elementos importantes para la robustez de un producto *software*. Las directrices que se han seguido en MOLDEAS para el control de errores es el uso de las excepciones, para ello, se ha creado una simple jerarquía de excepciones que capture los posibles errores dentro de la aplicación. Se han considerado dos tipos de excepciones, ver Figura 6.6:

  Excepciones no chequeadas: controlan errores graves producidos en tiempo de ejecución e inesperados en el modelo.

  Excepciones chequeadas: controlan errores graves que se pueden producir en tiempo de ejecución.

- Clase para la configuración del registro y traza de la aplicación. Esta característica de verdadero valor para una aplicación, permite agilizar los procesos de depuración y su integración en sistemas de mayor calado. En todas las clases del sistema MOLDEAS candidatas a generar mensajes de registro, se ha utilizado la biblioteca *Log4j* para la gestión del registro de la aplicación. Se trata de una herramienta Java para la gestión del registro y traza de acuerdo a un serie de niveles: *DEBUG, INFO, WARN, ERROR y FATAL*. Esta división por niveles permite configurar los mensajes del sistema dependiendo del estado en el que se encuentre: desarrollo, depuración, producción, etc.

- Otras clases de utilidad para acciones transversales como la carga de documentos XML, ejecución de consultas SPARQL, filtros de lenguaje natural para las bibliotecas de Apache Lucene y Solr.

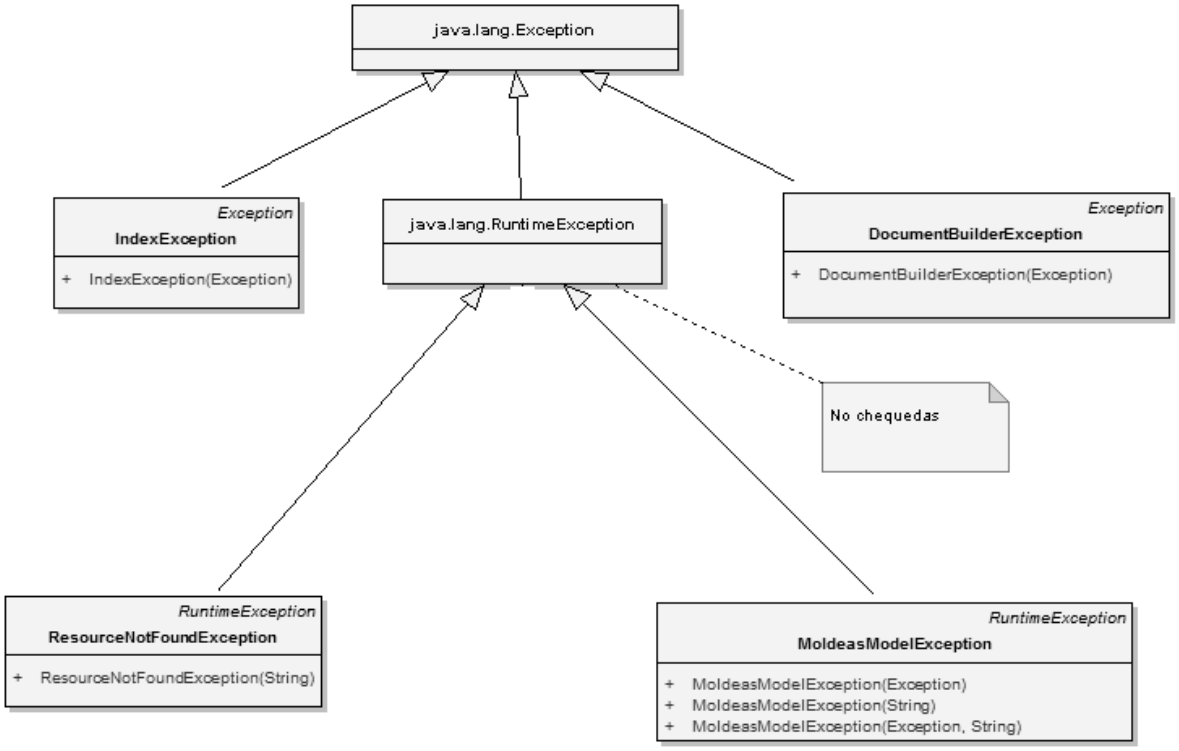


Figura 6.6: Diagrama Clases de excepciones en el sistema MOLDEAS.

6.4.2 Diseño de `moldeas-transformer`

El objetivo de este módulo es facilitar el soporte al proceso específico de producción de datos enlazados para las distintas entidades de información provenientes de los anuncios de licitación. En el enfoque de este trabajo se han tenido en cuenta los datos a transformar y promocionar a la iniciativa de *Linked Data* para agilizar las distintas tareas del modelo de ciclo de vida de datos enlazados propuestos. Existen ciertas tareas que se pueden acometer directamente con las herramientas existentes como Google Refine, para las que su funcionamiento es correcto cuando los datos son homogéneos y tan sólo requieren la ejecución de una serie de *mapeos* o reglas de transformación para utilizar los datos de entrada como valores en las tripletas RDF a generar. Sin embargo, dependiendo del tamaño del *dataset* de entrada a transformar, de la homogeneidad del mismo y de las operaciones posteriores a realizar como la adición de metainformación, la reconciliación de entidades o la simple serialización del modelo RDF en distintos formatos, hace necesario la implementación de un proceso personalizado que ejecute estas tareas de forma específica, ya que la dificultad de expresarlas en las herramientas de propósito general las convierte en tremendamente complicadas. Es por ello, que se han diseñado e implementado en este módulo una serie de funcionalidades de carácter específico para cada una de las entidades a transformar, teniendo presentes las características de las mismas. Para ello, se han considerado los siguientes puntos:

- La transformación de la información propia de los anuncios de licitación conlleva la promoción de una gran cantidad de datos, que pueden ser tomados de distintas fuentes como fichero CSV, MSExcel, XML o desde una base de datos. En este sentido, y debido a dos variables importantes, el tamaño del *dataset* y la diversidad de los formatos de entrada, se ha optado por la implementación de una serie de adaptadores que permiten procesar la información de entrada de forma homogénea de modo que la generación en RDF se convierte en un proceso transparente de la fuente de datos y se puede asegurar la validez de los datos transformados. En cuanto

a la reconciliación de entidades, para este conjunto de datos no se considera necesaria ya que se realiza específicamente en el catálogo de clasificaciones de productos, sin embargo el enriquecimiento del *dataset* de entrada si se ve afectado, ya que se han añadido nuevas propiedades (latitud y longitud) a los códigos NUTS para facilitar posteriores procesos como el de búsqueda de anuncios de licitación.

Por tanto, los anuncios de licitación se transforman mediante un proceso Java *ad-hoc* que cubre todas las tareas del ciclo de vida referentes a la producción de datos enlazados.

- En cuanto al catálogo de clasificaciones y teniendo en cuenta las características de las mismas, tamaño relativamente pequeño y formato de entrada homogéneo, se ha optado por un enfoque híbrido, en el cual en el primer paso se realiza la transformación de los datos originales mediante la herramienta Google Refine de forma que una vez obtenidos los datos en RDF se realiza una serie de pasos específicos para la reconciliación de entidades (enlazado de las distintas clasificaciones con el CPV 2008) y la adición de metainformación. Para ello, se ha implementado un proceso Java capaz de tomar como fuente de datos RDF e ir accediendo y enriqueciendo cada una de las descripciones de productos y servicios disponibles, basándose en la construcción de un reconciliador de entidades específico con las bibliotecas de Apache *Lucene* y *Solr*.
- Por último y en los datos particulares de organizaciones, países y personas se ha utilizado de nuevo un enfoque híbrido en el cual para la transformación inicial se ha optado por la herramienta de Google Refine, enriqueciendo y añadiendo a los datos RDF ya generados, las tripletas de información necesarias para cumplir con la especificación de recursos RDF realizada.

En general, este módulo consta de una serie de procesos independientes que son capaces, para cada una de las grandes entidades de información provenientes de los anuncios de licitación pública, de procesar los datos en distintos formatos, realizar el proceso de reconciliación de entidades y enriquecer los *datasets* de entrada mediante la adición de metainformación.

6.4.3 Diseño de `moldeas-api`

Atendiendo a las consideraciones generales sobre diseño, se explicarán las decisiones de diseño más relevantes para la construcción de un API que provee los métodos necesarios para el consumo de datos enlazados provenientes de la información de los anuncios de licitación, incluyendo las clasificaciones de productos, países y organizaciones.

Este módulo del sistema MOLDEAS se encarga de entregar la funcionalidad básica, a través de un fichero JAR, para dar soporte a los siguientes servicios:

- Consumo de los datos enlazados provenientes de los anuncios de licitación, clasificaciones de productos y organizaciones desde un entorno de programación y, en este caso, desde el lenguaje Java.
- Construcción de un sistema de recuperación de información, capaz de generar consultas en SPARQL para ser ejecutadas en el *endpoint* en el cual se encuentran almacenados todos los datos.

Para dar soporte a estos dos grandes servicios se han establecido una serie de paquetes, ver Figura 6.8, en los cuales se realizan las funciones necesarias que permiten la ejecución de estos servicios. En este dimensionado por paquetes se establecen las siguientes funcionalidades:

- El paquete `dao`, en el que se encuentran las interfaces que definen las operaciones necesarias para el acceso a los datos enlazados de los anuncios de licitación.

- El paquete `impl`, en el que se implementan las interfaces anteriores mediante clases que realizan el *mapeo* real desde los objetos Java de lógica de negocio, a los datos enlazados que se encuentran en un *dataset* RDF. En este caso, los datos se cargan a través de consultas en SPARQL que se pueden ejecutar contra un modelo RDF en memoria o bien remotamente en un *endpoint*. El objetivo de este paquete es proveer la carga de objetos Java con los datos enlazados provenientes de los anuncios de licitación, por ello para cada una de las entidades identificadas se suministra un interfaz de acceso que genera las consultas adecuadas en SPARQL para extraer los datos del *dataset* RDF. Este paquete conforma por tanto el nivel de acceso primario y básico a los datos enlazados, constituyendo la primera capa de lógica y acceso a datos del sistema MOLDEAS. Como ejemplo se presenta el acceso a datos para los códigos CPV mediante un diagrama de clases, ver Figura 6.7.

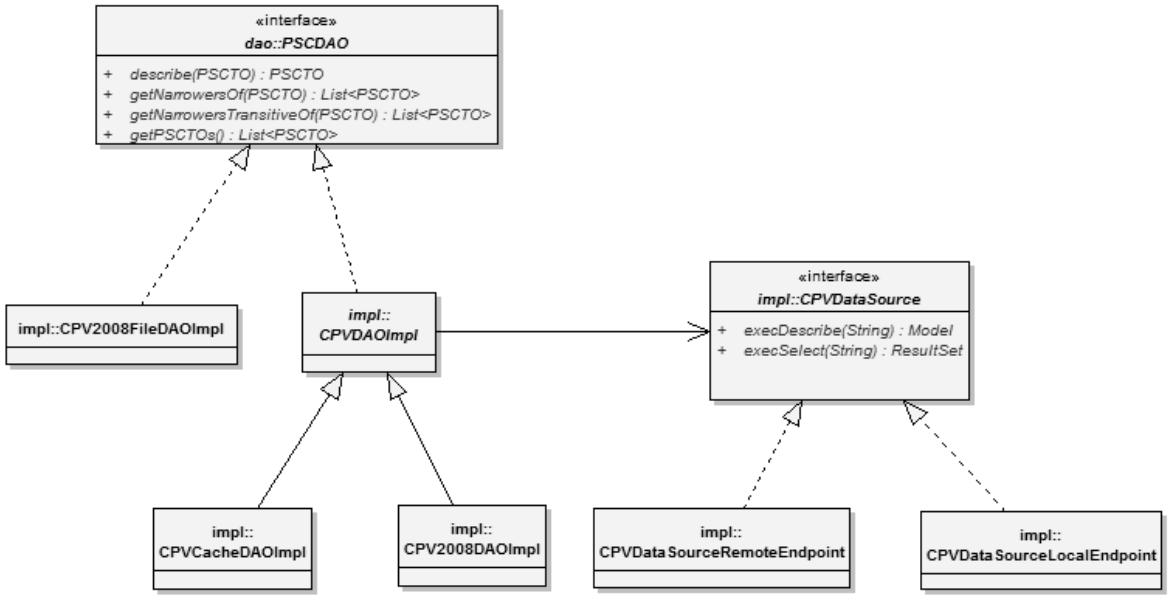


Figura 6.7: Diagrama de Clases del acceso a datos (CPV) en `moldeas-api`.

- El paquete `appserv`, en el que se encuentran las clases con los servicios de negocio propios del sistema MOLDEAS, principalmente el servicio de recuperación de información y construcción de consultas en SPARQL a partir de un perfil de búsqueda del usuario. Adicionalmente, se encuentran clases relativas a la gestión de los objetos con la información de los anuncios de licitación.
- El paquete `enhancers`, en el que se establecen las interfaces para la expansión de las variables de información correspondientes a un perfil de búsqueda de anuncios de licitación. Entre los métodos especificados para realizar la expansión de consultas se encuentran aquellos relativos a *Spreading Activation* a través de la biblioteca ONTOSPREAD [266], los basados en un motor de recomendación como Apache Mahout, los basados en un motor de búsqueda sintáctica como Apache Lucene y los relativos a las variables numéricas.
- El paquete `psc`, en el que se implementan las interfaces específicas relativas a la expansión de la variable de información de los anuncios de licitación correspondiente a los códigos de una clasificación de productos, en concreto del CPV 2008. Los métodos aquí utilizados son los relativos a la aplicación y configuración de las bibliotecas ONTOSPREAD, Apache Mahout, Lucene y Solr.
- El paquete `standalone`, en el que se implementan las interfaces específicas para otras variables de información secundarias como la información geográfica o la cuantía del anuncio de

licitación, en este caso de nuevo el principal componente utilizado es la biblioteca Apache Mahout configurada a través de distintos ficheros generados mediante el análisis del histórico de publicación de anuncios de licitación. Cada una de las implementaciones dependiendo de la información que se pretenda manejar debe tomar sus datos de una tabla previamente generada en el módulo de `moldeas-transformer`.

- El paquete `ranking`, en el que se específica y se realiza una primera implementación de los operadores de agregación para establecer un orden los anuncios de licitación recuperados del *dataset* RDF, ya sea través de un modelo en memoria o de un *endpoint* de SPARQL. Para cada una de las variables de información presentes en los anuncios de licitación se establece una ponderación, cuando los anuncios de licitación son recuperados tras el proceso de expansión, se establece una puntuación mediante una función lineal que establece un valor según el anuncio extraído contenga más o menos elementos de los conjuntos relativos a los códigos CPV, NUTS, etc.

- El paquete `business`, en el que se establece el interfaz de negocio para interactuar con los servicios del API, de esta manera se suministra un único punto de acceso como una caja negra, favoreciendo la abstracción del número de parámetros y las operaciones.

Una vez revisados los principales paquetes de `moldeas-api`, cabe realizar una descripción más completa del sistema de recuperación de información ya que el sistema de expansión de consultas se implementa a través de un patrón de diseño denominado `Chain of Responsibility` en el cual a través de distintas iteraciones y modificaciones de la consulta inicial se consigue un perfil enriquecido para ser traducido a una consulta en SPARQL. Evidentemente, el objeto contenedor del perfil de búsqueda de anuncios de licitación puede ser transformado a cualquier lenguaje de consulta tipo SQL o bien a una consulta de un motor de búsqueda sintáctica, esta abstracción se consigue de nuevo gracias a un uso correcto de interfaces. En primer lugar cabe especificar qué tipo de información y datos contiene un anuncio de licitación que como ya se ha repasado en el Capítulo 5 y concretamente en la Sección 5.1, consta de una serie de variables de información conteniendo códigos del tipo de licitación, información geográfica, metainformación relativa a la cuantía del contrato, el perfil del contrante, la duración, etc. En general, y de acuerdo a la información disponible y su tipo se pueden establecer una serie de métodos para realizar la expansión de una consulta o perfil de búsqueda.

- Información basada en conjuntos y modelada a través de una jerarquía, es el supuesto de los códigos CPV y de la información geográfica disponible en NUTS. En este caso los métodos que se pueden aplicar para expandir la consulta son los siguientes:

  1. Directo. Se navega a través de la jerarquía modelada buscando los elementos más específicos y estableciendo un peso determinado para cada uno de ellos.
  2. Búsqueda Sintáctica. De acuerdo a las descripciones de los códigos se realiza una búsqueda en el *dataset* correspondiente, buscando no sólo aquellos códigos que encajen perfectamente, sino también aquellos similares, mediante técnicas de procesamiento de lenguaje natural.
  3. Motor de Recomendación. Realizando un procesamiento previo del histórico de anuncios de licitación y su información, se generan una serie de ficheros en los cuales se alinean los códigos CPV con la información relativa a la localización, cuantía o fecha, para que de acuerdo a una serie de códigos de entrada se genere una serie de información de salida.
  4. `Spreading Activation`. En este caso se configura esta técnica para que tome un conjunto de conceptos de entrada, en general de códigos CPV, para obtener tras la ejecución del algoritmo un conjunto de salida ponderado.

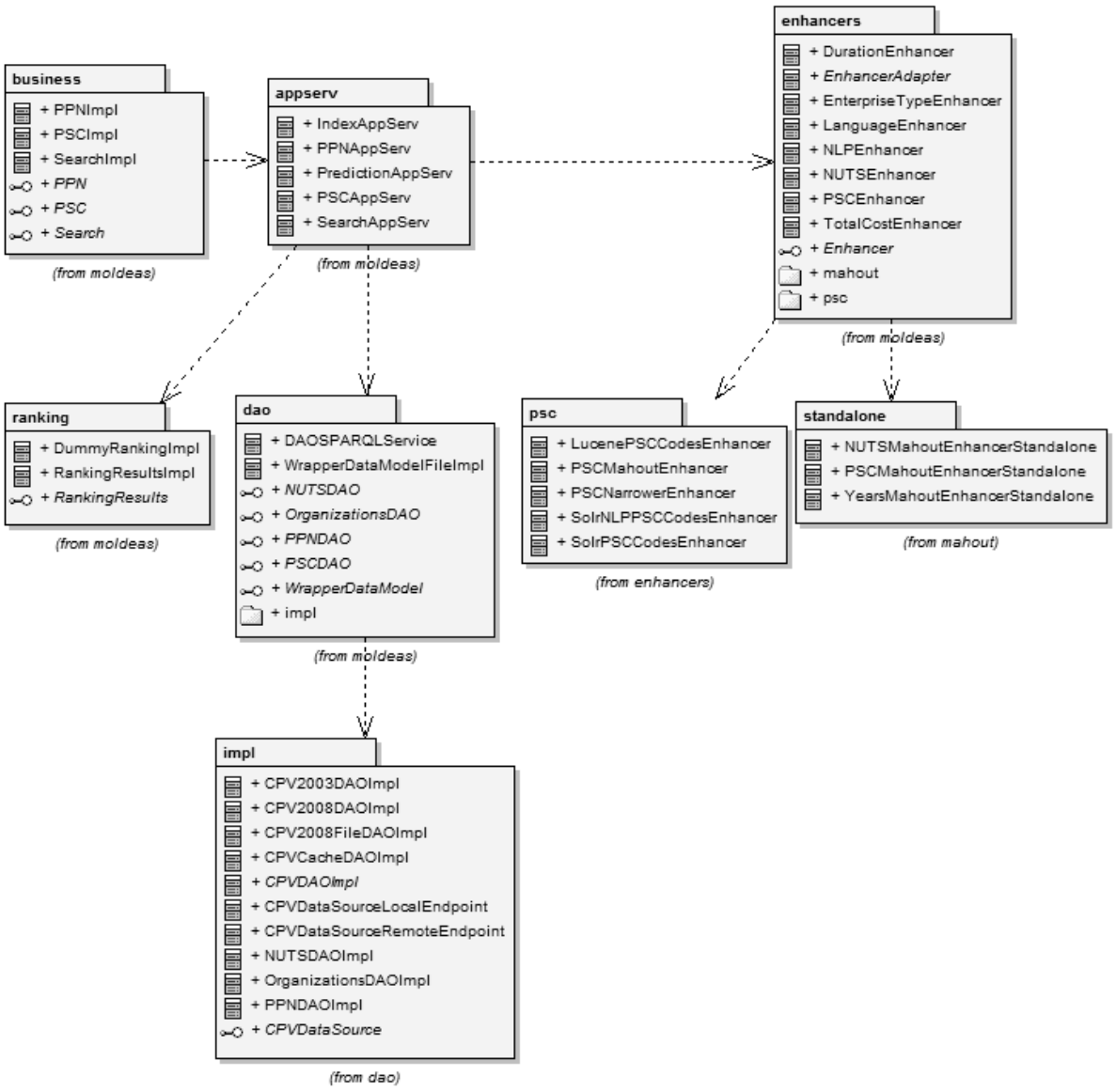


Figura 6.8: Diagrama de Paquetes relevantes del componente `moldeas-api`.

- Información numérica o asimilable. Se trata el caso de la duración del contrato, la cuantía del mismo, el radio de acción geográfico, etc. En este caso los métodos que se pueden aplicar para expandir la consulta son los siguientes:

  1. Basado en el usuario. El perfil de búsqueda del propio usuario define los rangos en los cuales se debe hallar una variable numérica sin necesidad de realizar ningún proceso de expansión automático.
  2. Uso de lógica borrosa. Se ha valorado el uso de estas técnicas para establecer un intervalo en el cual ponderar la pertenencia de un valor a un conjunto. No obstante, en esta primera versión del demostrador público estas técnicas implementadas a través de la biblioteca *JFuzzyLogic* tan sólo se han probado y no se han introducido como parte del proceso de búsqueda.

El objetivo final de estos métodos es obtener para cada una de las variables de información de un perfil de anuncio de licitación, una puntuación que permite ejecutar un operador de agregación para así establecer un orden entre los anuncios de licitación extraídos tras la generación de la consulta.

La importancia de este procedimiento por fases reside en que cada uno de los pasos y métodos de expansión se modelan mediante interfaces y son ejecutados mediante un proceso iterativo de enriquecimiento, como se puede ver en la Figura 6.9, que permite la adición de nuevos métodos de forma sencilla, simplemente reconfigurando las implementaciones de los interfaces a través del fichero de configuración de Spring. El flujo y la comunicación entre las distintas clases que intervienen en el proceso de expansión y recuperación de información se presenta a través de la Figura 6.10, en el cual es preciso señalar como las responsabilidades son compartidas a través de los distintos objetos situados en distintas capas lógicas.

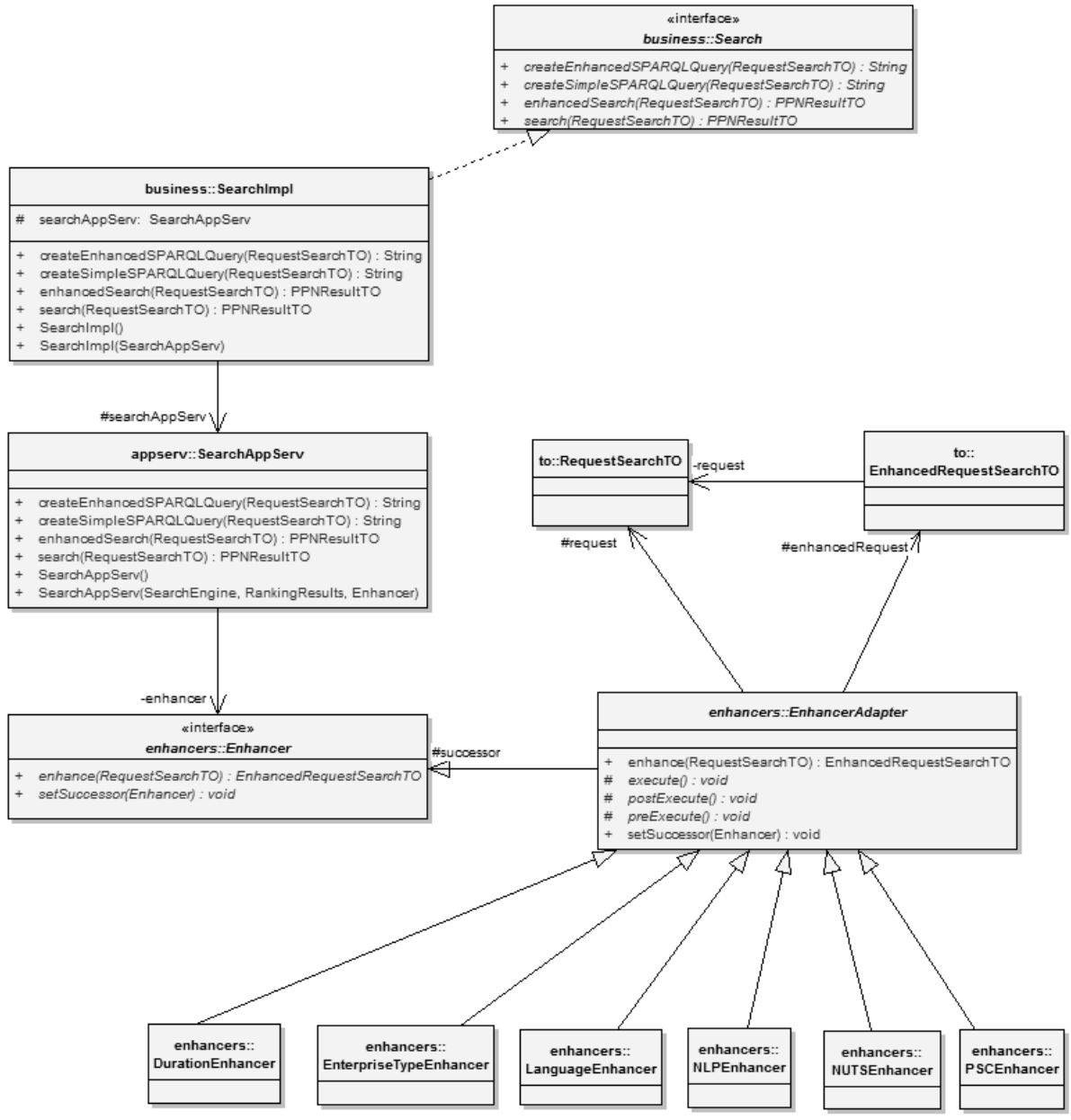


Figura 6.9: Diagrama de Clases del sistema de búsqueda en `moldeas-api`.

Finalmente, en este apartado cabe señalar los patrones de diseño, ver Tabla 6.1, que se han aplicado para obtener un sistema flexible y escalable en el cual nuevas implementaciones de los interfaces sean fácilmente integrables proporcionando un API para el consumo de datos enlazados y recuperación de información de los anuncios de licitación.

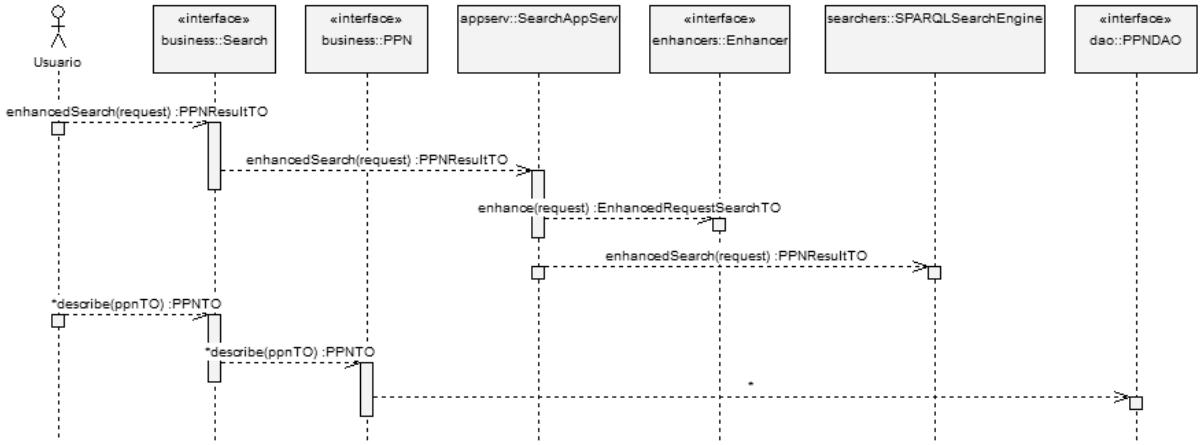


Figura 6.10: Diagrama de Secuencia de la búsqueda en `moldeas-api`.

Patrones de Diseño	
Nombre	**Aplicación**
Adapter	Implementación para el acceso y transformación de los datos de los anuncios de licitación.
Chain of Responsibility	Implantación del sistema de enriquecimiento de consultas.
DAO	Abstracción de acceso a la base de datos y de conocimiento.
Template Method	Implementación por omisión de métodos en el patrón *Chain of Responsibility*.
Factory Method	Creación de objetos provenientes de Spring.
Template Method	Funciones con llamadas a métodos de interfaces, por ejemplo en los `Enhancers`.
Transfer Object	Comunicación de información entre los distintos módulos y capas.
Singleton	Creación de objetos de acceso a datos, etc., provenientes de Spring.
Otros	Relativos al diseño por capas de aplicaciones J2EE.

Tabla 6.1: Principales Patrones de Diseño utilizados en MOLDEAS.

6.4.4 Diseño de `moldeas-test`

El objetivo de este módulo es abstraer las pruebas de los módulos anteriores, especialmente de `moldeas-api`, para así proveer un sistema separado de ejecución de *tests* en el cual se puedan realizar las siguientes acciones:

- Creación de configuraciones del API de MOLDEAS a través de distintos ficheros de Spring que permiten probar de forma automática la combinación de los distintos métodos de expansión de consultas así como verificar sus resultados.
- Validación de los recursos RDF generados de acuerdo a unas reglas de validación extraídas de las tablas de validación establecidas en el Apéndice C.

De esta manera se ofrece un módulo separado con un doble sentido, para la realización de pruebas de caja negra de los servicios disponibles en el API de MOLDEAS, así como para la propia validación de los datos enlazados. La configuración de este módulo se realiza a través de un fichero XML *ad-hoc* en el cual se cargan las configuraciones para ejecutar los *tests* pertinentes (en forma de plantilla) mediante Junit. No obstante, en la Sección 6.5 se realiza una descripción más detallada de las pruebas llevadas a cabo.

6.4.5 Diseño de `moldeas-web`

El objetivo principal de diseño de este módulo es definir y diseñar un herramienta de acceso gráfico a los servicios del sistema MOLDEAS. Es importante destacar que este cliente gráfico se realiza con el objetivo de facilitar un demostrador público ya que el uso de la biblioteca de `moldeas-api` es perfectamente válido desde cualquier programa así como desde el interfaz de servicios REST, que también se incluye en este módulo. Por tanto, este módulo surge para abarcar las siguientes funcionalidades:

1. Proveer un interfaz de servicios REST que sirva tanto para exponer los servicios de negocio, como para ejemplificar las llamadas del API de MOLDEAS.

   El interfaz de servicios REST diseñado, simplemente añade una capa extra de indirección a los servicios de negocio suministrados en el API de MOLDEAS, se duplica la signatura de los métodos en una nueva capa simplificando las llamadas. La idea de realizar un interfaz REST surge para dar soporte a la tendencia actual de publicación de servicios web HTTP y también para ejemplificar las llamadas al API (construcción de parámetros, gestión de los resultados, etc.), en cualquier caso, cualquier cliente siempre puede utilizar la versión de `moldeas-api` empaquetada individualmente. Para realizar este API de servicios REST se ha utilizado la biblioteca para Java-Jersey REST 0.8 y la descripción de los mismos se presenta, parcialmente, en la Figura 6.11 en formato WADL.

2. Suministrar un interfaz gráfico que ejemplifique las llamadas a los servicios REST y que a su vez sirva como demostrador público para el acceso a los datos de los anuncios de licitación y para la recuperación de información. La construcción de este interfaz se ha desarrollado utilizando bibliotecas de Jquery y HTML facilitando la creación de interfaces enriquecidos de programación sencilla.

3. Aunar las herramientas externas de publicación y acceso a datos enlazados en un sólo punto de entrada. En este sentido, el interfaz web creado también contempla ejemplos de consultas en SPARQL a realizar mediante la herramienta SNORQL y enlaces a ejemplos de recursos que son consultados a través de un *Linked Data front-end* como Pubby.

```
<?xml version="1.0"?>
<application xmlns="http://research.sun.com/wadl/2006/10">
  <resources base="http://156.35.31.156/moldeas/services/">
    ...
    <resource path="/moldeas">
      <resource path="search">
        <method name="GET">
          <request>
            <param xmlns:xs="http://www.w3.org/2001/XMLSchema" type="xs:string" style
                ="query" name="cpvCodes"/>
            <param xmlns:xs="http://www.w3.org/2001/XMLSchema" type="xs:string" style
                ="query" name="nutsCodes"/>
            ...
            <param xmlns:xs="http://www.w3.org/2001/XMLSchema" type="xs:string" style
                ="query" name="maxDuration"/>
          </request>
          <response>
            ...
          </response>
        </method>
      </resource>
      <resource path="search/code/{code}">
        <param xmlns:xs="http://www.w3.org/2001/XMLSchema" type="xs:string" style="
            template" name="code"/>
        <method name="GET">
          <response>
          ...
          </response>
        </method>
      </resource>
     <resource path="search/enhanced/{code}">
        <param xmlns:xs="http://www.w3.org/2001/XMLSchema" type="xs:string" style="
            template" name="code"/>
        <method name="GET">
          <response>
        ...
          </response>
        </method>
      </resource>
      <resource path="search/createQuery/code/{code}">
        <param xmlns:xs="http://www.w3.org/2001/XMLSchema" type="xs:string" style="
            template" name="code"/>
        <method name="GET">
          <response>
          ...
          </response>
        </method>
      </resource>
      <resource path="search/createQuery/enhanced/{code}">
        <param xmlns:xs="http://www.w3.org/2001/XMLSchema" type="xs:string" style="
            template" name="code"/>
        <method name="GET">
          <response>
            ...
          </response>
        </method>
      </resource>
    </resource>
        ...
   </resources>
</application>
```

Figura 6.11: Interfaz REST en formato WADL.

6.4.5.1 Conceptos y Diseño de la Interacción Gráfica

En los últimos años, el aumento en el uso de las nuevas tecnologías de la información ha conllevado un cambio fundamental en la sociedad que provoca la necesidad de adecuarse a nuevas relaciones que se establecen entre la persona y la máquina. Esta necesidad de interacción supone que la comunicación entre persona y máquina adquiera una especial relevancia en nuestro tiempo. Se maniesta un desafío consistente en proporcionar de significado a los productos y servicios ofrecidos, intentando dotarlos de estructura y comprensión cercana a la escala humana.

Debido a la complejidad interna de los ordenadores, la comunicación entre los entes debe realizarse a un nivel conceptual, predominando la comunicación visual, gráfica y verbal como medio para llegar a la comprensión. La idea subyacente está referida a la necesidad de encontrar el modelo ideal que permita a la persona interaccionar con las máquinas como si fueran un humano más.

En esta situación se hace especialmente relevante la generación de interfaces de usuario que se adecúen a las necesidades de las personas, por ello las reglas propuestas por *Sneiderman* suponen un primer acercamiento a la interacción persona ordenador y una métrica para evaluar la bondad de un interfaz de usuario.

En este sentido, desde los anales de la historia los hombres han ideado formas e instrumentos (lenguajes de símbolos, de texto, gráficos, etc.), tanto para la comunicación de sus experiencias, como para reproducir la realidad. La forma de expresión depende en buena medida del ambiente tecnológico y de los lenguajes utilizados entre el emisor y el intérprete para decodificar los mensajes.

La riqueza de expresión de un lenguaje se basa en el principio de concordancia entre los entes participantes del acto comunicativo, para que se establezca este acto deben estar en correcta relación. Desde el punto de vista de un interfaz gráfico estos factores quedan patentes por la dificultad que entraña ofrecer un entorno en el cual tanto: percepción, interpretación y comprensión sean lo suficientemente claros para que el ser humano no se sienta en un entorno extraño.

Los seres humanos a través de su sistema nervioso son capaces de percibir lo que sucede en su entorno y actuar en consecuencia, puede considerarse que el hombre es una fuente de información, percibe y recibe estímulos externos que le permiten adaptar sus respuestas.

Comprender el comportamiento humano como sistema de comunicación supone una gran dificultad, tanto en capacidad para interpretar, como para decodificar los mensajes, haciendo imposible conocer su respuesta exacta ante la recepción de un mensaje. El problema de la comunicación entre personas y máquinas no consiste en reducir el comportamiento humano al de una máquina, sino analizar las necesidades y capacidades para adaptar la presentación de información a su nivel, acercándose lo más posible al proceso cognitivo de la persona.

La visión es nuestro medio natural de percepción de los mensajes, limitada por condiciones físicas, por ello cobra especial relevancia para la distribución de información en un documento, dependiendo la compresión de la persona de la bondad de la misma. No obstante, este modelo a seguir no es completamente válido ya que habría que tener en cuenta las necesidades de las personas con discapacidades de distinto tipo, intentado adaptar el contenido al sujeto participante de la comunicación.

Utilizando los distintos sentidos de los que disponemos como fuentes de adquisición de información debemos ser capaces de procesar y comprender esta información, esta capacidad puede denominarse *entendimiento humano*, mediante el que determinamos nuestros actos a partir de las percepciones sensoriales. De todos los estímulos percibidos, sólo permanecen aquellos que generan un mayor impacto, este será un factor muy relevante en el diseño de un interfaz, por ejemplo cuando se selecciona una combinación de colores.

Debe tenerse en cuenta, que aunque el funcionamiento general humano es similar no puede ser predecible, ni aplicable por extensión, se cometería en un grave error de interpretación del com-

portamiento, este punto es especialmente interesante a la hora de adaptar un interfaz gráfico a las distintas personas, estableciendo perfiles.

Las 8 reglas de Oro de *Sneiderman* nos ayudan a este objetivo, intentando simplificar la comunicación persona-máquina. A continuación, se comentan estas reglas:

Esforzarse por la consistencia. Es el principio por el cual los elementos y relaciones deben ser presentados de forma idéntica e inequívocamente. Este concepto es aplicable a:

- Tipografía, por ejemplo enfatizar siempre con el mismo estilo.
- Iconos, la misma acción debe estar representada por el mismo icono.
- Comandos y menús, la información que se utilice en los menús deberá ser representativa de su acción interna.
- Funcionamiento multiplataforma, esta cualidad se pone de manifiesto actualmente, ya que se accede a la misma información utilizando diferentes medios: navegadores, móviles, etc.
- Percepción del sistema igual para todos los usuarios.
- Estructura el sistema de forma que se adapten al entorno de trabajo del usuario.

Por lo tanto la consistencia nos remite a obtener una representación de los objetos coherente en su significado, formas y métodos en el mismo contexto. También puede entenderse la consistencia como la relación que se establece entre una metáfora de representación y su uso real.

Permitir a los usuarios frecuentes utilizar atajos: se considera un usuario frecuente a una persona que interacciona con un interfaz gráfico con soltura, debido sobre todo a la experiencia en su uso. Este tipo de usuarios, a medida que avanza su manejo del interfaz, se convierten en expertos, exigiendo mayores prestaciones a la aplicación desde distintos puntos de vista: 1) adaptabilidad y 2) velocidad de ejecución.

Esta característica a tener en cuenta en los interfaces, se aplica no sólo a los usuarios expertos sino a la necesidad de que un interfaz sea adaptable a las nuevas necesidades que se manifiestan en el usuario a través del uso.

Por otro lado, los atajos deben ser fáciles de recordar y mejorar el rendimiento, por ejemplo sería contraproducente un atajo de teclado de cuatro teclas ya que sólo el posicionamiento correcto de los dedos sería más penalizador que utilizar un ratón.

Ofrecer retroalimentación informativa. Debido al concepto de acto comunicativo, resulta importante que cuando un usuario realiza una acción se vea recompensada con una respuesta del sistema. De esta forma, se consigue que se establezca la metáfora de la comunicación entre el emisor (usuario) y el receptor (máquina), permitiendo una comunicación fluida.

Este objetivo es relevante en cuestiones relacionadas con el establecimiento del éxito o no de la operación realizada por el usuario, por ejemplo si en un navegador no se dispone de conexión a internet y se solicita el acceso a una página, el proceso de espera sería tedioso, utilizando una barra de progreso de carga se consigue realimentar el estado de la acción solicitada e iniciada por el usuario.

Diseñar diálogos para fundamentar el cierre. Con esta característica se consigue acotar el inicio de una secuencia de acciones, fijando su inicio, desarrollo y final. Respetando este punto se consigue que el usuario se sienta satisfecho, ya que sabe si la acción que ha realizado ha seguido la secuencia correcta, utilizando todos los pasos previstos hasta llegar a un fin.

Ofrecer prevención de errores y manejo de errores simples. La gestión de errores es importante en cualquiera de los aspectos dentro de una aplicación informática. En el contexto de un interfaz

de usuario queda patente la necesidad y las ventajas que conlleva disponer de un interfaz que minimice los errores, involuntarios o no, provocados por los usuarios.

Un escenario de uso podría ser la introducción de una fecha, generalmente para facilitar la interacción con el usuario se genera una máscara de entrada o bien un ejemplo que permita al usuario cambiarlo para así facilitar sus tareas.

El objetivo, es que el usuario se preocupe de sus tareas y no del funcionamiento del interfaz, es decir, que el manejo del mismo no suponga un trabajo extra.

No obstante, este objetivo no siempre se puede conseguir, de ahí que cuando se produzcan errores no críticos de la aplicación, tales como errores en la introducción de datos o una secuencia de pasos inválidos para realizar una acción, sería conveniente que el sistema fuera capaz de ofrecer un mecanismo tutor para enseñar al usuario al manejo sin errores. Un claro ejemplo podría ser el famoso "ayudante" de Microsoft Word que si bien realiza su función, en ocasiones llega a resultar molesto.

Los tutores dentro de las aplicaciones para aleccionar a los usuarios, y prevenir errores, se deben diseñar con sumo cuidado y acotando muy bien su ámbito de acción ya que su efecto puede ser el contrario del pretendido. En estos casos, el menos es más puede resultar útil y un simple texto desplegable con un ejemplo puede ser suficiente.

Permitir inversión de las acciones. Esta característica ofrece la admirable ventaja de ofrecer confianza al usuario en su diálogo con el interfaz gráfico. Suponiendo que un usuario novel realiza una serie de acciones secuenciales y antes de grabar los cambios se percata de que ha cometido un fallo, podría pensar en distintos escenarios, así:

- Tranquilidad, el interfaz permite deshacer cambios, se puede retroceder para corregir lo que se ha realizado de forma incorrecta.
- Desconsuelo, después de estar X tiempo realizando una tarea se tiene que deshacer debido a un insignificante error, tiempo perdido.

Parece claro que el sentimiento de tranquilidad es más deseable en un usuario, así que es importante cumplir esta característica.

Desde otro punto de vista, el usuario se sentirá más tranquilo y seguro en el uso del interfaz si sabe que tiene más de una oportunidad cuando utiliza las funciones de la aplicación, de otra forma un sentimiento de angustia podría invadirle produciendo errores en el usuario y empeorando su rendimiento en el uso de la aplicación.

El sentido de estos cambios debería ser tanto hacia adelante como hacia atrás, ya que sino se desaprovecha la potencia de esta propiedad.

Apoyar el control de los usuarios: en la tarea de satisfacer al usuario, éste debe percibir que el sistema está ejecutándose bajo su supervisión y respondiendo a sus acciones. Todo aquello que sea público al usuario deberá permitir su control, siempre dentro de unos límites, ya que de otra forma quizás se obtuviera un sentimiento de insatisfacción.

Debería evitarse por ejemplo introducción de datos tediosa, comportamiento extraño de la aplicación (acciones o comandos que no funcionan o no responden) o imposibilidad de obtener información (sabemos que se ha introducido pero se desconoce cómo acceder a ello).

Reducir la carga de la memoria a corto plazo: esta necesidad surge por una limitación en la memoria humana para procesar información a corto plazo, por ello, cuando un interfaz emita mensajes al usuario deberán ser cortos de modo que pueda comprender rápidamente la información que se envía desde la máquina.

También es conveniente no tener demasiadas ventanas abiertas dentro de un entorno, ya que pueden despistar al usuario y no saber dónde está exactamente.

El diseño de un interfaz gráfico para el sistema MOLDEAS es una tarea relativamente sencilla, ya que el problema real consiste en suministrar un interfaz habitual al usuario en el cual pueda seleccionar las características de su perfil de búsqueda. En cualquier caso, se optado por un interfaz gráfico lo más sencillo, intuitivo, consistente y uniforme posible.

6.5 Pruebas del Sistema MOLDEAS

La metodología de la Programación Extrema [15] (XP), que ha servido de guía para el desarrollo del sistema MOLDEAS, hace hincapié en la importancia de las pruebas. La realización de ciclos de desarrollo breves (incorporando la funcionalidad perseguida en pequeñas versiones totalmente funcionales) habilita la posibilidad de concentrarse en la validación, en cada momento, de un segmento reducido del código o de la funcionalidad, dependiendo del nivel de la prueba. Pero también exige la existencia de un mecanismo que permita realizar los ensayos de forma automática; de lo contrario, la obligación de repetirlos frecuentemente los convertiría en extremadamente tediosos.

6.5.1 Pruebas Caja Blanca

Son pruebas que se realizan para verificar la validez de las operaciones internas del programa, intentando garantizar que todos los caminos de ejecución del programa han sido probados. Para realizar las pruebas de caja blanca se han seleccionado las pruebas unitarias por su importancia dentro la metodología de Programación Extrema. La descripción de las pruebas unitarias se amplia a través de los siguientes puntos.

Validación con pruebas unitarias: representan el nivel más bajo de prueba del *software*, permitiendo al desarrollador obtener retroalimentación instántanea del correcto comportamiento (o no) de un componente determinado.

Necesidad de las pruebas unitarias: como buena práctica de programación el desarrollador debería codificar la prueba antes que la propia funcionalidad, para así pensar los resultados antes de obtenerlos. Como norma general, un buen test siempre debería fallar la primera vez. Por otra parte, facilitan la depuración, si un test falla se dispone de un ámbito muy reducido en el cual localizar el error, incrementándose la productividad de los desarrolladores.

Además, desde el punto de vista de la gestión de un proyecto un buen conjunto de pruebas unitarias se pueden utilizar como documentación del código fuente, así, por ejemplo si se introducen nuevos desarrolladores éstos pueden comprender fácilmente la funcionalidad que se está realizando y probando.

Codificación de Pruebas: las pruebas unitarias deben seguir una serie de criterios:

- Prueban funcionalidad determinada, un método.
- Ejecución rápida, si no fuera así los test sólo se ejecutarían cada cierto tiempo repercutiendo en la integración del código.
- Se debe determinar sobre qué clases se van a generar casos de prueba. No todas las clases y métodos son candidatos a ser probados.
- Son independientes de otras pruebas.
- Deben ser independientes de terceros, por ejemplo de una base de datos.

***Refactoring*:** despues de que un componente haya pasado un test unitario, ya pueden aplicarse las técnicas de *refactoring* para mejorar la calidad de código. Este proceso, conjunto de prueba y recodificación, ya puede realizarse porque se dispone de una manera de probar la funcionalidad de forma automática. En las primeras etapas de desarrollo será cuando se haga más intenso el uso de pruebas unitarias ya que se está codificando una nueva funcionalidad en cada momento.

6.5.1.1 Junit

Herramienta Java para la realización de pruebas unitarias de programas, ejecuta test de pruebas de forma automática realizando un registro de los resultados obtenidos. El gran éxito de esta herramienta se debe a su facilidad de uso en comparación con el trabajo que realiza.

La realización de las pruebas de los programas normalmente resultan costosas temporalmente y muchas veces no se almacenan ni se documentan. Con esta herramienta, se consigue que la prueba de programas rebaje su coste, además de liberarnos de una parte de la construcción de programas que en muchos casos resulta pesada y poco eficiente.

Por tanto, con Junit se aumenta el nivel de calidad en el test de los programas, entendiendo por calidad en el test una mejora en la realización de las pruebas en cuanto a casos contemplados, número y registro de las mismas y facilidad de uso.

Las ventajas de uso de Junit se pueden resumir en los siguientes puntos:

- Facilidad de uso.
- Creación de conjuntos de pruebas para una clase.
- Combinación con otras herramientas de gestión de proyectos como Maven.
- Código libre.
- Bien documentado.
- Pruebas a diferentes niveles y capas.
- Extensible.

Cabe señalar que existen extensiones de Junit que permiten realizar pruebas en distintos ámbitos, siempre y cuando se haya seguido un diseño apropiado. De esta manera existen *frameworks* basados en Junit para probar:

- XmlUnit, para documentos XML.
- DbUnit, para probar el estado de la base de datos.
- EasyMock, para generar las clases que basadas en las interfaces de negocio permitan probar los servicios y funcionalidad de la aplicación.
- HttpUnit, JWebUnit o Cactus para las distintas capas de una arquitectura cliente/servidor.

El contexto del sistema MOLDEAS se han realizado 140 *tests* enclavados dentro de 48 clases Java específicas para Junit.

6.5.2 Pruebas de Caja Negra

Son pruebas que se centran en los requisitos funcionales del *software*. Existen diferentes tipos como, la prueba de los valores límites o partición equivalente. Para realizar este tipo de pruebas se utilizan posibles entradas al sistema y se verifica que cumple con los resultados esperados. Las pruebas realizadas para la validación del sistema de recuperación de anuncios de licitación se pueden considerar incluidas dentro de este tipo de pruebas de caja negra y teniendo en cuenta que el sistema MOLDEAS trabaja con una configuración externa, su funcionamiento dependerá de la misma y no del programa en sí, es decir, desde un punto de vista de software la implementación realizada representa un conjunto de configuraciones a través de Spring. Considerando el carácter investigador de este estudio, resultan más relevantes las pruebas de validación científica que las propias de caja negra, más apropiadas desde un punto de vista de aplicación de *software*.

6.5.3 Aportaciones Finales sobre las pruebas

La codificación de pruebas unitarias antes de implementar cierta funcionalidad puede resultar tediosa e incluso se considera como un retraso en el desarrollo, sin embargo disponer de un método eficaz para realizar pruebas ayuda en posteriores etapas a la codificación de determinada función y asegura una cierta calidad en el *software*.

Por ejemplo, se realiza un determinado método para una cierta tarea, para ejecutar la prueba se utiliza un método *main* en la propia clase donde se ha definido y una vez que se comprueba que todo funciona correctamente, se elimina ese método *main*, otorgando a esa funcionalidad la etiqueta de éxito. Si con posterioridad, nuevas situaciones obligan a modificar el código de esa clase porque no se habían contemplado todos los requisitos funcionales, pueden surgir las siguientes cuestiones: ¿quién está dispuesto a recodificar esa función?, ¿cuál era su comportamiento?, ¿para qué utilizaba esa variable? y ¿cómo se prueba que mantiene su funcionalidad anterior (necesario para otros módulos) y que además incorpora la nueva funcionalidad?. La respuesta es sencilla, se volverá a codificar un método *main*, creando un nuevo caso de prueba y prácticamente se delega en la "suerte" la entrega de un comportamiento correcto tanto para la versión anterior como para las nuevas funcionalidades. En conclusión, una prueba unitaria hubiera facilitado el desarrollo, ya que tan sólo se debería generar un caso de prueba para la nueva funcionalidad. Pero en cambio, se ha perdido tiempo: codificando métodos que posteriormente serán eliminados, pensando en qué hacía la función, para qué se utiliza cierta variable, etc.

En definitiva, las pruebas unitarias facilitan el desarrollo de *software* de calidad, en contraposición de los engendros (no desarrollos) de *software* creados al margen de la realización de pruebas.

6.5.4 Métricas de código fuente

La calidad del *software*, en cuanto a código fuente, es uno de los requisitos que se debe exigir a cualquier aplicación. Normalmente, no se plantea como uno de los requisitos no funcionales prioritarios y esto provoca que aunque se haya realizado un buen análisis e incluso un buen diseño, el código generado no es lo suficientemente bueno y es el origen de todos los posibles errores y en consecuencia no se cumple con los requisitos funcionales establecidos. La definición de calidad de *software*, de forma genérica, es mucho más extensa:

> *Concordancia con los requisitos funcionales y de rendimiento explícitamente establecidos con los estándares de desarrollo explícitamente documentados y con las características implícitas que se espera de todo software desarrollado profesionalmente*. R.S. Pressman (1992).

> *El conjunto de características de una entidad que le confieren su aptitud para satisfacer las necesidades expresadas y las implícitas*. ISO 8402 (UNE 66-001-92).

Estas definiciones de calidad se refieren sobre todo a la metodología utilizada y al grado de satisfacción de los requisitos funcionales del producto *software*.

Las métricas que aquí se presentan constituyen medidas de calidad del código fuente, ya que se estima que los requisitos funcionales han sido cubiertos al obtener un prototipo experimental que ha sido probado a través de distintos conjuntos de pruebas.

A continuación, se presentan una serie de métricas de calidad de código fuente (sólo del componente *moldes-api*), obtenidas mediante el *plugin* Metrics de Eclipse, que permite generar una serie de medidas, ver Tabla 6.2, significativas del código fuente, basadas en las definiciones realizadas en el libro *Object-Oriented Metrics, measures of Complexity* de *Brian Henderson-Sellers*, Prentice Hall, 1996.

Estas métricas miden, principalmente, valores de cohesión y acoplamiento que permiten establecer si el código está bien estructurado. Sin valorar detalles precisos de los tipos y niveles de cohesión y acoplamiento, estas dos características se pueden definir como:

> *Cohesión, consecuencia del ocultamiento de la información. Un módulo con cohesión (Pressman en "Ingeniería del Software: Un enfoque práctico") realiza solamente una tarea, requiriendo poca interacción con el resto de los procedimientos que se realizan en el resto del sistema de software.*

Según *Sommerville* en "Requirements Engineering. Processes and Techniques" la cohesión es deseable debido a que una unidad (componente) representa una parte simple de solución. Si es necesario cambiar el sistema, la parte correspondiente está en un solo lugar y lo que se desee hacer con el mismo estará encapsulado en él. La meta, según *Lawrence* en "Ingeniería del *software*: Teoría y práctica", es lograr que los componentes sean lo más cohesivos posible.

> *Acoplamiento, está relacionado con la cohesión. Es un indicador de la fuerza de interconexión entre los componentes o elementos de la arquitectura*

Los sistemas altamente acoplados tienen una fuerte interconexión, lo que se refleja en una gran dependencia entre los componentes. Los sistemas poco acoplados, por otro lado, tienen poca relación entre sus componentes o elementos. El objetivo, según (*Lawrence*), es mantener el acoplamiento en el nivel más bajo posible; la conectividad sencilla entre módulos da como resultado un *software* que es más fácil de comprender y menos propenso al "efecto onda" producido cuando los errores aparecen en una posición y se propagan a lo largo del sistema.

Métricas de código fuente		
Tipo	**Descripción (EN)**	**Descripción (ES)**
NSM	*Number of Static Methods*	Número total de métodos estáticos.
TLOC	*Total Lines of Code*	Número total de líneas de código. No incluye comentarios ni líneas en blanco.
CA	*Afferent Coupling*	Número de clases fuera del paquete que dependen de clases dentro del paquete.
RMD	*Normalized Distance*	$\|RMA + RMI - 1\|$, este número debe ser pequeño. Cercano a 0 indica un buen diseño de paquetes.
NOC	*Number of Classes*	Número total de clases.
SIX	*Specialization Index*	Indice de especialización :$NORM * DIT / NOM$
RMI	*Instability*	$CE/(CA + CE)$

Métricas de código fuente		
Tipo	**Descripción (EN)**	**Descripción (ES)**
NOF	*Number of Atributtes*	Número de atributos definidos en un ámbito.
NOP	*Number of Packages*	Número total de paquetes.
MLOC	*Method Lines of Code*	Número de líneas de código por método. No incluye comentarios ni espacios en blanco.
WMC	*Weighted methods per Class*	Suma de las complejidades ciclomáticas de McCabe de todos los métodos de una clase.
NORM	*Number of Overriden Methods*	Número total de métodos sobreescritos. No incluye los de la clase "Object".
NSF	*Number of Static Attributes*	Número total de atributos estáticos.
NBD	*Nested Block Depth*	Número de bloques anidados.
NOM	*Number of Methods*	Número de métodos definidos en un ámbito.
LCOM	*Lack of Cohesion of Methods*	Cohesión de una clase. Si $m(A)$ es el número de métodos que acceden a un atributo A, se calcula la media de $m(A)$ para todos los atributos, se resta el número de métodos m y se divide el resultado entre $(1 - m)$. Un valor bajo indica clase con alta cohesión (preferible) y alto, baja cohesión y se debería dividir en n subclases. (*Henderson-Sellers*)
VG	*McCabe Cyclomatic Complexity*	Complejidad ciclomática.
PAR	*Number of Parameters*	Número de parámetros.
RMA	*Abstractness*	Número de clases e interfaces abstractas en un paquete divido entre el total de tipos en el paquete.
NOI	*Number of Interfaces*	Número de interfaces.
CE	*Efferent Coupling*	Número de clases dentro del paquete que dependen de clases fuera del paquete.
NSC	*Number of Children*	Número de hijos de un tipo o clase.
DIT	*Depth of Inheritance Tree*	Distancia de una clase en la jerarquía de herencia.

Tabla 6.2: Métricas del código fuente

Valores Métricas de código fuente						
ID	*Total*	*Media*	*Desv. est.*	*Max*	*Rango*	*Ámbito*
NSM	103	0,656	1,967	16	⊙	Tipo.
TLOC	8090	⊙	⊙	⊙	⊙	⊙
CA	⊙	6,545	15,455	86	⊙	Paquete.
RMD	⊙	0,331	0,33	1	⊙	Paquete.
NOC	157	4,758	4,164	23	⊙	Paquete.
SIX	⊙	0,074	0,329	3	⊙	Tipo.
RMI	⊙	0,648	0,351	1	⊙	Paquete.
NOF	158	1,006	1,684	9	⊙	Tipo.
NOP	33	⊙	⊙	⊙	⊙	⊙
MLOC	4235	5,639	7,025	47	⊙	Método
WMC	1167	7,433	9,055	52	⊙	Tipo.
NORM	39	0,248	0,711	3	⊙	Tipo.
NSF	91	0,58	1,506	15	⊙	Tipo.
NBD	⊙	1,304	0,723	7	⊙	Método.
NOM	648	4,127	4,403	22	⊙	Tipo.
LCOM	⊙	0,129	0,26	1	⊙	Tipo.
VG	⊙	1,554	1,736	25	⊙	Método autogenerado por Eclipse.
PAR	⊙	0,606	0,756	4	5	Método.
RMA	⊙	0,082	0,176	0,714	⊙	Paquete.
NOI	14	0,424	1,016	5	⊙	Paquete.
CE	⊙	3,576	2,535	11	⊙	Paquete.
NSC	20	0,127	1,144	14	⊙	Media y máximo por tipo.
DIT	⊙	1,242	0,569	4	⊙	Tipo.

Tabla 6.3: Valores de Métricas del código fuente de `moldeas-api`

6.5.5 Aportaciones Finales sobre Métricas de Código Fuente

Todas las medidas obtenidas, ver Tabla 6.3, están dentro de los rangos fijados por la herramienta, aquellos valores que no son aplicables se señalan con el símbolo ⊙. Sobre los valores obtenidos hay que destacar los siguientes puntos:

- La complejidad ciclomática de McCabe es siempre aceptable, obteniendo una media de 1,554 por debajo de 10, y con un máximo de 25 en un método generado por Eclipse (`equals`).
- En LCOM o medida de cohesión de una clase se obtiene una media 0,129. En esta medida el valor adecuado es cercano a 0, por lo que se puede concluir que las clases tienen una excelente valoración en cohesión. El valor máximo es de 0,26 que resulta de nuevo un excelente valor.
- Las medidas de acoplamiento CA y CE de nuevo proporcionan buenos valores para todas las clases.

Con estas medidas, se ha creado un producto cuyo código fuente posee las características deseadas de alta cohesión y bajo acoplamiento. Por otra parte, esta situación no es extraordinaria ya

que el uso de patrones de diseño y de buenas prácticas de programación, como el binomio de pruebas unitarias y *refactoring*, aseguran calidad en el código fuente implementado, no obstante, toda aplicación y por extensión su código fuente son siempre candidatos a ser mejorados.

6.6 Utilizando el Sistema MOLDEAS

6.6.1 Acciones desde el Interfaz Gráfico

Siguiendo las directrices y conceptos de diseño de la interacción que se han repasado y señalado como importantes en las anteriores secciones, cabe ahora destacar las principales acciones a realizar en el demostrador público del sistema MOLDEAS.

1. Pantalla inicial del sistema MOLDEAS, ver Figura 6.12 en el cual el usuario puede seleccionar y completar las variables de información de su perfil de búsqueda para la recuperación de información. La metáfora utilizada se basa en una simulación del tradicional "carrito de la compra" en la que el usuario puede seleccionar distintos códigos CPV, NUTS, establecer los rangos para algunas variables, etc. Se ha intentado simplificar al máximo este interfaz para evitar que una simple búsqueda se convierta en una tarea tediosa completando un amplio formulario.

2. Una vez que el usuario ha seleccionado su perfil de búsqueda se ejecuta el proceso de recuperación de información, presentando los resultados en forma tabular y mediante el uso de Exhibit, ver Figura 6.13. El usuario puede modificar su perfil de búsqueda y las consultas se ejecutarán automáticamente una vez se haya detectado algún cambio. De la misma forma, el usuario puede filtrar los resultados, ordenarlos por relevancia, por sector, etc., e incluso visualizar la región en la que se han publicado los anuncios de licitación. La idea subyacente a esta interacción reside en que el usuario disponga del completo control de la presentación de los resultados obtenidos.

3. En los resultados de búsqueda todas las URIs presentadas son referenciables siguiendo las buenas prácticas de *Linked Data* que han guiado la promoción de los datos a esta iniciativa. Es por ello, que los datos propios del *dataset* RDF se pueden consultar y navegar a través de sus relaciones mediante el uso de un *Linked Data Frontend*, en este caso Pubby, ver Figura 6.14.

4. Finalmente y con el objetivo de ejemplificar el uso de SPARQL y las posibilidades de consultas al `endpoint` se han creado una serie de consultas representativas para que los usuarios más técnicos tengan la posibilidad de configurar y crear sus propias consultas, ejecutándolas y obteniendo los resultados directamente. Para suministrar esta funcionalidad se ha utilizado la herramienta SNORQL, ver Figura 6.15.

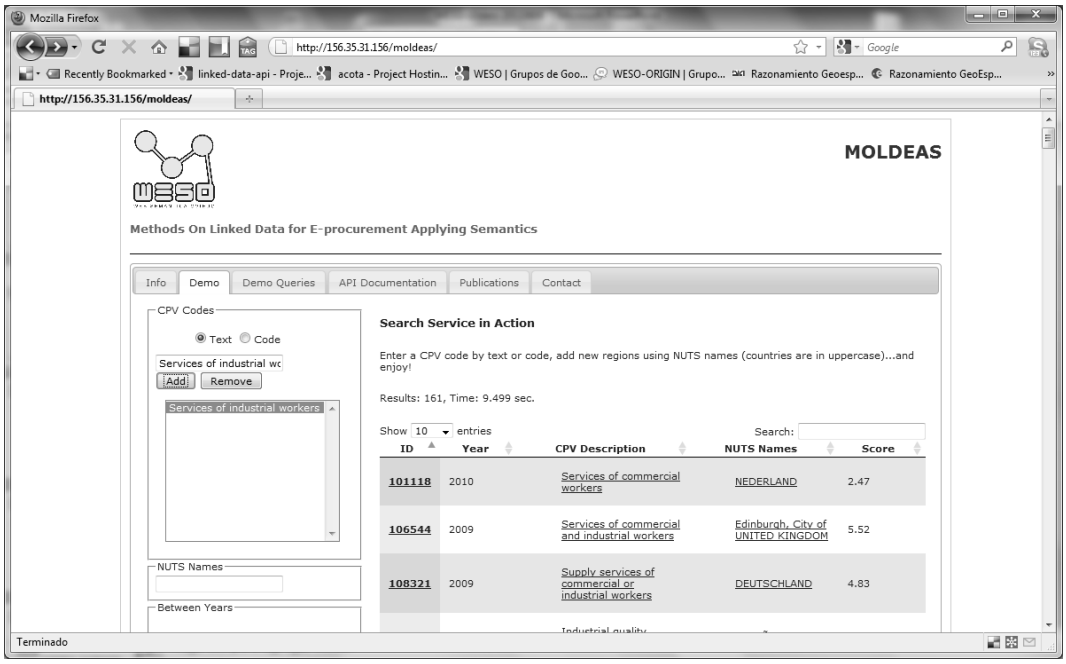


Figura 6.12: Ejemplo de pantalla inicial en `moldeas-web`.

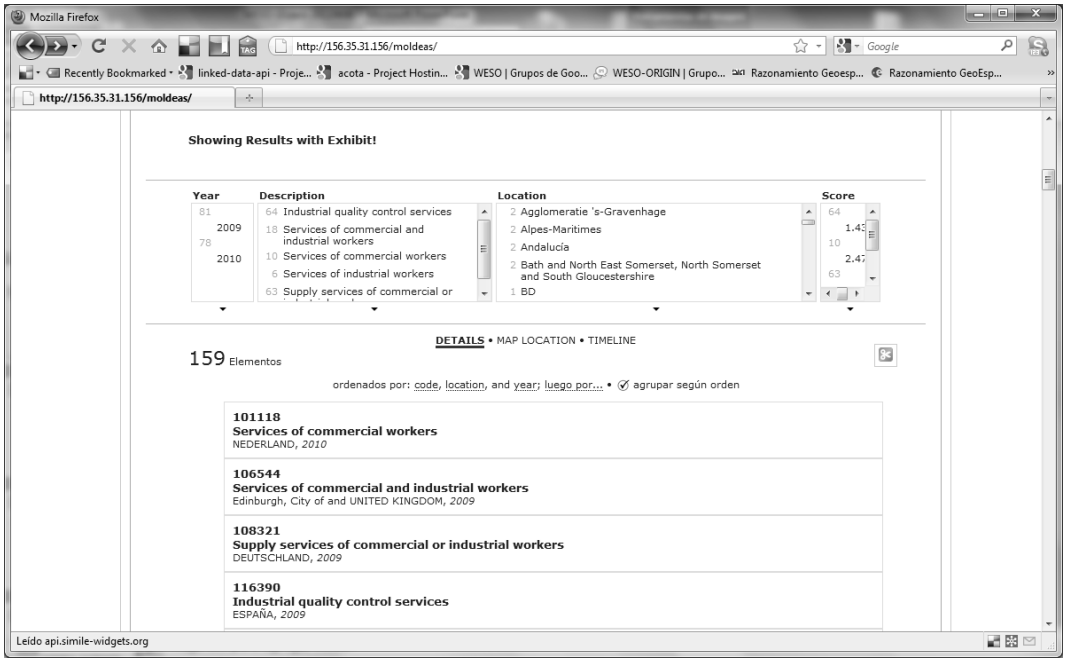


Figura 6.13: Ejemplo de pantalla de resultados en `moldeas-web`.

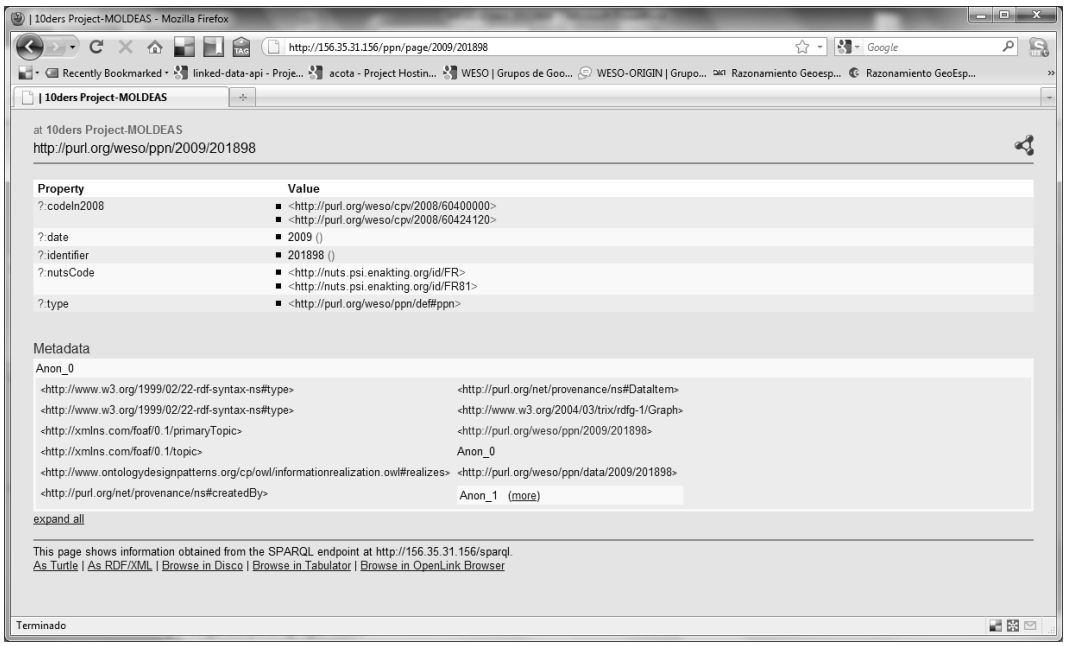


Figura 6.14: Acceso a los datos enlazados mediante Pubby.

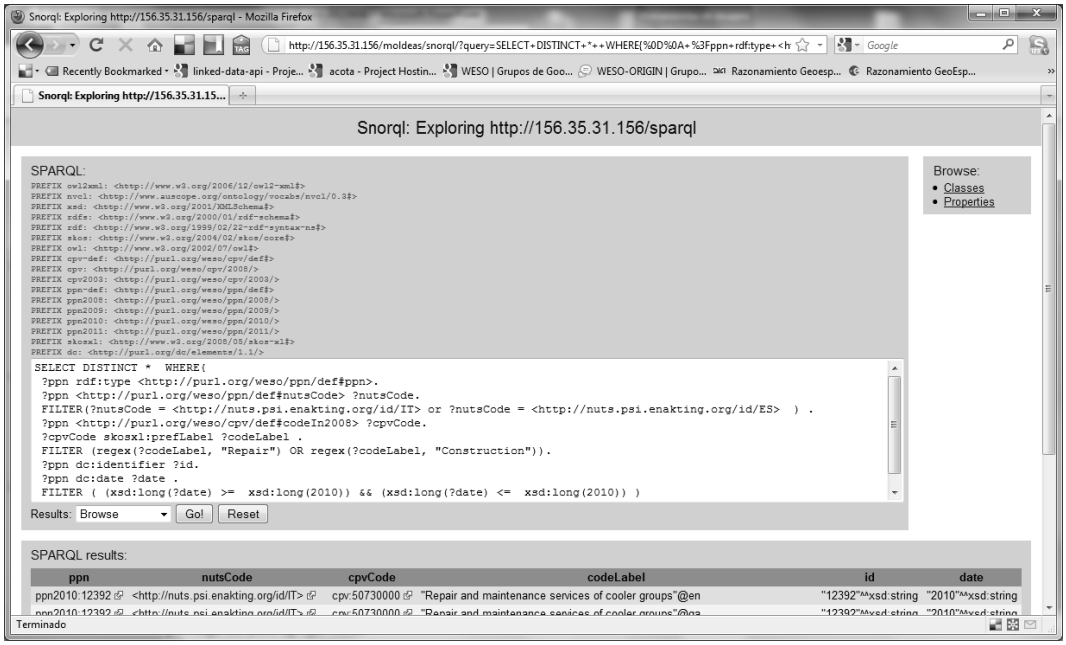


Figura 6.15: Consulta a los datos enlazados mediante SNORQL.

Capítulo 7

Experimentación y Validación

> Los errores causados por los datos inadecuados son mucho menores que los que se deben a la total ausencia de datos.
>
> *Citas Célebres*
> CHARLES BABBAGE

A lo largo de los capítulos anteriores se han repasado los conceptos fundamentales que atañen al proceso de contratación pública electrónica, la casuística, beneficios y métodos disponibles para la aplicación de los principios de *Open Data*, *Linked Data* y Web Semántica a distintos dominios de conocimiento. Por otra parte, se ha realizado una definición de un ciclo de vida de datos enlazados describiendo procesos, métodos y tareas cuya aplicación conlleva una promoción sostenible de datos a las iniciativas previamente citadas, como se ha señalado en el Capítulo 5. Finalmente, con el objetivo de suministrar soporte a este ciclo de vida se han implementado una serie de componentes desde un punto de vista de la ingeniería. La siguiente etapa, por tanto, corresponde a la validación del trabajo realizado mediante experimentos en los cuales se pueda cuantificar el enfoque realizado. En este sentido, la experimentación posible es muy variada pero atendiendo a la hipótesis de partida y a los componentes implementados es posible identificar los siguientes experimentos:

1. Experimento para la validación de los datos transformados desde un punto de vista cuantitativo y cualitativo, comparando las versiones actuales de los datos disponibles con la versión generada tras la aplicación de los métodos semánticos a las entidades implicados en los anuncios de licitación de un proceso de contratación pública.

2. Experimento para la verificación de la construcción de un sistema de recuperación de información utilizando una fuente de datos basada en *Linked Data* y tecnologías semánticas.

3. Experimento para la evaluación del rendimiento del sistema construido desde un punto de vista cuantitativo.

El diseño de estos experimentos permite por una parte, la verificación de la hipótesis objeto de estudio en este trabajo y, por otra parte, la validación de la construcción de sistemas basados en *Linked Data* y tecnologías semánticas.

7.0.2 Consideraciones Generales

A continuación es preciso señalar que todos los experimentos llevados a cabo se realizan bajo la ejecución de un plan detallado, las etapas a seguir serán las siguientes:

1. Definición de los objetivos del experimento.
2. Selección de una regla de asignación de las unidades experimentales a las condiciones de estudio.
   - Cualitativos: tipo de entorno hardware y software, etc. En este caso, se pueden fijar las condiciones *hardware* y *software*:
     - Servidor. El software consta de un repositorio RDF Virtuoso de OpenLink (version 06.01.3127) desplegado sobre una máquina virtual de Virtual Box (version 4.0.6) sobre un sistema operativo Linux 2.6.35-22-server #33-Ubuntu 2 SMP x86_64 GNU/Linux Ubuntu 10.10 con 2GB de RAM y 30GB HardDisk, la máquina física utilizada como anfitrión es un PC DELL con la misma configuración software que la máquina virtual.
     - Cliente. En este caso el cliente es un Notebook DELL XPS M1330 con un sistema operativo Linux 2.6.32-24-generic #43-Ubuntu SMP x86_64 GNU/Linux con 4GB de RAM, las distintas réplicas son ejecutadas a través de la herramienta Java, Junit.
   - Cuantitativos: tamaño de la muestra, de la memoria y número de posibilidades de expresar una consulta.
3. Especificación de las medidas de trabajo en cuanto a la respuesta.
4. Ejecución de un experimento piloto.
5. Especificación de un modelo.
6. Esquematización de los pasos a seguir.
7. Determinación del tamaño muestral.
8. Revisión de las decisiones anteriores.

7.1 Experimento sobre la aplicación de *Linked Data* a las Licitaciones Públicas

7.1.1 Diseño del experimento sobre la aplicación de *Linked Data* a las Licitaciones Públicas

Con el objeto de la validación de la hipótesis de partida surge la necesidad de comprobar si tanto el ciclo de vida utilizado para la promoción de los datos concernientes a los anuncios de licitación y sus derivados como las clasificaciones de productos y las organizaciones ha sido convenientemente ejecutado. Las ventajas de uso de *Linked Data*, *Open Data* y la aplicación de tecnologías semánticas ha sido convenientemente revisada en la Sección 3, por lo que aplicando los principios promulgados bajo esta iniciativa y realizando la correspondiente evaluación de los mismos con resultados positivos se estima, evidentemente, que las ventajas adicionales se obtienen de forma automática. Por ejemplo, en el caso de suministrar los datos bajo los protocolos y estándares actuales de Internet se asegura la interoperabilidad y facilidad de acceso, con lo que utilizando HTTP URIs este principio y sus correspondientes ventajas quedarían perfectamente demostradas.

Por otra parte, cabe mencionar que la evaluación de los datos ha sido abordada bajo distintos paradigmas, como la calidad. En este contexto existen muchos trabajos que tratan de definir métricas sobre cómo determinar si unos datos son lo suficientemente "buenos", pero tanto en las tradicionales bases de datos como en el emergente paradigma de datos enlazados, las herramientas, vocabularios y especificaciones para establecer la calidad de los datos se centran en evaluar el cumplimiento de ciertas características definidas por los autores pero que en ningún caso son de obligado cumplimiento, tratándose en cierto modo de buenas prácticas. En este sentido, el enfoque seguido en este experimento se basa en la recopilación de buenas prácticas, patrones de diseño, condiciones de los conjuntos de datos para pertenecer a una determinada iniciativa, experiencia adquirida, etc., en las cuales se fijan los criterios que posteriormente pueden ser expresadas en los vocabularios y herramientas disponibles para obtener los valores que estas tecnologías suministran. En este sentido, se dispone de herramientas como WIQA [36, 38] o LINK-QA [280] que ofrecen de un lenguaje de patrones para valorar la pertenencia de una tripleta a un determinado grafo, por ejemplo evaluando si una propiedad tiene un determinado valor. Por otro lado, se ha propuesto el vocabulario LODQ [240] para la creación de métricas enclavadas en una ordenación mayor pero cuya implantación se encuentra en una etapa muy temprana. Es necesario destacar que en su mayor parte los criterios de evaluación requieren una validación semi-manual por lo que el empleo de herramientas, aunque es conveniente debe ser valorado para no incurrir en un exceso de esfuerzo mayor que la propia validación. En el caso objeto de estudio, el enfoque seguido se centra en el cumplimiento de criterios para pertenecer a las iniciativas *Linked Data*, *Open Data* y a la suma de ambas. En cuanto a la corrección en los valores de los datos se supone, que todos los datos proceden de fuentes oficiales y en el caso concreto de los anuncios de licitación, han sido previamente validados por la empresa Gateway S.C.S. No obstante, hay que destacar que la calidad de datos en *Linked Data* es una línea de investigación interesante y emergente en la cual se está empezando a avanzar pero que no se ha declarado estratégica en estos primeros años, considerando el número de trabajos existentes hasta la fecha.

En el caso objeto de estudio, el punto de partida son conjuntos de datos generados por las Administraciones Públicas y ya procesados por una empresa, por lo que parcialmente la fuente de los mismos se considera fiable en cuanto a requerimientos de calidad, adicionalmente y debido a la aplicación de un proceso de transformación de datos utilizando herramientas apropiadas para ello y mediante la consecución de tareas determinadas permite asegurar, en un alto grado, el mantenimiento de la calidad, minimizando los posibles errores de transformación. Finalmente la evaluación a través de las tablas de validación permite asegurar que los datos generados cumplen criterios de accesibilidad, calidad, etc., igualmente éstos se pueden expresar en las herramientas y vocabularios citados, en su gran mayoría, por lo que en último término la aportación de los mismos es simplemente la validación automática y la extracción de métricas o estadísticas de uso basadas en criterios de clusterización, etc., fijados por las propias herramientas.

El diseño de este experimento, por tanto, se planifica desde dos puntos de vista:

Cuantitativo. La pregunta principal planteada en este caso se refiere a la posibilidad del uso de datos enlazados para facilitar el acceso a un mayor número de recursos relacionados con los anuncios de licitación y, en consecuencia, a la información y datos disponibles en los mismos.

Cualitativo. La cuestión planteada en este caso se centra en valorar si el proceso de generación y transformación de datos a la iniciativa de datos enlazados se puede valorar respecto a una serie de criterios y buenas prácticas.

Adicionalmente la evaluación debe asegurar que se cumplen los principios básicos de *Linked Data* y *Open Data* con lo que se aseguran los beneficios y ventajas promulgados por estas iniciativas.

7.1.1.1 Punto de vista Cuantitativo

Una vez presentadas las directrices básicas del experimento es necesario realizar la exposición del plan detallado y diseño del mismo. Para el primer caso, la evaluación cuantitativa debe suministrar la prueba de que el uso de datos enlazados permite facilitar el acceso a un mayor número de anuncios de licitación en general, y a una mayor cantidad de datos e información relativa al proceso de licitación pública electrónica en particular. Con ello, se demuestra que el uso de este enfoque favorece una mayor expresividad en la consulta de información y los datos. Las etapas a seguir serán las siguientes:

1. Definición de los objetivos del experimento. Las preguntas a responder por el experimento serán:
   - ¿Cuál es la expresividad actual, en términos de número de conceptos para realizar consultas, para el acceso a la información de anuncios de licitación?
   - ¿Cuál es la ventaja de uso de un modelo RDF para la expresión y recuperación de la información de los anuncios de licitación?
   - ¿Cómo los datos enlazados favorecen el aumento de expresividad en la ejecución de consultas y por tanto facilitan la recuperación de los anuncios de licitación?
   - ¿Cuál es el beneficio real del uso de datos enlazados para representar la información?
   - ¿Se incurre en algún error al aumentar la expresividad?
2. Selección de una regla de asignación de las unidades experimentales a las condiciones de estudio. En este caso, la unidad experimental de este estudio será un repositorio RDF en el cual se encuentra albergada la información sobre los anuncios de licitación, las clasificaciones de productos y las organizaciones.
   - Cualitativos: tipo de entorno *hardware* y *software*, expresividad de las consultas.
   - Cuantitativos: tamaño de la muestra, de la memoria y número de posibilidades de expresar una consulta.

   En cuanto a la caracterización de los factores cualitativos, respecto al entorno hardware y software se utiliza el definido en la Sección 7.0.2.

   Sobre la expresividad de las consultas cabe mencionar que no es admisible el establecimiento de una comparación entre un sistema de búsqueda sintáctico tradicional basado en texto libre con uno basado en un vocabulario controlado, ya sea mediante la concatenación de filtros u otro tipo de acotación de la búsqueda. Es por ello que como referente para realizar la comparación se toma un sistema de búsqueda que permite seleccionar los códigos CPV 2008 sobre los cuales realizar búsquedas de anuncios de licitación, la justificación de esta decisión reside en que la mayoría de los actuales sistemas de búsqueda huyen del texto libre (obviamente los resultados son menos ajustados) y sólo permiten la introducción de varios códigos CPV para la recuperación de documentos. Este es el caso de la herramienta Euroalert.net de Gateway S.C.S. o de otros sistemas como Licitaciones.es, etc., igualmente en el ámbito de los datos enlazados, la propuesta realizada en el proyecto LOTED se basa en la introducción de códigos para generar consultas en SPARQL. Dada esta situación, la comparación se realizará entre una versión básica de búsqueda, que acepta el CPV 2008 como vocabulario controlado, 10357 términos, sobre los datos que se han transformado en el Capítulo 5, $1M$ de anuncios de licitación y 50000 organizaciones.
3. Especificación de las medidas de trabajo en cuanto a la respuesta. En este caso la medida utilizada para valorar los resultados será en magnitud del número de nuevos términos/conceptos que se pueden utilizar al añadir un nuevo vocabulario controlado en el campo de las clasificaciones de productos y enlazarlo con el CPV 2008.

4. Ejecución de un experimento piloto. Con el objetivo de ejemplificar el modelo y diseño del experimento se realiza un ejemplo a pequeña escala para facilitar la compresión del mismo y realizar una validación previa del enfoque seguido.

5. Especificación de un modelo. En este sentido, la consecución del experimento permite establecer una fórmula sencilla para la predicción de la ganancia expresiva al añadir un nuevo vocabulario de negocio y enlazarlo con el CPV 2008 y *ProductOntology*. De esta forma, se suministra un valor a la ganancia expresiva del uso de datos enlazados en este dominio.

6. Esquematización de los pasos a seguir. Las etapas a cumplir por este experimento, una vez que las clasificaciones de productos se encuentran transformadas, residen en la extracción de consultas en SPARQL para establecer el número de enlaces entre las mismas y la realización de un `script` para la generación de datos finales. Específicamente se realizan los siguientes pasos:

   a) Estado inicial: clasificaciones de productos ya transformadas y disponibles a través de un *endpoint* de SPARQL.

   b) Extracción de los términos de una PSC relacionados con el CPV 2008. Para ello se pueden realizar las siguientes consultas en SPARQL, ver Figura 7.2, generando dos ficheros: `psc-1` y `psc-2` con los resultados de la consultas respectivamente.

```
SELECT DISTINCT ?psck WHERE {
  ?psck rdf:type pscs:PSCConcept.
  ?psck skos:inScheme <http://purl.org/weso/pscs/sitc/v4/resource/ds> .
  ?psck skos:closeMatch ?cpvk.
}

SELECT DISTINCT ?psck WHERE {
  ...
  ?psck skos:exactMatch ?cpvk.
}
```

Figura 7.1: Consultas en SPARQL de extracción de términos de una PSC relacionados el CPV 2008.

   c) Extracción de los términos una PSC, incluida el CPV 2008, relacionados con *ProductOntology* y de los términos en *ProductOntology* respecto a una PSC, generando dos ficheros: `psc-5` y `cpv-2008-6`.

```
SELECT DISTINCT ?psck WHERE {
  ...
  ?psck pscs:relatedMatch ?pok.
}

SELECT DISTINCT ?pok WHERE {
  ...
  ?psck pscs:relatedMatch ?pok.
}
```

Figura 7.2: Consultas en SPARQL de extracción de términos de una PSC con *ProductOntology*.

   d) Generación del fichero obtenido de las consultas realizadas en el Paso 2 con los enlaces reales de una PSC al CPV 2008. La extracción del número de enlaces se realiza mediante el comando:

```
cat psc-1 psc-2 | sort | uniq | wc -l
```

e) Generación del fichero intersección con los elementos de *ProductOntology* provenientes de una PSC y el CPV 2008. La intersección se realiza mediante el comando:

```
awk 'NR==FNRa[$0];next $0 in a' psc-5 cpv-2008-6 >psc-5-6
```

f) Generación del listado final de términos únicos de una PSC que tienen enlace directo con el CPV 2008 o bien a través de *ProductOntology*. Para la realización de esta operación se ha implementado un sencillo programa en Java que toma el fichero anterior, `psc-5-6`, el fichero RDF con los datos de la PSC y genera un listado de conceptos pertenecientes a la PSC, fichero `psc-po-cpv`, que tienen relación con un elemento de *ProductOntology* presente en el fichero de entrada.

g) Extracción del número final de elementos que realmente se encuentran enlazados entre la PSC y el CPV 2008, directamente o bien a través de *ProductOntology*, esta operación se realiza mediante el comando:

```
cat psc-1 psc-2 psc-po-cpv | sort | uniq | wc -l
```

h) Adición de todos los valores recogidos para la creación de porcentajes mediante un hoja de cálculo.

7. Determinación del tamaño muestral. Ya se han señalado en los aspectos cuantitativos del experimento

8. Revisión de las decisiones anteriores.

7.1.1.2 Punto de vista Cualitativo

De acuerdo a la introducción realizada, el objetivo de este experimento se centra en la evaluación y grado de cumplimiento de buenas prácticas sobre el uso de datos enlazados, principios de *Linked Data* y *Open Data*, patrones de diseño y características para la adición de *datasets* RDF a la nube de datos enlazados abiertos y a registros tipo CKAN. Esta evaluación permite discernir qué enfoques cumplen estos requisitos y principios para así determinar en qué grado de cumplimiento se encuentran unos datos respecto a las iniciativas ya mencionadas. Este enfoque permite establecer si el ciclo de vida de generación de datos ha asegurado la calidad de los mismos en cuanto a grado de cumplimiento de criterios. A continuación se detalla el plan de experimentación en los siguientes pasos:

1. Definición de los objetivos del experimento. Las preguntas a responder por el experimento serán:

   - ¿El ciclo de vida seguido y los datos generados certifican la aplicación de buenas prácticas de *Linked Data*?
   - ¿El ciclo de vida seguido y los datos generados certifican el cumplimiento de los principios de *Linked Data*?
   - ¿Qué nivel del modelo de 5 $\star$ se puede establecer?
   - ¿Qué porcentaje de patrones de diseño se han aplicado en los datos generados?
   - ¿Los datos generados pueden pertenecer a la nube de datos enlazados abiertos?
   - ¿Los datos generados pueden pertenecer a un registro CKAN?
   - ¿El ciclo de vida seguido y los datos generados certifican el cumplimiento de los principios de *Open Data*?
   - ¿Se puede asegurar que los datos son enlazados y abiertos?
   - ¿Qué beneficios se obtienen del cumplimiento de estos objetivos?

2. Selección de una regla de asignación de las unidades experimentales a las condiciones de estudio. En este caso, la unidad experimental de este estudio será cada uno de los *dataset* RDF generados para los anuncios de licitación públicos, el catálogo de clasificaciones de productos y las organizaciones.

   - Cualitativos: tipo de entorno hardware y software, grado de cumplimiento de los criterios.
   - Cuantitativos: tamaño de la muestra, número de criterios en cada una de las tablas de validación, selección de enfoques a comparar.

   En cuanto a la caracterización de los factores cualitativos, respecto al entorno hardware y software se utiliza el definido en la Sección 7.0.2.

   Sobre el grado de cumplimiento de criterios se establece un valor numérico y un porcentaje sobre el mismo para cada una de las tablas de criterios creadas y para cada uno de los enfoques evaluados. De esta forma, se da respuesta a las preguntas planteadas en el punto anterior, especificando en qué nivel de cumplimiento se encuentra cada uno de los enfoques.

   Respecto a los aspectos cuantitativos, el desarrollo del experimento consiste en la evaluación de 196 criterios, como una encuesta sobre los enfoques seleccionados, distribuidos en distintas tablas que permitan establecer correctamente los aspectos cualitativos. El diseño de estas tablas se presenta en el Apéndice C con las preguntas y la respuesta ideal de cada una de ellas. Con este enfoque se permite realizar un repaso pormenorizado de los criterios y realizar así una evaluación intensiva. Aunque algunos de los enfoques pertenecen a distintas disciplinas y han sido realizados con distintos objetivos a los presentes en la iniciativa de datos enlazados, es posible establecer ciertas características en común, por lo que la comparación es posible si se remite al acceso a la información y a los datos. Por tanto, en cuanto a los enfoques seleccionados, se comparan los más representativos para cada uno de los *datasets* RDF, de esta forma se han seleccionado los siguientes:

   a) Anuncios de licitación públicos, teniendo en cuenta que la publicación de estos anuncios es transversal desde la Unión Europea hasta organismos de carácter local, publicándose en los correspondientes boletines oficiales se han agrupado y escogido los siguientes:
      - Boletines y Publicaciones oficiales: TED y BOE. El objetivo es aunar la forma en la que los boletines oficiales publican la información y los datos de licitación. Se han seleccionado estas dos fuentes porque son oficiales y representan gran parte de la casuística: 1) información restringida con un modelo formal (XML Schema) pero con necesidad de suscripción (TED) y 2) información total pero en peores condiciones de publicación (BOE). Teniendo en cuenta que el comportamiento de los boletines oficiales es similar, la selección de uno de ellos se considera representativa, además, en los últimos años el BOE ha avanzado enormemente en la provisión de servicios de búsqueda, uso de capacidades avanzadas de PDF y mejora de la generación de HTML, por lo que dispone de un conjunto de características verdaderamente interesantes en términos de evaluación.
      - Plataformas de contratación: Plataforma de Contratación del Estado. Este enfoque supone un avance a la simple publicación en boletines oficiales, ya que los datos y la información están disponibles bajo un modelo formal (XML Schema), pero ocurre en cierta manera lo mismo que en TED, es necesaria una suscripción para el acceso completo. No obstante, teniendo en cuenta la proliferación de este tipo de plataformas, tanto públicas como privadas (por ejemplo opXML del proyecto "10ders Information Services") es necesario y conveniente incluirlas dentro del proceso de comparación.
      - Servicios de terceros que reutilizan la información oficial, se han agrupado los suministrados por Euroalert.net y Licitaciones.es. La característica de estos servicios es similar, permiten la búsqueda y acceso a las licitaciones de forma restringida, a partir de este punto, para realizar la consulta de datos e información de forma completa

se debe disponer de una suscripción. En general, no es posible conocer el modelo de datos utilizado ni la tecnología subyacente, pero desde el punto de vista del acceso a la información son muy relevantes ya que suponen servicios de negocio.

- LOTED. Se trata del primer gran esfuerzo por aunar tecnología semántica y datos enlazados reutilizando la información abierta publicada en TED a través de los RSS. En cuanto a comparación, sin duda este enfoque es un buen reflejo para el realizado con MOLDEAS.

b) Clasificaciones de Productos, en este caso las posibilidades están restringidas a las publicaciones oficiales en formato CSV o MSExcel por lo que será el único caso en común para todas las clasificaciones. No obstante, en cuanto al CPV y otras clasificaciones tipo NAICS, existen servicios de consulta *on-line* en forma de sistemas de búsqueda o directorios para navegar por la jerarquía.

c) Organizaciones, como subconjunto de la información proveniente de los anuncios de licitación públicos se utilizan fuentes similares de información, entre las cuales se encuentran:

- Boletines y Publicaciones oficiales: TED y BORME.
- Plataforma de Contratación del Estado.
- Servicios de terceros.
- Bases de datos administradas por terceros, en este caso, existen algunas empresas que comercializan una base de datos con información corporativa pero sin publicar la información y los datos vía web, tal es el caso de la Cámara de Comercio.
- *OpenCorporates*, por analogía con LOTED, se trata del primer gran esfuerzo por promocionar la información corporativa mediante la aplicación de los principios de *Open Data* y *Linked Data*, convirtiéndose en un enfoque pionero y de referencia a nivel mundial.

3. Especificación de las medidas de trabajo en cuanto a la respuesta. En este experimento específico los valores para cada uno de los criterios evaluados podrán ser de los siguientes tipos:

   a) Valor positivo, $\oplus$, si es un criterio que debe tener y se cumple.

   b) Valor negativo, $\ominus$, si es un criterio que debe tener y no se cumple.

   c) Valor no aplicable, $\odot$, si es un criterio que se desconoce, que se solapa con otro (por ejemplo en la tabla de cumplimiento del registro CKAN o bien los datos son *published-by-producer* o por *published-by-third-party*) o no está asociado a ese enfoque.

   La suma de todas las valoraciones permite establecer un grado de cumplimiento. Adicionalmente, las tablas de validación se han diseñado para que aporten la mayor información posible agrupando los criterios similares. Es conveniente destacar que el número de valores no aplicables nunca debe ser menor que el propuesto en el modelo de referencia.

4. Ejecución de un experimento piloto. Con el objetivo de probar el diseño de las tablas de validación se ha realizado un experimento inicial en el cual se ha detectado: 1) que algunos criterios no estaban correctamente divididos y no permitían discernir ciertos factores, por ejemplo inicialmente los principios de *Linked Data* y los criterios del modelo 5 $\star$ estaban juntos y se ha decidido separarlos para facilitar la evaluación posterior. Igualmente, los principios de *Open Data* estaban unidos a los beneficios y ventajas de este enfoque, por lo que también se han separado en dos tablas. En resumen, de la versión inicial se han generado dos tablas nuevas. 2) Se han incluido nuevos criterios respecto a la versión inicial para incluir características de metadatos en la evaluación como el sellado en el tiempo de los datos y la información. En este caso, el experimento piloto ha servido para detectar un diseño incorrecto de la encuesta e incluir nuevos criterios.

5. Especificación de un modelo. En este caso se establece un modelo de respuestas ideal como referencia, a partir del cual se puede evaluar el grado de cumplimiento. No obstante, la combinación de las distintas valoraciones es conveniente realizarla por separado y no mezclar las distintas encuestas con el objetivo de no introducir ruido y falsos positivos en los resultados.

6. Esquematización de los pasos a seguir. Las etapas a cumplir para la realización de los cuestionarios por cada uno de los enfoques seleccionados es el siguiente:

   a) Establecimiento del modelo de referencia, con los valores admitidos.

   b) Revisión uno a uno de los criterios, consultando cada una de las fuentes de información y datos para cada enfoque.

   c) Agregación de los resultados y valoraciones.

   d) Extracción de estadísticas, contraste de hipótesis, validación y evaluación.

7.1.2 Ejecución del experimento sobre la aplicación de *Linked Data* a las Licitaciones Públicas

7.1.2.1 Punto de vista Cuantitativo

La ejecución de este experimento requiere la expresión formal de los datos para establecer desde un punto de vista cuantitativo el porcentaje de ganancia en expresividad pasando de utilizar un vocabulario controlado para la realización de consultas, a una serie de los mismos mediante datos enlazados. En este caso, la recuperación de información entre la búsqueda habitual (texto libre) y un vocabulario controlado difiere en que el universo del discurso es en el primer caso infinito y en el segundo finito, determinado por el número de términos del vocabulario. En los anuncios de licitación se parte de las siguientes premisas:

- Todo anuncio de licitación está etiquetado al menos con un código/término perteneciente al CPV 2008.
- El uso de los códigos CPV es obligatorio en todos los anuncios de licitación a nivel europeo, como se ha señalado en la Sección 2.10.
- La única información obligatoria, además del emisor del anuncio de licitación, son los códigos CPV 2008.

Bajo estas circunstancias la muestra sobre la que se realiza el experimento cuenta con las siguientes características:

- Se cuenta con una base documental $\mathcal{D}$ constituida por 1 millón de anuncios de licitación.
- La clasificación de productos (vocabulario controlado $\mathcal{V}$ para contrato públicos) del CPV 2008 está formado $\#\mathcal{V} = 10357$ códigos/términos distintos.
- Cada uno de los documentos $d \in \mathcal{D}$ está etiquetado con al menos un código $v \in \mathcal{V}$.

Una vez caracterizado el entorno de pruebas, cabe formular cómo se calcula la ganancia de añadir un nuevo vocabulario mediante *Linked Data*.

- El nuevo vocabulario controlado $\mathcal{V}_{psc}$ dispone $\#\mathcal{V}_{psc}$ términos.

- El enlazado de datos que genera este vocabulario respecto a uno objetivo $\mathcal{V}$ presenta los siguientes casos:
  1. Enlace 1-1, es decir, existen términos $v^k_{psc} \in \mathcal{V}_{psc}$ que se enlazan con un sólo término $v^k \in \mathcal{V}$ generando K enlaces.
  2. Enlace 1-n, es decir, existen términos $v^k_{psc} \in \mathcal{V}_{psc}$ que se enlazan con varios términos $v^k \in \mathcal{V}$ generando K_n enlaces.
- El resultado de esta operación es un conjunto de pares de enlaces $\{(v^0_{psc}, v^0), (v^k_{psc}, v^k)..(v^n_{psc}, v^n)\}$.
- Dada esta situación el vocabulario inicial $\mathcal{V}$ ha sido incrementado en todos los elementos $v^k_{psc} \in \mathcal{V}_{psc}$ para los cuales existe un enlace con algún elemento $v^k \in \mathcal{V}$.
- El número de nuevos términos en los que el vocabulario $\mathcal{V}$ es incrementado es igual al conjunto de elementos v^k_{psc}, denominado $\mathcal{V}'_{psc}$, con algún enlace a $\mathcal{V}$ (estrictamente sin repeticiones por definición de conjunto).
- El porcentaje de ganancia de expresividad, en el sentido de cuántos nuevos términos permiten pasar de $\mathcal{V}_{psc}$ a $\mathcal{V}$, es el siguiente:

  Definición 7.1.1. *Porcentaje de Ganancia*

  $$\% = \langle (\#\mathcal{V}'_{psc} + \#\mathcal{V})/\#\mathcal{V} \rangle * 100 - 100 \tag{7.1}$$

- Finalmente, y en el caso concreto de las clasificaciones de productos/taxonomías, existe la posibilidad de que elementos $v^j_{psc} \in \mathcal{V}_{psc}$ no estén directamente relacionados con el conjunto $\mathcal{V}$ pero de acuerdo a la jerarquía de términos en $\mathcal{V}_{psc}$ se pueda pensar que si v^j_{psc} está relacionado con v^k_{psc} a través de una relación r_k, entonces $v^j_{psc} \in \mathcal{V}'_{psc}$. Este tipo de enfoque es crítico debido a que provoca una situación recursiva e infinita, uno de los problemas actuales de *Linked Data* en la ejecución de consultas, pero debido a que el vocabulario $\mathcal{V}_{psc}$ es finito y las relaciones entre sus términos es de jerarquía, se puede establecer un valor máximo para el porcentaje de ganancia que ocurrirá cuando $\mathcal{V}_{psc} \equiv \mathcal{V}'_{psc}$.

Como ejemplo para la validación de este enfoque se utilizarán los siguientes conjuntos:

- Sea $\mathcal{V} = \{1, 2, 3\}$ y $\mathcal{V}_{psc} = \{A, B, C, D, E\}$.
- El conjunto de pares enlaces es el siguiente: $\{(A, 1), (B, 2), (C, 1)(C, 2)\}$.
- Por tanto, el conjunto $\mathcal{V}'_{psc} = \{A, B, C\}$ y el porcentaje de ganancia en expresividad será el siguiente:

$$\% = \langle (3 + 3)/3 \rangle = 200 - 100 = 100 \tag{7.2}$$

- Lo que implica que el número de términos final es el doble que el inicial, aumentando así la expresividad en un 100 %.

Adicionalmente e introduciendo un nivel más de indirección mediante las PSCs con *ProductOntology* se da el siguiente caso:

- Existe un conjunto $\mathcal{V}''_{psc}$ que representa la unión de todos los $\mathcal{V}'_{psc}$ correspondientes a los enlaces entre una determinada PSC y *ProductOntology*, estrictamente conteniendo el elemento de *ProductOntology* de cada par.

- Existe un conjunto $\mathcal{V}'_{cpv}$ que representa el conjunto de pares de enlaces entre el CPV 2008 y *ProductOntology*, estrictamente conteniendo el elemento de *ProductOntology* de cada par.
- La intersección entre $\mathcal{V}''_{psc} \sqcap \mathcal{V}'_{cpv}$ representa el conjunto de todos enlaces mediante los cuales se puede realizar el salto entre una PSC al CPV 2008 a través de *ProductOntology*. Los elementos de cada PSC que permitan acceder al CPV 2008 de esta forma y no se encuentren en el conjunto $\mathcal{V}'_{psc}$ de cada PSC, señalado en la Tabla 7.1, deben ser también considerados como nuevos términos para la realización de consultas generando un nuevo conjunto $\mathcal{V}''_{psc}$.

Continuando con el ejemplo anterior esta situación se resolvería de la siguiente manera:

- Sea $\mathcal{V}'_{psc} = \{A, B, C\}$ el conjunto inicial de enlaces.
- Sea $\mathcal{V}''_{psc} = \{po_1, po_2\}$ ya que existe un conjunto de pares de enlaces: $\{(D, po_1), (C, po_2)\}$ entre la clasificación PSC y *ProductOntology*.
- Sea $\mathcal{V}'_{cpv} = \{po_1\}$ ya que existe un conjunto de pares de enlaces: $\{(cpv_1, po_1)\}$ entre el CPV y *ProductOntology*.
- El conjunto $\mathcal{V}''_{psc} \sqcap \mathcal{V}'_{cpv} = \{po_1\}$ lo que implica que el elemento D debe ser añadido al conjunto de elementos que tienen enlaces $\mathcal{V}'''_{psc} = \mathcal{V}'_{psc} \cup \{D\}$.

De acuerdo a estas definiciones, en la Tabla 7.1 se presentan los porcentajes reales y máximos obtenidos tras la realización del enlazado de las clasificaciones de productos con el CPV 2008, conjunto $\mathcal{V}$ con $\#\mathcal{V} = 10357$.

$\mathcal{V}_{psc}$	$\#\mathcal{V}_{psc}$	$\#\mathcal{V}'_{psc}$	$\#\mathcal{V}'''_{psc}$	% real	% real con PO	% máximo
CPV 2003	8323	462	8312	4,46	80,25	80,36
CN 2012	14552	2390	2390	23,07	23,07	140,50
CPC 2008	4408	4402	4403	42,50	42,51	42,56
CPA 2008	5429	5399	5410	52,12	52,23	52,41
ISIC v4	766	765	765	7,38	7,38	7,39
NAICS 2007	2328	2300	2300	22,20	22,20	22,47
NAICS 2012	2212	2186	2186	21,10	21,10	21,35
SITC v4	4017	3811	3820	36,79	36,88	38,78
Total						
$\star$	42035	21715	29586	209,66	285,66	405,86
Añadiendo enlaces entre CPV 2008 y *Product Ontology*						
ProductOntology	∞	10000	N/A	96,55	96,55	∞
Total con vocabulario de *Product Ontology*						
$\star$	∞	31715	39586	306,21	382,21	∞

Tabla 7.1: Porcentaje de Ganancia Real y Máxima al enlazar las Clasificaciones de Productos con el CPV 2008.

Adicionalmente, los resultados agregados de este experimento, se visualizan gráficamente en las siguientes gráficas, ver Figuras 7.3 y 7.4.

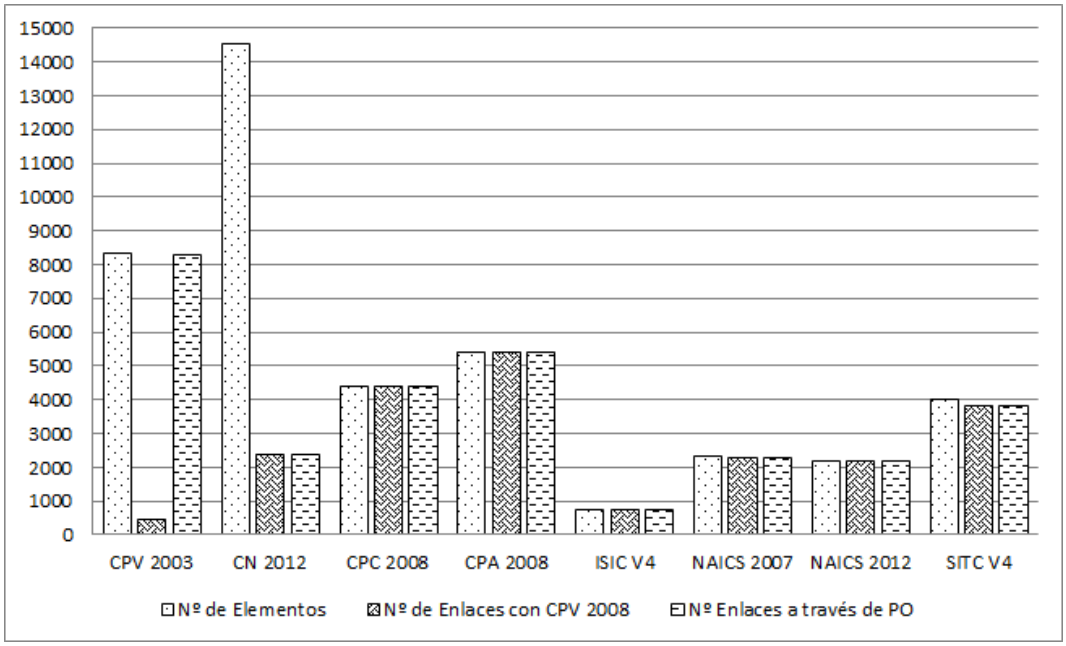


Figura 7.3: Gráfica de Número de Elementos y Enlaces entre las PSCs y el CPV 2008.

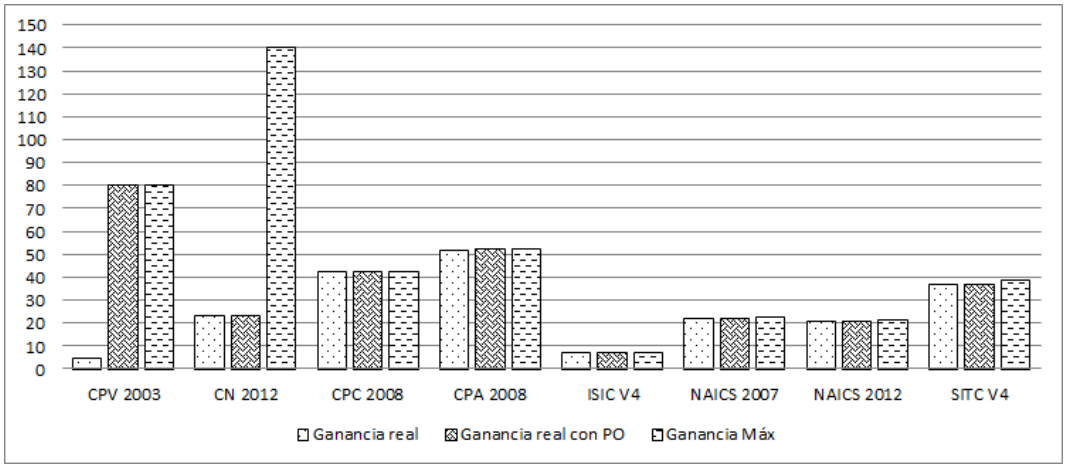


Figura 7.4: Gráfica de Ganancia en expresividad.

7.1.2.2 Punto de vista Cualitativo

La ejecución de este experimento requiere la evaluación uno a uno de los criterios diseñados en las distintas tablas para su posterior evaluación. Para ello, se han acumulado en una hoja de cálculo las respuestas de cada uno de los enfoques y agregado todos los resultados en las Tablas 7.2 y 7.3, también de forma gráfica en las Figuras 7.5, 7.6 y 7.7 se pueden observar los resultados obtenidos. De esta forma, se recoge por un lado la valoración y por otra la representación gráfica respecto al modelo de referencia. En el Apéndice C se presenta la valoración para el modelo de referencia como ejemplo de evaluación y para la visualización del diseño de las tablas de validación.

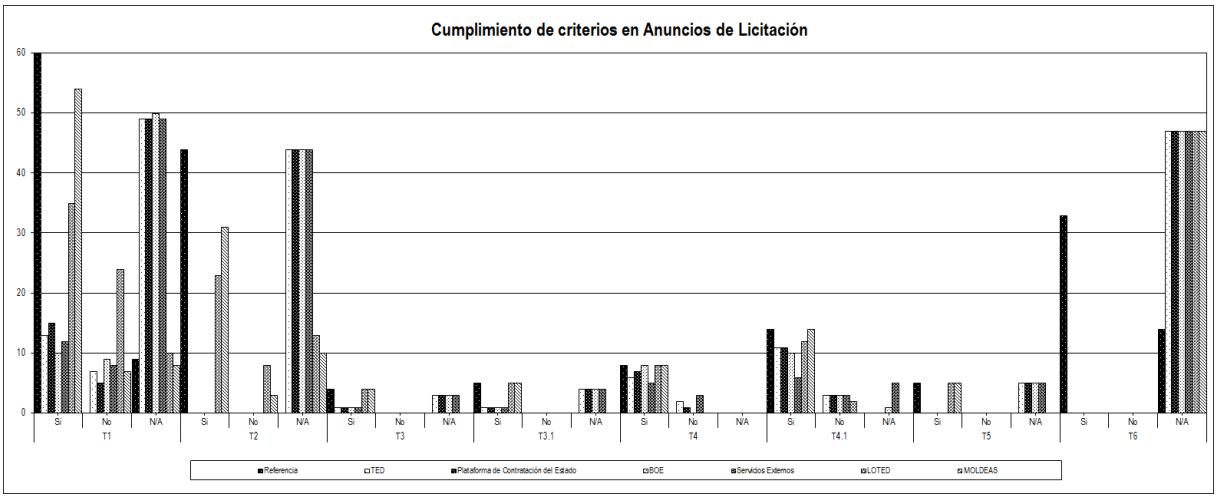


Figura 7.5: Gráfica del Grado de Cumplimiento de Criterios en Anuncios de Licitación.

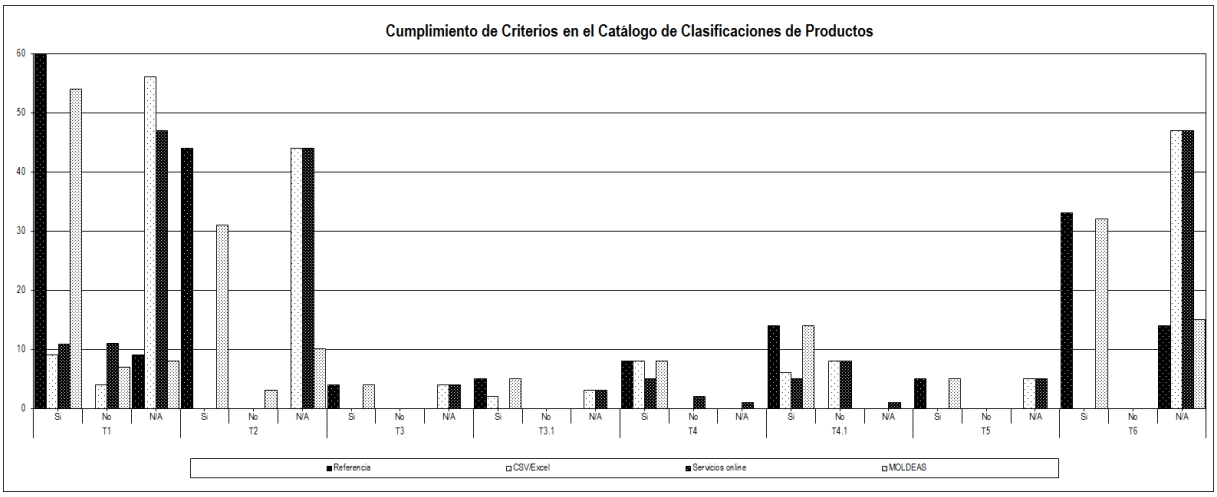


Figura 7.6: Gráfica del Grado de Cumplimiento de Criterios en el Catálogo de Clasificaciones de Productos.

En la Tabla 7.4 se presenta la suma de todos los criterios para cada una de las tablas de validación, con el objetivo de ofrecer una visión simplificada de la evaluación. Con el objetivo de representar esta información de forma gráfica y poder realizar una comparación visual se han elaborado las siguientes gráficas tanto para los anuncios de licitación, ver Figura 7.8, como para el catálogo de clasificaciones de productos, ver Figura 7.9, y los datos relativos a las organizaciones, ver Figura 7.10.

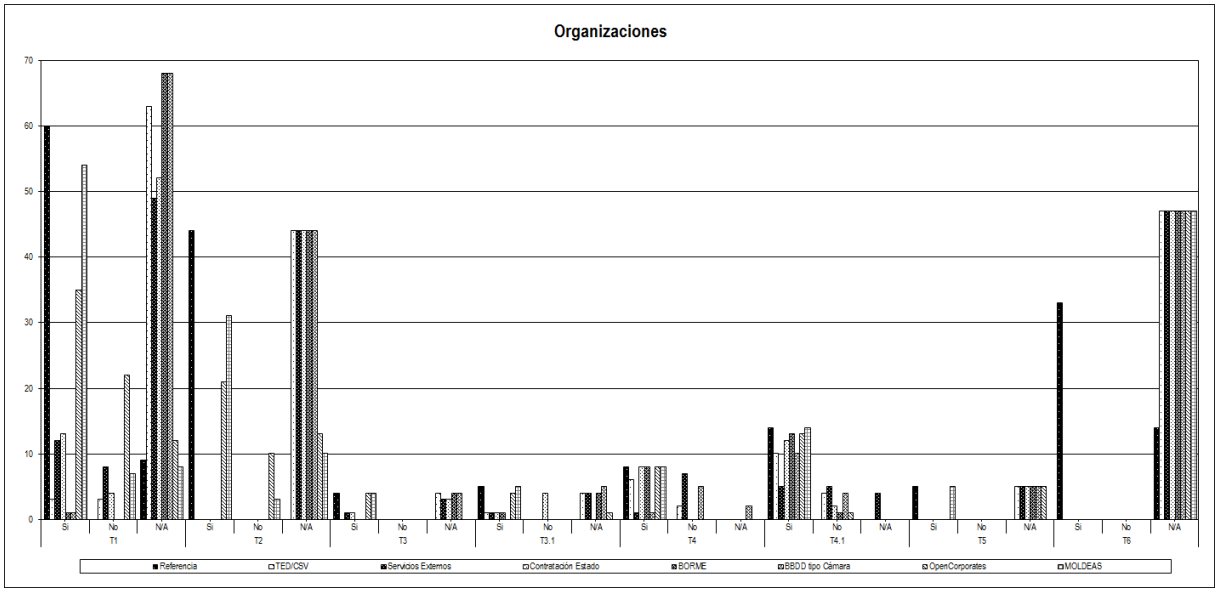


Figura 7.7: Gráfica del Grado de Cumplimiento de Criterios en las Organizaciones.

Versión	T^1	T^2	T^3	T^3_1	T^4	T^4_1	T^5	T^6
Anuncios de Licitación								
TED	65	$\odot$	100	100	75	78,57	$\odot$	$\odot$
Plataforma de Contratación	75	$\odot$	100	100	87,5	78,57	$\odot$	$\odot$
BOE	52,63	$\odot$	100	100	100	76,92	$\odot$	$\odot$
Servicios Externos	60	$\odot$	100	100	62,5	66,66	$\odot$	$\odot$
LOTED	59,32	74,19	100	100	100	85,71	100	$\odot$
MOLDEAS	88,52	91,18	100	100	100	100	100	$\odot$
Catálogo de Clasificaciones de Productos								
CSV/MSExcel	69,23	$\odot$	$\odot$	100	100	42,86	$\odot$	$\odot$
Servicios on-line	50	$\odot$	$\odot$	$\odot$	71,43	38,46	$\odot$	$\odot$
MOLDEAS	88,52	100	100	100	100	100	100	100
Organizaciones								
TED	50	$\odot$	$\odot$	100	75	71,43	$\odot$	$\odot$
Plataforma de Contratación	76,47	$\odot$	100	20	100	85,71	$\odot$	$\odot$
BORME	100	$\odot$	$\odot$	100	100	92,85	$\odot$	$\odot$
Servicios Externos	60	$\odot$	100	100	12,5	50	$\odot$	$\odot$
BBDD externa	100	$\odot$	$\odot$	$\odot$	16,67	71,43	$\odot$	$\odot$
OpenCorporates	61,40	67,74	100	100	100	92,86	100	$\odot$
MOLDEAS	88,52	91,18	100	100	100	100	100	$\odot$

Tabla 7.3: Tabla agregada de Validación Conjunta con Porcentajes $\oplus$ entre aplicables.

Versión	⊕	⊖	⊙	Total	% ⊕ entre aplicables
Referencia	173	0	23	196	100
Anuncios de Licitación					
TED	32	12	152	≡	72,72
Plataforma de Contratación	35	9	152	≡	79,54
BOE	30	12	154	≡	71,42
Servicios Externos	25	14	157	≡	64,10
LOTED	92	34	70	≡	73,01
MOLDEAS	121	10	65	≡	92,36
Catálogo de Clasificaciones de Productos					
CSV/MSExcel	25	12	159	≡	67,56
Servicios on-line	21	21	154	≡	50
MOLDEAS	166	7	23	≡	93,86
Organizaciones					
TED	20	9	167	≡	68,96
Plataforma de Contratación	35	10	151	≡	77,77
BORME	23	1	172	≡	95,83
Servicios Externos	20	20	156	≡	50
BBDD externa	12	9	175	≡	57,14
OpenCorporates	85	33	78	≡	72,03
MOLDEAS	121	10	65	≡	92,36

Tabla 7.4: Tabla agregada de Validación Conjunta con Valores Totales.

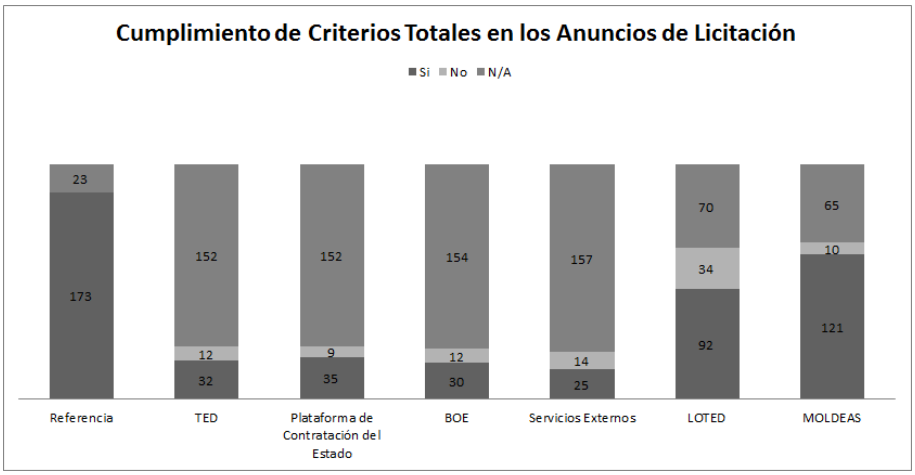


Figura 7.8: Gráfica del Grado de Cumplimiento Total de Criterios en Anuncios de Licitación.

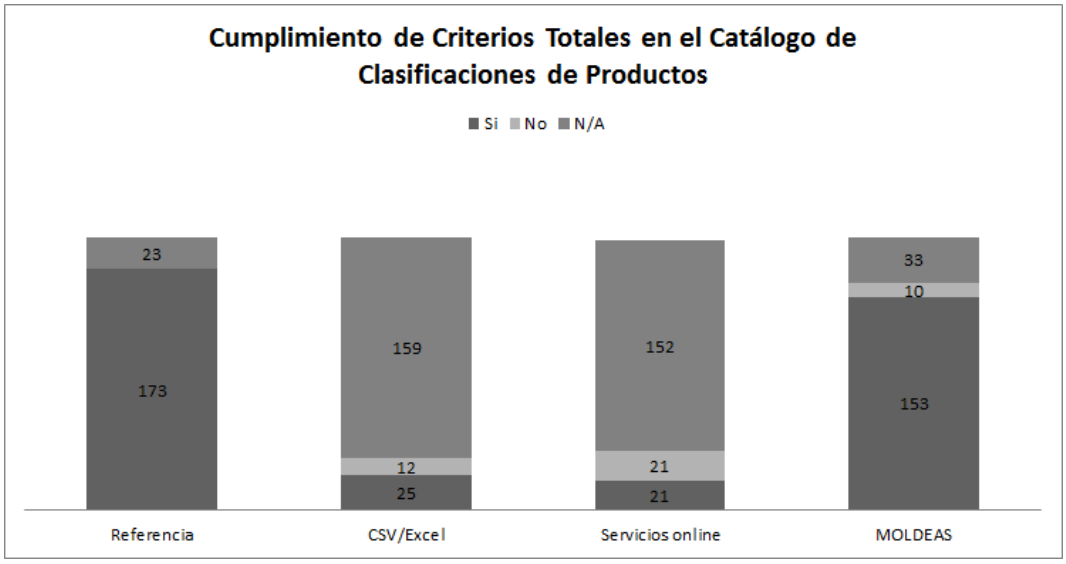


Figura 7.9: Gráfica del Grado de Cumplimiento Total de Criterios en el Catálogo de Clasificaciones de Productos.

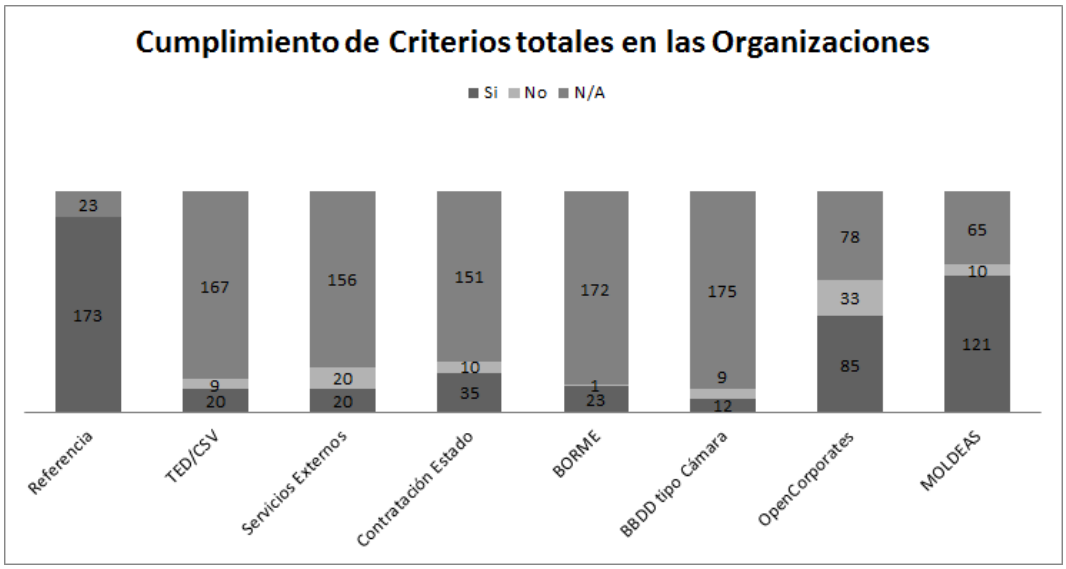


Figura 7.10: Gráfica del Grado de Cumplimiento Total de Criterios en las Organizaciones.

7.1.3 Validación del experimento sobre la aplicación de *Linked Data* a las Licitaciones Públicas

7.1.3.1 Punto de vista Cuantitativo

El experimento diseñado y probado permite la verificación del porcentaje de ganancia al añadir y enlazar un vocabulario controlado con el CPV 2008. La validación de este experimento debe plantearse asumiendo las siguientes premisas:

- La selección del CPV 2008 como vocabulario matriz para el acceso a los anuncios de licitación pública en el ámbito de la Unión Europea.
- La condición de que todos los anuncios de licitación pública deben estar etiquetados obligatoriamente con este vocabulario.

- La restricción de que el número de términos en el CPV 2008 es limitado.
- La diferencia de recuperación de información entre un sistema basado en texto libre en contraposición con los basados en un vocabulario controlado.
- La diversificación de vocabularios controlados, taxonomías, etc., existentes para el etiquetado de productos y servicios en diferentes ámbitos y con objetivos dispares: trazabilidad, extracción de estadísticas, etc.
- La necesidad de realizar una combinación efectiva entre las clasificaciones de productos para asegurar la comparación entre los mismos.
- El número de documentos recuperados es completo si se utiliza todo el vocabulario del CPV 2008.

Bajo estas condiciones, el paradigma basado en *Linked Data* permite que, a través de una correcta aplicación de un ciclo de vida para la generación de las clasificaciones de productos en RDF y su enlazado posterior, se suministre el soporte necesario para cumplir con las premisas fijadas aumentando la capacidad expresiva del vocabulario de entrada para la recuperación de información de los anuncios de licitación pública en el ámbito europeo, aumentando así la capacidad expresiva de los posibles clientes.

Desde el punto de vista del análisis de los resultados obtenidos y señalados en la sección anterior mediante la Tabla 7.1, se pueden extraer los siguientes puntos clave:

- La ganancia máxima ideal de expresividad se obtendrá cuando el vocabulario a añadir al conjunto de entrada contenga el mayor número de términos, es el caso de la clasificación de productos CN 2012. No obstante, el porcentaje de ganancia real dependerá siempre del número de términos que han sido enlazados al vocabulario matriz, en este caso el CPV 2008.
- Hay que tener en cuenta que los enlaces generados desde una clasificación al CPV 2008 se encuentran bajo un umbral μ según el cual se admite la creación del enlace entre un concepto de una clasificación y el CPV 2008. Este umbral puede variar de una aplicación a otra y dependiendo del sistema de reconciliación de entidades utilizado será más o menos exacto. El objetivo reside en la posibilidad de aumentar el vocabulario de entrada pudiendo realizar consultas bajo un determinado umbral *mu* de concreción del enlace.
- Todos los enlaces generados entre las clasificaciones de productos y el CPV 2008 $1503 + 462 = 1965$ son exactos, porque han sido realizados por fuentes oficiales. Esto supone un incremento del 8,65 % y 6,64 % (a través de *ProductOntology*) en el número de términos nuevos posibles en el vocabulario de entrada, perfectamente enlazados sin ningún umbral de ambigüedad, simplemente realizando su transformación a *Linked Data* y creando los enlaces entre sí de forma automática. Los porcentajes restantes de 91,35 % y 93,36 % respectivamente, indican enlaces automáticos, esta situación es habitual en el ámbito de *Linked Data* y la única forma de decrementar estos valores consiste en la intervención manual para la validación de los enlaces con expertos del dominio.
- Respecto a la ganancia en términos porcentuales se fija un máximo de 405,86 %, un real de 209,66 % y un real a través de *ProductOntology* de 285,66 %, lo que supone respectivamente un incremento de la expresividad de 5, 3 y cerca de 4 veces más, en el número de nuevos términos a utilizar en el vocabulario de entrada debido al uso de *Linked Data* respecto a la versión actual.
- La selección de una nueva clasificación para ser enlazada con el CPV 2008 se debería basar en los siguientes criterios:

  1. Número de términos.

2. Ganancia ideal y real.

No obstante, se establece un nuevo parámetro consistente en el número de enlaces exactos que se pueden establecer entre el vocabulario a añadir y el matriz, facilitando así la comprobación de la calidad de la adición del nuevo vocabulario. Esta situación pese a que responde a un entorno ideal, la realidad es que en el ámbito de datos enlazados se trabaja bajo unos ciertos umbrales de incertidumbre que son admitidos por la propia comunidad. Evidentemente, dependiendo del tipo de aplicación este punto debe ser considerado en mayor o menor medida.

- En cuanto al uso de un vocabulario intermedio como es *ProductOntology* su uso se convierte en estratégico por su carácter internacional. En general, no existe una mejora excesivamente reseñable respecto al enlazado de términos directo o indirecto salvo en el CPV 2003, en el cual sólo se habían contemplado los enlaces exactos, y que supone un incremento del número de términos enlazados desde el 4,46 % al 80,25 %. Esto implica que el proceso de enlazado inicial entre una clasificación y el CPV 2008 ha sido lo suficientemente bueno como para no ser necesario un paso intermedio mediante otro vocabulario que pudiera realizar el descubrimiento de nuevos enlaces.

- Respecto al enlazado entre el CPV 2008 y *ProductOntology* si bien supone una ganancia apreciable en cuanto a expresividad de entrada, 10000 nuevos términos posibles, simplemente se ha implementado y señalado pero no se considera estrictamente adecuado aunque aumente considerablemente la expresividad, ya que el vocabulario de entrada posible de *ProductOntology* es $infty$ por construcción del mismo, por lo que se prefiere su uso como puente entre una PSC y el CPV 2008 y no se compara con las demás clasificaciones de productos en cuanto a características: vocabulario controlado, realizado por una entidad, etc.

- Finalmente, es conveniente tener presente el problema de la recursión y posibilidad de convertir un vocabulario controlado de entrada (finito) en infinito por la suposición de nuevos enlaces entre elementos que directamente no están enlazados. Esta cuestión es perjudicial en dominios concretos y cerrados, como en el caso de los anuncios de licitación pública europea, pero si es beneficiosa (o admisible) en un entorno abierto para la realización de consultas de extracción de información sobre una base de datos infinita como es la Web, siempre teniendo presente el tiempo de la federación de consultas, etc., como se ha presentando en secciones anteriores.

7.1.3.2 **Punto de vista Cualitativo**

El experimento y evaluación realizada permiten realizar una calificación y comparación entre los distintos enfoques disponibles para la publicación y acceso a la información y datos referente a los anuncios de licitación, incluyendo el catálogo de clasificaciones de productos y las organizaciones. El diseño de las tablas, como se ha señalado, busca la valoración de criterios correspondientes a la iniciativa *Linked Data* y *Open Data*. De esta forma, si las valoraciones son completas en términos de valores positivos para las Tablas T^3, T_1^3 y T^4 se puede asegurar que los datos evaluados cumplen todos los criterios para ser considerados datos enlazados y abiertos, además de situarse con la máxima valoración en el modelo de 5 $\star$. Teniendo en cuenta esta premisa, aquellos enfoques que se sitúen bajo estas premisas estarán facilitando el acceso a la información de los anuncios de licitación mediante la aplicación completa de estas iniciativas y por lo tanto, se puede asegurar que todos los beneficios y ventajas que se proponen bajo estos paradigmas se cumplen, por ejemplo los presentados en la Tabla T_1^4. Además de esta valoración principal de los criterios de *Linked Data* y *Open Data* se han propuesto otras tablas con el objetivo de validar la calidad de los datos producidos en el sentido de la información y metainformación que contienen, el modelo de producción y publicación utilizado y las facilidades de consumo de los datos suministrados, estas características se ven reflejadas en la Tabla T^1. Adicionalmente y al igual que en el diseño del *software*, es conveniente la aplicación de

patrones de diseño con el objetivo de resolver problemas comunes mediante un enfoque sistemático, esta cuestión conlleva la valoración de los patrones de diseño utilizados al modelar los datos enlazados que se propone en la Tabla T^2.

Finalmente, como uno de los puntos clave para la difusión de los datos enlazados y cumpliendo la motivación inicial de esta iniciativa de reutilización de información, es importante la validación en cuanto a la posibilidad de pertenencia a la nube de datos enlazados, Tabla T^5, así como a los registros CKAN, Tabla T^6. Si bien los criterios de estas últimas tablas no son obligatorios en el sentido de *Linked Data*, si que tienen una importancia capital cuando nos referimos a datos enlazados abiertos, ya que carece de sentido realizar un esfuerzo para la apertura y enlazado de datos si la meta final no reside en su posterior reutilización.

A continuación se realiza una validación pormenorizada para cada uno de los conjuntos de datos y enfoques evaluados, se ha agrupado la valoración respecto a las tablas realizadas ya que la justificación de los resultados es similar en cada uno de los conjuntos de datos:

1. Atendiendo a las calificaciones de cada uno de los enfoques se puede observar como varios de ellos tienen características propias de los procesos implicados en la iniciativa de *Linked Data* en cuanto a la publicación de datos y el diseño del acceso a la información a través de URIs, de ahí que los distintos enfoques obtengan ciertos criterios positivos en la Tabla T^1 pese a que no se consideren bajo el paradigma de datos enlazados. Sin embargo, el esfuerzo realizado en el proyecto LOTED se ve reflejado en un grado aceptable de cumplimiento de los criterios evaluados, la negatividad de otros queda justificada por la ausencia de metainformación en los *datasets*. En el caso particular del enfoque propuesta por este trabajo, MOLDEAS, las calificaciones son positivas en un grado muy alto debido a la aplicación de un ciclo de vida para la generación de datos enlazados como un proceso de ingeniería, no obstante los puntos negativos obtenidos se refieren a cuestiones relacionadas con la documentación adicional necesaria que debería ser labor del experto en el dominio y dirigida a diferentes usuarios (técnicos, finales, etc.).

2. En cuanto a la validación de acuerdo a los patrones de diseño presentes en la Tabla T^2, los enfoques que utilizan datos enlazados si reflejan la aplicación de los mismos, si bien no se considera necesario la aplicación de todos ellos, si es conveniente aplicarlos en la mayor medida posible.

3. De nuevo la Tablas T^3, T^3_1 van dirigidas a los enfoques basados en datos enlazados principalmente, por lo que aquellos esfuerzos que no se encuentran bajo esta paradigma resultan ampliamente damnificados en su evaluación, cubriendo tan sólo alguna características básica.

4. Sin embargo, la calificación de la mayor parte de todos los enfoques en lo referente a los principios de *Open Data* en la Tablas T^4, T^4_1 es verdaderamente positiva debido principalmente a la naturaleza de la información y datos que es intrínsecamente abierta. Aquellos enfoques que ofrecen servicios de negocios se ven penalizados debido a que utilizan datos abiertos pero su estrategia o política posterior con los mismos restringe su uso y tan sólo cumplirán estos requisitos si los servicios de suscripción fueran libres respecto a la información, manteniendo su coste respecto al servicio construido sobre los datos, por ejemplo análisis, predicción o estadísticas. Evidentemente, esta situación sería ideal y también deben obtener rentabilidad del coste de recuperación de información.

5. Las Tablas T^5, T^6 son orientadas nuevamente a las iniciativas basadas en datos enlazados ya que carecen de sentido para las demás debido a su naturaleza. En este sentido, tanto los enfoques realizados por MOLDEAS y LOTED son candidatos a pertenecer a la nube de datos enlazados y a los registros CKAN sin ningún tipo de inconveniente.

En general, la tendencia que se observa en las tablas es que la información y datos relativa a los anuncios de licitación, clasificaciones de productos y organizaciones tiene un carácter abierto, como confirman los resultados, e incluso disponen de buenos fundamentos para que la realización del esfuerzo de transformación a datos enlazados sea asumible tanto por las Administraciones Públicas como por terceros, así lo demuestra el trabajo presentado en este documento a través del enfoque realizado en MOLDEAS. La aplicación de los principios de *Linked Data* y *Open Data* permiten asegurar ventajas y beneficios en el acceso a la información y manejo de los datos tanto para las personas físicas como para los componentes software, facilitando su consumo y permitiendo impulsar la creación de nuevos servicios de valor añadido con la información relevante de los anuncios de licitación y su entorno. De esta manera, el esfuerzo realizado por las empresas actuales de recuperar información de fuentes de datos heterogéneas y con distintos formatos como PDF, HTML, etc., es ampliamente mitigado debido al uso de datos enlazados.

7.1.4 Evaluación del experimento sobre la aplicación de *Linked Data* a las Licitaciones Públicas

7.1.4.1 Punto de vista Cuantitativo

La ejecución y validación del experimento del uso de datos enlazados en la recuperación de información de los anuncios de licitación aumentando la expresividad del vocabulario de entrada (CPV 2008) permite dar respuesta a las preguntas establecidas en el diseño del experimento.

- ¿Cuál es la expresividad actual, en términos de número de conceptos para realizar consultas, para el acceso a la información de anuncios de licitación?

  La versión inicial y los sistemas disponibles en Internet hacen uso del vocabulario CPV 2008 para la creación de consultas sobre los anuncios de licitación disponiendo de una expresividad como vocabulario de entrada de 10357 términos. Debido a que este vocabulario es normativo a nivel europeo, en el sentido de que cualquier anuncio de licitación debe contener al menos un código CPV 2008, se puede establecer que utilizando el conjunto completo de términos del vocabulario de entrada se puede acceder a la base documental completa de los anuncios de licitación, independientemente de la tecnología utilizada.

- ¿Cuál es la ventaja de uso de un modelo RDF para la expresión y recuperación de la información de los anuncios de licitación?

  Además de las ventajas intrínsecas del uso de RDF y la realización práctica del enfoque de *Linked Data* sobre las clasificaciones de productos y la información de los anuncios de licitación pública europea, el principal beneficio reside en el enlazado de datos desde otro vocabulario de entrada al CPV 2008 permitiendo el incremento del número de términos del vocabulario de entrada que inicialmente sólo contiene a los términos del CPV 2008 y mediante esta técnica se incrementa notablemente, aumentando el universo del discurso para la recuperación de información de los anuncios de licitación.

- ¿Cómo los datos enlazados favorecen el aumento de expresividad en la ejecución de consultas y por tanto facilitan la recuperación de los anuncios de licitación?

  Según la respuesta realizada en la pregunta anterior y la validación presentada en la Sección 7.1.3.1, el uso de datos enlazados beneficia la creación de consultas enriquecidas respecto al vocabulario de entrada permitiendo al cliente la utilización de un mayor número de términos convirtiendo el conocimiento del CPV 2008 en opcional. En la validación realizada y con la transformación de las clasificaciones de productos se obtiene un incremento entre 3 y 4 veces

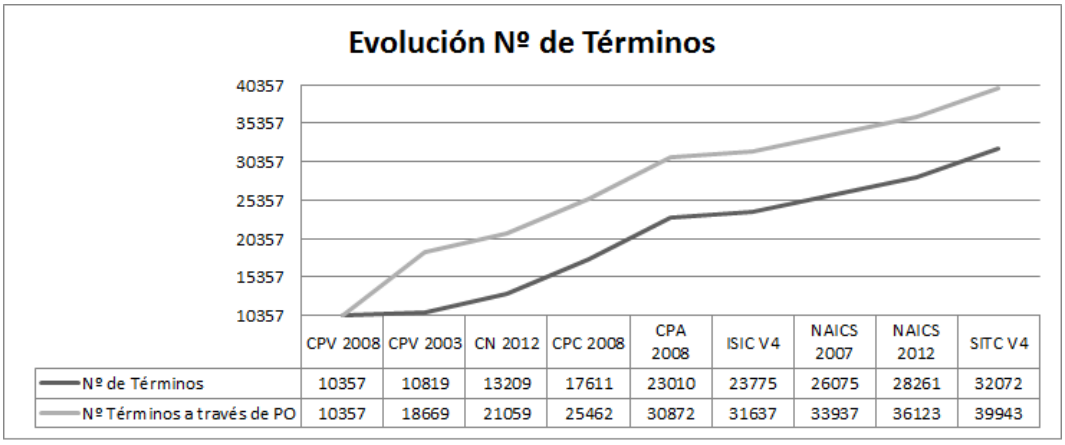


Figura 7.11: Evolución Número de Términos.

más respecto al número de términos posibles iniciales provenientes del CPV 2008, la evolución del número de términos se puede observar en la Figura 7.11.

- ¿Cuál es el beneficio real del uso de datos enlazados para representar la información?

  El beneficio real reside en la representación de información y datos mediante un modelo estándar (RDF) que permite el uso de lenguajes (SPARQL) y protocolos (HTTP) estándar para su acceso facilitando la recuperación de información. Evidentemente, el contraste respecto a las actuales versiones en HTML, PDF, MSExcel, etc., y sistemas de recuperación opacos suministradores que generan servicios de negocio sobre ellos, es crítico, ya que simplemente el manejo de información de distintas fuentes y formatos requiere un esfuerzo que se solventa en gran medida con el uso de datos enlazados. Es necesario resaltar que el enfoque realizado en este documento es posible debido a la posibilidad de utilizar *Open Data* y mejorar éstos con la iniciativa de *Linked Data*. Adicionalmente, se establece una fórmula para el cálculo de la ganancia al introducir una nueva clasificación enlazada con el CPV 2008.

- ¿Se incurre en algún error al aumentar la expresividad?

  El enlazado de datos consta de una variable de incertidumbre que si bien es aceptada por la comunidad, debe ser tenida en cuenta para la realización de aplicaciones de carácter crítico. En este sentido y siguiendo con la validación realizada, al menos, se obtiene un incremento entre el 6,64 % (con *ProductOntology*) y 8,65 % exacto del vocabulario controlado de entrada. En el caso particular de recuperación de documentos estos porcentajes son muy relevantes pero incluso la introducción de términos "relacionados" bajo un umbral de incertidumbre es conveniente ya que el funcionamiento del sistema se mantiene y se aumenta la expresividad de entrada para la realización de consultas. Por todo ello, pese a que este problema de enlazado de recursos y reconciliación de entidades está asociado a la iniciativa de *Linked Data*, sólo se puede solventar con el esfuerzo de expertos en el dominio con el consiguiente coste.

En resumen, el enfoque seguido en este experimento y en el estudio de este documento es extrapolable a un entorno o dominio en el cual exista un vocabulario controlado para la recuperación de documentos, existan diferentes vocabularios provenientes de distintas fuentes candidatos a interoperar con el inicial y en consecuencia a servir para el aumento del universo de discurso en ese dominio.

7.1.4.2 **Punto de vista Cualitativo**

La ejecución y validación del experimento del grado de cumplimento de criterios referentes a datos enlazados abiertos permite dar respuesta a las preguntas planteadas en el diseño del experimento.

- ¿El ciclo de vida seguido y los datos generados certifican la aplicación de buenas prácticas de *Linked Data*?

  A la vista de la evaluación realizada de los distintos enfoques ha quedado patente que en la actualidad tanto la información como los datos disponibles son abiertos en su mayor parte e incluso son utilizados para la construcción de servicios comerciales pero el esfuerzo requerido para llevar a cabo estos productos se ve reflejado en el precio de los mismos. La aplicación de los principios de *Linked Data* puede rebajar ampliamente este coste de acceso a la información promoviendo la competitividad y creando una administración más transparente. En este sentido el trabajo realizado en MOLDEAS partiendo de la definición de un ciclo de vida (procesos, métodos y tareas) y del conjunto de datos referente al dominio de *e-Procurement* permite asegurar los criterios de *Linked Data* además de cumplir las características más deseables en cuanto producción, publicación y consumo de datos.

- ¿El ciclo de vida seguido y los datos generados certifican el cumplimiento de los principios de *Open Data*?

  De la misma forma que en la pregunta anterior el modelo de ciclo de vida seguido ha permitido conservar el carácter de datos abiertos intrínseco a la información y datos relativas al entorno de los anuncios de licitación como se ha reflejado en las tablas de validación referentes a *Open Data*.

- ¿Qué nivel del modelo de 5 $\star$ se puede establecer?

  Consecuentemente por la aplicación del ciclo de vida y las valoraciones realizadas en los criterios relativos al nivel dentro del modelo de 5 $\star$ se refleja la consecución de la calificación más alta. En este punto cabe realizar una reflexión sobre el nivel de estrellas que deben tener los datos abiertos enlazados, especialmente los referentes a las Administraciones Públicas. En muchos casos y especialmente desde un punto de vista de un desarrollador es suficiente la publicación de datos a través de RSS (3 $\star$), esta situación es consecuencia directa de la simplicidad de este formato tanto en representación como por su madurez, ahora bien, la publicación de datos bajo 5 $\star$ confiere nuevas cualidades a los mismos como un modelo formal subyacente, la posibilidad de reconciliación automática, realización de procesos de razonamiento, etc., que para aplicaciones de gestión de la información avanzada resultan de gran interés. En conclusión, desde un punto de vista práctico el nivel de 3 $\star$ sería suficiente pero las ventajas adicionales de los datos 5 $\star$ debe conducir a los propietarios de datos a exigir este nivel, aunque el esfuerzo inicial sea mayor la recompensa a medio/largo plazo debe ser superior, más aún teniendo en cuenta la tendencia actual de apertura y enlazado de datos en diferentes dominios.

- ¿Qué porcentaje de patrones de diseño se han aplicado en los datos generados?

  Los patrones de diseño en disciplinas como la Ingeniería de *Software* permiten dar solución a problemas comunes de una forma sistemática y probada. En muchas ocasiones el exceso de uso de patrones implica añadir una complejidad extra a un determinado *software* para el cual ya se ha fijado una fecha de caducidad, viéndose así su capacidad de extensión cercenada desde el inicio. Con ello, se pretende ejemplificar que si bien la complejidad de introducir patrones de diseño en el dominio del *software* requiere un esfuerzo, permite obtener beneficios posteriores. En el caso de los datos enlazados ocurre de la misma forma, además de ofrecer una solución a problemas comunes también pretende que el calado y la perduración en el tiempo de los datos

enlazados sea la mayor posible, facilitando así su reutilización. Es por ello que en MOLDEAS se ha aplicado el mayor número de patrones posible siempre teniendo presente la valoración del coste/beneficio y la no obligatoriedad en su aplicación, de esta forma se puede fijar un porcentaje de uso en X %.

- ¿Los datos generados pueden pertenecer a la nube de datos enlazados abiertos?

  De acuerdo a los resultados en la evaluación tanto los datos particulares concernientes a los anuncios de licitación como los del catálogo de las clasificaciones de productos y las organizaciones son candidatos a formar parte de la nube de datos enlazados. Hasta el momento y como resultado colateral a este trabajo el catálogo de clasificaciones de productos ya ha sido incluido formalmente en la nube de datos enlazados, los referentes a los anuncios de licitación y las organizaciones debido a que han sido cedidos por la empresa Gateway S.C.S dentro de la ejecución del proyecto "10ders Information Services" se ha pospuesto su inclusión oficial.

- ¿Los datos generados pueden pertenecer a un registro CKAN?

  En relación a la respuesta de la pregunta anterior la metainformación necesaria para pertenecer a este tipo de registro está disponible para cada uno de los *dataset* RDF, por lo que es posible realizar el registro en cualquier portal de datos enlazados basado en CKAN.

- ¿Se puede asegurar que los datos son enlazados y abiertos?

  La evaluación realizada en las Tablas T^3, T_1^3 y T^4 verifica que se cumplen los requisitos, principios y criterios necesarios para que los datos realizados bajo MOLDEAS sean considerados como abiertos y enlazados.

- ¿Qué beneficios se obtienen del cumplimiento de estos objetivos?

  El enfoque realizado en MOLDEAS trasciende más allá de la simple transformación de datos siguiendo unas directrices, se ha establecido un ciclo de vida perfectamente definido que permite la gestión eficaz de la información y de los datos consiguiendo que un proceso considerado dramático, en algunos casos, se convierta en un proceso de ingeniería cuantificable. Es por ello que además de las ventajas inherentes a los datos enlazados abiertos es necesario añadir los propios generados del trabajo realizado en este estudio. En particular y desde el punto de vista de los datos enlazados abiertos se destacan los siguientes beneficios:

  - Mejora en el acceso a la información y datos, integración de distintas fuentes de datos, servicios de acceso basados en lenguajes formales.
  - Aumento de la visión global de los datos, expresividad y estructuración.
  - Aplicación intensiva de estándares.
  - Incremento del conocimiento en el dominio de la contratación pública electrónica.
  - Creación de un proceso de gestión del ciclo de vida de los datos enlazados.
  - Impulso de la reutilización de la información y datos, mayor poder de redistribución
  - Minimización de restricciones tecnológicas. Impulso de la reutilización automática.
  - Minimización de aspectos discriminatorios en el uso de la información y de los datos.
  - Aumento de la transparencia, inclusión y responsabilidad, mayor conocimiento de la información de procedencia.
  - Incremento en la concienciación de los avances en la gestión e integración de datos.
  - Alineación con las actuales propuestas estratégicas de futuro.
  - ...

Versión	T^1			T^2			T^3			T_1^3			T^4			T_1^4			T^5			T^6		
	$\oplus$	$\ominus$	$\odot$	$\oplus$	$\ominus$	$\odot$	$\oplus$	$\ominus$	$\odot$	$\oplus$	$\ominus$	$\odot$	$\oplus$	$\ominus$	$\odot$	$\oplus$	$\ominus$	$\odot$	$\oplus$	$\ominus$	$\odot$	$\oplus$	$\ominus$	$\odot$
Referencia	**60**	**0**	**9**	**44**	**0**	**0**	**4**	**0**	**0**	**5**	**0**	**0**	**8**	**0**	**0**	**14**	**0**	**0**	**5**	**0**	**0**	**33**	**0**	**14**
Anuncios de Licitación																								
TED	13	7	49	0	0	44	1	0	3	1	0	4	6	2	0	11	3	0	0	0	5	0	0	47
Plataforma de Contratación	15	5	49	0	0	44	1	0	3	1	0	4	7	1	0	11	3	0	0	0	5	0	0	47
BOE	10	9	50	0	0	44	1	0	3	1	0	4	8	0	0	10	3	1	0	0	5	0	0	47
Servicios Externos	12	8	49	0	0	44	1	0	3	1	0	4	5	3	0	6	3	5	0	0	5	0	0	47
LOTED	35	24	10	23	8	13	4	0	0	5	0	0	8	0	0	12	2	0	5	0	0	0	0	47
MOLDEAS	54	7	8	31	3	10	4	0	0	5	0	0	8	0	0	14	0	0	5	0	0	0	0	47
Catálogo de Clasificaciones de Productos																								
CSV/MSExcel	9	4	56	0	0	44	0	0	4	2	0	3	8	0	0	6	8	0	0	0	5	0	0	47
Servicios on-line	11	11	47	0	0	44	0	0	4	0	0	5	5	2	1	5	8	1	0	0	5	0	0	47
MOLDEAS	54	7	8	31	3	10	4	0	0	5	0	0	8	0	0	14	0	0	5	0	0	32	0	15
Organizaciones																								
TED	3	3	63	0	0	44	0	0	4	1	0	4	6	2	0	10	4	0	0	0	5	0	0	47
Plataforma de Contratación	13	4	52	0	0	44	1	0	3	1	4	0	8	0	0	12	2	0	0	0	5	0	0	47
BORME	1	0	68	0	0	44	0	0	4	1	0	4	8	0	0	13	1	0	0	0	5	0	0	47
Servicios Externos	12	8	49	0	0	44	1	0	3	1	0	4	1	7	0	5	5	4	0	0	5	0	0	47
BBDD externa	1	0	68	0	0	44	0	0	4	0	0	5	1	5	2	10	4	0	0	0	5	0	0	47
OpenCorporates	35	22	12	21	10	13	4	0	0	4	0	1	8	0	0	13	1	0	0	0	5	0	0	47
MOLDEAS	54	7	8	44	0	0	4	0	0	5	0	0	8	0	0	14	0	0	5	0	0	0	0	47

Tabla 7.2: Tabla agregada de Validación Conjunta con Valores Parciales.

7.2 Experimento sobre el Sistema MOLDEAS

7.2.1 Diseño del experimento sobre el Sistema MOLDEAS

Uno de los puntos clave para el triunfo de la iniciativa de *Linked Data* consiste en asegurar que las aplicaciones puedan consumir los datos para la construcción de servicios de valor añadido aprovechando el modelo semántico subyacente. En este sentido, en el caso objeto de estudio de este documento y tras el diseño e implementación del sistema MOLDEAS, en concreto del componente `moldeas-api` y teniendo en cuenta la ejecución del proyecto "10ders Information Services" se consideró la necesidad de completar el proceso de consumo mediante el diseño de una serie de métodos que suministren servicios de recuperación de información sobre las licitaciones.

Inicialmente, este sistema se encuadra dentro de la tecnología existente para la recuperación de información, pudiendo así considerarse como un sistema experto o bien un buscador híbrido de licitaciones. En ambos casos, la casuística y la gran cantidad y variedad de información contemplada en los documentos de los anuncios de licitación convierten a un sistema de recuperación de estos documentos en realmente un objeto de estudio de gran profundidad. Por ello, en el desarrollo de esta tesis se ha optado por la realización de un demostrador público que haciendo uso de datos enlazados y de las capacidades de las tecnologías semánticas sea capaz de abordar los siguientes puntos:

- Consumir los datos enlazados desde un lenguaje de programación.
- Crear un sistema de recuperación de información.

Bajo estos dos grandes objetivos se ha diseñado una primera versión de un prototipo experimental prestando especial interés en su diseño para ser extendido en siguientes iteraciones.

7.2.1.1 Recuperación de Información

Este término más conocido como *Information Retrieval* [211] es el área de conocimiento sobre la tecnología para la adquisición, representación, almacenamiento, organización y acceso a recursos de información. El objetivo de la misma se basa en ofrecer al usuario una serie de métodos para la obtención de documentos de su interés de acuerdo a una consulta. Como ejemplo de este tipo de sistemas se encuentran los motores de búsqueda sintácticos basados en la búsqueda de cadenas de texto coincidentes y que han demostrado en la última década su increíble capacidad para localizar y recuperar información en los contenidos presentes en la web. El uso de algoritmos como *PageRank* [245], en el caso de Google, para establecer la relevancia de un documento combinado con el refinamiento de las técnicas estadísticas de búsqueda de texto han aumentado la potencia y precisión de la búsqueda en la Web. Por eso, el uso de semántica como potenciador de los sistemas de búsqueda no trata de suplantar la búsqueda tradicional sino ofrecer soluciones en dominios verticales en los cuales el uso de estas técnicas puede ser mejorado.

En general, en un dominio vertical el espacio de búsqueda y la terminología utilizada por un usuario es más complicada por lo que la formulación de las consultas con criterios adecuados se convierte en un problema [213] en el momento de obtener resultados satisfactorios.

En el caso que nos ocupa de recuperación de anuncios de licitación, la aplicación de un algoritmo como *PageRank* es relativamente no aplicable ya que no se dispone de enlaces entre los documentos presentes y las relaciones entre los anuncios son conceptuales, clasificadas de acuerdo un tipo de contrato u otras variables de información como el emisor, la región, etc. La propuesta diseñada se basa en la aplicación de semántica para la recuperación de los anuncios de licitación combinando tanto técnicas semánticas como otras que exploten las capacidades semánticas y los sistemas de recomendación.

Un anuncio de licitación se puede considerar como un documento en el cual existen dos tipos de propiedades:

Extensionales. Estas propiedades son independientes del contenido del texto y se refieren propiamente a metadatos del documento como el tipo de anuncio, el contratante o la localización.

Intensionales. El contenido propio de un documento, es decir, el discurso, conjunto de oraciones que constituyen una unidad de significado

La formalización y la explotación de las propiedades de los documentos difiere según el caso. Para el tratamiento de las propiedades extensionales se suelen utilizar herramientas del ámbito de la gestión documental y utilizadas ampliamente en biblioteconomía y archivística utilizando vocabularios controlados para su gestión. En cambio, en cuanto a las propiedades intensionales aunque la teoría del análisis del discurso ha alcanzado una madurez importante en los últimos años todavía su representación semántica es bastante compleja.

En el caso objeto de estudio de este documento y en lo relativo a los anuncios de licitación queda de manifiesto que las propiedades extensionales permiten actualmente dar soporte a los sistemas de recuperación de información. A través del concepto de metadato y el uso de un vocabulario controlado como el CPV es posible clasificar los anuncios de licitación de forma facetada de acuerdo a un sistema homogéneo. De esta manera, se obtienen estructuras lógicas basadas en conceptos que permiten la recuperación efectiva de información. De acuerdo al modelo realizado en el Capítulo 5 se observa como los metadatos propios de los anuncios de licitación son modelados permitiendo la realización de procesos de clasificación y búsqueda. El uso de este tipo de propiedades facilita enormemente la recuperación de información para este dominio ya que el esfuerzo realizado por facilitar estos procesos compensa y en cierta manera evita el uso de propiedades intensionales que en la mayor parte de los casos inducen a la recuperación de elementos de información no deseados.

7.2.1.2 Tareas, Modelos y Métodos de Recuperación de Información

En general, existen dos grandes tipos de sistemas para el procesamiento de elementos de información: los de recuperación de información y los sistemas gestores de bases de datos orientadas al manejo de datos estructurados [60]. La diferencia entre ambos enfoques reside en que en el primero de los casos el sistema intenta obtener la mejor relación entre documentos relevantes recuperados y documentos relevantes, mientras que en el segundo caso existe un álgebra definido que devuelve todos los resultados de acuerdo al mismo.

Se pueden establecer una serie de tareas en los sistemas de recuperación de información:

- Recuperación *ad-hoc*. El paradigma sería un conjunto de documentos fijo sobre el cual se ejecutan diferentes consultas.
- Categorización o clasificación de documentos. En este caso, se trata de proporcionar propiedades intensionales a los documentos para su recuperación y filtrado.
- *Clustering* de documentos. En el caso de que no se puedan agrupar documentos mediante propiedades intensionales o vocabularios controlados se utilizan distintas técnicas para agrupar documentos con características comunes.
- Segmentación de documentos. Consiste en la división de un documento en unidades de información coherentes para ser tratado con una mejor eficacia.

Evidentemente las tareas de la recuperación de información guían hacia los distintos de modelos de representación interna, entre los cuales se puede destacar:

- *Bag-of-Terms*. Cada documento se representa a través de una serie de términos índices, por ejemplo los documentos de una biblioteca. Se basa en el principio de composicionalidad, según el cual un documento está formado por únicamente por los términos índices y en consecuencia se puede asumir que ese documento trata sobre ese tema.

  Este tipo de representación es eficiente pero carece de semántica entre los términos salvo que, como en el caso objeto de estudio, el vocabulario controlado utilizado ya contenga una semántica definida, en este caso una taxonomía con relaciones de especificidad.

- Términos con pesos. La semántica asociada a un documento con un número determinado de términos no es completa en el sentido de que no todos los términos índices utilizados tienen el mismo peso. Por ello, la evolución natural del modelo anterior consiste en asignar un peso w_{ij} para cada término i en un documento j.

  Existen dos puntos clave para la asignación de pesos en un documento: su frecuencia de aparición y su distribución dentro del conjunto de todos los documentos, lo que permite realizar dos suposiciones:

  1. Los términos que más se repiten deben ser considerados como representativos, grado de significatividad. En este caso, esta medida se refiere a la frecuencia.
  2. El número de documentos en los que aparece un determinado término implica una medida del grado de discriminación, por lo que aquellos que aparecen en muchos documentos deberían obtener un peso menor. En este segundo caso la medida se refiere a la frecuencia inversa.

  Si N es el número de documentos total y n_i el número de documentos en los que aparece un término t_i, se puede definir la frecuencia de un término t_i en un documento j como tf_{ij}. La frecuencia inversa de un término se calcula como $idf_i = log(N/n_i)$.

  Estas fórmulas deben ser tenidas en cuenta en el momento del cálculo de los pesos y en la mayoría de los casos se asume que los términos no están correlados con el objetivo de que la asignación de pesos sea independiente.

Una vez que se ha proporcionado una visión sintética de los modelos de representación cabe especificar los modelos de recuperación más habituales:

- Modelo Booleano. Se trata del modelo de recuperación más sencillo y está basado en teoría de conjuntos y álgebra de Bool. El acceso a la recuperación de la información en este modelo se basa en la ejecución de una consulta con distintos términos estableciendo una operación booleana entre los mismos para la recuperación de los documentos que cumplan dicha expresión. El conjunto total de documentos se divide por tanto en relevantes y no relevantes.

  Este método se ha utilizado inicialmente debido a su sencillez tanto desde un punto de vista formal como de implementación, se ejecuta con una gran eficiencia. El principal problema de la expresión de consultas mediante un conjunto de términos y operadores booleanos recae en que usuarios no expertos encuentran dificultades para la creación de consultas que impide rescatar los documentos realmente relevantes para los usuarios. Otra de las desventajas de este modelo reside en que no existe un encaje parcial de la consulta respecto a los documentos al tratarse de un modelo binario, por lo tanto, un documento o cumple con la expresión de la consulta entera o no.

- Modelo Vectorial. Este modelo surge para abordar los problemas encontrados en el modelo anterior para la realización de encajes parciales y la asignación de relevancia de un documento a una consulta. Para ello, se utiliza una representación tanto de las consultas como de los documentos mediante un vector en el cual se indican los términos y pesos. Si tanto el vector de la

consulta como el del documento están próximos, desde un punto de vista geométrico, se puede asumir que el documento es relevante.

La mejora de este modelo reside en que no sólo se basa en la creación de dos conjuntos como el modelo booleano sino que se basa en la similitud de los términos utilizados en la consulta y la definida en los documentos. Con todo ello, el modelo vectorial ha obtenido buenos resultados tanto desde un punto de vista formal como en la práctica.

- Modelo Probabilístico. Una vez repasados dos de los grandes modelos de recuperación de información basados en conjuntos y en geometría, cabe citar este modelo basado en términos de teoría de probabilidades. La teoría que subyace a este modelo consiste en asignar la probabilidad de que un documento sea relevante para una determinada consulta.

El repaso de los métodos de representación de información y los modelos de recuperación permiten establecer una primera visión de la complejidad de un sistema de recuperación de información con el cual caracterizar el entorno para la búsqueda de anuncios de licitación públicos.

7.2.1.3 Búsqueda y Generación de Consultas

El proceso de búsqueda de documentos tras la definición del método de representación y el modelo de recuperación debe además acometer determinadas tareas para el procesamiento del lenguaje de natural.

La principal tarea consiste en la extracción de términos para el indexado. La estrategia habitual para llevar a cabo esta tarea se basa en técnicas de transformación de los términos iniciales mediante procesamiento del lenguaje natural [301] con el objetivo de eliminar texto superfluo y normalizar los términos realmente interesantes para el indexado. La ejecución básica de esta tarea consiste en la eliminación de las denominadas *stop-words*, normalización de mayúsculas y minúsculas, aplicación de técnicas de *stemming*, etc. La aplicación de estas técnicas puede obtener diferentes términos de indexado conllevando un comportamiento diferente en el sistema de recuperación de información.

Una vez que el índice de búsqueda se ha generado conteniendo para cada documento un conjunto de términos índices obtenidos tras el análisis léxico de los elementos presentes en el documento ya se está en disposición de realizar búsquedas y recuperar información sobre este índice. De forma habitual el proceso a seguir consiste en la introducción de una consulta que es tratada inicialmente mediante el mismo tipo de análisis utilizado para la generación de términos de indexado. Dependiendo del modelo de recuperación utilizado se obtendrán resultados o documentos con una determinada relevancia para la consulta. Como mejora a este proceso se utilizan técnicas de expansión de consultas que de forma automática o semi-automática realizan transformaciones en la consulta inicial para obtener una consulta enriquecida con nuevos términos. Estas técnicas se basan principalmente en el uso de tesauros para la obtención de relaciones de sinonimia, antonimia, hiperonimia e hiponimia, etc., para la construcción de una nueva consulta, también se suele utilizar la realimentación del usuario y el estudio de los registros de las consultas para optimizar la generación de consultas extendidas, en ambos casos tesauros como Wordnet [138] son relevantes y existen enfoques híbridos [172]. Finalmente, otro tipo de sistemas de búsqueda realizan un enfoque compuesto combinando técnicas de búsqueda sintáctica añadiendo semántica para la expansión de conceptos y generación de consultas avanzadas, este tipo de enfoque es considerado como búsqueda semántica previa alineación de la consulta con conceptos de una ontología.

En este sentido y en los últimos años con la aparición de las ontologías como elemento vertebrador del conocimiento han aparecido técnicas como *Spreading Activation* (implementadas en la biblioteca ONTOSPREAD) para la exploración de conceptos [59,261] en redes semánticas, con el objetivo de facilitar dos funciones: 1) la selección de conceptos y 2) la ponderación de términos adicionales

de búsqueda de acuerdo a medidas de similaridad [139]. El objetivo final de aplicación de estos enfoques es la construcción de sistemas de búsqueda híbridos [34, 237, 264, 283] basados en semántica, en realidad se pueden alinear con los tradicionales enfoques de expansión de consultas comentados anteriormente.

7.2.1.4 Medidas de Evaluación

El éxito de un sistema de recuperación de información vendrá determinado por la evaluación para la comprobación del ajuste de los documentos recuperados de acuerdo a las expectativas del usuario. En general, se pueden evaluar múltiples aspectos: eficiencia, en cuanto al coste espacio-temporal; efectividad, cantidad de documentos relevantes recuperados; esfuerzo, en el momento de la construcción de consultas y usabilidad.

Entre las medidas más importantes se encuentra la efectividad del sistema ya que permite establecer la bondad de la recuperación y comparar distintos modelos de recuperación de información. Entre este tipo de métricas se establce la siguiente Tabla 7.5, característica de los documentos recuperados. Este tipo de medidas son ampliamente utilizadas en las clasificaciones binarias en las cuales un individuo pertenece o no a un conjunto.

	Actual class (expected)	
Predicted class (observation)	tp (*true positive*)	fp (*false positive*)
	fn (*false negative*)	tn (*true negative*)

Tabla 7.5: Clasificación de resultados de la Recuperación de Información.

Precisión. Capacidad del sistema para recuperar sólo los documentos relevantes de acuerdo a una consulta. La intepretación de esta medida será mejor cuanto más cercana a 1 se encuentre ya que indicará que la proporción de todos los documentos o recursos relevantes extraídos de la consulta respecto al número total de documentos extraídos.

$$Precision = \frac{tp}{tp+fp}$$
$$Precision = \frac{D_{relevantes} \cap D_{recuperados}}{D_{recuperados}}$$

Recall. Capacidad del sistema para todos los documentos relevantes de acuerdo a una consulta. Dependiendo del contexto de uso también se denomina *Sensivity*, por ejemplo en clasificación binaria. La interpretación de esta medida es evidente ya que un valor cercano a 1 indica que se han devuelto todos los documentos relevantes a la consulta por ello se debe combinar con la precisión.

$$Recall = \frac{tp}{tp+fn}$$
$$Recall = \frac{D_{relevantes} \cap D_{recuperados}}{D_{relevantes}}$$

Precisión y *recall* también se pueden combinar para la obtención de la medida $F - measure$, calculada como la media armónica.

Accuracy. Es la proporción de resultados positivos (tanto positivos como negativos) de la población observada. Es un valor de medición positivo.

$$Accuracy = \frac{tp+tn}{tp+fp+fn+tn}$$

Specifity. Es la capacidad del sistema para detectar resultados negativos. Por ejemplo, en un sistema de diagnóstico médico un valor alto de esta medida implica una alta probabilidad de presencia de la enfermedad diagnosticada.

$$Specifity = \frac{tn}{tn+fp}$$

Estas medidas permiten valorar y comparar la bondad de un sistema diagnóstico, un sistema de recuperación de información, etc., en los cuales resulta de vital de importancia la mejora constante para cumplir con las expectativas del usuario.

7.2.1.5 Caracterización de la Recuperación de Información en Anuncios de Licitación

La especificación del experimento sobre la recuperación de información en anuncios de licitación se ha abordado con los dos grandes objetivos previamente mencionados de consumir los datos enlazados producidos e implementar un demostrador público que permita realizar búsquedas sobre la información disponible. La recuperación de documentos de los anuncios de licitación se puede abordar desde diferentes puntos de vista, de acuerdo a los modelos y métodos de representación y de recuperación en la sección anterior se establece que el modelo de representación es similar a *Bag-of-Terms* ya que el contenido completo de los anuncios no está disponible y siempre se disponen de códigos CPV para etiquetar los mismos, el tipo de propiedades utilizadas en este caso serían intensionales y de acuerdo a los metadatos propios de un anuncio de licitación. En cuanto al método de recuperación se encontraría entre un híbrido entre el modelo booleano y el vectorial dependiendo del sistema del tipo de consulta que se pretenda proveer al usuario.

El sistema de recuperación de información diseñado se basa en un repositorio RDF en el cual se almacenan las descripciones RDF de cada uno de los anuncios de licitación suministrando a través de un *endpoint* de SPARQL el acceso a los mismos.

El experimento diseñado busca realizar las pruebas sobre el componente `moldeas-api` teniendo en cuenta los métodos desarrollados para la expansión de consultas. Las variables de información presentes en los anuncios de licitación y su correlación se pueden explotar de muchas formas distintas. No obstante, para este experimento se ha seleccionado el tipo de licitación, es decir, los códigos CPV de la misma para comprobar los métodos de recuperación de información debido principalmente a las siguientes causas:

- Es la variable de información cuya presencia está asegurada en todos los anuncios. Otras como el emisor, fechas, etc., también se pueden utilizar para la recuperación de información pero su importancia es ciertamente inferior en cuanto a trascendencia respecto al tipo de contrato público.
- Dentro de la ejecución del proyecto "10ders Information Services" se han solicitado consultas de los clientes del servicio "Euroalert.net" y la información proporcionada se basa en los códigos CPV exclusivamente.

Sintetizando el experimento que se realizará evaluará los códigos CPV generados (o recuperados) mediante el sistema MOLDEAS en contraposición con los suministrados por "10ders Information Services" y utilizando las medidas de evaluación Precisión (P), *Recall* (R), *Accuracy* (A) y *Especificidad* (S).

Una vez determinados los puntos clave a tener en cuenta para la evaluación de la recuperación de información es conveniente la caracterización propia del experimento que consta de las siguientes etapas:

1. Definición de los objetivos del experimento. Las preguntas a responder por el experimento serán:
   - ¿Es posible implementar un sistema de recuperación de información utilizando datos enlazados?
   - ¿Es posible explotar las relaciones semánticas establecidas para mejorar la recuperación de información?
   - ¿Cuál es el mejor enfoque para la recuperación de información en los anuncios de licitación?
   - ¿Cómo afectan los resultados en la implementación actual del sistema MOLDEAS?
2. Selección de una regla de asignación de las unidades experimentales a las condiciones de estudio. En este caso, la unidad experimental de este estudio será un repositorio RDF en el cual se encuentran almacenados $1M$ de anuncios de licitación y las clasificaciones de productos, especialmente el CPV 2008 con 10357 términos.
   - Cualitativos: tipo de entorno hardware y software, códigos CPV generados.
   - Cuantitativos: tamaño de la muestra y nº de códigos CPV iniciales.

   En este caso y centrándose en el tamaño de la muestra se han obtenido las consultas presentadas en la Tabla 7.6, de las cuales se tiene la siguiente información: consulta proporcionada por un cliente, Q_{str}, y códigos CPV seleccionados por los expertos de dominio de "Euroalert.net", Q_{cpv}. En este caso y debido al gran número de códigos CPV tan sólo se muestra la cantidad de los mismos esquivando el código y la descripción. Realmente y desde un punto de vista de la recuperación de información las medidas de evaluación dependen tan sólo del número de códigos que se hayan conseguido generar adecuadamente no así su descripción.

Q_i	**Consulta de Usuario-Q_{str}**	**Nº de Códigos CPV relevantes-#Q^i_{cpv}**
Q_1	"Comprehensive overview over all environmental technologies renewable energy products"	463
Q_2	"Tendering of public works: housing, hospitals, roads, housing developments, station drinking water treatment, reforestation"	35
Q_3	"Prefabricated buildings"	7
Q_4	"Games for children (parks swings, slides, land of play equipment in the public sphere"	26
Q_5	"Vital signs monitor"	277
Q_6	"Museum and exhibition and product launch services"	1
Q_7	"Voltmeters, instruments measuring electrical quantities, Ammeters, Instruments for checking physical characteristics, hygrometers, thermometers, measuring equipment and control, leak detector, Analyzers, Cable Splicing insulated cable joints kits, screwdrivers, hand tools , screwdriver"	117

Q_i	**Consulta de Usuario-Q_{str}**	**Nº de Códigos CPV relevantes-#Q^i_{cpv}**
Q_8	"Conservation Maintenance of pavements for roads, airfields, bridges, tunnels"	13
Q_9	"Wood poles, Wooden sleepers , Lattice towers"	10
Q_{10}	"Architectural, construction, engineering and inspection services"	173
Q_{11}	"Medical practice and related services"	13

Tabla 7.6: Consultas suministradas en el proyecto "10ders Information Services".

Respecto a los métodos empleados para la generación de códigos CPV, ver Tabla 7.7, es necesario destacar que uno de los mismos está basado en búsqueda sintáctica, M^1, y además sirve como primer filtro para la ejecución de los siguientes métodos ya que el primer conjunto de códigos CPV se obtiene de esta forma para ser utilizado como parámetro de entrada en los siguientes métodos. En el caso del segundo método se intentan explotar las relaciones de la jerarquía del CPV (`skos:broader`) para obtener un nuevo conjunto de salida. De la misma forma se procede en el método M^3 pero en este caso se utiliza la técnica de *Spreading Activation* a través de la implementación de ONTOSPREAD para obtener el conjunto de salida final. Finalmente, se ha utilizado un motor de recomendación para la obtención de códigos CPV similares a los de entrada en el método M^4, en este caso se supone que el comportamiento difiere de los anteriores métodos ya que la semántica de un sistema de búsqueda o de expansión de consulta difiere del comportamiento de un motor de recomendación, en el primer caso se busca recuperar exactamente un recurso mientras que en el segundo caso, dado un recurso se buscan similares. Este enfoque, también se ha utilizado para interpretar si el sistema de casamiento de consultas y usuarios se puede asociar a un sistema de búsqueda o a un motor de recomendación.

Método	**Descripción**	**Tecnología**
M^1	Se indexan las descripciones de los códigos CPV realizando un proceso de búsqueda sintáctica de las consultas preparadas para así obtener un conjunto de códigos CPV resultado. Similar a la reconciliación de entidades llevada a cabo para las clasificaciones de productos.	Apache Lucene y Solr
M^2	Se extraen una serie de códigos CPV candidatos, atendiendo a las relaciones de la jerarquía se obtiene un nuevo conjunto de códigos CPV de salida ponderado.	M^1 + ponderación jerarquía
M^3	Se extraen una serie de códigos CPV candidatos, atendiendo a las relaciones de la jerarquía y mediante *Spreading Activation* conjunto de códigos CPV de salida ponderado.	M^1 + ONTOSPREAD
M^4	Se extraen una serie de códigos CPV candidatos, atendiendo al histórico de las relaciones entre códigos del 1 M de anuncios se obtiene con el motor de recomendación un nuevo conjunto de códigos CPV de salida ponderado.	M^1 + Apache Mahout

Tabla 7.7: Métodos de generación de códigos CPV.

3. Especificación de las medidas de trabajo en cuanto a la respuesta. Para cada uno de los métodos a evaluar se recogen los códigos CPV devueltos con el objetivo de calcular los elementos de la Tabla 7.5 en contraste con los disponibles en las consultas preparadas. De esta forma y atendiendo a las características de la recuperación de anuncios de licitación se recuperarán tantas licitaciones como códigos coincidentes existan entre la consulta Q^i_{cpv} formada por los códigos CPV generados por los expertos del dominio y el conjunto $Q^{M^i}_{cpv}$ generado por los respectivos métodos.

4. Ejecución de un experimento piloto. Con el objetivo de cubrir todo el proceso transversalmente se ha realizado una primera iteración con una sola consulta extrayendo los códigos CPV para los distintos métodos y realizando los cálculos *PRAS* tanto para la validación de las fórmulas como la comparación con los resultados esperados.

5. Especificación de un modelo. En este caso, se podría pensar en la realización de un contraste de hipótesis para establecer y asegurar que un método mejora o no la generación de códigos CPV, sin embargo y teniendo en cuenta que la muestra de consultas es de tan sólo 11 no es considerada significativa para el uso de distribuciones estadísticas y el tipo de test a practicar para este tipo de experimento correspondería a un Test De Wilcoxon en el caso de que las muestras sean pareadas.

6. Esquematización de los pasos a seguir. La ejecución del experimento sigue los siguientes pasos:

   a) Para cada una de las consultas del usuario Q_{str} identificadas a través de Q_i se aplica un método M_i de recuperación de códigos CPV.

   b) Cada uno de los métodos M_i se configura para que devuelva al menos $\#Q^i_{cpv}$ elementos para así poder comparar conjuntos de cardinalidades iguales. En algunos casos los métodos utilizados no son capaces de generar $\#Q^i_{cpv}$ elementos debido a la baja representatividad de los últimos códigos generados y al gran número de códigos presentes en Q^i_{cpv}. Se ha desestimado utilizar la media del número de elementos de cada conjunto Q^i_{cpv} debido a su excesiva divergencia.

   c) Cada conjunto resultado $Q^{M^i}_{cpv}$ se compara con el conjunto de resultado esperado Q^i_{cpv} a través de un *script* en AWK, ver Figura 7.12, para obtener las medidas tp, fp, fn y tn (mediante hoja de cálculo).

   d) Finalmente con estos valores se generan los valores *PRAS* para cada uno de los métodos M_i y cada consulta de entrada Q_i.

```
awk 'NR==FNR{a[$0];next} $0 in a' $1 $2 > "$1–$2–tp"
TP=‘ cat "$1–$2–tp" | wc –l ‘
FP=‘awk 'NR==FNR { a[$0]; next } !($0 in a)' "$1" "$2" | wc –l ‘
FN=‘awk 'NR==FNR { a[$0]; next } $0 in a { delete a[$0]; next } 1; END
  { for (b in a) print b }' "$1" "$1–$2–tp" | wc –l ‘
echo "TP, FP, FN"
echo $TP "," $FP "," $FN
rm –f "$1–$2–tp"
```

Figura 7.12: Extracción de valores tp, fp y fn.

7. Determinación del tamaño muestral. Como ya se ha comentado anteriormente, en el apartado de factores cuantitativos, se utilizarán los datos generados durante el proceso de promoción de datos a *Linked Data* de los anuncios de licitación, de las clasificaciones de productos y de las consultas preparadas suministradas por el servicio de “Euroalert.net”.

8. Revisión de las decisiones anteriores.

7.2.2 Ejecución del experimento sobre el Sistema MOLDEAS

La ejecución del plan diseñado para el experimento arroja los resultados que se presentan a través de la Tabla 7.8, en la cual para cada una de las consultas de usuario Q_{str} identificadas a través de Q_i, se ha ejecutado la generación de códigos CPV a través de los distintos métodos, M^i, para obtener un conjunto de salida de códigos CPV $Q_{cpv}^{M^i}$ permitiendo establecer las métricas de evaluación: Precisión, *Recall*, *Accuracy* y Especificidad.

7.2.3 Validación del experimento sobre el Sistema MOLDEAS

La realización del experimento sobre la recuperación de códigos CPV y en consecuencia recuperación de información de los contratos públicos es motivado por el objetivo de ayudar al experto del dominio a decidir los códigos CPV que se ajustan a las consultas en lenguaje natural de los clientes o usuarios. La dificultad de selección de códigos reside en varios puntos que la experiencia ha dejado patente:

- Extensión del posible conjunto de entrada, 10357 términos del vocabulario controlado CPV 2008. La necesidad de facilitar la expresividad del conjunto de términos de entrada viene determinada por el aumento de las posibilidades de expresar consultas y enlazar códigos de forma automática, como se ha presentado en la Sección 7.1.3. Ahora bien, un conjunto de entrada más amplio, no implica mayor dificultad para seleccionar los códigos CPV, sino que el punto clave consiste en establecer a partir de un código CPV sus posibles códigos relacionados, independientemente de la extensión del vocabulario de entrada. En este sentido, el experto del dominio debe conocer cómo se etiquetan los anuncios de licitación y estimar las intenciones del usuario ya que en la mayoría de los casos no existe relación entre la expresión del usuario (intuición sobre cómo se han etiquetado los anuncios de licitación) y el etiquetado real de los anuncios. En este contexto, se pone de manifiesto la intervención del experto y las herramientas que faciliten la selección de códigos de forma automática, un sistema en el cual existan más relaciones permite establecer conjuntos de códigos CPV más completos.

- Explotación de las relaciones jerárquicas de la taxonomía del CPV 2008. En general, la búsqueda sintáctica tradicional se basa en el encaje de cadenas de texto preprocesadas mediante distintas técnicas (*stopwords*, *stemming*, *n-grams*, etc.) con el objetivo de facilitar el acceso al mayor número de documentos o recursos de acuerdo a una consulta de entrada. Tanto el proceso de indexación de recursos como el preprocesamiento de la consulta permite que la comparación se realice en las mismas condiciones. No obstante, la explotación de relaciones más allá del lenguaje natural se delega en la propia aplicación antes de proveer la verdadera consulta de entrada. El enriquecimiento o expansión de la consulta a través de la navegación por relaciones semánticas presentes en taxonomías o vocabularios controlados permite obtener un conjunto de términos estrechamente relacionados y validados ya que han sido consensuados por una comunidad. Este enfoque de expansión en el caso del CPV 2008 se considera perfectamente aplicable ya que no es delegado en una taxonomía genérica y externa como puede ser Wordnet, sino que se utilizan las relaciones entre los distintos códigos, que como se ha visto en la estructura del CPV 2008 no son necesariamente de herencia sino de grado de especificidad. De esta forma, la extensión del conjunto de entrada mediante la navegación por la jerarquía de la taxonomía se considera válida para obtener un conjunto enriquecido que pueda ser trasladado al sistema de búsqueda final, en este caso a una consulta sobre un motor sintáctico en el cual se han indexado las descripciones de los códigos CPV 2008.

- Diferenciación de comportamiento entre búsqueda y recomendación. La distinción entre estos dos enfoques tecnológicos para la recuperación de información reside en que los sistemas de búsqueda están orientados principalmente a la obtención de respuestas concretas para consultas de usuario planteadas como cadenas de texto para necesidades específicas. En general este enfoque también es correcto para la obtención de recursos relacionados con las consultas del usuario, aunque se basen en el procesamiento de lenguaje natural en muchos casos las respuestas obtenidas son bien aceptadas por el usuario. En el caso de recomendación el comportamiento es significativamente diferente ya que en este caso se buscan recursos relacionados no directamente obtenidos por una consulta de usuario sino más bien de acuerdo a un comportamiento histórico o la explotación de relaciones. En ambos casos, el objetivo es suministrar un mecanismo para el filtrado de la información. No obstante, en los últimos años pese al gran triunfo de los sistemas de búsqueda, los sistemas de recomendación se han convertido en su gran competidor y la tendencia dicta que se pueden convertir en su sucesor natural. Sin embargo, la deformación provocada por el uso de sistemas de búsqueda provoca que en algunos casos las respuestas proporcionadas por un sistema de recomendación no dispongan de la acogida apropiada por los usuarios. En este sentido, la recomendación de códigos CPV de acuerdo al histórico presente en los anuncios de licitación se plantea estratégica para futuras iteraciones de un sistema de soporte a la decisión sobre qué códigos CPV seleccionar para la mejora de la recuperación de información proveniente de anuncios de licitación.

Las evidencias destacadas de estos puntos conducen la validación del experimento con la consiguiente extracción de conclusiones:

- Las consultas, Q_i, preparadas por los expertos de acuerdo a clientes reales dejan patente que la búsqueda de códigos CPV se basa en un encaje sintáctico inicial mediante el cual se establece un conjunto de códigos potenciales a partir de los cuales se seleccionan todos los elementos jerárquicamente inferiores, aumento del grado de especificidad, como nuevos elementos a buscar. Esta situación se refleja en la cardinalidad de cada uno de los conjuntos de códigos CPV de las consultas Q_i. De esta forma si uno de los elementos de entrada fuera una división se obtienen todos los descendientes, grupos, clases y categorías, aumentando el espectro de códigos válidos enormemente. La disposición de un conjunto esperado de resultados tan numeroso implica por una parte, que la búsqueda real de anuncios de licitación se puede ver distorsionada y por otra parte, que es más sencillo que los métodos de generación de códigos CPV obtengan una precisión mayor. En el punto opuesto se encuentran aquellas consultas muy específicas en las cuales tan sólo se espera un código de salida, en este caso, la precisión de los métodos se ve menoscabada y en muchos casos en un encaje directo es más que suficiente.

- En general el método más preciso es M^1, correspondiente a un motor de búsqueda sintáctico basado en Apache Lucene y Solr mediante la aplicación de filtros sobre el texto tanto para el indexado de los elementos del CPV 2008 como para el texto de la consulta del usuario. De nuevo, esta situación evidencia que el punto anterior sobre el grado de cumplimiento de las expectativas del experto del dominio sobre la búsqueda y recomendación de códigos CPV se basa en técnicas puramente sintácticas. Evidentemente, la afinación de las técnicas de procesamiento de descripciones de códigos de clasificaciones de productos realizadas para la reconciliación de entidades y enlazado con otras clasificaciones se ha visto recompensado por su posterior reutilización para la búsqueda de códigos en general. No obstante, las aspiraciones del experto de dominio no se ven totalmente colmadas y existe un porcentaje importante de mejora para conseguir la máxima precisión con valor 1. Si bien en la mayoría de los sistemas de búsqueda actuales está situación se admite, un estudio posterior debería sondear con los expertos del dominio si los códigos sugeridos realmente son falsos positivos o bien son códigos que tendrían sentido para la consulta de entrada.

- El método M^2 hace uso de las mismas técnicas que el anterior para obtener un conjunto inicial de conceptos similar a M^1 para a partir de ese momento obtener la jerarquía completa para cada uno de los códigos repartiendo un valor de activación entre los hijos. Por ello, se debería mantener la precisión del método M^1 pero en este caso se ve decrementada debido principalmente a dos causas: 1) reordenación de la ponderación de los códigos y 2) la selección de códigos no sólo está basada en la navegación jerárquica. Esta situación es tremendamente crítica ya que si bien el experto tiende a situarse en una jerarquía y a partir de ese nivel seleccionar todos los descendientes, dependiendo del caso y de la extensa casuística de los anuncios de licitación, la experiencia conlleva a seleccionar códigos no relacionados directamente en ningún nivel. Por ejemplo, en los anuncios de licitación relativos a equipamiento de un hospital se encuentran códigos relacionados con la adquisición de *software* y de colchones cuya regla es difícilmente representable con una sola taxonomía. Algún tipo de información extra o el dominio en general podrían ayudar a solventar esta información suministrando así un sistema experto con una precisión elevada.

- El método M^3 utiliza las técnicas propuestas en M^1 para realizar el salto entre las descripciones textuales de una consulta de usuario a los códigos CPV de la taxonomía. Este conjunto inicial es utilizado por la técnica de *Spreading Activation* para descubrir y enriquecer el conjunto de códigos de salida. No obstante, los resultados no son especialmente satisfactorios debido de nuevo a que los resultados esperados en las consultas preparadas parten de una creación basada en el encaje de descripciones textuales y conocimiento experto de dominio de difícil formalización de forma genérica en una técnica como *Spreading Activation*. Una posible propuesta consiste en establecer un contexto de búsqueda en el cual las técnicas avanzadas sean más "informadas" de cuáles son los caminos a recorrer para simular el conocimiento experto.

- Finalmente, el método M^4 cuya ejecución es similar a los anteriores pero que difiere en la técnica de extracción de códigos CPV tampoco obtiene mejoras respecto al método básico M^1. Si bien este método reutiliza la información histórica de un millón de anuncios de licitación no consigue aproximarse en cuanto a precisión al conocimiento experto. La causa de esta situación se debe precisamente a que el comportamiento esperado de un sistema de búsqueda respecto a uno de recomendación difiere diametralmente. En este caso y teniendo en cuenta que las consultas preparadas se basan en encajes textuales, la recomendación como técnica de aproximación de códigos CPV se aprecia que no es apropiada. No obstante, las posibilidades de variación de parámetros en los algoritmos de recomendación tienen un carácter tan extenso que si podrían acercarse en implementaciones posteriores.

- La implementación de estos métodos para la generación de códigos CPV a partir de descripciones textuales de búsqueda de anuncios de licitación pone de manifiesto la posibilidad de consumo de datos enlazados para la construcción de servicios de soporte a la decisión para la recuperación de información de anuncios de licitación. Esta primera aproximación constituye un demostrador público de consumo de *Linked Data* y sienta las bases para el perfeccionamiento de la técnica de *Spreading Activation* en este ámbito y de los motores de recomendación como futuro para la recuperación de información. No obstante y de igual forma que en otros muchos contextos la búsqueda tradicional basada en técnicas sintácticas sigue manteniéndose como principal herramienta para obtener los resultados deseados de un usuario. Si bien esta situación puede deberse al comportamiento derivado de los sistemas de búsqueda existentes, la tendencia es obtener búsquedas más exactas y no directamente relacionadas con la consulta inicial, intentando discernir las intenciones del usuario y no ceñirse a una serie de términos.

7.2.4 Evaluación del experimento sobre el Sistema MOLDEAS

La ejecución y validación del experimento mediante la comparación de un conjunto de resultados esperado Q^i_{cpv} y un conjunto de resultados obtenido, $Q^{M^i}_{cpv}$, a través de distintos métodos permite dar respuesta a las preguntas planteadas al inicio del experimento.

- ¿Es posible implementar un sistema de recuperación de información utilizando datos enlazados?

  Evidentemente, el tipo y formato de una fuente de datos no debe ser impedimento para la construcción de servicios en un dominio determinado. En este caso particular, los datos enlazados provenientes de los anuncios de licitación pueden ser perfectamente consumidos desde un lenguaje de programación como Java mediante el acceso a un *endpoint* de SPARQL y realizando las consultas pertinentes para cargar los objetos de negocio con la información y datos necesarios. Hasta este punto, el consumo de datos enlazados no representa ninguna mejora respecto a otro tipo de almacenamiento y representación, la ventaja estratégica reside en su flexibilidad para añadir e integrar nuevos datos y a la posibilidad de explotación de las relaciones establecidas entre los distintos recursos. Sin embargo, el esfuerzo para el consumo de datos todavía es elevado ya que es complicado mantener esta flexibilidad de representación en su consumo desde un lenguaje de programación. Las APIs y bibliotecas disponibles facilitan esta labor pero distan de las herramientas disponibles para los sistemas de gestión de bases de datos relacionales. No obstante, este esfuerzo es recompensado ya que los servicios generados están verdaderamente más informados y son capaces de explotar eficientemente la información y los datos para la construcción de servicios de valor añadido. Además en el nuevo contexto de la *Web of Data* este requisito de consumo será prácticamente indispensable para cualquier aplicación. Finalmente, el consumo de datos enlazados presenta algunas cuestiones abiertas como la confianza y la verificación de la procedencia de los datos que deben ser abordadas para impulsar este proceso.

- ¿Es posible explotar las relaciones semánticas establecidas para mejorar la recuperación de información?

  En el demostrador público del sistema MOLDEAS se permite la consulta y recuperación de anuncios de licitación de acuerdo a un método preestablecido y ciertos parámetros de entrada. La experimentación realizada con los distintos métodos permite adecuar este demostrador a los resultados obtenidos pero la explotación de las relaciones semánticas no se considera suficientemente relevante, al menos desde el punto de vista de un experto de dominio, por lo que aunque el uso de semántica implica una serie de ventajas intrínsecas, su aplicación actual en comparación con las expectativas del usuario deben ser mejoradas para conseguir una diferenciación real entre usar o no semántica en las aplicaciones según es entendida en las iniciativas de Web Semántica y *Linked Data*. No sólo se trata de la posibilidad de formular consultas muy complejas navegando por las relaciones, sino más bien de comprender cómo se ha de utilizar la semántica para dar respuesta a los problemas del usuario, actualmente el uso de semántica se centra en facilitar la interoperabilidad e integración de aplicaciones que en el suministro de servicios de valor añadido para el gran público. La incorporación de las grandes compañías de *software* y servicios a esta iniciativa provocará un vuelco a esta situación, impulsando y facilitando el uso de técnicas semánticas para el usuario final.

- ¿Cuál es el mejor enfoque para la recuperación de información en los anuncios de licitación?

  Según los resultados obtenidos y de acuerdo al comportamiento de los expertos del dominio, queda patente que el método M^1 es el más cercano a los resultados esperados. La explicación, como ya se ha detallado en párrafos anteriores, se debe principalmente al grado de implantación que la búsqueda tradicional posee. No obstante, tan sólo se han utilizado 11 consultas

preparadas (no se disponía de una muestra mayor) por lo que los resultados bajo estas condiciones y sin un contraste de hipótesis estadístico, deben ser tomados como una guía en la cual la valoración con una muestra de la población relevante e informativa podrían variar.

- ¿Cómo afectan los resultados en la implementación actual del sistema MOLDEAS?

  Este experimento sirve como prueba del consumo de datos enlazados en un dominio y de la aplicación de técnicas tanto tradicionales como procedentes de otros ámbitos para la recuperación de códigos CPV. La principal conclusión que se debe extraer, consiste en el planteamiento de que un sistema experto para la recuperación de anuncios de licitación públicos tiene un gran interés debido a diversos factores: variables de información múltiples, correlaciones entre las mismas, cantidad de datos a procesar, etc., que pueden ser optimizados a través de las pruebas con distintos algoritmos y técnicas. Es por ello que las próximas iteraciones de MOLDEAS se centrarán en optimizar estas técnicas en el campo de las licitaciones para que puedan ser promocionadas a otros dominios de carácter general. Con ello, se consigue realizar una demostración pública y abrir una nueva e interesante línea de investigación en el campo de los sistemas expertos para la recomendación de anuncios de licitación.

Q_i	M^1				M^2				M^3				M^4			
	P	**R**	**A**	**S**	**P**	**R**	**A**	**S**	**P**	**R**	**A**	**S**	**P**	**R**	**A**	**S**
Q_1	0,15	0,08	0,94	0,98	0,15	0,15	0,92	0,96	0,12	0,06	0,94	0,98	0,06	0,06	0,68	0,81
Q_2	0,09	0,09	0,99	1,00	0,06	0,06	0,99	1,00	0,03	0,03	0,99	1,00	0,03	0,03	0,99	1,00
Q_3	0,14	0,14	1,00	1,00	0,14	0,14	1,00	1,00	0,14	0,14	1,00	1,00	0,00	0,00	1,00	1,00
Q_4	0,19	0,19	1,00	1,00	0,00	0,00	0,99	1,00	0,12	0,12	1,00	1,00	0,00	0,00	0,99	1,00
Q_5	0,12	0,01	0,97	1,00	0,01	0,01	0,95	0,97	0,08	0,01	0,97	1,00	0,03	0,03	0,95	0,97
Q_6	1,00	1,00	1,00	1,00	0,00	0,00	1,00	1,00	1,00	1,00	1,00	1,00	0,10	0,67	0,98	0,98
Q_7	0,20	0,20	0,98	0,99	0,09	0,09	0,98	0,99	0,15	0,16	0,98	0,99	0,03	0,03	0,98	0,99
Q_8	0,08	0,08	1,00	1,00	0,08	0,08	1,00	1,00	0,08	0,08	1,00	1,00	0,00	0,00	1,00	1,00
Q_9	0,50	0,50	1,00	1,00	0,00	0,00	1,00	1,00	0,30	0,38	1,00	1,00	0,00	0,00	1,00	1,00
Q_{10}	0,39	0,39	0,98	0,99	0,42	0,42	0,98	0,99	0,34	0,35	0,98	0,99	0,16	0,16	0,97	0,99
Q_{11}	0,23	0,23	1,00	1,00	0,23	0,23	1,00	1,00	0,15	0,17	1,00	1,00	0,00	0,00	1,00	1,00
Medias Totales de Métricas PRAS																
Total	0,28	0,26	0,99	1,00	0,11	0,11	0,98	0,99	0,23	0,23	0,99	1,00	0,03	0,03	0,96	0,98

Tabla 7.8: Resultados PRAS de las consultas suministradas en el proyecto “10ders Information Services”.

7.3 Experimento sobre el Rendimiento del Sistema MOLDEAS

7.3.1 Diseño del experimento sobre el Rendimiento del Sistema MOLDEAS

El objetivo de un experimento, especialmente los estadísticos, es averiguar si unos determinados factores influyen en una determinada variable de interés para así poder cuantificar dicha influencia. En el caso de esta sección la variable a estudiar es el rendimiento de las consultas SPARQL contra el *endpoint* utilizado, si bien no es el experimento principal de este documento ya que conllevaría la comparación con distintos proveedores de repositorios RDF, si que ha surgido la necesidad de abordar esta cuestión por la latencia en las consultas. La metodología utilizada para el diseño del experimento es la repetición, en condiciones indistinguibles, con el objetivo de minimizar la variabilidad del mismo y que el error experimental sea lo más pequeño posible.

El tratamiento del rendimiento de las consultas permite estudiar si las mejoras introducidas mejoran el tiempo de ejecución de las mismas partiendo de un caso básico sin ninguna optimización y a continuación, a través de distintas repeticiones y la introducción de optimizaciones extraer aquel conjunto de mejoras que permiten extraer la influencia de las características incorporadas, seleccionando así la combinación con mejores resultados. Con este objetivo, se ejecutan las distintas iteraciones del experimento para a través de este proceso empírico detectar cambios significativos en la respuesta y comprender el comportamiento del rendimiento en las consultas. Por lo tanto, en este caso el experimento se realiza para encontrar las condiciones con las que se consigue la optimización más crítica y comparar las respuestas en tiempo de ejecución con la variación de las variables, en contraposición, no es objetivo del experimento obtener un modelo matemático que permita realizar predicciones futuras.

Un aspecto importante de este experimento es que la evaluación del mismo se realiza mediante los resultados obtenidos mediante una correcta planificación, esquivando el uso de datos históricos para no incurrir en errores debidos a la aleatoriedad. Por otra parte, durante la ejecución de un experimento suelen surgir distintos de variabilidades que deben ser tomadas en cuenta:

- Sistemática y planificada, cuyo origen viene determinada por las diferencias sistemáticas al variar las condiciones, variables del experimento. Este tipo de variabilidad es conveniente y debe ocurrir para poder evaluar el experimento.
- Debida a la naturaleza del problema y del experimento, se trata de variaciones debidas al ruido, error de medida, etc., siendo impredecibles e inevitables. En el caso del experimento objeto de estudio se origina por cuestiones relacionadas con el tiempo de latencia en la red, etc. No obstante, esta variabilidad es medible y se puede extrapolar para la toma de decisiones.
- Sistemática y no planificada, cuyo origen se debe a causas desconocidas, generando resultados sesgados. En el caso de estudio este tipo de variabilidad ha sido resuelta realizando las distintas réplicas en bloques de ejecución.

Una vez determinadas los puntos clave a tener en cuenta para la evaluación de los resultados es conveniente la caracterización propia del experimento que consta de las siguientes etapas:

1. Definición de los objetivos del experimento. Las preguntas a responder por el experimento serán:
    - ¿Cuáles son las mejoras que se pueden aplicar sobre una consulta en SPARQL para mejorar el tiempo de ejecución?
    - ¿Cuál es la combinación de mejoras que obtiene un mejor tiempo de respuesta?

- ¿Cuál es el coste de la combinación de estas mejoras?
- ¿Existe algún elemento externo de configuración que implique un incremento en el tiempo de ejecución de las consultas?
- ¿Cómo afectan los resultados en la implementación actual del sistema MOLDEAS?

2. Selección de una regla de asignación de las unidades experimentales a las condiciones de estudio. En este caso, la unidad experimental de este estudio será un repositorio RDF con un determinado orden de ejecución de cada uno de los tratamientos (variedad de consultas, casos de test T_k) para obtener las medidas de las observaciones, de esta forma se deben atender a los siguientes factores:

   - Cualitativos: tipo de entorno hardware y software, tipo de consulta en SPARQL con variaciones.
   - Cuantitativos: tamaño de la muestra, de la memoria y número de réplicas.

   En cuanto a la caracterización de los factores cualitativos, respecto al entorno hardware y software se utiliza el definido en la Sección 7.0.2. Sobre las consultas en SPARQL, existen trabajos sobre la optimización de las mismas [32, 54, 269] para la consulta eficiente de diferentes repositorios RDF [303], en este caso el tipo de consulta se centra en dos principales variables de información de los anuncios de licitación como son los códigos CPV (tipo) y NUTS (localización), para ello tras la ejecución de método de expansión de conceptos a partir de un código inicial se obtienen una serie de códigos recomendados para su posterior transformación en una consulta en SPARQL, la lista completa de los códigos se presenta en la Tabla 7.9. Es conveniente destacar que el método de expansión es independiente de esta prueba ya que simplemente se necesitan códigos independientemente de su calidad y su descripción. La selección de códigos se ha llevado a cabo contemplando los tipos presentes en el CPV (División, Grupo, Clase y Categoría) estén representados ($Q_{1.,8}$) además de proveer una consulta aleatoria Q_9, sobre los códigos NUTS se han escogido aquellos representativos de países con el objetivo de generalizar y obtener un mayor nuevo de resultados.

   En referencia a los aspectos cuantitativos del experimento, el tamaño de la muestra consta de nueve consultas ($Q_{1.,9}$), el número de réplicas para evitar el ruido y la aleatoriedad previamente comentada es de 3 y la situación de partida para cada una de ellas se produce bajo las mismas condiciones: reinicio de la máquina y 18 consultas de entrenamiento (1 simple y 1 expandida) antes de la ejecución de una réplica para la obtención de las observaciones de tiempo de ejecución. Referente al tamaño de la muestra, se utilizará el *endpoint* descrito anteriormente y con la carga de tripletas generada a partir de la aplicación de los métodos semánticos a los anuncios de licitación y siguiendo la distribución por grafos nombrados especificada en el Sección 5.2.1.5.

   Por otra parte, las características, F_i, a evaluar en combinación se presentan en la siguiente Tabla 7.10, en resumen se parte de una consulta simple (1 código CPV y 1 código NUTS) para obtener un tiempo de referencia, a partir de esta primera referencia se aplican las distintas optimizaciones, generando los distintos tratamientos (T_i), ver Tabla 7.11, para el establecimiento de la mejor combinación posible.

ID	Descripción
F_1	Consulta simple: 1 código CPV y 1 código NUTS
F_2	F_1 con uso de la clausula `LIMIT` de SPARQL
F_3	Consulta expandida: n códigos CPV y n código NUTS
F_4	Rescritura de las consultas SPARQL: `FILTER`, etc.
F_5	Uso de grafos nombrados en la consulta SPARQL: claúsula `FROM`
F_6	Separación de las consultas en SPARQL en simples (F_1)

ID	Descripción
F_7	Consultas simples distribuidas con 5 hilos (1 por código CPV)

Tabla 7.10: Características de las consultas en SPARQL.

3. Especificación de las medidas de trabajo en cuanto a la respuesta. El valor que se toma es el tiempo de ejecución en segundos de la consulta en SPARQL.

4. Ejecución de un experimento piloto. Evidentemente esta tarea se ha realizado con una pequeña muestra de consultas, sólo un año de anuncios de licitación pero ejecutando todos los tratamientos con el fin de validar el enfoque propuesto, obtener los resultados de toma de tiempos y preparar el procesamiento del registro de tiempos para obtener las medidas finales.

5. Especificación de un modelo. En este caso, no se utilizará un modelo matemático para aproximar los valores como ya se ha explicado anteriormente debido a que el objetivo es optimizar la realización de consultas, en cuanto a tiempo de ejecución, no la predicción futura.

6. Esquematización de los pasos a seguir.

7. Determinación del tamaño muestral. Como ya se ha comentado anteriormente, en el apartado de factores cuantitativos, se utilizarán los datos generados durante el proceso de promoción de datos a *Linked Data* de los anuncios de licitación y de las clasificaciones de productos.

8. Revisión de las decisiones anteriores.

7.3.2 Ejecución del experimento sobre el Rendimiento del Sistema MOLDEAS

La aplicación del plan previsto en la anterior sección arroja un conjunto de valores para cada uno de los tratamientos (T_k) y réplicas de los tiempos de ejecución de las consultas de acuerdo a las distintas características (F_i). En la Tabla 7.12 se reflejan las medias de tiempo y ganancia agregadas de cada uno de los tratamientos realizados. Por otra parte, se presentan igualmente las medidas pormenorizadas de las observaciones realizadas y agregadas de las 3 réplicas ejecutadas, ver Tablas 7.13 y 7.14. El cálculo de la ganancia se realiza mediante la fórmula: $t_{old}/t_{new} - 1 * 100$, tomando como referencia para T_2 el tiempo de T_1 y para los tratamientos $T_{4.,10}$ el tiempo de T_3. Finalmente, la ejecución de todos los tratamientos con las distintas observaciones y réplicas ha implicado la realización de 5751 consultas en SPARQL más las de entrenamiento para cada una de las réplicas.

Adicionalmente, los resultados agregados de este experimento, respecto a los tratamientos de referencia T_1 y T_3, se visualizan gráficamente en las siguientes gráficas, ver Figuras 7.13 y 7.14 tomando como referencia el tiempo de ejecución y Figura 7.15 presentando la ganancia en porcentaje respecto a T_3.

ID Consulta	Código CPV inicial	Códigos CPV expandidos	Códigos NUTS
Q_1	15331137	48611000, 48611000, 50531510, 15871210	UK, PL, RO
Q_2	50531510	34144100, 44212211, 44212212, 50531500	ES, FR, DE
Q_3	34144100	44212211, 31140000, 31140000, 34144100	PL, CZ, RO
Q_4	64122000	64216120, 79571000, 15871210, 64121000	BE, SE, DE
Q_5	79320000	75241000, 75100000, 75000000, 60112000	UK, FR, AT
Q_6	44100000	44110000, 44170000, 44190000, UB03	NL, SE, DE
Q_7	31000000	33141000, 39000000, 44000000, 31600000	DE, IT, HU
Q_8	50000000	50512000, 50333100, 50530000, 50532300	UK, IR, FR
Q_9 (random)	15841400	15841300, 15511700, 44921210, 03131400	ES, FR, DK

Tabla 7.9: Descripción de los códigos en las consultas en SPARQL del experimento.

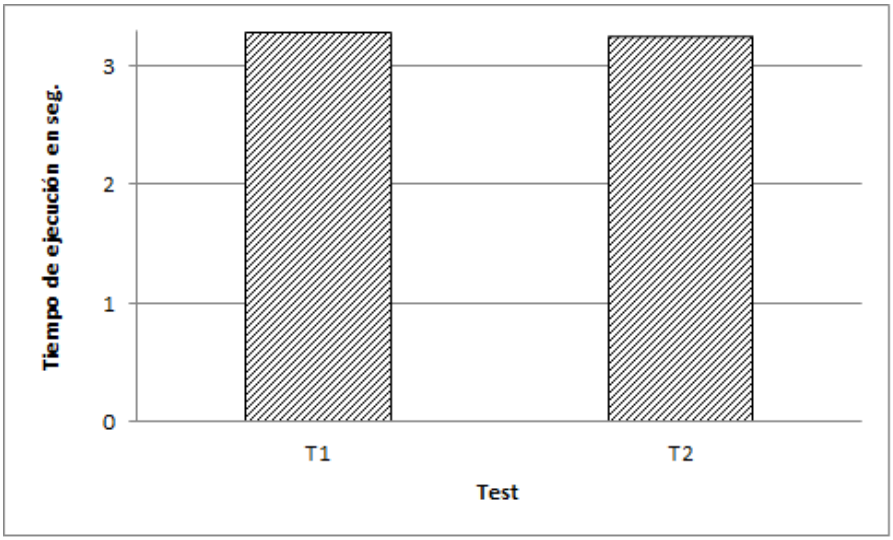


Figura 7.13: Gráfica de Tiempo de ejecución medio con referencia T_1.

Test/ Característica	F_1	F_2	F_3	F_4	F_5	F_6	F_7	**Nº consultas SPARQL**
T_1	⋆							1
T_2	⋆		⋆					1
T_3		⋆						1
T_4		⋆	⋆					1
T_5		⋆	⋆	⋆				1
T_6^1 (n códigos CPV y m códigos NUTS)		⋆	⋆	⋆	⋆	⋆		4
T_6^2 (≡)		⋆	⋆	⋆	⋆	⋆	⋆	4
T_7^1 (1 código CPV y m códigos NUTS)		⋆	⋆	⋆		⋆		5
T_7^2 (≡)		⋆	⋆	⋆		⋆	⋆	5
T_8^1 (≡)		⋆	⋆	⋆	⋆	⋆		20
T_8^2 (≡)		⋆	⋆	⋆	⋆	⋆	⋆	20
T_9^1 (1 código CPV y 1 código NUTS)		⋆	⋆	⋆		⋆		15
T_9^2 (≡)		⋆	⋆	⋆		⋆	⋆	15
T_{10}^1 (≡)		⋆	⋆	⋆	⋆	⋆		60
T_{10}^2 (≡)		⋆	⋆	⋆	⋆	⋆	⋆	60

Tabla 7.11: Descripción de cada uno de los tratamientos, características de optimización.

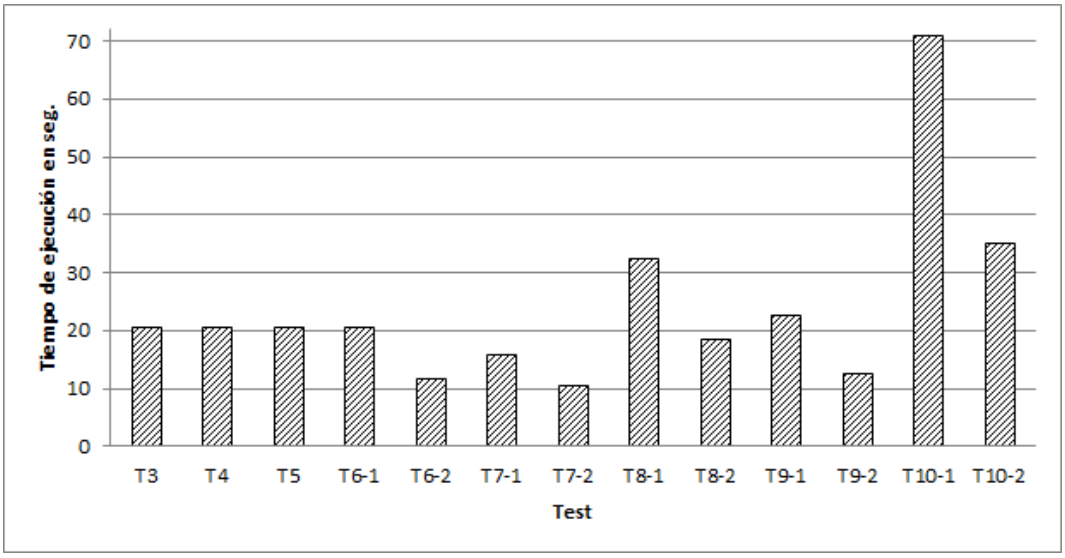


Figura 7.14: Gráfica de Tiempo de ejecución medio con referencia T_3.

Test	$\bar{X}$ **Tiempo (seg.)**	$\bar{X}$ **Ganancia (%)**
T_1	3,21	N/A
T_2	3,25	1,21
T_3	20,548	N/A
T_4	20,552	$-0,02$
T_5	20,545	$-0,01$
T_6^1	20,52	0,14
T_6^2	11,80	74,37
T_7^1	15,81	30,58
T_7^2	10,51	96,54
T_8^1	32,33	$-36,11$
T_8^2	18,45	11,21
T_9^1	22,53	$-8,77$
T_9^2	12,61	63,36
T_{10}^1	71,01	$-70,97$
T_{10}^2	35,08	$-40,42$

Tabla 7.12: Tiempo de ejecución (seg.) y ganancia (%).

ID Consulta/ Test	T_1	T_2	T_3	T_4	T_5	T_6^1	T_6^2
Q_1	3,21 (NA)	3,13 (2,46)	19,35 (NA)	19,53 ($-0,91$)	19,45 (0,49)	19,24 (0,58)	11,16 (73,49)
Q_2	3,15 (NA)	3,14 (0,52)	23,56 (NA)	23,68 ($-0,52$)	23,68 (0,52)	24,06 ($-2,10$)	13,24 (77,95)
Q_3	3,14 (NA)	3,14 (0,11)	18,56 (NA)	18,47 (0,46)	18,44 ($-0,60$)	18,69 ($-0,71$)	10,08 (84,14)
Q_4	3,16 (NA)	3,13 (0,79)	18,52 (NA)	18,32 (1,08)	18,50 ($-0,10$)	18,20 (1,77)	10,19 (81,73)
Q_5	3,29 (NA)	3,17 (3,82)	22,69 (NA)	22,85 ($-0,72$)	22,69 (0,01)	22,62 (0,31)	14,61 (55,31)
Q_6	3,21 (NA)	3,17 (1,35)	18,55 (NA)	18,45 (0,51)	18,48 ($-0,35$)	18,22 (1,78)	10,00 (85,46)
Q_7	3,39 (NA)	3,39 (0,25)	19,01 (NA)	19,19 ($-0,93$)	19,12 (0,55)	18,99 (0,09)	11,08 (71,56)
Q_8	3,94 (NA)	3,98 ($-0,95$)	23,44 (NA)	23,15 (1,26)	23,24 ($-0,82$)	23,50 ($-0,29$)	14,72 (59,24)
Q_9	3,17 (NA)	3,09 (2,59)	22,23 (NA)	22,32 ($-0,40$)	22,27 (0,19)	22,26 ($-0,15$)	12,32 (80,45)

Tabla 7.13: Tiempo de ejecución (seg.) y ganancia (%). Parte 1.

ID Consulta/ Test	T_7^1	T_7^2	T_8^1	T_8^2	T_9^1	T_9^2	T_{10}^1	T_{10}^2
Q_1	15,71 (23,18)	10,62 (82,25)	33,27 (−41,84)	18,75 (3,24)	21,58 (−10,33)	12,38 (56,28)	69,13 (−72,00)	33,41 (−42,07)
Q_2	15,64 (50,57)	9,90 (137,84)	32,26 (−26,98)	17,96 (25,66)	26,27 (−10,34)	15,50 (51,97)	73,62 (−68,00)	37,50 (−29,49)
Q_3	15,54 (19,44)	9,66 (92,13)	32,04 (−42,09)	17,94 (3,29)	20,32 (−8,68)	10,44 (77,68)	68,95 (−73,09)	32,95 (−50,52)
Q_4	15,87 (16,68)	9,99 (85,45)	32,07 (−42,24)	18,43 (3,24)	20,48 (−9,54)	10,33 (79,23)	68,85 (−73,10)	34,27 (−43,79)
Q_5	16,18 (40,22)	11,92 (90,39)	32,46 (−30,10)	19,02 (23,12)	24,28 (−6,57)	14,40 (57,55)	74,02 (−69,35)	37,08 (−33,79)
Q_6	15,76 (17,73)	9,60 (93,21)	32,11 (−42,23)	18,24 (−2,48)	20,51 (−9,55)	10,77 (72,23)	72,28 (−74,34)	32,80 (−49,98)
Q_7	15,93 (19,31)	11,23 (69,35)	32,21 (−40,99)	18,51 (4,23)	21,12 (−10,01)	11,27 (68,67)	69,02 (−72,46)	33,55 (−42,04)
Q_8	16,12 (45,43)	12,11 (93,54)	32,55 (−28,01)	19,50 (26,61)	25,33 (−7,48)	16,16 (45,05)	72,72 (−67,77)	38,27 (−30,16)
Q_9	15,58 (42,66)	9,89 (124,71)	31,99 (−30,51)	17,74 (14,01)	23,76 (−6,46)	13,76 (61,57)	70,76 (−68,59)	36,40 (−41,92)

Tabla 7.14: Tiempo de ejecución (seg.) y ganancia (%). Parte 2.

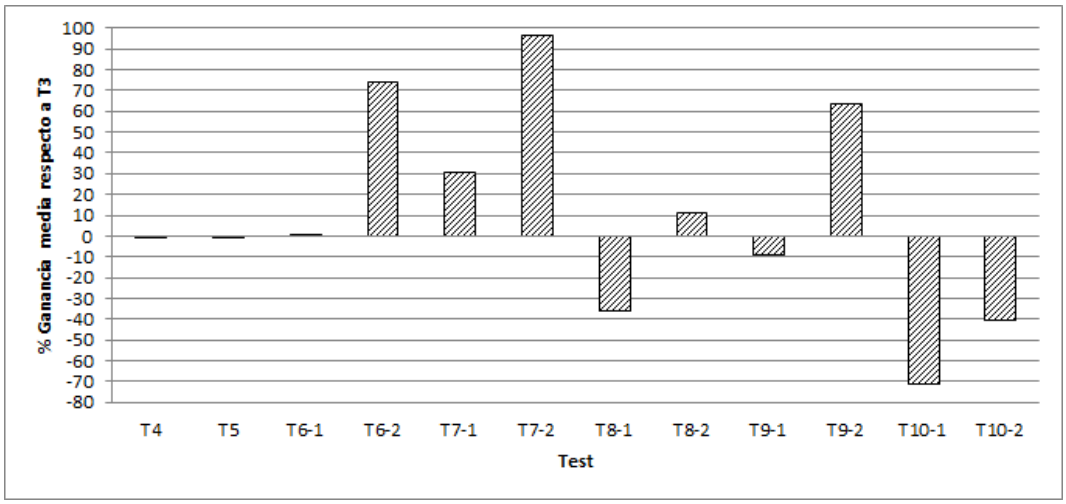


Figura 7.15: Gráfica de Ganancia media con referencia T_3 en (%).

7.3.3 Validación del experimento sobre el Rendimiento del Sistema MOLDEAS

La realización del experimento de rendimiento de las consultas en SPARQL es motivado por el exceso de latencia en las consultas al repositorio RDF. Como ya se ha mencionado no es objetivo la comparación de distintos repositorios RDF para comprobar cuál obtiene un mejor rendimiento para las consultas realizadas sino probar cuáles son las optimizaciones en las consultas en SPARQL y en el código fuente del demostrador que permiten acelerar el tiempo de la consulta independientemente del proveedor del servicio. Por ello, las características a combinar F_i para obtener un mejor tiempo de respuesta permiten establecer la mejor conjunción de las mismas de acuerdo al estudio de los tiempos de ejecución obtenidos.

De acuerdo a los resultados obtenidos se establecen los siguientes puntos clave y conclusiones:

- Los tipos de consulta generados a partir de F_1 y F_3 permiten establecer un valor de referencia en tiempo de ejecución para su posterior comparación con las mejoras introducidas. De esta manera se establece un primer valor de referencia para la comparación de consultas simples (1 código CPV y 1 código NUTS) entorno a los 3 segundos y otro para consultas expandidas, n códigos CPV y n código NUTS, en torno a los 20 segundos. Estos valores son un por una parte motivadores del experimento de mejora y por otra permiten evidenciar la ganancia con la introducción de nuevas características
- Existen algunas de las características, F_2 y F_4, que no implican mejora en el tiempo de ejecución.
- En el caso particular de F_3, uso de la clausula LIMIT, su valor se fija en $10,000$ y carece de representatividad debido a que los resultados de la consulta ya son previamente filtrados y no superan esta cifra.
- La reescritura de consultas, F_4, habitualmente implica mejoras en el tiempo de ejecución pero no se ha apreciado una mejora crítica debido a que este tipo de optimizaciones obtienen una mejora crítica cuando las consultas en SPARQL realizan operaciones con valores opcionales, clausula OPTIONAL, por ejemplo cuando se trasladan consultas del tipo NATURAL JOIN e INNER JOIN en SQL a SPARQL, en el caso de estudio estos valores no se han utilizado ya que las consultas generadas por el sistema no incluyen encaja de patrones mediante clausulas opcionales.
- En cuanto a la característica F_5 sobre el uso de grafos nombrados se obtiene una mejora relevante ya que la selección de los grafos sobre los cuales se realizan las consultas permite acotar el número de tripletas, no es lo mismo realizar sobre el grafo RDF completo presente en el *endpoint* que sobre un subconjunto del mismo. Sin embargo, hay que destacar que la ejecución de múltiples consultas (separando por grafo) provoca que el tiempo de ejecución sea incrementado.
- Sin duda la mejora más intensa en cuanto a tiempo de ejecución se produce en los casos F_6 y F_7 como se puede ver en los tratamiento T_6^2, T_7^1, T_7^2 y T_9^2 ya que el tiempo de ejecución de consultas simples es relativamente bajo y, evidentemente, la distribución de consultas permite la obtención de un mejor tiempo de ejecución global. No obstante, en los tratamientos T_{10}^1 and T_{10}^2 esta supuesta mejora se convierte en un inconveniente ya que el número de consultas generadas es muy alto y la distribución mediante 5 hilos no permite realizar una mejora.
- Teniendo en cuenta los puntos anteriores cabe destacar que el tipo de códigos CPV no tiene ninguna implicación para la ejecución de la consulta. Sin embargo, el número de códigos CPV si implica un incremento en el tiempo de ejecución, del orden de 3 segundos adicionales por cada código CPV nuevo, en cambio el uso de varios códigos NUTS no supone un incremento en el tiempo de ejecución y en el experimento la ejecución de un código CPV con varios NUTS es similar a la combinación $n - m$ de los mismos.

A la vista de las conclusiones extraídas y de los resultados obtenidos se comprueba que la mejora combinación de características se genera en el tratamiento T_7^2 con un tiempo de ejecución de 10,55 segundos y una ganancia respecto a T_3 del 95,73 %, la descripción de esta mejora implica que se trata de una consulta expandida (F_3) en la cual se utilizan clausulas LIMIT (F_2), se reescriben las consultas mediante FILTERs (F_4), generando consultas simples en SPARQL (F_6) de 1 código CPV y m códigos NUTS y, finalmente, se distribuyen con un hilo por código CPV. Por otra parte, el peor resultado se genera en el tratamiento T_{10}^1 con un tiempo de ejecución de 71,63 segundos y una pérdida en ganancia del $-71{,}17\,\%$, esta situación se debe a la extrema generación de consultas simples que si bien son rápidas en tiempo de ejecución por separado, el gran número de las mismas generadas implica que la supuesta ganancia se convierta en pérdida.

7.3.4 Evaluación del experimento sobre el Rendimiento del Sistema MOLDEAS

La ejecución y validación del experimento mediante la extracción de una serie de conclusiones permiten dar respuesta las preguntas establecidas en el diseño del experimento.

- ¿Cuáles son las mejoras que se pueden aplicar sobre una consulta en SPARQL para mejorar el tiempo de ejecución?

  Existen diversas mejoras aplicables a las consultas en SPARQL dependiendo del tipo de encaje de tripletas que se realice como son el uso de la clausula LIMIT, FILTER o los grafos nombrados, siendo independientes del proveedor del *endpoint*. No obstante, estas mejoras no son siempre son relevantes para todos los casos de estudio y en el caso particular estudiado su relevancia no es tan crítica como la separación de la consulta expandida en simples y su distribución.

- ¿Cuál es la combinación de mejoras que obtiene un mejor tiempo de respuesta?

  En el caso particular estudiado las características del tratamiento T_7^2 generan el mejor tiempo de ejecución utilizando consultas simples, con uso de clausulas LIMIT y FILTER en SPARQL y distribuyendo la ejecución de las mismas.

- ¿Cuál es el coste de la combinación de estas mejoras?

  La generación de consultas a partir de una consulta expandida es un proceso trivial que no genera sobrecarga en tiempo de ejecución en el código. El esfuerzo recae sobre la programación distribuida con hilos pero teniendo en cuenta las capacidades que suministran los lenguajes de programación actuales (Java) este tipo de práctica queda minimizada. Por tanto, el coste de implantación de esta solución es bajo.

- ¿Existe algún elemento externo de configuración que implique un incremento en el tiempo de ejecución de las consultas?

  El incremento del tiempo de ejecución se puede abordar desde dos punto de vista: 1) el nuevo código fuente implica una lógica con una alta complejidad temporal, como se ha señalado la generación de consultas simples no resulta crítico y la programación distribuida, con su intrínseca dificultad, queda minimizada debido a las facilidades de los lenguajes de programación, en este caso Java y 2) el cambio del número de peticiones y consultas implica una sobrecarga en la red, conllevando un exceso en el tiempo de ejecución. En este caso, esta situación es minimizada tratándose de un entorno local y en el caso de un entorno global la situación tampoco se considera crítica debido a que las consultas están limitadas en tamaño y se facilitaría algún sistema de caché de consultas y resultados [47].

- ¿Cómo afectan los resultados en la implementación actual del sistema MOLDEAS?

  La realización de este experimento ya ha generado una refactorización en el código del componente *moldeas-api* para atacar la división de consultas expandidas en simples y la ejecución

paralela de consultas, por lo que realmente este experimento ha servido para mejorar el tiempo de ejecución del sistema y el código fuente del mismo. Sobre la configuración necesaria para adoptar esta solución, el uso del *framework* Spring facilita los cambios en las implementaciones de un interfaz, por lo que, la dificultad para integrar nuevas implementaciones está minimizado con el uso de este tipo de tecnología.

Finalmente y como síntesis de esta sección cabe destacar que la ejecución de consultas sobre conjuntos de datos de cierta envergadura conlleva la aparición de problemas de escalabilidad y rendimiento que están siendo abordados en la comunidad de *Linked Data* mediante la aplicación de técnicas distribuidas, igualmente es conveniente la realización de un esfuerzo para que las consultas en SPARQL no supongan un cuello de botella en las aplicaciones que hagan uso de tecnologías semánticas, para ello los estudios [260] sobre el rendimiento [32, 269] de SPARQL, su optimización, etc., son claves. Actualmente, en la última especificación de SPARQL 1.1 y el emergente uso de consultas federadas [54] requieren especial atención para el correcto despliegue de aplicaciones semánticas nativas.

Capítulo 8

Conclusiones y Trabajo Futuro

> Todo lo que tiene un comienzo,
> tiene un final...
>
> *Citas Célebres*
> MATRIX REVOLUTIONS

8.1 Cumplimiento de Objetivos

En primer lugar para realizar la validación del estudio realizado en este documento cabe valorar el grado de consecución de los objetivos fijados al inicio de este trabajo y que se han señalado en la Sección 1.3:

Panorámica de la contratación pública electrónica. En concreto el objetivo número 1 se definía como una tarea en la que se debía "Estudiar, analizar y valorar las capacidades actuales del dominio de la contratación pública", en el Capítulo 2 se ha realizado una síntesis de las actuales tendencias en contratación pública electrónica, específicamente en el ámbito de la Unión Europea, realizando una valoración de las acciones estratégicas y proyectos que se están llevando a cabo a nivel europeo y que en muchos casos se ven reflejadas en las acciones nacionales y regionales. La importancia de la contratación pública electrónica como proceso crítico dentro de las Administraciones Públicas ha quedado patente debido a las posibilidades de generación de servicios de negocio y las cantidades presupuestarias manejadas para la realización de servicios y obras a la ciudadanía y en consecuencia para el impulso del mercado. Por otra parte, es importante prestar atención a la problemática que surge en el momento de facilitar el acceso a la información de las licitaciones a nivel europeo como las características multingües, la dispersión de información, la multiplicación de fuentes de información heterogéneas, etc., que provocan un descenso en la competitividad ya que los proveedores de la Administración, dependiendo de la situación, son reducidos a un ámbito concreto. Con todo ello, se pone de manifiesto la necesidad de suministrar un entorno ágil y flexible para la gestión de la información y datos con los objetivos tecnológicos de impulsar la interoperabilidad e integración de servicios dentro de la propia Administración y por otra parte, mejorar la imagen corporativa de la misma en cuanto a transparencia y servicios a los ciudadanos y empresas. El repaso realizado en el capítulo mencionado consolida tanto los problemas como las soluciones y estrategias actuales proporcionando un dominio ampliamente motivado para la investigación y la innovación.

Panorámica de *Open Data*, *Linked Data* y Web Semántica. Uno de los puntos clave de estudio e innovación en este trabajo consiste en la aplicación de los principios promulgados por la iniciativa de Web Semántica y en concreto de *Linked Data* al dominio de la contratación pública electrónica. Además, teniendo en cuenta la tendencia actual de apertura de datos impulsada por el movimiento *Open Data* se ha realizado un conveniente repaso de la tecnología, iniciativas, proyectos, etc., relacionados con estos movimientos, a la vez que contrastado los beneficios, ventajas, etc., con el objetivo de adquirir una perspectiva realista de las aspiraciones de estas iniciativas y su impacto en los servicios actuales así como la trascendencia de su aplicación a un sector como la contratación pública electrónica. Esta panorámica permite obtener una visión general de estos enfoques y su posible potencial para la resolución de problemas de gestión e integración de información y datos en un espectro temporal desde la actualidad hasta medio/largo plazo.

Ciclo de vida para *Linked Data*. Atendiendo al objetivo definido como "Definir los métodos basados en semántica para producir, publicar, consumir y validar la información de los anuncios de licitación siguiendo las directrices de *Open Data* y *Linked Data*" y tras la revisión de los trabajos en estas áreas de conocimiento, se ha realizado una descripción de un de ciclo de vida de datos enlazados que se alinea con los enfoques existentes definiendo un ciclo de vida basado en la abstracción en 3 niveles de los procesos, métodos y tareas a realizar para la consecución cuantificable de la promoción de datos siguiendo las iniciativas de *Linked Data* y *Open Data*.

Sistema MOLDEAS. La denominación de MOLDEAS engloba tanto el ciclo de vida como las herramientas que se han desarrollado para dar soporte a los procesos contemplados en el ciclo de vida de datos enlazados. De esta manera se ha implementado un conjunto de componentes como activos experimentales para el cumplimiento del ciclo de vida y la demostración del consumo de datos enlazados. El sistema MOLDEAS se convierte de esta manera en un primer y gran esfuerzo para la provisión de herramientas de carácter genérico en el ámbito de *Linked Data*, aunando tecnología ya existente y mejorando las mismas como por ejemplo ONTOSPREAD, y particularizando su aplicación al dominio de la contratación pública electrónica. Con todo ello, se completa el capítulo propio de ingeniería de este estudio y se da respuesta a los objetivos marcado de "Definir los algoritmos y procesos para dar soporte a la aplicación de los métodos basados en semántica a la información de los anuncios de licitación" e "Implementar y reutilizar los componentes software necesarios para dar soporte a los métodos semánticos".

Aplicación de MOLDEAS a *e-Procurement*. Una vez que se disponen de los datos necesarios sobre anuncios de licitación en general y los relativos a las clasificaciones de productos y organizaciones la aplicación de MOLDEAS se ha reflejado para la promoción de datos con un enfoque cuantitativo y cualitativo, permitiendo contrastar los procesos definidos en el ciclo de vida. De esta manera se da cumplimiento al gran objetivo de "Aplicar los métodos semánticos definidos al contexto de los anuncios de licitación pública".

Experimentación y Validación. Los experimentos realizados se concretan en 4 acciones diferentes:

- Elaboración y diseño de tablas de validación de *Linked Data* y *Open Data* de acuerdo a la documentación actual y la propia experiencia del autor y supervisor.
- Contraste de la elaboración de un demostrador público con capacidad para el consumo de *Linked Data* con el objetivo de verificar que la información y datos pueden ser consumidos por nuevos servicios manteniendo y mejorando el comportamiento actual y sentando la base para un estudio profundo de un posible sistema experto basado en tecnología semántica para el dominio de la contratación pública electrónica.
- Demostración del aumento de la expresividad mediante el uso de *Linked Data* en sistemas de recuperación de información basados en vocabularios controlados.

- Mejora del rendimiento de consultas en SPARQL mediante la aplicación de determinadas características obtenidas tras la valoración de la documentación existente y la experimentación empírica.

Los resultados de los experimentos han sido validados y evaluados con el objetivo de suministrar una demostración científica al estudio realizado en este documento y cumplir con la tarea de "Establecer un conjunto de prueba y validación de los componentes implementados para verificar la corrección, validez y rendimiento de los métodos propuestos".

Impacto y difusión. En cualquier proyecto o trabajo de investigación es absolutamente necesario dar difusión y crear concienciación de los avances realizados tanto a la comunidad científica como industrial. Por ello, esta tarea transversal se ha cubierto satisfactoriamente mediante la realización de distintas publicaciones, realizando la componente de investigación en un proyecto como "10ders Information Services" en colaboración con empresas de servicios y tecnología y participando activamente en la comunidad investigadora mediante el establecimiento de colaboraciones con personas e instituciones de carácter internacional que ha conllevado la elaboración conjunta de publicaciones y la pertenencia a comités de programas y revisión de artículos en diferentes revistas, conferencias y talleres. La justificación pormenorizada de estas actividades se ha señalado en los Apéndices A y B cumpliendo así con el objetivo "Difundir, formar y transferir la tecnología y conocimiento generado tanto a las comunidades científicas como industriales".

Transversalmente a estos objetivos específicos todo el trabajo realizado en las definiciones teóricas, desarrollo de activos experimentales y diseño de experimentos, se ha enmarcado bajo una apuesta fiel y constante por la aplicación y uso de los estándares dando así acogida a la meta de "Promover el uso de estándares y la reutilización de información y modelos de conocimiento compartido".

8.2 Principales Aportaciones

8.2.1 Aportaciones Científicas

El estudio desarrollado en este trabajo sobre la aplicación los principios de *Open Data* y las tecnologías semánticas, concretamente mediante la aplicación de la iniciativa *Linked Data* al dominio de la contratación pública electrónica, ha servido para la obtención de las siguientes aportaciones desde el punto de vista científico:

Repaso del estado actual de *e-Procurement* y aplicación de tecnologías semánticas. En múltiples ocasiones la selección de una tecnología para ser aplicada sobre un dominio es un decisión crítica ya que en un entorno de producción no es posible en muchas ocasiones dedicar esfuerzo y tiempo a comprobar iniciativas basadas en la investigación e innovación. Por ello, el estudio realizado conlleva un valor adicional desde un punto de vista estratégico para el campo de la contratación pública electrónica al haber conjugado el repaso de este dominio y enlazarlo con el trabajo relacionado bajo el paradigma de la Web Semántica y *Linked Data*.

Ciclo de vida para *Linked Data*. Actualmente con el éxito y las expectativas generadas en torno a los datos enlazados abiertos, se ha puesto de manifiesto la necesidad de gestionar el ciclo de vida de los datos enlazados y no limitarse únicamente a la exposición de las bases de datos existentes cumplimiento unas ciertas directrices. Esta situación queda reflejada en las múltiples guías que aparecen a lo largo de la geografía de cómo abrir y enlazar datos, en algunos casos utilizando un enfoque *top-down*, como en el Reino Unido, y en otros casos como España aplicando un sistema *bottom-up*. El objetivo final es disponer de una serie de guías y buenas prácticas para

minimizar el esfuerzo en la apertura y enlazado de datos. De la misma, en las distintas acciones de investigación como proyectos, actividad en universidades y centros tecnológicos, etc., se está suministrando un gran cantidad de información sobre cómo abordar esta situación. Sin embargo, la definición de un claro ciclo de vida basada en la abstracción de procesos, métodos y tareas no se ha realizado de forma completa y los enfoques existentes si bien se postulan como recetas o guías prácticas todavía no han alcanzado un nivel de normalización a través de alguna institución como el W3C. Con la meta fijada en gestionar el ciclo de vida de datos enlazados, reaprovechar el conocimiento generado en las propuestas anteriores y la experiencia particular en el desarrollo de estas actividades se ha definido un ciclo de vida que se ha aplicado a la información y datos concernientes a los anuncios de licitación públicos.

Sistema MOLDEAS. La denominación bajo este nombre cubre el desarrollo de activos experimentales y la aplicación de tecnología existente, con mejoras en la misma, para la realización de los procesos del ciclo de vida de datos enlazados en el dominio de los anuncios de contratación pública. De esta manera se han definido modelos e implementado algoritmos con las siguientes capacidades:

- Modelo de información y datos para los anuncios de licitación pública, el catálogo de clasificaciones de productos y las organizaciones basado en estándares y vocabularios comunes.
- Producción de datos enlazados de acuerdo a los modelos definidos.
- Reconciliación de entidades específica para la información y datos anteriores.
- Establecimiento de un modelo de publicación de datos enlazados abiertos.
- Implementación de un conjunto de componentes para el consumo y explotación de datos desde una aplicación externa aplicando tecnología existente y mejorando la propia como ONTOSPREAD, propagación sobre instancias no sólo sobre clases.
- Diseño de criterios de validación de los datos generados y evaluación semi-automática de los mismos.
- Demostración de la mejora cuantitativa y cualitativa en el acceso a la información y datos concernientes a los anuncios de licitación.

Generación de *know-how*. La realización de este trabajo ha permitido la generación de nuevo conocimiento sobre un campo tan relevante como la administración electrónica y en concreto sobre el proceso de contratación pública. De esta forma, es posible abordar problemas similares con el enfoque de MOLDEAS ya que si bien su escenario de aplicación es la contratación se puede extrapolar y aplicar a otros dominios. El modelo de trabajo, investigación e innovación realizado ha permitido la creación de soluciones genéricas que se han demostrado efectivas en un dominio concreto pero sin menoscabar su relevancia y aplicación a otros sectores, esta flexibilidad permite ofrecer información detalle para futuros esfuerzos en otros campos.

8.2.2 Aportaciones Tecnológicas

Desde un punto de vista tecnológico la realización de este trabajo se ha manifestado en la utilización de múltiples herramientas provenientes de distintos proveedores para la implementación de diversas tareas. En concreto, las aportaciones tecnológicas en este ámbito se sintetizan en las siguientes acciones:

- Adaptación de la entrada de biblioteca Apache Mahout para el consumo de datos enlazados.
- Adición de nuevas capacidades a la biblioteca ONTOSPREAD.

- Realización de algoritmos de reconciliación de entidades basados en Apache Lucene y Solr.
- Creación de un API para el consumo y explotación de datos enlazados mediante la aplicación de diferentes métodos de expansión de consultas.
- Creación de un sistema de test y validación de datos enlazados semi-automático.

8.3 Conclusiones Científicas

El estudio de los datos enlazados abiertos en el campo de las licitaciones públicas ha puesto de manifiesto las necesidades de este entorno y la necesidad de realizar una investigación profunda para impulsar la contratación pública electrónica dada su relevancia e impacto en la sociedad. Por ello, el enfoque basado en tecnologías semánticas y datos enlazados abiertos permite la culminación práctica de ciertas necesidades como la mejora de acceso y la reutilización de la información. Sin embargo, la conversión de un entorno tan amplio debe tomarse como una actividad iterativa en la cual con la aplicación de buenas prácticas se consiga dar respuesta y solución a los inconvenientes que presenta. Una vez que la problemática de la contratación pública electrónica está motivada convenientemente para que las tecnologías basadas en semántica puedan influir y mejorar este entorno, surgen nuevas situaciones que se han reflejado a lo largo del trabajo y que a continuación se detallan.

- Todos los procesos implicados en el tratamiento de datos enlazados deben asegurarse suministrando mecanismos para su validación con el objetivo de facilitar su reutilización posterior en condiciones de ausencia de incertidumbre.
- El uso de estándares es clave para minimizar los problemas de integración e interoperabilidad.
- El tratamiento de grandes cantidades de datos conlleva prestar especial atención a los algoritmos diseñados y a la construcción de consultas sobre los mismos.
- La necesidad de modelar la información y datos para su posterior reutilización en un dominio extenso debe realizarse con el suficiente grado de especificidad pero teniendo presenta la posibilidad de extensibilidad, por lo que es conveniente no realizar grandes modelos poco usables y que tan sólo el autor o autores pueden manejar. En este sentido la estrategia a seguir debe basarse en la generación de un marco de trabajo común y la creación sostenible y escalable de conocimiento con el objetivo de impulsar su reutilización.
- Los principios de *Linked Data* y *Open Data* verdaderamente ayudan a favorecer la reutilización de información si siguen unas directrices de producción, publicación y consumo adecuadas. Sin duda, la realización práctica de la Web Semántica se ve reflejada bajo la iniciativa de *Linked Data* y su éxito actual, queda patente en general y en el caso objeto de estudio de este documento.
- El maremágnum de tecnología, enfoques, algoritmos, vocabularios, conjuntos de datos, etc., dentro de la iniciativa de *Linked Data* convierte la toma de ciertas decisiones y la adopción de soluciones en un trabajo tedioso en el cual la experiencia cobra una especial relevancia. La necesidad de catalogación y valoración de todos estos trabajos y esfuerzos se hace cada vez más patente.
- Existen tareas como la reconciliación de entidades que todavía se hayan en etapa temprana de desarrollo y en las cuales muchas aportaciones son posibles ya que en la actualidad se basan principalmente en técnicas de procesamiento de lenguaje natural.
- El uso de datos enlazados permite mejorar el acceso a la información facilitando una mayor expresividad en la realización de consultas sobre grandes conjuntos de datos.

- El rendimiento, la completitud, el tiempo de ejecución finito y la consulta de grandes bases de datos en un entorno distribuido es una línea de investigación especialmente relevante y abierta en el campo de la Web de Datos.
- Desde un punto de vista de la gestión de la información y datos la Web Semántica se ha convertido en un pieza clave pero la creación de algoritmos explotando las capacidades de un entorno mejor informado todavía no se ha impulsado convenientemente y se reduce a la adaptación a algoritmos y técnicas perfectamente probadas como la minería de datos o el procesamiento de lenguaje natural. La consecuencia principal de esta situación recae en que el esfuerzo realizado para la mejora de la información y datos finalmente no se ve recompensada por algoritmos más avanzados.
- La explotación de la información mediante técnicas de razonamiento no siempre es necesaria y dependiendo del dominio, como en el caso de la publicación de información y datos de los anuncios de licitación, es conveniente centrar los esfuerzos en el modelo de datos para su posterior reutilización que en construir un base lógica muy sólida.
- Los procesos administrativos dentro de la administración electrónica deben necesariamente aprovechar las ventajas de las tecnologías de información para ser más eficientes, la tendencia en los últimos años y la planificación estratégica para los próximos años así lo corrobora, más aún teniendo en cuenta el carácter global de las comunicaciones e intercambio de información.
- Las primeras versiones de datos enlazados abiertos liberados desde las distintas organizaciones, entidades, etc., tendían a basar su éxito en el número de tripletas RDF, generando en muchos casos datos superfluos. Una vez superada esta fiebre sobre la cantidad de tripletas RDF el esfuerzo se centra en la publicación de los datos necesarios y suficientes bajo condiciones de calidad que aseguren su reutilización.

Estos puntos de clave extraídos del trabajo realizado deben unirse necesariamente a los generados tras la realización y evaluación de los experimentos realizados que a continuación se listan:

- El uso de datos enlazados permite aumentar la expresividad de las consultas facilitando el acceso a la información desde un punto de vista cuantitativo y cualitativo. Esta situación se refleja tanto en entornos utilizando vocabularios controlados, como la contratación pública, como en entornos abiertos, la propia Web de Datos.
- El beneficio real de los datos enlazados abiertos reside en la representación de información y datos mediante un modelo estándar (RDF) que permite el uso de lenguajes (SPARQL) y protocolos (HTTP) estándar para su acceso facilitando la recuperación de información.
- El enlazado de datos consta de una variable de incertidumbre que si bien es aceptada por la comunidad, debe ser tenida en cuenta para la realización de aplicaciones de carácter crítico.
- En la actualidad tanto la información como los datos disponibles son abiertos en su mayor parte e incluso son utilizados para la construcción de servicios comerciales pero el esfuerzo requerido para llevar a cabo estos productos se ve reflejado en el precio de los mismos por lo que el uso de datos enlazados simplifica y favorece la reutilización de información.
- En muchos casos no se necesario llegar a un modelo de 5 $\star$ ya que esto provoca la necesidad de adaptación de los desarrolladores a estas prácticas, por lo que la tendencia actual es por un parte generar los datos con el mayor nivel posible de estrellas pero facilitando la labor para su consumo, por ejemplo utilizando como formato de representación JSON.
- La aplicación de patrones de diseño en la planificación de los datos a transformar permite dar una respuesta homogénea a problemas recurrentes de representación de información y datos.

- La creación de un sistema experto para la recuperación de información de los anuncios de licitación es un dominio en el cual se puede profundizar variando una gran cantidad de parámetros tanto de configuración de los algoritmos como valorando la implicación de introducir nuevas variables de información.
- La división de consultas en SPARQL sobre grandes conjuntos de datos en otras más pequeñas y su distribución permite la mejora del rendimiento ya que se simplifican dos factores importantes: 1) el número de tripletas encajadas y 2) el tamaño del conjunto de datos sobre los que la consulta es ejecutada.
- El trabajo desarrollado en la nueva especificación de SPARQL da respuesta a algunos de los problemas o ausencias de funcionalidad encontradas en la especificación inicial como es el uso de agregados, *subqueries* o federación de consultas. Claramente la nueva especificación se acerca a lenguajes de consulta ya maduros como SQL, este esfuerzo sin duda es clave para la adopción de datos enlazados.
- Existen características en SPARQL deseables como la priorización del encaje de tripletas para la ordenación de resultados que todavía no se han incluido en las especificaciones pero si están presentes en productos como Virtuoso y que resultan de gran valor para la construcción de aplicaciones.

La extracción de estas conclusiones desde un punto de vista científica deja patente la conveniencia de aplicación de los datos enlazados abiertos al dominio de la contratación pública electrónica así como deja abiertas diversas líneas de investigación tanto para la propia administración electrónica como para la investigación a realizar en este campo.

8.4 Conclusiones Tecnológicas

El desarrollo de activos experimentales para dar soporte a los procesos indicados en el ciclo de vida de datos enlazados ha puesto de manifiesto tanto ventajas como inconvenientes en la tecnología actual para afrontar estas tareas, entre las cuales se destacan:

- De la misma forma que el número de vocabularios, catálogos de datos, etc., constituyen una fuente de reutilización relevante pero cuya clasificación no permite a usuarios no técnicos la selección de los mismos, también queda patente en la tecnología asociada a los datos enlazados, ya que la diversidad de herramientas constituye un amplio abanico de oportunidades en los cuales la incertidumbre sobre cuál usar conlleva simplemente a indagar las aplicadas en los proyectos de éxito sin tener en cuenta más factores. Esto implica un cierto grado de monopolio ya que la selección de tecnología se convierte en una simple consulta sin realizar una verdadera evaluación pero por otro lado la aplicación de este tipo de tecnología y herramientas justifica cualquier juicio posterior por terceros.
- Durante mucho tiempo el uso de tecnología relacionada con semántica se convertía en un proceso de depuración, esta cuestión ha sido parcialmente solucionada debida a la inversión realizada a través de proyectos de investigación y a la participación de grandes proveedores de *software* en su desarrollo.
- La proliferación de múltiples utilidades para la realización de diversas tareas conlleva una falta de integración de las mismas por lo que la intervención manual es inevitables. Plataformas integradas como la propuesta en el proyecto LOD2 o bien el *Linked Media Framework* permiten mejorar la adopción de la tecnología.

- Las actuales herramientas para la gestión y explotación de información no están preparadas para trabajar intrínsecamente con datos enlazados o con semántica, por lo que el esfuerzo para el consumo de datos enlazados y la creación de nuevos servicios y aplicaciones reutilizando soluciones existentes conlleva un esfuerzo adicional en el cual, en la mayoría de los casos, se pierde parcialmente las ventajas de reutilizar información y datos provenientes de modelos compartidos.

8.5 Futuras Líneas de Investigación y Trabajo

Las posibilidades de abrir nuevas líneas de investigación que tomen como semilla el trabajo realizado en este estudio son múltiples y se pueden abordar desde distintos puntos de vista:

- Las concernientes a la Web Semántica y *Linked Data* en general.
- Las referentes a la administración electrónica y en particular al proceso administración de contratación pública electrónica.

8.5.1 Visión Científica

La posibilidad de acciones en una línea de investigación son múltiples, la selección de las más interesantes no es una tarea sencilla y puede estar condicionada por diferentes factores: dificultad, ventaja estratégica o disponer de un escenario de aplicación correcto. Entre las más relevantes surgidas tras la realización de esta trabajo se encuentran:

- Realización de un sistema de catalogación de vocabulario y conjuntos de datos basado en diferentes métricas que permitan tanto a un usuario de dominio como desarrollador seleccionar áquel más conveniente a sus necesidades.
- Mejora de los algoritmos de reconciliación de entidades para llevar a efecto uno de los puntos clave de *Linked Data* como es la identificación de recursos iguales tanto para la realización de consultas como para su procesamiento mediante procesos de razonamiento.
- Establecimiento de un conjunto de métricas que permitan establecer y asegurar la calidad de los datos.
- Mejora del rendimiento de las consultas federadas sobre grandes bases de datos en continuo crecimiento y evolución. Creación de un benchmark basado en el conjunto de datos concerniente a los anuncios de licitación.
- Adaptación y mejora de las técnicas y algoritmos basados en semántica a entornos con características de tiempo real.
- Estudio de la aplicación de la investigación e innovación a otras etapas de la contratación pública electrónica, como la decisión de ofertas, y en general a la administración electrónica.
- Estudio de las posibilidades de realización de un sistema experto específico para la recuperación de información de anuncios de licitación teniendo en cuenta la mayor cantidad de descriptores posibles.
- Estudio de los operadores de agregación que permiten establecer un orden en la recuperación de documentos.

- Mejora de la capitalización del conocimiento experto en el dominio de contratación pública electrónica.
- Estudio de las posibilidades de creación de un sistema predictivo de anuncios de licitación de acuerdo a la información histórica de los anuncios de licitación.
- Investigación en aspectos relativos al consumo de datos como base para el impulso de los datos enlazados:
  - Mejora de la escalabilidad tanto para el indexado como para la extracción de datos a gran escala.
  - Procesamiento de consultas federadas de forma eficiente.
  - Búsqueda sobre fuentes de datos heterogéneas.
  - Descubrimiento automático de *datasets*.
  - Gestión de *datasets* dinámicos como los provenientes de sensores, sistemas reactivos, etc.
  - Calidad de los datos: procedencia, valores, etc.
  - Usabilidad en la interacción con datos enlazados.

8.5.2 Visión Tecnológica

Desde un punto de vista tecnológico, las mejoras y líneas de actuación futuras se centran en dos grandes grupos: 1) las referentes al sistema MOLDEAS como demostrador público y 2) las concernientes a tecnología de base para la Web Semántica y *Linked Data*. Es conveniente, por tanto, destacar las siguientes:

- Mejora del sistema de consumo de datos enlazados de MOLDEAS, personalización y prueba intensiva de los algoritmos disponibles.
- Mejora del sistema de visualización y consumo de datos enlazados desde el punto de vista del usuario final.
- Continuación del desarrollo del sistema de validación de datos enlazados. En este sentido se ha planificado la realización de acciones determinadas como un proyecto fin de carrera para impulsar su desarrollo a corto/medio plazo.
- Contribuir con nuevas herramientas a la comunidad de *Linked Data* para la mejora del consumo de datos enlazados desde lenguajes de programación y reutilización sencilla de información y datos. Por ejemplo, facilitando la traslación de los datos provenientes de un repositorio RDF a objetos de un modelo de negocio y viceversa.
- ...

Como se ha podido comprobar las líneas de actividad son múltiples tanto desde un punto de vista científico como tecnológico. Evidentemente también es necesario considerar dos factores importantes tras la realización de una investigación profunda en un dominio concreto:

- Capitalización del conocimiento y de la propiedad industrial e intelectual mediante la realización de patentes y de publicaciones.

- Creación de nuevas oportunidades de negocio y servicios a través de la constitución de *spin-off*s y *startup*s que culminen la cadena de valor de la investigación y desarrollo mediante la transferencia de tecnología a empresas de base tecnológica y la generación de nuevos servicios de negocio. En este sentido, se tiene presente tanto la transferencia de tecnología semántica, ya han surgido múltiples empresas tanto a nivel nacional como internacional que hacen uso comercial de esta tecnología, como la implementación de servicios de valor añadido explotando los datos abiertos enlazados, por ejemplo capitalizando la información sobre organizaciones. Si bien la competencia en el mercado constituye una barrera de entrada, la adición de semántica supone un avance respecto a los actuales servicios por lo que es posible crear mejores servicios, más informados, e impulsar la integración e interoperabilidad de las aplicaciones.
- Continuación de la investigación, desarrollo e innovación mediante la realización de propuestas a los principales programas de investigación competitivos para que la aplicación y extensión del conocimiento generado siga en evolución y con posibilidad de resolver la problemática presente en distintos dominios.

Finalmente, con el emergente flujo de información y datos constante y heterogéneo, las técnicas de procesamiento eficiente y explotación de la información y datos para la creación de nuevos servicios constituye una línea de actividad desde diferentes puntos de vista, que la semántica y los datos enlazados son capaces de abordar en condiciones de éxito.

Apéndice A

Impacto y Difusión

> Se alcanza el éxito convirtiendo cada paso en una meta y cada meta en un paso.
>
> *Citas Célebres*
> C.C. CORTÉZ

Las actividades de difusión y de explotación forman parte significativa de las actividades transversales que hay que realizar en la promoción de un trabajo de investigación y se configuran como fundamentales para la colaboración con otras personas o entidades desde un punto de vista científico como para iniciar el proceso de transferencia tecnológica y su posible explotación en el mercado.

El principal objetivo de este tipo de actividades es crear concienciación de los resultados conseguidos durante el estudio y ejecución de los trabajos de investigación tanto a nivel de calidad científica como de su posible explotación industrial. Para llevar a cabo este objetivo se deben seleccionar los canales apropiados de distribución así como las entidades que deben hacer eco de los avances realizados. Para conseguir el impacto deseado estas actividades de difusión deben ejecutarse en paralelo con la investigación para demostrar los avances y el carácter práctico atrayendo tanto las comunidades científicas como industriales.

En general las actividades de difusión y demostración se pueden plantear en dos planos diferentes:

Interna. Permitir que la comunicación entre los miembros participantes en los trabajos de investigación puedan colaborar de forma ágil. Para ello se han utilizado: listas de correo, un wiki y herramientas de desarrollo basadas en el control de versiones. De esta forma, se involucra y se obtiene realimentación tanto del Director de la tesis, crucial en este caso, como de los participantes en el marco del proyecto 10ders, ver Sección 1.1.4.

Externa. Difundir el trabajo realizado para la creación de concienciación pública. Habitualmente la participación en eventos como conferencias, *workshops*, etc. o el envío de artículos a revistas son los medios más utilizados para este tipo de difusión. Hoy en día y con la eclosión de las redes sociales, el uso de sitios como Twitter, Facebook o Linkedin pueden aportar un gran valor a este tipo de difusión consiguiendo un gran impacto real. Para ello se han realizado actividades de distinto tipo: a) Formacionales: celebración de cursos relacionados con la tecnología utilizada, b) Creación de un demostrador público. c) Creación de un sitio web y presencia en las principales redes sociales como canales de distribución de la información y sitios de referencia para consultas. d) Realización de publicaciones científicas e industriales (aplicación) en conferencias

y revistas tanto nacionales como internacionales. e) Informe de resultados y conocimiento a través de las distintas plataformas tecnológicas a las que pertenece el grupo WESO, en proceso.

Finalmente, la difusión de este trabajo busca integrarse en la comunidad científica situándose como referencia en la problemática tratada, *Linking Open Data* en el campo de la contratación pública, así como en el tejido empresarial proporcionando un nuevo enfoque para la gestión de información y conocimiento de la información de carácter público. En las siguientes secciones detallaremos de forma concisa algunas de las actividades realizadas.

A.0.3 Contacto a través de Correo Electrónico y Página Web

Con el objetivo de difundir los resultados parciales que se iban consiguiendo en el desarrollo de los trabajos y estudios realizados durante la elaboración de tesis se realizaron algunos artículos (en formato *blog*) para su posterior publicación a través de las redes sociales. Para ello se conformo la siguiente estructura:

- Página web con un demostrador público del trabajo realizado.
- Uso de la cuenta de Twitter del grupo WESO.
- Uso de la página web personal del autor (con correo electrónico) y del grupo WESO para publicitar los avances.

En este sentido y tras la publicación del clasificaciones de productos como *Linked Data* hemos obtenido la siguiente realimentación:

- Richard Cyganiak Linked *data technologist at the LiDRC*, DERI Galway, Ireland.

Hi Jose,

This looks great! I think it's more appropriate to add the "umbrella" dataset at pscs-catalogue to the LOD Cloud, given that all the individual classifications are published in the same SPARQL endpoint and look more like parts of one large dataset to me. (This is a judgment call, so please let us know if you think that any particular sub-datasets should be highlighted as individual datasets – we would consider it.)

So for now I've added pscs-catalogue to the lodcloud group and it will be in the next update of the LOD Cloud diagram.

The nomenclator-asturias-2010 dataset fulfils the technical criteria, so I'm adding it as well. Good job! ;-)

I've done a change to both of the CKAN records: The links:xxx extra field is supposed to use the name of the respective CKAN record, so it should be links:dbpedia to indicate the number of DBpedia links. I've fixed this for pscs-catalogue and nomenclator-asturias-2010, but you may want to still fix it for the individual sub-datasets.

I've also updated the description of pscs-catalogue to list the individual datasets. Feel free to change this further as you see fit.

Finally, I see that the blog post about the nomenclator dataset lists a number of further links that are not present in the CKAN entry. If there's no particluar reason for omitting them, then I encourage you to add those links as well to the CKAN record.

All the best, and keep up the good work, Richard

- Jindřich Mynarz, University of Economics Prague, miembro del consorcio LOD2 Project [126].

Hi Jose,

thanks for our Sunday Twitter discussion. It's always a pleasure to find someone who works on a closely related topic as you do! ...

Of the classifications you have mentioned, CPV and NACE are of most interest to me. We have started using you CPV codes to describe the public contracts we scraped from the Czech central public contracts repository. Your effort with NACE interests me because I'm working on a conversion of NACE rev. 2 to RDF. We have an agreement with the Czech Statistical Office to provide us with some of their classifications that we can convert to RDF and expose as linked data. One of these classification is NACE rev. 2, to which Czech Statistical Office has added one level of the most specific concepts (i.e., the activities in NACE have more narrower concepts as leaf nodes). It would make sense to me just to release this Czech extension of NACE and link to either to your converted NACE or to EUROSTAT's NACE in RDF on their metadata server. Unfortunately, there are some issues with the data dump I've got from the Czech Statistical Office that block progress on this work.

..

- Andreas Radinger, Professur für Allgemeine BWL, insbesondere E-Business e-business and web science research group. Universität der Bundeswehr München.

> *Hi Jose,*
>
> *congratulations for the very nice dataset at http://www.josemalvarez.es/web/2011/11/05/cpv/. One suggestion: The lemma of the URIs of the ProductOntology are case-sensitive. ... and perhaps you can also fix the correct spelling in some cases? ... Best regards, Andreas*

- Dominique Guardiola, empresa QuiNode en Francia.

> *Thanks for this great work, this is just what I was looking for Just to mention that stores are complaining about the rdf: namespace not being defined in the 2008 dump.*

Por otra parte y a través de los mecanismos que provee Twitter, ya existen trabajos [112] para evaluar el impacto de las publicaciones utilizando este herramienta, como los *retweets* o las menciones de otros usuarios podemos destacar los siguientes mensajes:

- Stefano Bertolo. Project Officer at European Commission DG INFSO/E2.

> *Common Procurement Vocabulary released as #linkeddata using #goodrelations and #productontology via @wesoviedo @mfhepp*

- Jindřich Mynarz.

> *Improving performance of SPARQL for querying EU public procurement notices. http://slidesha.re/tIIUmw by @wesoviedo*

> *@wesoviedo @chema_ar Great! It seems superior to our attempt at publishing CPV as #linkeddata: http://bit.ly/mTMmry*

> *Methods On Linked Data for E-procurement Applying Semantics - demo http://purl.org/weso/moldeas/ by @wesoviedo. #linkeddata*

- Nelson Piedra. Computer Sciences School, Director at UTPL - Universidad Técnica Particular de Loja. Ecuador.

> *Product Scheme Clas.: CPV, CN, CPA, etc. (beta release) as #LinkedData #goodrelations &... http://shar.es/o30mV @Euroalert (via @wesoviedo)*

En esta línea del uso de Twitter también hemos figurado dos veces en las "historias del día" que recoge Ralph Hodgson, Co-founder/CTO de TopQuadrant:

- 5 de noviembre de 2011 (http://paper.li/ralphtq/1316980404/2011/11/05)
- 16 de noviembre de 2011 (http://paper.li/ralphtq/1316980404/2011/11/16)

Finalmente, el desarrollo del trabajo de investigación de esta tesis ha permitido establecer colaboración a través de distintos contactos:

- Jindřich Mynarz. Colaboración en la elaboración de una ontología para formalizar la información contenida en los anuncios de licitación.
- Giner Alor-Hernández (Instituto Tecnológico de Orizaba, México). Elaboración de artículos de científicos para el caso de *e-Procurement* pero más centrado en las cadenas suministros industriales.

- Otras colaboraciones con entidades a nivel regional y nacional, por ejemplo reaprovechando el trabajo de la Universidad de Zaragoza en el campo de la georreferenciación para mejorar al acceso contextualizado a las licitaciones.

Todos estos resultados nos animan a seguir trabajando en esta línea de actuación en el campo de los anuncios de licitación y a promocionar la investigación y resultados realizados.

A.0.4 Contribución a las iniciativas *Linked Data* y a *Linking Open Data*

El trabajo realizado también se ha plasmado en la contribución de los datos generados a las iniciativas de *Linked Data* y *Linking Open Data*. En concreto, el catálogo de clasificaciones de productos se ha difundido a través del registro CKAN provisto por "TheDataHub.org" y solicitado su incorporación a la nube de datos enlazados de forma exitosa. De esta manera, en la próxima actualización de este diagrama, ampliamente usado y extendido en muchas de las presentaciones relacionadas con esta iniciativa, estará presente el catálogo de clasificaciones de productos.

Para consultar esta información se puede acceder a la siguiente URL: `http://thedatahub.org/dataset/pscs-catalogue`, en la cual el catálogo está descrito y definido convenientemente. Igualmente, cada una de las clasificaciones de productos están disponibles individualmente.

Se ha seleccionado este catálogo para su difusión debido a que los datos son abiertos y están disponibles en las fuentes oficiales, respecto a la información y datos concernientes a los anuncios de licitación y organizaciones se ha pospuesto su adición ya que han sido procesados, generados y cedidos de forma conveniente por la empresa Gateway S.C.S. dentro de la ejecución del proyecto "10ders Information Services".

No obstante es necesario marcar siempre como objetivo la difusión de los datos promocionados a la iniciativa de *Linked Data* para facilitar su posterior reutilización, colaborando en el despegue de este nuevo entorno que constituye la *Web of Data*.

Apéndice B

Trabajos publicados

> Si no quieres perderte en el
> olvido tan pronto como estés
> muerto y corrompido,
> escribe cosas dignas de leerse, o
> haz cosas dignas de escribirse.
>
> *Citas Célebres*
> BENJAMIN FRANKLIN

Como ya hemos introducido en el Apéndice A la difusión de los resultados de investigación es una pieza clave para validar y continuar el trabajo en una determinada línea de investigación. Es por ello que la realización de publicaciones científicas y también más orientadas a la industria deben ser contempladas desde el principio para conseguir un impacto real y de calidad en las distintas comunidades. En las siguientes secciones se detallan algunas de las publicaciones más destacadas realizadas en colaboración con distintas instituciones y empresas que resultan del trabajo de esta tesis.

Para facilitar la lectura de las publicaciones realizadas a continuación se dispone de una lista, ver Tabla B.1, de las distintas personas con las que se ha colaborado ordenadas por afiliación.

Autor	**Afiliación**
José Emilio Labra Gayo	Universidad de Oviedo
Patricia Ordoñez De Pablos	Universidad de Oviedo
Francisco Cifuentes Silva	Universidad de Oviedo
Jorge González Lorenzo	Universidad de Oviedo
Pablo Abella Vallina	Universidad de Oviedo
Weena Jiménez Nácero	Universidad de Oviedo
Miguel García Rodríguez	Universidad de Oviedo y Fundación CTIC
Antonio Campos López	Fundación CTIC
Diego Berrueta Muñoz	Fundación CTIC
Luis Polo Paredes	Fundación CTIC
Emilio Rubiera Azcona	Fundación CTIC
Jose Luis Marín	Gateway S.C.S
Ángel Marín	Gateway S.C.S
Mai Rodríguez	Gateway S.C.S
Ramón Calmeau	EXIX TI

Autor	Afiliación
Giner Alor-Hernández	Instituto Tecnológico de Orizaba en México
Cauthemoc Sánchez	Instituto Tecnológico de Orizaba en México
Jaime Alberto Guzman Luna	Universidad Nacional de Colombia
Cristina Casado Lumbreras	Universidad Carlos III de Madrid
Alejandro Rodríguez González	Universidad Carlos III de Madrid
Ricardo Colomo Palacios	Universidad Carlos III de Madrid

Tabla B.1: Colaboración con autores en las publicaciones.

B.0.5 Revistas Internacionales con Índice de Impacto

1. Jose María Alvarez Rodríguez, José Emilio Labra Gayo, Francisco Cifuentes Silva, Giner Alor-Hernández, Cauthemoc Sánchez y Jaime Alberto Guzman Luna. *Towards a Pan-European E-Procurement platform to Aggregate, Publish and Search Public Procurement Notices powered by Linked Open Data: The MOLDEAS Approach*. International Journal of Software Engineering and Knowledge Engineering (IJSEKE) Focused Topic Issue on Consuming and Producing Linked Data on Real World Applications. 2011. Factor de Impacto: 0,262.

2. Jose María Alvarez Rodríguez, José Emilio Labra Gayo, Patricia Ordoñez De Pablos. *Survey of New Trends on e-Procurement Applying Semantics*. International Journal of Computers in Industry Focused Topic Issue on New Trends on *e-Procurement* Applying Semantics. 2014. Factor de Impacto: 1,620. Como parte de la aceptación de esta revista para ser editores invitados del *Special Issue* titulado como *New Trends on e-Procurement Applying Semantic* se nos emplaza por parte de los editores a realizar esta artículo. La semilla del mismo es el propio documento de la tesis.

B.0.6 Editor Invitado de *Special Issues* en Revistas Internacionales

1. Jose María Alvarez Rodríguez, José Emilio Labra Gayo, Patricia Ordoñez De Pablos. *New Trends on e-Procurement Applying Semantics*. International Journal of Computers in Industry Focused Topic Issue on New Trends on *e-Procurement* Applying Semantics. 2014. Factor de Impacto: 1,620.

B.0.7 Revistas Internacionales

1. Jose María Alvarez Rodríguez, José Emilio Labra Gayo, Patricia Ordoñez De Pablos. *An Extensible Framework to Sort Out Nodes in Graph-based Structures Powered by the Spreading Activation Technique: The ONTOSPREAD approach*. International Journal of Knowledge Society Research (IJKSR). 2011. Aceptado para publicación.

2. Jorge González Lorenzo, José Emilio Labra Gayo y Jose María Alvarez Rodríguez. *A MapReduce implementation of the Spreading Activation algorithm for processing large knowledge bases based on semantic networks*. International Journal of Knowledge Society Research (IJKSR). Aceptado para publicación.

3. Jose María Alvarez Rodríguez, José Emilio Labra Gayo, Ramón Calmeau, Ángel Marín y Jose Luis Marín. *Query Expansion Methods and Performance Evaluation for Reusing Linking Open Data*

of the European Public Procurement Notices. Current Topics in Artificial Intelligence. 14th Conference of the Spanish Association for Artificial Intelligence, CAEPIA 2011, La Laguna, Spain, November 8-11, 2011, Selected Papers.

4. Jose María Alvarez Rodríguez, José Emilio Labra Gayo, Ramón Calmeau, Ángel Marín y Jose Luis Marín.*Innovative Services to ease the Access to the Public Procurement Notices using Linking Open Data and Advanced Methods based on Semantics*. International Journal of Electronic Government. Aceptada para publicación, marzo 2012.

B.0.8 Capítulos de Libros

1. Jose Luis Marín, Mai Rodríguez, Ramón Calmeau, Ángel Marín, Jose María Alvarez Rodríguez y José Emilio Labra Gayo. *Euroalert.net: aggregating public procurement data to deliver commercial services to SMEs*. "E-Procurement Management for Successful Electronic Government System". IGI Global. 2012.

2. Jose María Alvarez Rodríguez, Luis Polo Paredes, Emilio Rubiera Azcona, José Emilio Labra Gayo y Patricia Ordoñez De Pablos. *Enhancing the Access to Public Procurement Notices by Promoting Product Scheme Classifications to the Linked Open Data Initiative*. "Cases on Open-Linked Data and Semantic Web Applications". IGI Global. 2012.

B.0.9 Conferencias Internacionales

1. Jose María Alvarez Rodríguez, José Emilio Labra Gayo, Patricia Ordoñez De Pablos. *An Extensible Framework to Sort Out Nodes in Graph-based Structures Powered by the Spreading Activation Technique: The ONTOSPREAD approach*. The 4th World Summit on the Knowledge Society (WSKS 2011). 2011.

2. Jorge González Lorenzo, José Emilio Labra Gayo y Jose María Alvarez Rodríguez. *A MapReduce implementation of the Spreading Activation algorithm for processing large knowledge bases based on semantic networks*.The 4th World Summit on the Knowledge Society (WSKS 2011). 2011.

3. Jose María Alvarez Rodríguez, José Emilio Labra Gayo, Ramón Calmeau, Ángel Marín y Jose Luis Marín. *Innovative Services to ease the Access to the Public Procurement Notices using Linking Open Data and Advanced Methods based on Semantics*. 5th International Conference on Methodologies & Tools enabling e Government (MeTTeG 2011). Camerino-Italia 2011.

B.0.10 *Workshops* Internacionales

1. Jose María Alvarez Rodríguez, Luis Polo Paredes, Pablo Abella Vallina, Weena Jiménez Nácero y José Emilio Labra Gayo. *Application of the Spreading Activation Technique for Recommending Concepts of well-known ontologies in Medical Systems*. SATBI 2011. ACM BCB (Chicago IL, USA).

2. Jose Luis Marín, Mai Rodríguez, Ramón Calmeau, Ángel Marín, Jose María Alvarez Rodríguez y José Emilio Labra Gayo. *Euroalert.net: Building a pan-European platform to aggregate public procurement data and deliver commercial services for SMEs powered by open data*. in Workshop Share-PSI.eu. 2011.

3. Jose María Alvarez Rodríguez, Emilio Rubiera Azcona y Luis Polo Paredes. *Promoting Government Controlled Vocabularies for the Semantic Web: the EUROVOC Thesaurus and the CPV Product Classification System*. SIEDL 2008. ESWC 2008.

B.0.11 *Posters* Internacionales

1. Jose María Alvarez Rodríguez, José Emilio Labra Gayo, Ramón Calmeau, Ángel Marín y Jose Luis Marín. *MOLDEAS-Methods On Linked Data for E-procurement Applying Semantics*. The Seventh Reasoning Web Summer School, Galway, Ireland at the Digital Research Enterprise Institute (DERI) from August 23-27, 2011.

2. Jose María Alvarez Rodríguez y José Emilio Labra Gayo. *Semantic Methods for Reusing Linking Open Data of the European Public Procurement Notices*. PhD Symposium ESCW 2011. Se aceptó como póster pero finalmente se declinó ir a presentar este trabajo.

B.0.12 *Workshops* Nacionales

1. Jose María Alvarez Rodríguez, José Emilio Labra Gayo, Ramón Calmeau, Ángel Marín y Jose Luis Marín. *Query Expansion Methods and Performance Evaluation for Reusing Linking Open Data of the European Public Procurement Notices*. Workshop Tecnologías Linked Data y sus Aplicaciones en España, CAEPIA 2011, La Laguna, Spain, November 8-11, 2011.

2. Francisco Cifuentes Silva, José Emilio Labra Gayo y Jose María Alvarez Rodríguez. *An architecture and process of implantation for Linked Data environments*. Workshop Tecnologías Linked Data y sus Aplicaciones en España, CAEPIA 2011, La Laguna, Spain, November 8-11, 2011.

3. Jose María Alvarez Rodríguez, Emilio Rubiera Azcona y Luis Polo Paredes. *Generación automática de ontologías en SKOS de clasificaciones estándar de productos: Common Procurement Vocabulary (CPV)*. CEDI 2007. CAEPIA 2007.

B.0.13 Relacionados con Web Semántica y *Linked Data*

1. Cristina Casado Lumbreras, Alejandro Rodríguez González, Jose María Alvarez Rodríguez y Ricardo Colomo Palacios. *PsyDis: towards a diagnosis support system for psychological disorders.*. International Journal Expert Systems With Applications. 2012. Factor de Impacto: 1,926. Aceptado para publicación.

2. Jose María Alvarez Rodríguez y Antonio Campos López. *Integration and Interoperability on Service Oriented Architectures using Semantics*. PhD Symposium. ICWE 2009. San Sebastián, Spain.

3. Miguel García Rodríguez, Jose María Alvarez Rodríguez, Diego Berrueta Muñoz y Luis Polo Paredes. *Declarative Data Grounding Using a Mapping Language*. Communications of SIWN. ISSN 1757-4439. April 2009.

4. Miguel García Rodríguez, Jose María Alvarez Rodríguez, Diego Berrueta Muñoz y Luis Polo Paredes. *Declarative Data Grounding Using a Mapping Language*. 3rd Complex Distributed Systems (CODS 2009). Leipzig, Germany International Conference.

Como resumen de estas publicaciones se puede comentar que la mayoría se han realizado en la fase final de la tesis debido a la consolidación de la investigación realizada y al disponer de fondos para su difusión. También añadir que la publicación de los artículos ha sido acompañada de las presentaciones consiguientes en los lugares de aceptación y en la mayoría de los casos los artículos presentados han recibido la correspondiente invitación para su posterior publicación en revistas bajo las condiciones establecidas por los organizadores sobre extensión y contenido.

Finalmente, estos trabajos sirven como punto de partida y experiencia para la realización de nuevas publicaciones, teniendo como objetivo la mejora continua en calidad y contemplando aspectos de investigación de la tesis que se consideran de gran valor para este trabajo y de los cuales se intentará sacar la mayor realimentación posible a través de su publicación en la comunidad científica.

B.0.14 Otros

Debido a la participación del equipo WESO de la Universidad de Oviedo en la "Red Temática Española de Linked Data" se ha presentado este trabajo en la reunión plenaria celebrado en junio de 2011 en las instalaciones de la Universidad de Politécnica de Madrid.

Por otra parte, se han redactado varios entregables [265] de los distintos proyectos de investigación en los que se ha participado y dirigido proyectos fin de carrera como "Sistema Dinámico y Automático de Generación de Preguntas desde Fuentes Heterogéneas basada en Formatos de Publicación de la Web Semántica" presentado por Lucía Otero García en marzo de 2010 para la obtención del título de Ingeniería Técnica en Informática en la Escuela Politécnica de Ingeniería de Gijón.

Además, como parte de la actividad institucional y profesional desarrollada por los distintos miembros del grupo se han realizado demostraciones a distintas empresas y organismos públicos con el objetivo de concienciar de las posibilidades del enfoque de *Linked Data* para la resolución de ciertos problemas como pueden ser la integración de fuentes datos heterogéneas o la mejora de los sistemas de búsqueda.

Apéndice C

Tablas de Validación

> Si no conozco una cosa, la investigaré.
>
> *Citas Célebres*
> LOUIS PASTEUR

C.1 Consideraciones sobre las Tablas de Validación

El objetivo de esta sección es presentar las tablas de validación que se utilizan para realizar la evaluación del grado de cumplimiento de criterios relativos a las iniciativas de *Linked Data* y *Open Data* para verificar que verdaderamente los datos generados se pueden enclavar bajo estos enfoques asegurando por tanto los beneficios y ventajas propugnados por los mismos. La elaboración y diseño de estas tablas ha considerado el trabajo relacionado presentado en la Sección 3.2.6 así como la experiencia personal en el desarrollo de distintas actividades relacionadas con estas iniciativas.

Los criterios presentados en estas tablas pueden adquirir tres valores:

- Valor positivo, $\oplus$, si es un criterio que debe tener y se cumple.
- Valor negativo, $\ominus$, si es un criterio que debe tener y no se cumple.
- Valor no aplicable, $\odot$, si es un criterio que se desconoce, se solapa con otro o bien o no está asociado al enfoque evaluado.

Para ejemplificar una validación en las tablas siguientes se realiza la valoración de un conjunto de datos de referencia en general, y *datasets* RDF en particular. El grado de cumplimiento de criterios ideal se fija en un total de 173 criterios positivos ($\oplus$) y 23 no aplicables ($\odot$) realizando una encuesta total de 196 criterios de valoración. De esta forma, es posible delimitar si una enfoque pertenece a la iniciativa de *Open Data*, *Linked Data*, a la nube de datos enlazados abiertos o a un registro CKAN, adicionalmente se valora el grado en el que los procesos de producción, publicación, consumo, validación y realimentación de los conjuntos de datos se facilitan.

C.1.1 Tabla de Validación T^1

En la Tabla C.1 se recogen criterios de validación relativos a los procesos de producción, publicación y consumo de datos enlazados, realizando un especial hincapié en el diseño de URIs, modelado

de información y datos utilizando tecnologías semánticas así como factores relativos a la descripción de la metainformación de los recursos. Se ha realizado la valoración ideal de un *dataset* RDF de referencia que contaría con 60 criterios positivos (⊕) y 9 no aplicables (⊙).

ID	Pregunta	Cumplimiento
1	***Uso de URIs***	
1.1	¿Las URIs utilizadas permiten acceder a los recursos, *Minting HTTP URIs*?	⊕
1.2	¿El *namespace* utilizado en las URIs está bajo nuestro control?	⊕
1.3	¿Se utiliza el esquema HTTP?	⊕
1.4	¿Las URIs siguen las directrices de *Cool URIs*?	⊕
1.5	¿Se utilizan *hash URIs*?	⊙
1.6	¿Se utilizan *slash URIs*?	⊕
1.7	¿Se utilizan *param URIs*?	⊙
1.8	¿No se incluyen detalles de implementación en la URI?	⊕
1.9	¿Se utilizan claves primarias para identificar los recursos, ID URIs?	⊕
1.10	¿Se utilizan nombres para identificar a los recursos, *Meaningful URIs*?	⊙
1.11	¿Se ha definido una URI base para los recursos?	⊕
1.12	¿Se ha definido una URI base para el modelo formal?	⊕
1.13	¿Se ha definido un esquema de URIs para el *dataset* RDF?	⊕
1.14	¿Se ha definido un esquema de URIs para el modelo formal?	⊕
1.15	¿Se incluye metainformación en las URIs?	⊕
2	***Descripción de recursos en RDF***	
2.1	¿Se utilizan vocabularios como SKOS, RDFS, OWL para modelar el dominio?	⊕
2.2	¿Se reutilizan vocabularios de acuerdo a su uso y actualización?	⊕
2.3	¿Se utilizan anotaciones de RDFS o SKOS?	⊕
2.4	¿Se relacionan las clases y propiedades con otras ya existentes?	⊕
2.5	¿Se reutilizan las clases y propiedades ya existentes?	⊕
2.6	¿Se definen nuevas clases y propiedades?	⊕
2.7	¿Se enriquecen las descripciones RDF con otros *datasets* RDF?	⊕
2.8	¿Se describe parcialmente los recursos de otros *datasets* RDF a los que se enlaza desde el actual?	⊖
2.9	¿Se añade metainformación a cada recurso RDF individualmente?	⊕

ID	Pregunta	Cumplimiento
2.10	¿Son las descripciones de los recursos RDF navegables?	⊕
2.11	¿Se provee información útil del recurso RDF a través de la URI?	⊕
2.12	¿Son las URIs reutilizadas referenciables?	⊕
3	***Descripción del dataset RDF***	
3.1	¿Se ha definido una URI para el *dataset* RDF?	⊕
3.2	¿Se utiliza el vocabulario voID para describir el *dataset* RDF?	⊕
3.3	¿Se utiliza el *Semantic Sitemaps* para describir el *dataset* RDF?	⊙
3.4	¿Se provee información de *provenance*?	⊕
3.5	¿Se provee licencia de uso, información sobre derechos de copia, etc.?	⊕
3.6	¿Se provee información sobre la autoría?	⊕
3.7	¿Se proveen ejemplos de uso del *dataset* RDF?	⊕
3.8	¿Se utilizan anotaciones de RDFS o SKOS?	⊕
3.9	¿Se utilizan anotaciones multilingües RDFS o SKOS?	⊕
4	***Otros***	
4.1	¿Se utilizan herramientas automáticas para la producción de RDF?	⊕
4.2	¿Se consolidan parte de los datos en RDF?	⊖
4.3	¿Se realiza la reconciliación de entidades de forma automática?	⊖
4.4	¿Los datos son dinámicos?	⊙
4.5	¿Los datos son estáticos?	⊕
4.6	¿El tamaño del *dataset* es del orden de millones de tripletas?	⊕
4.7	¿El tamaño del *dataset* es del orden de billones de tripletas?	⊙
5	***Publicación de Linked Data***	
5.1	¿Se publica un volcado de los datos en RDF?	⊖
5.2	¿Se utiliza algún método de publicación de datos enlazados?	⊕
5.3	¿Se provee algún lenguaje de consulta formal?	⊕
5.4	¿Se provee un *endpoint de SPARQL*?	⊕
5.5	¿Se provee negociación de contenido?	⊕
5.6	¿Se pueden referenciar los recursos desde otros documentos tipo HTML?	⊕
5.7	¿Se provee un *Linked Data Frontend*?	⊕
5.8	¿Se provee metainformación del *dataset* RDF?	⊕

ID	Pregunta	Cumplimiento
5.9	¿Se difunde el *dataset* RDF en distintos medios tipo CKAN, Prefix.cc, etc.?	$\oplus$
5.10	¿Se publica algún API o servicio web para la consulta de los datos?	$\oplus$
5.11	¿Existe alguna restricción en la consulta de los datos?	$\ominus$
5.12	¿Se establece algún mecanismo de privacidad?	$\ominus$
5.13	¿Se informa del tamaño del *dataset*?	$\oplus$
5.14	¿Se publican los datos RDF como un fichero estático?	$\odot$
5.15	¿Se publican los datos RDF bajo demanda desde base de datos?	$\odot$
5.16	¿Se publican los datos RDF en un repositorio?	$\oplus$
5.17	¿Se publican los datos RDF bajo demanda desde una aplicación?	$\oplus$
5.18	¿Se publican los datos con información temporal y evolución en el tiempo?	$\ominus$
5.19	¿Se proveen ejemplos para depuración y consumo?	$\oplus$
5.20	¿Se proveen alias a ciertos directorios o nombres?	$\oplus$
5.21	En caso de error, ¿se devuelve algún recurso por omisión?	$\ominus$
5.22	¿Se utilizan protocolos estándar?	$\oplus$
5.23	¿Se provee algún mecanismo de realimentación?	$\ominus$
5.24	¿Se provee documentación sobre los datos publicados?	$\oplus$
5.25	¿Se proveen estadísticas de los datos publicados?	$\oplus$
5.26	¿Se utilizan mecanismos de sellado en el tiempo o similares?	$\oplus$

Tabla C.1: T^1-Tabla de Validación de Características *Linked Data*.

C.1.2 Tabla de Validación T^2

En la Tabla C.2 se recogen criterios de validación relativos a los a los patrones de diseño utilizados para la generación de datos enlazados. Si bien su uso no es obligatorio, la idea subyacente coincide con los patrones de diseño en ingeniería del software en los cuales se ofrecen soluciones para casuísticas habituales con el objetivo de mejorar tanto los procesos de producción, publicación, consumo, validación y realimentación. Se ha realizado la valoración ideal de un *dataset* RDF de referencia que contaría con 44 criterios positivos ($\oplus$), incluyendo la aplicación de todos los patrones, se ha tomado esta decisión para obtener una medida del grado de aplicación de los mismos ya que en muchos casos la información generada en un *dataset* RDF puede utilizar unos u otros no es conveniente plantear su uso como disyuntivo.

ID	Pregunta	Cumplimiento
1	***Identifier Patterns***	
1.1	*Hierarchical URIs*	⊕
1.2	*Literal Keys*	⊕
1.3	*Natural Keys*	⊕
1.4	*Patterned URIs*	⊕
1.5	*Proxy URIs*	⊙
1.6	*Shared Keys*	⊙
1.7	*URL Slug*	⊙
2	***Modelling Patterns***	
2.1	*Custom Datatype*	⊕
2.2	*Index Resources*	⊙
2.3	*Label Everything*	⊕
2.4	*Link Not Label*	⊕
2.5	*Multi-Lingual Literal*	⊕
2.6	*N-Ary Relation*	⊙
2.7	*Ordered List*	⊙
2.8	*Ordering Relation*	⊙
2.9	*Preferred Label*	⊕
2.10	*Qualified Relation*	⊕
2.11	*Reified Statement*	⊙
2.12	*Topic Relation*	⊕
2.13	*Typed Literal*	⊕
3	***Publishing Patterns***	
3.1	*Annotation*	⊕
3.2	*Autodiscovery*	⊕
3.3	*Document Type*	⊕
3.4	*Edit Trail*	⊖
3.5	*Embedded Metadata*	⊕
3.6	*Equivalence Links*	⊕
3.7	*Link Base*	⊕
3.8	*Materialize Inferences*	⊙
3.9	*Named Graphs*	⊕
3.10	*Primary Topic Autodiscovery*	⊕
3.11	*Progressive Enrichment*	⊕
3.12	*See Also*	⊕
4	***Application Patterns***	
4.1	*Assertion Query*	⊕
4.2	*Blackboard*	⊕
4.3	*Bounded Description*	⊕
4.4	*Composite Descriptions*	⊕
4.5	*Follow Your Nose*	⊖
4.6	*Missing Isn't Broken*	⊖
4.7	*Parallel Loading*	⊖
4.8	*Parallel Retrieval*	⊖

ID	Pregunta	Cumplimiento
4.9	*Resource Caching*	⊖
4.10	*Schema Annotation*	⊕
4.11	*Smushing*	⊕
4.12	*Transformation Query*	⊖

Tabla C.2: T^2-Tabla de Validación de *Linked Data Patterns*.

C.1.3 Tabla de Validación T^3

En la Tabla C.3 se recogen los principios de *Linked Data* que sirven para valorar si un enfoque se encuentra dentro de esta iniciativa. Se ha realizado la valoración ideal de un *dataset* RDF de referencia que contaría con 4 criterios positivos (⊕).

ID	Pregunta	Cumplimiento
1.1	*Use URIs as names for things*	⊕
1.2	*When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)*	⊕
1.3	*Include links to other URIs*	⊕
1.4	*Use HTTP URIs*	⊕

Tabla C.3: T^3-Tabla de Validación de Principios de *Linked Data*.

C.1.4 Tabla de Validación T_1^3

En la Tabla C.4 y en combinación con la anterior, se recogen los criterios relacionados con el modelo de ⋆ propuesto por Tim Berners-Lee y de esta forma se puede establecer el nivel de datos enlazados de un determinado *dataset* RDF, el valor de referencia máximo contaría con 5 criterios positivos (⊕), uno por cada ⋆.

ID	Pregunta	Cumplimiento
1.1	⋆	⊕
1.2	⋆⋆	⊕
1.3	⋆⋆⋆	⊕
1.4	⋆⋆⋆⋆	⊕
1.5	⋆⋆⋆⋆⋆	⊕

Tabla C.4: T_1^3-Tabla de Validación del Modelo ⋆.

C.1.5 Tabla de Validación T^4

En la Tabla C.5 se recogen los principios de *Open Data* que sirven para valorar si un enfoque se encuentra dentro de esta iniciativa. Se ha realizado la valoración ideal de un *dataset* RDF de referencia, o en general un conjunto de datos públicos, contaría con 8 criterios positivos (⊕).

ID	Pregunta	Cumplimiento

ID	Pregunta	Cumplimiento
	8 Principios	
1.1	*Complete*	⊕
1.2	*Primary*	⊕
1.3	*Timely*	⊕
1.4	*Accessible*	⊕
1.5	*Machine processable*	⊕
1.6	*Non-Discriminatory*	⊕
1.7	*Non-Proprietary*	⊕
1.8	*License-free*	⊕

Tabla C.5: T^4-Tabla de Validación de Principios de *Open Data*.

C.1.6 Tabla de Validación T_1^4

En la Tabla C.6 se recogen beneficios y ventajas de la aplicación de los principios de *Open Data*. Se ha realizado la valoración ideal de un *dataset* RDF, o en general de un conjunto de datos públicos, de referencia contaría con 14 criterios positivos (⊕).

ID	Pregunta	Cumplimiento
	Producción	
1.1	¿Se ha definido una misión y estrategia para la apertura de los datos?	⊖
1.2	¿Los datos proceden de una fuente segura?	⊖
1.3	¿Se puede conocer la procedencia de los datos?	⊕
1.4	¿Existe algún mecanismo para asegurar la calidad de los datos?	⊕
	Ventajas	
2.1	¿Facilitan los datos la inclusión?	⊕
2.2	¿Mejoran la transparencia?	⊕
2.3	¿Existe alguna responsabilidad sobre los datos?	⊖
	Beneficios	
3.1	¿Pueden las aplicaciones servirse de estos datos para generar servicios, reutilización?	⊕
3.2	¿Se pueden generar múltiples vistas de los datos?	⊕
3.3	¿Se pueden integrar con otras fuentes de datos?	⊕
	Consumo	
4.1	Uso de anotaciones	⊕
4.2	¿Se provee un API o servicio web de consumo?	⊕

ID	Pregunta	Cumplimiento
4.3	¿Se provee algún mecanismo de sindicación para obtener los datos?	⊖
4.4	¿Existe algún modelo formal o especificación de los datos publicados?	⊕

Tabla C.6: T_1^4-Tabla de Validación sobre Características de *Open Data*.

C.1.7 Tabla de Validación T^5

En la Tabla C.7 se recogen los criterios para pertenecer a la nube de datos enlazados y abiertos. Estos criterios son complementarios a los señalados en las tablas anteriores en el sentido del grado de cumplimiento de *Linked Data* y *Open Data*. Sin embargo, la importancia de esta valoración reside en calibrar el grado de reutilización de información y la capacidad de difusión del *dataset* RDF. Es por ello que se fija una valoración de referencia de 5 criterios positivos (⊕).

ID	Pregunta	Cumplimiento
1	¿Son los recursos RDF accesibles mediante HTTP o HTTPS?	⊕
2	¿Se provee negociación de contenido?	⊕
3	¿El *dataset* contiene más de 1000 tripletas?	⊕
4	¿Se provee, al menos, 50 enlaces a *datasets* ya disponibles en el diagrama?	⊕
5	¿Se provee acceso al *dataset* completo?	⊕

Tabla C.7: T^5-Tabla de Validación sobre Características para pertenecer a *The Linking Open Data Cloud* de los Anuncios de Licitación.

C.1.8 Tabla de Validación T^6

En la Tabla C.8 se recogen los criterios para pertenecer a un registro CKAN, complementario y obligatorio a la pertenencia a la nube de datos enlazados y abiertos. Este valoración cobra especial relevancia si el *dataset* RDF va a ser publicado para su reutilización por terceros, en el caso de datos enlazados a nivel privado no tendría tanta trascendencia. La valoración de un *dataset* RDF de referencia se fija en 33 criterios positivos (⊕) y 14 criterios positivos (⊙).

ID	Pregunta	Cumplimiento
1	**Standard CKAN fields**	
1.1	*Name*	⊕
1.2	*Title*	⊕
1.3	*URL*	⊕
	Enhanced CKAN fields	
1.4	*Version*	⊕
1.5	*Notes*	⊖
1.6	*Author*	⊕
1.7	*Author email*	⊕
1.8	*License*	⊕

ID	Pregunta	Cumplimiento
	Custom CKAN fields	
1.9	*shortname*	
1.10	*license_link*	⊕
1.11	*sparql_graph_name*	⊖
1.12	*namespace*	⊕
1.13	*triples*	⊕
1.14	*links:xxx*	⊕
2	**CKAN tags**	
2.1	*<topic>*	⊕
	Metainformation CKAN tags	
2.2	*format-<prefix>*	⊕
2.3	*no-proprietary-vocab*	⊙
2.4	*deref-vocab*	⊕
2.5	*no-deref-vocab*	⊙
2.6	*vocab-mappings*	⊕
2.7	*no-vocab-mappings*	⊙
2.8	*provenance-metadata*	⊕
2.9	*no-provenance-metadata*	⊙
2.10	*license-metadata*	⊕
2.11	*no-license-metadata*	⊙
2.12	*published-by-producer*	⊙
2.13	*published-by-third-party*	⊕
2.14	*limited-sparql-endpoint*	⊖
2.15	*lodcloud.nolinks*	⊖
2.16	*lodcloud.unconnected*	⊖
2.17	*lodcloud.needsinfo*	⊖
2.18	*lodcloud.needsfixing*	⊖
3	**CKAN resource links**	
3.1	*Download Page*	⊕
3.2	*meta/sitemap*	⊙
3.3	*api/sparql*	⊕
3.4	*meta/void*	⊕
3.5	*application/rdf+xml*	⊕
3.6	*text/turtle*	⊙
3.7	*application/x-ntriples*	⊙
3.8	*application/x-nquads*	⊙
3.9	*meta/rdf-schema*	⊕
3.10	*example/rdf+xml*	⊕
3.11	*example/turtle*	⊙
3.12	*example/ntriples*	⊙
3.13	*example/rdfa*	⊙
3.14	*mapping/{format}*	⊖

Tabla C.8: T^6-Tabla de Validación para registrar el *dataset* en CKAN de los Anuncios de Licitación.

Glosario

10ders Proyecto de investigación cofinanciado por el Ministerio de Industria, Turismo y Comercio dentro del plan Avanza 2 con código TSI-020100-2010-919. 1.1, 2.9, A.0

A2A *Administration to Administration*. 1.0, 3.3
AEAT Agencia Estatal de Administración Tributaria. 2.1
AIISO *Academic Institution Internal Structure Ontology*. 3.4
ANT *Pure Java build tool*. 1.4
API *Application Programming Interface*. 3.2, 4.7, 5.2, 5.3, 6.3, 6.4, C.1
ATOM *Atom Publishing Protocol*. 3.2

B2B *Business to Business*. 1.0, 2.9, 3.3
B2C *Business to Client*. 1.0, 3.3
BBDD Base de Datos. 7.1
Big Data *Data sets that grow so large that they become awkward to work with using on-hand database management tools*. 3.1
BOE BoletínOficial del Estado. 2.1, 5.1, 7.1
BOPA Boletín Oficial del Principado de Asturias. 1.5
BORME Boletín Oficial del Registro Mercantil. 7.1

CIIU Clasificación Industrial Internacional Uniforme. 2.10
CKAN *It is the world's leading open-source data portal platform*. 4.10, 7.1, A.0, C.1
CN *Combined Nomenclature*. 2.10, 5.2
CODICE Componentes y Documentos Interoperables para la Contratación Electrónica. 2.2, 2.9, 5.1
COLD *Consuming Linking Open Data*. 3.2
CPA Clasificación de Productos por Actividades. 2.10, 5.2
CPC Clasificación Central de Productos. 2.10, 5.2
CPV *Common Procurement Vocabulary*. 2.10, 3.4, 5.2, 6.4, 6.6, 7.1–7.3
CSV *Comma-separated values*. 3.2, 4.4, 4.5, 5.1–5.3, 6.4, 7.1

DAML+OIL *DAML+OIL Ontology*. 3.1
DAO Patrón de Diseño *Data Access Object*. 6.4
DARPA *Defense Advanced Research Projects Agency*. 3.1
DG-INFSO *The Information Society and Media Directorate General*. 2.8
DG-MARKT *The Internal Market and Services Directorate General*. 2.8
DG-TAXUD *Taxation and Customs Union*. 2.8
DIGIT *Directorate-General for Informatics*. 2.8
DL *Description Logics*. 3.1
DOAP *Description Of A Project RDF vocabulary*. 3.1
DOUE Diario Oficial de la Unión Europea. 1.1, 2.1, 5.1
DTD *Document Type Definition*. 3.1

e-Procurement Contratación Pública Electrónica. 0.0, 1.0, 2.3, 2.9, 2.12, 3.3, 3.4, 7.1, 8.1, 8.2, A.0, B.0

ebXML *Electronic Business using eXtensible Markup Language*. 3.4
Eclipse *A universal toolset for development*. 1.4
EEE Islandia, Liechtenstein y Noruega. 2.8
eGovernment Administración Pública Electrónica. 2.8, 3.2
ePRIOR *electronic PRocurement Invoicing and ORdering*. 2.8
ERP *Enterprise resource planning*. 3.4
Estados Estados Miembros. 1.0, 2.1, 2.3, 2.4, 2.7, 2.8
ETL *Extract, transform and load*. 4.4
Europea Unión Europea. 1.0, 1.1, 2.1, 2.3, 2.6–2.8, 7.1, 8.1

F2F *Face to Face*. 3.2
FOAF *Friend Of A Friend RDF vocabulary*. 3.1, 3.4, 5.3
FP7 *The Seventh Framework Programme*. 3.4
FTP *File Transfer Protocol*. 5.3

GIS Sistema de Información Geográfica. 3.2

HTML *HyperText Markup Language*. 1.5, 3.1, 3.2, 3.4, 4.4, 7.1, C.1
HTTP *Hypertext Transfer Protocol*. 2.9, 3.1, 3.2, 4.1, 4.6, 6.2, 6.4, 7.1, 8.3, C.1
HTTP URI URIs basadas en el esquema HTTP. 3.2, 4.4, 7.1, C.1
HTTPS *Hypertext Transfer Protocol Secure*. C.1

IDE *Integrated Development Environment*. 6.2
IRI *Internationalized Resource Identifier*. 1.6, 3.1, 4.1
ISIC *International Standard Industrial Classification of All Economic Activities*. 2.10, 5.2
ISO *International Organization for Standardization*. 2.9, 3.4, 5.3, 6.5

J2EE *Java 2 Platform, Enterprise Edition*. 6.2
JAR *Java ARchive*. 6.4
JAXB *Java Architecture for XML Binding*. 3.1
JSON *JavaScript Object Notation*. 3.1–3.3, 4.7

KOS *Knowledge Organization Systems*. 5.2

LDOW *Linked Data On the Web*. 3.2
LOD *Linking Open Data*. 3.2
LODQ *LOD Data Quality Vocabulary*. 7.1
Lucene *Ultra-fast Search Library and Server*. 5.2, 6.2, 6.4, 7.2, 8.2

Mahout *Scalable machine learning and data mining*. 1.4, 6.2, 8.2
Maven *Apache build manager for Java projects*. 1.4
MIME *Multipurpose Internet Mail Extensions*. 3.2, 4.6
MOLDEAS *Methods On Linked Data for E-procurement Applying Semantics*. 1.6, 4.6, 5.1, 5.3, 6.2–6.6, 7.1–7.3, 8.1, 8.2, 8.5

N3 *Notation 3 Logic*. 3.1, 3.2, 4.4, 4.6
NACE *Nomenclature générale des activités économiques dans les Communautés européennes*. 2.10
NAICS *North American Industry Classification System*. 2.10, 5.2
NAPCS *North American Product Classification System*. 2.10
NUTS Nomenclatura de las Unidades Territoriales Estadísticas. 4.4, 5.3, 6.4, 6.6, 7.3

OASIS *Advancing open standards for the information society*. 3.4
OG *Open Government*. 2.5
OGD *Open Government Data*. 1.0

OIL *Ontology Inference Layer*. 3.1
OKFN *Open Knowledge Foundation*. 3.2
OMG *Object Management Group*. 3.3, 3.4
ONU Organización de las Naciones Unidas. 2.10
opXML Vocabulario XML de contratos públicos del proyectos "10ders Information Services". 2.9, 3.4, 5.1, 7.1
ORM *Object-Relational Mapping*. 4.7
OWL *Web Ontology Language*. 3.1–3.4, 4.1, 5.1–5.3, 6.2, 6.4

PDF *Portable Document Format*. 2.10, 3.1, 7.1
PEPPOL *Pan-European Public Procurement Online*. 2.6, 2.8
PIB Producto Interior Bruto. 1.0, 2.1
PIN *Prior Information Notice*. 2.1
POJO *Plain Old Java Object*. 6.2
POST Petición sobre HTTP según RFC 2616. 2.9
PRODCOM *PRODuction COMmunautaire*. 2.10
PSC *Product Scheme Classifications*. 5.2, 7.1
PSI Información del Sector Público. 1.0, 3.2
Pubby *A Linked Data Frontend for SPARQL Endpoints*. 1.4, 3.2, 4.6, 6.2
PYME Pequeña y Mediana Empresa. 1.0, 1.1, 2.7

RAMON *Eurostat's Metadata Server*. 2.10
RDF *Resource Description Framework*. 1.0, 1.5, 3.1–3.4, 4.1, 4.2, 4.4–4.10, 5.1–5.3, 6.2, 6.4, 6.6, 7.0–7.3, 8.3, 8.5, C.1
RDF Schema *Resource Description Framework Schema*. RDF(S) ó RDFS. 3.1
RDF/XML *It is a syntax, defined by the W3C, to express (i.e. serialize) an RDF graph as an XML document*. 3.1, 3.2, 4.4
RDFa *Bridging the Human and Data Webs*. 3.1, 3.2, 3.4
REEEP *Renewable Energy and Energy Efficiency Partnership*. 3.2
Refine *A power tool for working with messy data, cleaning it up, transforming it from one format into another*. 1.4, 4.4, 5.2, 5.3, 6.4
REST *Representational State Transfer*. 1.5, 2.9, 3.2, 4.4, 4.6, 5.1, 6.2, 6.4
RFC *Request For Comments*. 4.1
RIF *Rule Interchange Format*. 1.5, 3.1, 3.3, 3.4
ROLEC Registro Oficial de Licitadores y de Empresas Clasificadas. 2.1
RSS *Really Simple Syndication*. 1.5, 3.1, 3.2, 3.4, 7.1

SaaS *Software as a Service*. 1.0, 3.3
SBC Sistema Basado en el Conocimiento. 3.1
SBVR *Semantics of Business Vocabulary and Business Rules*. 3.3, 3.4
SCOR *Supply-Chain Operations Reference-Model*. 3.4
SDOW *Social Data On the Web*. 3.2
SGML *Standard Generalized Markup Language*. 3.1
SIOC *Semantically-Interlinked Online Communities*. 3.1, 3.4
SITC *Standard International Trade Classification*. 2.10, 5.2
SKOS *Simple Knowledge Organization System*. 3.1, 3.2, 4.4, 5.1–5.3, C.1
SKOS-XL *SKOS eXtension for Labels*. 4.4, 5.1, 5.2
SNORQL *A SPARQL Explorer*. 1.4, 6.2, 6.4, 6.6
SOA *Service Oriented Architecture*. 2.9
SOAP *Simple Object Access Protocol*. 1.5, 2.9, 3.1, 3.2, 4.4, 4.6
Solr *Ultra-fast Search Library and Server*. 5.2, 6.2, 6.4, 7.2, 8.2
SPARQL *Query Language for RDF*. 3.1–3.4, 4.1, 4.2, 4.4, 4.6, 4.7, 4.9, 5.1–5.3, 6.2, 6.4, 6.6, 7.1–7.3, 8.1, 8.3, C.1

SPARUL *SPARQL with Updates*. 3.3, 4.9
SQL *Structured Query Language*. 4.5, 8.3
STORK *Secure idenTity acrOss euRope linKed*. 2.8
SWRL *Semantic Web Rule Language*. 3.1

TARIC *Integrated Tariff of the European Communities*. 2.10
TED *Tenders Electronic Daily*. 2.2, 2.6, 2.9, 3.4, 5.3, 7.1
TGSS Tesorería General de la Seguridad Social. 2.1
TRIOO *Keeping the Semantics of Data Safe and Sound into Object-Oriented Software*. 4.7
Turtle *Terse RDF Triple Language*. 3.1, 3.2

UBL *Universal Business Language*. 2.9
UNE Una Norma Española. 6.5
UNESCO *United Nations Educational, Scientific and Cultural Organization*. 2.10
UNSPSC *United Nations Standard Products and Services Code*. 2.10, 3.4
URI *Uniform Resource Identifier*. 1.6, 3.1, 3.2, 4.1, 4.4, 4.6, 5.1–5.3, 6.4, 6.6, 7.1, C.1
URL *Uniform Resource Locator*. 4.1, 5.3, 6.4, A.0, C.1

voID *Vocabulary of Interlinked Datasets*. 3.2, 4.4, 5.1–5.3, C.1

W3C *World Wide Web Consortium*. 1.5, 3.1–3.4, 4.1, 4.4, 4.10, 5.1, 5.2, 8.2
WADL *Web Application Description Language*. 6.4
WCAG *Web Content Accessibility Guidelines*. 3.2
WESO Grupo de Investigación de Web Semántica Oviedo. 1.1, 1.5, A.0, B.0
WIQA *Web Information Quality Assessment Framework*. 7.1
WSDL *Web Services Description Language*. 2.9, 3.1
WSML *Web Service Modeling Language*. 3.1, 3.4
WSMO *Web Service Modeling Ontology*. 3.1

XBRL *eXtensible Business Reporting Language*. 2.9, 3.4
XHTML *eXtensible HyperText Markup Language*. 3.1, 3.2
XML *eXtensible Markup Language*. 2.9, 3.1, 3.3, 3.4, 4.1, 4.4, 4.5, 6.4
XML Schema *XML vocabulary with the purpose of defining a class of XML documents*. 2.9, 3.1, 4.4, 5.1–5.3, 7.1
XP *Extreme programming*. 6.5
XPath *The XML Path Language*. 3.1, 4.5
XQuery *XQuery is a standardized language for combining documents, databases, Web pages and almost anything else*. 4.5
XSL *Extensible Stylesheet Language*. 3.1
XSLT *Extensible Stylesheet Language Transformations*. 3.1

Referencias

[1] Suzanne Acar, Jose Manuel Alonso, y Kevin Novak. Improving Access to Government through Better Use of the Web. W3C note, W3C, 2009. http://www.w3.org/TR/2009/NOTE-egov-improving-20090512/.

[2] Ben Adida y Mark Birbeck. RDFa Primer, Bridging the Human and Data Webs. W3C working group note, W3C, 2008. http://www.w3.org/TR/xhtml-rdfa-primer/.

[3] Keith Alexander, Richard Cyganiak, Michael Hausenblas, y Jun Zhao. Describing Linked Datasets with the VoID Vocabulary. W3C interest group note, W3C, 2011. http://www.w3.org/TR/void/.

[4] Giner Alor-Hernández, Juan Miguel Gómez Berbís, y Alejandro Rodríguez González et al. HYDRA: A Middleware-Oriented Integrated Architecture for e-Procurement in Supply Chains. *T. Computational Collective Intelligence*, 1:1–20, 2010.

[5] Deepak Alur, John Crupi, y Dan Malks. *Core J2EE Patterns: Best Practices and Design Strategies*. Sun Microsystems, 2003.

[6] Andrea Matteini Andreas Schultz, Robert Isele, Christian Bizer, y Christian Becker. *LDIF - Linked Data Integration Framework*, páginas 1–4. 2011.

[7] Anupriya Ankolekar, Markus Krötzsch, Thanh Tran, y Denny Vrandecic. The Two Cultures: Mashing up Web 2.0 and the Semantic Web. En *In: Proceedings of the 16th International Conference on World Wide Web. 2007 MAY 7-8*. ACM Press, 2007.

[8] Samur Araujo, Jan Hidders, Daniel Schwabe, y Arjen P De Vries. SERIMI – Resource Description Similarity , RDF Instance Matching and Interlinking. *Work*, 2011.

[9] Michael Armbrust, Armando Fox, Rean Griffith, Anthony D. Joseph, Randy H. Katz, Andrew Konwinski, Gunho Lee, David A. Patterson, Ariel Rabkin, y Matei Zaharia. Above the Clouds: A Berkeley View of Cloud Computing. Informe técnico, University of California, Berkeley, 2009.

[10] Emilio Rubiera Azcona. Recursos digitales de lexicografía especializada: Avances en Nuevas Tecnologías. Informe técnico, Facultad de Filología, Universidad de Oviedo, Sep 2007.

[11] Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, y Peter F. Patel-Schneider, editores. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, 2003. ISBN 0-521-78176-0.

[12] José Manuel Barroso y Mario Monti. New strategy for the single market. at the service of europe's economy and society, mayo 2010. http://ec.europa.eu/bepa/pdf/monti_report_final_10_05_2010_en.pdf.

[13] Mark Bartel, John Boyer, Barb Fox, Brian LaMacchia, y Ed Simon. XML Signature Syntax and Processing (Second Edition). W3C recommendation, W3C, 2008. http://www.w3.org/TR/xmldsig-core/.

[14] Florian Bauer y Martin Kaltenböck. *Linked Open Data: The Essentials: A Quick Start Guide for Decision Makers*, tomo 1. 2012. ISBN 978-3902796059.

[15] Kent Beck. *Extreme Programming Explained: Embrace Change*. Addison-Wesley Professional, 1999.

[16] Dave Beckett. RDF/XML Syntax Specification (Revised). W3C recommendation, W3C, 2008. http://www.w3.org/TR/rdf-syntax-grammar/.

[17] Dave Beckett, Tim Berners-Lee, y Eric Prud'hommeaux. Turtle-Terse RDF Triple Language. W3C working draft, W3C, 2011. http://www.w3.org/TR/turtle/.

[18] Richard Benjamins y Dieter Fensel. The Ontological Engineering Initiative-KA, 1998.

[19] Richard Benjamins, Dieter Fensel, y Asunción Gómez Pérez. Knowledge Management through Ontologies. En *Proceedings of the Second International Conference on Practical Aspects Knowledge Management*, páginas 29–30. 1998.

[20] Daniel Bennett y Adam Harvey. Publishing Open Government Data. W3C working draft, W3C, 2009. http://www.w3.org/TR/gov-data/.

[21] Bettina Berendt, Andreas Hotho, y Gerd Stumme. Towards Semantic Web Mining. En Ian Horrocks y James Hendler, editores, *The Semantic Web — ISWC 2002*, tomo 2342 de *Lecture Notes in Computer Science*, páginas 264–278. Springer Berlin / Heidelberg, 2002. ISBN 978-3-540-43760-4. 10.1007/3-540-48005-6_21.

[22] Tim Berners-Lee. Cool URIs don't change, 1998.

[23] Tim Berners-Lee. Linked data, jul 2006.

[24] Tim Berners-Lee. Bag of Crisps at Gov 2.0 Expo, Washington, nov 2010. http://techpresident.com/blog-entry/tim-berners-lee-and-bag-crisps.

[25] Tim Berners-Lee, Yuhsin Chen, Lydia Chilton, Dan Connolly, Ruth Dhanaraj, James Hollenbach, Adam Lerer, y David Sheets. Tabulator: Exploring and Analyzing linked data on the Semantic Web. En *Proceedings of the 3rd International Semantic Web User Interaction*. 2006.

[26] Tim Berners-Lee y Dan Connolly. Notation3 (N3): A readable RDF syntax. W3C team submission, W3C, 2011. http://www.w3.org/TeamSubmission/n3/.

[27] Tim Berners-Lee, Roy Fielding, y Larry Masinter. RFC 3986, Uniform Resource Identifier (URI): Generic Syntax. Request For Comments (RFC), 2005.

[28] Tim Berners-Lee, Roy Fielding, y Larry et al. Masinter. RFC 2616, Hypertext Transfer Protocol – HTTP/1.1, 1999.

[29] Tim Berners-Lee, Mark Fischetti, y Michael Dertouzos. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its Inventor*. Harper San Francisco, 1999. ISBN 1402842937.

[30] Tim Berners-Lee, Wendy Hall, James Hendler, Kieronx O'Hara, Nigel Shadbolt, y Daniel Weitzner. A Framework for Web Science. *Foundations and Trends in Web Science*, 1(1):1–130, 2006.

[31] Tim Berners-Lee, James Hendler, y Ora Lassila. The Semantic Web. *Scientific American*, 284(5):34–43, May 2001.

[32] Abraham Bernstein, Christoph Kiefer, y Markus Stocker. OptARQ: A SPARQL Optimization Approach based on Triple Pattern Selectivity Estimation. Informe técnico, University of Zurich, Department of, 2007.

[33] Diego Berrueta, Sergio Fernández, y Iván Frade. Cooking HTTP content negotiation with Vapour. En *In Proceedings of 4th workshop on Scripting for the Semantic Web 2008 (SFSW2008). co-located with ESWC2008*. 2008.

[34] Diego Berrueta, José Emilio Labra, y Luis Polo. Searching over Public Administration Legal Documents Using Ontologies. En *Proc. of JCKBSE 2006*, páginas 167–175. 2006. ISBN 978-1-58603-640-9.

[35] Diego Berrueta, Jon Phipps, Alistair Miles, Thomas Baker, y Ralph Swick. Best Practice Recipes for Publishing RDF Vocabularies. W3c working group note, aug 2008.

[36] Christian Bizer. *Quality-Driven Information Filtering in the Context of Web-Based Information Systems*. Tesis Doctoral, 2007.

[37] Christian Bizer y Richard Cyganiak. *D2R Server - Publishing Relational Databases on the Semantic Web*. Citeseer, 2006.

[38] Christian Bizer y Richard Cyganiak. Quality-driven information filtering using the WIQA policy framework. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(1):1–10, jan 2009. ISSN 15708268.

[39] Christian Bizer, Richard Cyganiak, y Tom Heath. How to publish Linked Data on the Web, 2007.

[40] Christian Bizer, Richard Cyganiak, Oliver Maresch, y Tobias Gauss. The WIQA Browser V0.3. Information Quality Assessment enabled Semantic Web Browser, diciembre 2011. http://virtuoso.openlinksw.com/.

[41] Christian Bizer, Andreas Harth Hannes M¨hleisen, y Steffen Stadtm¨ller. Web Data Commons: Extracting Structured Data from the Common Web Crawl, marzo 2012. http://webdatacommons.org/.

[42] Christian Bizer, Tom Heath, Tim Berners-Lee, y Michael Hausenblas. Linked Data on the Web (LDOW2011), 2011. http://events.linkeddata.org/ldow2011/.

[43] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, y Sebastian Hellmann. DBpedia - A crystallization point for the Web of Data. *Web Semant.*, 7:154–165, September 2009. ISSN 1570-8268.

[44] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, y Sebastian Hellmann. DBpedia : a crystallization point for the web of data. 2009.

[45] Christian Bizer, Julius Volz, Georgi Kobilarov, y Martin Gaedke. Silk - A Link Discovery Framework for the Web of Data. En *18th International World Wide Web Conference*. apr 2009.

[46] Inc. Black Duck Software. Ohloh Committed to Code, febrero 2012.

[47] Roi Blanco, Edward Bortnikov, Flavio Junqueira, Ronny Lempel, Luca Telloli, y Hugo Zaragoza. Caching search engine results over incremental indices. En *Proceeding of the 33rd ACM SIGIR'10 Conference*, SIGIR '10, páginas 82–89. ACM, New York, NY, USA, 2010. ISBN 978-1-4503-0153-4.

[48] Harold Boley, Gary Hallmark, Michael Kifer, Adrian Paschke, Axel Polleres, y Dave Reynolds. RIF Core Dialect. W3C recommendation, W3C, 2010. http://www.w3.org/2005/rules/wiki/Core.

[49] Peter Boncz, Torsten Grust, Maurice Keulen, Stefan Manegold, Jan Rittinger, y Jens Teubner. MonetDB/XQuery: a fast XQuery processor powered by a relational engine. En *In SIGMOD*, páginas 479–490. 2006.

[50] Don Box, David Ehnebuske, y Gopal Kakivaya. Simple Object Access Protocol (SOAP) 1.1. W3C note, W3C, 2000. Http://www.w3.org/TR/2000/NOTE-SOAP-20000508/.

[51] Tim Bray, Jean Paoli, C. M. Sperberg-McQueen, Eve Maler, François Yergeau, y John Cowan. Extensible Markup Language (XML) 1.1 (Second Edition). W3C recommendation, W3C, 2006. Http://www.w3.org/TR/xml11/.

[52] Dan Brickley y R.V. Guha. RDF Vocabulary Description Language 1.0: RDF Schema. W3C recommendation, W3C, 2004.

[53] Jos de Bruijn, Marc Ehrig, Cristina Feier, Francisco Martín-Recuerda, François Scharffe, y Moritz Weiten. Ontology mediation, merging and aligning. En *Semantic Web Technologies*. Wiley, 2006.

[54] Carlos Buil-Aranda, Marcelo Arenas, y Oscar Corcho. Semantics and Optimization of the SPARQL 1.1 Federation Extension. En *Proc. of the 8th ESWC'11*, ESWC'11. 2011.

[55] A.S. Butt, S. Khan, y K. Latif. Comparative evaluation of native RDF stores. En *Emerging Technologies (ICET), 2010 6th International Conference on*, páginas 321–326. oct. 2010.

[56] Jean-Paul Calbimonte, Hoyoung Jeung, Oscar Corcho, y Karl Aberer. Semantic sensor data search in a large-scale federated sensor network. En *SSN*, páginas 14–29. 2011.

[57] Ben Caldwell, Michael Cooper, Loretta Guarino, y Gregg Vanderheiden. Web Content Accessibility Guidelines (WCAG) 2.0. W3C recommendation, W3C, 2008. http://www.w3.org/TR/WCAG20/.

[58] Jeremy Carroll, Christian Bizer, Pat Hayes, y Patrick Stickler. Named graphs, provenance and trust. En *In WWW '05: Proceedings of the 14th international conference on World Wide Web*, páginas 613–622. ACM, 2005.

[59] H. Chen y T. Ng. An Algorithmic Approach to Concept Exploration in a Large Knowledge Network (automatic thesaurus consultation): Symbolic Branch-and-Bound search vs. connectionist Hopfield net activation. *J. Am. Soc. Inf. Sci.*, 46(5):348–369, 1995. ISSN 0002-8231.

[60] Peter Pin-Shan Chen. The entity-relationship model, toward a unified view of data. *ACM Trans. Database Syst.*, 1(1):9–36, marzo 1976. ISSN 0362-5915.

[61] Roberto Chinnici, Jean-Jacques Moreau, Arthur Ryman, y Sanjiva Weerawarana. Web Services Description Language (WSDL) version 2.0 part 1: Core language. W3C recommendation, W3C, 2007. Http://www.w3.org/TR/wsdl.

[62] Erik Christensen, Francisco Curbera, Greg Meredith, y Sanjiva Weerawarana. Web Services Description Language (WSDL) 1.1. W3C note, W3C, 2001. Http://www.w3.org/TR/wsdl.

[63] Francisco Cifuentes-Silva, Jose María Alvarez, y José Emilio Labra. An architecture and process of implantation for Linked Data environments. 2011.

[64] Francisco Adolfo Cifuentes-Silva, Christian Sifaqui, y José Emilio Labra Gayo. Towards an architecture and adoption process for linked data technologies in open government contexts: a case study for the Library of Congress of Chile. En *I-SEMANTICS*, páginas 79–86. 2011.

[65] James Clark, editor. *XSL Transformations (XSLT)*. W3C, 1999.

[66] James Clark y Steve DeRose. XML Path Language (XPath) Version 1.0. W3C recommendation, W3C, 1999. Http://www.w3.org/TR/xpath.

[67] European Commision. Reglamento (CE) no 213/2008 de la Comisión, de 28 de noviembre de 2007, noviembre 2007. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32008R0213:es:NOT.

[68] European Commision. ICT for Government and Public Services. Action Plan 2011 - 2015., diciembre 2010. http://ec.europa.eu/information_society/activities/egovernment/policy/index_en.htm.

[69] European Commision. ePractice.eu the professional community of eGovernment, agosto 2011. www.epractice.eu.

[70] European Commission. Manual del vocabulario común de contratos públicos, enero 2008. http://simap.europa.eu/codes-and-nomenclatures/codes-cpv/cpv_2008_guide_es.pdf.

[71] European Commission. e-Notices, diciembre 2011. http://simap.europa.eu/enotices.

[72] European Commission. eCERTIS, diciembre 2011. http://ec.europa.eu/markt/ecertis.

[73] European Commission. Pan-European Public Procurement Online (PEPPOL, diciembre 2011. http://www.peppol.eu.

[74] European Commission. Standard forms for public procurement, diciembre 2011. http://simap.europa.eu/buyer/forms-standard/index_en.htm.

[75] European Commission. Digital Agenda: Turning government data into gold., diciembre 2012. http://europa.eu/rapid/pressReleasesAction.do?reference=IP/11/1524&format=HTML&aged=0&language=en&guiLanguage=en.

[76] European Commmision. Plan de acción para la aplicación del marco jurídico de la contratación pública electrónica. SEC(2004)-1639, diciembre 2004. http://ec.europa.eu/internal_market/publicprocurement/docs/eprocurement/actionplan/actionplan_es.pdf.

[77] European Commmision. Evaluation of the 2004 Action Plan for Electronic Public Procurement, diciembre 2010. http://ec.europa.eu/internal_market/consultations/docs/2010/e-procurement/evaluation-report_en.pdf.

[78] Oscar Corcho y Asunción Gómez Pérez. Integrating E-Commerce Standards and Initiatives in a Multi-Layered Ontology.

[79] CTO Council's cross Government Enterprise Architecture. Designing URI Sets for the UK Public Sector. Draft, UK Government, 2010.

[80] Fundación CTIC. Centro Tecnológico de las Tecnologías de la Información y Comunicaciones, mayo 2005. http://www.fundacionctic.org.

[81] Fundación CTIC y Universidad de Oviedo. Use Case: Enhancing Web Searches within the Principality of Asturias, 2007. http://www.w3.org/2001/sw/sweo/public/UseCases/CTIC/.

[82] Richard Cyganiak y Anja Jentzsch. The Linking Open Data cloud diagram, noviembre 2011. http://richard.cyganiak.de/2007/10/lod/.

[83] Richard Cyganiak, Holger Stenzhorn, Renaud Delbru, Stefan Decker, y Giovanni Tummarello. Semantic Sitemaps: Efficient and flexible access to datasets on the Semantic Web. En *In Proceedings of the 5th European Semantic Web Conference*, páginas 690–704. 2008.

[84] Aba-Sah Dadzie y Matthew Rowe. Approaches to visualising Linked Data: A survey. *Semantic Web*, 2(2):89–124, 2011.

[85] Aba-Sah Dadzie y Matthew Rowe. Approaches to visualising Linked Data: A survey. *Semantic Web*, 2(2):89–124, 2011.

[86] TaskForce Linking Open Data DataSets. Guidelines for Collecting Metadata on Linked Datasets in CKAN. W3c working group note, dec 2011.

[87] Dave Reynolds. The Web Service Modeling Ontology (WSMO). Informe técnico, Epimorphics Ltd, 2010.

[88] Ministerio de Administraciones Públicas. Ley 6/1997, de 14 de abril, de Organización y Funcionamiento de la Administración General del Estado., abril 1997. http://noticias.juridicas.com/base_datos/Admin/l6-1997.html.

[89] Ministerio de Administraciones Públicas del Gobierno de España. Métrica v.3, febrero 2012.

[90] Jos de Bruijn, Marc Ehrig, Cristina Feier, Francisco Martín-Recuerda, François Scharffe, y Moritz Weiten. Ontology mediation, merging and aligning. En *Semantic Web Technologies*. Wiley, UK, 2006.

[91] Jos de Bruijn, Holger Lausen, Axel Polleres, y Dieter Fensel. The web service modeling language: An overview. En *Proc. of the European Semantic Web Conference*. 2006.

[92] Plan Regional de Ciencia Tecnología e Innovación 2006-2009 del Principado de Asturias. PRAVIA-Plataforma de Recursos de Acceso Virtual a la Información del Sector Asegurador (código IE05-172), enero 2007.

[93] Plan Regional de Ciencia Tecnología e Innovación 2006-2009 del Principado de Asturias. SAITA, junio 2009.

[94] Ministerio de Industria Turismo y Comercio. Factura Electrónica, noviembre 2011. http://www.facturae.es/.

[95] 7º Programa Marco de la Unión Europea. ONTORULE-ONTOlogies meet business RULEs (código 231875), enero 2009. http://ontorule-project.eu.

[96] Ayuntamiento de Oviedo. Ejemplo de Perfil de Contrante, noviembre 2011. http://www.oviedo.es/index.php/es/el-ayuntamiento/perfil-de-contratante.

[97] Ministerio de Presidencia. Ley 37/2007, de 16 de noviembre, sobre reutilización de la información del sector público., noviembre 2007. http://www.boe.es/aeboe/consultas/bases_datos/doc.php?id=BOE-A-2007-19814.

[98] Ministerio de Presidencia. Real Decreto 1495/2011, de 24 de octubre, por el que se desarrolla la Ley 37/2007, de 16 de noviembre, sobre reutilización de la información del sector público, para el ámbito del sector público estatal., noviembre 2011. http://www.boe.es/diario_boe/txt.php?id=BOE-A-2011-17560.

[99] Ayuntamiento de Zaragoza. Datos que me gustaría reutilizar, diciembre 2011. http://zaragoza.uservoice.com/.

[100] Jeffrey Dean y Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. páginas 137–150.

[101] Stefan Decker, Dieter Fensel, Frank van Harmelen, Ian Horrocks, Sergey Melnik, Michel C. A. Klein, y Jeen Broekstra. Knowledge Representation on the Web. En *Description Logics*, páginas 89–97. 2000.

[102] Dirección General del Patrimonio del estado. Plataforma de Contratación del Estado, noviembre 2011. http://contrataciondelestado.es/wps/portal/plataforma.

[103] Helena Deus. A survey of Linked Data Quality Metrics, febrero 2012.

[104] European Commission Internal Market Directorate-General. Study on the evaluation of the Action Plan for the implementation of the legal framework for electronic procurement (Phase II). Analysis, assessment and recommendation, julio 2010. http://ec.europa.eu/internal_market/consultations/docs/2010/e-procurement/siemens-study_en.pdf.

[105] Leigh Dodds y Ian Davis. Linked Data Patterns. A pattern catalogue for modelling, publishing, and consuming Linked Data, agosto 2011. http://patterns.dataincubator.org/book/.

[106] Angel Esteban-Gil, Francisco García Sánchez, Rafael Valencia-García, y Jesualdo Tomás Fernández-Breis. SocialBROKER: A collaborative social space for gathering semantically-enhanced financial information. *Expert Syst. Appl.*, 39(10):9715–9722, 2012.

[107] DOUE L 313 de 28 de Noviembre 2009 European Commision. Reglamento (CE) no 1150/2009 de la Comisión, de 10 de noviembre de 2009, noviembre 2009. http://eur-lex.europa.eu/JOHtml.do?uri=OJ:L:2009:313:SOM:ES:HTML.

[108] DOUE L 314 de 1 de Diciembre 2009 European Commision. Reglamento (CE) no 1177/2009 de la Comisión, de 30 de noviembre de 2009, noviembre 2009. http://eur-lex.europa.eu/JOHtml.do?uri=OJ:L:2009:314:SOM:ES:HTML.

[109] DOUE L 74/1 de 15 de Marzo 2008 European Commision. Reglamento (ce) nº 2195/2002 de la comisión, de 28 de noviembre de 2007, noviembre 2007. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:074:0001:0375:ES:PDF.

[110] Unit C4 – Economic Dimension of Public Procurement; E-Procurement European Commision. Consultation on the Green paper on expanding the use of e-Procurement in the EU, abril 2010. http://ec.europa.eu/internal_market/consultations/2010/e-procurement_en.htm.

[111] Directorate-General for Informatics European Commission. The eProcurement Map. a map of activities having an impact on the development of european interoperable eprocurement solutions, agosto 2011. http://www.epractice.eu/en/library/5319079.

[112] Gunther Eysenbach. Can Tweets Predict Citations? Metrics of Social Impact Based on Twitter and Correlation with Traditional Metrics of Scientific Impact, 2011. http://www.jmir.org/2011/4/e123/.

[113] David C. Fallside y Priscilla Walmsley. XML Schema Part 0: Primer Second Edition. W3C recommendation, W3C, 2004. Http://www.w3.org/TR/xmlschema-0/.

[114] Dieter Fensel, Ying Ding, y Borys Omelayenko. Product Data Integration in B2B E-commerce. *IEEE-Intelligent E-Business*, páginas 54–59, 2001.

[115] Dieter Fensel, Frank Van Harmelen, Ian Horrocks, y Deborah L. Mcguinness. OIL: An Ontology Infrastructure for the Semantic Web. *IEEE Intelligent Systems*, páginas 38–45, 2001.

[116] Sergio Fernández, Diego Berrueta, Miguel García, y José Emilio Labra. Trioo - Keeping the Semantics of Data Safe and Sound into Object-oriented Software. En *ICSOFT (1)*, páginas 311–320. 2010.

[117] Javier D. Fernández, Miguel A. Martínez-Prieto, Claudio Gutierrez, y Claudio Gutierrez. Binary RDF Representation for Publication and Exchange (HDT). W3C member submission, W3C, 2011. http://www.w3.org/Submission/HDT/.

[118] Sergio Fernández. RDFohloh, a RDF wrapper of Ohloh. En *In 1st workshop on Social Data on the Web (SDoW2008), co-located with ISWC2008*. 2008.

[119] Roy T. Fielding. *Architectural styles and the design of network-based software architectures*. Tesis Doctoral, 2000.

[120] Apache Foundation. Apache Solr, febrero 2012.

[121] Open Data Foundation. Open Data Foundation. http://www.opendatafoundation.org/.

[122] Open Knowledge Foundation. CKAN - the Open Source Data Hub Software, diciembre 2011. http://ckan.org/.

[123] Open Knowledge Foundation. the Data Hub — The easy way to get, use and share data, noviembre 2011. http://thedatahub.org/.

[124] Martin Fowler. *Refactoring: Improving the Design of Existing Code*. Addison-Wesley, Boston, MA, USA, 1999. ISBN 0-201-48567-2.

[125] MS. Fox, M. Barbuceanu, y M. Grüninger. An Organisation Ontology for Enterprise Modelling: Preliminary Concepts for Linking Structure and Behaviour. En *Computers in Industry*, páginas 123–134. 1995.

[126] LOD2 European Project FP7. LOD2 – Creating Knowledge out of Interlinked Data., agosto 2010. http://lod2.eu.

[127] PlanetData European Project FP7. PlanetData. A European Network of Excellence on Large-Scale Data Management, febrero 2012. http://planet-data.eu/.

[128] Gouvernement Français. La politique de la France en matière d'ouverture des données publiques., diciembre 2011. http://www.data.gouv.fr/Articles/La-politique-de-la-France-en-matiere-d-ouverture-des-donnees-publiques.

[129] Erich Gamma, Richard Helm, Ralph Johnson, y John Vlissides. *Design Patterns*. Addison-Wesley Professional, January 1995. ISBN 0201633612.

[130] Ángel García-Crespo, Juan Miguel Gómez Berbís, Ricardo Colomo Palacios, y Francisco García Sánchez. Digital libraries and Web 3.0. The CallimachusDL approach. *Computers in Human Behavior*, 27(4):1424–1430, 2011.

[131] Ignacio García-Manotas, Eduardo Lupiani-Ruiz, Francisco García Sánchez, y Rafael Valencia-García. Populating Knowledge Based Decision Support Systems. *IJDSST*, 2(1):1–20, 2010.

[132] S.L. Gateway Strategic Consultancy Services. Euroalert.Net. Alertas de concursos públicos y subvenciones para PYMES, 1998. http://euroalert.net/.

[133] S.L. Gateway Strategic Consultancy Services. Gateway Servicios de Consultoría Estratégica, 1998. http://gateway-scs.es/.

[134] Jose Emilio Labra Gayo, Juan Manuel Cueva Lovelle, Raúl Izquierdo Castanedo, Aquilino Adolfo Juan Fuente, , María Cándida Luengo Díez, y Francisco Ortín Soler. *Intérpretes y Diseño de Lenguajes de Programación*. Servicio de Publicaciones de la Universidad de Oviedo, apr 2004.

[135] Adrian Giurca y Emilian Pascalau. JSON Rules. En Grzegorz J. Nalepa y Joachim Baumeister, editores, *KESE*, tomo 425 de *CEUR Workshop Proceedings*. CEUR-WS.org, 2008.

[136] Juan Miguel Gómez, Francisco García Sánchez, Rafael Valencia-García, Ioan Toma, y Carlos García Moreno. SONAR: A Semantically Empowered Financial Search Engine. En *IWINAC (1)*, páginas 405–414. 2009.

[137] Asunción Gomez-Perez, Mariano Fernandez-Lopez, y Oscar Corcho, editores. *Ontological Engineering: With Examples from the Areas of Knowledge Management, E-Commerce and the Semantic Web (2nd edition)*. Springer-Verlag, Heidelberg, 2007. ISBN 1846283965.

[138] Zhiguo Gong, Chan Wa Cheang, y Leong Hou U. Multi-term Web Query Expansion Using WordNet. *Database and expert systems applications 17th international conference DEXA 2006 Kraków Poland September 48 2006 proceedings*, página 379, 2006.

[139] S. Gouws, G-J. Van Rooyen, y HE. Engelbrecht. Measuring Conceptual Similarity by Spreading Activation over Wikipedia's Hyperlink Structure. En *Proceedings of the 2nd Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, páginas 46–54. Beijing, China, August 2010.

[140] UK Government. Opening Up Government-Ideas, diciembre 2011. http://data.gov.uk/ideas.

[141] UK Government. Openning Up Government, diciembre 2011. http://data.gov.uk/.

[142] Bernardo Cuenca Grau. Combination and Integration of Ontologies on the Semantic Web. Informe técnico, Universidad de Valencia, jul 2005. PhD Dissertation.

[143] Memento Development Group y Library of Congress. Memento Adding Time to the Web, noviembre 2011. http://www.mementoweb.org/.

[144] W3C Semantic Web Interest Group. SweoIG/TaskForces/CommunityProjects/LinkingOpenData, noviembre 2011. http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData.

[145] Thomas R. Gruber. A translation approach to portable ontology specifications. *Knowl. Acquis.*, 5(2):199–220, June 1993. ISSN 1042-8143.

[146] Thomas R. Gruber. Towards Principles for the Design of Ontologies Used for Knowledge Sharing. En N. Guarino y R. Poli, editores, *Formal Ontology in Conceptual Analysis and Knowledge Representation*. Kluwer Academic Publishers, Deventer, The Netherlands, 1993.

[147] Rosa Guerequeta y Antonio Vallecillo. *Técnicas de Diseño de Algoritmos*. Servicio de Publicaciones de la Universidad de Málaga, 1998. ISBN 84-7496-666-3.

[148] Volker Haarslev y Ralf Müller. RACER System Description. En Rajeev Goré, Alexander Leitsch, y Tobias Nipkow, editores, *Automated Reasoning*, tomo 2083 de *Lecture Notes in Computer Science*, páginas 701–705. Springer Berlin / Heidelberg, 2001.

[149] Andreas Harth, Aidan Hogan, Stephan Kotoulas, y Jacopo Urbani. Scalable integration and processing of linked data. En *Proc. of the 20th international conference companion on World wide web*, WWW '11, páginas 295–296. ACM, New York, NY, USA, 2011. ISBN 978-1-4503-0637-9.

[150] Olaf Hartig, Christian Bizer, y J.C. Freytag. Executing SPARQL queries over the Web of Linked Data. En *In 8th International Semantic Web Conference (ISWC)*. 2009.

[151] Olaf Hartig, Andreas Harth, y Juan Sequeda. First International Workshop on Consuming Linked Data (COLD2010), noviembre 2010. http://people.aifb.kit.edu/aha/2010/cold/.

[152] Olaf Hartig y Jun Zhao. Publishing and Consuming Provenance Metadata on the Web of Linked Data. En *IPAW*, páginas 78–90. 2010.

[153] Erik Hatcher y Otis Gospodnetic. *Lucene in Action (In Action series)*. Manning Publications Co., Greenwich, CT, USA, 2004. ISBN 1932394281.

[154] Michael Hausenblas, Robert Grossman, Andreas Harth, y Philippe Cudré-Mauroux. Large-Scale Linked Data Processing: Cloud Computing to the Rescue? En *2nd International Conference on Cloud Computing and Services Science*. 2012.

[155] Michael Hausenblas, Boris Villazón-Terrazas, y Bernadette Hyland. GLD Life cycle. W3C government linked data group, W3C, 2011. http://www.w3.org/2011/gld/wiki/GLD_Life_cycle.

[156] Patrick Hayes, editor. *RDF Semantics*. W3C Recommendation. World Wide Web Consortium, febrero 2004.

[157] Tom Heath y Christian Bizer. *Linked Data: Evolving the Web into a Global Data Space*, tomo 1. Morgan & Claypool, 2011.

[158] James Hendler y Deborah L. McGuinness. The DARPA Agent Markup Language. *IEEE Intelligent Systems*, 15(6):67–73, 2000.

[159] Martin Hepp. eClassOWL: A Fully-Fledged Products and Services Ontology in OWL. En *Poster Proceedings of the 4th International Semantic Web Conference (ISWC2005)*. Galway, Ireland, 2005.

[160] Martin Hepp. A Methodology for Deriving OWL Ontologies from Products adn Services Categorization Standards. (1):72–99, 2006.

[161] Martin Hepp. The True Complexity of Product Representacion in the Semantic Web. En *Proceedings of the 14th European Conference on Information System (ECIS 006)*. Gothenburg, Sweden, 2006.

[162] Alan Hevner y Samir Chatterjee. *Design Research in Information Systems: Theory and Practice*. Springer Publishing Company, Incorporated, 1st edición, 2010. ISBN 1441956522, 9781441956521.

[163] Ian Hickson. HTML5-A vocabulary and associated APIs for HTML and XHTML. W3C working draft, W3C, 2011. http://www.w3.org/TR/html5/.

[164] Pascal Hitzler, Markus Krötzsch, Bijan Parsia, Peter F. Patel-Schneider, y Sebastian Rudolph. OWL 2 Web Ontology Language: Primer. Informe técnico, 2009. http://www.w3.org/TR/owl2-primer/.

[165] Aidan Hogan, Andreas Harth, y Axel Polleres. Scalable Authoritative OWL Reasoning for the Web. *International Journal on Semantic Web and Information Systems*, 5(2):49–90, 2009.

[166] Aidan Hogan, Andreas Harth, Jürgen Umbrich, Sheila Kinsella, Axel Polleres, y Stefan Decker. Searching and Browsing Linked Data with SWSE: The Semantic Web Search Engine. *Journal of Web Semantics (JWS)*, accepted, 2011. In press.

[167] Aidan Hogan, Jeff Z. Pan, Axel Polleres, y Stefan Decker. SAOR: Template Rule Optimisations for Distributed Reasoning over 1 Billion Linked Data Triples. En *International Semantic Web Conference (1)*, páginas 337–353. 2010.

[168] Aidan Hogan, Jurgen Umbrich, Andreas Harth, Richard Cyganiak, Axel Polleres, y Stefan Decker. An empirical survey of Linked Data conformance. *Journal of Web Semantics (to appear)*, 2012.

[169] Ian Horrocks, , Peter F. Patel-Schneider, Harold Boley, et al. SWRL: A Semantic Web Rule Language Combining OWL and RuleML. W3C member submission, W3C, 2004. Http://www.w3.org/Submission/SWRL/.

[170] Ian Horrocks y Peter F. Patel-Schneider. KR and reasoning on the semantic web: OWL. En John Domingue, Dieter Fensel, y James A. Hendler, editores, *Handbook of Semantic Web Technologies*, capítulo 9, páginas 365–398. Springer, 2011. ISBN 978-3-540-92912-3.

[171] White House. Open Government Initiative, enero 2009. http://www.whitehouse.gov/open.

[172] Lisa Hunemark. Query expansion using search logs and WordNet. Informe técnico, Uppsala University, mar 2010. Master's thesis in Computational Linguistics.

[173] Bernadette Hyland, Boris Villazón Terrazas, y Sarven Capadisli. Cookbook for Open Government Linked Data . W3C task force-government linked data group, W3C, 2011. http://www.w3.org/2011/gld/wiki/Linked_Data_Cookbook.

[174] Takeshi Imamura, Blair Dillaway, y Ed Simon. XML Encryption Syntax and Processing. W3C recommendation, W3C, 2002. http://www.w3.org/TR/xmlenc-core/.

[175] Amazon Inc. Amazon Inc. http://www.amazon.com/.

[176] Facebook Inc. Facebook Linked Data Support, octubre 2011. http://graph.facebook.com.

[177] Freebase Inc. Freebase is an open, Creative Commons licensed graph database, noviembre 2011. http://www.freebase.com/.

[178] Google Inc. Google Refine, a power tool for working with messy data, febrero 2012.

[179] Linkedin Inc. Linkedin Inc. http://www.linkedin.com/.

[180] Open Link Software Inc. Virtuoso Universal Server, diciembre 2011. http://virtuoso.openlinksw.com/.

[181] Oracle Inc. Oracle Spatial Resource Description Framework (RDF). 10g Release 2 (10.2). Informe técnico, Oracle, julio 2005. http://download.oracle.com/otndocs/tech/semantic_web/pdf/rdfrm.pdf.

[182] Talis Inc. Talis Platform, noviembre 2011. http://www.talis.com/platform/.

[183] Linked Data Initiative. Linked Data - Connect Distributed Data across the Web, noviembre 2011. http://linkeddata.org/.

[184] Open Government Data Initiative. Open Government Data Principles., diciembre 2007. https://public.resource.org/8_principles.html.

[185] Ian Jacobs y Norman Walsh. Architecture of the World Wide Web, Volume One. World wide web consortium, recommendation rec-webarch-20041215, W3C, dec 2004.

[186] Jim Jacobs y Alexander Linden. Semantic Web Technologies Take Middleware to Next Level. Tr-109295, Gartner Inc., 2002.

[187] BOE de 2 de agosto de 2011 Núm. 184 Jefatura del Estado. Ley 24/2011, de 1 de agosto, agosto 2011. http://boe.es/boe/dias/2011/08/02/pdfs/BOE-A-2011-13239.pdf.

[188] BOE de 31 de octubre de 2007 Núm. 261 Jefatura del Estado. Ley 30/2007, de 30 de octubre, de Contratos del Sector Público., octubre 2007. http://noticias.juridicas.com/base_datos/Admin/l30-2007.html.

[189] BOE de 9 de agosto de 2010 Núm. 192 Jefatura del Estado. Ley 34/2010, de 5 de agosto, agosto 2010. www.boe.es/boe/dias/2010/08/09/pdfs/BOE-A-2010-12765.pdf.

[190] Hoyoung Jeung, Sofiane Sarni, Ioannis Paparrizos, Saket Sathe, Karl Aberer, Nicholas Dawes, Thanasis G. Papaioannou, y Michael Lehning. Effective Metadata Management in Federated Sensor Networks. En *Proceedings of the 2010 IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing*, SUTC '10, páginas 107–114. IEEE Computer Society, Washington, DC, USA, 2010. ISBN 978-0-7695-4049-8.

[191] Joinup y EEuropean Commission. Share and reuse open source software, semantic assets and other interoperability solutions for public administrations., febrero 2012. http://joinup.ec.europa.eu/.

[192] Marcelo Arenas Jorge Pérez y Claudio Gutierrez. Semantics and complexity of SPARQL. *ACM Trans. Database Syst.*, 34:16:1–16:45, sep 2009. ISSN 0362-5915.

[193] Michael Kifer, Jos de Bruijn, Harold Boley, y Dieter Fensel. A Realistic Architecture for the Semantic Web. En *RuleML*, páginas 17–29. 2005.

[194] Michael Kifer y Georg Lausen. F-logic: a higher-order language for reasoning about objects, inheritance, and scheme. *SIGMOD Rec.*, 18:134–146, June 1989. ISSN 0163-5808.

[195] G. Klyne y JJ. Carroll. Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C recommendation, W3C, 2004. Http://www.w3.org/TR/rdf-concepts/.

[196] Dimitri Konstantas, Jean-Paul Bourrières, Michel Léonard, y Nacer Boudjlida. *Interoperability of Enterprise Software and Applications*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005. ISBN 1846281512.

[197] Naomi Korn y Charles Oppenheim. Licensing Open Data: A Practical Guide, junio 2011. http://discovery.ac.uk/files/pdf/Licensing_Open_Data_A_Practical_Guide.pdf.

[198] Naomi Korn y Charles Oppenheim. Licensing Open Data: A Practical Guide. Informe técnico, jun 2011. http://sca.jiscinvolve.org/wp/files/2010/12/SCA_BP_Open_Licences_Dec10_v1-02.pdf.

[199] Dirk Krafzig, Karl Banke, y Dirk Slama. *Enterprise SOA: Service-Oriented Architecture Best Practices (The Coad Series)*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2004. ISBN 0131465759.

[200] Atif Latif, Muhammad Tanvir Afzal, Denis Helic, Klaus Tochtermann, y Hermann A. Maurer. Discovery and Construction of Authors' Profile from Linked Data (A case study for Open Digital Journal). En *LDOW*. 2010.

[201] Jens Lehmann. DL-Learner: Learning Concepts in Description Logics. *Journal of Machine Learning Research (JMLR)*, 10:2639–2642, 2009.

[202] Joerg Leukel. Automating product classification: The. En *Proceedings of the IADIS International Conference e-Society 2004 (ES 2004)*. Ávila, Spain, 2004.

[203] Joerg Leukel. Standardization of product Ontologies in B2B relationships, On the role of ISO 13584. En *Proceedings of the Tenth Americas Conference on Information Systems*, páginas 4084–4091. New York, United States, 2004.

[204] Joerg Leukel y Gregory Maniatopoulos. A comparative analysis of product classification in public vs. private e-procurement. *The Electronic Journal of e-Government (EJEG)*, (4):201–212, 2005.

[205] Open Knowledge Foundation Limited. Open Knowledge Foundation. http://okfn.org/.

[206] Sergio Fernández López. SWAML, Publicación de Listas de Correo en Web Semántica. Informe técnico, Escuela Universitaria de Ingeniería Técnica Informática Oviedo, December 2006. Bachelor's thesis.

[207] Fadi Maali y Richard Cyganiak. Re-using Cool URIs : Entity Reconciliation Against LOD Hubs. *Library*, página 8, 2011.

[208] Fadi Maali y Richard Cyganiak. RDF Extension for Google Refine, febrero 2012.

[209] Fadi Maali, John Erickson, y Phil Archer. Data Catalog Vocabulary (DCAT). Informe técnico, 05 de abril de 2012. http://www.w3.org/TR/vocab-dcat/.

[210] Alejandro Mallea, Marcelo Arenas, Aidan Hogan, y Axel Polleres. On Blank Nodes. En *International Semantic Web Conference*, páginas 421–437. 2011.

[211] Christopher D. Manning, Prabhakar Raghavan, y Hinrich Schtze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008. ISBN 0521865719, 9780521865715.

[212] Jose Luis Marín y José Emilio Labra. Doing Business by selling free services. En P. Ordóñez et al., editor, *Web 2.0: The Business Model*, part 6, páginas 89–102. Springer, 2009.

[213] Sergei Maslov y S. Redner. Promise and Pitfalls of Extending Google's PageRank Algorithm to Citation Networks, 2009.

[214] W3C Members. Best Practices Discussion Summary. W3C government linked data working group first f2f, W3C, 2011. http://www.w3.org/2011/gld/wiki/Best_Practices_Discussion_Summary.

[215] W3C Members. Government Linked Data Working Group Charter. W3C group, W3C, 2011. http://www.w3.org/2011/gld/charter.

[216] W3C Members. Provenance Working Group. W3C group, W3C, 2011. http://www.w3.org/2011/prov/wiki/Main_Page.

[217] Alistair Miles y Sean Bechhofer. SKOS Simple Knowledge Organisation Systems. W3C recommendation, W3C, 2009. http://www.w3.org/TR/skos-reference/.

[218] BOE de 1 de mayo de 2008 Núm. 105 Ministerio de Economía y Hacienda. Orden EHA/1220/2008 de 30 de abril, mayo 2009. http://www.boe.es/boe/dias/2008/05/01/pdfs/A22260-22264.pdf.

[219] BOE de 10 de junio de 2010 Núm. 141 Ministerio de Economía y Hacienda. Orden EHA/1490/2010, de 28 de mayo, junio 2010. www.boe.es/boe/dias/2010/06/10/pdfs/BOE-A-2010-9186.pdf.

[220] BOE de 15 de mayo de 2009 Núm. 118 Ministerio de Economía y Hacienda. Real Decreto 817/2009, de 8 de mayo, mayo 2009. http://www.boe.es/boe/dias/2009/05/15/pdfs/BOE-A-2009-8053.pdf.

[221] BOE de 16 de noviembre de 2011 Núm. 276 Ministerio de Economía y Hacienda. Real Decreto Legislativo 3/2011, de 14 de noviembre, noviembre 2011. http://www.boe.es/boe/dias/2011/11/16/pdfs/BOE-A-2011-17887.pdf.

[222] BOE de 22 de marzo de 2011 Núm. 69 Ministerio de Economía y Hacienda. Real Decreto 300/2011, de 4 de marzo, marzo 2011. www.boe.es/boe/dias/2011/03/22/pdfs/BOE-A-2011-5192.pdf.

[223] BOE de 29 de diciembre de 2010 Núm. 313 Ministerio de Economía y Hacienda. Orden EHA/3497/2009, de 23 de diciembre, diciembre 2009. www.boe.es/boe/dias/2009/12/29/pdfs/BOE-A-2009-21048.pdf.

[224] BOE de 5 de marzo de 2011 Núm. 55 Ministerio de Economía y Hacienda. Ley 2/2011, de 4 de marzo, de Economía Sostenible, marzo 2011. www.boe.es/boe/dias/2011/03/05/pdfs/BOE-A-2011-4117.pdf.

[225] BOE de 8 de marzo de 2010 Núm. 58 Ministerio de Economía y Hacienda. Resolución de 3 de marzo de 2010, marzo 2010. http://www.boe.es/boe/dias/2010/03/08/pdfs/BOE-A-2010-3807.pdf.

[226] Turismo y Comercio Ministerio de Industria. MyMobileWeb (código TSI-020400-2010-118), enero 2008. http://mymobileweb.morfeo-project.org/.

[227] Turismo y Comercio Ministerio de Industria. EzWeb (código FIT-340503-2007-2), agosto 2009. http://ezweb.morfeo-project.org/.

[228] Turismo y Comercio Ministerio de Industria y Plan Avanza. PRIMA-Plataforma de Recursos de Información y Movilidad para el sector Asegurador (código TSI-020302-2009-17), agosto 2009. http://prima.morfeo-project.org/.

[229] Turismo y Comercio Ministerio de Industria y Plan Avanza2. 10ders Information Services(código TSI-020100-2010-919), noviembre 2010. http://rd.10ders.net/wiki/Main_Page.

[230] Luc Moreau y Paolo Missier. The PROV Data Model and Abstract Syntax Notation. W3C working draft, W3C, 2011. http://www.w3.org/TR/2011/WD-prov-dm-20111018/.

[231] Boris Motik. Kaon2 - scalable reasoning over ontologies with large data sets. *ERCIM News*, 2008(72), 2008.

[232] Boris Motik. Semantics and Reasoning Algorithms for a Faithful Integration of Description Logics and Rules. En *RR*, página 12. 2008.

[233] Boris Motik, Rob Shearer, y Ian Horrocks. Hypertableau Reasoning for Description Logics. *Journal of Artificial Intelligence Research*, 36:165–228, 2009.

[234] Martin Nally y Steve Speicher. Toward a Basic Profile for Linked Data., diciembre 2011. http://www.ibm.com/developerworks/rational/library/basic-profile-linked-data/index.html.

[235] Martin Nally, Steve Speicher, John Arwe, y Arnaud Le Hors. Linked Data Basic Profile 1.0. W3C Member Submission, W3C, 2012. http://www.w3.org/Submission/2012/02/.

[236] NETFINGER, NETJUICE, y SENDA. EXIS-TI Sistemas de Información y Consultoría, 2003. http://www.exis-ti.com/.

[237] Jian-Yun Nie. Query Expansion and Query Translation as Logical Inference. *J. Am. Soc. Inf. Sci. Technol.*, 54(4):335–346, 2003. ISSN 1532-2882.

[238] Mark Nottingham y Robert Sayre. RFC 4287, The Atom Syndication Format, 2005.

[239] Fridman Natasha Noy y M. A. Musen. PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment. En *Proc. 17th Intl. Conf. on Artificial Intelligence (AAAI 2000)*, páginas 450–455. Austin, Texas, USA, July–August 2000.

[240] Alvaro Graves Olye Rickson. LOD Data Quality Vocabulary: LODQ. Working Draft, Linking Open Government Data, 2011. http://logd.tw.rpi.edu/lodq.

[241] Borys Omelayenko y Dieter Fensel. An Analysis of B2B Catalogue Integration Problems. En *Proceedings of the International Conference on Enterprise Information Systems (ICEIS-2001)*. Setúbal,Portugal, 2001.

[242] W3C Workshop on Data y Services Integration. Linked Enterprise Data Patterns Workshop, noviembre 2011. http://www.w3.org/2011/09/LinkedData/.

[243] Open Street Map Org. Open Street Map, noviembre 2011. http://www.openstreetmap.org/.

[244] Larry Page y Sergey Brin. Google Inc. http://www.google.com/.

[245] Lawrence Page, Sergey Brin, Rajeev Motwani, y Terry Winograd. The PageRank citation ranking: Bringing order to the web. Informe técnico, Stanford Digital Library Technologies Project, Stanford University, Stanford, CA, USA, noviembre 1998.

[246] European Parliament y European Council. Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information., noviembre 2003. http://ec.europa.eu/information_society/policy/psi/docs/pdfs/directive/psi_directive_en.pdf.

[247] Alexandre Passant, Sergio Fernández, Jhon Breslin, y Uldis Bojars. Social Data On the Web (SDOW2011), 2011. http://sdow.semanticweb.org/2011.

[248] Peter F. Patel-Schneider y Ian Horrocks. OWL Web Ontology Language Overview 1.1. W3C member submission, W3C, 2006. Http://www.w3.org/Submission/owl11-overview/.

[249] Isidro Peñalver-Martínez, Rafael Valencia-García, y Francisco García Sánchez. Ontology-Guided Approach to Feature-Based Opinion Mining. En *NLDB*, páginas 193–200. 2011.

[250] Emmanuel Pietriga, Christian Bizer, David Karger, y Ryan Lee. Fresnel - a browser-independent presentation vocabulary for RDF. En *In: Proceedings of the Second International Workshop on Interaction Design and the Semantic Web*, páginas 158–171. Springer, 2006.

[251] Mateja Podlogar. Chapter III e-Procurement Success Factors: Challenges and Opportunities for a Small Developing Country. *e-Procurement in Emerging Economies Theory and Cases, Editors Pani and Agrahari, Group Publishing*, 2007.

[252] Axel Polleres. From SPARQL to rules (and back). En *Proceedings of the 16th international conference on World Wide Web*, WWW '07, páginas 787–796. ACM, New York, NY, USA, 2007. ISBN 978-1-59593-654-7.

[253] PoolParty. Metadata Management | Enterprise Search | Text mining | Data integration, diciembre 2011. http://poolparty.biz/.

[254] Jeffrey Pound, Peter Mika, y Hugo Zaragoza. Ad-hoc object retrieval in the Web of Data. En *Proc. of the 19th WWW*, WWW '10, páginas 771–780. ACM, New York, NY, USA, 2010. ISBN 978-1-60558-799-8.

[255] LATC Project. LATC – Linked Open Data Around-The-Clock., diciembre 2011. http://latc-project.eu/.

[256] LOD2 Project. LOD Stats, febrero 2012.

[257] LOTED Project. LOTED – Linked Open Tenders Electronic Daily., diciembre 2011. http://loted.eu:8081/LOTED1Rep/.

[258] SWEO Community Project. Linking Open Data on the Semantic Web. W3C task force, W3C, 2011. http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/CommonVocabularies.

[259] Eric Prud'hommeaux y Andy Seaborne. SPARQL Query Language for RDF. W3C recommendation, W3C, 2007. Http://www.w3.org/TR/rdf-sparql-query/.

[260] Jorge Pérez, Marcelo Arenas, y Claudio Gutierrez. Semantics of SPARQL. TR/DCC-2006-17, Universidad de Chile, 2006. http://www2.ing.puc.cl/~jperez/papers/sparql_semantics.pdf.

[261] Y. Qiu y HP. Frei. Concept-based query expansion. En *Proceedings of SIGIR-93*, páginas 160–169. Pittsburgh, US, 1993.

[262] Norbert Rainer. El sistema revisado de nomenclaturas y clasificaciones internacionales. *CIIU Rev.3, Nace Rev.1, CPA, Unión Europea*, (2):51–65, 1997.

[263] Dave Reynolds. An organization ontology. Working Draft, W3C, 2012. http://www.w3.org/TR/vocab-org/.

[264] C. Rocha, D. Schwabe, y MP. de Aragão. A Hybrid Approach for Searching in the Semantic Web. En *WWW*, páginas 374–383. 2004.

[265] Jose María Alvarez Rodríguez. Other research works, 2011. http://www.josemalvarez.es/web/category/publications/research-works/.

[266] Jose María Alvarez Rodríguez, José Emilio Labra Gayo, y Patricia Ordoñez de Pablos. ONTOSPREAD: A Framework for Supporting the Activation of Concepts in Graph-based Structures through the Spreading Activation technique. En *Proceedings of WSKS 2011. Miltiadis et al.*. 2011.

[267] Sebastian Rudolph, Markus Krötzsch, y Pascal Hitzler. Terminological Reasoning in SHIQ with Ordered Binary Decision Diagrams. En *AAAI*, páginas 529–534. 2008.

[268] Leo Sauermann, Richard Cyganiak, y Max Völkel. Cool URIs for the Semantic Web. Technical Memo TM-07-01, DFKI GmbH, feb 2007. Written by 29.11.2006.

[269] M. Schmidt, M. Meier, y G. Lausen. Foundations of SPARQL query optimization. En *Proc. of the 13th International Conference on Database Theory*, ICDT '10, páginas 4–33. ACM, New York, NY, USA, 2010. ISBN 978-1-60558-947-3.

[270] Volker Schmitz y Joerg Leukel. Findings and Recommendations from a Pan-European Research Project: Comparative Analysis of E-Catalog Standards. *International Journal of IT Standards and Standardization Research (JITSR), Idea Group*, (2):51–65, 2005.

[271] Volker Schmitz, Joerg Leukel, y Frank-Dieter Dorloff. Do E-Catalog Standards Support Advanced Processes in B2B E-Commerce? Findings from the CEN/ISSS Workshop eCAT. En *HICSS '05: Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05) - Track 7*, página 162.3. IEEE Computer Society, Washington, DC, USA, 2005. ISBN 0-7695-2268-8-7.

[272] Frantisek Simancik, Yevgeny Kazakov, y Ian Horrocks. Consequence-Based Reasoning beyond Horn Ontologies. En *IJCAI*, páginas 1093–1098. 2011.

[273] Evren Sirin y Bijan Parsia. SPARQL-DL: SPARQL Query for OWL-DL. En *In 3rd OWL Experiences and Directions Workshop (OWLED-2007*. 2007.

[274] Evren Sirin, Bijan Parsia, Bernardo Cuenca Grau, Aditya Kalyanpur, y Yarden Katz. Pellet: A practical OWL-DL reasoner. *Web Semantics Science Services and Agents on the World Wide Web*, 5(2):51–53, 2007.

[275] Ramakrishna Soma y V K Prasanna. Parallel Inferencing for OWL Knowledge Bases. *2008 37th International Conference on Parallel Processing*, páginas 75–82, 2008.

[276] John F. Sowa. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Course Technology, August 1999. ISBN 0534949657.

[277] Sparallax. Faceted browsing interface for SPARQL endpoint, diciembre 2011. http://sparallax.deri.ie/.

[278] RSS Specifications. RSS Specifications., diciembre 2011. http://www.rss-specifications.com/rss-specifications.htm.

[279] Manu Sporny, Chris Matheus, Nathan Rixham, Andy Seaborne, Thomas Steiner, Matteo Brunati, David Wood, Pat Hayes, Sandro Hawke, y Nicholas Humfrey. JSON RDF Task Force. W3C task force, W3C, 2011. http://www.w3.org/2011/rdf-wg/wiki/TF-JSON.

[280] Claus Stadler, Christophe Guéret, Jens Lehmann, Paul Groth, y Anja Jentzsch. D1.4.1 First Deployment of QA Module. Deliverable of European Project, LATC Project, 2011.

[281] Claus Stadler, Jens Lehmann, Konrad Höffner, y Sören Auer. LinkedGeoData: A Core for a Web of Spatial Open Data. *Semantic Web Journal*, 2011.

[282] Rudi Studer, V. Richard Benjamins, y Dieter Fensel. Knowledge Engineering: Principles and Methods. *Data Knowledge Engineering*, 25(1-2):161–197, 1998.

[283] J. Suchal. On finding power method in spreading activation search. En Viliam Geffert, Juhani Karhumäki, Alberto Bertoni, Bart Preneel, Pavol Návrat, y Mária Bieliková, editores, *SOFSEM (2)*, páginas 124–130. Safarik University, Slovakia, 2008. ISBN 978-80-7097-697-5.

[284] Chris Taggart y Rob McKinnon. OpenCorporates-The Open Database Of The Corporate World, diciembre 2011. http://opencorporates.com/.

[285] TenForce. D1.4 – Early LOD2 Stack prototype. Deliverable of European Project, LOD2 Project, 2011. http://static.lod2.eu/Deliverables/LOD2_D1.4_Early_Stack_Prototype.pdf.

[286] Herman J Ter Horst. Completeness, decidability and complexity of entailment for RDF Schema and a semantic extension involving the OWL vocabulary. *Web Semantics Science Services and Agents on the World Wide Web*, 3:79–115, 2005.

[287] Sebastian Tramp, Philipp Frischmuth, y Norman Heino. OntoWiki – a Semantic Data Wiki Enabling the Collaborative Creation and (Linked Data) Publication of RDF Knowledge Bases. En Oscar Corcho y Johanna Voelker, editores, *Demo Proceedings of the EKAW 2010*. oct 2010.

[288] Triplify. Expose Semantics!, diciembre 2011. http://triplify.org/.

[289] Dmitry Tsarkov y Ian Horrocks. FaCT++ Description Logic Reasoner: System Description. En *Proc. of the Int. Joint Conf. on Automated Reasoning (IJCAR 2006)*, tomo 4130 de *Lecture Notes in Artificial Intelligence*, páginas 292–297. Springer, 2006.

[290] Giovanni Tummarello, Richard Cyganiak, Michele Catasta, Szymon Danielczyk, Renaud Delbru, y Stefan Decker. Sig.ma: Live views on the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 8(4):355–364, noviembre 2010. ISSN 15708268.

[291] Giovanni Tummarello, Renaud Delbru, y Eyal Oren. Sindice.com: Weaving the Open Linked Data. En Karl Aberer, Key-Sun Choi, Natasha Fridman Noy, Dean Allemang, Kyung-Il Lee, Lyndon J. B. Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, y Philippe Cudré-Mauroux, editores, *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007*, tomo 4825 de *Lecture Notes in Computer Science*, páginas 552–565. Springer, 2007. ISBN 978-3-540-76297-3.

[292] Jacopo Urbani, Spyros Kotoulas, Jason Maassen, Frank van Harmelen, y Henri Bal. OWL Reasoning with WebPIE: Calculating the Closure of 100 Billion Triples. En Lora Aroyo, Grigoris Antoniou, Eero Hyvönen, Annette ten Teije, Heiner Stuckenschmidt, Liliana Cabral, y Tania Tudorache, editores, *Proceedings of the 7th Extended Semantic Web Conference (ESWC)*, tomo 6088 de *Lecture Notes in Computer Science*, páginas 213–227. Springer, Berlin, Heidelberg, mayo 2010.

[293] Jacopo Urbani, Spyros Kotoulas, Eyal Oren, y Frank van Harmelen. Scalable Distributed Reasoning Using MapReduce. En *International Semantic Web Conference*, páginas 634–649. 2009.

[294] Jacopo Urbani, Jason Maassen, y Henri Bal. Massive Semantic Web data compression with MapReduce. *Processing*, páginas 795–802, 2010.

[295] Piek Vossen. EuroWordNet, General Document. Informe técnico, University of Amsterdam, 1999.

[296] W3C. Semantic Web Case Studies and Use Cases, nov 2011. http://www.w3.org/2001/sw/sweo/public/UseCases/s.

[297] Jimmy Wales y Larry Sanger. Wikipedia. http://www.wikipedia.org/.

[298] Jesse Weaver y James Hendler. Parallel Materialization of the Finite RDFS Closure for Hundreds of Millions of Triples. En Abraham Bernstein, David Karger, Tom Heath, Lee Feigenbaum, Diana Maynard, Enrico Motta, y Krishnaprasad Thirunarayan, editores, *The Semantic Web - ISWC 2009*, tomo 5823 de *Lecture Notes in Computer Science*, páginas 682–697. Springer Berlin / Heidelberg, 2009. ISBN 978-3-642-04929-3.

[299] Max Weber. *Economy and Society: An Outline of Interpretive Sociology (2 volume set)*. University of California Press, 1978.

[300] Marc Wick. The GeoNames geographical database, noviembre 2011. http://www.geonames.org/.

[301] Graham Wilcock. *Introduction to Linguistic Annotation and Text Analytics*. Morgan & Claypool Publishers, 1st edición, 2009. ISBN 1598297384, 9781598297386.

[302] WSMO Working Group. The Web Service Modeling Ontology (WSMO), 2006.

[303] Y. Yan, C. Wang, y A. Zhou. Efficiently querying rdf data in triple stores. En *Proc. of the 17th WWW*, WWW '08, páginas 1053–1054. ACM, New York, NY, USA, 2008. ISBN 978-1-60558-085-2.

[304] Yang Yu, Donald Hillman, Basuki Setio, y Jeff Heflin. A Case Study in Integrating Multiple E-commerce Standards via Semantic Web Technology. En *Proceedings of the 8th International Semantic Web Conference*, ISWC '09, páginas 909–924. Springer-Verlag, Berlin, Heidelberg, 2009. ISBN 978-3-642-04929-3.

[305] Mark Zuckerberg. Facebook. http://www.facebook.com/.