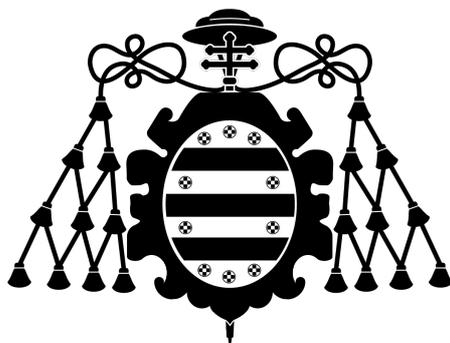


---

# UNIVERSIDAD DE OVIEDO



Escuela de  
Ingeniería  
Informática  
Universidad de Oviedo

ESCUELA DE INGENIERÍA INFORMÁTICA

**TRABAJO FIN DE MÁSTER**

“HERRAMIENTA WEB PARA EL ANÁLISIS DE PORTALES  
BASADA EN APACHE NUTCH”

**AUTOR:** Sergio Rozas García  
**DIRECTOR:** Daniel Fernández Lanvin

Julio 2015

---

# Resumen

Este Trabajo Fin de Máster consiste en el desarrollo de una herramienta informática que facilite el análisis y la comparación de distintos portales web especificados por el usuario mediante el escrutinio de los distintos recursos web que dichos sitios albergan.

Se desarrolla este Trabajo de Fin de Máster en el marco del Máster Universitario en Ingeniería Web y está destinado a ser utilizado por personal del Departamento de Informática de la Universidad de Oviedo para realizar tareas de investigación y prueba de portales web.

La herramienta desarrollada es una herramienta web construida sobre la plataforma *Java Enterprise Edition* con *Spring Framework* que se apoya principalmente en el software de araña web (en Inglés, *crawler* o *web crawler*) Nutch de la Fundación Apache. Este software es altamente escalable, robusto y extensible y permite el procesado por lotes de gran cantidad de recursos web gracias a que se ejecuta sobre la plataforma de computación distribuida Apache Hadoop (basada también en Java).

Uno de los principales objetivos mediante el uso de Apache Nutch es dejar la herramienta preparada para, si así lo requieren los usuarios, realizar un despliegue sobre un clúster de servidores de tal modo que se dé soporte a un gran volumen de descarga de portales web. Para el Trabajo Fin de Máster se realizará el despliegue sobre un único nodo lo cuál resulta suficiente para satisfacer los propósitos fijados. Otro objetivo prioritario es proporcionar una arquitectura software ampliable que permita definir a posteriori, mas algoritmos de análisis en función de las necesidades futuras.

Además se utiliza el software de gestión de bases de datos distribuidas Apache Cassandra, también de la Fundación Apache, para dar soporte a la persistencia de los recursos web obtenidos por Apache Nutch. Este software de persistencia es de tipo NoSQL e implementa una aproximación al modelo de almacenamiento BigTable de Google.

Adicionalmente, los recursos web descargados se indexan en un motor de búsqueda; Apache Solr, por si pudiera ser de utilidad acceder al código fuente de los recursos web y a diversos metadatos que Apache Nutch recoge a posteriori.

# Abstract

This End Of Master's Degree Project develops a tool for easing the analysis and comparison of web sites by scrutinizing the web resources those web sites hold.

This project is developed within the Master's Degree in Web Engineering by the University of Oviedo and it was elaborated as a tool to be used by the personnel of the IT Department of the University of Oviedo in order to do research & testing tasks over different web sites.

The tool was built using technologies around the Java Enterprise Edition platform and Pivotal's Spring framework. The whole system relies on web crawling software Nutch from Apache Foundation. This web crawler was developed with high scalability, robustness & extensibility in mind. It also provides high capacity batch processing due to it's execution cycle runs over Apache Hadoop (both software run in Java).

One of the main goals by using Apache Nutch is to get Apache Hadoop distributed computing capabilities and, if necessary, perform a deployment over a machine cluster. This scenario may be faced when a large amount of crawling operations are required (by having a large list of URLs to be analysed). For demonstration purposes, this project will be deployed on a single node Hadoop setting that will be powerful enough to fulfil this system's basic usage.

Another important goal, is to deliver a software architecture that is able to get expanded by the coding of further analysis algorithms when the staff of the University of Oviedo that will use this system requires to do so.

In order to support the web crawler, a persistence backend will be used to save the web resources. In this project, Cassandra, from Apache Foundation is used. Apache Cassandra is a wide-column NoSQL distributed database management system whose internal schema is an approximation to Google's BigTable storage model.

In addition, this project will involve the usage of Solr from Apache Foundation. Apache Solr is a search engine software capable of indexing the web resources downloaded by Apache Nutch in order to be able to get the source code and other metadata from them.

# Palabras clave

- Análisis de portales web.
- Computación distribuida.
- Java Enterprise Edition.
- Araña Web.
- Software libre.
- Bases de datos distribuidas.
- Motor de búsqueda.

# Keywords

- Web site analysis.
- Distributed computing
- Java Enterprise Edition.
- Web crawler.
- Free software.
- Distributed database.
- Search engine.

# Índice de contenido

Resumen.....	2
Abstract.....	3
Palabras clave.....	4
Keywords.....	5
Índice de figuras.....	12
1 Memoria.....	19
1.1 Esbozo del proyecto.....	19
1.2 Resumen de la motivación.....	20
1.3 Resumen de los objetivos.....	20
1.4 Resumen del alcance.....	21
2 Introducción.....	23
2.1 Justificación del Proyecto.....	23
2.2 Objetivos del Proyecto.....	24
2.3 Estudio de la situación actual.....	25
2.3.1 Evaluación de alternativas de solución.....	27
3 Aspectos técnicos.....	31
3.1 Aplicación web.....	31
3.1.1 Java Enterprise Edition.....	31
3.1.2 Patrón arquitectónico MVC.....	32
3.1.3 Pivotal Spring Framework.....	33
3.1.4 FreeMarker.....	34
3.1.4 Hibernate.....	35
3.1.5 Apache Hadoop.....	35
3.1.6 Apache Nutch.....	37
3.1.7 Apache Cassandra.....	40
3.1.8 Apache Solr.....	41
4 Planificación del proyecto.....	43
4.1 Planificación inicial.....	43
4.2 Cambios en la planificación.....	44
5 Análisis.....	47
5.1 Definición del sistema.....	47
5.1.1 Determinación del alcance.....	47
5.2 Requisitos del sistema.....	48
5.2.1 Obtención de los requisitos del sistema.....	48
5.2.2 Identificación de actores.....	51
5.2.3 Especificación de casos de uso.....	51
5.2.3.1 Gestión de usuarios y grupos.....	52
5.2.3.2 Gestión de servicios.....	53
5.2.3.3 Gestión de trabajos.....	54

5.2.3.4 Vista principal del sistema.....	56
5.3 Identificación de los subsistemas en fase de análisis.....	57
5.3.1 Descripción de los subsistemas.....	57
5.3.2 Descripción de interfaces entre subsistemas.....	58
5.4 Diagrama de clases preliminar de análisis.....	59
5.4.1 Descripción de las clases.....	60
5.4.1.1 Subsistema de usuarios.....	60
5.4.2 Subsistema de servicios.....	61
5.4.3 Subsistema de análisis.....	65
5.4.4 Subsistema de trabajos.....	69
5.4.5 Subsistema WebKumo Connector.....	70
5.5 Análisis de caso de uso y escenarios.....	71
5.5.1 Crear usuario.....	71
5.5.2 Modificar usuario.....	72
5.5.3 Eliminar usuario.....	74
5.5.4 Listar usuarios.....	75
5.5.5 Crear grupo.....	76
5.5.6 Eliminar grupo.....	77
5.5.7 Modificar grupo.....	78
5.5.8 Listar grupos.....	79
5.5.9 Listar grupos.....	80
5.5.8 Arrancar servicio.....	81
5.5.9 Detener servicio.....	82
5.5.10 Listar servicios.....	83
5.5.11 Crear trabajo.....	84
5.5.12 Modificar trabajo.....	85
5.5.13 Eliminar trabajo.....	86
5.5.14 Listar trabajos.....	88
5.5.15 Ver trabajo.....	89
5.5.16 Ejecutar trabajo.....	90
5.5.17 Ver situación general.....	92
5.6 Análisis de interfaces de usuario.....	93
5.6.1 Descripción de la interfaz.....	93
5.6.2 Descripción de comportamiento de la interfaz.....	98
5.6.3 Diagrama de navegabilidad.....	98
5.7 Especificación del plan de pruebas.....	100
5.7.1 Pruebas para casos de uso por subsistemas.....	100
5.7.1.1 Subsistema de usuarios.....	100
5.7.1.2 Subsistema de servicios.....	104
5.7.1.3 Subsistema de trabajos.....	105
5.7.1.4 Subsistema de vista general.....	106
5.7.1.5 Subsistema de WebKumo Connector.....	107
5.7.2 Pruebas de sistema.....	107

5.7.3 Pruebas de accesibilidad y usabilidad.....	108
6 Diseño del sistema.....	109
6.1 Arquitectura del sistema.....	109
6.1.1 Diagramas de paquetes.....	109
6.1.1.1 Paquete 'business'.....	110
6.1.1.2 Paquete 'connector'.....	110
6.1.1.3 Paquete 'view'.....	110
6.1.1.4 Paquete 'persistence'.....	110
6.1.1.5 Paquete 'conf'.....	111
6.1.1.6 Paquete 'model'.....	111
6.2 Diagrama de componentes.....	112
6.3 Diagrama de despliegue.....	113
6.3.1 Agente de usuario.....	113
6.3.2 Aplicación web WebKumo.....	114
6.3.3 Base de datos.....	114
6.3.4 Plataforma de computación.....	114
6.3.5 Araña web.....	114
6.3.6 Persistencia de la araña web.....	115
6.3.7 Motor de búsqueda.....	115
6.4 Diseño de clases.....	116
6.4.1 Paquete org.rz.webkumo.view.....	116
6.4.2 Paquete org.rz.webkumo.business.....	117
6.4.3 Paquete org.rz.webkumo.persistence.....	118
6.4.4 Paquete org.rz.webkumo.conf.....	119
6.4.5 Paquete org.rz.webkumo.connector.....	120
6.4.6 Paquete org.rz.webkumo.model.exceptions.....	120
6.4.7 Paquete org.rz.webkumo.model.analyses.....	121
6.4.8 Paquete org.rz.webkumo.model.users.....	122
6.4.9 Paquete org.rz.webkumo.model.sites.....	122
6.4.10 Paquete org.rz.webkumo.model.jobs.....	123
6.4.11 Paquete org.rz.webkumo.model.enums.....	124
6.4.12 Paquete org.rz.webkumo.model.services.....	125
6.5 Diagramas de interacción y de estados.....	126
6.5.1 Caso de uso 'Crear usuario'.....	126
6.5.2 Caso de uso 'Eliminar usuario'.....	127
6.5.3 Caso de uso 'Listar usuarios'.....	128
6.5.4 Caso de uso 'Ejecutar trabajo'.....	129
6.5.5 Caso de uso 'Iniciar sesión'.....	130
6.6 Diseño de la base de datos.....	131
6.6.1 Descripción del SGBD usado.....	131
6.6.2 Integración del SGBD.....	132
6.6.3 Diagrama Entidad-Relación.....	133
6.7 Diseño de la interfaz.....	134

6.7.1 Aspectos comunes.....	134
6.7.2 Pantalla de inicio de sesión.....	137
6.7.3 Pantalla de vista de situación general.....	137
6.7.4 Pantalla del panel de trabajos.....	138
6.7.5 Pantalla del información de trabajos.....	138
6.7.6 Pantalla de creación/modificación de un trabajo.....	139
6.7.7 Pantalla del panel de servicios.....	140
6.7.8 Pantalla del panel del usuario.....	140
6.7.9 Pantalla de creación/modificación de un usuario.....	142
6.7.10 Pantalla de creación/edición de grupo.....	143
6.8 Especificación técnica del plan de pruebas.....	144
6.8.1 Pruebas unitarias.....	144
6.8.2 Pruebas de sistema.....	144
6.8.3 Pruebas de usabilidad y accesibilidad.....	145
6.8.3.1 Pruebas de usabilidad.....	146
6.8.3.1.1 Cuestiones de carácter general.....	146
6.8.3.1.2 Actividades guiadas.....	146
6.8.3.1.3 Preguntas cortas y observaciones.....	147
6.8.3.1.4 Cuestionario para el responsable de pruebas.....	150
6.8.3.2 Pruebas de accesibilidad.....	151
6.8.4 Pruebas de rendimiento.....	151
7 Implementación del sistema.....	152
7.1 Estándares y normas seguidos.....	152
7.1.1 JavaDoc.....	152
7.1.2 HTML5.....	152
7.1.3 CSS.....	153
7.1.4 WCAG 2.0.....	153
7.2 Lenguajes de programación.....	154
7.3 Herramientas utilizadas en el desarrollo.....	156
7.4 Creación del sistema.....	158
7.4.1 Problemas encontrados.....	158
7.4.2 Descripción detallada de las clases.....	160
8 Desarrollo de las pruebas.....	161
8.1 Pruebas unitarias.....	161
8.1.1 Pruebas sobre usuarios y grupos.....	161
8.1.2 Pruebas sobre algoritmos de análisis.....	162
8.2 Pruebas del sistema.....	164
8.2.1 Pruebas de gestión de usuarios.....	164
8.2.2 Pruebas de servicios.....	169
8.2.3 Pruebas de trabajos.....	171
8.2.4 Pruebas de vista general.....	175
8.2.5 Prueba de WebKumo Connector.....	175
8.3 Pruebas de usabilidad y accesibilidad.....	177

8.3.1 Pruebas de usabilidad.....	177
8.3.1.1 Perfiles de los sujetos de prueba.....	177
8.3.1.2 Cuestionarios previos.....	178
8.3.1.3 Sujeto de pruebas 1: Cuestionarios de actividades guiadas.....	179
8.3.1.4 Sujeto de pruebas 1: Cuestionario del evaluador.....	182
8.3.1.5 Sujeto de pruebas 2: Cuestionarios de actividades guiadas.....	184
8.3.1.6 Sujeto de pruebas 2: Cuestionario del evaluador.....	186
8.3.1.7 Análisis según guía de usabilidad.....	188
8.3.1.8 Cambios derivados de las pruebas.....	192
8.3.2 Pruebas de accesibilidad.....	193
8.3.2.1 Revisión preliminar.....	193
8.3.2.2 Evaluación de conformidad.....	193
8.3.2.2.1 Revisión preliminar de la página de situación general.....	194
8.3.2.2.2 Revisión de la página de vista de un trabajo.....	200
8.3.2.2.3 Revisión de la página de servicios.....	206
8.3.2.2.4 Revisión de la página de nuevo trabajo.....	210
8.3.2.2.5 Revisión del panel del administrador.....	215
8.3.2.3 Comentarios sobre la revisión preliminar.....	220
8.3.2.3.1 Conformidad de la página de situación general.....	222
8.3.2.3.2 Conformidad de la página de ficha de un trabajo.....	225
8.3.2.3.3 Conformidad de la página del panel de servicios.....	228
8.3.2.3.4 Conformidad de la página del creación de un trabajo.....	231
8.3.2.3.5 Conformidad de la página del panel del administrador.....	234
8.3.2.4 Corrección de problemas y declaración de conformidad general. .	237
8.4 Pruebas de rendimiento.....	243
9 Manuales.....	246
9.1 Manual de instalación.....	246
9.1.1 Instalación de Apache Hadoop.....	246
9.1.2 Instalación de Apache Cassandra.....	252
9.1.3 Instalación de Apache Solr.....	254
9.1.4 Instalación de Apache Nutch.....	256
9.1.5 Contenedor de aplicaciones J2EE.....	267
9.2 Manual de ejecución.....	269
9.2.1 Apache Hadoop.....	269
9.2.2 Apache Solr.....	270
9.2.3 Apache Cassandra.....	270
9.2.4 Aplicación web WebKumo.....	272
9.2.4.1 Initialization.properties.....	273
9.2.4.2 Datasource.properties.....	273
9.2.4.3 Infrastructure.properties.....	274
9.2.4.4 Webmvc.properties.....	277
9.2.4.5 Implementación de los servicios.....	277
9.3 Manual del usuario.....	279

9.3.1 Inicio de sesión.....	279
9.3.2 Cierre de sesión.....	280
9.3.3 Panel del usuario (administrador).....	281
9.3.4 Creación de usuario (administrador).....	283
9.3.5 Edición de usuario (administrador).....	285
9.3.6 Eliminación de un usuario (administrador).....	286
9.3.7 Creación de un grupo (administrador).....	288
9.3.8 Modificación de un grupo (administrador).....	289
9.3.9 Borrado de un grupo (administrador).....	290
9.3.10 Panel del usuario (regular).....	290
9.3.11 Modificación del usuario (regular).....	291
9.3.12 Consulta de la situación general del sistema.....	292
9.3.13 Panel de servicios (administrador).....	293
9.3.14 Arranque/Parada de un servicio (administrador).....	294
9.3.15 Panel de trabajos.....	295
9.3.16 Crear un trabajo.....	296
9.3.17 Consulta de un trabajo.....	299
9.3.18 Modificar un trabajo.....	303
9.3.19 Borrar un trabajo.....	303
9.4 Manual del programador.....	305
9.4.1 Creación de la nueva clase de análisis.....	305
9.4.2 Parámetros personalizados.....	309
9.4.3 Creación de una nueva clase de informe.....	310
9.4.4 Creación de una nueva clase de campo de informe.....	311
9.4.5 Gestión de dependencias.....	313
9.4.6 Integración de nuevos análisis en Apache Nutch.....	315
10 Conclusiones y ampliaciones.....	317
10.1 Conclusiones.....	317
10.2 Ampliaciones.....	318
11 Presupuesto.....	319
12 Referencias bibliográficas.....	321
12.1 Referencias en internet.....	321
12.2 Libros.....	326
13 Glosario.....	327
14 Contenido entregado.....	329

# Índice de figuras

Fig. 1: Descripción esquemática del patrón arquitectónico MVC.....	32
Fig. 2: Esquema de funcionamiento de FreeMarker.....	34
Fig. 3: Esquema de un clúster de Apache Hadoop típico.....	37
Fig. 4: Ciclo de ejecución básico de Apache Nutch.....	39
Fig. 5: Teorema CAP.....	41
Fig. 6: Casos de uso de gestión de los usuarios y los grupos de usuarios.....	53
Fig. 7: Casos de uso de gestión de servicios.....	53
Fig. 8: Casos de uso de gestión de trabajos.....	54
Fig. 9: Caso de uso de visualización de la situación general del sistema.....	56
Fig. 10: Diagrama de clases preliminar simplificado.....	59
Fig. 11: Diagrama de robustez para el alta de un usuario.....	72
Fig. 12: Diagrama de robustez del caso de uso de baja de usuario.....	75
Fig. 13: Diagrama de robustez del caso de uso de listado de usuarios.....	76
Fig. 14: Diagrama de robustez del caso de uso de inicio de sesión.....	80
Fig. 15: Diagrama de robustez del caso de uso de arranque de servicio.....	81
Fig. 16: Diagrama de robustez del caso de uso de parada de servicio.....	83
Fig. 17: Diagrama de robustez para el caso de uso de listado de servicios.....	84
Fig. 18: Diagrama de robustez del caso de uso para crear un trabajo nuevo.....	85
Fig. 19: Diagrama de robustez del caso de uso de eliminación de un trabajo.....	88
Fig. 20 Diagrama de robustez del caso de uso de vista de un trabajo.....	90
Fig. 21: Diagrama de robustez del caso de uso de ejecución de un trabajo.....	91
Fig. 22: Diagrama de robustez del caso de uso de vista de la situación general...	92
Fig. 23: Prototipo de la pantalla de inicio de sesión.....	93
Fig. 24: Prototipo de la pantalla de vista general.....	94
Fig. 25: Prototipo de la pantalla con información tabular (trabajos).....	95
Fig. 26: Prototipo de una pantalla con formulario de ejemplo (trabajo nuevo).....	96
Fig. 27: Prototipo de una pantalla de trabajo (trabajo nuevo).....	97
Fig. 28: Diagrama de navegabilidad de pantallas.....	98
Fig. 29: Diagrama de paquetes.....	109
Fig. 30: Diagrama de componentes.....	112
Fig. 31: Diagrama de despliegue.....	113
Fig. 32: Diagrama de clases del paquete de vista.....	116
Fig. 33: Diagrama de clases de la capa de negocio.....	117
Fig. 34: Diagrama de clases de la capa de persistencia.....	118
Fig. 35: Diagrama de clases de configuración.....	119
Fig. 36: Diagrama de clases del plugin WebKumo Connector.....	120
Fig. 37: Diagrama de clases del modelo de excepciones.....	120
Fig. 38: Diagrama de clases del modelo de análisis.....	121
Fig. 39: Diagrama de clases del modelo de usuarios.....	122

Fig. 40: Diagrama de clases del modelo de recursos web.....	122
Fig. 41: Diagrama de clases del modelo de trabajos de araña web.....	123
Fig. 42: Diagrama de clases de enumeraciones del modelo.....	124
Fig. 43: Diagrama de clases del modelo de servicios de infraestructura.....	125
Fig. 44: Diagrama de interacción para dar de alta un elemento (usuario).....	126
Fig. 45: Diagrama de interacción para la baja de un elemento (usuario).....	128
Fig. 46: Diagrama de interacción para listar múltiples elementos.....	128
Fig. 47: Diagrama de interacción del caso de uso de ejecución de un trabajo.....	129
Fig. 48: Diagrama de interacción para el caso de uso de inicio de sesión.....	130
Fig. 49: Diagrama Entidad-Relación del modelo de la base de datos.....	133
Fig. 50: Barra de navegación superior.....	134
Fig. 51: Botón desplegable del menú de navegación.....	135
Fig. 52: Botón de accionamiento en la parte derecha del título de la sección.....	135
Fig. 53: Pie de página.....	135
Fig. 54: Tabla utilizada en la interfaz.....	136
Fig. 55: Migas de pan en la interfaz.....	136
Fig. 56: Pantalla de inicio de sesión.....	137
Fig. 57: Pantalla de vista general.....	137
Fig. 58: Pantalla del panel de trabajos.....	138
Fig. 59: Pantalla de información de un trabajo.....	138
Fig. 60: Resultados de un análisis de ejemplo (pantalla de información de un trabajo).....	139
Fig. 61: Pantalla de creación de un trabajo.....	139
Fig. 62: Pantalla del panel de servicios.....	140
Fig. 63: Pantalla del panel del usuario administrador con la tabla de usuarios....	140
Fig. 64: Pantalla del panel del usuario administrador con la tabla de grupos.....	141
Fig. 65: Pantalla del panel del usuario regular.....	141
Fig. 66: Pantalla de creación y/o edición de un usuario para administradores.....	142
Fig. 67: Pantalla de edición de un usuario regular.....	142
Fig. 68: Pantalla de creación y/o edición de un grupo.....	143
Fig. 69: Ejecución exitosa de los tests unitarios de usuarios y grupos.....	161
Fig. 70: Ejecución exitosa de los test de análisis.....	162
Fig. 71: Ejecución exitosa de las pruebas de funcionalidad de inicio de sesión....	164
Fig. 72: Ejecución exitosa de las pruebas de funcionalidad relacionada con usuarios.....	165
Fig. 73: Ejecución exitosa de las pruebas de sistema sobre los servicios.....	169
Fig. 74: Ejecución exitosa de las pruebas de sistema sobre trabajo.....	171
Fig. 75: Ejecución exitosa de la prueba funcional de ejecución de un trabajo.....	174
Fig. 76: Ejecución exitosa de la prueba de WebKumo Connector.....	175
Fig. 77: Página de situación general visualizada mediante un navegador de texto. .....	194
Fig. 78: Página de situación general sin imágenes.....	195
Fig. 79: Página de situación general con ancho de dispositivo móvil (720x1280px).	

.....	196
Fig. 80: Página de situación general sin estilos.....	197
Fig. 81: Página de situación general renderizada por Mozilla Firefox.....	198
Fig. 82: Página de situación general renderizada por Google Chrome.....	198
Fig. 83: Página de situación general renderizada por Microsoft Internet Explorer .....	199
Fig. 84: Página de ficha de trabajo visualizada mediante navegador de texto.....	200
Fig. 85: Página de ficha de trabajo sin imágenes.....	200
Fig. 86: Página de ficha de trabajo a una resolución típica de dispositivos móviles (720x1080px).....	201
Fig. 87: Página de ficha de trabajo sin hojas de estilos aplicadas.....	202
Fig. 88: Página de ficha de trabajo renderizada por Mozilla Firefox.....	203
Fig. 89: Página de ficha de trabajo renderizada por Google Chrome.....	204
Fig. 90: Página de ficha de trabajo renderizada por Microsoft Internet Explorer.....	205
Fig. 91: Página del panel de servicios visualizada en un navegador de texto.....	206
Fig. 92: Página del panel de servicios visualizada sin imágenes.....	207
Fig. 93: Página del panel de servicios a una resolución típica de dispositivos móviles (720x1280px).....	207
Fig. 94: Página del panel de servicios sin hojas de estilos aplicadas.....	208
Fig. 95: Página del panel de servicios renderizada por Mozilla Firefox.....	209
Fig. 96: Página del panel de servicios renderizada por Google Chrome.....	209
Fig. 97: Página del panel de servicios renderizada por Microsoft Internet Explorer .....	210
Fig. 98: Página de creación de un trabajo mostrada por un navegador de texto.....	210
Fig. 99: Página de creación de un trabajo renderizada sin imágenes.....	211
Fig. 100: Página de creación de un trabajo a una resolución típica de dispositivo móvil (720x1080px).....	212
Fig. 101: Página de creación de un trabajo sin hojas de estilo aplicadas.....	213
Fig. 102: Página de creación de un trabajo renderizada por Mozilla Firefox.....	214
Fig. 103: Página de creación de un trabajo renderizada por Google Chrome.....	214
Fig. 104: Página de creación de un trabajo renderizada por Microsoft Internet Explorer.....	215
Fig. 105: Página del panel del usuario del administrador vista en un navegador textual.....	215
Fig. 106: Página del panel del usuario del administrador sin imágenes.....	216
Fig. 107: Página del panel del administrador a una resolución típica de dispositivos móviles (720x1080px).....	217
Fig. 108: Página del panel del administrador sin hojas de estilos aplicadas.....	218
Fig. 109: Página del panel del administrador renderizada por Mozilla Firefox.....	219
Fig. 110: Página del panel del administrador renderizada por Google Chrome.....	219
Fig. 111: Página del panel del administrador renderizada por Microsoft Internet Explorer.....	220
Fig. 112: Validación sintáctica del código HTML de la página de situación general	

.....	222
Fig. 113: Validación sintáctica del código HTML de la página de información de un trabajo.....	225
Fig. 114: Validación sintáctica del código HTML de la página de panel de servicios .....	228
Fig. 115: Validación sintáctica del código HTML de la página de creación de un trabajo.....	231
Fig. 116: Validación sintáctica del código HTML de la página del panel del administrador.....	234
Fig. 117: Criterio WCAG 2.0 3.1.1 corregido.....	238
Fig. 118: Criterio 1.4.3 corregido para la página de situación general.....	238
Fig. 119: Criterio 1.4.3 corregido para la página de información de un trabajo...	239
Fig. 120: Criterio 1.4.3 corregido para la página de panel de servicios.....	239
Fig. 121: Criterio 1.4.3 corregido para la página de creación de un trabajo.....	240
Fig. 122: Criterio 1.4.3 corregido para la página del panel del administrador.....	240
Fig. 123: Criterio 2.2.1 corregido mediante el uso del atributo tabindex.....	241
Fig. 124: Criterio 2.4.5 corregido mediante el insertado de enlaces a otras páginas del sitio.....	241
Fig. 125: Criterio 3.1.2 corregido mediante el insertado de marcas de idioma.....	242
Fig. 126: Criterio 4.1.2 corregido mediante el uso de elementos label.....	242
Fig. 127 : Prueba de rendimiento para 100 campos.....	244
Fig. 128: Prueba de rendimiento para 1000 campos.....	244
Fig. 129: Prueba de rendimiento para 2500 campos.....	245
Fig. 130: Problemas de memoria en el IDE al ejecutar la prueba con 2500 campos. ....	245
Fig. 131: Descarga de Apache Hadoop.....	247
Fig. 132: Usuario y grupo añadido para Hadoop.....	247
Fig. 133: Variable de entorno establecida para Hadoop.....	247
Fig. 134: Generación de la llave RSA para SSH.....	248
Fig. 135: Copia de la llave SSH al fichero de llaves autorizadas.....	248
Fig. 136: Asignación del propietario del directorio de instalación de Apache Hadoop.....	249
Fig. 137: Creación de la jerarquía de directorios locales para HDFS.....	249
Fig. 138: Configuración del fichero hadoop-env.....	250
Fig. 139: Configuración del fichero core-site.....	250
Fig. 140: Configuración del fichero hdfs-site.....	251
Fig. 141: Formateo del namenode.....	251
Fig. 142: Arranque del nodo Hadoop.....	252
Fig. 143: Vista de la interfaz web del namenode en ejecución.....	252
Fig. 144: Descarga y descompresión de Apache Cassandra.....	253
Fig. 145: Proceso de arranque de Apache Cassandra.....	253
Fig. 146: Intérprete CQLSH en ejecución.....	254
Fig. 147 Descarga y descompresión de Apache Solr.....	255

Fig. 148: Copia del fichero schema para Apache Nutch.....	255
Fig. 149: Edición del fichero schema.xml.....	255
Fig. 150: Arranque de Apache Solr.....	256
Fig. 151: Interfaz web de Apache Solr.....	256
Fig. 152: Descarga y descompresión de Apache Nutch.....	257
Fig. 153: Configuración del fichero nutch-site.....	257
Fig. 154: Modificación del fichero ivy.xml.....	260
Fig. 155: Configuración de Apache Gora mediante su fichero properties.....	260
Fig. 156: Configuración de la versión del compilador de Java.....	261
Fig. 157: Copia de los ficheros de código de WebKumo Connector.....	261
Fig. 158: Configuración del fichero build.xml para los plugins de Apache Nutch.....	262
Fig. 159: Compilación exitosa de Apache Nutch.....	263
Fig. 160: Copia de las librerías JAR descargadas.....	263
Fig. 161: Creación del fichero de semillas.....	264
Fig. 162: Inyección de semillas.....	264
Fig. 163: Keyspace webpage presente en Apache Cassandra.....	264
Fig. 164: Ejecución de la fase de generación de Nutch.....	265
Fig. 165: Ejecución de la fase de descarga de Nutch.....	265
Fig. 166: Ejecución de la fase de procesado de Nutch.....	266
Fig. 167: Ejecución de la fase de actualización de Apache Cassandra.....	266
Fig. 168: Ejecución de la fase de indexado de Apache Nutch.....	266
Fig. 169: Documento principal del portal web de la Universidad de Oviedo indexado en Solr.....	267
Fig. 170: Descarga y descompresión de Apache Tomcat.....	267
Fig. 171: Configuración del usuario de Apache Tomcat.....	268
Fig. 172: Ejecución del script de inicio de Apache Hadoop.....	269
Fig. 173: Ejecución del script de parada de Apache Hadoop.....	269
Fig. 174: Ejecución del script de inicio de Apache Solr.....	270
Fig. 175: Ejecución del script de parada de Apache Solr.....	270
Fig. 176: Ejecución del script de arranque de Apache Cassandra.....	271
Fig. 177: Detención de Apache Cassandra mediante la parada de su proceso.....	271
Fig. 178: Configuración de Apache Tomcat en Netbeans.....	272
Fig. 179: Valores válidos de initialization.properties.....	273
Fig. 180: Fichero de configuración datasource.properties.....	274
Fig. 181: Factoría de servicios para Apache Nutch.....	278
Fig. 182: Formulario de inicio de sesión.....	279
Fig. 183: Credenciales incorrectas en el inicio de sesión.....	280
Fig. 184: Botón de cierre de sesión.....	280
Fig. 185: Mensaje de cierre de sesión.....	281
Fig. 186: Botón de acceso al panel de administración.....	281
Fig. 187: Vista del panel de administración.....	282
Fig. 188: Selector de páginas de tabla.....	282
Fig. 189: Cuadro de búsqueda en tabla.....	282

Fig. 190: Botón de ordenación de columnas y selector de registros por página..	283
Fig. 191: Formulario de creación de usuario.....	283
Fig. 192: Indicaciones de error en el formulario de creación de usuario.....	285
Fig. 193: Mensaje de éxito en la operación.....	285
Fig. 194: Acceso a la edición de un usuario.....	286
Fig. 195: Botón de borrado de usuario.....	286
Fig. 196: Mensaje de confirmación de borrado del usuario.....	287
Fig. 197: Mensaje de éxito y error en el borrado de un usuario.....	287
Fig. 198: Botón de creación de un grupo.....	288
Fig. 199: Formulario de creación de grupo.....	288
Fig. 200: Error en el formulario de creación de grupo.....	289
Fig. 201: Acceso al formulario de edición de grupo.....	289
Fig. 202: Botón de acceso al panel del usuario.....	290
Fig. 203: Panel del usuario regular.....	290
Fig. 204: Botón de edición del usuario regular.....	291
Fig. 205: Formulario de edición de un usuario regular.....	291
Fig. 206: Panel de vista general del sistema.....	292
Fig. 207: Botón de acceso al panel de servicios.....	293
Fig. 208: Vista del panel de servicios.....	293
Fig. 209: Botones de acciones sobre un servicio.....	294
Fig. 210: Indicador de operación en curso sobre un servicio.....	294
Fig. 211: Éxito en la operación de arranque de un servicio y su salida.....	295
Fig. 212: Botón de acceso al panel de trabajos.....	295
Fig. 213: Vista del panel de trabajos.....	296
Fig. 214: Indicador de progreso de un trabajo en ejecución.....	296
Fig. 215: Mensaje de error para un usuario sin permisos de escritura al intentar crear un trabajo.....	297
Fig. 216: Formulario de creación de un trabajo.....	297
Fig. 217: Mensaje de error en la creación de un trabajo.....	299
Fig. 218: Mensaje de confirmación en el panel de trabajos.....	299
Fig. 219: Acceso al trabajo.....	299
Fig. 220: Ficha del trabajo.....	300
Fig. 221: Panel de resultados de ejecución.....	301
Fig. 222: Vista del código fuente de un recurso indexado.....	301
Fig. 223: Mensaje de fallo de ejecución del trabajo.....	302
Fig. 224: Botón de carga de campos.....	302
Fig. 225: Mensaje que indica que ya se han cargado todos los campos.....	302
Fig. 226: Identificador de trabajo clickable para acceder a la edición de un trabajo. .....	303
Fig. 227: Botón de borrado de un trabajo.....	303
Fig. 228: Mensaje de confirmación de borrado de un trabajo.....	304
Fig. 228: Ejemplo de dependencia Maven para WebKumo.....	313
Fig. 229: Declaración de dependencias para WebKumo Connector mediante	

Apache Ivy.....	314
Fig. 230: Muestra de una dependencia en el repositorio central de Maven en sus distintas formas.....	315

# 1 Memoria

## 1.1 Esbozo del proyecto

Aproximadamente el 50% de la población mundial posee acceso a Internet según el informe del año 2014 de la *World Wide Web Foundation [WINDEX]* y por tanto una inmensa mayoría de esa población tiene acceso a la Web. Esto implica que existe un enorme número potencial de usuarios de sitios web que representan diversos perfiles, cada uno con sus particularidades.

Del dato del párrafo anterior se puede concluir fácilmente que la existencia de distintos perfiles de usuario (habituados al uso de un computador, con problemas visuales, novatos en la exploración de la web etc.) hace que un único diseño de los distintos portales web existentes no sea adecuado para cada tipo de perfil. Dentro de la Informática y mas concretamente de la Ingeniería Web, el estudio de la Usabilidad y la Accesibilidad de los distintos sitios web, ayudan a crear una Web que se adapte a los usuarios según sus características (discapacidades, facilidad de uso etc.).

Resulta obvio que debido al enorme tamaño de la Web (Google en 2008 ya indexaba **[GOOB]** mas de 1 trillón de URIs) es de gran utilidad para los administradores de sitios web, investigadores y demás profesionales analizar ciertos aspectos de distintos portales Web incluyendo aspectos como su internacionalización, su adaptación a dispositivos móviles etc. El problema es que el análisis manual de dichos portales Web es lento y por tanto caro.

De la necesidad anterior surge este Trabajo Fin de Máster. Personal del Departamento de Informática de la Universidad de Oviedo plantea al Director de este proyecto la necesidad de desarrollar una herramienta de evaluación de portales Web como parte de tareas de investigación.

El principal objetivo de la herramienta a desarrollar es analizar y comparar portales Web y presentar los resultados de dichas evaluaciones al usuario. Otros objetivos relevantes que se detallarán mas en profundidad en secciones posteriores son por ejemplo la flexibilidad del código, la posibilidad de construir la aplicación en torno a una plataforma de computación que soporte altas cargas de trabajo y la definición de criterios de parada a la hora de descargar los portales Web.

En cuanto al ámbito del proyecto, es decir, el trabajo realizado para llevar a cabo los objetivos del proyecto **[PMBOK]**, básicamente se refiere al desarrollo de la herramienta web (una aplicación web), el despliegue de la infraestructura que Apache Nutch requiere y la codificación de criterios de análisis y comparación precargados. Además de la gestión del proyecto en sí.

## 1.2 Resumen de la motivación

El presente proyecto tiene por objeto facilitar tareas a profesionales de las Tecnologías de la Información que requieran analizar algún aspecto técnico de portales web concretos que sean objeto de estudio. Temáticamente es un proyecto de relevancia debido a su utilidad y a su aplicación real futura.

En el ejercicio de las actividades de distintos investigadores y profesionales, y especialmente aquellos que trabajen en el campo de la Web, en ocasiones puede ser útil realizar un análisis de ciertos portales web evaluando los recursos que los conforman. Este proyecto intenta dar respuesta a esta necesidad mediante la creación de una herramienta sostenida por una plataforma de computación distribuida bien conocida que proporcione una alta capacidad; entendiendo dicha capacidad como la posibilidad de realizar las tareas para una gran cantidad de portales web.

Las posibilidades de análisis y comparación que ofrece la herramienta son altamente útiles para la automatización de los procesos que de otra manera se llevarían a cabo de manera manual con el alto coste tanto temporal como económico que ello supondría.

Además de los motivos anteriores, se encuentra la natural motivación de ser este proyecto necesario para la consecución del título de Máster Universitario en Ingeniería Web así como la introducción a las tecnologías llamadas de *Big Data*; sistemas informáticos que procesan una gran cantidad de datos, como son las adoptadas en el marco de este Trabajo Fin de Máster.

## 1.3 Resumen de los objetivos

Principalmente se pretende con este proyecto crear una herramienta que permita, como se explicó en la motivación del proyecto, que permita analizar y comparar características técnicas de portales web.

Es vital (uno de los puntos en los que más hace hincapié el cliente) desde el punto de vista de la codificación, la flexibilidad del código desarrollado; principalmente para la posterior extensión de la herramienta por parte del equipo cliente de cara a analizar otras características de los portales Web que se requieran evaluar.

Por lo anterior se hace imprescindible aplicar Patrones de Diseño Software bien conocidos que permitan ampliar con facilidad los criterios de análisis y comparación mediante la construcción de una arquitectura Software extensible.

Entre las características a analizar (o comparar entre varios sitios web) que sería de interés se encuentran la internacionalización de los portales, la adaptación de la interfaz a dispositivos móviles, la validez sintáctica del código HTML, la presencia de conectores de redes sociales, el servidor web utilizado etc. Algunos de estos criterios de análisis se proporcionan precargados en la aplicación.

Otro importante horizonte que se plantea es el desarrollo de la herramienta sostenida por un sistema que soporte gran volumen de descargas de sitios. Se define este objetivo de cara a, si se requiere, realizar el despliegue de la herramienta y el software que la sustenta en un cluster donde poder ejecutar los trabajos de descarga y análisis de forma distribuida.

## 1.4 Resumen del alcance

Puede condensarse el alcance de este proyecto en los puntos detallados a continuación.

- 1) Despliegue de la plataforma de descarga de portales web, esto es, la araña web Apache Nutch; la plataforma de computación distribuida sobre la que se ejecuta la araña, Apache Hadoop; y el software de persistencia para la araña, Apache Cassandra.
- 2) Desarrollo de la herramienta web, bautizada como *WebKumo* (araña web en idioma Japonés) que es el punto de acceso para los usuarios a la funcionalidad.
- 3) Creación de la arquitectura interna de la aplicación web para dar soporte a la funcionalidad de análisis.
- 4) Realizar la integración entre la plataforma de araña web (la descrita en el punto 1) y la aplicación *WebKumo*.

5) Llevar a cabo las tareas concretas (especificadas como tareas en la *Estructura de Descomposición de Trabajo: EDT ó WBS*) elaborada como parte de la asignatura de Dirección y Gestión de Proyectos del Máster.

# 2 Introducción

## 2.1 Justificación del Proyecto

El Proyectante de este Trabajo Fin de Máster, en búsqueda de un proyecto que realizar para completar sus estudios de Máster contactó con el Director, que presentó la problemática existente y finalmente se acordó la puesta en marcha de dicho proyecto.

Se propuso este proyecto como respuesta al requerimiento de un grupo de miembros del Departamento de Informática de la Universidad de Oviedo los cuáles en el desarrollo de sus tareas de investigación y/o desarrollo de sus actividades profesionales en la Universidad, vieron la necesidad de crear una herramienta que permitiera analizar las características técnicas de diversos portales web. Es de gran ayuda en procesos, por ejemplo, de análisis de la usabilidad o la accesibilidad tener una herramienta que automatice la descarga y el análisis de los recursos web para hacer mas rápido dicho análisis.

Por otra parte, el desarrollo de la aplicación web permite aplicar muchos de los campos estudiados durante el Máster como son la programación web, los patrones de diseño, la usabilidad y accesibilidad, el procesamiento de lenguajes propios de la web etc. Lo anterior supone una justificación en cuanto a ejercicio académico aplicado a un proyecto real se refiere.

## 2.2 Objetivos del Proyecto

Los objetivos de este proyecto son los listados a continuación.

- 1) Realizar el análisis de portales web en función de algún criterio codificado en la aplicación. Dentro de este objetivo se incluye también la comparación de portales, entendiéndolo como la ejecución de análisis con algún elemento diferenciador que permita la comparación de distintas versiones de los recursos que albergan.
- 2) Generar una arquitectura software flexible que permita la adición de nuevos algoritmos de análisis.
- 3) Proveer una plataforma de descarga de recursos web de alta capacidad para la automatización posterior del análisis comentado en los puntos anteriores.
- 4) Facilitar el análisis de gran cantidad de los recursos web obtenidos mediante la plataforma de descarga.

## 2.3 Estudio de la situación actual

No existen, en la actualidad, soluciones ciento por ciento similares con respecto a la funcionalidad que se pretende cubrir en este proyecto. Es cierto que hay herramientas que coinciden parcialmente en la funcionalidad pero suelen ser típicamente servicios online de análisis de una característica concreta y no una aplicación cuyo código fuente pueda obtenerse y ampliar según las necesidades del análisis.

Como ejemplos que constituyen el caso mas típico de herramientas que cubren parte de la funcionalidad de análisis que WebKumo ofrece, se tienen:

- **Analizadores de accesibilidad:** Herramientas online que comprueban la accesibilidad de un sitio web mediante el análisis de las características técnicas de los recursos web que contienen. Típicamente se analizan características en base a una guía o recomendación de accesibilidad estandarizada o reconocida como por ejemplo las recomendaciones WCAG [WCAG] o WAI [WAI] del W3C [W3C] o la recomendación surgida de diversas entidades de investigación europeas UWEM [UWEM]. También los hay mas especializados, como analizadores de contraste basándose en los estilos computados o analizadores de estilos para web móvil. Como ejemplos de estas herramientas pueden verse los siguientes:
  - ◆ TAWDIS (Fundación CTIC) [TAW]
  - ◆ Achecker (Universidad de Toronto) [ACHECK]
  - ◆ Wave (WebAIM) [WAVE]
  - ◆ AccessColor (AccessKeys) [AKEYS]
  - ◆ HERA (Fundación SIDAR) [HERA]
  - ◆ Mobile Test (Google) [GOOMOV]
  
- **Analizadores de SEO [SEO]:** Estos servicios analizan características técnicas de los portales web para determinar su idoneidad respecto a lo que las arañas web de los distintos buscadores web comerciales esperan encontrar. El cumplimiento de ciertos aspectos técnicos redundante en una mejor posición en los resultados de búsqueda en dichos buscadores. El análisis se basa principalmente en la validez sintáctica de los recursos web, en el marcado semántico de los mismos y el grado de accesibilidad entre otros. Ejemplos de herramientas de análisis de este tipo (típicamente online) son las siguientes:

- ◆ **Analizador SEO (WooRank) [WOORNK]**
  - ◆ **Site Auditor (Raven Tools) [SAUDI]**
  - ◆ **SEO Toolkit (SeoBook): [STOOL]**
- 
- **Analizadores sintácticos:** En este caso, las herramientas realizan el análisis sobre la validez sintáctica de ciertos recursos web. El caso de aplicación mas típico es la validación del código HTML y CSS que conforman el portal web, siguiendo los estándares diversos existentes (XHTML, HTML5, CSS2, CSS3 etc.). Se pueden poner como ejemplos las siguientes herramientas:
    - ◆ **Validador HTML (W3C) [W3HTMLV]**
    - ◆ **Validador CSS (W3C) [W3CSSV]**
    - ◆ **NU Checker (Mozilla Foundation, Sivonen, Pieters) [NUVAL]**
    - ◆ **Validador RDF (W3C) [W3RDFV]**
- 
- **Analizadores de vulnerabilidades:** Este tipo de analizadores ejecutan tareas de comprobación de seguridad, como por ejemplo la existencia de formularios inseguros, búsqueda de *scripts* maliciosos conocidos, inyección de código, versiones vulnerables del servidor web etc. Este tipo de análisis podría ser parcialmente cubierto por WebKumo ya que se basan la identificación de elementos en el código fuente de los recursos web. Como ejemplos de este tipo de servicios se tienen:
    - ◆ **Scan My Server (Beyond Security) [SCANSVR]**
    - ◆ **Sucuri (Sucury.net) [SUCURI]**
    - ◆ **VirusTotal (VirtuaTotal) [VIRTOTAL]**
    - ◆ **Acunetix Scanner (Acunetix) [ACUNETIX]**
- 
- **Otras herramientas:** Además de las herramientas mas típicas descritas anteriormente, también existen herramientas de análisis mas especializadas como pueden ser analizadores de contraste sobre navegador, medición de rendimiento de carga, medición de rendimiento de red etc. En parte, algunas especializaciones de este tipo de herramientas podrían ser cubiertas por la funcionalidad de WebKumo siempre que comprendan el análisis de los recursos web (por ejemplo en ciertas mediciones de rendimiento, se podría utilizar WebKumo para comprobar si los recursos JavaScript están minimizados [**MINIFIC**] o no).

Se ha de tener en cuenta, que uno de los principales objetivos del proyecto, es la flexibilidad a la hora de definir criterios de análisis, esto es, incluir la posibilidad de definir análisis sobre recursos web de manera personalizada para atender a una necesidad concreta. Los análisis que se requerirán no pueden saberse con antelación (dependerán de los usuarios de la aplicación, que serán usuarios con alto grado de conocimiento técnico) aunque sí que se sabe de qué tipo serán (análisis de accesibilidad, análisis de internacionalización etc.).

Como conclusión relativa a la situación preexistente al desarrollo de esta herramienta, se extrajo que las herramientas existentes no cubren la necesidad que se plantea ya que suelen ser aplicaciones o servicios muy especializados en el análisis de un pequeño conjunto de características de los portales web. Además se suelen ofrecer como un servicio, no como un programa cuyo código pueda obtenerse y modificarse (con excepciones, como *NU.parser* o *TAWDIS* que son de código abierto) para adaptarse a las necesidades. Además si se optara por esta opción, sería muy costoso ya que con gran probabilidad habría que implementar cambios muy profundos en las herramientas si el requisito de análisis varía del fin de la herramienta debido a su alto grado de especialización.

Por lo anterior, se considera que la opción mas adecuada es el desarrollo de una herramienta de nuevo cuño.

### 2.3.1 Evaluación de alternativas de solución

A continuación se describen las distintas alternativas para dar respuesta a los requerimientos del proyecto.

- 1) **Aplicación tipo *hub* de servicios:** Debido a la heterogeneidad de los posibles análisis que se deseen realizar, una posible opción es realizar una aplicación que tome el papel de meta-analizador, esto es, una aplicación que utilice servicios especializados de análisis de terceros ya existentes (como los descritos en el punto 2.3) proporcionando una interfaz común en la medida de lo posible para ejecutar análisis sobre los portales web que el usuario desee.

✓ **Ventajas**

1. Diversidad de servicios online disponibles sin necesidad de programar los algoritmos de análisis.
2. Se elimina la restricción de análisis de los recursos web, pudiéndose así analizar aspectos accesorios (por ejemplo, rendimientos de red).
3. Infraestructura liviana y bajos requerimientos de recursos ya que la computación de los análisis los realizaría el servicio de terceros.
4. Los servicios de terceros interoperan principalmente mediante servicios web, que son tecnologías bien conocidas y documentadas.

x **Inconvenientes**

1. Limitación a los análisis que provean terceros.
2. Servicios limitados por sus términos de uso y sus posibles cambios.
3. Dependencia fuerte de APIs de terceros (mediante servicios web muy probablemente) cuya disponibilidad no suele estar garantizada y que además, suelen incluir restricciones de uso.
4. Algunos de los servicios son comerciales (no gratuitos) o bien gratuitos pero limitados en dicha modalidad.
5. Probabilidad de que los servicios no se adapten totalmente al requerimiento de análisis que el usuario tenga.
6. Es posible que para un cierto requerimiento de análisis, no exista un servicio de terceros que lo cubra o bien no disponga de una interfaz pública para su uso.
7. Diseño y mantenimiento del sistema complejo (cada proveedor de servicio proveerá su propio método de interconexión, distinto de todos los demás).
8. Poco o ningún control sobre los problemas que puedan surgir durante el análisis.

- 2) **Aplicación de análisis:** Otra posibilidad es la realización de una aplicación dedicada, entendiendo 'de análisis' como que el análisis se realiza dentro de la misma y no mediante servicios de terceros. Se requeriría la construcción de una arquitectura software que permita ser ampliada mediante la codificación de criterios de análisis para adaptarse a los criterios que el usuario requiera. Sería conveniente proveer este tipo de aplicación con una serie de análisis precargados típicos.

✓ **Ventajas**

1. Flexibilidad proveniente de la definición personalizada de los criterios de análisis para los recursos web.
2. Control de las operaciones de descarga de los recursos web ya que se realizarían mediante una araña web controlable programáticamente.
3. Facilidad para la integración de la aplicación con software adicional (como por ejemplo para el indexado de los resultados en un software de buscador) ya que se dispone de control sobre todas las fases de ejecución.
4. Al ejecutar los análisis en la propia aplicación, posibilidad de escoger las herramientas o librerías para el análisis mas convenientes.
5. Posibilidad de controlar todos los procesos de tal modo que se pueda tener un mejor control de los posibles errores en el análisis.

x **Inconvenientes**

1. Las tareas de análisis se realizarían sobre la propia máquina por lo que aumentaría la carga de trabajo y los requerimientos.
2. Necesidad de codificar los distintos algoritmos de análisis.
3. Con probabilidad, se necesitará adquirir conocimientos específicos sobre alguna plataforma que realice la descarga de los recursos web.

- 3) **Aplicación de análisis distribuido:** Partiendo del planteamiento del punto anterior de una aplicación de análisis, se podría descargar el análisis en un sistema de computación distribuido. Esto es ideal para dar dar respuesta al requisito de dar soporte al análisis sobre un gran número de portales web ya que se dispondría de un nodo central donde se ejecutaría la aplicación y los análisis se distribuirían entre distintos nodos (si así se configura la infraestructura).

Adicionalmente a las ventajas descritas para la alternativa del punto anterior, se añaden.

✓ **Ventajas**

1. No se realizaría la computación del análisis en una máquina única por lo que se reduciría la carga.
2. El software de computación distribuido suele permitir procesar una gran cantidad de información (en función de los nodos del clúster de computación).
3. Fiabilidad ante fallos. Si los análisis se ejecutaran en una misma máquina y esta fallase, no se podría ejecutar nada mas mientras que si la computación es distribuida, es probable que siempre haya alguna máquina disponible para computar el análisis.

4. Escalabilidad en la capacidad de computación mediante la adición de nodos al clúster según se requiera.

**x Inconvenientes**

1. Necesidad de codificar los distintos algoritmos de análisis.
2. Con probabilidad, se necesitará la adquisición de conocimientos de tecnologías como la del software que realice la descarga de los recursos.
3. Complejidad del despliegue y configuración de sistemas distribuidos.
4. El modelo de computación distribuido hace que el diseño y la codificación de aquellas partes del código que se vayan a ejecutar de forma distribuida sea mas compleja.
5. Necesidad de múltiples máquinas para afrontar grandes cargas de trabajo (nodos del clúster).
6. Dependencia de comunicaciones por red entre máquinas del clúster de computación (si existen varios nodos).

Todas las alternativas podrían materializarse tanto en aplicaciones de escritorio, como web, pero debido a la naturaleza de los estudios en los que se enmarca este proyecto parece mas pertinente el desarrollo de una aplicación web.

Se considera que para satisfacer los objetivos y requisitos, es necesario dar soporte a una carga de trabajo elevada de análisis (especialmente cuando se requiera realizar un análisis sobre un gran número de portales web) por lo que para ello se hace necesario adoptar la solución de la aplicación propia con ejecución de análisis sobre plataforma de computación distribuida.

# 3 Aspectos técnicos

## 3.1 Aplicación web

Como se especificó en el apartado 2.4, se consideró que era conveniente el desarrollo de una aplicación web como interfaz principal entre el usuario y la funcionalidad de análisis que este proyecto cubre. El cometido de la aplicación web, que no es otra cosa que un programa informático con el que se interactúa a través de un navegador web, es dar acceso a los usuarios a las funcionalidades de análisis que cubre el proyecto. WebKumo, como ya se comentó con anterioridad, es el nombre que se ha dado a dicha aplicación web desarrollada.

### 3.1.1 Java Enterprise Edition

El proyecto se ha desarrollado en el lenguaje de programación Java, utilizando la plataforma de desarrollo *Java Enterprise Edition (JEE)* que es una plataforma bien conocida de desarrollo de software de aplicaciones empresariales orientadas a la web desarrollada por Sun Microsystems (ahora Oracle).

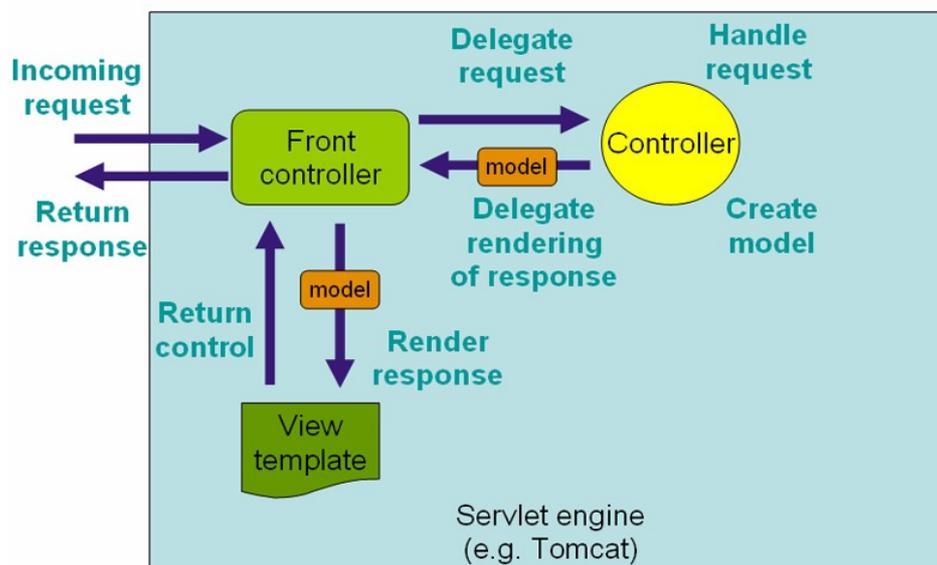
Esta plataforma, que es sobre la que principalmente se desarrollan las actividades realizadas durante los estudios de este máster, proporciona una serie de APIs para desarrollar software con el foco puesto en la modularidad, la escalabilidad y la seguridad. Es una plataforma de estilo CoC **[COC]** en la que se prima la configuración programática de las distintas APIs mediante anotaciones Java.

Se han utilizado entre otras, las APIs que gestionan ciclos de peticiones HTTP (Java Servlet API **[JSERVLET]**), la persistencia mediante software ORM **[ORM]** (Java Persistence API **[JPA]**) o la seguridad de la aplicación web (Java Authentication And Authorization API **[JAAZ]**).

### 3.1.2 Patrón arquitectónico MVC.

Al implementar una aplicación web utilizando la plataforma JEE, casi es imprescindible comentar el patrón arquitectónico clásico MVC **[MVC]** (*Modelo-Vista-Controlador*) que es el que se ha utilizado para diseñar la arquitectura software de la aplicación web desarrollada con el fin de crear una aplicación de sencillo mantenimiento.

Este patrón software se fundamenta en el principio de separación de responsabilidades, consistente en la diferenciación de tres partes bien definidas en la aplicación. El *modelo* consiste en un conjunto de información y reglas básicas que definen el ámbito de la aplicación (típicamente formado por datos almacenados en una base de datos y una serie de entidades o clases). El *controlador* (o *controladores*) se encarga de recibir las peticiones por parte del usuario y tomar las acciones oportunas para completar su petición. La vista finalmente proporciona la interacción efectiva entre el usuario mediante el accionamiento de distintos componentes de interfaz así como mediante la presentación de información del modelo o de resultados de acciones procesadas por la aplicación.



**Fig. 1:** Descripción esquemática del patrón arquitectónico MVC.

### 3.1.3 Pivotal Spring Framework

Por lo general, al implementar una aplicación utilizando la plataforma JEE, no se suele realizar directamente utilizando las API de la especificación (aunque sí que se utilizan algunos de sus componentes directamente). En este proyecto, la aplicación web se ha apoyado en el conjunto de frameworks de desarrollo web Spring de Pivotal.

Se trata de un framework de aplicaciones y un contenedor IoC **[IOC]** no estrictamente orientado a la web que pone el acento en proveer una manera sencilla de utilizar la plataforma *JEE*. Además es un framework abierto a la integración con otros componentes y tecnologías típicas del ecosistema Java como diversas implementaciones de JPA **[JPA]**, servicios web, JNDI **[JNDI]** e incluso otros frameworks de desarrollo de aplicaciones web como Struts **[STRU]**.

El framework está compuesto por una serie de componentes que permiten desarrollar una estructura consistente a lo largo de toda la aplicación que incluyen aspectos como la gestión de la persistencia de la aplicación, programación orientada a aspectos, implementación propia del patrón MVC, desarrollo de aplicaciones Big-Data etc.

Concretamente, para el desarrollo de este proyecto se utilizaron los siguientes componentes del framework Spring.

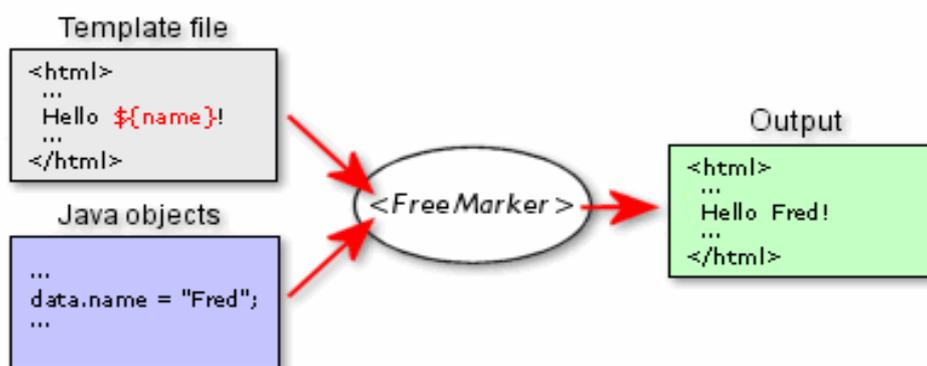
- ◆ **Spring Framework:** El núcleo del framework que proporciona la funcionalidad común.
- ◆ **Spring MVC:** Para el desarrollo de aplicaciones con arquitectura MVC.
- ◆ **Spring Data:** Gestión de la persistencia de la aplicación.
- ◆ **Spring Security:** Gestión de la seguridad de la aplicación y la autenticación.

### 3.1.4 FreeMarker

La parte de vista, en la aplicación web MVC desarrollada, se ha implementado utilizando un motor de plantillas llamado FreeMarker orientado a la generación dinámica de las vistas en una aplicación MVC y a la generación de plantillas textuales en general. Posee integración con el framework Spring utilizado en la aplicación web y está publicado bajo la licencia de software libre Apache versión 2.0.

La misión básica de las plantillas es combinar los datos que los controladores exponen a las vistas con el código HTML de la vista en sí. Para ello proporciona una serie de etiquetas propias (del estilo `<#etiqueta>`) que se mezclan con el código HTML nativo para realizar operaciones básicas relacionadas como iteraciones, sentencias condicionales, operaciones sobre cadenas de texto etc. Además proporciona otras características útiles, como mecanismos de internacionalización o utilización de librerías de etiquetas de JSP **[JSP]** (para casos puntuales).

FreeMarker permite sustituir completamente los documentos JSP aportando ciertas ventajas significativas como la no necesidad de compilar las plantillas, el anidamiento de plantillas en tiempo de ejecución, llamadas a funciones en etiquetas, manejo de valores nulos y cadenas de texto vacías, soporte para JSON **[JSON]** etc.



**Fig. 2: Esquema de funcionamiento de FreeMarker.**

### 3.1.4 Hibernate

Además de gestionar mediante Spring Data la persistencia de la aplicación web, se utilizó el software de mapeo objeto-relacional **[ORM]** Hibernate para la persistencia de las entidades que maneja la aplicación web. Este software, implementación de la *Java Persistence API* **[JPA]**, permite mediante anotaciones Java (o configuración XML **[XML]**) marcar las entidades que se han de persistir y cómo se deben persistir en una base de datos. Se ofrece bajo licencia de software libre LGPL y es desarrollado principalmente por Red Hat.

El principal cometido es realizar la traducción entre el modelo de clases típico del modelo de datos de una aplicación Java y las tablas con las que correspondería dicho modelo de datos en un modelo relacional como el que utilizan los SGBD **[SGBD]** mas habituales (como *MySQL* ó *PostgreSQL*). Permite las operaciones típicas CRUD **[CRUD]** sobre entidades mediante una API programática (definida por *JPA*) así como un lenguaje de propio *HQL* **[HQL]** para consultas mas complejas y la definición programática de consultas mediante *Criteria Queries* **[CRITQ]**.

### 3.1.5 Apache Hadoop

En el apartado 2.4 se optó por desarrollar una aplicación de análisis en la que los análisis se ejecutaran de forma distribuida, para ello se requiere de un software para dicho fin. Apache Hadoop es una plataforma de computación distribuida escrita en Java capaz de procesar volúmenes muy grandes de datos sin requerir hardware especializado. Además comprende un sistema de archivos distribuido *HDFS* **[HDFS]** que los nodos del clúster que ejecute Hadoop puedan usar y un software para gestionar los trabajos en los nodos (*YARN*).

El funcionamiento básico de Hadoop es la ejecución de trabajos (principalmente en Java) escritos según el modelo de programación distribuida MapReduce **[MRED]**. Estos trabajos son gestionados por Hadoop siendo distribuidos automáticamente entre los nodos.

En la actualidad existen dos ramas mantenidas de Apache Hadoop, la rama 1.X que es la inicial, basada en Hadoop 0.20 y la rama 2.X basada en Hadoop 0.23 que incorpora *YARN* para la gestión de nodos en lugar de MapReduce1 que es el software de gestión de las tareas de computación de la rama 1.X.

Ambas ramas son capaces de ejecutar trabajos MapReduce pero la rama 2.X está abierta a la posibilidad de ejecutar otros modelos de computación como *BSP* [BSP]. Para este proyecto se utilizó la última versión de la rama 1.X, la **1.2.0** por motivos que se describirán mas adelante (ver punto siguiente 3.1.6).

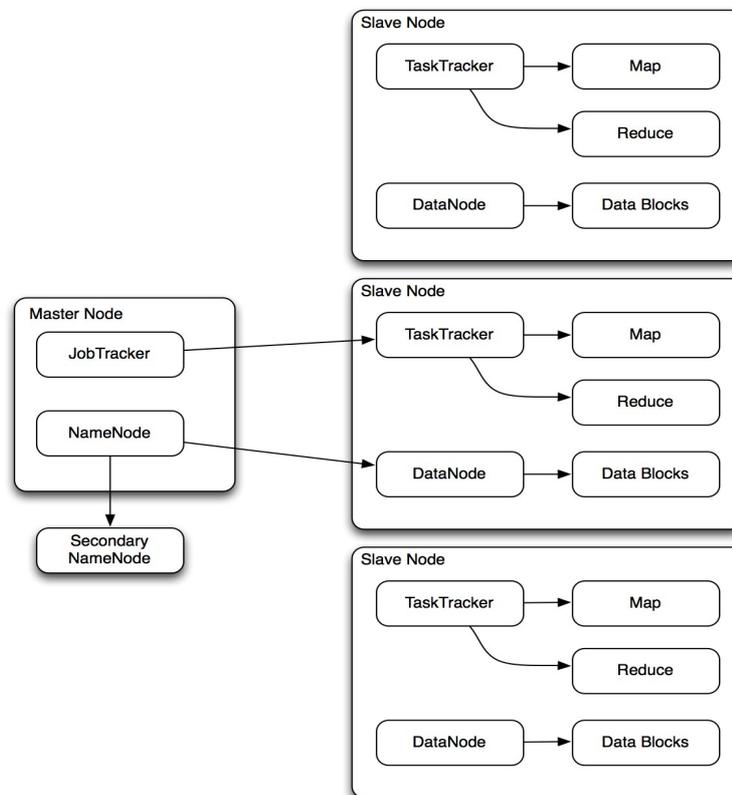
Apache Hadoop basa buena parte de su capacidad de cómputo en la proximidad de los datos [YAH01], procurando enviar así las tareas de computación a aquellos nodos que contienen los datos (en el sistema de archivos distribuido HDFS) . Este principio se utiliza también en la replicación de datos, intentando replicarlos en aquellos nodos a los que se les ha asignado los trabajos que usan dichos datos. Si no fuera posible enviar el trabajo al mismo nodo, Hadoop intentará enviarlo a otro próximo a él (los nodos están etiquetados según *rack* ó *switch*).

Como se puede deducir de lo ya comentado, Apache Hadoop está pensado para gestionar un clúster de máquinas. Los distintos modos de funcionamiento [HADO1] incluyen un modo con un nodo único con sistema de archivos local (*Standalone*), un nodo único con sistema de archivos distribuido (*Pseudo-distributed*) y un modo con un clúster de múltiples máquinas [HADO2] (Cluster mode). El desarrollo de este proyecto ha sido llevado a cabo en un modo *Pseudo-distributed* con una máquina proporcionada por la Escuela de Ingeniería Informática.

Los componentes principales que componen Apache Hadoop son:

- ◆ **JobTracker:** Se encarga de distribuir los trabajos a distintas máquinas del clúster siguiendo el principio de proximidad de los datos.
- ◆ **TaskTracker:** Nombre dado a cualquier nodo del clúster que acepta trabajos.
- ◆ **Namenode:** Es el equivalente al JobTracker pero para el sistema de archivos distribuido HDFS. Coordina a los nodos en su acceso al sistema de archivos y lleva la traza de la localización de los distintos bloques de datos a lo largo del clúster.
- ◆ **Datanode:** Equivale al TaskTracker pero para el sistema de archivos. Es un nodo que está disponible para operaciones HDFS sobre él.
- ◆ **Secondary Namenode:** Realiza operaciones de soporte al Namenode como la actualización de los *logs* de operaciones y la actualización persistente de las localizaciones de los ficheros en el clúster.

La ejecución de un trabajo de análisis, a tenor de lo anteriormente visto, consistiría en la remisión al nodo maestro del clúster de Apache Hadoop un trabajo MapReduce (empaquetado como *JAR [JAR]*) que el nodo maestro de Apache Hadoop se encargaría de distribuir junto con sus datos en el clúster.



**Fig. 3: Esquema de un clúster de Apache Hadoop típico.**

### 3.1.6 Apache Nutch

Para poder llevar a cabo el análisis de los recursos web, se precisa de una herramienta que realice la descarga de los mencionados recursos. Una posible opción sería escribir trabajos *MapReduce* de Hadoop en los que se utilice alguna librería liviana (como Apache HTTP Components [*HTCOM*]) para ejecutar la descarga de los recursos web. Sin embargo existe una solución mas apropiada, que es la utilización de Apache Nutch, software libre desarrollado por la Fundación Apache.

Apache Nutch es un software de araña web o *crawler*, es decir, realiza la descarga de los distintos recursos web que componen un portal en base a una lista inicial de URIs llamadas semillas.

Se trata de un software que se ejecuta sobre la plataforma Apache Hadoop (de hecho, Apache Hadoop surge debido a diversos problemas encontrados para ejecutar Apache Nutch), por lo que se puede ejecutar de forma distribuida en un clúster aprovechando las ventajas que aporta en cuanto a la cantidad de información que puede procesar. También es altamente escalable, configurable y modular gracias a su sistema de *plugins* **[NUTCH1]** que permite añadir funcionalidad al proceso de descarga de manera relativamente sencilla. Los documentos descargados (o al menos sus metadatos) son almacenados en un sistema de persistencia (típicamente) no relacional. Opcionalmente se puede realizar el indexado de los documentos en un software de buscador.

Los sistemas de persistencia que soporta Apache Nutch mas típicamente utilizados son Apache Hbase **[HBASE]**, Apache Cassandra **[CASSA]** y Apache Accumulo **[ACCUM]**. Como sistemas de buscador soportados por Nutch para realizar el indexado, el mas comúnmente utilizado es Apache Solr **[SOLR1]** pero también se dispone de soporte para otros software de buscador populares como Elasticsearch **[ESEAR]**.

Gracias a que se ejecuta como un trabajo de Apache Hadoop, se elimina la necesidad de escribir manualmente trabajos *MapReduce*, restando así complejidad al desarrollo y reduciendo la posibilidad de errores en ese sentido derivados de la poca experiencia en la utilización de dicho modelo de programación.

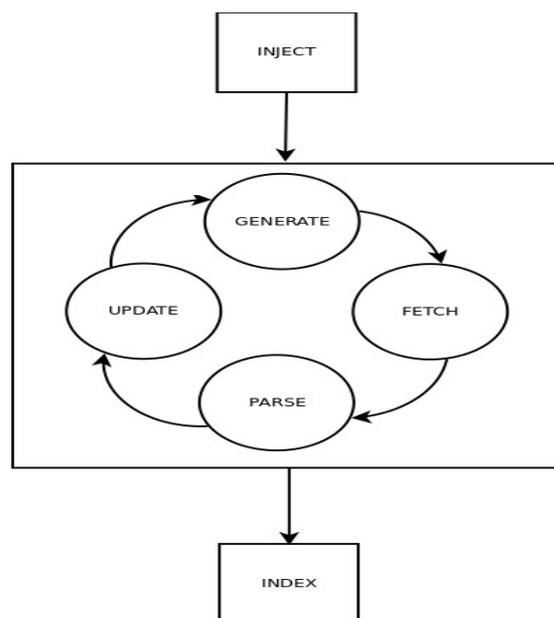
Como ocurría en el caso de Apache Hadoop, existen dos ramas de desarrollo de Nutch, la rama 1.X y la rama 2.X. La diferencia entre ambas es que la rama 2.X gestiona la persistencia de los elementos descargados utilizando Apache Gora **[GORA]**, un software análogo a un *ORM* pero indicado para sistemas de persistencia no relacionales. En cambio, la rama 1.X integra la gestión de la persistencia embebida en el propio Apache Nutch, dificultando la flexibilidad a la hora de almacenar nuevos campos o simplemente cambiar un sistema de persistencia por otro. Salvo que se requiera por algún motivo especial el uso de la rama 1.X, es recomendable por lo que se acaba de comentar, usar la rama 2.X. que es la que se ha utilizado en este proyecto (en concreto la **versión 2.2.1**).

La **versión de Apache Hadoop** que se recomienda **[NUTCH2]** para ejecutar la versión de Nutch utilizada es la **versión 1.2.0**, de la rama 1.X. Se empaquetan junto con Nutch, los ficheros JAR básicos para dar la posibilidad de ejecución del software mediante línea de comandos independientemente de otras instalaciones de Hadoop.

Uno de los puntos fuertes de Apache Nutch es su flexibilidad gracias a su sistema de conectores o *plugins*. Estos conectores son piezas software que se integran en las distintas fases del ciclo de ejecución de Nutch, permitiendo realizar operaciones según las necesidades del desarrollador.

En este proyecto, se desarrolló uno de estos *plugin*, denominado **WebKumo Connector** que inyecta en la fase de procesado de los recursos (*parsing*) los algoritmos de análisis que se hayan codificado previamente de tal modo que estos análisis tienen acceso a los recursos (HTML, CSS y JavaScript principalmente, aunque también podrían descargarse y procesarse recursos no textuales mediante Apache Tika [**TIKA**], aunque esto queda fuera del ámbito del proyecto) descargados para aplicar sobre su código fuente las comprobaciones pertinentes.

El sistema de funcionamiento de Nutch se basa en un ciclo de ejecución consistente en una serie de trabajos *MapReduce* individuales. Existen múltiples opciones de combinar estas fases de ejecución, siendo el ciclo básico para realizar una operación de *crawling*, la que sigue:



**Fig. 4: Ciclo de ejecución básico de Apache Nutch**

Las fases se detallan a continuación:

1. **Inject:** Fase de inyección en la que Apache Nutch toma o obtiene las URIs iniciales (semillas) a partir de las cuáles se va a realizar el proceso de *crawling*.

2. **Generate:** En esta fase, se genera el trabajo para Apache Hadoop (empaquetado en un JAR) en función de las semillas y diversas opciones de configuración que se pueden definir.
3. **Fetch:** Fase de descarga de los recursos en sí.
4. **Parse:** Fase de procesado de los recursos descargados. En esta fase se procesa el código fuente de los recursos extrayendo sus metadatos y los enlaces que contienen.
5. **Update:** Se actualiza el sistema de persistencia subyacente con la información descargada y procesada.
6. **Index:** Los campos configurados para ser indexados, son enviados al motor de búsqueda.

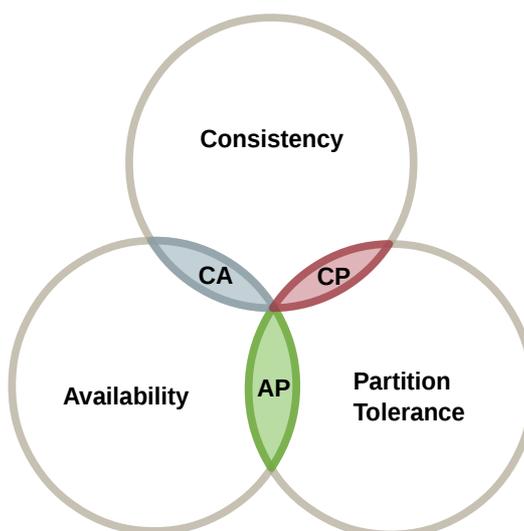
Este es un proceso incremental en el que se itera por las fases 2, 3, 4 y 5. En la primera iteración se descarga el recurso que se encuentra en las URIs semilla, se obtienen (en la fase *parse*) los enlaces a otros recursos web que se encuentran en el código fuente. Estos enlaces son añadidos (en la fase *update*) a la cola de descargas de tal modo que en la siguiente iteración, la generación del nuevo trabajo (fase *generate*) incluya los nuevos recursos detectados. De este modo se consigue descargar el portal completo en sucesivas iteraciones.

Aunque existen otras alternativas de software de araña web, Apache Nutch aporta muchas ventajas gracias a su diseño modular y a su ejecución sobre la famosa plataforma de computación distribuida Apache Hadoop; permitiendo así realizar de forma distribuida la descarga y el análisis de los recursos, facilitando de este modo la escalabilidad general y la capacidad de cómputo del sistema de procesado.

### 3.1.7 Apache Cassandra

Como se comentó en el punto anterior, Apache Nutch requiere un software de persistencia donde almacenar los datos y/o metadatos (según configuración) de los recursos web que se vayan obteniendo. Para este proyecto se escogió Apache Cassandra, software libre desarrollado por la Fundación Apache, como *backend* de persistencia para Apache Nutch.

Apache Cassandra es un sistema de gestión de bases de datos NoSQL **[NOSQL1]** de columna ancha **[NOSQL2]** (aunque mezcla aspectos de base de datos documental y tabular), altamente escalable incluyendo *sharding* **[SHARD]** (escalabilidad horizontal), tolerante a fallos y con una alta disponibilidad. Es una base de datos no relacional de tipo AP (prima la disponibilidad y el particionamiento o *sharding*, frente a la consistencia) según el teorema CAP **[CAP]**.



**Fig. 5: Teorema CAP.**

Se escogió este sistema de persistencia NoSQL debido a su mayor rendimiento en operaciones de escritura **[CASBEN]** (la operación más frecuente ya que se trata de descargar recursos web y almacenar información al respecto mucho más que leer su contenido desde el *backend* de persistencia) frente a sus competidores más directos (Apache Hbase y Apache Accumulo) para su utilización en Apache Nutch.

### 3.1.8 Apache Solr

Aunque no es estrictamente necesario para el funcionamiento de Apache Nutch, puede ser conveniente indexar los resultados de las descargas (no los análisis) en un motor de búsqueda. En este proyecto se ha utilizado Apache Solr, parte del proyecto Lucene desarrollado por la Fundación Apache para dicho fin. Está disponible bajo licencia de software libre Apache versión 2.0.

Apache Solr permite búsquedas por texto completo, es altamente escalable y tolerante a fallos (si se requiere se puede desplegar de forma distribuida mediante Solr Cloud **[SOLR2]**), proporciona indexado en casi tiempo real, búsqueda por facetas y es modular mediante un sistema de conectores o *plugins*, entre otras prestaciones. Se integra fácilmente con Apache Nutch, el cuál posee soporte nativo para indexar las descargas en este motor de búsqueda.

En este proyecto se utiliza para ofrecer de manera simplificada el código fuente de los recursos descargados al usuario en caso de requerir un examen manual de alguno de los mismos. También proporciona una interfaz web de uso sencillo para la indagación manual en los datos indexados.

# 4 Planificación del proyecto

## 4.1 Planificación inicial

Como parte de estos estudios de Máster, en la asignatura *Gestión y Dirección de Proyectos Web*, se realizó la planificación inicial del desarrollo del proyecto. Se incluyeron diversos elementos en dicha planificación, como son el desglose de tareas, la planificación temporal de las mismas, un cálculo aproximado del presupuesto en base a horas de trabajo y recursos y otras actividades adicionales relacionadas (como la hipotética aplicación de las pautas de PMBok **[PMBOK]** al proyecto o un análisis de riesgos).

La planificación inicial comprendía la planificación desde el planteamiento del proyecto en Julio de 2014 hasta la prevista defensa del Trabajo Fin de Máster en Julio de 2015. Entre medias se planificaron reuniones, tareas de consideración de alternativas, configuraciones de despliegue, tareas de diseño, tareas de análisis, de codificación, de elaboración de documentación, de pruebas y otras tareas accesorias (administrativas, impresión de documentos etc.). Para facilitar la partición en tareas se establecieron hitos intermedios en distintas fases, como al terminar la consideración de alternativas técnicas, al finalizar ejecuciones de prueba de las distintas alternativas, en el fin de la codificación, pruebas etc.

Dentro de lo inicialmente previsto se contemplaba un receso en el desarrollo de un mes desde el 9 de enero de 2015 hasta el 15 de febrero con motivo de la realización de prácticas en empresa como parte del plan de estudios del Máster.

Esta planificación está constituida por 62 tareas, con un plazo de ejecución del proyecto estimado de 371 días y un presupuesto estimado de 55,000€.

Esta planificación fue entregada junto al anteproyecto en Enero de 2015 por lo que se evita reproducir aquí de nuevo dicha planificación al completo. Por si fuera requerida **se entregará anexa dicha planificación.**

## 4.2 Cambios en la planificación

La planificación inicial se vio afectada por un cambio en la planificación de las prácticas en empresa que finalmente fueron cambiadas al período de entre el día 2 de febrero y el día 30 de marzo, obligando a replanificar parcialmente los esquemas inicialmente planteados. Se modificó el calendario del proyecto eliminando el período 10 de Enero a 15 de Febrero, no laborable y estableciendo en su lugar el período 2 de Febrero a 30 de Marzo como laborable pero sólo con 4 horas de jornada.

Por suerte este cambio no fue insalvable para el desarrollo del proyecto ya que aún así se logró dedicar al menos 4 horas diarias de trabajo en ese período. Además se habían sobrestimado la duración de ciertas tareas como por ejemplo la 42 (y su subtarea 43) de implementación de la capa de persistencia de la aplicación que pasó de 20 a 12 días de duración o la tarea 50 (y su subtarea 51) de codificación de algoritmos de análisis que pasó de estimarse en 30 días a hacerlo en 20 a medida que se iba realizando la codificación de los algoritmos.

Por otra parte surgieron nuevas tareas que habían quedado fuera de la planificación inicial como la tarea 46 (y su subtarea 47) consistente en la preparación de una instancia local de la infraestructura que soporta este proyecto (Apache Hadoop, Apache Nutch etc.) en el ordenador de desarrollo como forma conveniente de realizar pruebas y depuraciones de forma rápida.

Otro cambio relevante fue el diseño de la interfaz gráfica, inicialmente previsto para el final de la codificación (mientras tanto se tendría una estructura básica pero no se iría profundizando en ella). Finalmente se fue realizando en paralelo a las otras tareas a medida que se fue codificando la aplicación en sí (tarea 52 y subtarea 53).

Se consideró también pertinente la eliminación de la tarea 67 (en la planificación inicial) consistente en la inclusión de un apartado que mostrara los términos de uso de la aplicación. Finalmente se eliminó esta tarea debido a que la aplicación no está pensada para un uso público por lo que no existen unos términos generales a redactar.

Por otra parte se añadió la tarea de codificación no prevista en la planificación inicial (con el número 49 en la replanificación) consistente en el desarrollo del conector o plugin de Apache Nutch, *WebKumo Connector*. No se incluyó esto en la planificación inicial porque en aquel momento no se tenía claro que la funcionalidad de análisis fuera a estar integrada en la propia araña web.

Adicionalmente se había establecido incorrectamente la duración de la elaboración de ciertas tareas en la fase final del proyecto. Esta planificación es incorrecta ya que los plazos administrativos fijan límite para la entrega del proyecto y su documentación el día 6 de Julio (para la defensa estimada en la convocatoria de Julio). Inicialmente estas tareas estaban previstas para finalizar el día de la defensa del proyecto, estimada hacia el 17 de Julio de 2015.

Estas tareas son (se especifica el número de tarea **en la planificación inicial**):

- ◆ **Tarea 51 (y subtareas):** La documentación debe entregarse junto con el resto del proyecto antes de la fecha límite, no siendo posible la extensión de dichas tareas hasta el día previsto de la defensa.
- ◆ **Tarea 68:** La elaboración del presupuesto final ha de ser incluida en la documentación por lo que es imposible que pueda finalizarse después del día 6 de Julio de 2015. Además se movió como subtarea de documentación en lugar de subtarea de finalización.
- ◆ **Tareas 61 a 65:** Estas tareas comprenden parte del plan de pruebas que deben estar incluidas en la documentación por lo que deben finalizarse antes de la fecha límite.

Además de por la fecha límite de entrega de documentación del día 6 de Julio, también estaba mal programada la parte final (accesoria, de trámites) denominada "Finalización". Se tenía planificada la defensa para el día 27 de Julio (de forma aproximada), sin embargo, el plazo de defensa de la convocatoria es entre el día 13 y 17 de Julio.

Continuando con los cambios introducidos, se modificó la tarea 61 (58 en la planificación inicial) que contemplaba la corrección de problemas detectados al realizar pruebas unitarias. Como la realización de las mismas es continuo a lo largo de toda la codificación, se modificó esta tarea ya que debía ser llevada a cabo paralelamente durante todo el período de desarrollo.

Se advirtieron a medida que fue avanzando el proyecto, otras tareas adicionales que habían perdido su razón de ser. Se eliminó la tarea que se había previsto para imprimir la documentación (en esta convocatoria basta con entregarla en formato digital). También se mostraron prescindibles las tareas de planificación y ejecución de pruebas software de integración ya que finalmente el proyecto fue llevado a cabo por un sólo proyectante.

Otros cambios a destacar, tienen que ver con la gestión de recursos. En la planificación inicial se habían previsto 150€ de costo de material para la impresión y encuadernación de los tomos de documentación. Finalmente este costo se eliminó ya que el año en el que se presenta este proyecto la normativa ha variado y ya no es necesaria dicha impresión.

Así mismo, relacionado con los recursos, se añadieron nuevos roles al proyectante en la planificación. En la estimación inicial, el proyectante tomaba un rol único independientemente de las tareas que realizara. Se añadieron los perfiles de programador, analista, probador de software (*Quality Assurance*), director de proyecto, administrador de sistemas y formador. Se asignaron estos recursos a sus tareas correspondientes.

Los calendarios que se han aplicado a la planificación y que determinan el número de horas junto con las tareas son:

- ◆ **Verano (17/07/14 al 05/09/14):** Estimación de 4 horas de trabajo diarias.
- ◆ **Periodo lectivo (08/09/14 al 19/12/14):** Estimación de 2 horas diarias de trabajo.
- ◆ **Navidad (22/12/14 al 07/01/15):** No laborable.
- ◆ **Prácticas en empresa (02/02/15 al 30/03/15):** Estimación de 4 horas de trabajo diarias.
- ◆ **Semana santa (02/04/15 al 12/04/15):** No laborable.

Se adjunta igualmente el documento de replanificación como anexo al proyecto.

# 5 Análisis

## 5.1 Definición del sistema

En este apartado se cubrirá la parte de análisis del proyecto incluyendo aspectos como el alcance del proyecto, requisitos, casos de uso, subsistemas, interfaces y planes de prueba.

### 5.1.1 Determinación del alcance

Concretando lo ya expuesto anteriormente en esta documentación, este proyecto comprenderá una **aplicación web**, denominada **WebKumo** como se ha visto en apartados anteriores, además del despliegue de la plataforma de software que comprende lo visto en el apartado 3. Como base para el proyecto, se revela suficiente el despliegue de un solo nodo de los servicios software indicados, estos son, **un nodo de Apache Hadoop 1.2.0** en modo *Pseudo-Distributed*, **una instalación de Apache Nutch 2.2.1**, un nodo de **Apache Solr 4.10.3** y un nodo de **Apache Cassandra 2.0.2**. Además del despliegue de software anterior, se cubrirá también el desarrollo de **WebKumo Connector, un conector o plugin para Apache Nutch** que ejecute los análisis inyectados desde la aplicación web WebKumo. De estos elementos se extraerán posteriormente los subsistemas.

## 5.2 Requisitos del sistema

### 5.2.1 Obtención de los requisitos del sistema

Se definieron mediante la condensación de lo extraído en las reuniones con el cliente los requisitos que lo desarrollado en este proyecto debe cumplir. El planteamiento inicial es que el producto software será utilizado para diversos fines de análisis sin especificar de forma precisa cuáles inicialmente, pero sí esbozando el tipo de análisis típico (análisis de la internacionalización del un portal, cuantificación de imágenes en un sitio etc.).

Los usuarios finales serán personal de la Universidad de Oviedo que utilizarán lo desarrollado en este proyecto como base para realizar tareas de análisis automático de cara a proyectos de investigación y otras actividades docentes relacionadas.

Se especifican a continuación los **requisitos funcionales** en forma de tabla.

CÓDIGO	REQUISITO	DESCRIPCIÓN
R1.1	Roles de usuario	Se disponen dos perfiles de usuario, uno básico con uso restringido y otro administrador con plenos poderes.
R1.2	Agrupación de usuarios	Deben existir grupos de usuarios para separar los trabajos ejecutados por ellos.
R1.3	Gestión de permisos de usuarios	Los usuarios deben tener al menos restricciones para ver trabajos, crear trabajos y ver trabajos de otros usuarios de su mismo grupo.
R1.4	Dar de alta usuarios	Los usuarios con rol administrador deben poder dar de alta usuarios.
R1.5	Dar de baja usuarios	Los usuarios con rol administrador deber poder dar de baja usuarios.
R1.6	Modificar usuarios	Los usuarios con rol administrador deber poder modificar la información de usuarios existentes. Además los usuarios regulares deben tener la posibilidad de modificar sus propios datos (excepto su rol o pertenencia a grupo).
R1.7	Listar usuarios	Los usuarios con rol administrador deben poder

		ver la lista completa de usuarios dados de alta en el sistema.
R1.8	Listar grupos	Los usuarios con rol administrador deben poder ver la lista completa de grupos de usuarios dados de alta en el sistema.
R1.9	Dar de alta grupos	Los usuarios con rol administrador deben poder dar de alta grupos de usuarios.
R1.10	Modificar grupos de usuarios	Los usuarios con rol administrador deben poder modificar la información de los distintos grupos de usuarios.
R2.1	Listar servicios	Los servicios de los que depende la aplicación (araña web por ejemplo) deben poder listarse.

<b>CÓDIGO</b>	<b>REQUISITO</b>	<b>DESCRIPCIÓN</b>
R2.2	Arrancar servicios	Los servicios que puedan arrancarse, deben de poder arrancarse desde la aplicación informando al usuario del resultado de la operación.
R2.3	Parar servicios	Los servicios que así lo permitan, deben de poder detenerse desde la aplicación informando al usuario del resultado de la operación.
R3.1	Situación general	Debe permitirse obtener información sobre el sistema en general de forma rápida (tanto de servicios como de trabajos) mediante algún tipo de listado o panel.
R4.1	Dar de alta un trabajo de análisis	Los usuarios que tengan permisos para ello, deberán tener la posibilidad de crear nuevos trabajos de análisis mediante un formulario validado.
R4.2	Modificar un trabajo de análisis	Los parámetros de un trabajo que no haya sido ejecutado deben poder ser modificados.
R4.3	Ejecutar un trabajo de análisis	Los trabajos creados deben poder ser remitidos para su ejecución.
R4.5	Listar trabajos	Los trabajos visibles para cada usuario deben poderse listar mostrando su información mas relevante.
R4.6	Visualizar trabajos	Se debe permitir ver el estado actual de los trabajos así como el resultado de ejecución de los mismos. Los resultados deben ser mostrados mediante un informe con un porcentaje de cumplimiento de las características analizadas,

		puntuación total, descripción textual del mismo o similar. En la medida de lo posible, se presentará desglose detallado del informe de análisis.
R4.7	Visualizar el código fuente de los recursos analizados.	Se debe permitir el acceso rápido al código fuente de los recursos analizados.
R4.8	Añadir nuevos algoritmos de análisis.	Es necesario genera una arquitectura software adecuada que permita crear nuevos algoritmos de análisis mediante su codificación. Debido a la naturaleza heterogénea de los análisis a realizar, debe facilitarse el uso de librerías o software de terceros que los nuevos análisis pudieran requerir.

CÓDIGO	REQUISITO	DESCRIPCIÓN
R4.9	Criterios de parada	A la hora de ejecutar análisis, deben de establecerse unos criterios de parada en la descarga y/o análisis de los recursos. Al menos debe poder definirse criterios de parada por dominio y por profundidad en el portal.
R4.10	Análisis de ejemplo	Al menos debe proveerse un análisis implementado para ejemplificar el funcionamiento de la arquitectura.

En cuanto a los **requisitos no funcionales**, son los que se siguen:

1. **Requisitos de usuario:** No serán necesarios especiales conocimientos técnicos a la hora de utilizar la aplicación, sin embargo, por el tipo de usuario objetivo de este proyecto se tratará de usuarios con altos conocimientos técnicos. Sí que se requerirán conocimientos de programación Java para implementar nuevos algoritmos de análisis y de administración básica de sistemas. No serán necesarios conocimientos del funcionamiento de la infraestructura que ejecuta los trabajos.
2. **Requisitos tecnológicos:** Se precisa que la plataforma se ejecute sobre un sistema operativo GNU/Linux para su despliegue en un servidor. Preferiblemente distribuciones basadas en Debian.
3. **Requisitos de usabilidad:** Sin requisitos especiales en este sentido. Los usuarios de esta aplicación serán, con toda probabilidad, usuarios con un alto grado de conocimientos técnicos que no requieren especiales atenciones en el apartado de usabilidad.

4. **Requisitos de seguridad:** La aplicación requerirá autenticación para su uso. Adicionalmente sería recomendable implementar medidas de seguridad para prevenir los tipos de ataque mas comunes (inyección SQL [**SQLIN**] o ataque CSRF [**CSRF**] como ejemplo).

Como requisito adicional no funcional, la interfaz de usuario estará internacionalizada en Español e Inglés.

## 5.2.2 Identificación de actores

Considerando los componentes del sistema y la definición de requisitos, se tienen los siguientes actores:

1. **Usuario regular:** Persona que utilizará el sistema a través de la aplicación web WebKumo para realizar trabajos de análisis.
2. **Usuario administrador:** Persona con privilegios que además de realizar las operaciones de los usuarios regulares, podrá operar los servicios de los subsistemas y la gestión de usuarios.
3. **Backend de procesamiento:** Se considera la infraestructura de descarga y análisis que sustenta el sistema como un actor secundario. Esto comprende a Apache Hadoop, Apache Nutch, Apache Cassandra y Apache Solr.

## 5.2.3 Especificación de casos de uso

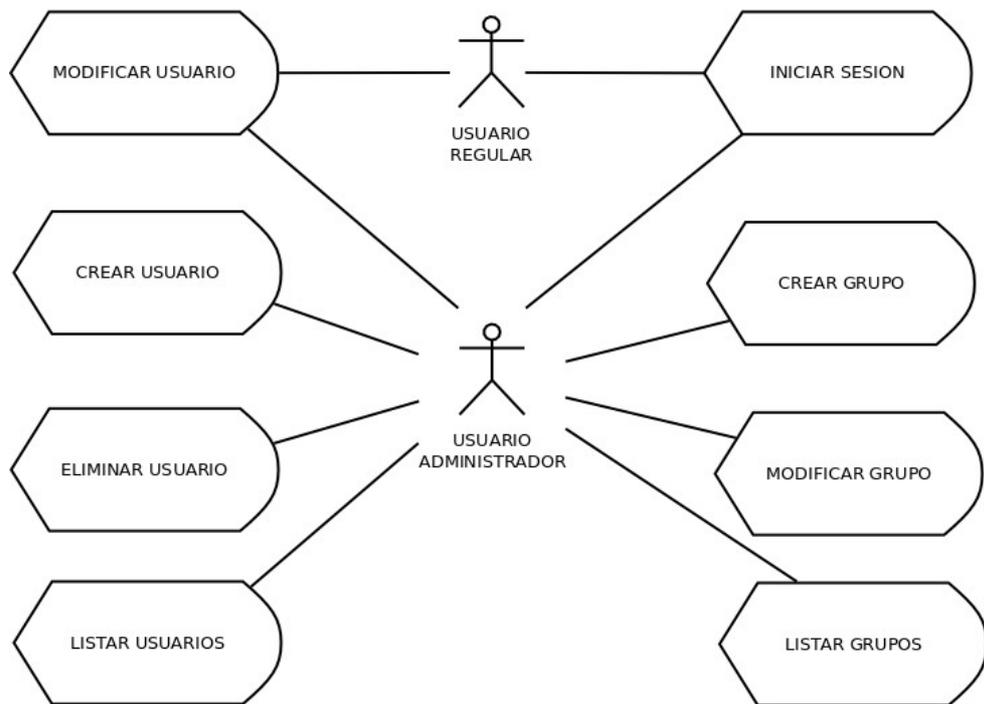
Se han identificado las siguientes agrupaciones de casos de uso:

- ◆ **Gestión de usuarios y grupos**
- ◆ **Gestión de servicios**
- ◆ **Gestión de trabajos**
- ◆ **Vista principal del sistema**

### 5.2.3.1 Gestión de usuarios y grupos

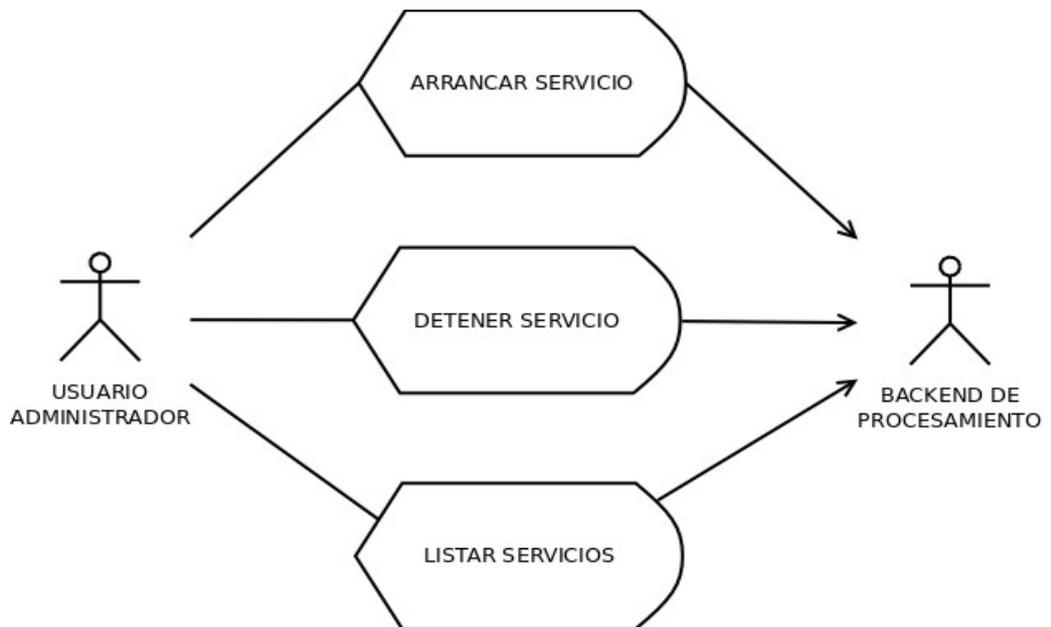
Se especifican ahora los casos de uso contenidos en el diagrama:

- ◆ **Crear Usuario:** Un usuario administrador rellenará un formulario de alta, que será remitido a la aplicación web y validado. Si todos los datos son correctos se dará de alta el usuario de manera efectiva. En caso contrario se mostrarán mensajes de ayuda al usuario para subsanar los errores.
- ◆ **Modificar usuario:** Se posibilitará la edición de los datos de un usuario mediante un formulario con validación. En caso de que los datos sean válidos se procederá a grabar los datos, en caso contrario se mostrará el oportuno mensaje de error.  
Los usuarios administradores podrán modificar cualquier otro usuario (además de sus propios datos). Los usuarios regulares únicamente tendrán permitido modificar sus propios datos.
- ◆ **Eliminar usuario:** Los usuarios administradores tendrán la opción de eliminar a cualquier otro usuario. Únicamente se podrá dar de baja un usuario si no tiene trabajos asociados.
- ◆ **Listar usuarios:** Se permitirá a los usuarios administradores el listado de todos los usuarios dados de alta en el sistema.
- ◆ **Crear grupo:** Al igual que ocurre con los usuarios, los administradores podrán generar nuevos grupos mediante formulario validado.
- ◆ **Modificar grupo:** Los usuarios con rol de administración deberán poder modificar la información de los grupos que estén dados de alto mediante un formulario validado.
- ◆ **Listar grupos:** Se debe permitir que los usuarios administradores listen los grupos que estén dados de alta en el sistema mediante lista, tabla o similar.
- ◆ **Iniciar sesión:** Para poder utilizar WebKumo el usuario debe identificarse mediante un formulario.



**Fig. 6:** Casos de uso de gestión de los usuarios y los grupos de usuarios.

### 5.2.3.2 Gestión de servicios

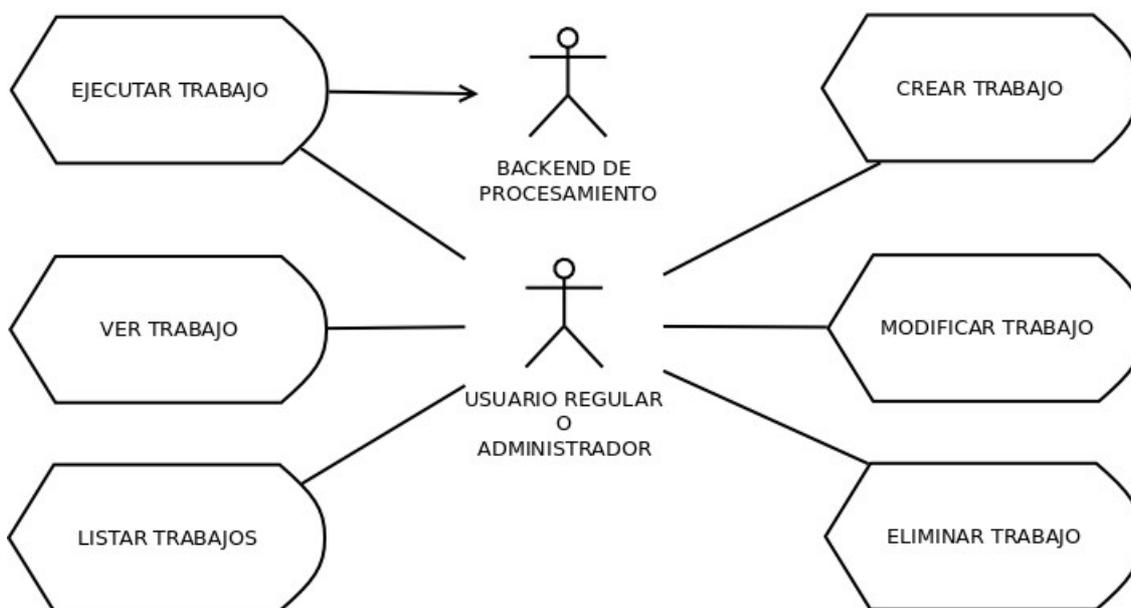


**Fig. 7:** Casos de uso de gestión de servicios

Se especifican los casos de uso contenidos en el diagrama a continuación:

- ◆ **Arrancar servicio:** Los usuarios administradores tendrán la posibilidad de arrancar aquellos servicios de la infraestructura que lo posibiliten. Se debe notificar al usuario del éxito o fallo en la operación.
- ◆ **Detener servicio:** Se deberá permitir la parada de aquellos servicios de la infraestructura que así lo permitan. No se podrán detener servicios si existen trabajos en ejecución. Al igual que en el caso del arranque del servicio, se debe indicar si la operación solicitada fue exitosa o no.
- ◆ **Listar servicios:** Existirá la opción de ver los servicios de infraestructura registrados y su estado (no se incluye en este caso de uso el requisito de ver información general sobre el sistema, esto se especifica como un caso de uso aparte).

### 5.2.3.3 Gestión de trabajos

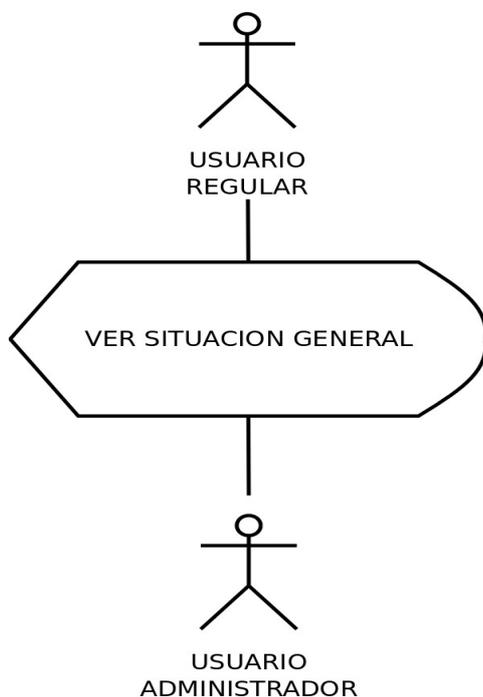


**Fig. 8: Casos de uso de gestión de trabajos**

A continuación se describen los casos de uso relacionados:

- ◆ **Crear trabajo:** Los usuarios administradores o los regulares que lo tengan permitido, crearán nuevos trabajos especificando los portales a analizar, el análisis y demás parámetros mediante formulario validado.
- ◆ **Modificar trabajo:** Aquellos usuarios que lo tengan permitido deberán poder modificar trabajos siempre que no hayan sido ejecutados. La modificación se realizará vía formulación con validaciones.
- ◆ **Eliminar trabajo:** Se dará la posibilidad de eliminar un trabajo previamente creado siempre que este no haya sido ejecutado todavía.
- ◆ **Ejecutar trabajo:** Para realizar el análisis de los portales web, se deberá permitir remitir los trabajos para su ejecución. Para ejecutar un trabajo, los servicios de infraestructura deben de estar arrancados, como es lógico.
- ◆ **Ver trabajo:** Es imprescindible que los usuarios puedan ver sus trabajos, incluyendo la información básica sobre los mismos (portales que analiza y los análisis que aplica por ejemplo) así como los resultados de la ejecución del trabajo si hubiera sido ejecutado ya.
- ◆ **Listar trabajos:** Los usuarios deben poder visualizar en forma de lista o similar los trabajos de análisis pertinentes con la información necesaria para identificarlos.

### 5.2.3.4 Vista principal del sistema



**Fig. 9: Caso de uso de visualización de la situación general del sistema**

Se describe el caso de uso de esta sección:

- ◆ **Ver situación general:** En relación al requisito R3.1, los usuarios podrán ver información general del sistema en forma de panel de usuario o similar. Incluirá información sobre los servicios y sobre los trabajos del usuario que esté usando la aplicación.

## 5.3 Identificación de los subsistemas en fase de análisis

### 5.3.1 Descripción de los subsistemas

Este proyecto tiene las siguientes áreas bien diferenciadas que pueden ser consideradas como subsistemas:

1. **Subsistema de usuarios:** Este subsistema proporciona la descripción y los mecanismos de los usuarios que utilizarán WebKumo y que ejecutarán los trabajos de análisis. En el modelo de datos se refleja en el paquete **[JPAC]** `Java org.rz.webkumo.model.users`.
2. **Subsistema de servicios:** Consiste en la gestión de los servicios, incluyendo su arranque, parada, atributos necesarios etc. Los componentes de este subsistema se agrupan en el paquete `org.rz.webkumo.model.services`.
3. **Subsistema de análisis:** Comprende no sólo los algoritmos de análisis sino también la generación de informes como manera de presentar los resultados de los análisis. El paquete Java que contiene este subsistema es `org.rz.webkumo.model.analyses` (y sus paquetes hijos) y el paquete `org.rz.webkumo.model.sites`.
4. **Subsistema de trabajos:** Este subsistema abarca la gestión de los trabajos de análisis de WebKumo. Lo forma el paquete `org.rz.webkumo.model.jobs`.
5. **Subsistema WebKumo Connector:** Este subsistema consiste en el conector plugin de Apache Nutch que realiza los análisis inyectados desde WebKumo en la fase *parse*. Este subsistema se encuentra en el paquete `org.rz.webkumo.connector`.
6. **Subsistema de infraestructura de procesamiento:** Consiste en el software que se ejecuta para dar soporte a las tareas de la araña web Apache Nutch. Comprende la plataforma de procesamiento distribuido Apache Hadoop, la araña web en sí, el *backend* de persistencia de la araña web Apache Cassandra y el motor de búsqueda que indexa lo que la araña web descarga, Apache Solr.

## 5.3.2 Descripción de interfaces entre subsistemas

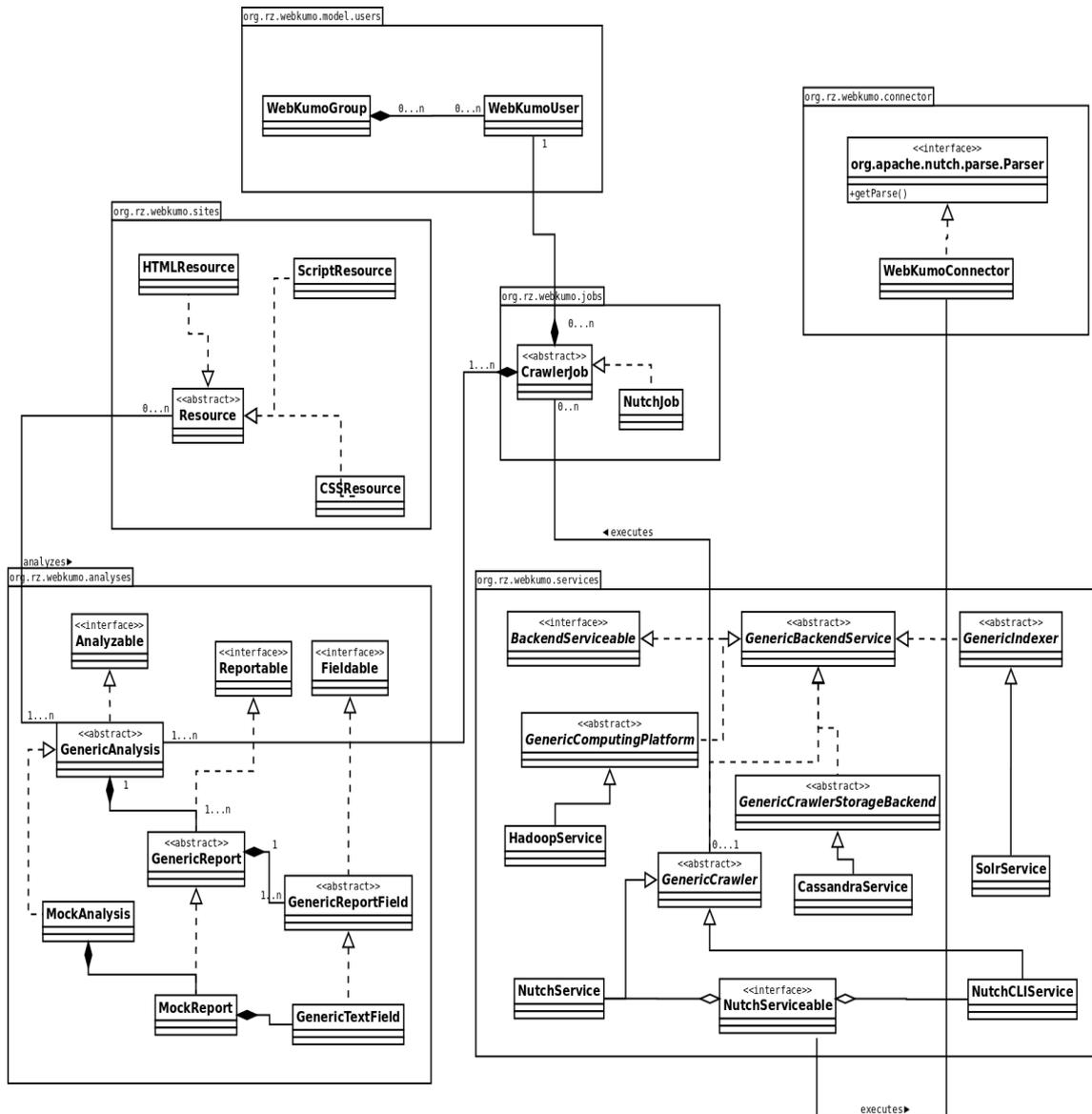
Los subsistemas 1, 2, 3 y 4 son parte de una misma aplicación J2EE por lo que se comunican entre sí dentro del contenedor web J2EE (Apache Tomcat, Oracle Glassfish etc.) mediante la propia API Java.

El subsistema 5, WebKumo Connector forma parte también de la aplicación por lo que se comunicará con el resto de subsistemas mediante la API Java. Sin embargo este subsistema no se ejecutará en el contenedor web anteriormente mencionado, sino que se ejecutará como trabajo MapReduce de Apache Nutch sobre el clúster de computación de Apache Hadoop.

Para el subsistema 6, las interfaces son las que siguen:

- ◆ **WebKumo con Apache Hadoop:** Se comunican mediante interfaz de línea de comandos **[HADO03]** para su puesta en marcha y detención.
- ◆ **WebKumo con Apache Nutch:** Se comunicarán mediante API Java **[NUTCH2]** y/o mediante interfaz de línea de comandos **[NUTCH3]**. Los resultados de la ejecución de algoritmos son serializados a JSON por parte de WebKumo Connector y escritos en el sistema de archivos distribuido HDFS de Apache Hadoop. Luego WebKumo toma de ese sistema de archivos los datos y los deserializa para traducirlos al modelo de clases y posteriormente persistirlos.
- ◆ **WebKumo con Apache Cassandra:** La comunicación se realiza vía API Java (para comprobar su estado usando un *driver* **[CASSA2]** específico) y mediante línea de comandos (para su arranque).
- ◆ **WebKumo con Apache Solr:** Se establece la comunicación mediante el servicio web REST **[SOLR3]** que Apache Solr proporciona.
- ◆ **Apache Hadoop con Apache Nutch:** Se relacionan mediante su API Java utilizando especialmente la utilidad *ToolRunner* **[HADO3]** de Apache Hadoop.
- ◆ **Apache Nutch con Apache Cassandra:** Se relacionan mediante comunicación vía red usando Apache Gora.
- ◆ **Apache Nutch con Apache Solr:** Se comunican mediante la API REST de Apache Solr.

## 5.4 Diagrama de clases preliminar de análisis



**Fig. 10: Diagrama de clases preliminar simplificado.**

El diagrama de la figura anterior es el modelo de clases preliminar necesario para sustentar la funcionalidad del sistema. Es un modelo, como su nombre indica, provisional, un borrador de la arquitectura software planteada. Se adjunta a esta documentación como anexo el diagrama original para su mejor inspección.

## 5.4.1 Descripción de las clases

Se detallan a continuación las clases contenidas en el diagrama de la figura 10. No se contemplan en las descripciones, por brevedad, los posibles métodos de acceso y modificación de atributos (*getters* y *setters*) ni demás métodos accesorios (*hashCode*, *equals* etc.).

### 5.4.1.1 Subsistema de usuarios

WebKumoUser	
<b>Descripción</b>	Entidad que representa un usuario del sistema.
<b>Responsabilidades</b>	Contener la información básica relativa a los usuarios y los métodos relacionados con su gestión.
<b>Atributos propuestos</b>	<ul style="list-style-type: none"> <li>◆ <b>login</b>: Nombre de usuario.</li> <li>◆ <b>name</b>: Nombre de pila del usuario.</li> <li>◆ <b>surname</b>: Apellidos del usuario.</li> <li>◆ <b>pass</b>: Contraseña de acceso.</li> <li>◆ <b>email</b>: Dirección de correo electrónico de contacto.</li> <li>◆ <b>role</b>: Rol del usuario (administrador o regular).</li> <li>◆ <b>permissions</b>: Permisos de usuario.</li> <li>◆ <b>groups</b>: Grupos a los que pertenece el usuario.</li> </ul>
<b>Métodos propuestos</b>	<ul style="list-style-type: none"> <li>● <b>getReadableJobs()</b>: Retorna los trabajos que el usuario puede ver (los suyos mas los de los usuarios de los grupos a los que pertenece si posee el permiso pertinente).</li> </ul>

WebKumoGroup	
<b>Descripción</b>	Entidad que representa un grupo del sistema.
<b>Responsabilidades</b>	Contener la información básica relativa a los grupos y los métodos relacionados con su gestión.
<b>Atributos propuestos</b>	<ul style="list-style-type: none"> <li>◆ <b>id</b>: Identificador de usuario.</li> <li>◆ <b>name</b>: Nombre del grupo.</li> <li>◆ <b>description</b>: Descripción del grupo (ej: grupo de investigación 1, usuarios del departamento etc.).</li> <li>◆ <b>members</b>: Usuarios que pertenecen al grupo.</li> </ul>

<b>Métodos propuestos</b>	<ul style="list-style-type: none"> <li>● <b>addMember()</b>: Añade un nuevo usuario al grupo.</li> <li>● <b>removeMember()</b>: Elimina un usuario del grupo.</li> </ul>
---------------------------	--

## 5.4.2 Subsistema de servicios

<b>BackendServiceable</b> <<interface>>	
<b>Descripción</b>	Interfaz base para todos los servicios de infraestructura.
<b>Responsabilidades</b>	Definir los métodos que todos los servicios deben implementar.
<b>Atributos propuestos</b>	Ninguno
<b>Métodos propuestos</b>	<ul style="list-style-type: none"> <li>● <b>start()</b>: Arranca el servicio (si el servicio se puede arrancar).</li> <li>● <b>stop()</b>: Detiene el servicio en ejecución (si el servicio se puede detener).</li> <li>● <b>isStartable()</b>: Define si el servicio se puede arrancar.</li> <li>● <b>isStoppable()</b>: Determina si el servicio se puede parar.</li> <li>● <b>isStarted()</b>: Obtiene información que determina si el servicio está arrancado.</li> <li>● <b>isAvailable()</b>: Determina si el servicio está disponible para su utilización.</li> </ul>

<b>GenericBackendService</b> <<abstract>>	
<b>Descripción</b>	Clase base para todos los servicios de infraestructura.
<b>Responsabilidades</b>	Definir atributos comunes a todos los servicios de infraestructura.
<b>Atributos propuestos</b>	<ul style="list-style-type: none"> <li>◆ <b>startable</b>: Booleano que determina si el servicio se puede arrancar o no (esto dependerá de la implementación específica del servicio).</li> <li>◆ <b>stoppable</b>: Ídem que el anterior atributo pero para comprobar si el servicio se puede detener o no.</li> <li>◆ <b>available</b>: Booleano cuyo valor será cierto si el servicio está disponible para su uso.</li> </ul>

	<ul style="list-style-type: none"> <li>◆ <b>output</b>: Cadena de texto conteniendo la salida o resultado de la operación de arranque o parada ejecutada sobre el servicio.</li> <li>◆ <b>type</b>: Tipo de servicio de infraestructura (araña web, motor de búsqueda etc.).</li> </ul>
<b>Métodos propuestos</b>	Ninguno.

<b>GenericComputingPlatform</b> <<abstract>>	
<b>Descripción</b>	Clase base para todos los servicios de computación distribuida.
<b>Responsabilidades</b>	Definir atributos comunes y métodos inherentes a los servicios de computación distribuida.
<b>Atributos propuestos</b>	<ul style="list-style-type: none"> <li>◆ <b>platform</b>: Descripción de la plataforma de computación.</li> </ul>
<b>Métodos propuestos</b>	Ninguno.

<b>GenericCrawler</b> <<abstract>>	
<b>Descripción</b>	Clase base para todos los servicios de araña web.
<b>Responsabilidades</b>	Definir atributos comunes y métodos inherentes a los servicios de araña web.
<b>Atributos propuestos</b>	<ul style="list-style-type: none"> <li>◆ <b>platform</b>: Descripción de la araña web.</li> </ul>
<b>Métodos propuestos</b>	Ninguno.

<b>GenericCrawlerStorageBackend</b> <<abstract>>	
<b>Descripción</b>	Clase base para todos los servicios de persistencia de la araña web.
<b>Responsabilidades</b>	Definir atributos comunes y métodos inherentes a los servicios de persistencia de la araña web.
<b>Atributos propuestos</b>	<ul style="list-style-type: none"> <li>◆ <b>platform</b>: Descripción de la infraestructura de persistencia.</li> </ul>
<b>Métodos propuestos</b>	Ninguno.

<b>GenericIndexer</b> <<abstract>>	
<b>Descripción</b>	Clase base para todos los servicios de motor de búsqueda.
<b>Responsabilidades</b>	Definir atributos comunes y métodos inherentes a los servicios de motor de búsqueda.
<b>Atributos propuestos</b>	◆ <b>platform</b> : Descripción del servicio de indexado.
<b>Métodos propuestos</b>	Ninguno.

<b>GenericNutchService</b> <<interface>>	
<b>Descripción</b>	Interfaz para todos los servicios de araña web basados en Apache Nutch.
<b>Responsabilidades</b>	Definir atributos las operaciones propias de los servicios proporcionados por Apache Nutch.
<b>Atributos propuestos</b>	Ninguno.
<b>Métodos propuestos</b>	<ul style="list-style-type: none"> <li>● <b>inject()</b>: Invoca las operaciones de la fase de inyección de semillas.</li> <li>● <b>generate()</b>: Invoca las operaciones de la fase de generación del trabajo Hadoop.</li> <li>● <b>fetch()</b>: Invoca las operaciones de la fase de descarga de los recursos web.</li> <li>● <b>parse()</b>: Invoca las operaciones de la fase de procesamiento de los recursos descargados.</li> <li>● <b>updateDb()</b>: Invoca las operaciones de la fase de actualización del <i>backend</i> de persistencia (Cassandra).</li> <li>● <b>Index()</b>: Invoca la fase de indexado en Apache Solr.</li> </ul>

<b>HadoopService</b>	
<b>Descripción</b>	Clase de implementación de servicio para Apache Hadoop mediante interfaz de línea de comandos.
<b>Responsabilidades</b>	Implementar las operaciones definidas por la interfaz <i>BackendServiceable</i> para esta implementación concreta de la plataforma de computación.
<b>Atributos propuestos</b>	<ul style="list-style-type: none"> <li>◆ <b>hadoopFolder</b>: Ruta al directorio de instalación local de Apache Hadoop.</li> <li>◆ <b>hadoopHost</b>: URI al nodo maestro del clúster de</li> </ul>

	<p>Apache Hadoop.</p> <ul style="list-style-type: none"> <li>◆ <b>hdfsPort</b>: Puerto de acceso al sistema de archivos distribuido HDFS:</li> <li>◆ <b>hadoopUser</b>: Usuario del sistema operativo con privilegios para ejecutar Apache Hadoop (ver manual de despliegue para mas información).</li> </ul>
<b>Métodos propuestos</b>	Los heredados de la interfaz de nivel superior.

<b>NutchService</b>	
<b>Descripción</b>	Clase de implementación de servicio para Apache Nutch mediante su API Java.
<b>Responsabilidades</b>	Implementar las operaciones definidas por la interfaz <i>BackendServiceable</i> para esta implementación concreta de la araña web.
<b>Atributos propuestos</b>	Los comunes a los servicios de infraestructura.
<b>Métodos propuestos</b>	Los heredados de la interfaz de nivel superior.

<b>NutchCommandLineService</b>	
<b>Descripción</b>	Clase de implementación de servicio para Apache Nutch mediante línea de comandos.
<b>Responsabilidades</b>	Implementar las operaciones definidas por la interfaz <i>BackendServiceable</i> para esta implementación concreta de la araña web.
<b>Atributos propuestos</b>	Los comunes a los servicios de infraestructura.
<b>Métodos propuestos</b>	Los heredados de la interfaz de nivel superior.

<b>CassandraService</b>	
<b>Descripción</b>	Clase de implementación de servicio para Apache Cassandra mediante interfaz de línea de comandos y mediante el <i>driver</i> Java.
<b>Responsabilidades</b>	Implementar las operaciones definidas por la interfaz <i>BackendServiceable</i> para esta implementación concreta de la infraestructura de persistencia de la araña web.
<b>Atributos propuestos</b>	Los comunes a los servicios de infraestructura.

<b>Métodos propuestos</b>	Los heredados de la interfaz de nivel superior.
---------------------------	---

<b>SolrService</b>	
<b>Descripción</b>	Clase de implementación de servicio para Apache Solr mediante interfaz de línea de comandos.
<b>Responsabilidades</b>	Implementar las operaciones definidas por la interfaz <i>BackendServiceable</i> para esta implementación concreta del motor de búsqueda.
<b>Atributos propuestos</b>	Los comunes a los servicios de infraestructura.
<b>Métodos propuestos</b>	Los heredados de la interfaz de nivel superior.

### 5.4.3 Subsistema de análisis

<b>Analyzable</b> <<interface>>	
<b>Descripción</b>	Interfaz de análisis de WebKumo.
<b>Responsabilidades</b>	Listar las operaciones que todo análisis debe implementar.
<b>Atributos propuestos</b>	Ninguno.
<b>Métodos propuestos</b>	<ul style="list-style-type: none"> <li>● <b>analyze()</b>: Realiza el análisis de los recursos que Apache Nutch haya descargado.</li> </ul>

<b>Reportable</b> <<interface>>	
<b>Descripción</b>	Interfaz de informe para los análisis ejecutados.
<b>Responsabilidades</b>	Determinar las operaciones que deben ser implementadas por cualquier informe de análisis.
<b>Atributos propuestos</b>	Ninguno.
<b>Métodos propuestos</b>	<ul style="list-style-type: none"> <li>● <b>calculateTotalScore()</b>: Realiza el análisis de los recursos que Apache Nutch haya descargado.</li> <li>● <b>prepareReport()</b>: Realiza operaciones adicionales tras la ejecución del análisis.</li> </ul>

<b>Fieldable</b> <<interface>>	
<b>Descripción</b>	Interfaz común para los distintos campos que componen un informe.
<b>Responsabilidades</b>	Detallar las operaciones comunes de cada campo de un informe.
<b>Atributos propuestos</b>	Ninguno.
<b>Métodos propuestos</b>	<ul style="list-style-type: none"> <li>● <b>toString()</b>: Se encarga de proporcionar la representación textual del campo de informe (nombre del recursos analizado, aspecto analizado, etc.).</li> </ul>

<b>GenericAnalysis</b> <<abstract>>	
<b>Descripción</b>	Clase base con los atributos comunes de cualquier análisis.
<b>Responsabilidades</b>	Contener los campos con los que todo análisis debe contar así como ciertos métodos comunes cuya implementación no varía.
<b>Atributos propuestos</b>	<ul style="list-style-type: none"> <li>◆ <b>reports</b>: Ruta al directorio de instalación local de Apache Hadoop.</li> <li>◆ <b>mimeTypes</b>: URI al nodo maestro del clúster de Apache Hadoop.</li> <li>◆ <b>description</b>: Descripción del análisis.</li> <li>◆ <b>jobs</b>: Referencia al trabajo que ha ejecutado esta instancia de un análisis.</li> <li>◆ <b>reportType</b>: Tipo de informe que el análisis utilizará.</li> </ul>
<b>Métodos propuestos</b>	La implementación por defecto de alguno de los métodos de la interfaz de nivel superior.

<b>GenericReport</b> <<abstract>>	
<b>Descripción</b>	Clase base con los atributos comunes de cualquier informe de análisis.
<b>Responsabilidades</b>	Contener los campos con los que todo informe debe contar así como ciertos métodos comunes cuya implementación no

	varía.
<b>Atributos propuestos</b>	<ul style="list-style-type: none"> <li>◆ <b>fields</b>: Conjunto de puntos o campos que conforman el informe.</li> <li>◆ <b>analysis</b>: Instancia del análisis asociada a la instancia del informe.</li> </ul>
<b>Métodos propuestos</b>	La implementación por defecto de alguno de los métodos de la interfaz de nivel superior.

<b>GenericReportField</b> <<abstract>>	
<b>Descripción</b>	Clase base con los atributos comunes de cualquier campo que forme parte de un informe de análisis. Estos campos puntúan el análisis.
<b>Responsabilidades</b>	Contener los campos con los que todo campo de informe debe contar así como ciertos métodos comunes cuya implementación no varía.
<b>Atributos propuestos</b>	<ul style="list-style-type: none"> <li>◆ <b>name</b>: Nombre del campo del informe.</li> <li>◆ <b>score</b>: Puntuación sobre 100 puntos que el campo aporta al total del informe.</li> <li>◆ <b>weight</b>: Peso del campo, esto es, su relevancia en el conjunto del informe.</li> <li>◆ <b>Value</b>: Valor textual del campo.</li> </ul>
<b>Métodos propuestos</b>	La implementación por defecto de alguno de los métodos de la interfaz de nivel superior.

<b>GenericTextField</b>	
<b>Descripción</b>	Campo de informe genérico que contiene información textual.
<b>Responsabilidades</b>	Informar sobre un aspecto concreto de un análisis (presencia o no de una característica en un recurso, valor de algún aspecto del recurso etc.).
<b>Atributos propuestos</b>	Ninguno.
<b>Métodos propuestos</b>	Los heredados de interfaces de nivel superior.

Las entidades `MockReport` y `MockAnalysis` que aparecen en el diagrama preliminar de clases son ejemplos de análisis e informe de análisis implementados.

<b>Resource</b> ««abstract»»	
<b>Descripción</b>	Clase base para todos los recursos web.
<b>Responsabilidades</b>	Contener atributos y operaciones comunes que todos los recursos web comparten.
<b>Atributos propuestos</b>	<ul style="list-style-type: none"> <li>◆ <b>sourceCode</b>: Cadena de texto que contiene el código fuente del recurso.</li> <li>◆ <b>uri</b>: Localizador unívoco universal del recurso.</li> <li>◆ <b>title</b>: Nombre del recurso.</li> <li>◆ <b>mimeType</b>: Tipo MIME del recurso.</li> </ul>
<b>Métodos propuestos</b>	Ninguno.

<b>HTMLResource</b>	
<b>Descripción</b>	Entidad que representa un recurso de hipertexto.
<b>Responsabilidades</b>	Representar el modelo de recurso HTML con los atributos y operaciones característicos.
<b>Atributos propuestos</b>	<ul style="list-style-type: none"> <li>◆ <b>outboundLinks</b>: Enlaces, elementos anchor que el recurso contiene.</li> </ul>
<b>Métodos propuestos</b>	<ul style="list-style-type: none"> <li>● <b>addOutboundLinks()</b>: Añade un nuevo enlace que se haya encontrado en el documento.</li> </ul>

<b>CSSResource</b>	
<b>Descripción</b>	Entidad que representa un recurso de hoja de estilo.
<b>Responsabilidades</b>	Representar el modelo de recurso CSS con los atributos y operaciones característicos.
<b>Atributos propuestos</b>	Los heredados de la clase abstracta <i>Resource</i> .
<b>Métodos propuestos</b>	Los heredados de la clase abstracta <i>Resource</i> .

<b>ScriptResource</b>	
<b>Descripción</b>	Entidad que representa un recurso JavaScript.
<b>Responsabilidades</b>	Representar el modelo de recurso JavaScript con los atributos y operaciones característicos.
<b>Atributos propuestos</b>	Los heredados de la clase abstracta <i>Resource</i> .
<b>Métodos propuestos</b>	Los heredados de la clase abstracta <i>Resource</i> .

#### 5.4.4 Subsistema de trabajos

<b>CrawlerJob</b> <<abstract>>	
<b>Descripción</b>	Entidad que representa un trabajo genérico que se ejecute en WebKumo.
<b>Responsabilidades</b>	Contener atributos y funcionalidad común a todos los trabajos que se ejecuten.
<ul style="list-style-type: none"> <li>◆ <b>Atributos propuestos</b></li> </ul>	<ul style="list-style-type: none"> <li>◆ <b>creationDate</b>: Fecha de creación del trabajo.</li> <li>◆ <b>executionDate</b>: Fecha en la que se ejecutó el trabajo.</li> <li>◆ <b>owner</b>: Usuario responsable del trabajo.</li> <li>◆ <b>seeds</b>: Listado de URIs (semillas) de los portales a analizar.</li> <li>◆ <b>analyses</b>: Listado de análisis a realizar sobre los portales especificados.</li> <li>◆ <b>customParameters</b>: Conjunto de parámetros personalizados que los análisis podrán utilizar para su computación.</li> <li>◆ <b>status</b>: Estado del trabajo (detenido, finalizado etc.).</li> </ul>
<b>Métodos propuestos</b>	Ninguno

NutchJob	
<b>Descripción</b>	Clase que representa un trabajo de análisis ejecutado con la araña web Apache Nutch.
<b>Responsabilidades</b>	Albergar la configuración con la que Apache Nutch debe ejecutar la descarga y el posterior análisis.
<b>Atributos propuestos</b>	<ul style="list-style-type: none"> <li>◆ <b>batchIds</b>: Conjunto de identificadores de trabajo por lotes de Apache Hadoop que se utilizaron durante la ejecución del trabajo.</li> <li>◆ <b>rounds</b>: Número de ciclos que Apache Nutch debe ejecutar.</li> <li>◆ <b>topN</b>: Número de recursos por iteración que Apache Nutch debe descargar y procesar.</li> </ul>
<b>Métodos propuestos</b>	Ninguno

### 5.4.5 Subsistema WebKumo Connector

WebKumoConnector	
<b>Descripción</b>	Clase extensión o <i>plugin</i> de Apache Nutch desarrollado para WebKumo.
<b>Responsabilidades</b>	Procesar los recursos web que Nutch descarga realizando un procesamiento básico de los recursos HTML, CSS y JavaScript y ejecutando los algoritmos de análisis que se hayan inyectado desde WebKumo. También escribe el resultado de los análisis en el sistema de archivos distribuido HDFS mediante JSON serializado.
<b>Atributos propuestos</b>	Ninguno.
<b>Métodos propuestos</b>	<ul style="list-style-type: none"> <li>● <b>getParse()</b>: Método de implementación de la interfaz <code>org.apache.nutch.parse.Parser</code> que es invocado por Apache Nutch durante la fase <i>parse</i> de ejecución.</li> </ul>

En el diagrama de clases preliminar se incluye la interfaz `org.apache.nutch.parse.Parser` para clarificar que el *plugin* WebKumo Connector implementa dicha interfaz.

## 5.5 Análisis de caso de uso y escenarios

A continuación se ejemplificarán los casos de uso mediante escenarios en los que los usuarios se puedan encontrar al realizar las tareas principales de la aplicación. Se descartan escenarios que no aportan nada al análisis y que son infrecuentes (pérdida de conectividad de red, errores internos del servidor etc.).

Para cada de uso identificado en los apartados anteriores se describirá el escenario mediante una tabla explicativa con los supuestos y posibles variaciones de escenarios. Además se proveerá para cada escenario un diagrama de robustez [ROBU].

### 5.5.1 Crear usuario

Crear usuario	
<b>Precondiciones</b>	El usuario debe haber iniciado sesión y ser usuario con rol administrador.
<b>Postcondiciones</b>	Un nuevo usuario estará dado de alta en el sistema (aparecerá en la base de datos).
<b>Actores</b>	Usuario administrador
<b>Descripción</b>	<ol style="list-style-type: none"> <li>1. El usuario administrador se dirigirá al formulario de creación de usuario,</li> <li>2. El usuario rellenará con los datos apropiados (nombre de usuario, rol, grupos a los que pertenece etc.).</li> <li>3. El usuario enviará el formulario tras lo se procederá al alta.</li> <li>4. Retornará al panel del administrador donde podrá ver en el listado el nuevo usuario creado.</li> </ol>
<b>Escenarios secundarios</b>	<ol style="list-style-type: none"> <li>1. El nombre de usuario (<i>login</i>) ya está registrado. <ul style="list-style-type: none"> <li>◆ Se mostrará un mensaje de error y el usuario será dirigido al paso 2.</li> </ul> </li> <li>2. El formulario contiene valores inválidos <ul style="list-style-type: none"> <li>◆ Se mostrará un mensaje de error con el/los campos erróneos y el usuario será dirigido al paso 2. Los valores introducidos se mantendrán.</li> </ul> </li> </ol>
<b>Excepciones</b>	<ol style="list-style-type: none"> <li>1. El usuario no tiene activada la ejecución JavaScript en</li> </ol>

	<p>su navegador.</p> <ul style="list-style-type: none"> <li>◆ Se realizará la validación únicamente por parte del servidor perdiendo los datos introducidos.</li> </ul> <p>2. La base de datos no está disponible</p> <ul style="list-style-type: none"> <li>◆ Se mostrará la página general de error de WebKumo.</li> </ul> <p>3. El usuario no es administrador.</p> <ul style="list-style-type: none"> <li>◆ No debería poder acceder a la página de alta de usuario pero si esta excepción ocurriese, al intentar darlo de alta se le redirigiría la una página donde se informa de un error HTTP 404.</li> </ul>
--	---

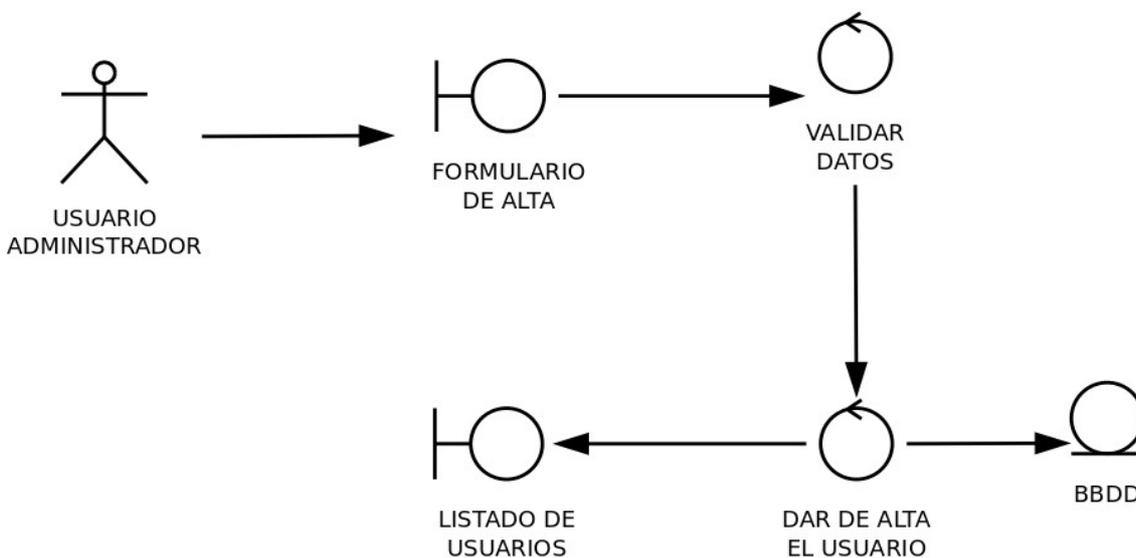


Fig. 11: Diagrama de robustez para el alta de un usuario.

### 5.5.2 Modificar usuario

Modificar usuario	
<b>Precondiciones</b>	Usuario con sesión iniciada y usuarios dados de alta.
<b>Postcondiciones</b>	Un usuario preexistente verá modificados sus datos (en la base de datos).
<b>Actores</b>	Usuario administrador o usuario regular
<b>Descripción</b>	1. El usuario administrador se dirigirá al formulario de

	<p>modificación de usuario,</p> <ol style="list-style-type: none"> <li>2. El usuario modificará los datos pertinentes (nombre de usuario, rol, grupos a los que pertenece etc.).</li> <li>3. El usuario enviará el formulario tras lo se procederá a la modificación.</li> <li>4. Retornará al panel del usuario donde podrá ver los cambios guardados.</li> </ol>
<b>Escenarios secundarios</b>	<ol style="list-style-type: none"> <li>1. El formulario contiene valores inválidos <ul style="list-style-type: none"> <li>◆ Se mostrará un mensaje de error con el/los campos erróneos y el usuario será dirigido al paso 2. Los valores introducidos se mantendrán.</li> </ul> </li> </ol>
<b>Excepciones</b>	<ol style="list-style-type: none"> <li>1. El usuario no tiene activada la ejecución JavaScript en su navegador. <ul style="list-style-type: none"> <li>◆ Se realizará la validación únicamente por parte del servidor perdiendo los datos introducidos.</li> </ul> </li> <li>2. La base de datos no está disponible <ul style="list-style-type: none"> <li>◆ Se mostrará la página general de error de WebKumo.</li> </ul> </li> <li>3. El usuario no es administrador e intenta editar a otro usuario distinto de él mismo. <ul style="list-style-type: none"> <li>◆ No debería poder acceder a la página de modificación de otro usuario pero si esta excepción ocurriese, al intentar la modificación se le redirigiría a una página donde se informa de un error HTTP 404.</li> </ul> </li> </ol>

El diagrama de robustez es prácticamente idéntico al del caso de uso anterior "Crear un usuario" pero con la salvedad de que los datos del usuario que se intenta modificar ya aparecerán en el formulario.

## 5.5.3 Eliminar usuario

<b>Eliminar usuario</b>	
<b>Precondiciones</b>	El usuario debe haber iniciado sesión y debe tener el rol de administrador. Debe haber usuarios dados de alta.
<b>Postcondiciones</b>	Un usuario preexistente será dado de baja (en la base de datos).
<b>Actores</b>	Usuario administrador.
<b>Descripción</b>	<ol style="list-style-type: none"> <li>1. El usuario administrador se dirigirá al panel del administrador donde visualizará el listado de usuario.</li> <li>2. El usuario administrador seleccionará el usuario a borrar.</li> <li>3. El usuario administrador verá la información del usuario con una opción para eliminarlo que accionará.</li> <li>4. El sistema dará de baja al usuario seleccionado dirigiendo de nuevo al usuario administrador a la vista del listado de usuarios donde comprobará que ya no existe.</li> </ol>
<b>Escenarios secundarios</b>	<ol style="list-style-type: none"> <li>1. El usuario tiene trabajos de los que es dueño. <ul style="list-style-type: none"> <li>◆ Se mostrará un mensaje de error y se evitará llevar a cabo la operación</li> </ul> </li> </ol>
<b>Excepciones</b>	<ol style="list-style-type: none"> <li>1. La base de datos no está disponible <ul style="list-style-type: none"> <li>◆ Se mostrará la página general de error de WebKumo.</li> </ul> </li> <li>2. El usuario no es administrador. <ul style="list-style-type: none"> <li>◆ No debería poder acceder a la página de modificación de otro usuario pero si esta excepción ocurriese, al intentar la modificación se le redirigirá a una página donde se informa de un error HTTP 404.</li> </ul> </li> </ol>

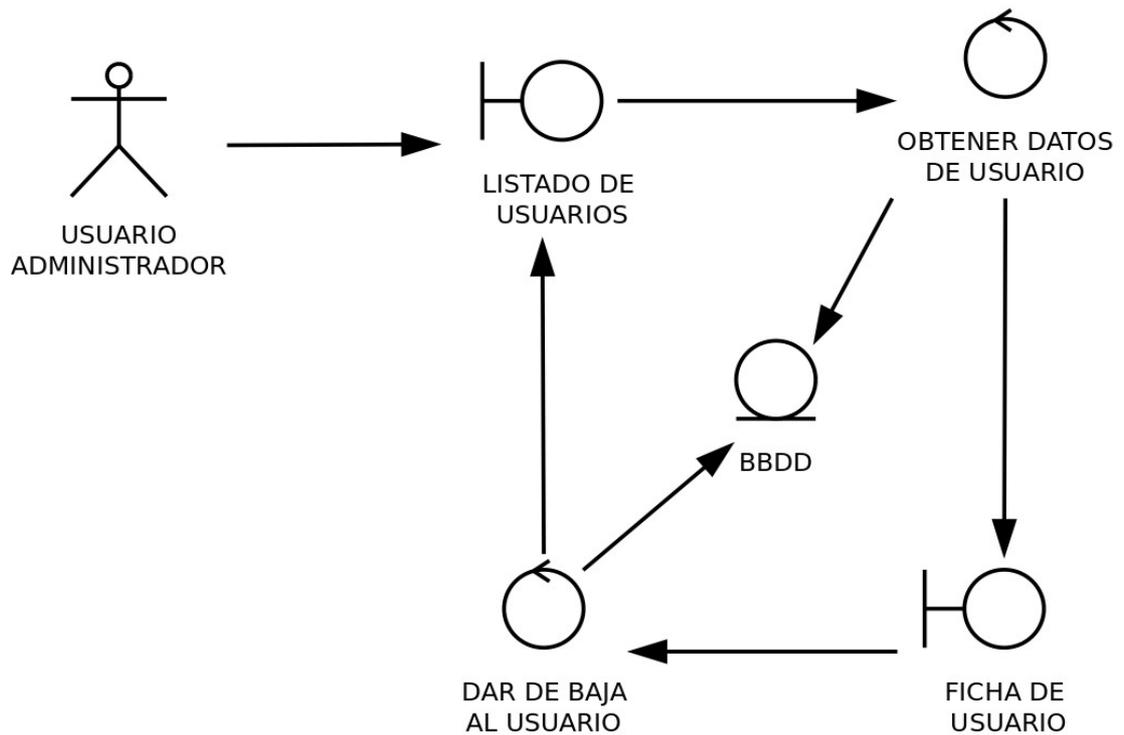


Fig. 12: Diagrama de robustez del caso de uso de baja de usuario.

### 5.5.4 Listar usuarios

Listar usuarios	
<b>Precondiciones</b>	El usuario debe haber iniciado sesión y debe tener el rol de administrador.
<b>Postcondiciones</b>	Ninguna que modifique el estado del sistema.
<b>Actores</b>	Usuario administrador.
<b>Descripción</b>	<ol style="list-style-type: none"> <li>1. El usuario administrador se dirigirá al panel del administrador donde visualizará el listado de usuarios.</li> </ol>
<b>Excepciones</b>	<ol style="list-style-type: none"> <li>1. La base de datos no está disponible                             <ul style="list-style-type: none"> <li>◆ Se mostrará la página general de error interno.</li> </ul> </li> <li>2. El usuario no es administrador.                             <ul style="list-style-type: none"> <li>◆ Se le redirigirá a una página donde se informa de un error HTTP 404.</li> </ul> </li> </ol>

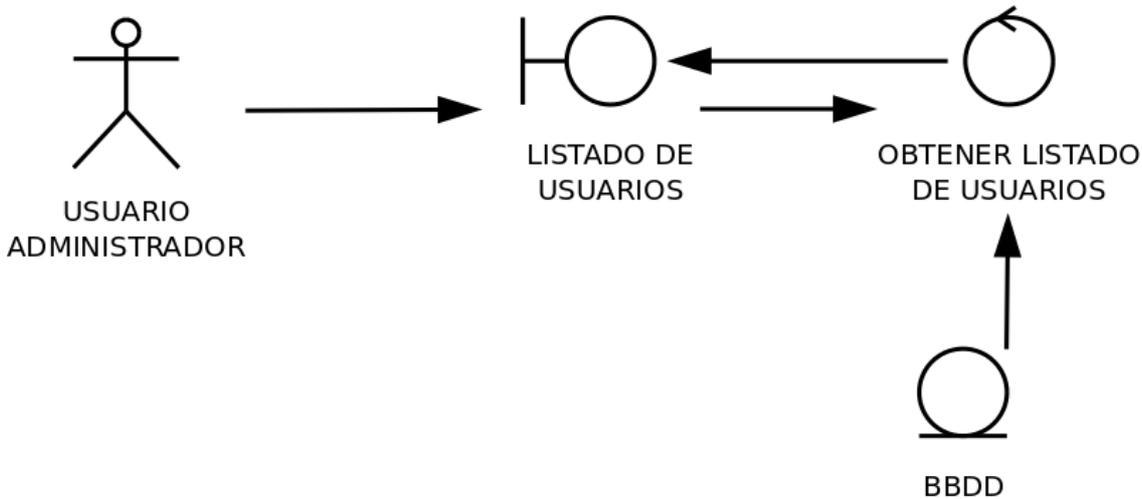


Fig. 13: Diagrama de robustez del caso de uso de listado de usuarios.

### 5.5.5 Crear grupo

Crear grupo	
<b>Precondiciones</b>	El usuario debe haber iniciado sesión y debe tener el rol de administrador.
<b>Postcondiciones</b>	El sistema contendrá un grupo de usuarios dado de alta mas.
<b>Actores</b>	Usuario administrador.
<b>Descripción</b>	<ol style="list-style-type: none"> <li>1. El usuario administrador se dirigirá al formulario de creación de grupo.</li> <li>2. El usuario rellenará con los datos apropiados (nombre, descripción etc.).</li> <li>3. El usuario enviará el formulario tras lo se procederá al alta.</li> <li>4. Retornará al panel del administrador donde podrá ver en el listado el nuevo grupo creado.</li> </ol>
<b>Escenarios secundarios</b>	<ol style="list-style-type: none"> <li>1. Ya existe un grupo con el mismo nombre que el grupo que se quiere dar de alta.               <ul style="list-style-type: none"> <li>◆ Se mostrará un mensaje de error y se evitará llevar a cabo la operación.</li> </ul> </li> </ol>
<b>Excepciones</b>	<ol style="list-style-type: none"> <li>1. La base de datos no está disponible</li> </ol>

	<ul style="list-style-type: none"> <li>◆ Se mostrará la página general de error de WebKumo.</li> </ul> <ol style="list-style-type: none"> <li>2. El usuario no tiene activada la ejecución JavaScript en su navegador. <ul style="list-style-type: none"> <li>◆ Se realizará la validación únicamente por parte del servidor perdiendo los datos introducidos.</li> </ul> </li> <li>3. El usuario no es administrador. <ul style="list-style-type: none"> <li>◆ No debería poder acceder al formulario de alta de grupo. Si ocurriese se le redirigiría a una página donde se informa de un error HTTP 404.</li> </ul> </li> </ol>
--	--

El diagrama de robustez es el mismo que para el caso de uso de "Crear usuario" con la salvedad que el formulario contiene campos distintos.

### 5.5.6 Eliminar grupo

Eliminar grupo	
<b>Precondiciones</b>	El usuario debe haber iniciado sesión y debe tener el rol de administrador. Además al menos un grupo de usuarios debe estar dado de alta.
<b>Postcondiciones</b>	El sistema contendrá un grupo de usuarios dado de alta menos.
<b>Actores</b>	Usuario administrador.
<b>Descripción</b>	<ol style="list-style-type: none"> <li>1. El usuario administrador se dirigirá al panel del administrador donde visualizará el listado de grupos.</li> <li>2. El usuario administrador seleccionará el grupo a borrar.</li> <li>3. El usuario administrador verá la información del grupo con una opción para eliminarlo que accionará.</li> <li>4. El sistema dará de baja al grupo seleccionado dirigiendo de nuevo al usuario administrador a la vista del listado de grupos donde no aparecerá ya.</li> </ol>
<b>Escenarios secundarios</b>	<ol style="list-style-type: none"> <li>1. El grupo tiene miembros. <ul style="list-style-type: none"> <li>◆ Se mostrará un mensaje de error y se evitará llevar a cabo la operación</li> </ul> </li> </ol>
<b>Excepciones</b>	<ol style="list-style-type: none"> <li>3. La base de datos no está disponible</li> </ol>

	<ul style="list-style-type: none"> <li>◆ Se mostrará la página general de error de WebKumo.</li> </ul> <p>4. El usuario no es administrador.</p> <ul style="list-style-type: none"> <li>◆ No debería poder acceder a la página de eliminación de grupos pero si esta excepción ocurriese, al intentar la baja se le redirigiría a una página donde se informa de un error HTTP 404.</li> </ul>
--	--

El diagrama de robustez es el mismo que para el caso de uso de "Eliminar usuario" con la salvedad que lo que se elimina es un grupo y no un usuario.

### 5.5.7 Modificar grupo

<b>Modificar grupo</b>	
<b>Precondiciones</b>	El usuario debe haber iniciado sesión y debe existir por lo menos un grupo dado de alta.
<b>Postcondiciones</b>	Un grupo preexistente verá modificados sus datos (en la base de datos).
<b>Actores</b>	Usuario administrador
<b>Descripción</b>	<ol style="list-style-type: none"> <li>1. El usuario administrador se dirigirá al formulario de modificación de grupo,</li> <li>2. El usuario modificará los datos pertinentes (nombre, descripción etc.).</li> <li>3. El usuario enviará el formulario tras lo se procederá a la modificación.</li> <li>4. Retornará al panel del usuario donde podrá ver los cambios guardados.</li> </ol>
<b>Escenarios secundarios</b>	<ol style="list-style-type: none"> <li>1. El formulario contiene valores inválidos <ul style="list-style-type: none"> <li>◆ Se mostrará un mensaje de error con el/los campos erróneos y el usuario será dirigido al paso 2.</li> </ul> </li> <li>2. Los valores introducidos se mantendrán.</li> </ol>
<b>Excepciones</b>	<ol style="list-style-type: none"> <li>1. El usuario no tiene activada la ejecución JavaScript. <ul style="list-style-type: none"> <li>◆ Se realizará la validación únicamente por parte del servidor perdiendo los datos introducidos.</li> </ul> </li> <li>2. La base de datos no está disponible <ul style="list-style-type: none"> <li>◆ Se mostrará la página general de error de</li> </ul> </li> </ol>

	<p>WebKumo.</p> <ol style="list-style-type: none"> <li>3. El usuario no es administrador. <ul style="list-style-type: none"> <li>◆ No debería poder acceder a la página de modificación de un grupo pero si esta excepción ocurriese, al intentar la modificación se le redirigiría a una página donde se informa de un error HTTP 404.</li> </ul> </li> </ol>
--	--

El diagrama de robustez es prácticamente idéntico al del caso de uso para “Crear un usuario” pero con la salvedad de que los datos del grupo que se intenta modificar ya aparecerán en el formulario.

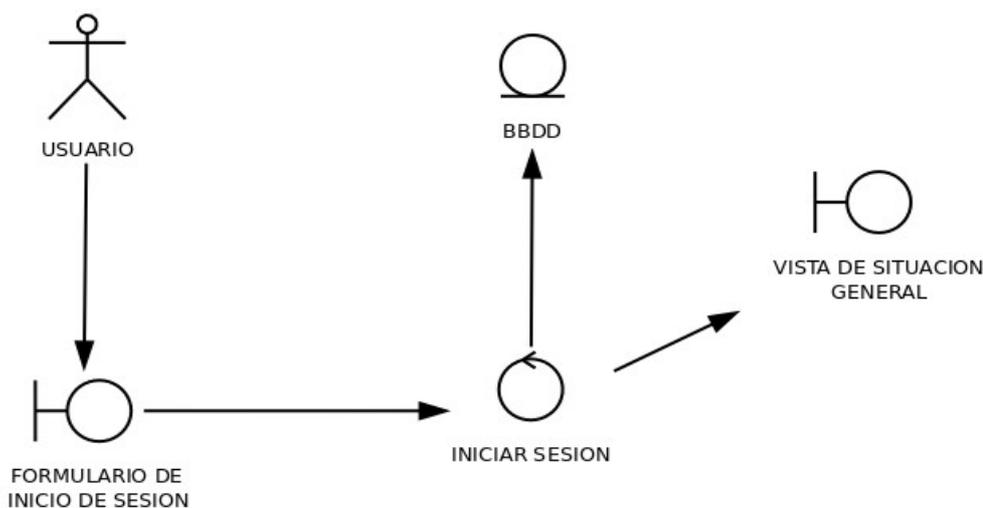
### 5.5.8 Listar grupos

<b>Listar grupos</b>	
<b>Precondiciones</b>	El usuario debe haber iniciado sesión.
<b>Postcondiciones</b>	Ninguna que afecte al estado de la aplicación.
<b>Actores</b>	Usuario administrador
<b>Descripción</b>	<ol style="list-style-type: none"> <li>1. El usuario administrador se dirigirá al panel del usuario donde, además de ver listados los usuarios, podrá ver listados los grupos.</li> </ol>
<b>Excepciones</b>	<ol style="list-style-type: none"> <li>1. La base de datos no está disponible <ul style="list-style-type: none"> <li>◆ Se mostrará la página general de error de WebKumo.</li> </ul> </li> <li>2. El usuario no es administrador. <ul style="list-style-type: none"> <li>◆ No debería poder acceder al panel de usuarios para administradores pero si esta excepción ocurriese, al intentar la modificación se le redirigiría a una página donde se informa de un error HTTP 404.</li> </ul> </li> </ol>

El diagrama de robustez es prácticamente idéntico al del caso de uso para “Listar usuarios” pero con la salvedad de que los datos que se muestran en forma tabular o de lista son los de los grupos.

### 5.5.9 Listar grupos

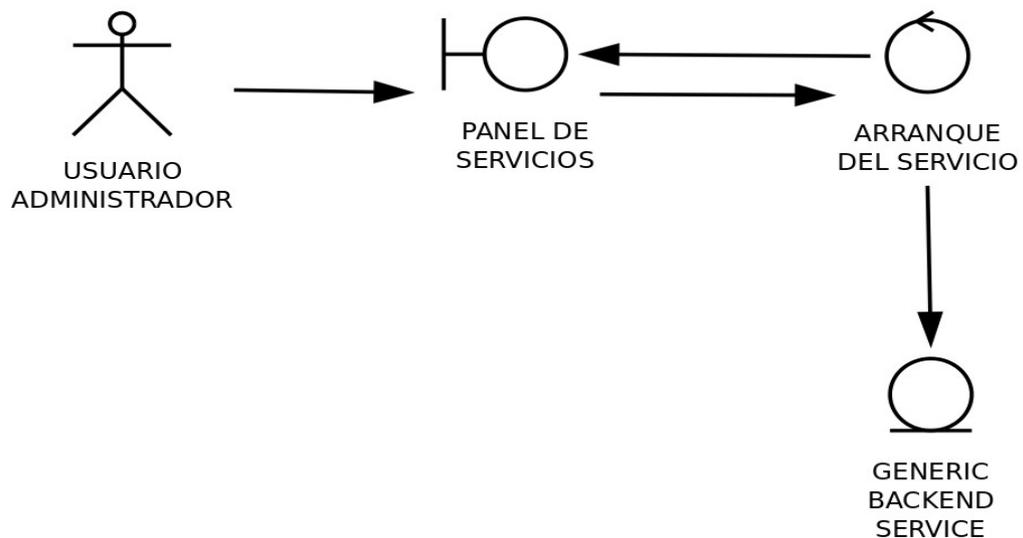
Iniciar sesión	
<b>Precondiciones</b>	El usuario no debe tener sesión activa en la aplicación.
<b>Postcondiciones</b>	El usuario que inicie sesión tendrá una sesión de aplicación en activo.
<b>Actores</b>	Usuario administrador o regular.
<b>Descripción</b>	<ol style="list-style-type: none"> <li>1. El usuario se dirigirá al punto de acceso de la aplicación donde se le mostrará el formulario de inicio de sesión.</li> <li>2. El usuario deberá entonces rellenar los campos del formulario con sus credenciales.</li> <li>3. La aplicación validará la información y dirigirá al usuario al panel de vista principal.</li> </ol>
<b>Escenarios secundarios</b>	<ol style="list-style-type: none"> <li>1. El usuario ya tiene sesión iniciada.                     <ul style="list-style-type: none"> <li>◆ Se dirigirá automáticamente al usuario al panel de vista principal.</li> </ul> </li> <li>2. El usuario introduce una combinación de usuario y contraseña incorrectas.                     <ul style="list-style-type: none"> <li>◆ Se notificará al usuario mediante un mensaje de error y se le devolverá al formulario de inicio de sesión.</li> </ul> </li> </ol>
<b>Excepciones</b>	<ol style="list-style-type: none"> <li>1. La base de datos no está disponible                     <ul style="list-style-type: none"> <li>◆ Se mostrará la página general de error de WebKumo.</li> </ul> </li> </ol>



**Fig. 14: Diagrama de robustez del caso de uso de inicio de sesión.**

### 5.5.8 Arrancar servicio

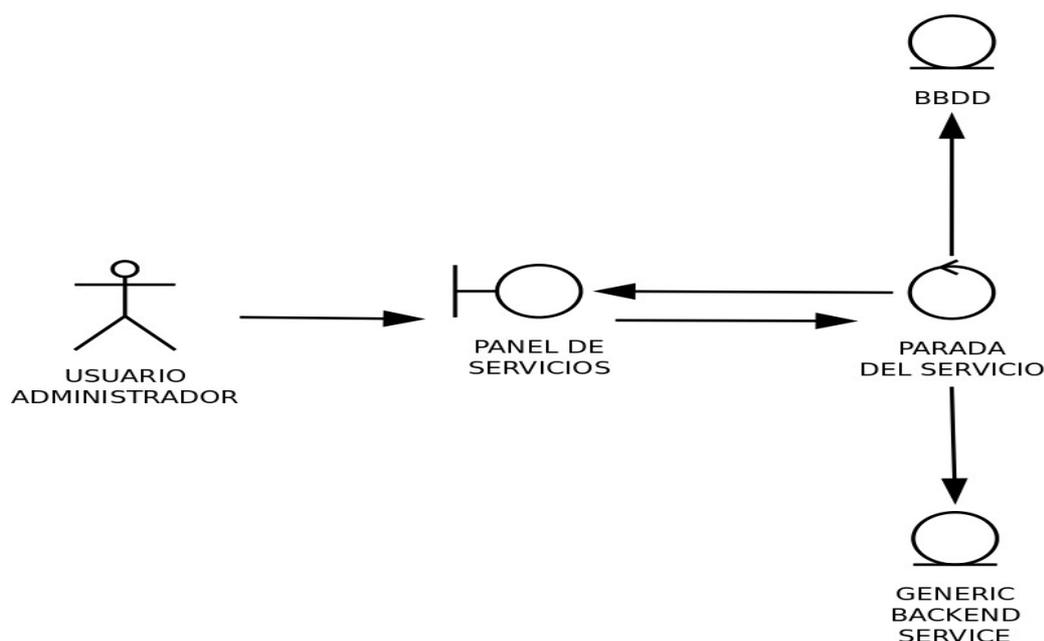
<b>Arrancar servicio</b>	
<b>Precondiciones</b>	El usuario debe haber iniciado sesión.
<b>Postcondiciones</b>	Un servicio de infraestructura que presta funcionalidad a WebKumo se iniciará.
<b>Actores</b>	Usuario administrador
<b>Descripción</b>	<ol style="list-style-type: none"> <li>1. El usuario administrador se dirigirá al panel de servicios.</li> <li>2. El usuario utilizará el botón pertinente para arrancar el servicio que desee.</li> <li>3. El servicio arrancará mostrando un mensaje al usuario con el resultado de dicha operación.</li> </ol>
<b>Escenarios secundarios</b>	<ol style="list-style-type: none"> <li>1. El servicio no pudo arrancar.                             <ul style="list-style-type: none"> <li>◆ Se indicará al usuario que no se ha podido iniciar el servicio con un mensaje. Además se incluirá información adicional para indicar el motivo de la imposibilidad de arrancar el servicio.</li> </ul> </li> </ol>
<b>Excepciones</b>	<ol style="list-style-type: none"> <li>1. El usuario no es administrador.                             <ul style="list-style-type: none"> <li>◆ No debería poder acceder a la vista de panel de servicios pero si esta excepción ocurriese, al intentar el acceso se le mostrará un error 404.</li> </ul> </li> </ol>



**Fig. 15: Diagrama de robustez del caso de uso de arranque de servicio.**

### 5.5.9 Detener servicio

<b>Detener servicio</b>	
<b>Precondiciones</b>	El usuario debe haber iniciado sesión y algún servicio debe estar iniciado.
<b>Postcondiciones</b>	Un servicio de infraestructura que presta funcionalidad a WebKumo se detendrá.
<b>Actores</b>	Usuario administrador
<b>Descripción</b>	<ol style="list-style-type: none"> <li>1. El usuario administrador se dirigirá al panel de servicios.</li> <li>2. El usuario utilizará el botón pertinente para detener el servicio que desee.</li> <li>3. El servicio se detendrá mostrando un mensaje al usuario con el resultado de dicha operación.</li> </ol>
<b>Escenarios secundarios</b>	<ol style="list-style-type: none"> <li>1. El servicio no pudo detenerse o bien existen trabajos en ejecución. <ul style="list-style-type: none"> <li>◆ Se indicará al usuario que no se ha podido detener el servicio con un mensaje. Además se incluirá información adicional para indicar el motivo de la imposibilidad de detener el servicio.</li> </ul> </li> </ol>
<b>Excepciones</b>	<ol style="list-style-type: none"> <li>1. El usuario no es administrador. <ul style="list-style-type: none"> <li>◆ No debería poder acceder a la vista de panel de servicios pero si esta excepción ocurriese, al intentar la modificación se le redirigiría la una página donde se informa de un error HTTP 404.</li> </ul> </li> </ol>



*Fig. 16: Diagrama de robustez del caso de uso de parada de servicio.*

## 5.5.10 Listar servicios

Listar servicios	
<b>Precondiciones</b>	El usuario debe haber iniciado sesión.
<b>Postcondiciones</b>	Ninguna que afecte al estado del sistema.
<b>Actores</b>	Usuario administrador
<b>Descripción</b>	<ol style="list-style-type: none"> <li>1. El usuario administrador se dirigirá al panel de servicios.</li> <li>2. El usuario podrá ver en el panel los servicios de infraestructura asociados y las opciones para gestionarlos.</li> </ol>
<b>Excepciones</b>	<ol style="list-style-type: none"> <li>1. El usuario no es administrador.                             <ul style="list-style-type: none"> <li>◆ No debería poder acceder a la vista de panel de servicios pero si esta excepción ocurriese, al intentar la modificación se le redirigiría a una página donde se informa de un error HTTP 404.</li> </ul> </li> </ol>

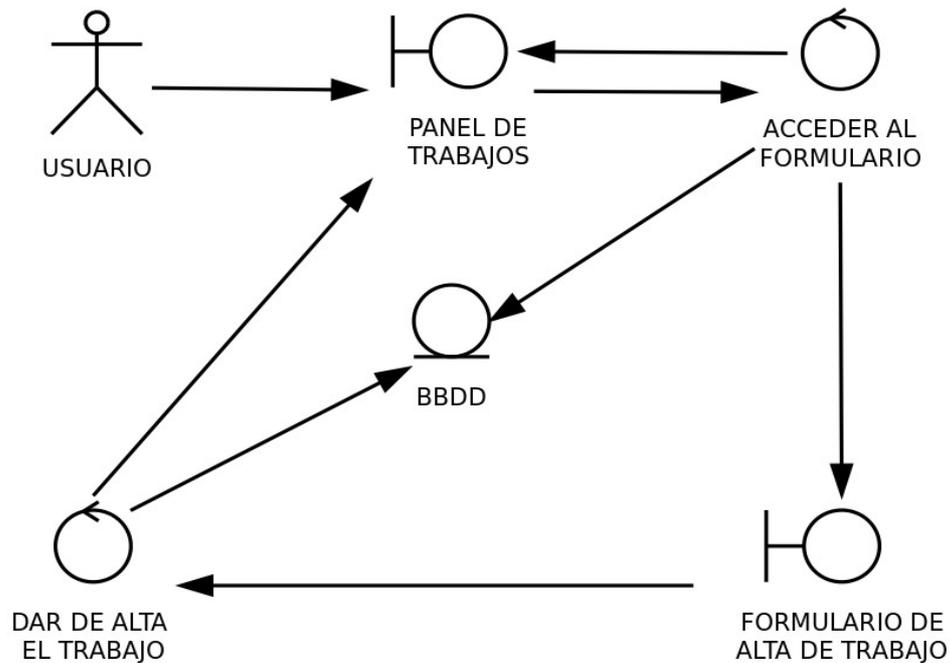


Fig. 17: Diagrama de robustez para el caso de uso de listado de servicios.

### 5.5.11 Crear trabajo

Crear trabajo	
<b>Precondiciones</b>	El usuario debe haber iniciado sesión.
<b>Postcondiciones</b>	El sistema tendrá dado de alta un nuevo trabajo (en la base de datos)
<b>Actores</b>	Usuario administrador o regular.
<b>Descripción</b>	<ol style="list-style-type: none"> <li>1. El usuario se dirigirá al panel de trabajos donde seleccionará la opción para crear un trabajo nuevo.</li> <li>2. El usuario rellenará el formulario de creación de un nuevo trabajo con los datos pertinentes (los análisis a realizar, las semillas, el nombre etc.).</li> <li>3. El usuario deberá enviar el formulario y entonces WebKumo creará el trabajo.</li> <li>4. El usuario será dirigido al panel de trabajos donde podrá visualizar el nuevo trabajo creado en el listado.</li> </ol>
<b>Escenarios secundarios</b>	<ol style="list-style-type: none"> <li>1. El formulario contiene valores inválidos               <ul style="list-style-type: none"> <li>◆ Se mostrará un mensaje de error con el/los campos erróneos y el usuario será dirigido al paso 2. Los valores introducidos se mantendrán.</li> </ul> </li> <li>2. El usuario no tiene permisos de escritura.               <ul style="list-style-type: none"> <li>◆ Se le indicará mediante mensaje de error que no puede crear un trabajo.</li> </ul> </li> </ol>
<b>Excepciones</b>	<ol style="list-style-type: none"> <li>1. La base de datos no está disponible</li> </ol>

	<ul style="list-style-type: none"> <li>◆ Se mostrará la página general de error interno.</li> </ul> <p>2. El usuario no tiene activada la ejecución JavaScript en su navegador.</p> <ul style="list-style-type: none"> <li>◆ Se realizará la validación únicamente por parte del servidor perdiendo los datos introducidos.</li> </ul>
--	--



**Fig. 18:** Diagrama de robustez del caso de uso para crear un trabajo nuevo.

### 5.5.12 Modificar trabajo

Modificar trabajo	
<b>Precondiciones</b>	El usuario debe haber iniciado sesión y al menos debe existir un trabajo registrado.
<b>Postcondiciones</b>	El sistema tendrá dado de alta un nuevo trabajo (en la base de datos)
<b>Actores</b>	Usuario administrador o regular.
<b>Descripción</b>	1. El usuario se dirigirá al panel de trabajos donde seleccionará la opción para crear un trabajo nuevo.

	<ol style="list-style-type: none"> <li>2. El usuario rellenará el formulario de creación de un nuevo trabajo con los datos pertinentes (los análisis a realizar, las semillas, el nombre etc.).</li> <li>3. El usuario deberá enviar el formulario y entonces WebKumo creará el trabajo.</li> <li>4. El usuario será dirigido al panel de trabajos donde podrá visualizar el nuevo trabajo creado en el listado.</li> </ol>
<b>Escenarios secundarios</b>	<ol style="list-style-type: none"> <li>1. El formulario contiene valores inválidos <ul style="list-style-type: none"> <li>◆ Se mostrará un mensaje de error con el/los campos erróneos y el usuario será dirigido al paso 2. Los valores introducidos se mantendrán.</li> </ul> </li> <li>2. El usuario no tiene permisos de escritura. <ul style="list-style-type: none"> <li>◆ Se le indicará mediante mensaje de error que no puede crear un trabajo.</li> </ul> </li> </ol>
<b>Excepciones</b>	<ol style="list-style-type: none"> <li>1. La base de datos no está disponible <ul style="list-style-type: none"> <li>◆ Se mostrará la página general de error de WebKumo.</li> </ul> </li> <li>2. El usuario no tiene activada la ejecución JavaScript en su navegador. <ul style="list-style-type: none"> <li>◆ Se realizará la validación únicamente por parte del servidor perdiendo los datos introducidos.</li> </ul> </li> </ol>

El diagrama de robustez es análogo al del caso anterior de “Crear trabajo” con la salvedad de que los campos del formulario ya aparecerán rellenos. Además en lugar de en el panel de trabajos solicitar la creación de uno nuevo, se debe seleccionar uno del listado.

### 5.5.13 Eliminar trabajo

<b>Eliminar trabajo</b>	
<b>Precondiciones</b>	El usuario debe haber iniciado sesión y al menos debe existir un trabajo registrado.
<b>Postcondiciones</b>	El sistema tendrá un trabajo registrado menos (en la base de datos)
<b>Actores</b>	Usuario administrador o regular.
<b>Descripción</b>	<ol style="list-style-type: none"> <li>1. El usuario se dirigirá al panel de trabajos donde</li> </ol>

	<p>seleccionará el trabajo que quiera eliminar del listado.</p> <ol style="list-style-type: none"><li>2. El usuario seleccionará en la pantalla del trabajo la opción para eliminarlo.</li><li>3. El sistema dará de baja el trabajo y dirigirá al usuario al panel de trabajos donde comprobará que ya no existe.</li></ol>
<b>Escenarios secundarios</b>	<ol style="list-style-type: none"><li>1. El trabajo ha sido ejecutado.<ul style="list-style-type: none"><li>◆ Se impedirá el borrado del trabajo indicándoselo al usuario mediante un mensaje.</li></ul></li><li>2. El usuario no posee permisos de escritura.<ul style="list-style-type: none"><li>◆ El usuario recibirá notificación de la imposibilidad de la eliminación y el sistema no se verá alterado.</li></ul></li></ol>
<b>Excepciones</b>	<ol style="list-style-type: none"><li>1. La base de datos no está disponible<ul style="list-style-type: none"><li>◆ Se mostrará la página general de error de WebKumo.</li></ul></li><li>2. El usuario intenta eliminar un trabajo que no es suyo.<ul style="list-style-type: none"><li>◆ No se le permitirá el acceso al trabajo, pero por seguridad se comprobará al intentar eliminarlo que el dueño del trabajo sea quien intenta darlo de baja. En caso contrario se dirigirá al usuario a una página de recurso no encontrado (HTTP 404).</li></ul></li></ol>

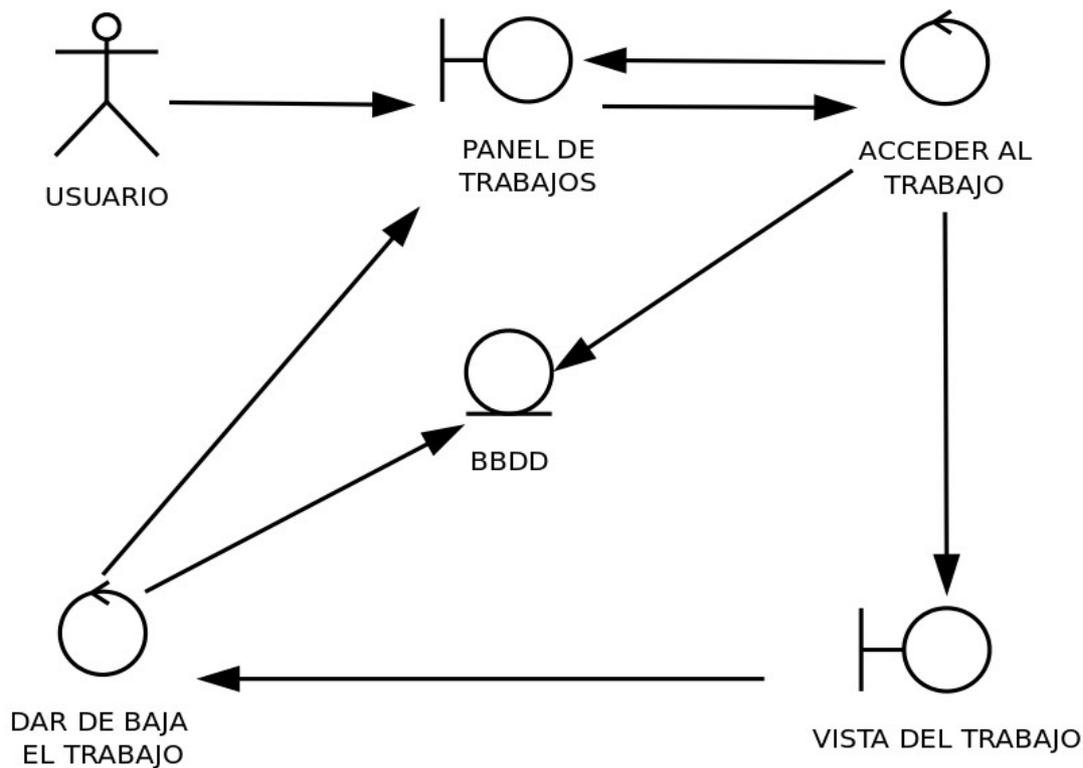


Fig. 19: Diagrama de robustez del caso de uso de eliminación de un trabajo.

### 5.5.14 Listar trabajos

Listar trabajos	
<b>Precondiciones</b>	El usuario debe haber iniciado sesión.
<b>Postcondiciones</b>	Ninguno que modifique el estado del sistema.
<b>Actores</b>	Usuario administrador o regular.
<b>Descripción</b>	<ol style="list-style-type: none"> <li>1. El usuario se dirigirá al panel de trabajos donde se le mostrará el listado de trabajos que tiene permiso para visualizar. Se incluirá la información básica del trabajo y su estado.</li> </ol>
<b>Excepciones</b>	<ol style="list-style-type: none"> <li>1. La base de datos no está disponible                         <ul style="list-style-type: none"> <li>◆ Se mostrará la página general de error de WebKumo.</li> </ul> </li> </ol>

El diagrama de robustez es, en este caso, idéntico al del caso de uso “Listar servicios”. La diferencia es que el panel al que se accede es al de trabajos en lugar de el de servicios y que la información mostrada es distinta.

### 5.5.15 Ver trabajo

<b>Ver trabajo</b>	
<b>Precondiciones</b>	El usuario debe haber iniciado sesión y al menos debe existir un trabajo dado de alta que el usuario tenga permisos para ver.
<b>Postcondiciones</b>	Ninguno que modifique el estado del sistema.
<b>Actores</b>	Usuario administrador o regular.
<b>Descripción</b>	<ol style="list-style-type: none"> <li>1. El usuario se dirigirá al panel de trabajos donde se le mostrará el listado de trabajos que tiene permiso para visualizar. Seleccionará uno de ellos.</li> <li>2. Se le dirigirá al usuario a la vista del trabajo seleccionado donde podrá ver sus parámetros y si ha sido ejecutado con éxito, los resultados de la ejecución.</li> </ol>
<b>Excepciones</b>	<ol style="list-style-type: none"> <li>1. La base de datos no está disponible <ul style="list-style-type: none"> <li>◆ Se mostrará la página general de error de WebKumo.</li> </ul> </li> <li>2. El usuario no tiene permisos para visualizar el trabajo. <ul style="list-style-type: none"> <li>◆ Si un usuario intenta acceder a la vista de un trabajo que no tiene permiso para visualizar, se le mostrará la vista de recurso no encontrado (HTTP 404).</li> </ul> </li> </ol>

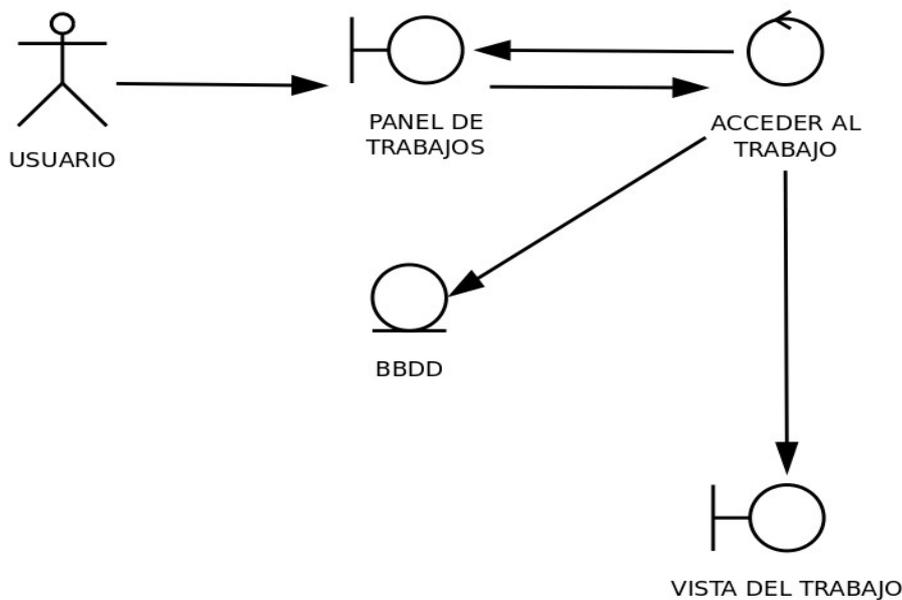
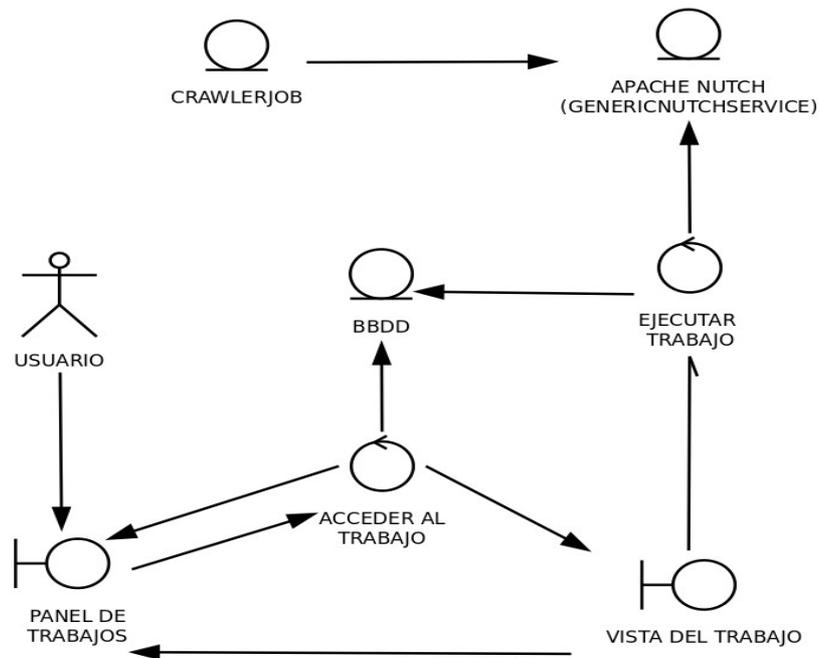


Fig. 20 Diagrama de robustez del caso de uso de vista de un trabajo.

### 5.5.16 Ejecutar trabajo

Ejecutar trabajo	
<b>Precondiciones</b>	El usuario debe haber iniciado sesión y al menos debe existir un trabajo dado de alta sin ejecutar que el usuario tenga permisos para ver.
<b>Postcondiciones</b>	Tras la ejecución de este caso de uso, un trabajo no ejecutado será remitido a la araña web para ser ejecutado y habrá cambiado su estado.
<b>Actores</b>	Usuario administrador o regular.
<b>Descripción</b>	<ol style="list-style-type: none"> <li>1. El usuario se dirigirá al panel de trabajos donde se le mostrará el listado de trabajos que tiene permiso para visualizar. Seleccionará uno de ellos.</li> <li>2. Se le dirigirá al usuario a la vista del trabajo seleccionado donde podrá ver sus parámetros y si ha sido ejecutado con éxito, los resultados de la ejecución.</li> <li>3. Seleccionará la opción de enviar su trabajo para ejecución.</li> <li>4. El usuario será dirigido al panel de trabajos donde</li> </ol>

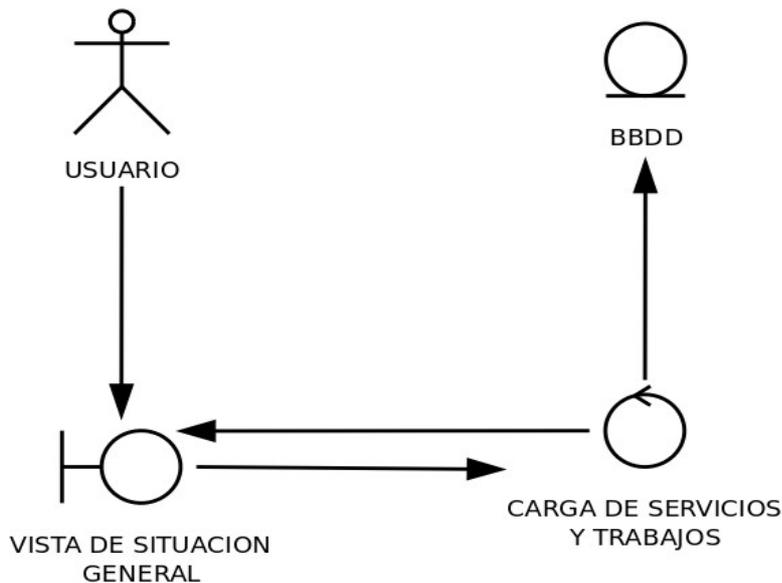
	verá el progreso del trabajo.
<b>Escenarios secundarios</b>	<ol style="list-style-type: none"> <li>Algunos de los servicios de infraestructura no está disponible. <ul style="list-style-type: none"> <li>Se indicará dicha situación al usuario.</li> </ul> </li> </ol>
<b>Excepciones</b>	<ol style="list-style-type: none"> <li>La base de datos no está disponible <ul style="list-style-type: none"> <li>Se mostrará la página general de error de WebKumo.</li> </ul> </li> <li>El usuario no tiene permisos para visualizar el trabajo. <ul style="list-style-type: none"> <li>Si un usuario intenta acceder a la vista de un trabajo que no tiene permiso para visualizar, se le mostrará la vista de recurso no encontrado (HTTP 404).</li> </ul> </li> </ol>



**Fig. 21: Diagrama de robustez del caso de uso de ejecución de un trabajo.**

### 5.5.17 Ver situación general

Ver situación general	
<b>Precondiciones</b>	El usuario debe haber iniciado sesión.
<b>Postcondiciones</b>	Ninguno que modifique el estado del sistema.
<b>Actores</b>	Usuario administrador o regular.
<b>Descripción</b>	<ol style="list-style-type: none"> <li>1. El usuario se dirigirá al panel de situación general donde se le mostrará el listado de los trabajos mas recientes que tiene permiso para visualizar. También se mostrará el estado de los servicios de infraestructura.</li> </ol>
<b>Excepciones</b>	<ol style="list-style-type: none"> <li>1. La base de datos no está disponible                     <ul style="list-style-type: none"> <li>◆ Se mostrará la página general de error de WebKumo.</li> </ul> </li> </ol>



**Fig. 22: Diagrama de robustez del caso de uso de vista de la situación general.**

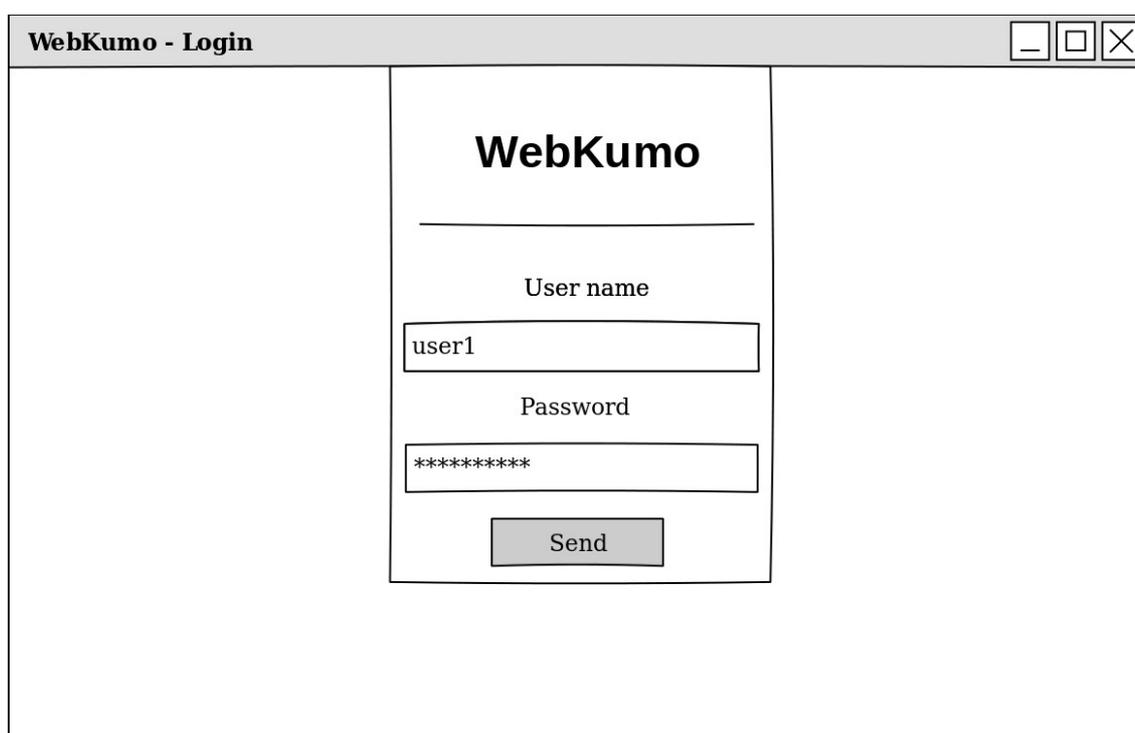
Este escenario podría encajar dentro del subsistema también de servicios ya que abarca tanto la situación general de estos como de los trabajos del usuario. Tampoco cabría definir un subsistema únicamente para este caso de uso por su pequeño tamaño.

## 5.6 Análisis de interfaces de usuario

En esta sección se describirán los principales aspectos de las interfaces de usuario que WebKumo dispondrá. Se ejemplificarán las interfaces mediante prototipos de cada tipo de pantalla que contendrán los elementos típicos de interfaces web procurando utilizar un patrón de diseño visual reconocido mediante el uso de barras de navegación, migas de pan y demás elementos.

### 5.6.1 Descripción de la interfaz

Para la interfaz de inicio de sesión se sugiere el siguiente esquema:



The image shows a wireframe of a login window titled "WebKumo - Login". The window contains a central form with the following elements: the title "WebKumo" at the top, a horizontal line separator, the label "User name" above a text input field containing "user1", the label "Password" above a text input field containing "\*\*\*\*\*", and a "Send" button at the bottom. The window has standard OS-style title bar controls (minimize, maximize, close) in the top right corner.

**Fig. 23: Prototipo de la pantalla de inicio de sesión**

Se trata de un formulario simple con dos campos utilizando un botón para el envío del mismo.

Tras el inicio de sesión se dirigirá a la vista de situación general cuyo aspecto preliminar es el que sigue:

WebKumo - Dashboard

**WebKumo** Jobs Services user1

**All services are running normally**

Service 1: **RUNNING**  
 Service 2: **RUNNING**  
 Service 3: **RUNNING**  
 Service 4: **RUNNING**

- Job #51569 FINISHED
- Job #0419 STOPPED
- Job #51 FINISHED
- Job #0895 FAILED
- Job #4419 RUNNING

Footer content Universidad de Oviedo

**Fig. 24: Prototipo de la pantalla de vista general**

Se tienen en la interfaz los siguientes elementos que serán comunes a todas las pantallas:

- ◆ **Barra de navegación:** En la parte superior se tiene una barra horizontal con enlaces o botones a las principales secciones de la aplicación.
- ◆ **Zona central:** Variará en función de la vista solicitada (vista general, servicios, trabajos, formularios de creación etc.).
- ◆ **Pie de página:** Ofrecerá información accesoria.

En el caso concreto de la pantalla de vista general se visualizará en una parte el estado de los servicios de infraestructura mediante información gráfica y textual y en otra parte un listado de los últimos trabajos visibles para el usuario.

A continuación se muestra un prototipo de pantalla para mostrar información de forma tabular. Se ejemplifica el prototipo con la pantalla que se utilizaría para ver el listado de trabajos y sus opciones de gestión.

	Name	Status	Owner	Creation date
#415	Example job 1	RUNNING	user1	05-06-2015
#696	Example job 2	STOPPED	user2	03-01-2015
#591	Example job 3	FINISHED	user 3	11-04-2015

**Fig. 25: Prototipo de la pantalla con información tabular (trabajos).**

La variación en esta pantalla es la zona central de contenido que en este caso consiste en la representación de los trabajos en forma de tabla. Esta misma pantalla es aplicable a la vista del panel del usuario administrador en la que se muestra la información sobre los usuarios registrados y los grupos creados en forma similar.

Otro tipo de pantalla cuyo prototipo se muestra a continuación es la de formulario, se ejemplifica con el formulario de creación de un trabajo. Esta pantalla de formulario es aplicable a otros formularios del sistema (creación de usuario, de grupo, etc.).

WebKumo - Jobs

WebKumo Dashboard Services user1

### Create a new job

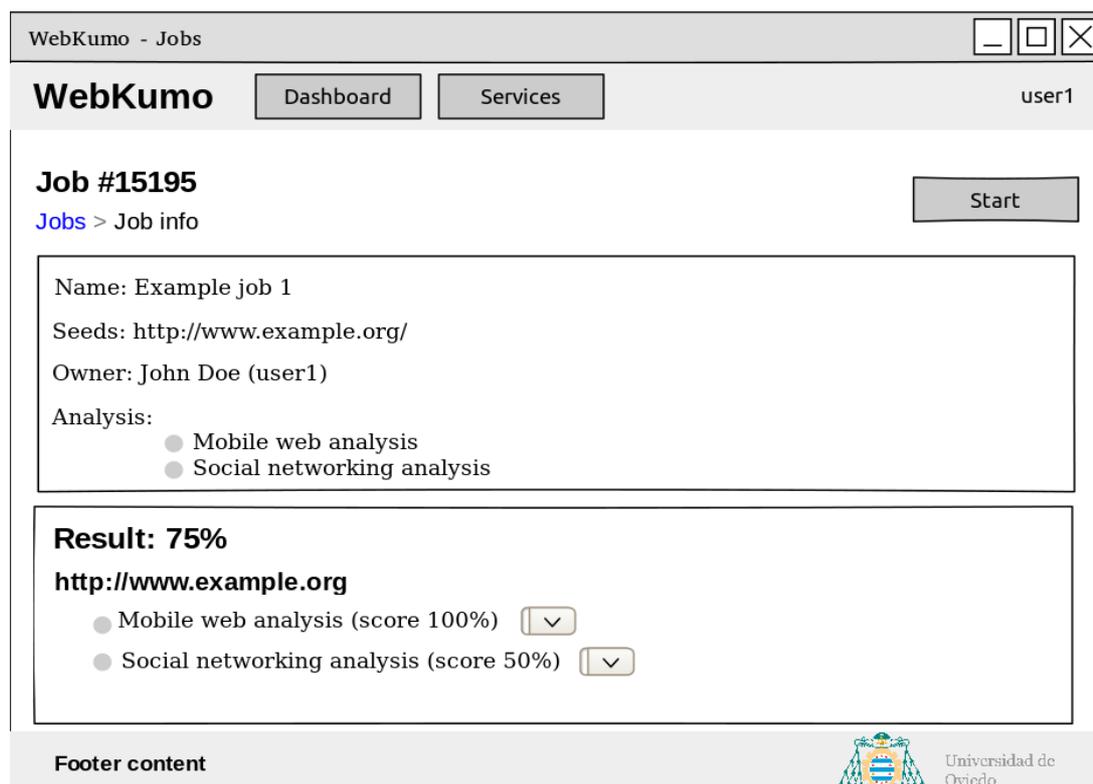
[Jobs](#) > Create a new job

Name:	<input type="text" value="Example job"/>
Description	<input type="text" value="Description for this job"/>
Resources / page	<input type="text" value="Example job"/>
Depth	<input type="text" value="Example job"/>
Seed	<input type="text" value="Example job"/> <input type="button" value="Browse"/>
Analyses	<input type="text" value="Analysis1 Analysis2"/> <input type="button" value="▲"/> <input type="button" value="▼"/>

Footer content  Universidad de Oviedo

**Fig. 26: Prototipo de una pantalla con formulario de ejemplo (trabajo nuevo).**

El último prototipo de pantalla es aquel en el que se muestra información para la vista de un trabajo y sus resultados. En ella en la parte superior se mostrará la información básica del trabajo (nombre, semillas que serán analizadas, análisis a realizar etc.). En la parte inferior se mostrará los resultados de la ejecución del trabajo ordenando dichos resultados por análisis y por dominio. Adicionalmente se detallarán los distintos campos que conforman el informe del análisis.



**Fig. 27: Prototipo de una pantalla de trabajo (trabajo nuevo).**

En cuanto a patrones generales de usabilidad se diseñarán las interfaces finales con criterios entre los que se encuentran:

- ◆ Interfaz sin ruido visual ni adornos recargantes.
- ◆ Logotipo de aplicación identificativo con enlace a la página principal.
- ◆ URLs sencillas.
- ◆ Etiquetado de campos de formulario.
- ◆ Los elementos `img` contendrán el atributo `alt`.
- ◆ Contraste adecuado entre fondo y contenido.

Se procurarán incorporar el mayor número posible de elementos que mejoren la usabilidad .

## 5.6.2 Descripción de comportamiento de la interfaz

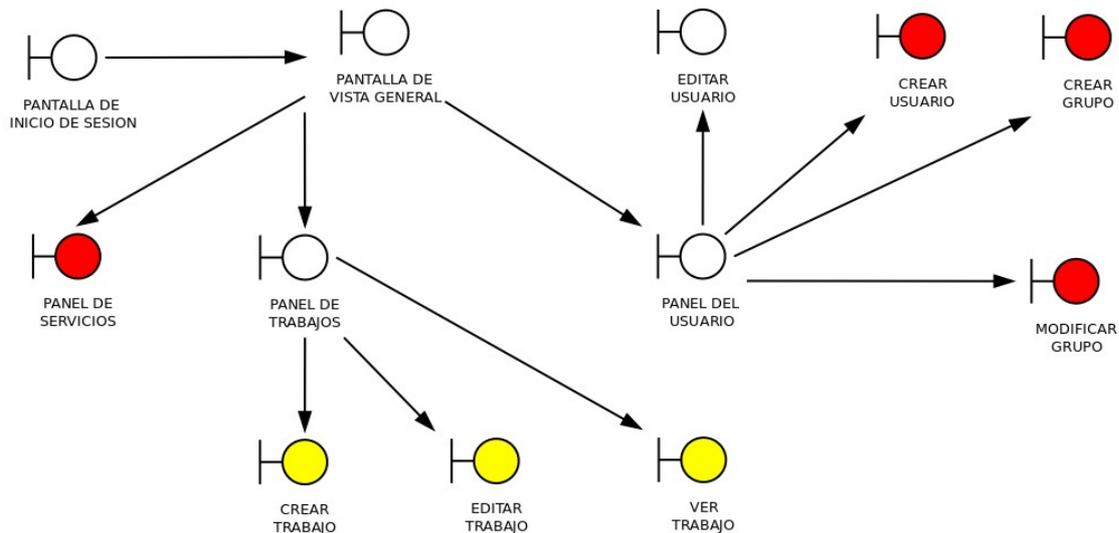
En esta sección se especificarán los convenios que la interfaz seguirá a la hora de interactuar con el usuario.

Cuando se trate de formularios será obligatoria la validación de los campos del mismo con el fin de depurar posibles errores en dichos campos. Se indicará mediante un panel con texto descriptivo del error o errores encontrados en el formulario. El panel debe estar resaltado (con color fuerte por ejemplo) para facilitar al usuario la búsqueda de errores.

Por otra parte, se incorporarán en aquellos campos en los que sea pertinente, sugerencias de valores válidos para la entrada.

Se utilizarán migas de pan o *breadcrumbs* para ayudar al usuario en procesos que involucren varias acciones. Cuando sea pertinente además, se indicará al menos la ocurrencia de errores al ejecutar una acción. Si el resultado de la operación es exitoso se podrá mostrar opcionalmente un mensaje confirmando el éxito en la acción llevada a cabo.

## 5.6.3 Diagrama de navegabilidad



**Fig. 28: Diagrama de navegabilidad de pantallas**

Como indicativo del diagrama de navegabilidad, aquellos nodos límite de color rojo estarán limitados únicamente a usuarios que posean el rol de administración. En el caso de los nodos límite rellenos de color amarillos, se trata de nodos en los que existirá alguna restricción en función de los permisos que el usuario que ha iniciado sesión tenga asignados.

## 5.7 Especificación del plan de pruebas

Se procederá en esta sección a esbozar el plan de pruebas que se llevará a cabo para mejorar la calidad del producto software mediante el desarrollo de comprobaciones manuales y automáticas que determinen si el comportamiento del sistema es el esperado.

Las pruebas a realizar se dividen según la siguiente clasificación:

1. **Pruebas unitarias:** Consistirá en la prueba automatizada de casos de uso específicos mediante la codificación de tests que probarán aspectos concretos de la implementación.
2. **Pruebas del sistema:** En este caso, las pruebas del sistema comprobarán la integración de todos los componentes del sistema verificando que se cumplan las funcionalidades que el sistema requiere mediante la prueba de funcionalidades.
3. **Pruebas de usabilidad:** Se determinarán en este plan las comprobaciones a llevar a cabo para asegurar que la experiencia de los usuarios con la aplicación es sencilla, esto es, que la interfaz no añada un grado de dificultad añadido al uso de la aplicación. Este punto se puede llevar a cabo mediante pruebas con usuarios reales.
4. **Pruebas de accesibilidad:** Para las pruebas de este tipo, se verificará que la interfaz de usuario cumpla alguna de las guías (total o parcialmente) de accesibilidad reconocidas de la industria.

### 5.7.1 Pruebas para casos de uso por subsistemas

#### 5.7.1.1 Subsistema de usuarios

INICIAR SESIÓN	
Prueba	Resultado esperado
Introducir credenciales correctas	Se debe iniciar sesión dirigiendo al usuario a la pantalla de vista general.

Introducir credenciales incorrectas.	Impedir el acceso mostrando un mensaje de error.
--------------------------------------	--

<b>CREAR USUARIO</b>	
<b>Prueba</b>	<b>Resultado esperado</b>
Crear un usuario con un nombre de usuario ( <i>login</i> ) que ya existe.	El sistema no contendrá un nuevo usuario y se le deberá indicar al usuario el error en el proceso.
Crear un usuario con valores inválidos.	El sistema no contendrá un nuevo usuario y se le deberá indicar al usuario el error en el proceso.
Crear un usuario con sus valores válidos y que no exista previamente.	El sistema contendrá un nuevo usuario registrado.
Crear un nuevo usuario (válido o no) por parte de un usuario no administrador.	El sistema no contendrá un nuevo usuario dado de alta y dirigirá al usuario a una página de recurso no encontrado.

<b>MODIFICAR USUARIO</b>	
<b>Prueba</b>	<b>Resultado esperado</b>
Cambiar el nombre del usuario por un nombre de usuario ( <i>login</i> ) que ya existe.	El usuario no se modificará y se deberá notificar un error en el proceso.
Modificar un usuario con valores inválidos.	El sistema no modificará el usuario en edición y se deberá indicar un error en el proceso.
Modificar un usuario con valores válidos y con un nombre de usuario que no exista previamente.	El sistema modificará el usuario preexistente con los nuevos valores.
Modificar un usuario con valores válidos por parte de un usuario no administrador.	El sistema no modificará el usuario y dirigirá al usuario a una página de recurso no encontrado.

<b>ELIMINAR USUARIO</b>	
<b>Prueba</b>	<b>Resultado esperado</b>
Eliminar un usuario con trabajos asociados	El sistema no dará de baja al usuario y mostrará un mensaje de error.
Eliminar un usuario que no tenga trabajos asociados.	El sistema contendrá un usuario menos.
Un usuario no administrador intentará eliminar otro usuario.	El sistema no dará de baja al usuario que se intenta eliminar y dirigirá al usuario a una página de recurso no encontrado.

<b>CREAR GRUPO</b>	
<b>Prueba</b>	<b>Resultado esperado</b>
Crear un grupo con un nombre de grupo que ya existe.	El sistema no contendrá un nuevo grupo y se le deberá indicar al usuario el error en el proceso.
Crear un grupo con valores inválidos.	El sistema no contendrá un nuevo grupo y se le deberá indicar al usuario el error en el proceso.
Crear un grupo con sus valores válidos y que no exista previamente.	El sistema contendrá un nuevo grupo registrado.
Crear un nuevo grupo (válido o no) por parte de un usuario no administrador.	El sistema no contendrá un nuevo grupo dado de alta y dirigirá al usuario a una página de recurso no encontrado.

<b>MODIFICAR GRUPO</b>	
<b>Prueba</b>	<b>Resultado esperado</b>
Cambiar el nombre del grupo por un nombre de grupo que ya existe.	El grupo no se modificará y se deberá notificar un error en el proceso.
Modificar un grupo con valores inválidos.	El sistema no modificará el grupo en edición y se deberá indicar un error en el proceso.

Modificar un grupo con valores válidos y con un nombre de grupo que no exista previamente.	El sistema modificará el grupo preexistente con los nuevos valores.
Modificar un grupo con valores válidos por parte de un usuario no administrador.	El sistema no modificará el grupo y dirigirá al usuario a una página de recurso no encontrado.

### ELIMINAR GRUPO

Prueba	Resultado esperado
Eliminar un grupo con usuarios miembros.	El sistema no dará de baja el grupo y mostrará un mensaje de error.
Eliminar un grupo que no tenga miembros asociados.	El sistema contendrá un grupo menos.
Un usuario no administrador intentará eliminar un grupo.	El sistema no dará de baja el grupo que se intenta eliminar y dirigirá al usuario a una página de recurso no encontrado.

### LISTAR USUARIOS

Prueba	Resultado esperado
Listar los usuarios.	El sistema obtendrá la información de los usuarios registrados y se los mostrará al usuario.
Un usuario no administrador intenta listar los usuarios.	Los usuarios no serán listados.

### LISTAR GRUPOS

Prueba	Resultado esperado
Listar los grupos.	El sistema obtendrá la información de los grupos registrados y se los mostrará al usuario.
Un usuario no administrador intenta listar los grupos.	Los grupos no serán listados.

### 5.7.1.2 Subsistema de servicios

<b>ARRANCAR SERVICIO</b>	
<b>Prueba</b>	<b>Resultado esperado</b>
Iniciar un servicio	El servicio arrancará mostrando un mensaje con el éxito de la operación.
Un usuario no administrador intenta arrancar un servicio.	El servicio no se arrancará.

<b>DETENER SERVICIO</b>	
<b>Prueba</b>	<b>Resultado esperado</b>
Detener un servicio	El servicio se detendrá mostrando un mensaje con el éxito de la operación.
Detener un servicio cuando existen trabajos en ejecución.	El servicio no parará y se mostrará un mensaje indicando el fallo de la operación.
Un usuario no administrador intenta detener un servicio.	El servicio no parará.

<b>LISTAR SERVICIOS</b>	
<b>Prueba</b>	<b>Resultado esperado</b>
Listar los servicios.	El sistema obtendrá la información de los servicios registrados y se los mostrará al usuario.
Un usuario no administrador intenta listar los servicios.	Los servicios no serán listados.

### 5.7.1.3 Subsistema de trabajos

<b>CREAR TRABAJO</b>	
<b>Prueba</b>	<b>Resultado esperado</b>
Crear un trabajo con valores inválidos (fuera de rango).	El sistema no se verá modificado y el usuario será notificado de dicho error.
Crear un trabajo con valores válidos.	El sistema contendrá un trabajo nuevo.
Un usuario sin permisos para ello, intentará crear un nuevo trabajo.	El sistema no contendrá un nuevo trabajo y el usuario será notificado de la imposibilidad de crear el trabajo.

<b>MODIFICAR TRABAJO</b>	
<b>Prueba</b>	<b>Resultado esperado</b>
Editar un trabajo con valores inválidos (fuera de rango).	El sistema no se verá modificado y el usuario será notificado de dicho error.
Editar un trabajo que se encuentra en ejecución o finalizado.	El sistema no salvará los cambios en el trabajo y el usuario será notificado.
Un usuario sin permisos para ello, intentará crear un nuevo trabajo.	El sistema no contendrá un nuevo trabajo y el usuario será notificado de la imposibilidad de crear el trabajo.

<b>ELIMINAR TRABAJO</b>	
<b>Prueba</b>	<b>Resultado esperado</b>
Eliminar un trabajo que no ha sido ejecutado todavía.	El sistema contendrá un trabajo menos.

Eliminar un trabajo que ya ha sido ejecutado.	El sistema no dará de baja el trabajo y el usuario será notificado del error.
Un usuario sin permisos para ello, intentará eliminar un trabajo existente.	El sistema no se verá modificado y el usuario será notificado de la imposibilidad de crear el trabajo.

<b>VER TRABAJO</b>	
<b>Prueba</b>	<b>Resultado esperado</b>
Visualizar la información de un trabajo.	El sistema obtendrá la información del trabajo y se la mostrará al usuario.
Visualizar la información de un trabajo por parte de un usuario sin permisos para ello.	El sistema indicará al usuario la condición del error.

<b>LISTAR TRABAJOS</b>	
<b>Prueba</b>	<b>Resultado esperado</b>
Obtener el listado de trabajos.	El sistema obtendrá la información de los trabajos que puede ver y se la mostrará al usuario.

### 5.7.1.4 Subsistema de vista general

<b>VER SITUACIÓN GENERAL</b>	
<b>Prueba</b>	<b>Resultado esperado</b>
Acceder a la vista general.	El sistema obtendrá la información de situación de los servicios y de los últimos trabajos del usuario y se los mostrará.

### 5.7.1.5 Subsistema de WebKumo Connector

WEBKUMO CONNECTOR	
Prueba	Resultado esperado
Analizar recursos.	El <i>plugin</i> generará los resultados de análisis para los recursos indicados.
Analizar recursos cuyo tipo MIME es desconocido.	El <i>plugin</i> ignorará los recursos sin generar resultados.

En cuanto al subsistema que conforma la infraestructura de procesamiento, la realización de pruebas unitarias sobre el software que lo compone, corresponde a sus desarrolladores.

### 5.7.2 Pruebas de sistema

USUARIOS Y GRUPOS	
Prueba	Resultado esperado
Crear un usuario.	Un usuario administrador creará un usuario, posteriormente podrá visualizar dicho usuario en el listado del panel del usuario y se probará a iniciar sesión en WebKumo con el nuevo usuario.
Crear un grupo	Se creará un nuevo grupo y se añadirán usuarios al mismo. Se comprobará mediante la ficha del usuario añadido, que aparece en el nuevo grupo.

SERVICIOS	
Prueba	Resultado esperado
Gestionar un servicio	Un usuario administrador arrancará y detendrá un servicio comprobando que dichas operaciones surten efecto.

TRABAJOS	
Prueba	Resultado esperado
Ejecutar un trabajo	Un usuario creará un nuevo trabajo que luego podrá visualizar en el panel de trabajos. A continuación lo ejecutará y cuando este finalice comprobará que se generaron resultados del análisis.

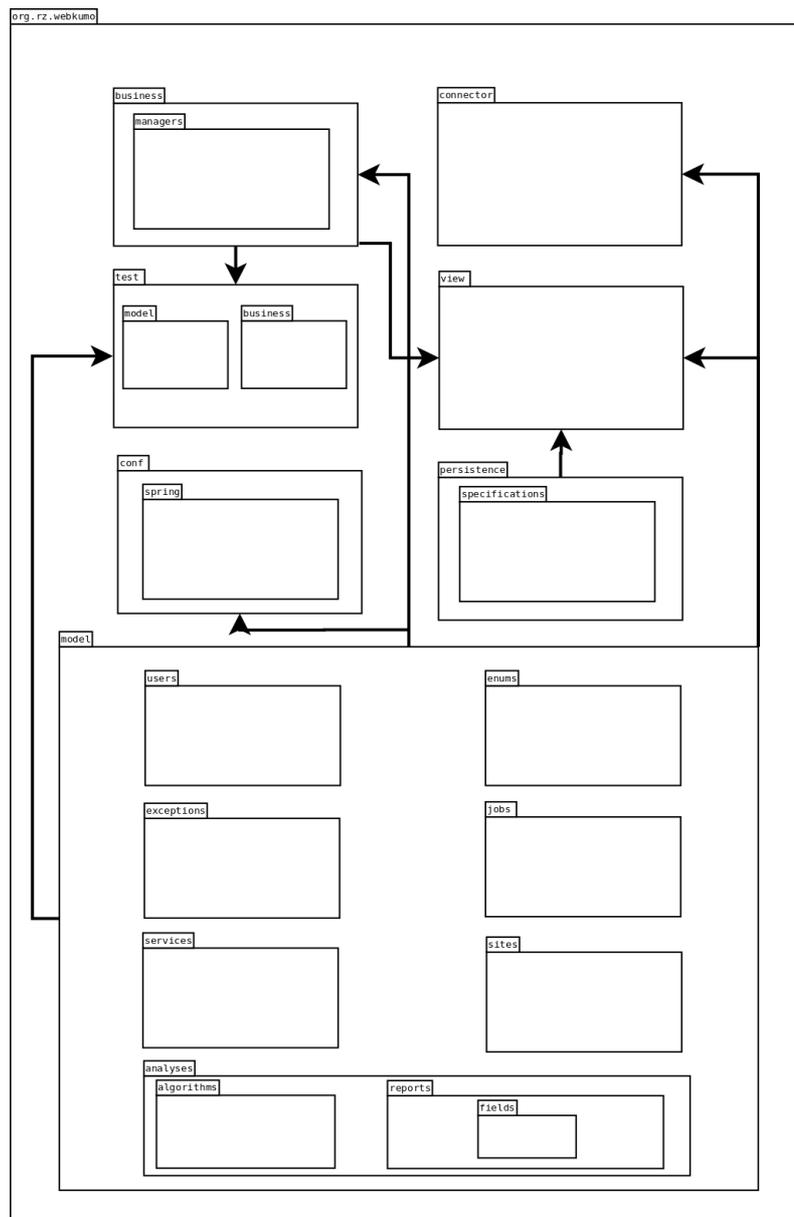
### 5.7.3 Pruebas de accesibilidad y usabilidad

Se diseñarán las interfaces de usuario atendiendo a los criterios más comunes de diseño usable. Para la evaluación de la interfaz se realizará una evaluación mediante formularios utilizando varios usuarios potenciales que se adapten al perfil objetivo del producto. Asimismo para el apartado de accesibilidad se verificarán los aspectos técnicos de las interfaces mediante la evaluación de la misma utilizando alguna guía reconocida.

# 6 Diseño del sistema

## 6.1 Arquitectura del sistema

### 6.1.1 Diagramas de paquetes



**Fig. 29: Diagrama de paquetes**

### 6.1.1.1 Paquete 'business'

Este paquete contiene la lógica de negocio, esto es, aquellas clases denominadas *managers* que se encargan de intermediar entre la vista (los controladores localizados en el paquete *view*) y el modelo. Se ha procurado dejar la capa que contiene la lógica de negocio lo mas fina posible para evitar el antipatrón de modelos de datos anémicos **[ANEM]**. El paquete contiene interfaces para definir las operaciones que cada *manager* debe contener. El subpaquete *managers* contiene las clases de implementación de las mencionadas interfaces.

### 6.1.1.2 Paquete 'connector'

En este caso, el paquete contiene la implementación del plugin para Apache Nutch que se encarga de realizar el procesamiento básico de los recursos HTML, CSS o JavaScript descargados. Además de eso, se encarga de ejecutar los análisis inyectados desde la aplicación WebKumo así como de escribir los resultados serializados en JSON en el sistema de archivos HDFS.

Adicionalmente se incluyen clases de utilidad del *plugin* original HTMLParser de Apache Nutch y ficheros XML de configuración.

### 6.1.1.3 Paquete 'view'

El paquete *view* contendrá la lógica de los controladores. Cada controlador procesa la petición del usuario invocando a los servicios de la capa de negocio en función de lo requerido. Además se encarga de controlar el flujo de navegación y la validación de formularios por parte del servidor.

### 6.1.1.4 Paquete 'persistence'

Este paquete contiene las clases encargadas de gestionar la persistencia del sistema. Se utilizaron interfaces para definir repositorios de entidades que permitan gestionar la persistencia de cada entidad del modelo de datos. El subpaquete *specifications* se utiliza para definir consultas programáticamente.

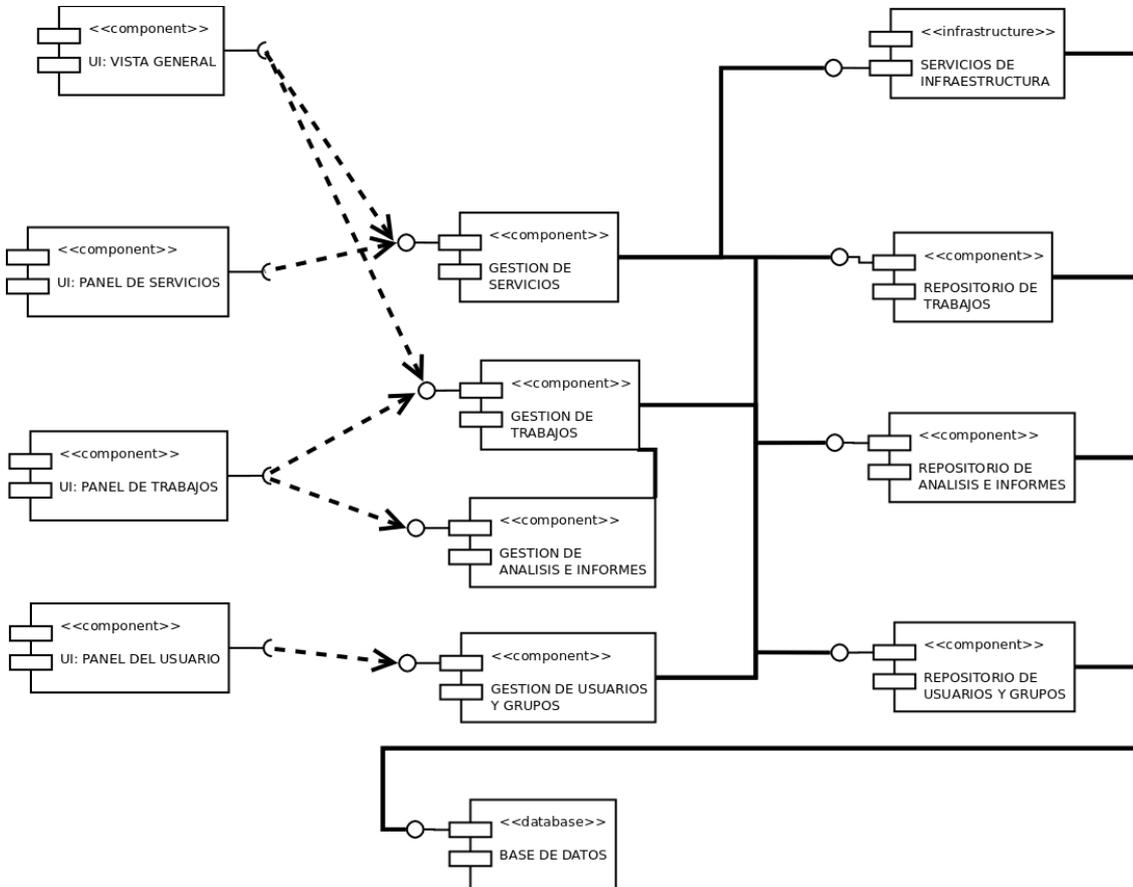
### 6.1.1.5 Paquete 'conf'

Este paquete contendrá las factorías para poder generar los distintos tipos de entidades con mayor control. Además contendrá un subpaquete *spring* donde se encontrará las clases de configuración del sistema para la aplicación web.

### 6.1.1.6 Paquete 'model'

El paquete *model* contiene las clases del dominio de datos. El subpaquete *reports* contiene las implementaciones concretas de los informes de los análisis. En el caso del subpaquete *algorithms* se trata de las implementaciones concretas de cada algoritmo de análisis. Por último el subpaquete *fields* contiene los subtipos de campo de informe que se puedan necesitar.

## 6.2 Diagrama de componentes

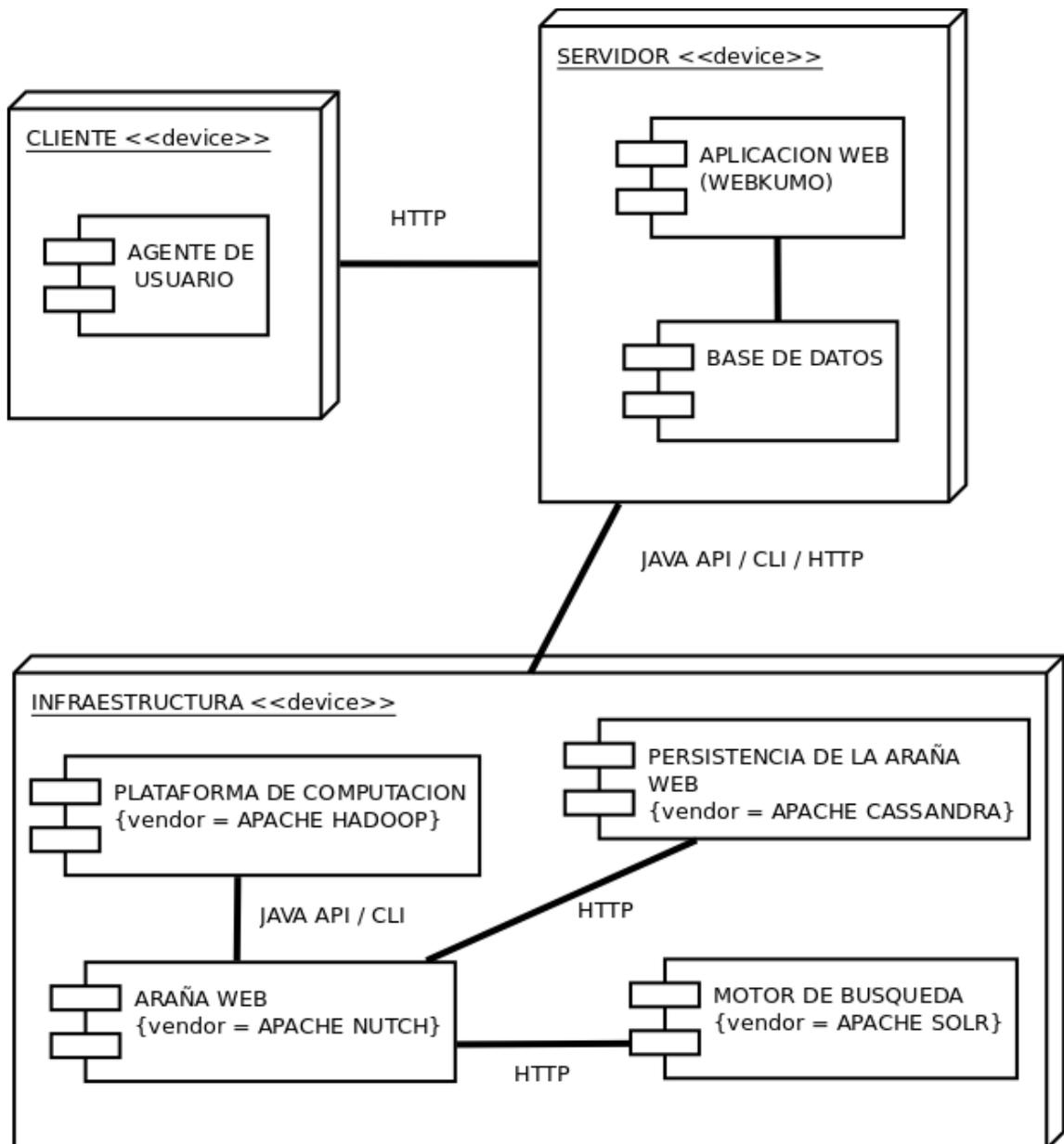


**Fig. 30: Diagrama de componentes**

En el diagrama se aprecian tres zonas diferenciadas, en la parte izquierda se tiene la interfaz web con la que los usuarios interactuarán. En la parte central se encuentran los componentes que condensan la lógica de negocio que responde a las peticiones de los usuarios. Finalmente en la parte derecha del esquema se encuentra la parte que se encarga de la persistencia de los datos.

Se han separado los componentes por funcionalidad sobre los principales subsistemas. Todo ello se encuentra sustentado por una base de datos dedicada a la persistencia de los datos de la aplicación web WebKumo (esta base de datos es independiente de la base de datos que utilizará la araña web).

## 6.3 Diagrama de despliegue



**Fig. 31: Diagrama de despliegue**

### 6.3.1 Agente de usuario

El agente de usuario es representado por el navegador web mediante el cuál los usuarios interactúan con el sistema a través de la aplicación web.

## 6.3.2 Aplicación web WebKumo

Este componente es el intermediario entre el usuario y la plataforma que conforma la infraestructura de análisis y descarga. Se trata de una aplicación web que se ejecutará sobre un servidor web. La aplicación web, además de recoger la interacción del usuario, realizará la gestión de usuarios y principalmente servirá para controlar la infraestructura de procesamiento así como interfaz que permitirá visualizar los resultados de los análisis.

La aplicación web se ejecutará en la misma máquina que el nodo principal (*master*) de la plataforma de computación distribuida.

## 6.3.3 Base de datos

Se trata de una base de datos relacional que almacenará la información relativa al funcionamiento de la aplicación web. Contendrá entre otros, los usuarios registrados, los grupos, los trabajos generados, los resultados de los análisis etc.

## 6.3.4 Plataforma de computación

Como se ha descrito en puntos anteriores de esta documentación, los trabajos de descarga y análisis se ejecutarán de manera distribuida en una plataforma que permita, si es requerido, un clúster de computación. Este componente representa tanto el software de computación distribuida como los nodos que vayan a formar finalmente el clúster.

## 6.3.5 Araña web

Es el software que se ejecutará sobre la plataforma de computación y que realizará los trabajos de descarga y análisis. Estará controlada desde la aplicación web la cuál le enviará trabajos para su ejecución.

### 6.3.6 Persistencia de la araña web

Software que almacena la información descargada por la araña web. Comprende al software de base de datos y los nodos del clúster en los que se despliegue.

### 6.3.7 Motor de búsqueda

Los recursos descargados por la araña web serán opcionalmente indexados en un motor de búsqueda que forma parte de la plataforma de computación. Al igual que en ocasiones anteriores, este punto contemplará tanto el software como el nodo del clúster que lo ejecuta.

## 6.4 Diseño de clases

En este apartado se mostrarán las distintas clases que conforman el sistema desarrollado. Se incluyen tanto las clases del dominio de datos como aquellas clases relativas a la gestión de la persistencia, la capa de negocio, la vista y demás elementos accesorios.

Debido al elevado número de clases que conforman el diagrama de clases, se omite presentar aquí un diagrama global que contenga todas las clases y sus relaciones. Se indicarán a continuación los diagramas de clases separados por paquetes Java. Como anexo a la documentación se incluye el diagrama de clases completo para su consulta.

### 6.4.1 Paquete org.rz.webkumo.view

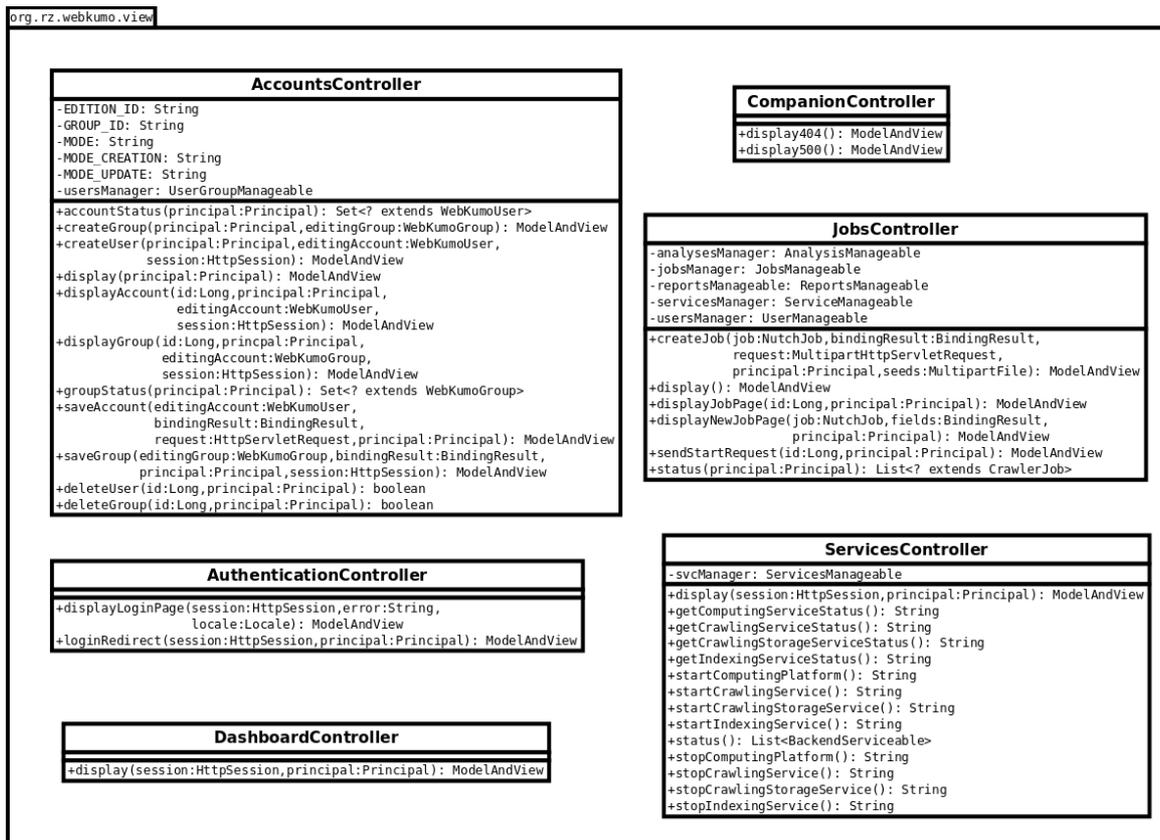


Fig. 32: Diagrama de clases del paquete de vista.

## 6.4.2 Paquete org.rz.webkumo.business

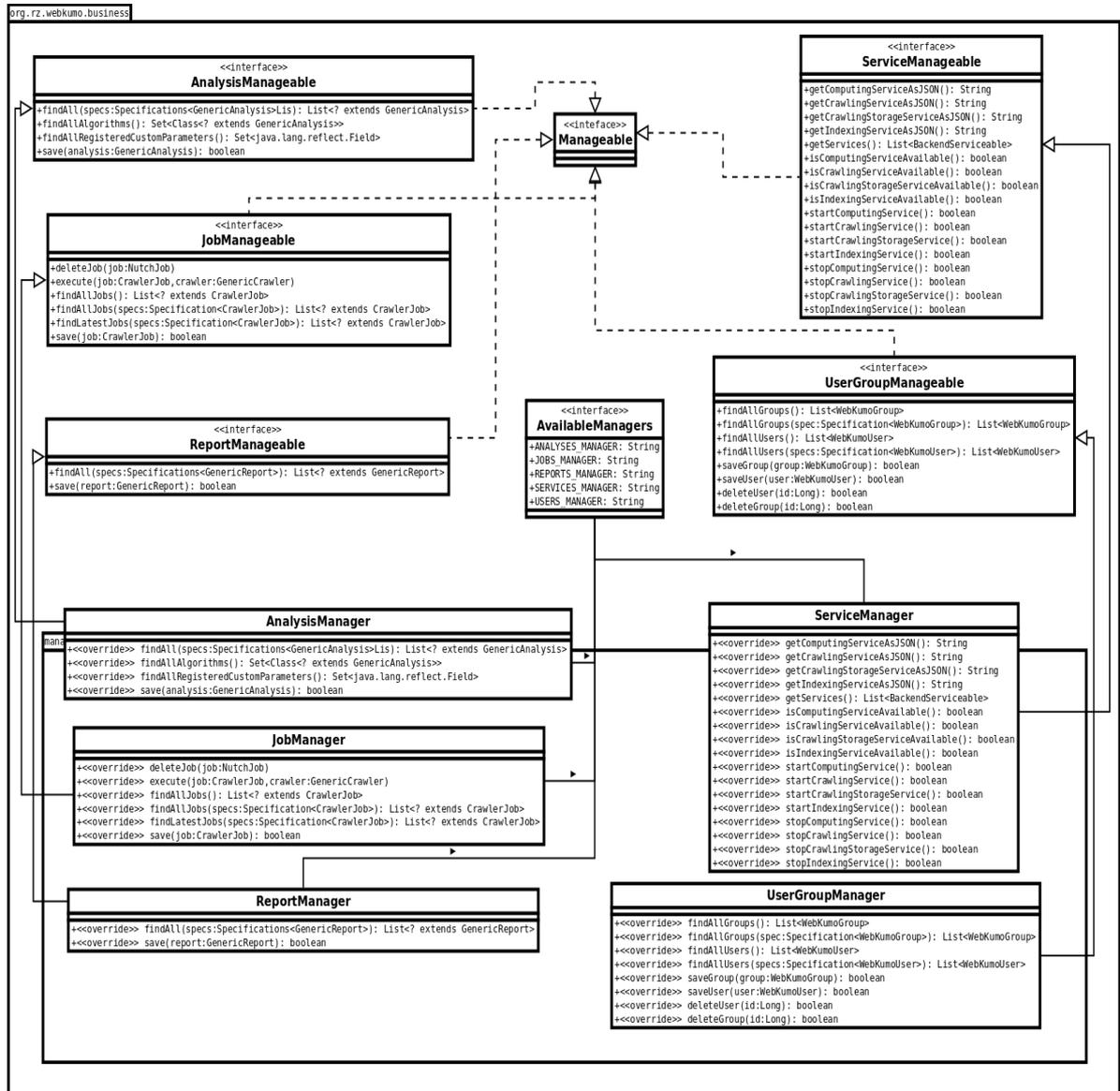


Fig. 33: Diagrama de clases de la capa de negocio

### 6.4.3 Paquete org.rz.webkumo.persistence

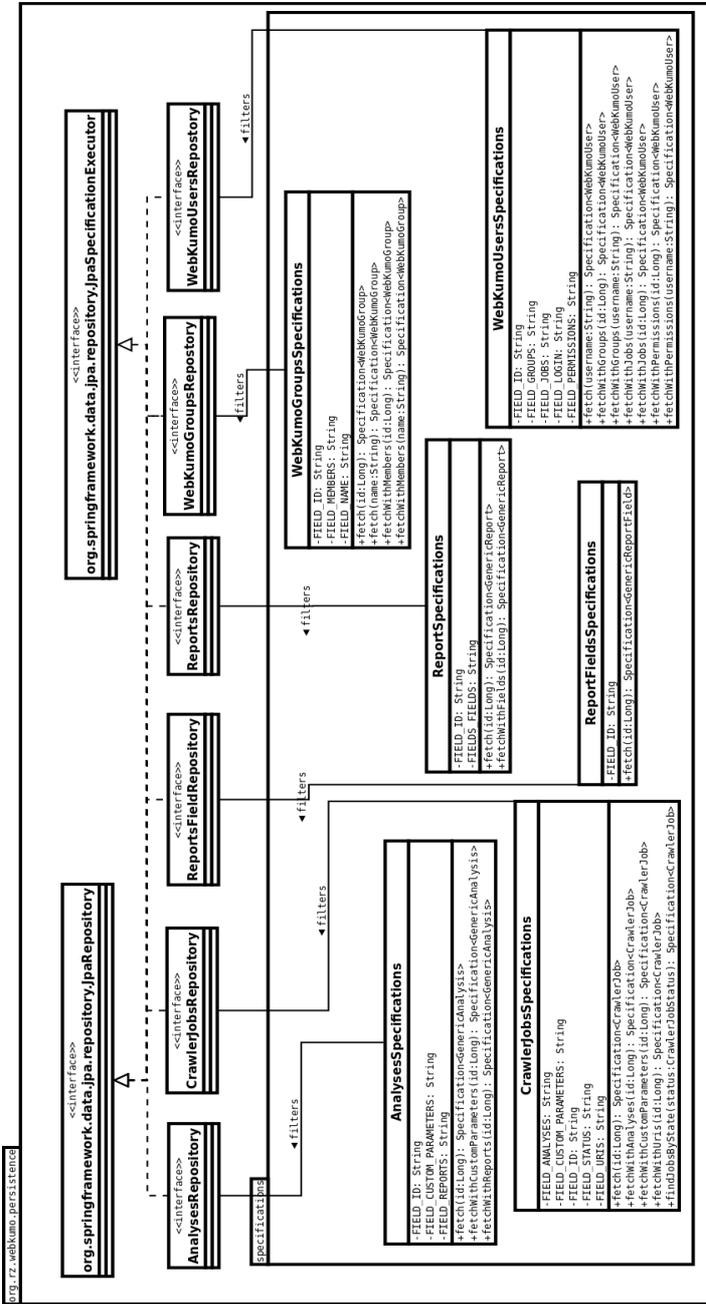


Fig. 34: Diagrama de clases de la capa de persistencia



### 6.4.5 Paquete org.rz.webkumo.connector

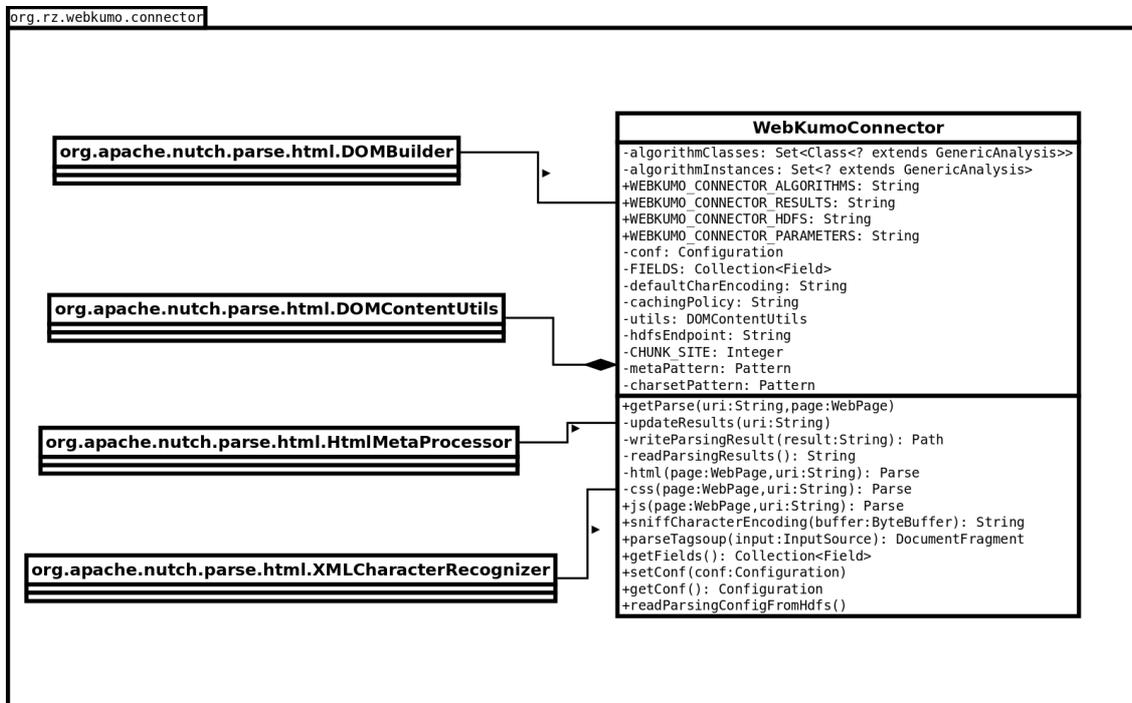


Fig. 36: Diagrama de clases del plugin WebKumo Connector

### 6.4.6 Paquete org.rz.webkumo.model.exceptions

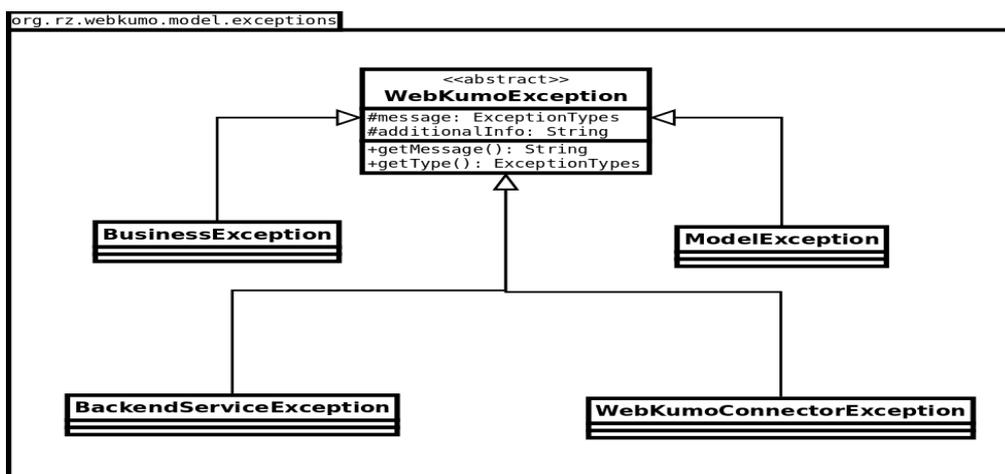


Fig. 37: Diagrama de clases del modelo de excepciones

## 6.4.7 Paquete org.rz.webkumo.model.analysises

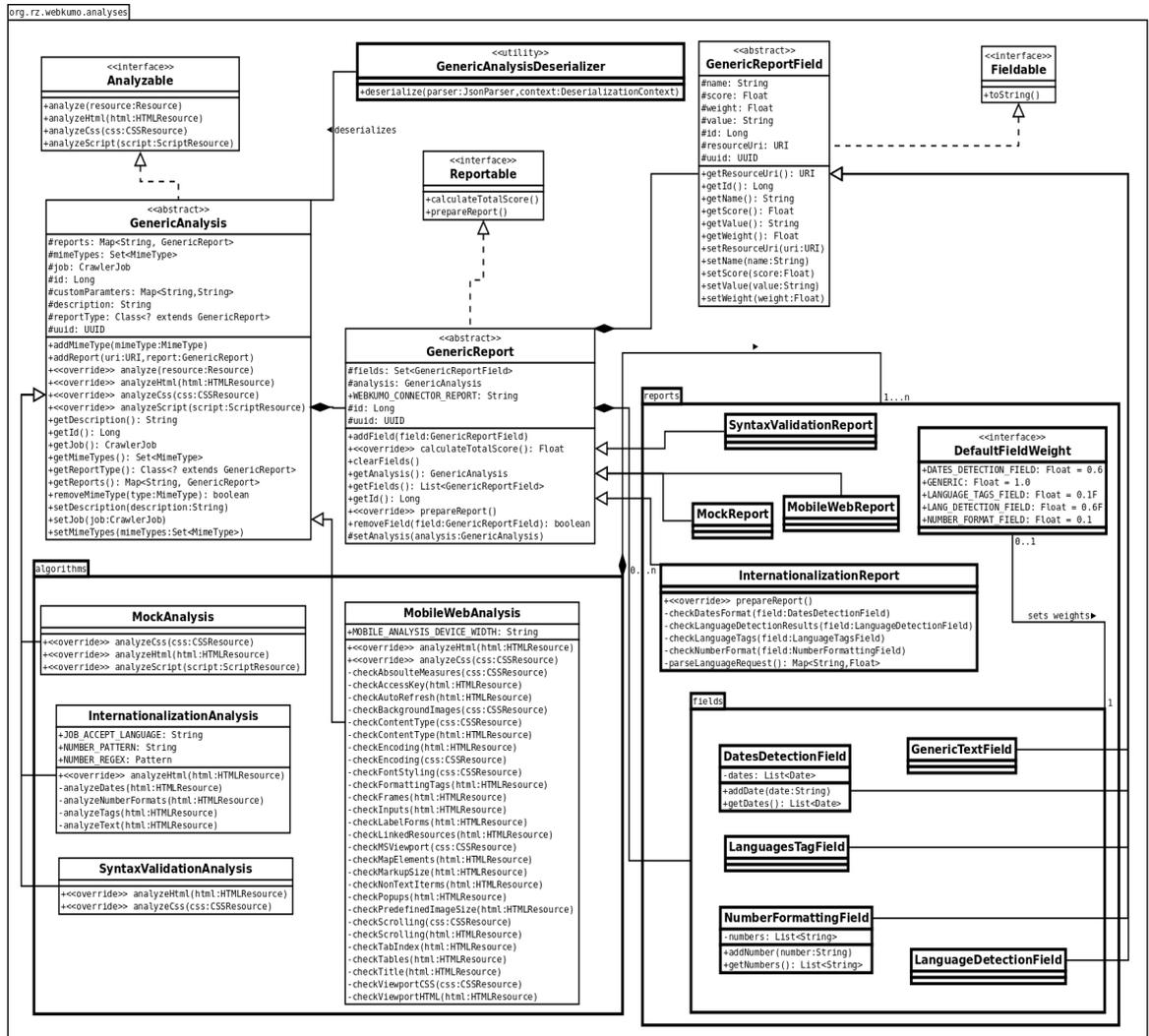


Fig. 38: Diagrama de clases del modelo de análisis.

### 6.4.8 Paquete org.rz.webkumo.model.users

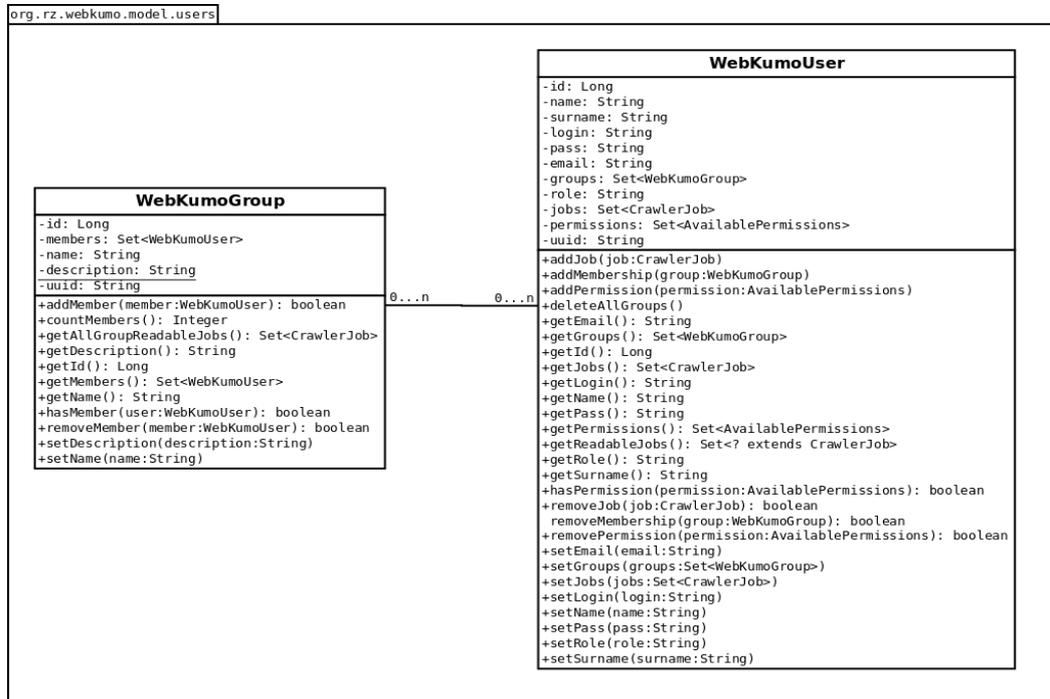


Fig. 39: Diagrama de clases del modelo de usuarios.

### 6.4.9 Paquete org.rz.webkumo.model.sites

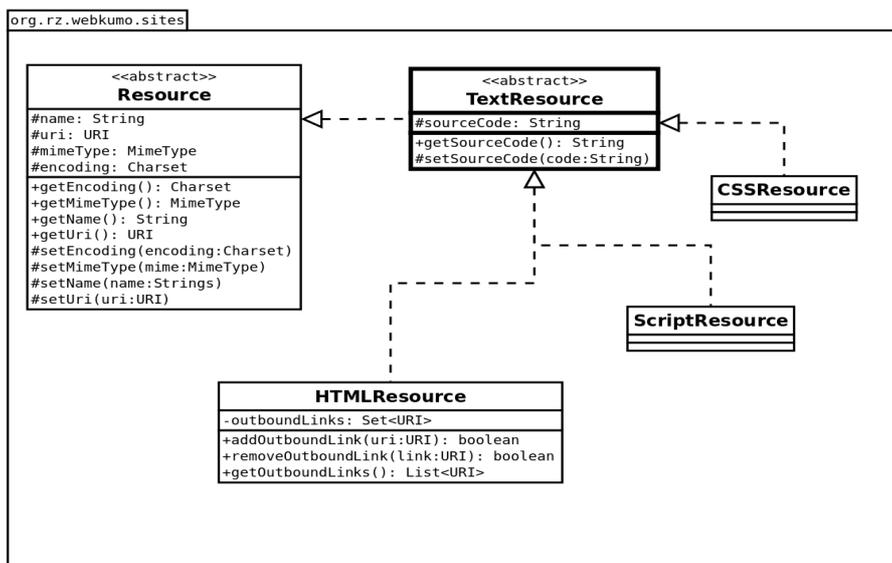


Fig. 40: Diagrama de clases del modelo de recursos web.

## 6.4.10 Paquete org.rz.webkumo.model.jobs

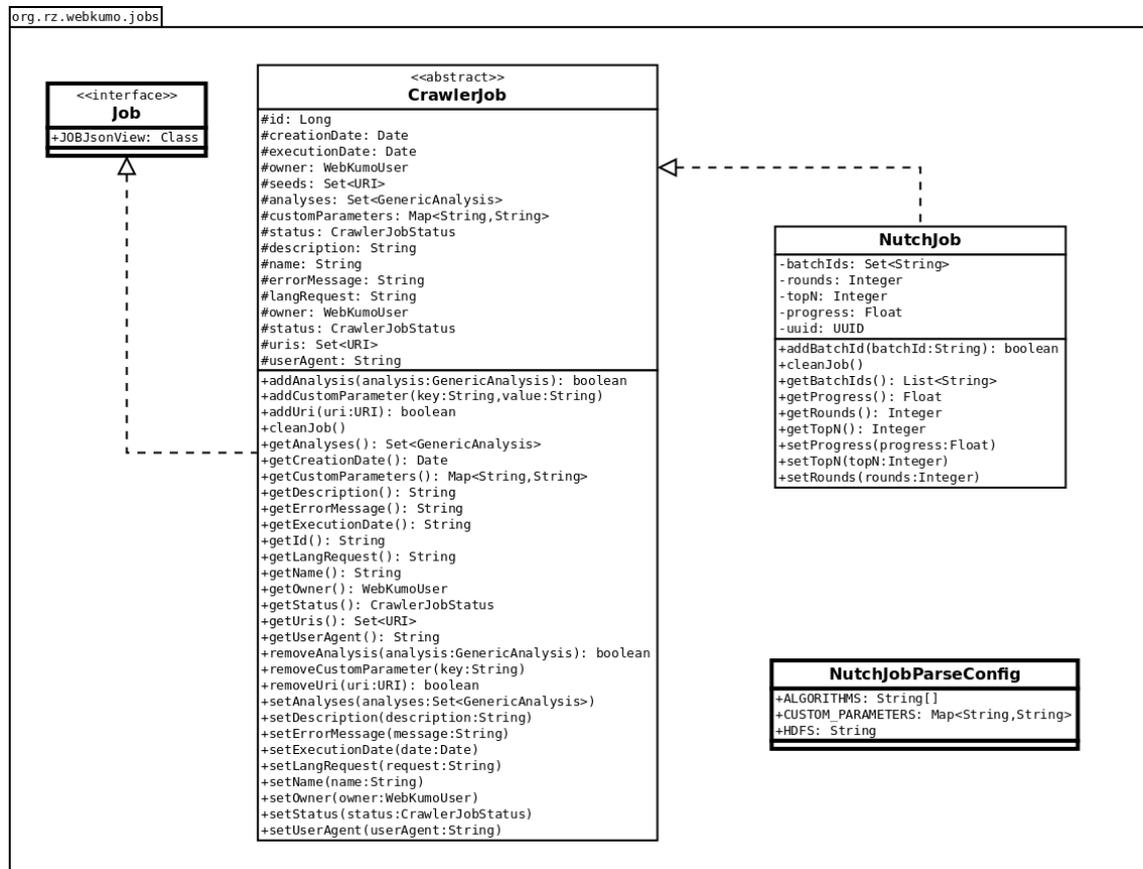


Fig. 41: Diagrama de clases del modelo de trabajos de araña web.

### 6.4.11 Paquete org.rz.webkumo.model.enums

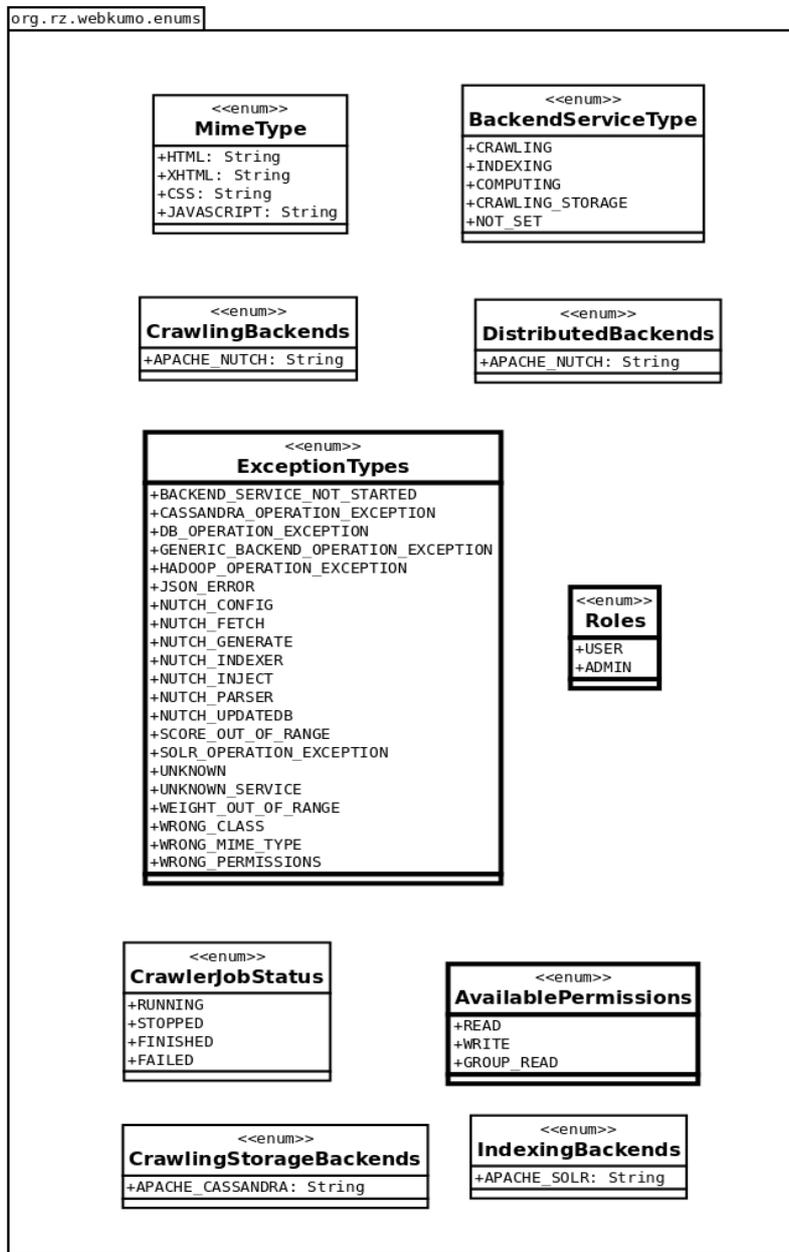


Fig. 42: Diagrama de clases de enumeraciones del modelo

## 6.4.12 Paquete org.rz.webkumo.model.services

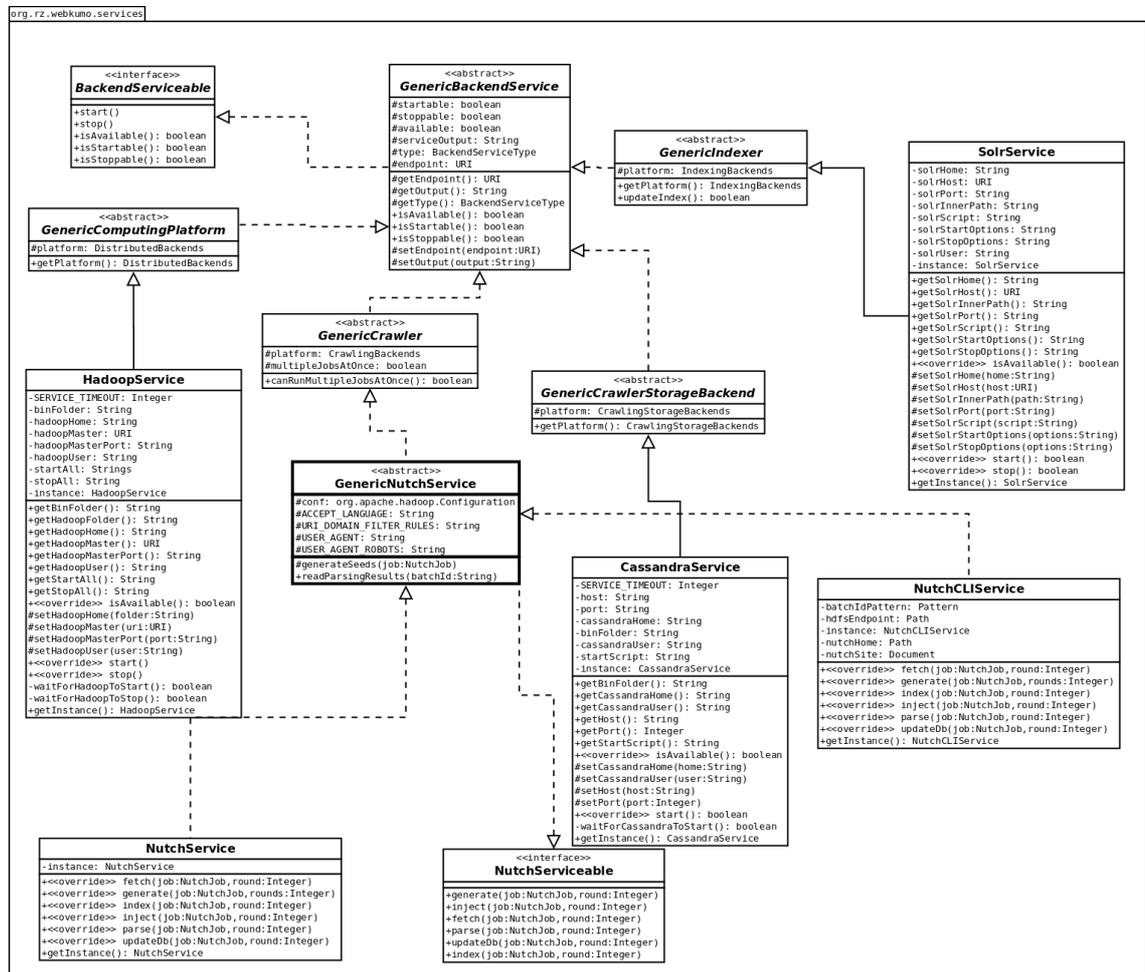


Fig. 43: Diagrama de clases del modelo de servicios de infraestructura.

## 6.5 Diagramas de interacción y de estados

Para clarificar cómo se ejecutan las interacciones entre subsistemas ante una operación solicitada por el usuario, se describirán diagramas de interacción para los casos de uso descritos anteriormente en la documentación.

Por simplicidad y para no repetir diagramas similares, aquellos casos de uso que se traduzcan en diagramas análogos serán agrupados indicándolo convenientemente.

### 6.5.1 Caso de uso 'Crear usuario'

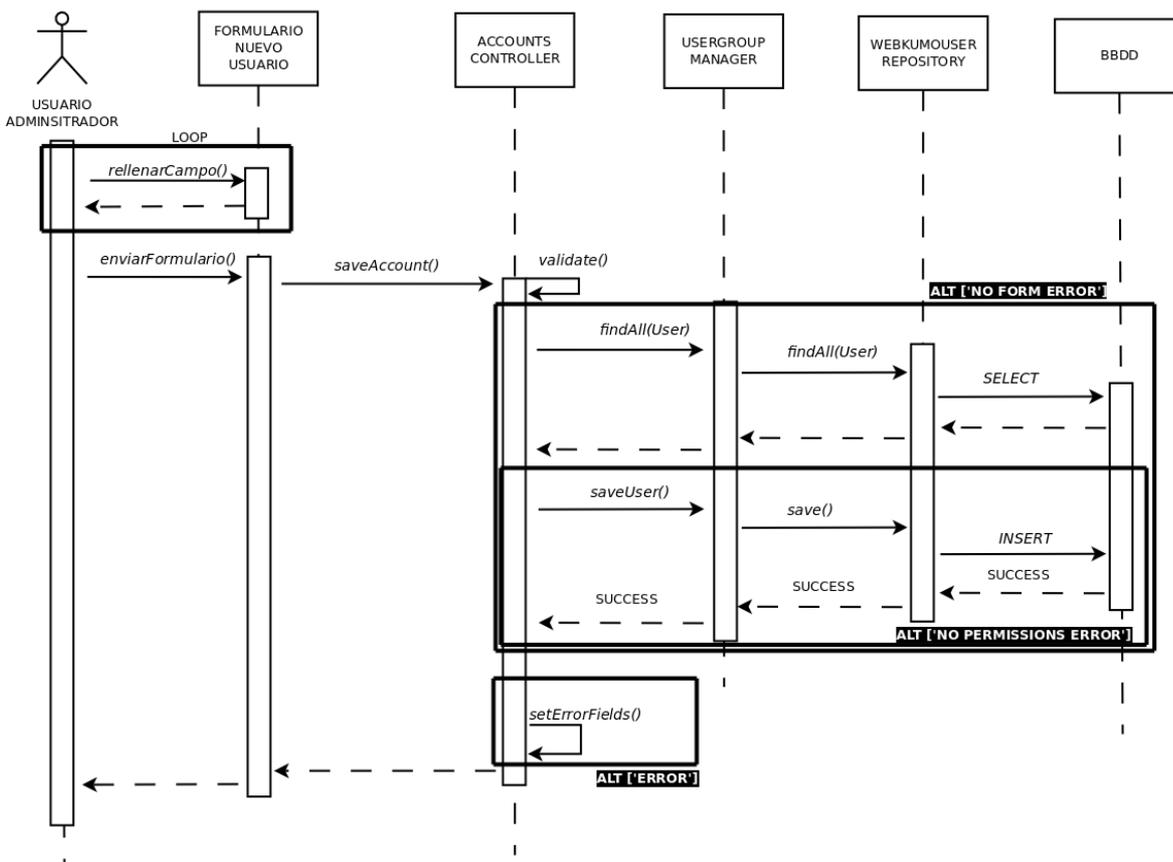


Fig. 44: Diagrama de interacción para dar de alta un elemento (usuario).

Para el caso de uso de alta se tiene una interacción de la vista con el controlador correspondiente, posteriormente este controlador se comunicará con la capa de negocio a través de la clase *manager* adecuada. Seguido, será la capa de negocio la que se comunicará con el repositorio de la entidad involucrada en el proceso de alta que se comunicará con la base de datos para realizar la persistencia.

Como elementos intermedios, se realiza la validación del formulario de entrada así como las comprobaciones pertinentes de permisos. En el diagrama puede comprobarse que sólo se lleva a cabo la operación si el usuario que ha iniciado sesión tiene permisos para realizar la operación y si el formulario tampoco tiene errores de validación; esto se representa mediante las alternativas en el diagrama. La alternativa que se ejecuta si no hay errores se ejecuta en el ejemplo tras realizar `findAll()`, el usuario que se busca en la base de datos posee permisos suficientes y además comprueba que no exista ya un usuario con el nombre que se quiere dar de alta. La alternativa

Este diagrama de interacción es aplicable también al **alta de un grupo**, a la **modificación de un usuario**, a la **modificación de un grupo**, al **alta de un trabajo** y a la **modificación de un trabajo**.

## 6.5.2 Caso de uso 'Eliminar usuario'

En este caso, el proceso de comunicación entre capas es similar al del caso anterior. En este caso, el proceso intermedio incluye la carga de una colección de entidades (los trabajos del usuario) para comprobar si se cumple alguna condición concreta que permita o descarte la operación. En la situación particular del caso de uso de ejemplo que da de baja un usuario, la condición a comprobar es que el usuario no haya creado trabajos.

Si no se cumple la condición requerida para efectuar la baja, se mostrará un mensaje de error, lo cuál se corresponde con la alternativa de error mostrada en el diagrama.

Se puede aplicar este diagrama de interacción al resto de procesos de baja contenidos en la aplicación, estos son la baja de un grupo y la baja de un trabajo. Variarían las entidades sobre las que se realizan las operaciones y las restricciones concretas a comprobar.

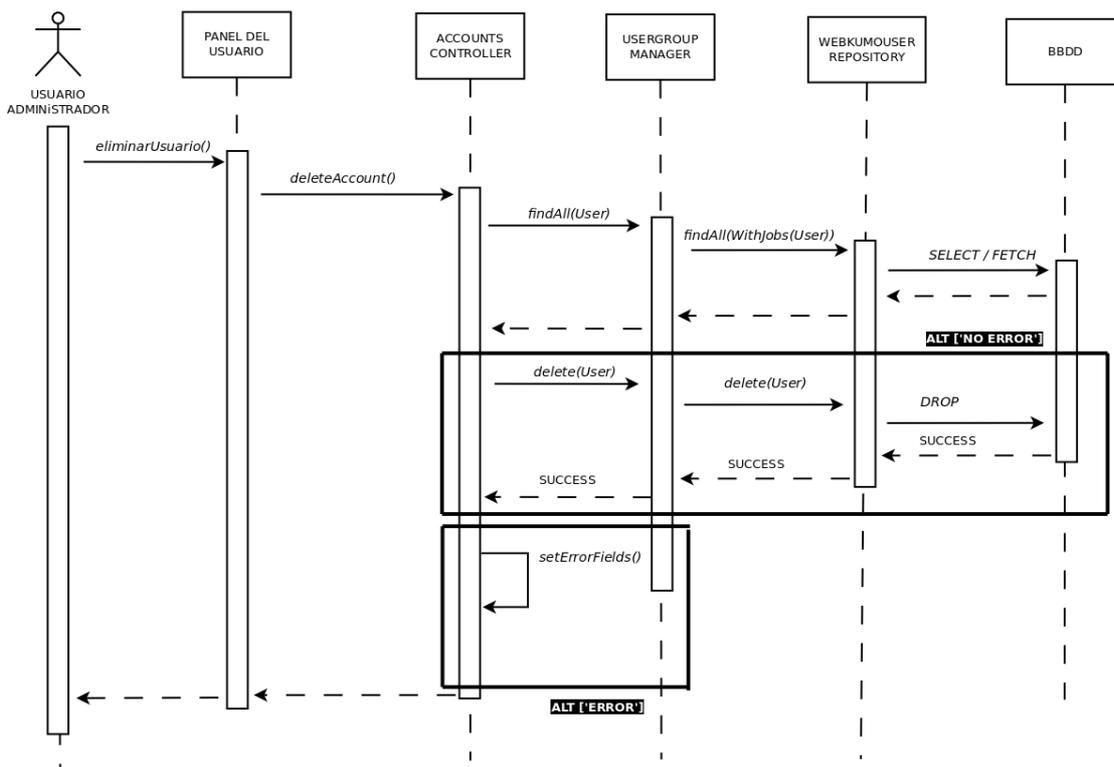


Fig. 45: Diagrama de interacción para la baja de un elemento (usuario).

### 6.5.3 Caso de uso 'Listar usuarios'

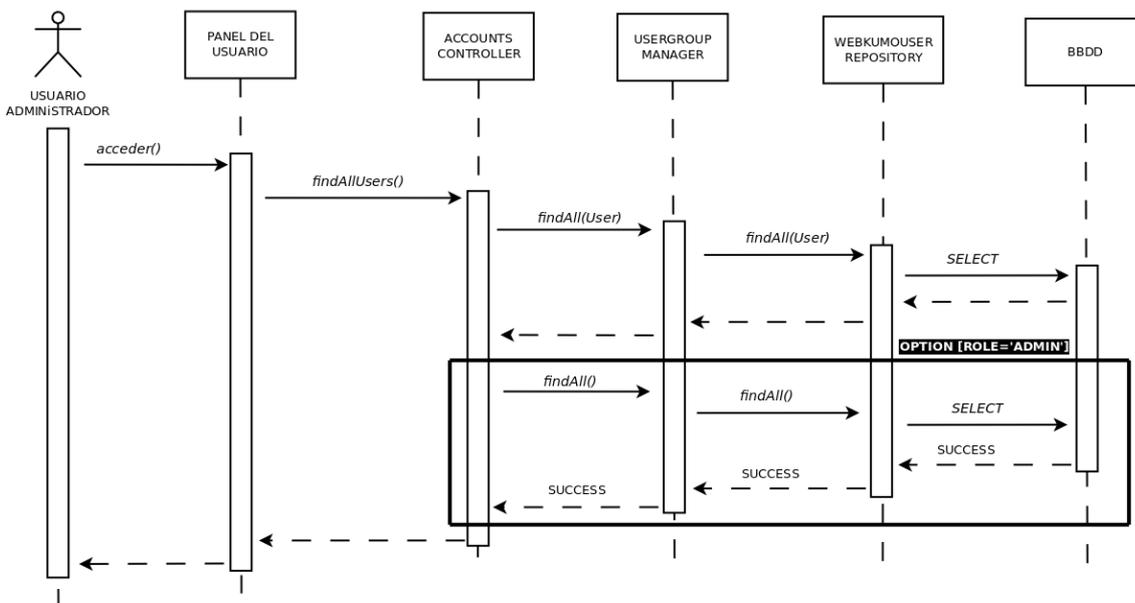


Fig. 46: Diagrama de interacción para listar múltiples elementos.

Este caso cubre la situación en la que se listan elementos de manera dinámica (como en el caso de AJAX **[AJAX]**) en la que no se toman acciones si existe una condición de error. Por ejemplo en el diagrama anterior se ejemplifica el listado de usuarios en el cuál si el usuario no es administrador, la petición no añadirá ningún elemento a la respuesta siendo esta vacía.

Se puede aplicar este diagrama de interacción al caso de uso de **vista general**, al de **listado de grupos**, **listado de trabajos**, al de **arrancar y detener servicio** (que para facilitar la fluidez de la interacción con el usuario se realizará de forma similar).

### 6.5.4 Caso de uso 'Ejecutar trabajo'

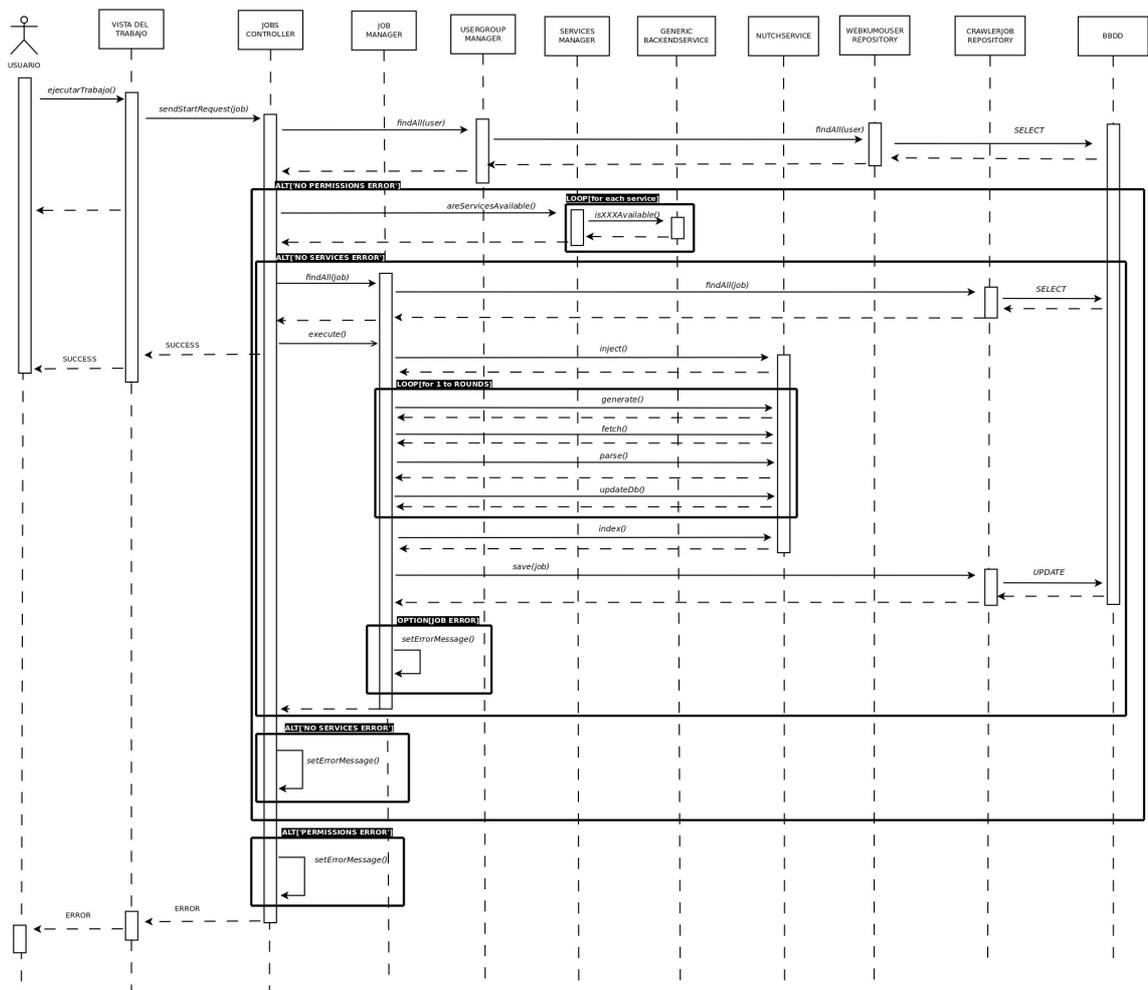


Fig. 47: Diagrama de interacción del caso de uso de ejecución de un trabajo.

El diagrama de interacción muestra la interacción entre sistemas al solicitar la ejecución de un trabajo. Se puede apreciar que el controlador de trabajos primeramente comprueba que el usuario que solicita la ejecución tenga permisos para realizar tal acción. A continuación se realiza una comprobación imprescindible, que todos los servicios de infraestructura que sostienen el análisis y descarga, estén disponibles para ejecutar el trabajo.

Tras las comprobaciones pertinentes, el trabajo se ejecutará de forma asíncrona (llamada a `execute()`) mediante el servicio de Apache Nutch que ejecutará el ciclo de ejecución de dicha araña web (se ha omitido la interacción con los sistemas concretos por simplicidad y porque es algo que no realiza WebKumo sino que es efectuado por Apache Nutch).

Tras la ejecución, se guardará el resultado de la ejecución y la interacción habrá finalizado. Como alternativas, si se produjeron errores durante las comprobaciones, se establecerá el mensaje de error correspondiente y se retornará.

### 6.5.5 Caso de uso 'Iniciar sesión'

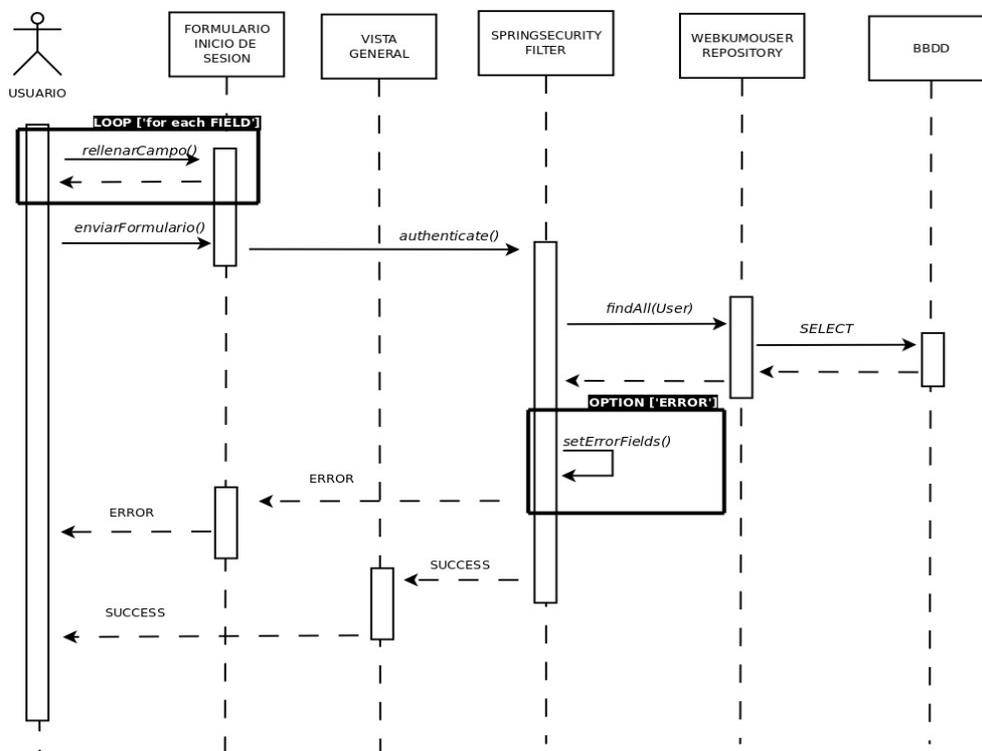


Fig. 48: Diagrama de interacción para el caso de uso de inicio de sesión.

## 6.6 Diseño de la base de datos

### 6.6.1 Descripción del SGBD usado

Para dar soporte a la persistencia de la aplicación web WebKumo (no confundir con la persistencia de la araña web que es soportada por Apache Cassandra) es necesario el uso de un SGBD **[SGBD]** que almacene la información relativa a los usuarios, los trabajos, los grupos etc.

En este caso se ha optado por utilizar **MySQL** (en su última versión disponible en el repositorio software de la máquina donde se realizará el despliegue es la 5.5.41). Este software de gestión de bases de datos es de tipo relacional **[BDREL1]** y está disponible bajo licencia de software libre GPL **[GPL]** (aunque también existen versiones propietarias).

Al no existir requisitos específicos que limiten el rango de elección del SGBD, se optó por MySQL por ser el software mas popular **[BDREL2]** en este ámbito que se ofrece de manera gratuita. Esto además asegura una gran cantidad de documentación y soporte por si fuera necesario. Entre las características de MySQL mas destacables, se encuentran:

- ◆ Garantía de cumplimiento de propiedades ACID **[ACID]** (mediante motor *InnoDB* que es el que se utiliza por defecto en versiones superiores a la 5.5).
- ◆ Soporte multiplataforma.
- ◆ Cacheo de consultas **[QUCAC]**.
- ◆ Replicación entre nodos para mejorar la escalabilidad.
- ◆ Soporte para subconsultas y procedimientos almacenados.
- ◆ Soporte para disparadores y cursores.

## 6.6.2 Integración del SGBD

Se integrará la aplicación WebKumo con el SGBD mediante el software de mapeo objeto-relacional (ORM **[ORM]** **Hibernate**). Este software gestiona automáticamente la persistencia de las entidades según la especificación JPA **[JPA]** (Java Persistence API) en la cuál se configuran las características de las entidades (mediante configuración XML o mediante anotaciones Java) de tal modo que se proporciona una API que permite gestionar la persistencia de las entidades sin conocer detalles sobre la implementación de la propia API ni del SGBD subyacente.

Además para integrar dicho software con la aplicación WebKumo que utiliza el framework *Spring*, se ha utilizado otra capa adicional llamada **Spring Data** que facilita la persistencia de las entidades configuradas proveyendo métodos predefinidos para dicho fin mediante repositorios. Para aportar además, mayor granularidad sobre lo que se consulta a la base de datos, Spring Data se maneja mediante el uso de **especificaciones [SPECI]**, que no son otra cosa que la interpretación de *Spring* de la API *Criteria [CRITE]* de JPA. De este modo, las consultas se definen programáticamente mediante la aplicación de filtros y mediante la especificación de colecciones a cargar proporcionando un alto grado de control sobre las mismas con las ventajas que incluye la definición de dichas consultas vía lenguaje Java (comprobación de tipos, expresividad etc.).

### 6.6.3 Diagrama Entidad-Relación

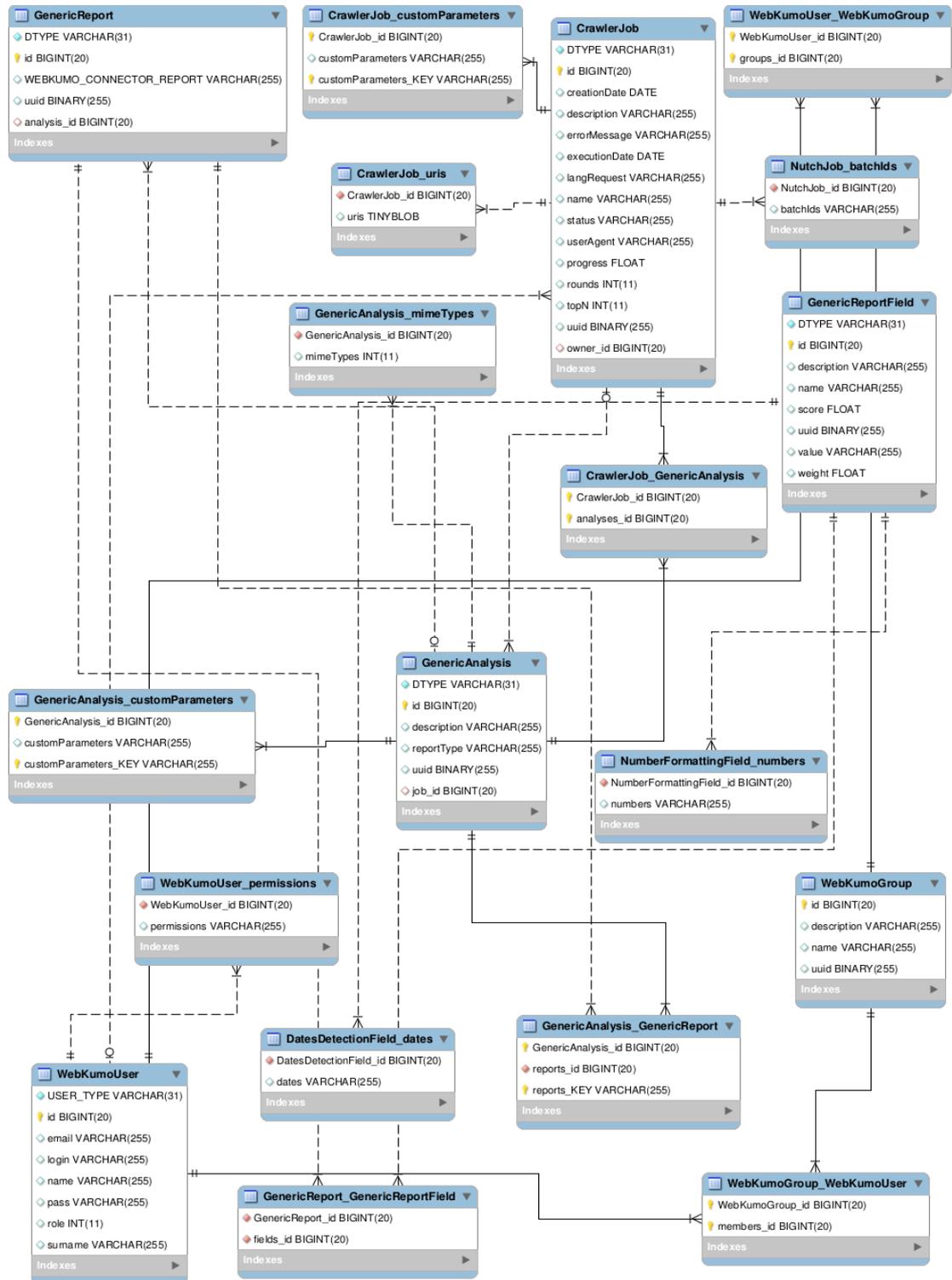


Fig. 49: Diagrama Entidad-Relación del modelo de la base de datos.

## 6.7 Diseño de la interfaz

En esta sección se mostrará la interfaz final del usuario mediante capturas de pantalla. Asimismo se indicarán los elementos comunes de interfaz utilizados para facilitar el acceso a la funcionalidad procurando tener un interfaz navegable y homogénea.

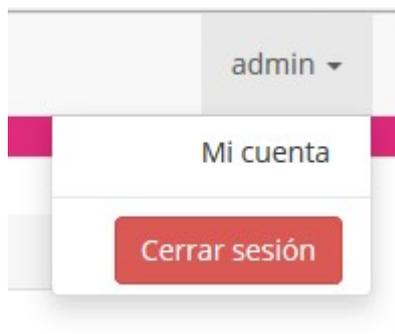
### 6.7.1 Aspectos comunes

Uno de los elementos comunes mas importantes que comparten todas las pantallas de la interfaz, es la barra superior de navegación. Tiene el aspecto siguiente:



**Fig. 50: Barra de navegación superior**

Esta barra de navegación está presente en todas las páginas de la interfaz y permite de forma rápida acceder a las distintas secciones de la aplicación. Cabe destacar del mencionado menú de navegación, la posibilidad de hacer click en el icono de WebKumo para retornar en cualquier momento a la página de vista de situación general (página principal de la aplicación). Además se dispone en la parte derecha de un botón desplegable que permite acceder al panel de usuarios (cuyo contenido varía en función del rol del usuario que haya iniciado sesión) así como finalizar la sesión actual:



**Fig. 51: Botón desplegable del menú de navegación.**

Otro elemento que se ha uniformizado a lo largo de la aplicación, es la ubicación de los botones que procesan una solicitud (no enlaces). Se establecen todos en la parte superior derecha de la sección de la página. Se ve un ejemplo a continuación:



**Fig. 52: Botón de accionamiento en la parte derecha del título de la sección.**

Todas las páginas de la aplicación web poseen en común, el pie de página con los créditos de la aplicación:

WebKumo - Herramienta web de análisis de portales basada en Apache Nutch  
2015 - Trabajo Fin de Máster (Máster Universitario en Ingeniería Web)  
Autor: Sergio Rozas García - Director: Daniel Fernández Lanvin



**Fig. 53: Pie de página.**

También cabe mencionar el uso de tablas en diversos lugares de la aplicación que poseen una funcionalidad y estilo similares:

**Lista de trabajos**

Mostrar  registros Buscar:

Identificador	Estado	Nombre del trabajo	Usuario propietario	Fecha de creación	Fecha de ejecución	Eliminar trabajo
#1	FINISHED	Sample job 10	user23	25/6/2015 2:00:00	---	<a href="#">✕</a>
#10	STOPPED	Sample job 24	user13	25/6/2015 2:00:00	---	<a href="#">✕</a>
#11	FINISHED	Sample job 15	user12	25/6/2015 2:00:00	---	<a href="#">✕</a>
#12	STOPPED	Sample job 25	user21	25/6/2015 2:00:00	---	<a href="#">✕</a>
#13	FINISHED	Sample job 16	user11	25/6/2015 2:00:00	---	<a href="#">✕</a>
#14	STOPPED	Sample job 26	user24	25/6/2015 2:00:00	---	<a href="#">✕</a>
#15	FINISHED	Sample job 17	user13	25/6/2015 2:00:00	---	<a href="#">✕</a>
#16	STOPPED	Sample job 27	user21	25/6/2015 2:00:00	---	<a href="#">✕</a>
#17	FINISHED	Sample job 18	user22	25/6/2015 2:00:00	---	<a href="#">✕</a>
#18	STOPPED	Sample job 28	user10	25/6/2015 2:00:00	---	<a href="#">✕</a>

Mostrando registros del 1 al 10 de un total de 50 registros Anterior [1](#) [2](#) [3](#) [4](#) [5](#) [Siguiente](#)

**Fig. 54: Tabla utilizada en la interfaz**

Las tablas utilizadas incluyen la disposición de la información mediante páginas seleccionables, la posibilidad de ordenar las columnas de forma ascendente o descendente, la búsqueda en toda la tabla y el establecimiento del número de registros a mostrar por página.

Otro aspecto que se puede encontrar en ciertas páginas son las denominadas “migas de pan” para facilitar la navegación en procesos que requieran varios pasos:



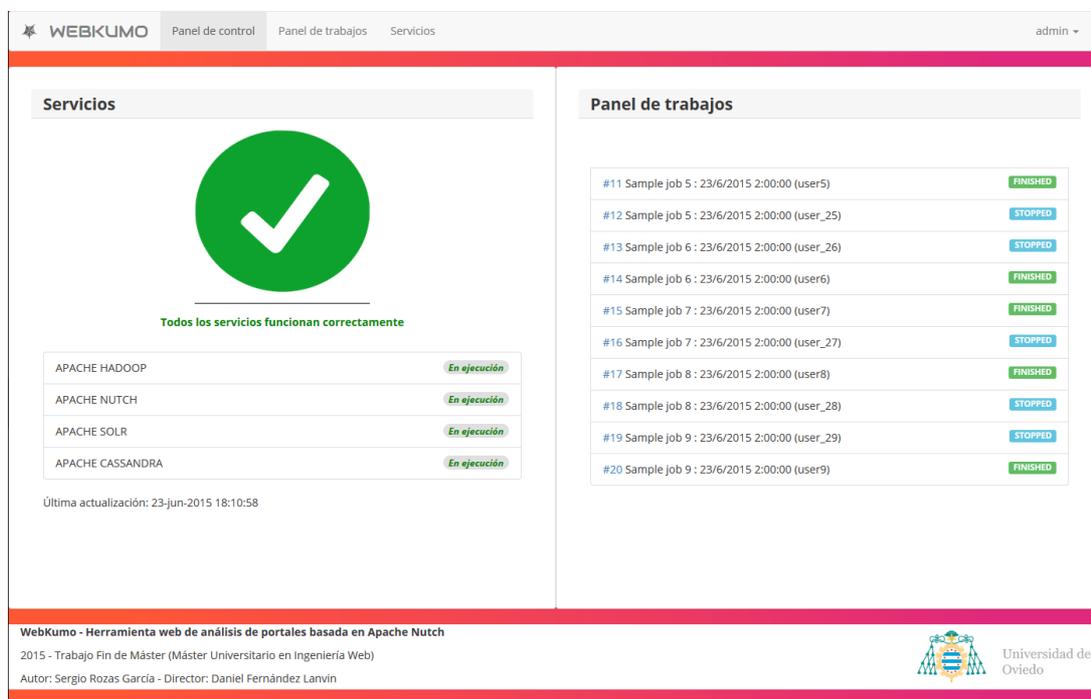
**Fig. 55: Migas de pan en la interfaz**

## 6.7.2 Pantalla de inicio de sesión



**Fig. 56: Pantalla de inicio de sesión.**

## 6.7.3 Pantalla de vista de situación general



**Fig. 57: Pantalla de vista general**

## 6.7.4 Pantalla del panel de trabajos

**Panel de trabajos** admin ▾

[Crear un trabajo de análisis](#)

**Lista de trabajos**

Mostrar  registros Buscar:

Identificador	Estado	Nombre del trabajo	Usuario propietario	Fecha de creación	Fecha de ejecución	Eliminar trabajo
#1	FINISHED	Sample job 10	user23	25/6/2015 2:00:00	---	✕
#10	STOPPED	Sample job 24	user13	25/6/2015 2:00:00	---	✕
#11	FINISHED	Sample job 15	user12	25/6/2015 2:00:00	---	✕
#12	STOPPED	Sample job 25	user21	25/6/2015 2:00:00	---	✕
#13	FINISHED	Sample job 16	user11	25/6/2015 2:00:00	---	✕
#14	STOPPED	Sample job 26	user24	25/6/2015 2:00:00	---	✕
#15	FINISHED	Sample job 17	user13	25/6/2015 2:00:00	---	✕
#16	STOPPED	Sample job 27	user21	25/6/2015 2:00:00	---	✕
#17	FINISHED	Sample job 18	user22	25/6/2015 2:00:00	---	✕
#18	STOPPED	Sample job 28	user10	25/6/2015 2:00:00	---	✕

Mostrando registros del 1 al 10 de un total de 50 registros Anterior **1** 2 3 4 5 Siguiente

**WebKumo - Herramienta web de análisis de portales basada en Apache Nutch**  
 2015 - Trabajo Fin de Máster (Máster Universitario en Ingeniería Web)  
 Autor: Sergio Rozas García - Director; Daniel Fernández Lanvin

 Universidad de Oviedo

**Fig. 58: Pantalla del panel de trabajos**

## 6.7.5 Pantalla del información de trabajos

**Trabajo # 1**

**Nombre del trabajo:** Sample Job 10

**Descripción del trabajo:** This is a sample job

**Propietario del trabajo:** user23 (John Doe)

**Fecha de creación:** 25-jun-2015

**Iteraciones:** 15

**Recursos por nivel:** 8

**Agente de usuario:** Sin establecer. Se utilizará el valor presente en nutch-site.xml

**Cabecera Accept-Language HTTP:** Sin establecer. Se utilizará el valor presente en nutch-site.xml

**Fecha de ejecución:** No ejecutado todavía

**Estado:** Finalizado

**URLs asociadas al trabajo**

- http://ingenieriainformatica.uniovi.es/
- http://epijijon.uniovi.es/

**Análisis a realizar**

- MockAnalysis (Análisis de prueba que extrae las URL y los nombres de los recursos.)

**Parámetros adicionales del trabajo:**

- TEST\_CUSTOM\_PARAMETER : VALUE

**Fig. 59: Pantalla de información de un trabajo**

**Results**

- MockAnalysis
  - Análisis de prueba que extrae las URL y los nombres de los recursos.
  - epigijon.uniovi.es
    - Porcentaje global de cumplimiento: 0%

```

1. http://www.0epigijon.uniovi.es (S:100 | W: 1)
2. http://www.19epigijon.uniovi.es (S:100 | W: 1)
3. http://www.5epigijon.uniovi.es (S:100 | W: 1)
4. http://www.3epigijon.uniovi.es (S:100 | W: 1)
5. http://www.4epigijon.uniovi.es (S:100 | W: 1)
6. http://www.17epigijon.uniovi.es (S:100 | W: 1)
7. http://www.24epigijon.uniovi.es (S:100 | W: 1)
8. http://www.2epigijon.uniovi.es (S:100 | W: 1)
9. http://www.27epigijon.uniovi.es (S:100 | W: 1)
10. http://www.10epigijon.uniovi.es (S:100 | W: 1)
11. http://www.15epigijon.uniovi.es (S:100 | W: 1)
12. http://www.23epigijon.uniovi.es (S:100 | W: 1)
13. http://www.11epigijon.uniovi.es (S:100 | W: 1)
14. http://www.14epigijon.uniovi.es (S:100 | W: 1)
15. http://www.34epigijon.uniovi.es (S:100 | W: 1)
16. http://www.25epigijon.uniovi.es (S:100 | W: 1)
17. http://www.1epigijon.uniovi.es (S:100 | W: 1)
18. http://www.22epigijon.uniovi.es (S:100 | W: 1)
19. http://www.6epigijon.uniovi.es (S:100 | W: 1)
20. http://www.28epigijon.uniovi.es (S:100 | W: 1)
21. http://www.36epigijon.uniovi.es (S:100 | W: 1)
22. http://www.13epigijon.uniovi.es (S:100 | W: 1)
23. http://www.18epigijon.uniovi.es (S:100 | W: 1)
24. http://www.33epigijon.uniovi.es (S:100 | W: 1)
25. http://www.16epigijon.uniovi.es (S:100 | W: 1)
26. http://www.30epigijon.uniovi.es (S:100 | W: 1)
27. http://www.29epigijon.uniovi.es (S:100 | W: 1)
            
```

**Fig. 60: Resultados de un análisis de ejemplo (pantalla de información de un trabajo)**

## 6.7.6 Pantalla de creación/modificación de un trabajo

**WEBKUMO** Panel de control Panel de trabajos Servicios admin

Panel de trabajos / Crear un trabajo de análisis

**Crear un trabajo de análisis** Enviar

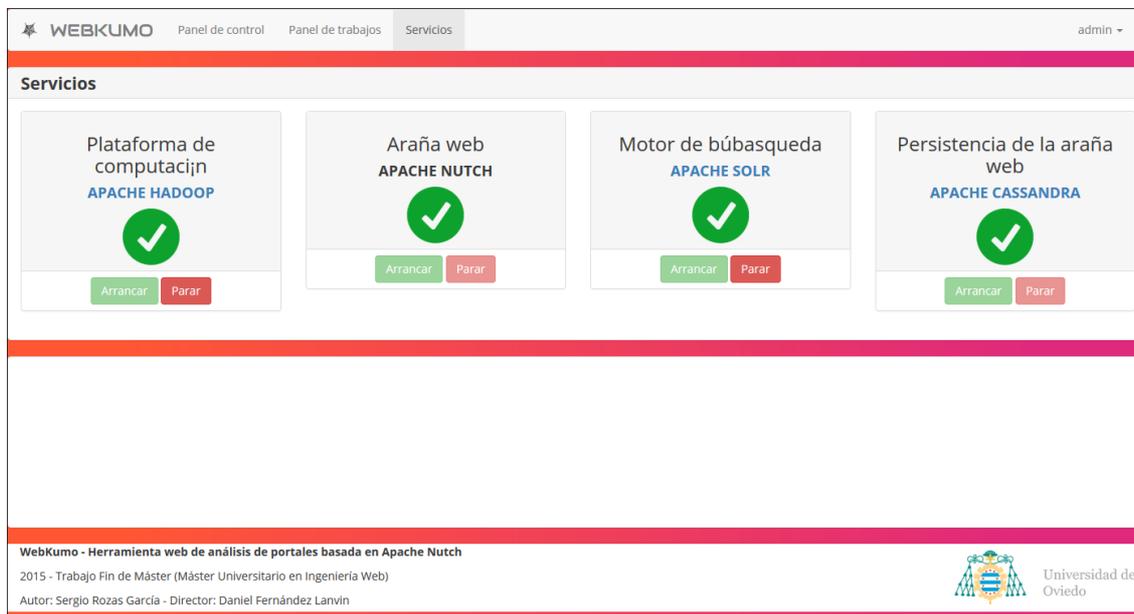
Nombre	Job 1				
Descripción	...				
Recursos por nivel	1				
Iteraciones	1				
Agente de usuario	Webkumo-0.1				
Cabecera Accept-Language HTTP	es-ES, en-us, en-gb, en;q=0.7,*;q=0.3				
Seeds	http://www.example.org				
Fichero de URLs semilla	<input type="button" value="Examinar..."/> No se ha seleccionado ningún archivo.				
Análisis a realizar	Ninguno				
Custom Job parameters	<table border="1"> <thead> <tr> <th>Clave del parámetro</th> <th>Valor del parámetro</th> </tr> </thead> <tbody> <tr> <td> </td> <td> </td> </tr> </tbody> </table>	Clave del parámetro	Valor del parámetro		
Clave del parámetro	Valor del parámetro				

WebKumo - Herramienta web de análisis de portales basada en Apache Nutch  
 2015 - Trabajo Fin de Máster (Máster Universitario en Ingeniería Web)  
 Autor: Sergio Rozas García - Director: Daniel Fernández Larvín

Universidad de Oviedo

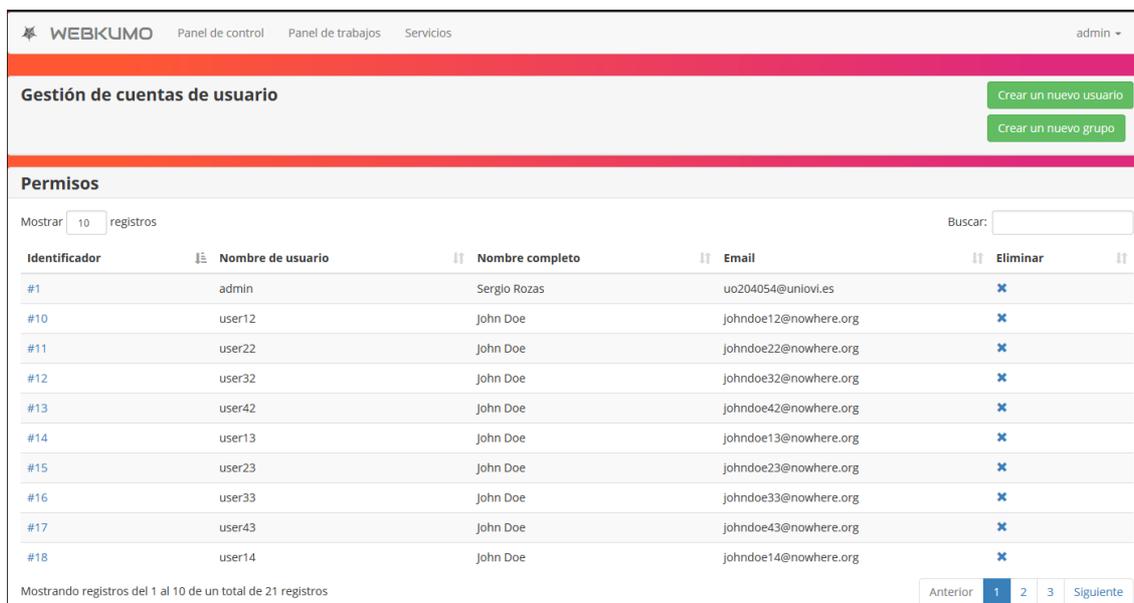
**Fig. 61: Pantalla de creación de un trabajo**

### 6.7.7 Pantalla del panel de servicios



**Fig. 62: Pantalla del panel de servicios**

### 6.7.8 Pantalla del panel del usuario



**Fig. 63: Pantalla del panel del usuario administrador con la tabla de usuarios**

**Grupos**

Mostrar  registros Buscar:

Id	Nombre del grupo	Descripción del grupo	Eliminar
#1	group1	Example Group 1	<a href="#">✕</a>
#10	group1	Example Group 1	<a href="#">✕</a>
#11	group2	Example Group 2	<a href="#">✕</a>
#12	group3	Example Group 3	<a href="#">✕</a>
#13	group1	Example Group 1	<a href="#">✕</a>
#14	group2	Example Group 2	<a href="#">✕</a>
#15	group3	Example Group 3	<a href="#">✕</a>
#2	group2	Example Group 2	<a href="#">✕</a>
#3	group3	Example Group 3	<a href="#">✕</a>
#4	group1	Example Group 1	<a href="#">✕</a>

Mostrando registros del 1 al 10 de un total de 15 registros Anterior **1** 2 Siguiente

---

**WebKumo - Herramienta web de análisis de portales basada en Apache Nutch**  
 2015 - Trabajo Fin de Máster (Máster Universitario en Ingeniería Web)  
 Autor: Sergio Rozas García - Director: Daniel Fernández Lanvín



Universidad de Oviedo

**Fig. 64: Pantalla del panel del usuario administrador con la tabla de grupos**

**WEBKUMO** Panel de control Panel de trabajos user10 ▾

**Gestión de cuentas de usuario** [Edit your account info](#)

**John Doe (johndoe10@nowhere.org)**

Permisos

- Pues ver tus propios trabajos
- Puedes crear nuevos trabajos

Groups

#1 group1: Example Group 1

---

**WebKumo - Herramienta web de análisis de portales basada en Apache Nutch**  
 2015 - Trabajo Fin de Máster (Máster Universitario en Ingeniería Web)  
 Autor: Sergio Rozas García - Director: Daniel Fernández Lanvín



Universidad de Oviedo

**Fig. 65: Pantalla del panel del usuario regular**

## 6.7.9 Pantalla de creación/modificación de un usuario

The screenshot shows the 'Editar información de usuario' form for an administrator. The form includes the following fields and options:

- Nombre de usuario: user1
- Nombre: John
- Apellidos: Doe
- Email: somebody@somedomain.org
- Contraseña and Repita contraseña: (empty)
- Grupos: Ninguno
- Rol:  Administrador,  Usuario regular
- Permisos:  Pues ver tus propios trabajos,  Pues crear nuevos trabajos,  Pues ver trabajos de otros usuarios que estén en los mismos grupos que tú.

The footer contains the text: 'WebKumo - Herramienta web de análisis de portales basada en Apache Nutch', '2015 - Trabajo Fin de Máster (Máster Universitario en Ingeniería Web)', 'Autor: Sergio Rozas García - Director: Daniel Fernández Lanvin', and the Universidad de Oviedo logo.

**Fig. 66: Pantalla de creación y/o edición de un usuario para administradores.**

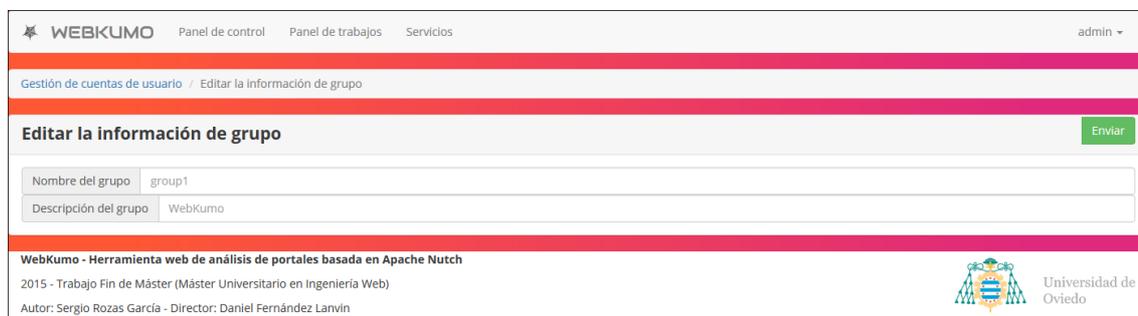
The screenshot shows the 'Editar información de usuario' form for a regular user. The form includes the following fields and options:

- Nombre: John
- Apellidos: Doe
- Email: johndoe10@nowhere.org
- Contraseña and Repita contraseña: (empty)

The footer contains the text: 'WebKumo - Herramienta web de análisis de portales basada en Apache Nutch', '2015 - Trabajo Fin de Máster (Máster Universitario en Ingeniería Web)', 'Autor: Sergio Rozas García - Director: Daniel Fernández Lanvin', and the Universidad de Oviedo logo.

**Fig. 67: Pantalla de edición de un usuario regular.**

## 6.7.10 Pantalla de creación/edición de grupo



The screenshot shows the WEBKUMO web interface. At the top, there is a navigation bar with the logo 'WEBKUMO' and links for 'Panel de control', 'Panel de trabajos', and 'Servicios'. The user is logged in as 'admin'. Below the navigation bar, there is a breadcrumb trail: 'Gestión de cuentas de usuario / Editar la información de grupo'. The main content area is titled 'Editar la información de grupo' and contains two input fields: 'Nombre del grupo' with the value 'group1' and 'Descripción del grupo' with the value 'Webkumo'. A green 'Enviar' button is located to the right of the input fields. At the bottom of the page, there is a footer with the text: 'WebKumo - Herramienta web de análisis de portales basada en Apache Nutch', '2015 - Trabajo Fin de Máster (Máster Universitario en Ingeniería Web)', and 'Autor: Sergio Rozas García - Director: Daniel Fernández Larvín'. The logo of the Universidad de Oviedo is also present in the footer.

**Fig. 68: Pantalla de creación y/o edición de un grupo.**

## 6.8 Especificación técnica del plan de pruebas

Esta sección contiene la descripción sobre cómo se han realizado las pruebas a lo largo de todo el ciclo de desarrollo del proyecto. Se especificarán los distintos tipos de pruebas indicando los aspectos técnicos mas relevantes.

### 6.8.1 Pruebas unitarias

Cuando se trata de pruebas unitarias, se prueban módulos concretos en solitario para comprobar que la tarea que tienen asignada se cumple con éxito. Esto se traduce en probar lo que suelen ser pequeños fragmentos de código (métodos o interfaces en función de su tamaño) concretos con el fin de asegurar que los caminos de ejecución mas triviales son correctos.

Aplicando esto a este proyecto, se puede comprobar que debido a la gran dependencia que tiene los componentes del sistema de la plataforma de la infraestructura de procesamiento (mas allá de probar funciones triviales las cuáles se ignoran por simplicidad, por ej: *getters* y *setters*), la frontera entre pruebas unitarias y del sistema queda muy difuminada.

Por ejemplo, no es posible la prueba de las fases de ejecución de Apache Nutch implementadas sin ejecutar un trabajo de forma real, salvo que se simulara la ejecución de Apache Nutch de algún modo.

Es por ello que se ha decidió limitar las pruebas de carácter unitario a las entidades de usuarios (*WebKumoUser*), grupos (*WebKumoGroup*) y a los algoritmos de análisis implementados (clases que extiendan la funcionalidad de *GenericAnalysis*). El resto de componentes (servicios y trabajos principalmente) se probarán mediante la comprobación de las funcionalidades del sistema.

### 6.8.2 Pruebas de sistema

Las pruebas de funcionalidad realizadas durante el desarrollo de este proyecto consistieron en probar casos de uso propios de la aplicación de tal modo que se verifique que el comportamiento es el esperado.

La plataforma de pruebas automatizadas escogida para realizar las pruebas, es **TestNG [TESNG]**, un framework de pruebas unitarias para Java siendo este uno de los más habituales dentro del mundo del desarrollo Java. Además TestNG se integra con facilidad con el framework *Spring* de desarrollo y su contexto de ejecución mediante la extensión de la clase `AbstractTestNGSpringContextTests` por parte de las clases de prueba que contengan los tests. Se proporciona así una forma sencilla de acceder a la mayoría de las características de Spring (y sus *sub-frameworks*) como son la inyección de dependencias, el acceso a servicios y uso de repositorios entre otros.

Las pruebas se ejecutaron mediante el uso del *goal* "test" de Maven **[MVN]**, que es el software de gestión de dependencias y de construcción del proyecto que se ha utilizado.

Como comentario adicional, se han dividido las pruebas por funcionalidades, teniendo así conjuntos de test para los controladores, para los usuarios del modelo, para los trabajos, para los análisis y para los servicios.

### 6.8.3 Pruebas de usabilidad y accesibilidad

Se debe procurar, para cualquier aplicación, que la interfaz de usuario sea usable; esto es, que la interfaz de usuario facilite lo más posible la experiencia al usuario evitando introducir limitaciones por la propia interfaz. Para lo anterior se deben de seguir una serie de criterios de diseño de dicha interfaz que serán puestos a prueba mediante estas comprobaciones.

Para la realización de pruebas de usabilidad se deben definir primero sus condiciones, que vienen dadas por los siguientes puntos.

1. **Usuarios:** Los usuarios que participaron en las pruebas son 2 de alto perfil técnico, acordes con el usuario objetivo de este proyecto.
2. **Lugar de realización:** Se realizaron las pruebas en el domicilio del proyectante.
3. **Metodología:** La metodología ejecutada es mediante el uso de formularios que tanto el evaluador (el propio proyectante) como los usuarios rellenarán para evaluar la experiencia. En todo momento los usuarios tendrán disponible el manual del usuario para su consulta.

### 6.8.3.1 Pruebas de usabilidad

#### 6.8.3.1.1 Cuestiones de carácter general

Se debe tener en cuenta que este producto software no está destinado al público general, sino que será utilizado por personas con alto grado de conocimiento técnico por lo que las preguntas de carácter general están enfocadas a tal público.

¿Con qué frecuencia usa un computador?	<ol style="list-style-type: none"> <li>1. Todos los días</li> <li>2. Varias veces a la semana</li> <li>3. Ocasionalmente</li> <li>4. Nunca o casi nunca</li> </ol>
¿Qué tipo de actividades realiza mediante un computador?	<ol style="list-style-type: none"> <li>1. Parte de mi trabajo</li> <li>2. Actividades de ocio</li> <li>3. Manejo básico de algún tipo de software</li> <li>4. Uso ocasional (correo electrónico o similar)</li> </ol>
¿Qué grado de conocimiento posee sobre Apache Hadoop o sobre otros sistemas de computación distribuida?	<ol style="list-style-type: none"> <li>1. Alto</li> <li>2. Medio</li> <li>3. Bajo</li> <li>4. Ninguno</li> </ol>
¿Con qué tipo de interfaz se siente más cómodo a la hora de trabajar?	<ol style="list-style-type: none"> <li>1. Interfaz web</li> <li>2. Interfaz de escritorio</li> <li>3. Interfaz de comandos.</li> <li>4. Otra (especificar cuál).</li> </ol>
Enumere tres ramas o campos de la informática que le susciten interés	<hr/> <hr/> <hr/>

#### 6.8.3.1.2 Actividades guiadas

Las actividades que los usuarios deberán realizar a modo de evaluación serán una mezcla de actividades más genéricas que puedan encontrarse en una aplicación web cualquiera con alguna más específica de el proyecto que esta documentación trata. De este modo se pretenden cubrir un espectro amplio sin dejar de lado los aspectos comunes a cualquier interfaz de gestión.

1. **Crear un usuario nuevo.**
2. **Eliminar un grupo**
3. **Modificar un trabajo.**
4. **Arrancar un servicio**
5. **Ejecutar un trabajo y comprobar su resultado.**

### 6.8.3.1.3 Preguntas cortas y observaciones

<b>FACILIDAD DE USO</b>	<b>Siempre</b>	<b>Casi siempre</b>	<b>A veces</b>	<b>Nunca</b>
¿Sabe en qué parte de la aplicación se encuentra?				
Si está en un proceso de varios pasos, ¿sabe con facilidad en cuál está?				
¿Sabe, ante un formulario, lo que significa cada campo con precisión?				
Ante una operación que requiere una espera, ¿sabe que dicha operación se está llevando a cabo?				
Cuando se produjo un error, ¿supo con facilidad que dicho error ocurrió?				
¿Le resultó difícil volver a una pantalla anterior en algún momento?				

<b>FACILIDAD DE USO</b>	<b>Siempre</b>	<b>Casi siempre</b>	<b>A veces</b>	<b>Nunca</b>
¿Fueron útiles los mensajes de error que aparecieron al introducir algún valor incorrecto en un formulario?				
¿Le fue difícil determinar cuál había sido el resultado de un trabajo?				

¿Le fue difícil conocer cuál era el estado de un trabajo?				
¿La forma de indicar el progreso de un trabajo le fue suficientemente explicativa?				

<b>FUNCIONALIDAD</b>	<b>Siempre</b>	<b>Casi siempre</b>	<b>A veces</b>	<b>Nunca</b>
¿Las tareas funcionaron como se esperaba?				
¿El tiempo de respuesta al acceder al resultado de un trabajo tras su ejecución fue excesivo?				
¿Se encontró con alguna pantalla que indicara que hubo un error interno o que el recurso no había sido encontrado?				
¿El trabajo ejecutado mostró algún mensaje de error tras su finalización?				
¿Pudo iniciar sesión con el usuario creado tras haber completado el proceso?				
¿El servicio se mostraba como iniciado tras enviar la orden?				
¿Seguía apareciendo el grupo eliminado en el panel del administrador tras haberlo borrado?				

<b>ASPECTOS GRÁFICOS</b>	<b>Siempre</b>	<b>Casi siempre</b>	<b>A veces</b>	<b>Nunca</b>
¿El tamaño del texto le fue impedimento?				
¿Algún texto mostraba caracteres extraños o ilegibles?				
¿El color del fondo le molestó de alguna manera?				

¿Supo identificar correctamente el botón que realizaba la acción que usted deseaba?				
¿El diseño de las páginas le resultó claro?				
¿Le ayudó a la hora de identificar un error que este se mostrara mediante un panel de color rojo?				

<b>DISEÑO</b>	<b>Sí</b>	<b>No</b>	<b>A veces</b>
¿La interfaz le pareció atractiva?			
¿Encontró fácil de usar la interfaz?			
¿Existió algún elemento gráfico (botones, flechas ...) que no supiera identificar?			
¿Supo identificar el menú de navegación con facilidad?			
¿La colocación de elementos para mostrar los usuarios del sistema le resultó adecuada?			
¿Le aportó algo el uso de iconos para indicar elementos con los que podía interactuar?			
¿Le fue sencillo seleccionar múltiples análisis mediante el selector en forma de lista?			
¿Encontró dificultoso la forma de ver la información en las tablas debido a su tamaño o color?			
¿Utilizó en algún momento el icono de la aplicación para acceder a la página principal?			

Además de estos formularios en forma de tabla, se debe recoger cualquier comentario adicional que el sujeto de pruebas desee expresar.

## 6.8.3.1.4 Cuestionario para el responsable de pruebas

ASPECTO OBSERVADO	Notas
El usuario duda cómo comenzar la tarea	
El usuario se pierde a la hora de seguir una tarea con pasos	
El usuario duda si la tarea ha finalizado o no	
El usuario comienza la tarea de forma rápida sin escrutar toda la página para saber por dónde tiene que comenzar	
El usuario identifica en el panel de servicios el accionamiento de arranque y/o parada con rapidez	
El usuario abandona el panel de servicios antes de que haya terminado el proceso de arranque/parada	
El usuario identifica el resultado de la operación con el servicio como un enlace y lo utiliza para mostrar la información de salida	
El usuario espera a que el trabajo termine (no identifica el lanzamiento del trabajo con el fin de la ejecución del mismo).	
El usuario se percató del fin del trabajo y accede a sus resultados	
El usuario identifica la opción para visualizar los campos del informe del trabajo detallados y lo utiliza	
El usuario pide ayuda para rellenar algún campo de algún formulario	
El usuario consulta el manual para operaciones triviales	
Tiempo que tarda el usuario en realizar cada tarea	
Errores de relevancia cometidos	
Otras observaciones	

### 6.8.3.2 Pruebas de accesibilidad

Para la parte de pruebas de accesibilidad, se realizará la evaluación de la interfaz web siguiendo las pautas WAI-W3C para WCAG 2.0 **[WCAG]** para un nivel de conformidad AA. Adicionalmente se comprobará la validez de los documentos HTML y CSS que componen la aplicación web.

La mayoría de las comprobaciones de accesibilidad se realizarán de manera manual debido a la imposibilidad de ejecutar herramientas de terceros de evaluación automática ya que la aplicación requiere autenticación.

Cabe destacar que debido a la naturaleza de la aplicación web, que será de uso interno para la Universidad, con un grupo muy reducido de usuarios, la accesibilidad no toma un papel tan importante que el que tendría en una aplicación abierta de uso público. Sin embargo se procurará alcanzar un nivel de conformidad WCAG 2.0 AA.

### 6.8.4 Pruebas de rendimiento

Aunque no existen requisitos concretos que impongan restricciones sobre el rendimiento de la aplicación (en parte debido a la complejidad de los sistemas de infraestructura y a la dependencia del despliegue concreto que se efectúe), sí que se identifica un claro escenario dentro de la aplicación web en la que pueden presentarse problemas de rendimiento.

En caso en el que se pueden producir mayores problemas de rendimiento es a la hora de mostrar los resultados del análisis de un trabajo. Esto puede surgir debido a que, ante un trabajo de cierta envergadura, se producirá una considerable cantidad de campos en el informe de resultado. Esto puede suponer un problema por lo que se vigilará mediante una prueba de rendimiento.

# 7 Implementación del sistema

## 7.1 Estándares y normas seguidos

### 7.1.1 JavaDoc

Para la elaboración de los comentarios que se incluyen en el código, se ha utilizado Javadoc **[JDOC]**. Mediante este tipo de comentarios se especifican los parámetros, tipo de retorno y excepciones que elevan los métodos de las clases. Los atributos de las clases también es explicitan mediante comentarios de este tipo. Adjunto al proyecto se aportará la documentación generada automáticamente mediante Javadoc.

### 7.1.2 HTML5

Aunque se genera mediante el sistema de plantillas Freemarker descrito en apartados anteriores, la implementación de las vistas mediante este sistema genera código de marcado de hipertexto HTML **[HTML5]**. HTML es el lenguaje de marcas mas habitual en el diseño de páginas web cuya funcionalidad básica es la de indicar el contenido y la estructura de los documentos. Es un estándar del W3C **[W3C]**.

Permite indicar textos, imágenes, hiperenlaces **[LINK]**, elementos de formulario, listas de elementos, tablas etc. describiendo así el contenido del recurso. La idea subyacente es que los navegadores web interpreten dichas marcas de contenido para proporcionar una representación gráfica del mismo sin que existe una manera única de realizar esta tarea.

Se han utilizado elementos típicos de la última revisión del estándar como son controles de formulario específicos, especificación de la codificación o del tipo de documento entre otros.

### 7.1.3 CSS

Las hojas de estilo en cascada **[CSS]** constituyen un estándar definido por el W3C **[W3C]** que pretenden definir el estilo de la presentación de un recurso web HTML, aunque su aplicación no se limita sólo a estos documentos.

Su utilización se basa en el uso de selectores mediante los cuáles, a los elementos DOM **[DOM]** seleccionados se les aplican ciertas reglas de estilo incluyendo tamaño del texto, colores y posicionamiento entre otros. Al separar estos aspectos de la estructura del documento se hace más fácil conseguir la accesibilidad del documento, se evita la repetición de código y se mejora el control sobre los estilos aplicados.

### 7.1.4 WCAG 2.0

Este estándar (**ISO/IEC 40500:2012**) consiste en una serie de recomendaciones **[WCAG]** del W3C **[W3C]** de cara a mejorar los niveles de accesibilidad de los recursos web.

Se basa en el cumplimiento de una serie de *criterios de éxito* en función de *niveles de conformidad* (A, AA y AAA) para alcanzar un grado mayor o menor de accesibilidad en el sitio. Además de los criterios, también proponen una serie de técnicas para alcanzar el cumplimiento de cada requisito.

Los ámbitos de actuación de WCAG son aquellos que implican cómo se percibe el sitio ante distintos usuarios, la operabilidad de los distintos elementos de interfaz, la comprensibilidad del contenido del sitio y la robustez con la que se ha diseñado la interfaz.

En este proyecto se han seguido las indicaciones pertinentes para alcanzar el nivel de conformidad doble A.

## 7.2 Lenguajes de programación

- ◆ **Java:** Lenguaje de propósito general, orientado a objetos que primariamente se utiliza para el desarrollo de aplicaciones sobre la plataforma de aplicaciones empresariales JEE descrita anteriormente en esta documentación. Se ejecuta sobre una máquina abstracta llamada máquina virtual de Java la cuál está implementada para una gran diversidad de plataformas por lo cuál aporta un alto grado de portabilidad. Para este proyecto se ha utilizado la versión 7 de lenguaje (mediante la implementación libre OpenJDK **[OJDK]**).
- ◆ **HTML 5 + Freemarker:** Para la interfaz de la aplicación se ha utilizado el motor de plantillas Freemarker que combina el lenguaje de etiquetas HTML con lógica para controlar la presentación de las interfaces de usuario. La parte correspondiente a HTML se trata del estándar para la creación de documentos web mediante nubes de etiquetas que definen la estructura del contenido de la interfaz. Se combina esta estructura con etiquetas de Freemarker, un lenguaje para el motor de plantillas homónimo que permite realizar un control básico de la interfaz mediante estructuras de control, bucles, acceso al modelo de datos etc.
- ◆ **JavaScript:** Algunas de las funcionalidades de la aplicación se controlan mediante este lenguaje ejecutado en el navegador del usuario. Se trata de un lenguaje interpretado, orientado a objetos principalmente utilizado para controlar la interacción del usuario con la aplicación web y realizar comunicaciones asíncronas (AJAX **[AJAX]**) con el fin de evitar la recarga completa de las páginas.  
Además, Javascript se ha utilizado principalmente mediante la librería **Jquery [JQUE]** (licencia libre MIT) que facilita la mayoría de operaciones ejecutadas sobre el árbol DOM **[DOM]** de los recursos HTML mediante una sintaxis con selectores. También se ha utilizado esta librería para las llamadas asíncronas entre el navegador y la aplicación web mediante AJAX.
- ◆ **CSS:** Mediante este lenguaje se han definido hojas de estilo en cascada para controlar el aspecto visual de los documentos HTML. Mediante el uso de selectores y reglas se define cómo debe ser la presentación de los elementos de marcado de los recursos HTML. Además estas hojas de estilo se han definido en ficheros separados, separando así la presentación del contenido del marcado.  
Además se ha utilizado un framework CSS para facilitar el desarrollo de la interfaz (ya que CSS suele tener un grado importante de dificultad para desarrollar una presentación estética y a la vez funcional sin perjudicar

usabilidad y/o accesibilidad). El framework utilizado es **Bootstrap [BOOTS]** (licencia libre MIT) que proporciona una serie de utilidades predefinidas mediante plantillas de elementos (para formularios, botones, fuentes etc.) las cuáles pueden integrarse con facilidad para conseguir el diseño visual deseado.

## 7.3 Herramientas utilizadas en el desarrollo

- ◆ **Linux Mint:** Distribución de GNU/Linux **[GNLIN]** que ha sido la plataforma base para el desarrollo y la prueba del sistema. Se optó por utilizar una plataforma Linux como fundamento debido a que el software de infraestructura (Apache Hadoop, Apache Nutch, Apache Cassandra y Apache Solr) están principalmente diseñados pensando en esta plataforma lo cuál facilita las tareas de desarrollo, despliegue y depuración **[LMINT]**.
- ◆ **Netbeans IDE:** Entorno de desarrollo multilenguaje especialmente pensado para Java. Soporta además de Java, HTML, CSS, JavaScript y demás lenguajes mediante un sistema de *plugins* modular. Además permite gestionar a través del mismo distintos SGBD y contenedores de aplicaciones web **[NBEAN]**.
- ◆ **Mozilla Firefox:** Navegador web libre desarrollado por la fundación Mozilla. Es el navegador que principalmente se ha utilizado (en su última versión) en el desarrollo de la aplicación destacando su inspector web que permite escrutar el código HTML, depurar código JavaScript, comprobar peticiones HTTP etc **[MZFIR]**.
- ◆ **Dia:** Software libre de creación de diagramas y esquemas. Soporta el modelado básico de distintos tipos de diagramas UML **[UML]**. También soporta la generación de diagramas de flujo, entidad-relación y otros tipos de esquemas (no sólo de modelado de software) **[DIA]**.
- ◆ **MySQLWorkbench:** Herramienta de gestión y análisis del SGBD MySQL así como de las bases de datos que contiene. También permite generar diagramas entidad relación para, a partir de ellos, crear esquemas de bases de datos **[MSQWO]**.
- ◆ **LibreOffice:** Conjunto de herramientas de elaboración de documentos libre que incluye un procesador de textos, software de hoja de cálculo y software de elaboración de presentaciones entre otros. Es la suite ofimática utilizada para confeccionar esta documentación **[LOFFI]**.
- ◆ **Apache Tomcat:** Contenedor de servlets y servidor web libre que implementa diversas especificaciones de J2EE. Es el servidor utilizado para realizar el despliegue de la aplicación web WebKumo **[TOMC1]**.
- ◆ **Natty:** Librería de detección Java de fechas en lenguaje natural **[NATTY]**.

- ◆ **DataTables:** Librería JavaScript con integración con JQuery y Bootstrap que permite paginar y filtrar información en una tabla con facilidad **[DTAB]**.
- ◆ **Java Language Detector:** Librería Java para la detección de idiomas de un texto. **[LANDE]**.
- ◆ **Nu.Parser:** Librería de procesado de código HTML utilizado para la validación del código HTML (en su versión para Java) **[NUPAR]**.
- ◆ **Phloc CSS Parser:** Librería de procesado de CSS utilizado para la validación del código fuente CSS en el análisis de sintaxis **[PHLOC]**.
- ◆ **Jsoup:** Librería de procesado de código HTML utilizado en el análisis de dispositivos móviles **[JSOUP]**.

## 7.4 Creación del sistema

### 7.4.1 Problemas encontrados

Se enumeran aquí una serie de problemas encontrados que destacaron durante el desarrollo del proyecto por su persistencia o complejidad.

1. La documentación referida al software de araña web Apache Nutch, es extremadamente pobre y/o desactualizada. Además la que existe suele referirse a la rama 1.X de Nutch la cuál contiene divergencias importantes respecto a la rama 2.X que es la que se ha utilizado en este proyecto . Adicionalmente, Apache Nutch no posee una comunidad de usuarios demasiado grande lo que implica que ante un problema, es bastante probable que no se encuentre documentado. En muchas ocasiones se tuvo que recurrir a examinar el código fuente de Apache Nutch (y a su no muy explicativo Javadoc) para clarificar su funcionamiento.
2. A la hora de configurar el framework Spring Security que se encarga de las tareas de autenticación y seguridad de la aplicación web, se produjo un problema persistente que requirió varios días para su solución. El problema que se presentaba era que Spring, en su inicialización, no era capaz de encadenar el filtro que Spring Security utiliza para proveer sus servicios. Se probaron múltiples soluciones sin éxito. El problema desapareció finalmente tras desplegar la aplicación en otro contenedor de aplicaciones (en Apache Tomcat 8, en la versión 7 era donde el problema aparecía).
3. El conector de Apache Nutch en el que se basó WebKumo Connector, incluye la utilización de un filtro en la fase de ejecución de procesamiento de los recursos (*parse*) definido en la clase `ParseFilters`. Este filtro provoca que el plugin WebKumo Connector falle en su ejecución debido a una excepción incontrolada. Parece que este problema tiene relación con la codificación de los recursos web ya que se ha observado en análisis de prueba en las que las semillas eran portales web asiáticos. Existen varios indicios **[NUTCH4]** de este problema (cuya raíz parece no estar en el filtro en sí, sino en la clase de utilidad `ParseUtil`) por otros usuarios, pero no existe ningún *bug* abierto.

4. Resulta difícil depurar el *plugin* de Apache Nutch desarrollado, WebKumo Connector debido a que parte de su funcionalidad reside en interactuar con Apache Nutch para realizar procesos previos de filtrado y procesado de los recursos, procesos los cuáles no son controlados por el plugin sino que son operaciones internas de Apache Nutch.  
Cuando se ejecuta mediante su API Java, el conector de Apache Nutch no es capaz de finalizar su ejecución (previsiblemente por incompatibilidad de librerías **[DHELL]**) pero tampoco imprime la traza de la excepción en los *logs* debido a un *bug* **[HADO4]** presente en la última versión de Apache Hadoop. Sin esta traza resulta prácticamente imposible encontrar el error debido a la complejidad del sistema.  
Este problema solo se presenta si se ejecuta Apache Nutch mediante su API de Java (NutchService), funcionando correctamente si se ejecuta mediante su interfaz de línea de comandos (NutchCommandLineService). Es por esto que **se recomienda el uso de WebKumo en conjunción con la implementación del servicio de Nutch vía interfaz de línea de comandos** (aspecto configurable mediante la factoría BackendServicesFactory) hasta que sea lanzada la versión 1.3.0 de Apache Hadoop que contendrá el arreglo.
5. Existe una gran confusión, debido en parte a la pobre documentación de Apache Nutch, sobre las dependencias que se requieren para ejecutar la plataforma Apache Nutch en sus distintas versiones. Se realizaron diversas pruebas con distintas versiones de Apache Nutch, Apache Hadoop y Apache Gora hasta que se encontró una combinación compatible (Nutch 2.2.1, Hadoop 1.2.0 y Gora 0.3). Como ejemplo de esta confusión, Apache Nutch 2.3 no funciona sobre Hadoop 1.20 porque incluye una versión de Gora mas actualizada la cuál provoca excepciones incontroladas debidas a un *bug* **[NUTCH5]**.
6. Una dificultad encontrada al construir la configuración de la aplicación web WebKumo mediante configuración programática, es la falta de ejemplificaciones claras sobre cómo afrontar dicha tarea mas allá de interminables referencias técnicas. Fue de mucha utilidad el seguimiento de un tutorial **[JACONF1]** de migración de la configuración basada en ficheros XML a configuración programática Java así como plantillas de código de terceros **[JACONF2]**.
7. En la implementación del arranque y/o parada mediante la interfaz web WebKumo, se utiliza el código de salida **[EXSTA]** del proceso para determinar si la operación fue exitosa o no (si el código de retorno es cero, la operación fue exitosa, si no se produjo un error). Este mecanismo estándar funciona correctamente en la mayoría de los casos, sin embargo,

en ocasiones Apache Hadoop retorna un código de retorno distinto de cero a pesar de que la operación de arranque o parada haya sido exitosa.

## 7.4.2 Descripción detallada de las clases

Esta sección pretende compilar todos los aspectos de las clases utilizadas en el diseño orientado a objetos de tal modo que se expliciten todos los atributos y métodos junto a sus descripciones.

Para evitar recargar innecesariamente este documento y gracias a la generación automática de la documentación de la API mediante Javadoc, se recogerá la descripción detallada de las clases como anexo a esta documentación. El formato en el que se proporciona esta descripción es mediante una serie de documentos HTML navegables.

Consultar la sección 14 de esta documentación para comprobar dónde se encuentran localizados los documentos generados.

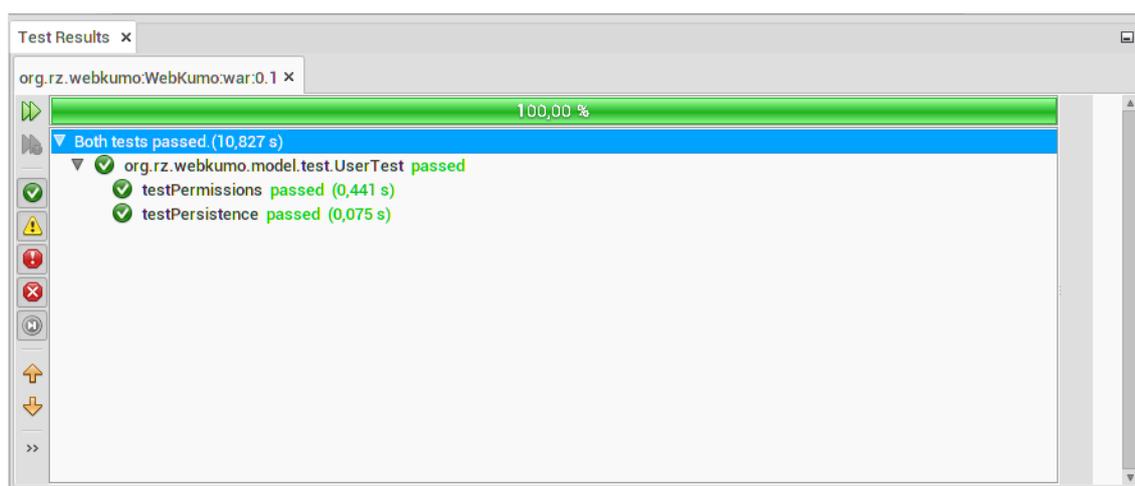
# 8 Desarrollo de las pruebas

## 8.1 Pruebas unitarias

Como se justificó en la especificación técnica del plan de pruebas, se decidió probar de manera unitaria únicamente las entidades de usuarios, grupos y algoritmos de análisis debido a la imposibilidad de realizar pruebas unitarias sobre el resto de entidades. Esto era debido a su alto grado de dependencia de la infraestructura de computación subyacente. Se describen a continuación los resultados de la ejecución de dichas pruebas.

### 8.1.1 Pruebas sobre usuarios y grupos

Primeramente se realizaron pruebas únicamente sobre el modelo (estas pruebas son más próximas al tipo de pruebas unitarias). Se refiere este apartado a la ejecución de los tests contenidos en `org.rz.webkumo.model.test`.

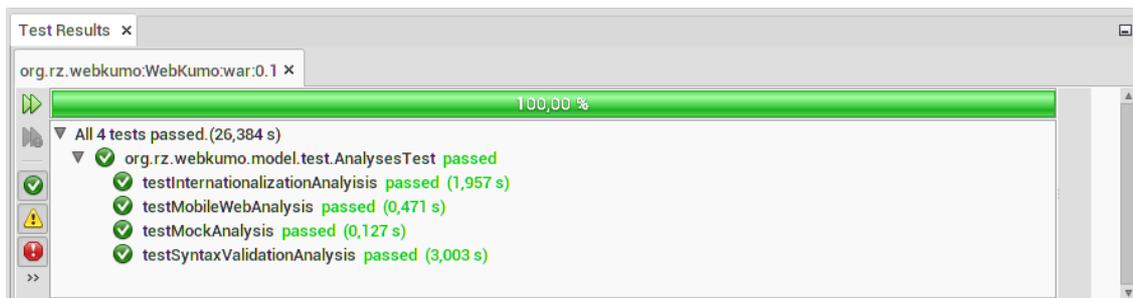


**Fig. 69:** Ejecución exitosa de los tests unitarios de usuarios y grupos.

UserTest	
testPermissions	<ul style="list-style-type: none"> <li>◆ <b>Se prueba:</b> La función que controla los trabajos que son visibles para un usuario.</li> <li>◆ <b>Resultado esperado:</b> <ul style="list-style-type: none"> <li>● Un usuario sin permisos para ver trabajos grupales sólo podrá ver sus propios trabajos.</li> <li>● Un usuarios con permisos de lectura sobre trabajos grupales podrá ver sus trabajos y los del grupo.</li> <li>● Un usuario sin permisos de lectura no podrá ver ningún trabajo.</li> <li>● Un usuario sin permisos de escritura no podrá ser dueño de ningún trabajo.</li> </ul> </li> <li>◆ <b>Resultado obtenido:</b> El esperado.</li> </ul>
testPersistence	<ul style="list-style-type: none"> <li>◆ <b>Se prueba:</b> Los usuarios y los grupos a los que pertenecen sean persistidos correctamente.</li> <li>◆ <b>Resultado esperado:</b> <ul style="list-style-type: none"> <li>● Los usuarios y grupos persistidos existan.</li> <li>● Un grupo contenga a sus miembros.</li> </ul> </li> <li>◆ <b>Resultado obtenido:</b> El esperado.</li> </ul>

### 8.1.2 Pruebas sobre algoritmos de análisis

Se realizaron pruebas unitarias sobre los algoritmos de análisis que se proporcionan. Para estas pruebas, se simularon recursos web que los análisis deberán comprobar.



**Fig. 70: Ejecución exitosa de los test de análisis.**

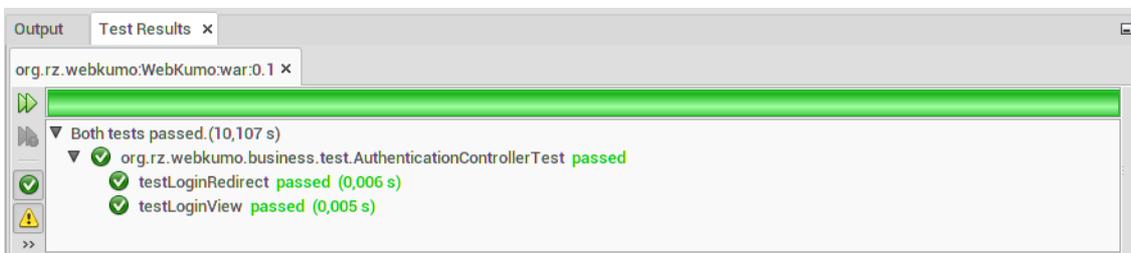
<b>AnalysesTest</b>	
testInternationalizationAnalysis	<ul style="list-style-type: none"> <li>◆ <b>Se prueba:</b> El análisis de internacionalización proporcione el resultado correcto.</li> <li>◆ <b>Resultado esperado:</b> <ul style="list-style-type: none"> <li>● Un recurso HTML en castellano con fechas y numerales en formato adecuado para el <i>Locale "es"</i> tendrá una puntuación de 100 puntos.</li> <li>● Un recurso HTML en inglés con fechas y numerales en formato adecuado para el <i>Locale "en"</i> tendrá una puntuación de 100 puntos.</li> </ul> </li> <li>◆ <b>Resultado obtenido:</b> El esperado.</li> </ul>
testMobileWebAnalysis	<ul style="list-style-type: none"> <li>◆ <b>Se prueba:</b> Un recurso HTML y otro CSS para comprobar si siguen las guías de adaptación a dispositivos móviles del W3C.</li> <li>◆ <b>Resultado esperado:</b> <ul style="list-style-type: none"> <li>● El recurso HTML adaptado a dispositivos móviles obtiene una puntuación de 100 puntos.</li> <li>● El recurso CSS adaptado a dispositivos móviles obtiene una puntuación de 100 puntos.</li> </ul> </li> <li>◆ <b>Resultado obtenido:</b> El esperado.</li> </ul>
testSyntaxValidationAnalysis	<ul style="list-style-type: none"> <li>◆ <b>Se prueba:</b> La validez sintáctica de recursos web HTML y CSS.</li> <li>◆ <b>Resultado esperado:</b> <ul style="list-style-type: none"> <li>● Dos recursos validos y dos inválidos provocarán que el análisis tenga una puntuación de 50 puntos.</li> </ul> </li> <li>◆ <b>Resultado obtenido:</b> El esperado.</li> </ul>

## 8.2 Pruebas del sistema

Esta sección se encargará de mostrar los resultados de las pruebas de funcionalidad ejecutadas. Se detallan los tests automatizados realizados a continuación.

### 8.2.1 Pruebas de gestión de usuarios

Se comienza probando el caso de uso de inicio de sesión.

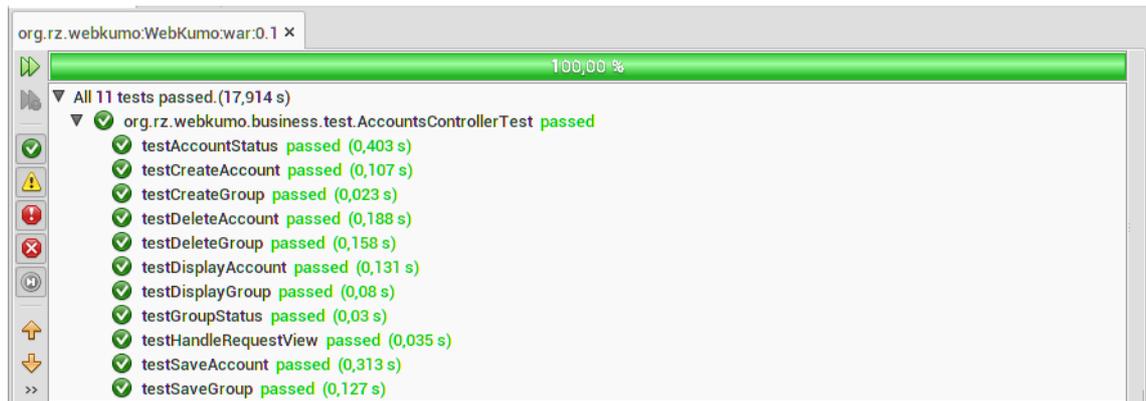


**Fig. 71: Ejecución exitosa de las pruebas de funcionalidad de inicio de sesión.**

AuthenticationControllerTest	
testLoginRedirect	<ul style="list-style-type: none"> <li>◆ <b>Se prueba:</b> Que la función de redirección de la pantalla de <i>login</i> sea correcta.</li> <li>◆ <b>Resultado esperado:</b> <ul style="list-style-type: none"> <li>● Un usuario que haya iniciado sesión, al acceder a la página de inicio de sesión será redirigido al panel de vista general.</li> <li>● Un usuario que haya iniciado sesión, al acceder a la página de inicio de sesión será redirigido al formulario para introducir sus credenciales.</li> </ul> </li> <li>◆ <b>Resultado obtenido:</b> El esperado.</li> </ul>

testLoginView	<ul style="list-style-type: none"> <li>◆ <b>Se prueba:</b> Que los mensajes de la vista de inicio de sesión se muestren correctamente.</li> <li>◆ <b>Resultado esperado:</b> <ul style="list-style-type: none"> <li>● Un usuario que haya cerrado sesión deberá poder ver un mensaje de despedida cuando se le dirija al formulario de inicio de sesión.</li> <li>● Un usuario que haya introducido credenciales incorrectas en el formulario de inicio debe poder ver un mensaje indicando el error.</li> </ul> </li> <li>◆ <b>Resultado obtenido:</b> El esperado.</li> </ul>
---------------	---

Ahora se muestran los resultados del resto de casos de uso de gestión de usuarios y grupos.



**Fig. 72:** Ejecución exitosa de las pruebas de funcionalidad relacionada con usuarios

AccountsControllerTest	
testHandleRequestView	<ul style="list-style-type: none"> <li>◆ <b>Se prueba:</b> Que se cargue correctamente el listado de usuarios en el panel del usuario.</li> <li>◆ <b>Resultado esperado:</b> <ul style="list-style-type: none"> <li>● Un usuario administrador podrá ver el listado de usuarios y grupos.</li> <li>● Un usuario regular no podrá ver el listado de usuarios y grupos.</li> </ul> </li> <li>◆ <b>Resultado obtenido:</b> El esperado.</li> </ul>

testAccountStatus	<ul style="list-style-type: none"> <li>◆ <b>Se prueba:</b> La generación del listado de usuarios para el panel del administrador.</li> <li>◆ <b>Resultado esperado:</b> <ul style="list-style-type: none"> <li>● Para un usuario administrador se generará la lista de usuarios.</li> <li>● Para un usuario regular no se generará la lista de usuarios.</li> </ul> </li> <li>◆ <b>Resultado obtenido:</b> El esperado.</li> </ul>
testCreateAccount	<ul style="list-style-type: none"> <li>◆ <b>Se prueba:</b> El acceso a la página de creación de usuario.</li> <li>◆ <b>Resultado esperado:</b> <ul style="list-style-type: none"> <li>● Para un usuario administrador se permitirá el acceso.</li> <li>● Para un usuario regular no se permitirá el acceso</li> </ul> </li> <li>◆ <b>Resultado obtenido:</b> El esperado.</li> </ul>
testCreateGroup	<ul style="list-style-type: none"> <li>◆ <b>Se prueba:</b> El acceso a la página de creación de grupo.</li> <li>◆ <b>Resultado esperado:</b> <ul style="list-style-type: none"> <li>● Para un usuario administrador se permitirá el acceso.</li> <li>● Para un usuario regular no se permitirá el acceso</li> </ul> </li> <li>◆ <b>Resultado obtenido:</b> El esperado.</li> </ul>
TestDisplayAccount	<ul style="list-style-type: none"> <li>◆ <b>Se prueba:</b> El acceso a la página de información y edición de un usuario.</li> <li>◆ <b>Resultado esperado:</b> <ul style="list-style-type: none"> <li>● Para un usuario regular se impedirá el acceso cuando intente acceder a la información de otro usuario distinto.</li> <li>● Para un usuario cualquiera que intenta acceder a la información de un usuario inexistente se muestra la página de error 404.</li> </ul> </li> </ul>

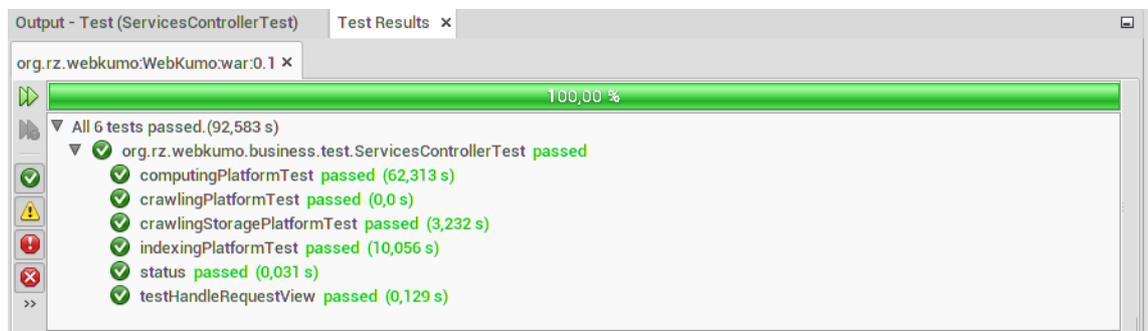
	<ul style="list-style-type: none"> <li>● Un usuario administrador podrá acceder a la información de cualquier otro usuario.</li> <li>● Un usuario regular podrá acceder a su propia información.</li> </ul> <p>◆ <b>Resultado obtenido:</b> El esperado</p>
testDisplayGroup	<p>◆ <b>Se prueba:</b> El acceso a la página de información y edición de un grupo.</p> <p>◆ <b>Resultado esperado:</b></p> <ul style="list-style-type: none"> <li>● Para un usuario regular se impedirá el acceso cuando intente acceder a la información de un grupo.</li> <li>● Para un usuario administrador que intenta acceder a la información de un grupo inexistente se muestra la página de error 404.</li> <li>● Un usuario administrador podrá acceder a la información de cualquier grupo.</li> </ul> <p>◆ <b>Resultado obtenido:</b> El esperado</p>
testGroupStatus	<p>◆ <b>Se prueba:</b> La generación del listado de grupos para el panel del administrador.</p> <p>◆ <b>Resultado esperado:</b></p> <ul style="list-style-type: none"> <li>● Para un usuario administrador se generará la lista de grupos.</li> <li>● Para un usuario regular no se generará la lista de grupos.</li> </ul> <p>◆ <b>Resultado obtenido:</b> El esperado.</p>
testSaveAccount	<p>◆ <b>Se prueba:</b> La persistencia de una cuenta de usuario.</p> <p>◆ <b>Resultado esperado:</b></p> <ul style="list-style-type: none"> <li>● Para un usuario regular se impedirá el guardado de la información de otro usuario distinto.</li> <li>● Para un usuario regular se permitirá el guardado de la información de sí mismo.</li> <li>● Un usuario administrador podrá guardar la</li> </ul>

	<p>información de cualquier usuario.</p> <ul style="list-style-type: none"> <li>● Un usuario que guarde los datos de un usuario con valores inválidos será redirigido a la vista 'account' y el modelo contendrá el objeto 'fields' que contiene los campos erróneos.</li> </ul> <p>◆ <b>Resultado obtenido:</b> El esperado.</p>
testSaveGroup	<p>◆ <b>Se prueba:</b> La persistencia de una grupo de usuarios.</p> <p>◆ <b>Resultado esperado:</b></p> <ul style="list-style-type: none"> <li>● Para un usuario regular se impedirá el guardado de la información un grupo.</li> <li>● Un usuario administrador podrá guardar la información de cualquier grupo.</li> <li>● Un usuario que guarde los datos de un grupo con valores inválidos será redirigido a la vista 'group' y el modelo contendrá el objeto 'fields' que contiene los campos erróneos.</li> </ul> <p>◆ <b>Resultado obtenido:</b> El esperado.</p>
testDeleteGroup	<p>◆ <b>Se prueba:</b> La funcionalidad de borrado de grupos.</p> <p>◆ <b>Resultado esperado:</b></p> <ul style="list-style-type: none"> <li>● Para un grupo con usuarios miembros, la baja no debe realizarse.</li> <li>● En el caso de un grupo sin usuarios miembros, se debe dar de baja efectivamente el grupo</li> </ul> <p>◆ <b>Resultado obtenido:</b> El esperado.</p>

testDeleteUser	<ul style="list-style-type: none"> <li>◆ <b>Se prueba:</b> La funcionalidad de borrado de un usuario.</li> <li>◆ <b>Resultado esperado:</b> <ul style="list-style-type: none"> <li>● Para un usuario con trabajos asociados, la baja no debe realizarse.</li> <li>● En el caso de un usuario sin trabajos asociados, se debe dar de baja efectivamente el usuario</li> </ul> </li> <li>◆ <b>Resultado obtenido:</b> El esperado.</li> </ul>
----------------	---

## 8.2.2 Pruebas de servicios

Otro de los casos de uso cuya funcionalidad se comprueba mediante pruebas del sistema es la de la gestión de los servicios. Los resultados de dichas pruebas automatizadas son los que siguen:



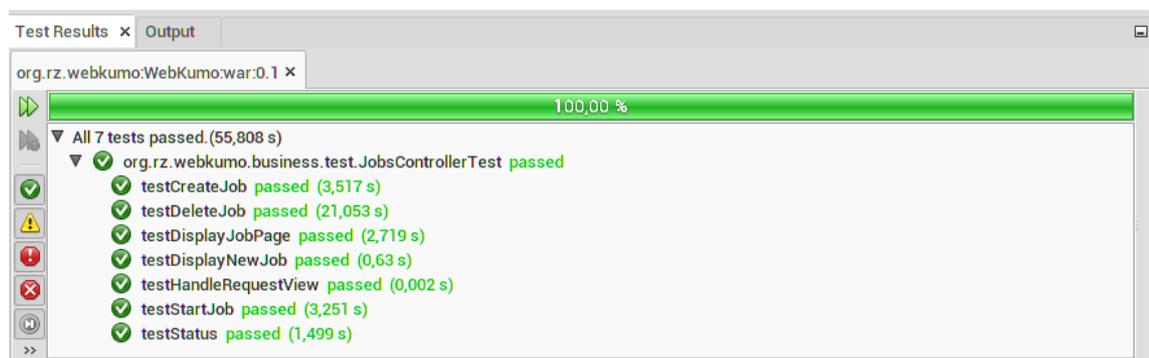
**Fig. 73: Ejecución exitosa de las pruebas de sistema sobre los servicios.**

<b>ServicesControllerTest</b>	
computingPlatformTest	<ul style="list-style-type: none"> <li>◆ <b>Se prueba:</b> El servicio de computación distribuida mediante su puesta en marcha, comprobación de disponibilidad y posterior detención.</li> <li>◆ <b>Resultado esperado:</b> <ul style="list-style-type: none"> <li>● El servicio se inicia correctamente.</li> <li>● El servicio se muestra disponible y así se muestra en su estado.</li> <li>● El servicio se detiene con éxito.</li> </ul> </li> <li>◆ <b>Resultado obtenido:</b> El esperado.</li> </ul>
crawlingPlatformTest	<ul style="list-style-type: none"> <li>◆ <b>Se prueba:</b> El servicio de araña web mediante su puesta en marcha, comprobación de disponibilidad y posterior detención.</li> <li>◆ <b>Resultado esperado:</b> <ul style="list-style-type: none"> <li>● El servicio se inicia correctamente.</li> <li>● El servicio se muestra disponible y así se muestra en su estado.</li> <li>● El servicio se detiene con éxito.</li> </ul> </li> <li>◆ <b>Resultado obtenido:</b> El esperado.</li> </ul>
crawlingStoragePlatformTest	<ul style="list-style-type: none"> <li>◆ <b>Se prueba:</b> El servicio de araña web mediante su puesta en marcha, comprobación de disponibilidad y posterior detención.</li> <li>◆ <b>Resultado esperado:</b> <ul style="list-style-type: none"> <li>● El servicio se inicia correctamente.</li> <li>● El servicio se muestra disponible y así se muestra en su estado.</li> <li>● El servicio se detiene con éxito.</li> </ul> </li> <li>◆ <b>Resultado obtenido:</b> El esperado.</li> </ul>
indexingPlatformTest	<ul style="list-style-type: none"> <li>◆ <b>Se prueba:</b> El servicio de araña web mediante su puesta en marcha, comprobación de disponibilidad y posterior detención.</li> <li>◆ <b>Resultado esperado:</b> <ul style="list-style-type: none"> <li>● El servicio se inicia correctamente.</li> </ul> </li> </ul>

	<ul style="list-style-type: none"> <li>● El servicio se muestra disponible y así se muestra en su estado.</li> <li>● El servicio se detiene con éxito.</li> </ul> <p>◆ <b>Resultado obtenido:</b> El esperado.</p>
status	<p>◆ <b>Se prueba:</b> La obtención del estado de todos los servicios (lo que constituye el listado).</p> <p>◆ <b>Resultado esperado:</b></p> <ul style="list-style-type: none"> <li>● Se deben obtener el estado de los 4 servicios.</li> </ul> <p>◆ <b>Resultado obtenido:</b> El esperado.</p>
testHandleRequestView	<p>◆ <b>Se prueba:</b> El acceso al panel de servicios.</p> <p>◆ <b>Resultado esperado:</b></p> <ul style="list-style-type: none"> <li>● Si el usuario es administrador se debe permitir el acceso al panel de servicios.</li> <li>● Si el usuario no es administrador se debe impedir el acceso al panel de servicios.</li> </ul> <p>◆ <b>Resultado obtenido:</b> El esperado.</p>

### 8.2.3 Pruebas de trabajos

Otro conjunto de casos de uso a probar es el de trabajos cuya prueba ha sido automatizada y cuyos resultados pueden verse a continuación:



**Fig. 74: Ejecución exitosa de las pruebas de sistema sobre trabajo.**

<b>JobsControllerTest</b>	
testHandleRequestView	<ul style="list-style-type: none"> <li>◆ <b>Se prueba:</b> Al solicitar el acceso al panel de trabajos, este se permite.</li> <li>◆ <b>Resultado esperado:</b> <ul style="list-style-type: none"> <li>● El el controlador permite el acceso al panel de trabajos</li> </ul> </li> <li>◆ <b>Resultado obtenido:</b> El esperado.</li> </ul>
testDisplayNewJob	<ul style="list-style-type: none"> <li>◆ <b>Se prueba:</b> El acceso a la página de creación de un nuevo trabajo.</li> <li>◆ <b>Resultado esperado:</b> <ul style="list-style-type: none"> <li>● Un usuario con permisos de escritura (o administrador) podrá acceder.</li> <li>● Un usuario sin permisos de escritura verá denegado el acceso.</li> </ul> </li> <li>◆ <b>Resultado obtenido:</b> El esperado.</li> </ul>
testCreateJob	<ul style="list-style-type: none"> <li>◆ <b>Se prueba:</b> La creación de un nuevo trabajo.</li> <li>◆ <b>Resultado esperado:</b> <ul style="list-style-type: none"> <li>● Un usuario que cree un trabajo con campos erróneos no podrá crearlo y será dirigido al formulario de nuevo con una serie de mensajes de error.</li> <li>● Un usuario que cree un trabajo con campos correctos tendrá como resultado el alta efectiva del trabajo.</li> <li>● Un usuario que intente crear un trabajo con campos correctos pero sin permisos de escritura será redirigido a una página de recurso no encontrado.</li> </ul> </li> <li>◆ <b>Resultado obtenido:</b> El esperado.</li> </ul>

testDisplayJobPage	<ul style="list-style-type: none"> <li>◆ <b>Se prueba:</b> El acceso a la información de un trabajo.</li> <li>◆ <b>Resultado esperado:</b> <ul style="list-style-type: none"> <li>● Un usuario con permisos para visualizar el trabajo verá aceptada su solicitud de acceso.</li> <li>● Un usuario sin permisos de lectura de trabajos grupales verá denegada su solicitud de acceso a un trabajo de otro usuario de su mismo grupo.</li> <li>● Un usuario sin ningún permiso de lectura no podrá acceder tampoco a la información del trabajo.</li> </ul> </li> <li>◆ <b>Resultado obtenido:</b> El esperado.</li> </ul>
testStartJob	<ul style="list-style-type: none"> <li>◆ <b>Se prueba:</b> El funcionamiento del envío de un trabajo para su ejecución.</li> <li>◆ <b>Resultado esperado:</b> <ul style="list-style-type: none"> <li>● Un usuario no podrá iniciar su trabajo debido a que los servicios de infraestructura no están iniciados.</li> <li>● Un usuario no podrá iniciar el trabajo de otro usuario distinto.</li> </ul> </li> <li>◆ <b>Resultado obtenido:</b> El esperado .</li> </ul>
testStatus	<ul style="list-style-type: none"> <li>◆ <b>Se prueba:</b> La obtención de los últimos trabajos que un usuario puede leer.</li> <li>◆ <b>Resultado esperado:</b> <ul style="list-style-type: none"> <li>● Para un usuario que sólo tiene permisos de lectura, se obtendrán sus propios trabajos.</li> <li>● Para un usuario que posee permisos de lectura grupal, se obtendrán sus trabajos y los del grupo.</li> <li>● Para un usuario sin permisos de lectura no se retornará ningún trabajo.</li> </ul> </li> <li>◆ <b>Resultado obtenido:</b> El esperado.</li> </ul>

testDeleteJob	<ul style="list-style-type: none"><li>◆ <b>Se prueba:</b> El borrado de un trabajo.</li> <li>◆ <b>Resultado esperado:</b><ul style="list-style-type: none"><li>● Un usuario regular que intente eliminar un trabajo de otro usuario distinto que no haya sido ejecutado, verá denegada su solicitud.</li><li>● Un usuario regular que intente eliminar un trabajo propio que no haya sido ejecutado, verá aceptada su solicitud.</li><li>● Un usuario (de cualquier tipo) que intente eliminar un trabajo finalizado, verá denegada su solicitud.</li><li>● Un usuario administrador que intente borrar cualquier trabajo que no haya sido ejecutado, verá aceptada su solicitud.</li></ul></li> <li>◆ <b>Resultado obtenido:</b> El esperado.</li></ul>
---------------	--

Para la prueba de trabajos, también se tiene como comprobación que cubre el caso de uso de “Ejecución de trabajo”, el siguiente test automatizado:



**Fig. 75: Ejecución exitosa de la prueba funcional de ejecución de un trabajo.**

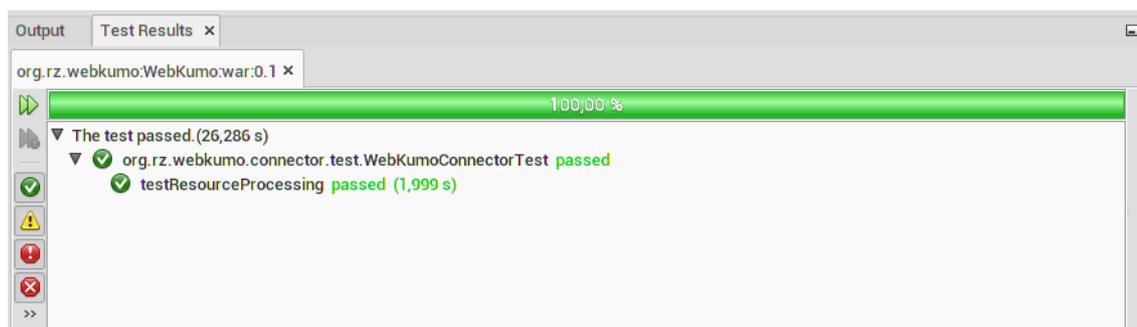
JobTest	
jobTest	<ul style="list-style-type: none"><li>◆ <b>Se prueba:</b> La ejecución completa de un trabajo sobre la infraestructura de procesamiento.</li><li>◆ <b>Resultado esperado:</b><ul style="list-style-type: none"><li>● Que se ejecute el trabajo de análisis y produzca resultados. Al finalizar, su estado no debe ser "fallido" ni "parado".</li></ul></li><li>◆ <b>Resultado obtenido:</b> El esperado.</li></ul>

## 8.2.4 Pruebas de vista general

El caso de uso de visualización de la situación general queda cubierto por las pruebas Status de `ServicesControllerTest` y de `JobsControllerTests` por lo que no es necesario realizar pruebas específicas.

## 8.2.5 Prueba de WebKumo Connector

El plugin de Apache Nutch desarrollado se puede probar mediante la implementación de una simulación de ejecución de su función principal `getParse()`. El resultado de la prueba es el que sigue:



**Fig. 76: Ejecución exitosa de la prueba de WebKumo Connector**

<b>WebKumoConnectorTest</b>	
testResourceProces sing	<ul style="list-style-type: none"><li>◆ <b>Se prueba:</b> El procesado de un recurso web de ejemplo.</li><li>◆ <b>Resultado esperado:</b><ul style="list-style-type: none"><li>● Que se ejecute sin producir ninguna excepción y que al finalizar se haya generado el fichero de resultados en el sistema de archivos distribuido HDFS.</li></ul></li><li>◆ <b>Resultado obtenido:</b> El esperado.</li></ul>

## 8.3 Pruebas de usabilidad y accesibilidad

### 8.3.1 Pruebas de usabilidad

A continuación se describen los resultados de las pruebas de usabilidad mediante los cuestionarios que se realizaron a los dos participantes. Las pruebas de usabilidad se realizaron mediante las tareas guiadas descritas en el apartado de descripción técnica de las pruebas de usabilidad. En todo momento los usuarios han podido consultar el manual del usuario.

#### 8.3.1.1 Perfiles de los sujetos de prueba

Los usuarios que participaron poseen el siguiente perfil:

Sujeto de pruebas 1	
Nombre	Adriano
Edad	25
Nivel de estudios	Superiores: Ingeniero Técnico en Informática de Sistemas
Actividad profesional actual	Ninguna

Sujeto de pruebas 2	
Nombre	Francesc
Edad	25
Nivel de estudios	Superiores: Ingeniero Técnico en Informática de Sistemas
Actividad profesional actual	Programador junior, Madrid.

## 8.3.1.2 Cuestionarios previos

<b>Sujeto de pruebas 1</b>	
¿Con qué frecuencia usa un computador?	<ol style="list-style-type: none"> <li>1. Todos los días</li> <li>2. Varias veces a la semana</li> <li>3. Ocasionalmente</li> <li>4. Nunca o casi nunca</li> </ol>
¿Qué tipo de actividades realiza mediante un computador?	<ol style="list-style-type: none"> <li>1. Parte de mi trabajo</li> <li>2. Actividades de ocio</li> <li>3. Manejo básico de algún tipo de software</li> <li>4. Uso ocasional (correo electrónico o similar)</li> </ol>
¿Qué grado de conocimiento posee sobre Apache Hadoop o sobre otros sistemas de computación distribuida?	<ol style="list-style-type: none"> <li>1. Alto</li> <li>2. Medio</li> <li>3. Bajo</li> <li>4. Ninguno</li> </ol>
¿Con qué tipo de interfaz se siente más cómodo a la hora de trabajar?	<ol style="list-style-type: none"> <li>1. Interfaz web</li> <li>2. Interfaz de escritorio</li> <li>3. Interfaz de comandos</li> <li>4. Otra (especificar cuál)</li> </ol>
Enumere tres ramas o campos de la informática que le susciten interés	<ol style="list-style-type: none"> <li>1. Diseño de hardware</li> <li>2. Sistemas empujados</li> <li>3. Administración de sistemas</li> </ol>

<b>Sujeto de pruebas 2</b>	
¿Con qué frecuencia usa un computador?	<ol style="list-style-type: none"> <li>1. Todos los días</li> <li>2. Varias veces a la semana</li> <li>3. Ocasionalmente</li> <li>4. Nunca o casi nunca</li> </ol>
¿Qué tipo de actividades realiza mediante un computador?	<ol style="list-style-type: none"> <li>1. Parte de mi trabajo</li> <li>2. Actividades de ocio</li> <li>3. Manejo básico de algún tipo de software</li> <li>4. Uso ocasional (correo electrónico o similar)</li> </ol>
¿Qué grado de conocimiento posee sobre Apache Hadoop o sobre otros sistemas de	<ol style="list-style-type: none"> <li>1. Alto</li> <li>2. Medio</li> <li>3. Bajo</li> </ol>

computación distribuida?	4. Ninguno
¿Con qué tipo de interfaz se siente más cómodo a la hora de trabajar?	<b>1. Interfaz web</b> 2. Interfaz de escritorio 3. Interfaz de comandos 4. Otra (especificar cuál)
¿Qué rama de la informática le suscita más interés?	<b>1. Automatización</b> <b>2. Informática industrial</b> <b>3. Redes</b>

### 8.3.1.3 Sujeto de pruebas 1: Cuestionarios de actividades guiadas

<b>FACILIDAD DE USO</b>	<b>Siempre</b>	<b>Casi siempre</b>	<b>A veces</b>	<b>Nunca</b>
¿Sabe en qué parte de la aplicación se encuentra?	<b>X</b>			
Si está en un proceso de varios pasos, ¿sabe con facilidad en cuál está?	<b>X</b>			
¿Sabe, ante un formulario, lo que significa cada campo con precisión?			<b>X</b>	
Ante una operación que requiere una espera, ¿sabe que dicha operación se está llevando a cabo?	<b>X</b>			
Cuando se produjo un error, ¿supo con facilidad que dicho error ocurrió?	<b>X</b>			
¿Le resultó difícil volver a una pantalla anterior en algún momento?			<b>X</b>	
¿Fueron útiles los mensajes de error que aparecieron al introducir algún valor incorrecto en un formulario?	<b>X</b>			
¿Le fue difícil determinar cuál había sido el resultado de un trabajo?				<b>X</b>

¿Le fue difícil conocer cuál era el estado de un trabajo?				<b>X</b>
¿La forma de indicar el progreso de un trabajo le fue suficientemente explicativa?	<b>X</b>			

<b>FUNCIONALIDAD</b>	<b>Siempre</b>	<b>Casi siempre</b>	<b>A veces</b>	<b>Nunca</b>
¿Las tareas funcionaron como se esperaba?	<b>X</b>			
¿El tiempo de respuesta al acceder al resultado de un trabajo tras su ejecución fue excesivo?				<b>X</b>
¿Se encontró con alguna pantalla que indicara que hubo un error interno o que el recurso no había sido encontrado?				<b>X</b>
¿El trabajo ejecutado mostró algún mensaje de error tras su finalización?				<b>X</b>
¿Pudo iniciar sesión con el usuario creado tras haber completado el proceso?		<b>X</b>		
¿El servicio se mostraba como iniciado tras enviar la orden?	<b>X</b>			
¿Seguía apareciendo el grupo eliminado en el panel del administrador tras haberlo borrado?				<b>X</b>

<b>ASPECTOS GRÁFICOS</b>	<b>Siempre</b>	<b>Casi siempre</b>	<b>A veces</b>	<b>Nunca</b>
¿El tamaño del texto le fue impedimento?				<b>X</b>
¿Algún texto mostraba caracteres extraños o ilegibles?			<b>X</b>	
¿El color del fondo le molestó de alguna manera?				<b>X</b>

¿Supo identificar correctamente el botón que realizaba la acción que usted deseaba?	<b>X</b>			
¿El diseño de las páginas le resultó claro?	<b>X</b>			
¿Le ayudó a la hora de identificar un error que este se mostrara mediante un panel de color rojo?	<b>X</b>			

<b>DISEÑO</b>	<b>Sí</b>	<b>No</b>	<b>A veces</b>
¿La interfaz le pareció atractiva?	<b>X</b>		
¿Encontró fácil de usar la interfaz?	<b>X</b>		
¿Existió algún elemento gráfico (botones, flechas ...) que no supiera identificar?			<b>X</b>
¿Supo identificar el menú de navegación con facilidad?	<b>X</b>		
¿La colocación de elementos para mostrar los usuarios del sistema le resultó adecuada?	<b>X</b>		
¿Le aportó algo el uso de iconos para indicar elementos con los que podía interactuar?	<b>X</b>		
¿Le fue sencillo seleccionar múltiples análisis mediante el selector en forma de lista?	<b>X</b>		
¿Encontró dificultoso la forma de ver la información en las tablas debido a su tamaño o color?		<b>X</b>	
¿Utilizó en algún momento el icono de la aplicación para acceder a la página principal?		<b>X</b>	

### 8.3.1.4 Sujeto de pruebas 1: Cuestionario del evaluador

ASPECTO OBSERVADO	Notas
El usuario duda cómo comenzar la tarea	No
El usuario se pierde a la hora de seguir una tarea con pasos	No
El usuario duda si la tarea ha finalizado o no	Parcialmente, al enviar el trabajo para su ejecución, duda de si se ha ejecutado ya o debe esperar. No tarda en ver la barra de progreso.
El usuario comienza la tarea de forma rápida sin escrutar toda la página para saber por dónde tiene que comenzar	Si.
El usuario identifica en el panel de servicios el accionamiento de arranque y/o parada con rapidez	Si.
El usuario abandona el panel de servicios antes de que haya terminado el proceso de arranque/parada	No.
El usuario identifica el resultado de la operación con el servicio como un enlace y lo utiliza para mostrar la información de salida	No. No pasa el puntero del ratón por el texto indicado.
El usuario espera a que el trabajo termine (no identifica el lanzamiento del trabajo con el fin de la ejecución del mismo).	Si.
El usuario se percató del fin del trabajo y accede a sus resultados	Si.
El usuario identifica la opción para visualizar los campos del informe del trabajo detallados y lo utiliza	Si.

El usuario pide ayuda para rellenar algún campo de algún formulario	Si.
El usuario consulta el manual para operaciones triviales	No.
Tiempo que tarda el usuario en realizar cada tarea	<ol style="list-style-type: none"> <li>1. 40s</li> <li>2. 15s</li> <li>3. 50s</li> <li>4. 45s (contando el tiempo que tardó el servicio en arrancar).</li> <li>5. 2m (contando el tiempo que tardó el trabajo en ejecutarse).</li> </ol>
Errores de relevancia cometidos	Ninguno
Otras observaciones	<ol style="list-style-type: none"> <li>1. Consultó el manual del usuario para comprobar el significado del campo de formulario de edición de un trabajo 'Recursos por nivel'.</li> <li>2. En la tarea de creación de usuario, dudó sobre cómo volver al panel del administrador. Finalmente lo hizo mediante el menú desplegable.</li> <li>3. En la operación de crear un usuario, al probar que se podía iniciar sesión con el mismo, no se pudo debido a que la aplicación había sido arrancada por error en modo "desarrollo" el sólo permite el inicio de sesión con unos usuarios predefinidos. Se subsanó y se repitió la prueba.</li> <li>4. En ocasiones aparecieron caracteres extraños en textos con tildes (¿problema de codificación?).</li> <li>5. En el formulario de creación/modificación de trabajo no supo identificar la utilidad que muestra los parámetros adicionales disponibles.</li> </ol>

### 8.3.1.5 Sujeto de pruebas 2: Cuestionarios de actividades guiadas

<b>FACILIDAD DE USO</b>	<b>Siempre</b>	<b>Casi siempre</b>	<b>A veces</b>	<b>Nunca</b>
¿Sabe en qué parte de la aplicación se encuentra?	<b>X</b>			
Si está en un proceso de varios pasos, ¿sabe con facilidad en cuál está?		<b>X</b>		
¿Sabe, ante un formulario, lo que significa cada campo con precisión?		<b>X</b>		
Ante una operación que requiere una espera, ¿sabe que dicha operación se está llevando a cabo?	<b>X</b>			
Cuando se produjo un error, ¿supo con facilidad que dicho error ocurrió?	<b>X</b>			
¿Le resultó difícil volver a una pantalla anterior en algún momento?	<b>X</b>			
¿Fueron útiles los mensajes de error que aparecieron al introducir algún valor incorrecto en un formulario?	<b>X</b>			
¿Le fue difícil determinar cuál había sido el resultado de un trabajo?				<b>X</b>
¿Le fue difícil conocer cuál era el estado de un trabajo?				<b>X</b>
¿La forma de indicar el progreso de un trabajo le fue suficientemente explicativa?	<b>X</b>			

<b>FUNCIONALIDAD</b>	<b>Siempre</b>	<b>Casi siempre</b>	<b>A veces</b>	<b>Nunca</b>
¿Las tareas funcionaron como se esperaba?	<b>X</b>			

¿El tiempo de respuesta al acceder al resultado de un trabajo tras su ejecución fue excesivo?				<b>X</b>
¿Se encontró con alguna pantalla que indicara que hubo un error interno o que el recurso no había sido encontrado?				<b>X</b>
¿El trabajo ejecutado mostró algún mensaje de error tras su finalización?				<b>X</b>
¿Pudo iniciar sesión con el usuario creado tras haber completado el proceso?	<b>X</b>			
¿El servicio se mostraba como iniciado tras enviar la orden?	<b>X</b>			
¿Seguía apareciendo el grupo eliminado en el panel del administrador tras haberlo borrado?				<b>X</b>

<b>ASPECTOS GRÁFICOS</b>	<b>Siempre</b>	<b>Casi siempre</b>	<b>A veces</b>	<b>Nunca</b>
¿El tamaño del texto le fue impedimento?				<b>X</b>
¿Algún texto mostraba caracteres extraños o ilegibles?				<b>X</b>
¿El color del fondo le molestó de alguna manera?				<b>X</b>
¿Supo identificar correctamente el botón que realizaba la acción que usted deseaba?	<b>X</b>			
¿El diseño de las páginas le resultó claro?	<b>X</b>			
¿Le ayudó a la hora de identificar un error que este se mostrara mediante un panel de color rojo?		<b>X</b>		

<b>DISEÑO</b>	<b>Sí</b>	<b>No</b>	<b>A veces</b>
¿La interfaz le pareció atractiva?	<b>X</b>		
¿Encontró fácil de usar la interfaz?	<b>X</b>		
¿Existió algún elemento gráfico (botones, flechas ...) que no supiera identificar?		<b>X</b>	
¿Supo identificar el menú de navegación con facilidad?	<b>X</b>		
¿La colocación de elementos para mostrar los usuarios del sistema le resultó adecuada?	<b>X</b>		
¿Le aportó algo el uso de iconos para indicar elementos con los que podía interactuar?		<b>X</b>	
¿Le fue sencillo seleccionar múltiples análisis mediante el selector en forma de lista?	<b>X</b>		
¿Encontró dificultoso la forma de ver la información en las tablas debido a su tamaño o color?		<b>X</b>	
¿Utilizó en algún momento el icono de la aplicación para acceder a la página principal?	<b>X</b>		

### 8.3.1.6 Sujeto de pruebas 2: Cuestionario del evaluador

<b>ASPECTO OBSERVADO</b>	<b>Notas</b>
El usuario duda cómo comenzar la tarea	No
El usuario se pierde a la hora de seguir una tarea con pasos	No
El usuario duda si la tarea ha finalizado o no	No.
El usuario comienza la tarea de forma rápida sin escrutar toda la página para saber por dónde tiene que comenzar	En la tarea de ejecutar un trabajo, el usuario duda momentáneamente por donde empezar. Buscaba un botón de inicio de trabajo desde el panel de trabajos sin acceder al trabajo en sí.
El usuario identifica en el panel	Si.

de servicios el accionamiento de arranque y/o parada con rapidez	
El usuario abandona el panel de servicios antes de que haya terminado el proceso de arranque/parada	No.
El usuario identifica el resultado de la operación con el servicio como un enlace y lo utiliza para mostrar la información de salida	Si.
El usuario espera a que el trabajo termine (no identifica el lanzamiento del trabajo con el fin de la ejecución del mismo).	Si.
El usuario se percata del fin del trabajo y accede a sus resultados	Si.
El usuario identifica la opción para visualizar los campos del informe del trabajo detallados y lo utiliza	Si.
El usuario pide ayuda para rellenar algún campo de algún formulario	No.
El usuario consulta el manual para operaciones triviales	No.
Tiempo que tarda el usuario en realizar cada tarea	<ol style="list-style-type: none"> <li>1. 1m 10s</li> <li>2. 25s</li> <li>3. 1ms</li> <li>4. 30s (contando el tiempo que tardó el servicio en arrancar).</li> <li>5. 3m40s (contando el tiempo que tardó el trabajo en ejecutarse).</li> </ol>
Errores de relevancia cometidos	Ninguno

Otras observaciones	<ol style="list-style-type: none"> <li>1. Comentó la necesidad de una forma de cancelar el proceso de edición de un trabajo.</li> <li>2. No entendió la utilidad de los parámetros personalizados de los análisis teniendo que recurrir al manual.</li> <li>3. Indicó que la aparición de un panel nuevo ya era suficiente para identificar que se había producido un error en un formulario, aunque el color rojo ayudaba a identificarlo mejor.</li> <li>4. Los iconos de credenciales incorrectas al iniciar sesión con el usuario nuevo no eran de utilidad y parecían tener un fin mas estético, recalcó.</li> </ol>
---------------------	---

### 8.3.1.7 Análisis según guía de usabilidad

Adicionalmente a las pruebas de usabilidad realizadas mediante la evaluación manual de los sujetos de prueba, se realizó la comprobación de ciertos puntos según la guía de usabilidad **[HASS]** siguiente:

Criterios	¿Cumplido?
¿Cuáles son los objetivos del sitio web? ¿Son concretos y bien definidos? ¿Los contenidos y servicios que ofrece se corresponden con esos objetivos?	<b>SI</b>
¿Tiene una URL correcta, clara y fácil de recordar? ¿Y las URL de sus páginas internas? ¿Son claras y permanentes?	<b>SI</b>
¿Muestra de forma precisa y completa qué contenidos o servicios ofrece realmente el sitio web?	<b>SI</b>
¿La estructura general del sitio web está orientada al usuario?	<b>SI</b>
¿El <i>look &amp; feel</i> general se corresponde con los objetivos, características, contenidos y servicios del sitio web?	<b>SI</b>

¿Es coherente el diseño general del sitio web?	<b>SI</b>
¿Es reconocible el diseño general del sitio web?	<b>SI</b>
¿El sitio web se actualiza periódicamente? ¿Indica cuándo se actualiza?	<b>SI</b>
¿Se muestra claramente la identidad de la empresa-sitio a través de todas las páginas?	<b>SI</b>
El Logotipo, ¿es significativo, identificable y suficientemente visible?	<b>SI</b>
El eslogan o <i>tagline</i> , ¿expresa realmente qué es la empresa y qué servicios ofrece?	<b>N/A</b>
¿Se ofrece algún enlace con información sobre la empresa, sitio web, 'webmaster',...?	<b>N/A</b>
¿Se proporciona mecanismos para ponerse en contacto con la empresa?	<b>N/A</b>
¿Se proporciona información sobre la protección de datos de carácter personal de los clientes o los derechos de autor de los contenidos del sitio web?	<b>N/A</b>
En artículos, noticias, informes... ¿Se muestra claramente información sobre el autor, fuentes y fechas de creación y revisión del documento?	<b>N/A</b>
¿El sitio web habla el mismo lenguaje que sus usuarios?	<b>SI</b>
¿Emplea un lenguaje claro y conciso?	<b>SI</b>
¿Es amigable, familiar y cercano?	<b>SI</b>
¿1 párrafo = 1 idea?	<b>N/A</b>
Los rótulos, ¿son significativos?	<b>SI</b>
¿Usa rótulos estándar?	<b>N/A</b>
¿Usa un único sistema de organización, bien definido y claro? No se deben mezclar diferentes. Los sistemas de organización son: alfabético, geográfico, cronológico, temático, orientado a tareas, orientado al público y orientado a metáforas.	<b>SI</b>

¿Utiliza un sistema de rotulado controlado y preciso?	<b>SI</b>
El título de las páginas, ¿Es correcto? ¿Ha sido planificado?	<b>SI</b>
La estructura de organización y navegación, ¿Es la más adecuada? ..	<b>SI</b>
En el caso de estructura jerárquica, ¿Mantiene un equilibrio entre Profundidad y Anchura?	<b>SI</b>
¿Los enlaces son fácilmente reconocibles como tales? ¿Su caracterización indica su estado (visitados, activos,...)? Los enlaces no sólo deben reconocerse como tales, sino que su caracterización debe indicar su estado, y ser reconocidos como una unidad	<b>NO</b>
En menús de navegación, ¿Se ha controlado el número de elementos y de términos por elemento para no producir sobrecarga memorística?	<b>SI</b>
¿Es predecible la respuesta del sistema antes de hacer clic sobre el enlace?	<b>SI</b>
¿Se ha controlado que no haya enlaces que no lleven a ningún sitio?	<b>NO</b>
¿Existen elementos de navegación que orienten al usuario acerca de dónde está y cómo deshacer su navegación?	<b>SI</b>
Las imágenes enlace, ¿se reconocen como clicables? ¿Incluyen un atributo 'title' describiendo la página de destino?	<b>N/A</b>
¿Se ha evitado la redundancia de enlaces?	<b>SI</b>
¿Se ha controlado que no haya páginas "huérfanas"?	<b>SI</b>
¿Se aprovechan las zonas de alta jerarquía informativa de la página para contenidos de mayor relevancia?	<b>SI</b>
¿Se ha evitado la sobrecarga informativa?	<b>SI</b>
¿Es una interfaz limpia, sin ruido visual?	<b>SI</b>

¿Existen zonas en "blanco" entre los objetos informativos de la página para poder descansar la vista?	<b>SI</b>
¿Se hace un uso correcto del espacio visual de la página?	<b>SI</b>
¿Se utiliza correctamente la jerarquía visual para expresar las relaciones del tipo "parte de" entre los elementos de la página?	<b>SI</b>
¿Se ha controlado la longitud de página?	<b>SI</b>
¿El tamaño de fuente se ha definido de forma relativa, o por lo menos, la fuente es lo suficientemente grande como para no dificultar la legibilidad del texto?	<b>SI</b>
¿El tipo de fuente, efectos tipográficos, ancho de línea y alineación empleadas facilitan la lectura?	<b>SI</b>
¿Existe un alto contraste entre el color de fuente y el fondo?	<b>SI</b>
¿Incluyen las imágenes atributos 'alt' que describan su contenido?	<b>N/A</b>
¿Es compatible el sitio web con los diferentes navegadores? ¿Se visualiza correctamente con diferentes resoluciones de pantalla?	<b>SI</b>
¿Puede el usuario disfrutar de todos los contenidos del sitio web sin necesidad de tener que descargar e instalar <i>plugins</i> adicionales?	<b>SI</b>
¿Se ha controlado el peso de la página?	<b>N/A</b>
¿Se puede imprimir la página sin problemas?	<b>NO</b>
¿Tiene el usuario todo el control sobre el interfaz?	<b>SI</b>
¿Se informa constantemente al usuario acerca de lo que está pasando?	<b>SI</b>
¿Se informa al usuario de lo que ha pasado?	<b>NO</b>
Cuando se produce un error, ¿se informa de forma clara y no alarmista al usuario de lo ocurrido y de cómo solucionar el problema?	<b>SI</b>

¿Posee el usuario libertad para actuar?	<b>SI</b>
¿Se ha controlado el tiempo de respuesta?	<b>SI</b>
¿Se ha evaluado adecuadamente la orientación del usuario? (Donde estoy, como volver, que he visitado, que va a pasar)	<b>NO</b>
¿Se ha usado correctamente la publicidad?	<b>N/A</b>

### 8.3.1.8 Cambios derivados de las pruebas

Tras la realización de las pruebas, teniendo en consideración lo extraído de los formularios y las observaciones, se decidió introducir los siguientes cambios en la interfaz:

1. Introducir indicaciones de los significados de los campos de 'Recursos por nivel', 'Iteraciones' y 'Parámetros adicionales' en los formularios de trabajos.
2. Introducir algunas migas de pan en lugares donde no estaban presentes (en los formularios de creación y/o modificación de elementos).
3. Revisar la configuración de la aplicación para evitar problemas de codificación (por ejemplo, ante la presencia de palabras con tilde).
4. Modificar la forma de mostrar la salida de las operaciones con servicios mediante un desplegable.
5. Evitar que el logotipo de WebKumo sea accionable desde la página de vista de situación general (enlace a sí misma).
6. Introducción de mensajes que indiquen que una operación fue exitosa (tras realizar operaciones con usuarios/grupos y trabajos).

## 8.3.2 Pruebas de accesibilidad

Esta sección tendrá como propósito, asegurar unas mínimas normas de accesibilidad web en la aplicación *WebKumo* desarrollada. Se efectuará la evaluación de la accesibilidad según las indicaciones WAI-W3C adaptado a WCAG 2.0 para un nivel de conformidad AA.

### 8.3.2.1 Revisión preliminar

La revisión preliminar constituye un primer acercamiento al análisis de la usabilidad del sitio mediante el análisis manual y semiautomático de las características básicas a analizar.

1. **Elección de un grupo significativo de páginas:** Estas serán las siguientes:
  - ◆ Página de vista de situación general (*dashboard*).
  - ◆ Página de vista de un trabajo y sus resultados.
  - ◆ Panel de servicios.
  - ◆ Página de creación de un trabajo.
  - ◆ Panel del usuario administrador.
2. **Examen de las páginas con un navegador de texto:** Se utilizará Lynx [**LYNX**].
3. **Visualización de las páginas con varios navegadores:** Se utilizarán tres navegadores gráficos, en concreto aquellos tres con mas cuota de mercado [**BROSH**]:
  - ◆ Microsoft Internet Explorer 11 [**MSIEX**]
  - ◆ Google Chrome 43 [**GOCHR**]
  - ◆ Mozilla Firefox 38 [**MZFIR**]
4. **Resultados**

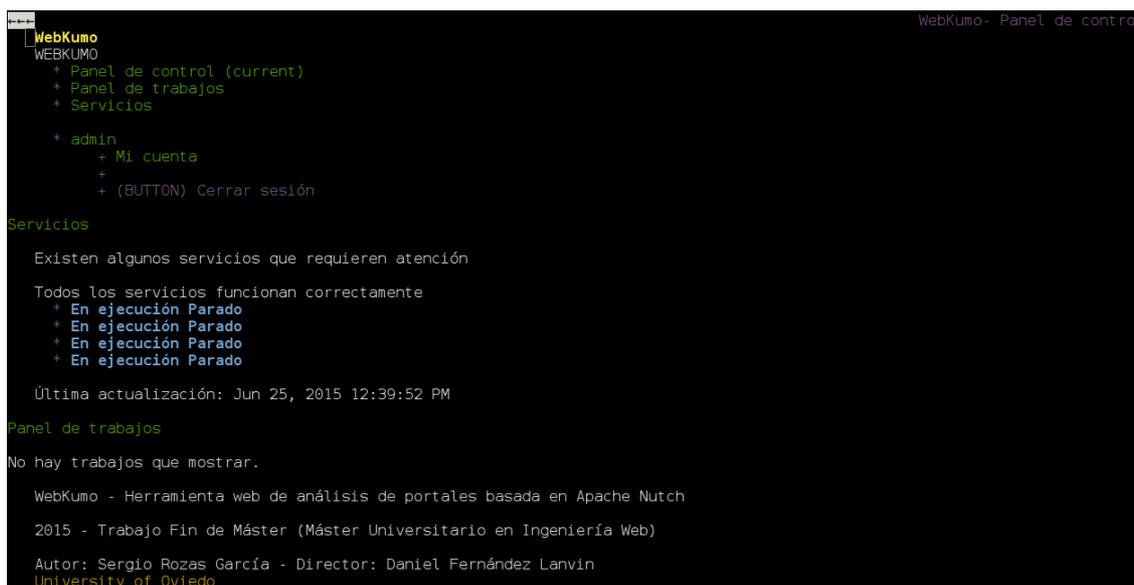
### 8.3.2.2 Evaluación de conformidad

1. **Alcance de la evaluación:** Se pretende alcanzar un nivel de conformidad AA según las normas WCAG 2.0.

Se evaluarán las páginas mas relevantes de la aplicación, siendo estas:

- ◆ Página de vista de situación general (*dashboard*).
  - ◆ Página de vista de un trabajo y sus resultados.
  - ◆ Panel de servicios.
  - ◆ Página de creación de un trabajo.
  - ◆ Panel del usuario administrador.
2. **Herramientas de evaluación automática:** Debido a que la herramienta desarrollada requiere autenticación, se hace imposible la ejecución de herramientas automáticas de evaluación. Se realizará entonces la validación del código HTML y CSS mediante el validador sintáctico **[W3HTMLV]** del W3C que permite como entrada el código fuente de los documentos.
  3. **Evaluación manual:** Se utilizará la lista de puntos a verificar del WCAG.

### 8.3.2.2.1 Revisión preliminar de la página de situación general.



```
WebKumo - Herramienta web de análisis de portales basada en Apache Nutch
2015 - Trabajo Fin de Máster (Máster Universitario en Ingeniería Web)
Autor: Sergio Rozas García - Director: Daniel Fernández Lanvin
University of Oviedo

WebKumo
WEBKUMO
+ Panel de control (current)
+ Panel de trabajos
+ Servicios

+ admin
+ Mi cuenta
+
+ (BUTTON) Cerrar sesión

Servicios
Existen algunos servicios que requieren atención

Todos los servicios funcionan correctamente
* En ejecución Parado
* En ejecución Parado
* En ejecución Parado
* En ejecución Parado

Última actualización: Jun 25, 2015 12:39:52 PM

Panel de trabajos
No hay trabajos que mostrar.

WebKumo - Herramienta web de análisis de portales basada en Apache Nutch
2015 - Trabajo Fin de Máster (Máster Universitario en Ingeniería Web)
Autor: Sergio Rozas García - Director: Daniel Fernández Lanvin
University of Oviedo
```

**Fig. 77:** Página de situación general visualizada mediante un navegador de texto.

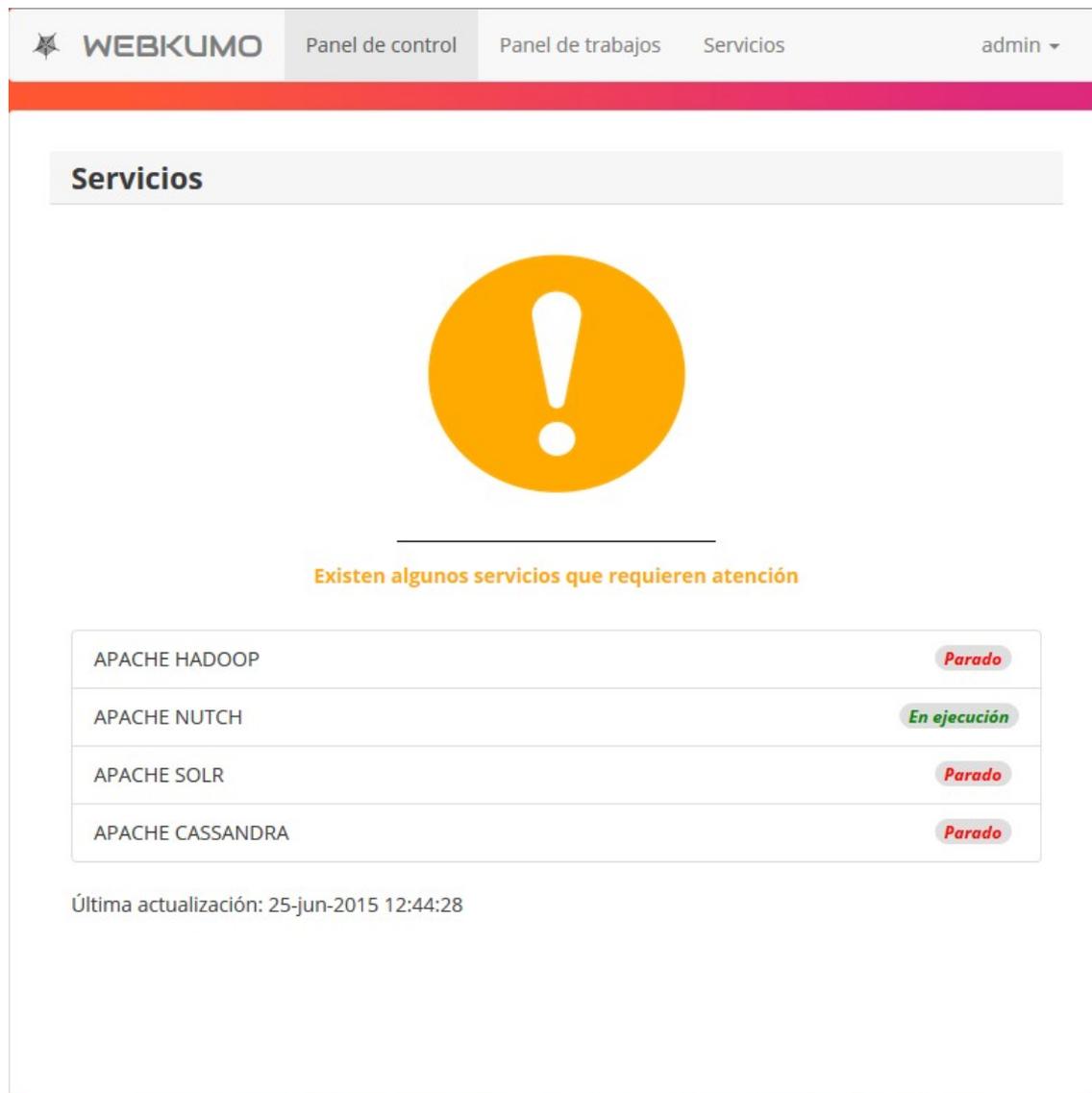
The screenshot displays the WebKumo dashboard with a navigation bar at the top containing 'WebKumo WEBKUMO', 'Panel de control', 'Panel de trabajos', 'Servicios', and a user profile 'admin'. The main content is split into two columns. The left column, titled 'Servicios', shows a warning 'Existen algunos servicios que requieren atención' and a table of service statuses: APACHE HADOOP (Parada), APACHE NUTCH (En ejecución), APACHE SOLR (Parada), and APACHE CASSANDRA (Parada). The right column, titled 'Panel de trabajos', lists 10 sample jobs with their completion status (FINISHED or STOPPED). The footer contains the title 'WebKumo - Herramienta web de análisis de portales basada en Apache Nutch', the author 'Sergio Rozas García', and the affiliation 'University of Oviedo'.

Servicios	
Existen algunos servicios que requieren atención	
Existen algunos servicios que requieren atención	
APACHE HADOOP	Parada
APACHE NUTCH	En ejecución
APACHE SOLR	Parada
APACHE CASSANDRA	Parada
Última actualización: 25-jun-2015 12:33:01	

Panel de trabajos	
#1 Sample Job 10 : 25/6/2015 2:00:00 (user22)	FINISHED
#2 Sample Job 20 : 25/6/2015 2:00:00 (user23)	STOPPED
#3 Sample Job 11 : 25/6/2015 2:00:00 (user20)	FINISHED
#4 Sample Job 21 : 25/6/2015 2:00:00 (user12)	STOPPED
#5 Sample Job 12 : 25/6/2015 2:00:00 (user11)	FINISHED
#6 Sample Job 22 : 25/6/2015 2:00:00 (user20)	STOPPED
#7 Sample Job 13 : 25/6/2015 2:00:00 (user12)	FINISHED
#8 Sample Job 23 : 25/6/2015 2:00:00 (user24)	STOPPED
#9 Sample Job 14 : 25/6/2015 2:00:00 (user12)	FINISHED
#10 Sample Job 24 : 25/6/2015 2:00:00 (user24)	STOPPED

WebKumo - Herramienta web de análisis de portales basada en Apache Nutch  
 2015 - Trabajo Fin de Máster (Máster Universitario en Ingeniería Web)  
 Autor: Sergio Rozas García - Director: Daniel Fernández Lanvín  
 University of Oviedo

**Fig. 78: Página de situación general sin imágenes.**



**Fig. 79: Página de situación general con ancho de dispositivo móvil (720x1280px).**

WEBKUMO

- [Panel de control \(current\)](#)
- [Panel de trabajos](#)
- [Servicios](#)
- [admin](#)
  - [Mi cuenta](#)
  - 
  -

[Servicios](#)



Existen algunos servicios que requieren atención

Todos los servicios funcionan correctamente

**Fig. 80: Página de situación general sin estilos.**

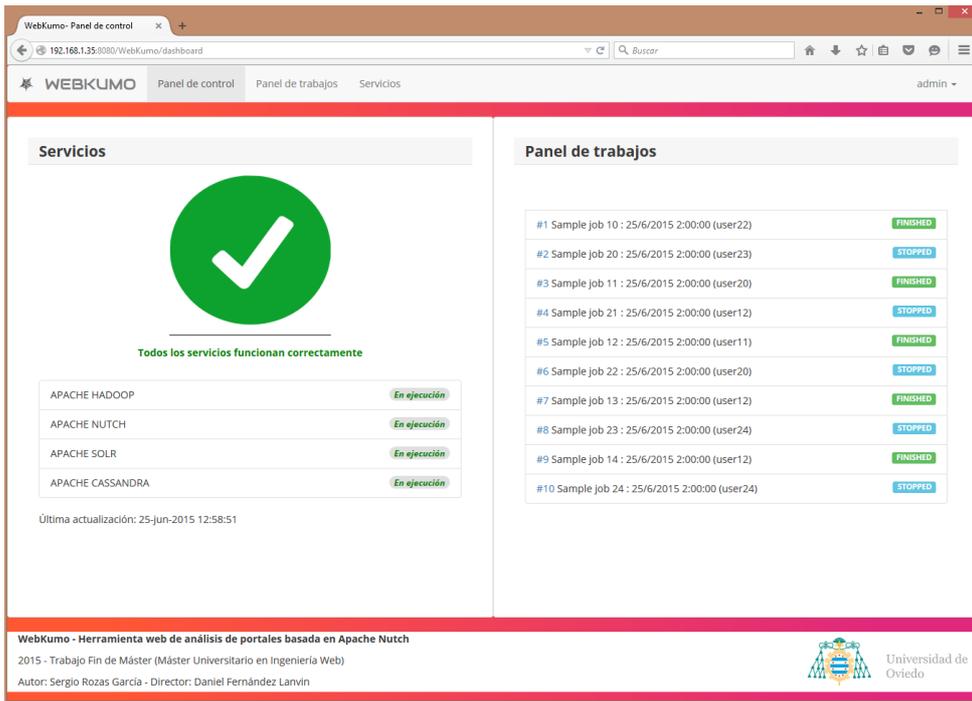


Fig. 81: Página de situación general renderizada por Mozilla Firefox

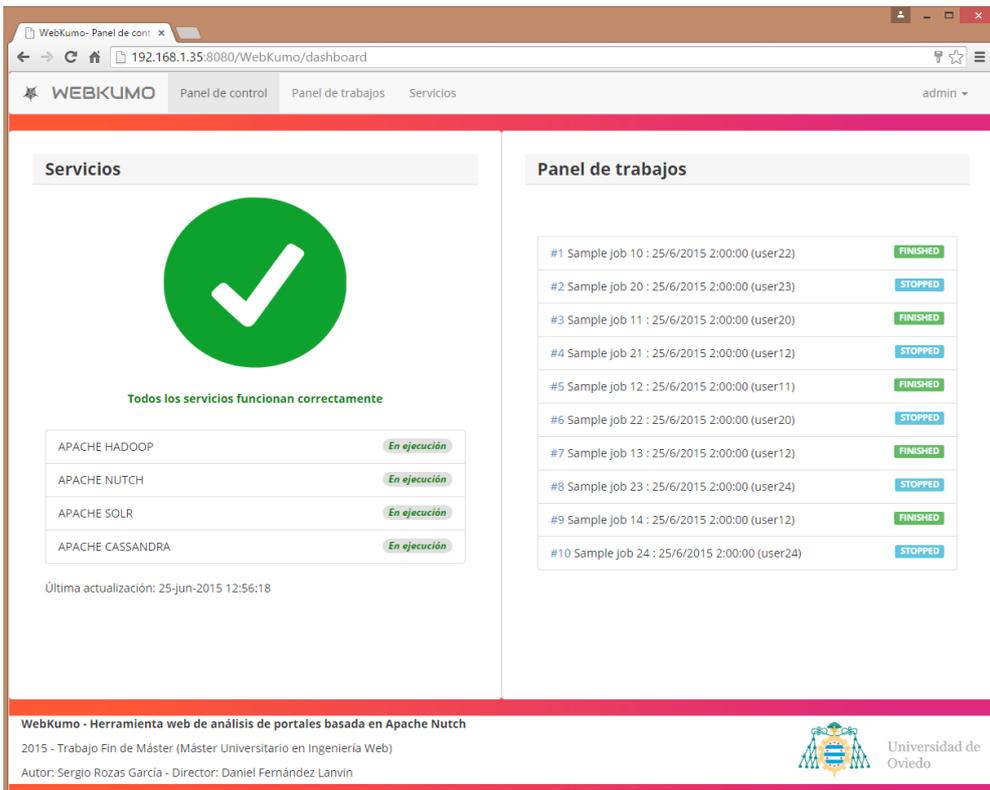


Fig. 82: Página de situación general renderizada por Google Chrome

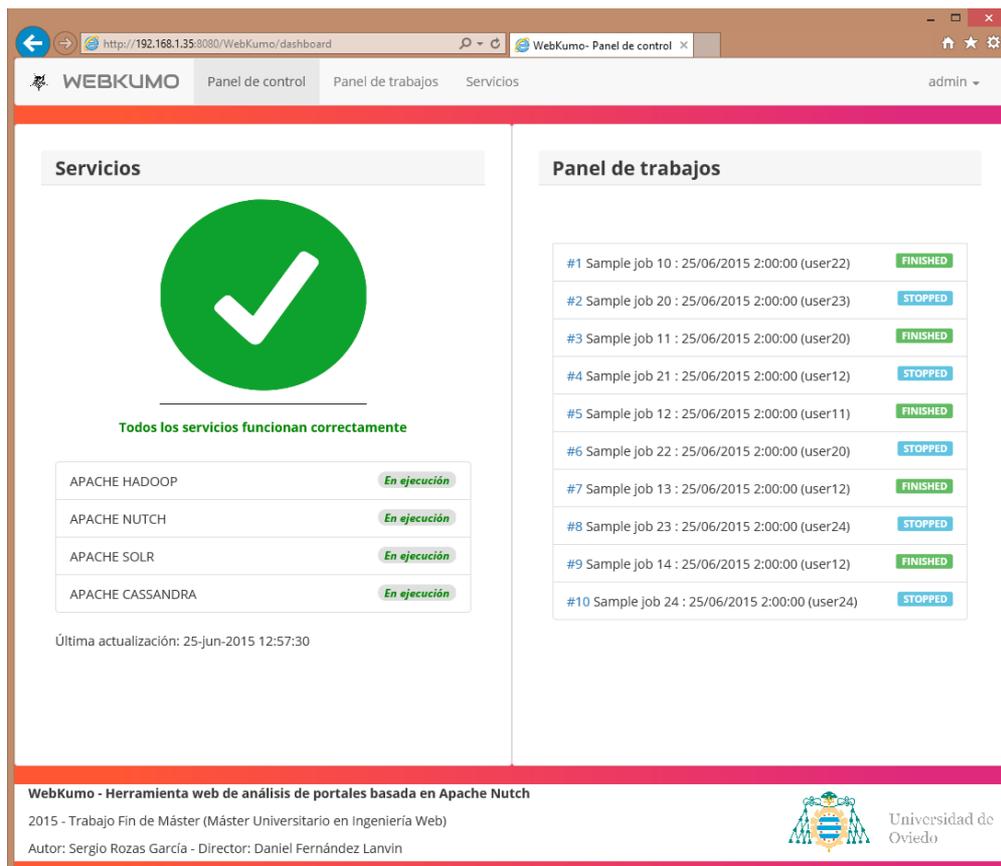


Fig. 83: Página de situación general renderizada por Microsoft Internet Explorer

### 8.3.2.2 Revisión de la página de vista de un trabajo

```

WebKumo
WEBKUMO
* Panel de control
* Panel de trabajos
* Servicios

* admin
  + Mi cuenta
  +
  + (BUTTON) Cerrar sesión

1. Panel de trabajos
2. Trabajo

Trabajo # 9

* Nombre del trabajo: Sample job 14
* Descripción del trabajo: This is a sample job
* Propietario del trabajo: user12 (John Doe)
* Fecha de creación: Jun 25, 2015
* Iteraciones: 7
* Recursos por nivel: 13
* Agente de usuario: Sin establecer. Se utilizará el valor presente en nutch-site.xml
* Cabecera Accept-Language HTTP: Sin establecer. Se utilizará el valor presente en nutch-site.xml
* Fecha de ejecución: No ejecutado todavía
* Estado: Finalizado
* URLs asociadas al trabajo
*
  1. http://ingenieriainformatica.uniovi.es/
  2. http://epigijon.uniovi.es/
* Análisis a realizar
*
  1. MockAnalysis (Análisis de prueba que extrae las URL y los nombres de los recursos.)
* Parámetros adicionales del trabajo:
*
  
```

**Fig. 84:** Página de ficha de trabajo visualizada mediante navegador de texto.

WebKumo WEBKUMO		Panel de control	Panel de trabajos	Servicios	admin ▾
Panel de trabajos / Trabajo					
<b>Trabajo # 9</b>					
<b>Nombre del trabajo:</b>	Sample job 14				
<b>Descripción del trabajo:</b>	This is a sample job				
<b>Propietario del trabajo:</b>	user12 (John Doe)				
<b>Fecha de creación:</b>	25-jun-2015				
<b>Iteraciones:</b>	7				
<b>Recursos por nivel:</b>	13				
<b>Agente de usuario:</b>	Sin establecer. Se utilizará el valor presente en nutch-site.xml				
<b>Cabecera Accept-Language HTTP:</b>	Sin establecer. Se utilizará el valor presente en nutch-site.xml				
<b>Fecha de ejecución:</b>	No ejecutado todavía				

**Fig. 85:** Página de ficha de trabajo sin imágenes.

 <b>WEBKUMO</b>	Panel de control	Panel de trabajos	Servicios	admin ▾
Panel de trabajos / Trabajo				
<b>Trabajo # 9</b>				
<b>Nombre del trabajo:</b> Sample job 14				
<b>Descripción del trabajo:</b> This is a sample job				
<b>Propietario del trabajo:</b> user12 (John Doe)				
<b>Fecha de creación:</b> 25-jun-2015				
<b>Iteraciones:</b> 7				
<b>Recursos por nivel:</b> 13				
<b>Agente de usuario:</b> Sin establecer. Se utilizará el valor presente en nutch-site.xml				
<b>Cabecera Accept-Language HTTP:</b> Sin establecer. Se utilizará el valor presente en nutch-site.xml				
<b>Fecha de ejecución:</b> No ejecutado todavía				
<b>Estado:</b> Finalizado				
<b>URLs asociadas al trabajo</b>				
1. http://ingenieriainformatica.uniovi.es/ 2. http://epigijon.uniovi.es/				
<b>Análisis a realizar</b>				
1. MockAnalysis (Análisis de prueba que extrae las URL y los nombres de los recursos.)				
<b>Parámetros adicionales del trabajo:</b>				
1. TEST_CUSTOM_PARAMETER : VALUE				

**Fig. 86:** Página de ficha de trabajo a una resolución típica de dispositivos móviles (720x1080px)

## WEBKUMO

- [Panel de control](#)
- [Panel de trabajos](#)
- [Servicios](#)
  
- [admin](#)
  - [Mi cuenta](#)
  - 
  -

1. [Panel de trabajos](#)
2. Trabajo

### Trabajo # 9

- Nombre del trabajo: Sample job 14
- Descripción del trabajo: This is a sample job
- Propietario del trabajo: user12 (John Doe)
- Fecha de creación: 25-jun-2015
- Iteraciones: 7
- Recursos por nivel: 13
- Agente de usuario: Sin establecer. Se utilizará el valor presente en nutch-site.xml
- Cabecera Accept-Language HTTP: Sin establecer. Se utilizará el valor presente en nutch-site.xml
- Fecha de ejecución: No ejecutado todavía
- Estado: Finalizado
- URLs asociadas al trabajo
  - 1. <http://ingenieriainformatica.uniovi.es/>
  - 2. <http://epigijon.uniovi.es/>
- Análisis a realizar
  - 1. MockAnalysis (Análisis de prueba que extrae las URL y los nombres de los recursos.)
- Parámetros adicionales del trabajo:
  - 1. TEST\_CUSTOM\_PARAMETER : VALUE

### Results

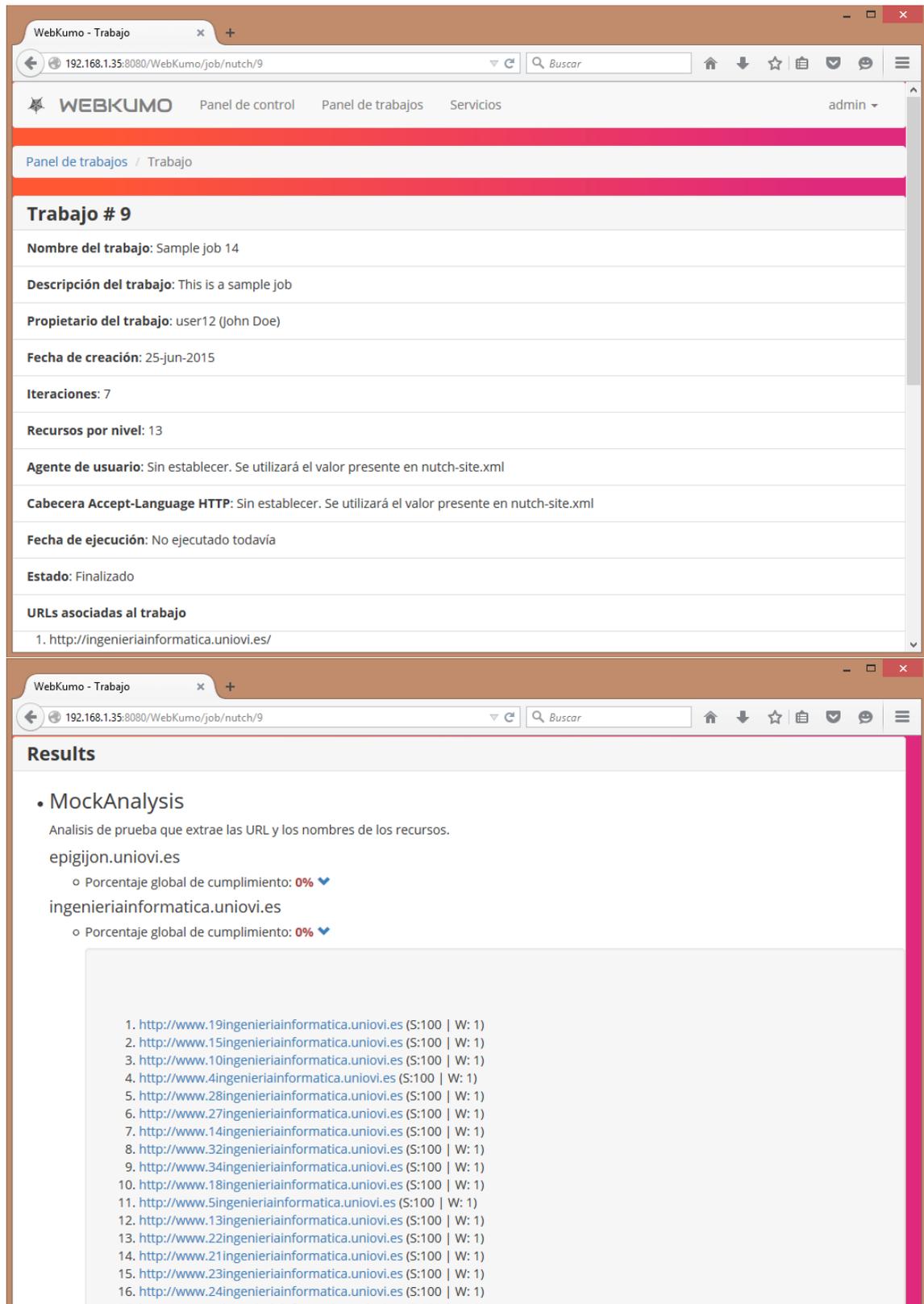
- **MockAnalysis**

Análisis de prueba que extrae las URL y los nombres de los recursos.

**epigijon.uniovi.es**

- Porcentaje global de cumplimiento: 0%

**Fig. 87: Página de ficha de trabajo sin hojas de estilos aplicadas**



**Fig. 88:** Página de ficha de trabajo renderizada por Mozilla Firefox

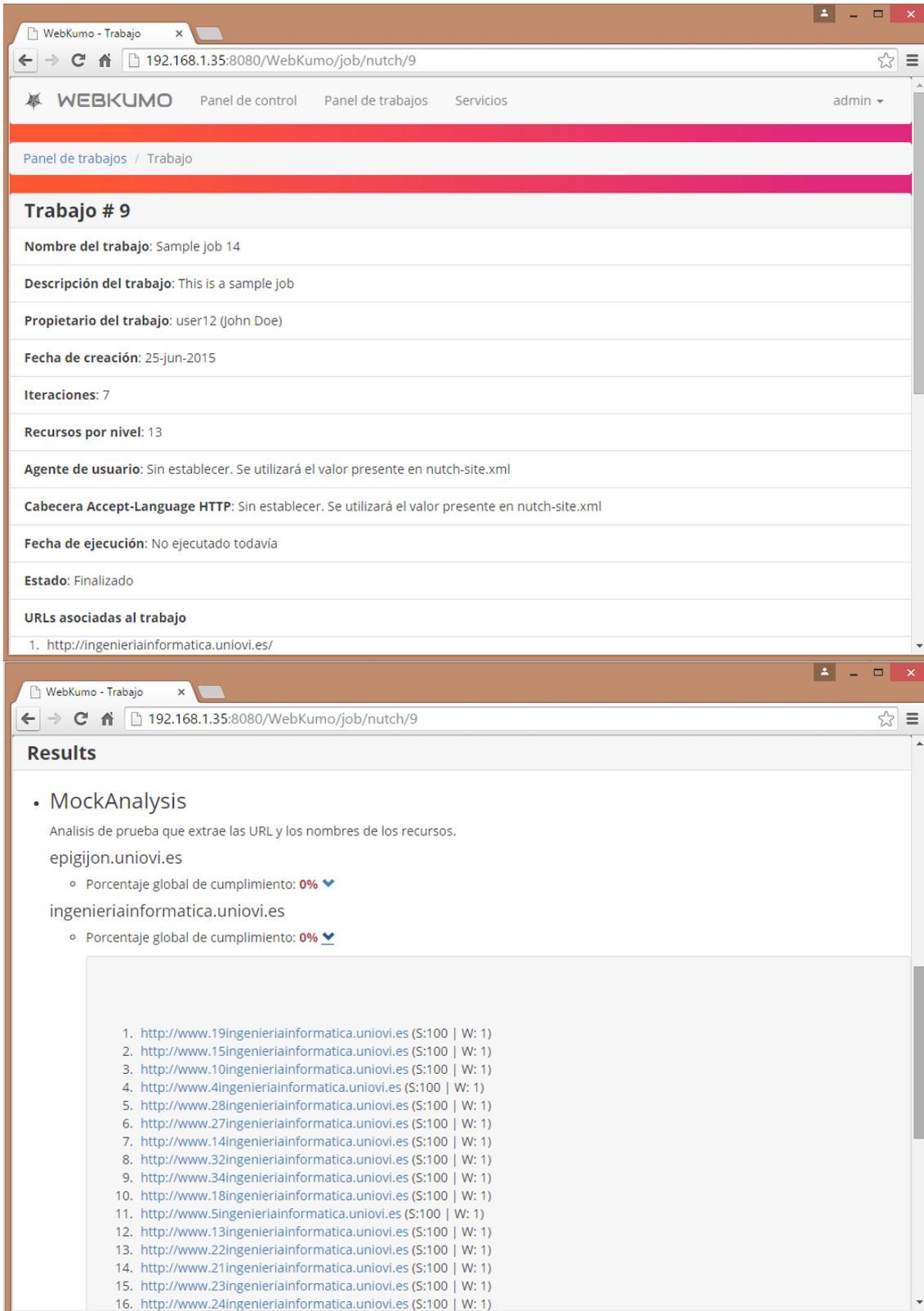


Fig. 89: Página de ficha de trabajo renderizada por Google Chrome

The screenshot displays the WebKumo interface in Microsoft Internet Explorer. The browser address bar shows the URL `http://192.168.1.35:8080/WebKumo/job/nutch/9`. The page header includes the WebKumo logo and navigation links: 'Panel de control', 'Panel de trabajos', and 'Servicios'. The user is logged in as 'admin'.

The main content area is titled 'Trabajo # 9' and contains the following details:

- Nombre del trabajo:** Sample job 14
- Descripción del trabajo:** This is a sample job
- Propietario del trabajo:** user12 (John Doe)
- Fecha de creación:** 25-jun-2015
- Iteraciones:** 7
- Recursos por nivel:** 13
- Agente de usuario:** Sin establecer. Se utilizará el valor presente en nutch-site.xml
- Cabecera Accept-Language HTTP:** Sin establecer. Se utilizará el valor presente en nutch-site.xml
- Fecha de ejecución:** No ejecutado todavía
- Estado:** Finalizado

Under the 'URLs asociadas al trabajo' section, a text input field contains the URL `http://192.168.1.35:8080/WebKumo/dashboard/es/`.

The 'Results' section is titled 'MockAnalysis' and includes the following information:

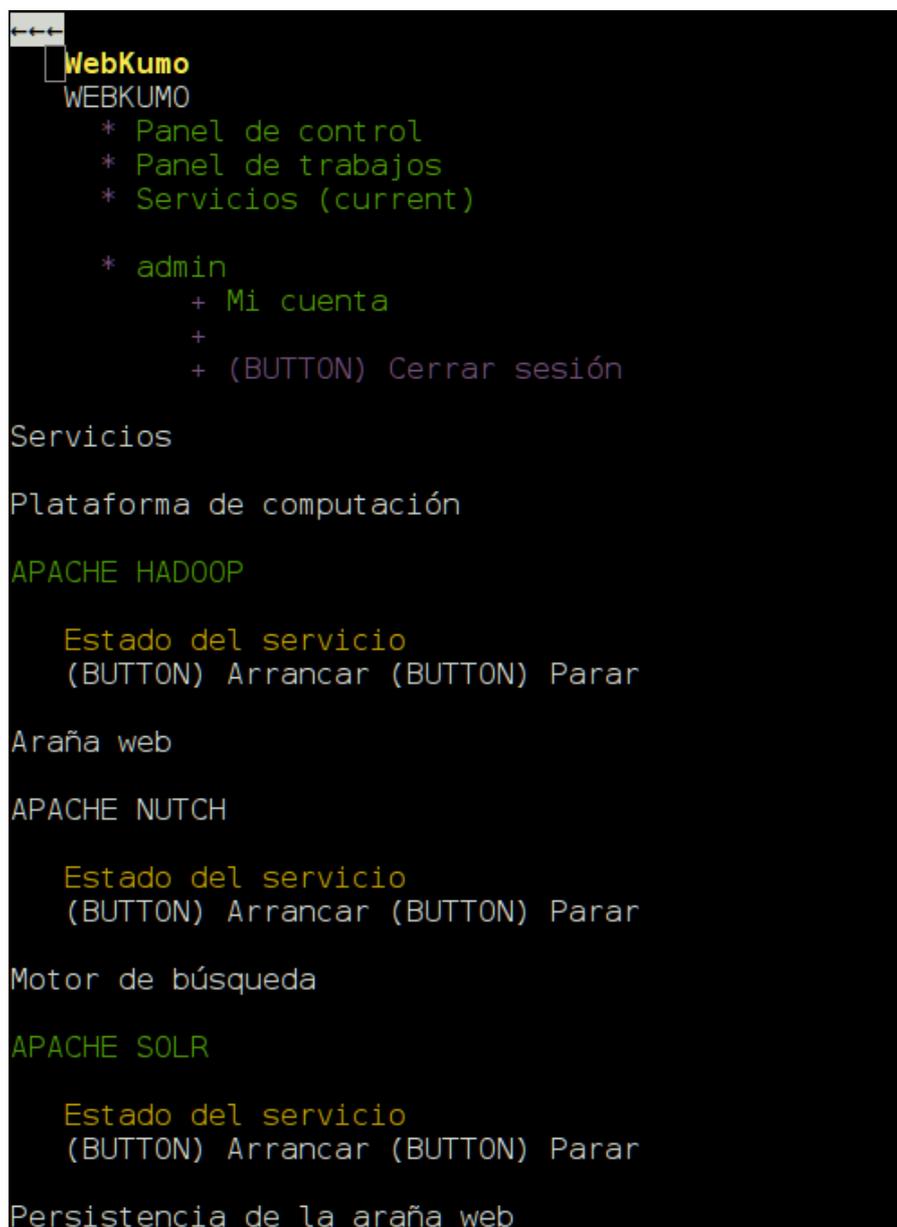
- Analisis de prueba que extrae las URL y los nombres de los recursos.
- epigijon.uniovi.es
- Porcentaje global de cumplimiento: 0%

A list of 20 URLs is displayed, each followed by its status and weight:

1. `http://www.0epigijon.uniovi.es` (S:100 | W: 1)
2. `http://www.19epigijon.uniovi.es` (S:100 | W: 1)
3. `http://www.5epigijon.uniovi.es` (S:100 | W: 1)
4. `http://www.3epigijon.uniovi.es` (S:100 | W: 1)
5. `http://www.4epigijon.uniovi.es` (S:100 | W: 1)
6. `http://www.17epigijon.uniovi.es` (S:100 | W: 1)
7. `http://www.24epigijon.uniovi.es` (S:100 | W: 1)
8. `http://www.2epigijon.uniovi.es` (S:100 | W: 1)
9. `http://www.27epigijon.uniovi.es` (S:100 | W: 1)
10. `http://www.10epigijon.uniovi.es` (S:100 | W: 1)
11. `http://www.15epigijon.uniovi.es` (S:100 | W: 1)
12. `http://www.23epigijon.uniovi.es` (S:100 | W: 1)
13. `http://www.11epigijon.uniovi.es` (S:100 | W: 1)
14. `http://www.14epigijon.uniovi.es` (S:100 | W: 1)
15. `http://www.34epigijon.uniovi.es` (S:100 | W: 1)
16. `http://www.25epigijon.uniovi.es` (S:100 | W: 1)
17. `http://www.1epigijon.uniovi.es` (S:100 | W: 1)
18. `http://www.22epigijon.uniovi.es` (S:100 | W: 1)
19. `http://www.6epigijon.uniovi.es` (S:100 | W: 1)
20. `http://www.28epigijon.uniovi.es` (S:100 | W: 1)

Fig. 90: Página de ficha de trabajo renderizada por Microsoft Internet Explorer

### 8.3.2.2.3 Revisión de la página de servicios



```
WebKumo
WEBKUMO
* Panel de control
* Panel de trabajos
* Servicios (current)

* admin
  + Mi cuenta
  +
  + (BUTTON) Cerrar sesión

Servicios

Plataforma de computación

APACHE HADOOP

  Estado del servicio
  (BUTTON) Arrancar (BUTTON) Parar

Araña web

APACHE NUTCH

  Estado del servicio
  (BUTTON) Arrancar (BUTTON) Parar

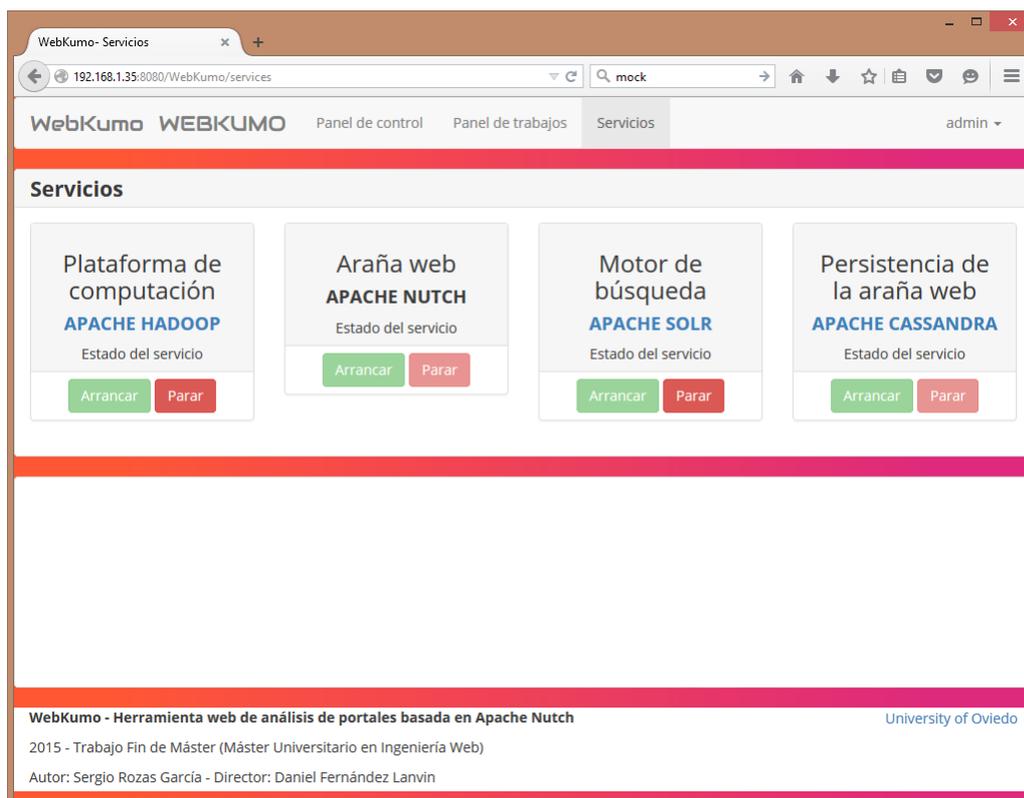
Motor de búsqueda

APACHE SOLR

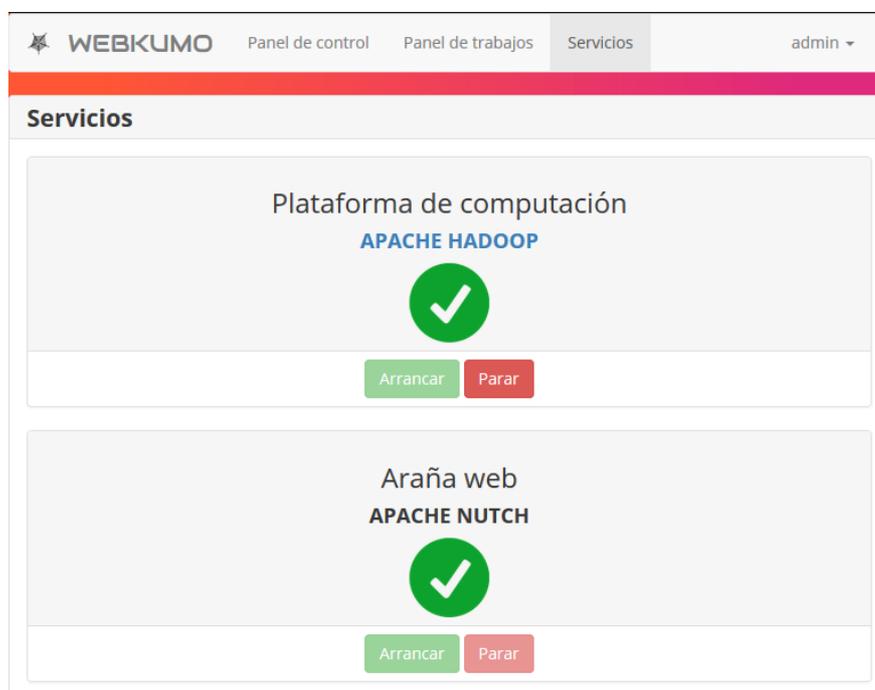
  Estado del servicio
  (BUTTON) Arrancar (BUTTON) Parar

Persistencia de la araña web
```

**Fig. 91:** *Página del panel de servicios visualizada en un navegador de texto*



**Fig. 92: Página del panel de servicios visualizada sin imágenes**



**Fig. 93: Página del panel de servicios a una resolución típica de dispositivos móviles (720x1280px)**

Servicios

Plataforma de computación

[APACHE HADOOP](#)



Arrancar Parar

*Fig. 94: Página del panel de servicios sin hojas de estilos aplicadas*

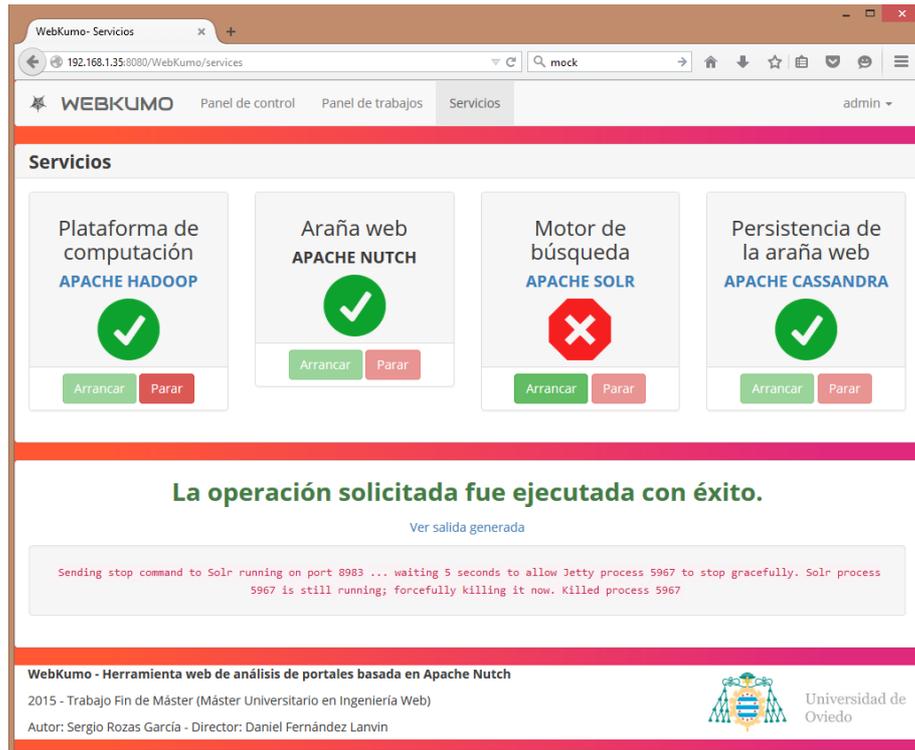


Fig. 95: Página del panel de servicios renderizada por Mozilla Firefox

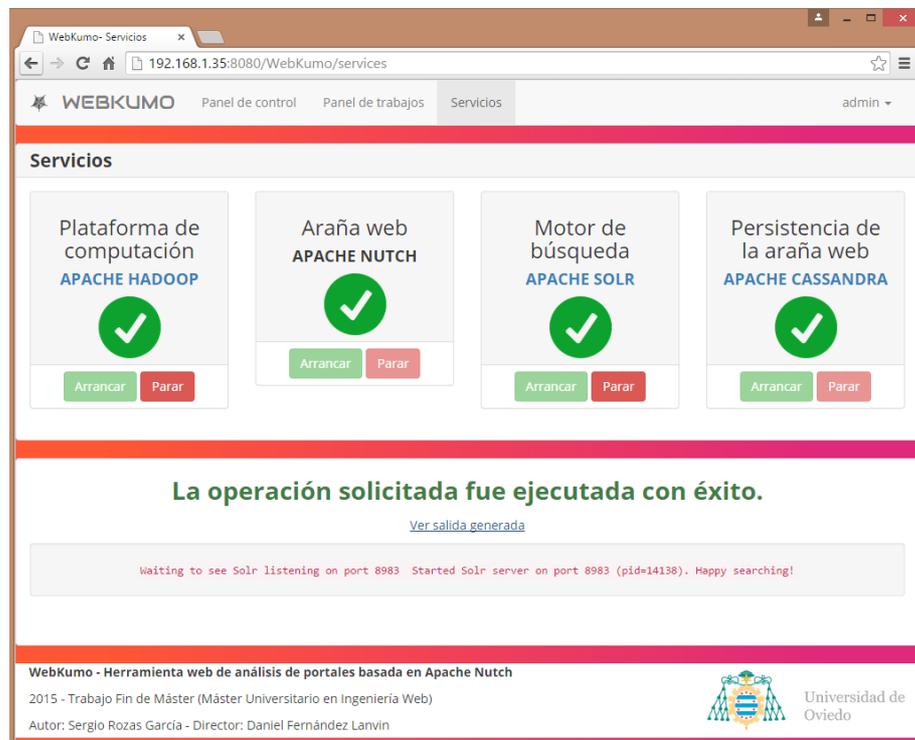


Fig. 96: Página del panel de servicios renderizada por Google Chrome

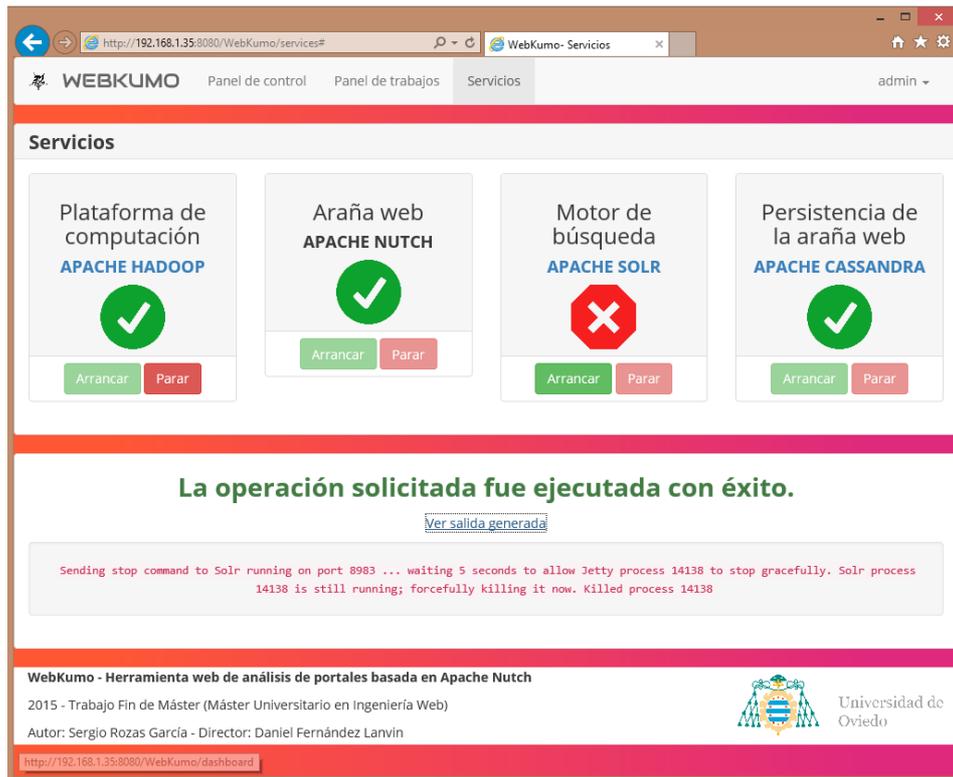


Fig. 97: Página del panel de servicios renderizada por Microsoft Internet Explorer

### 8.3.2.2.4 Revisión de la página de nuevo trabajo

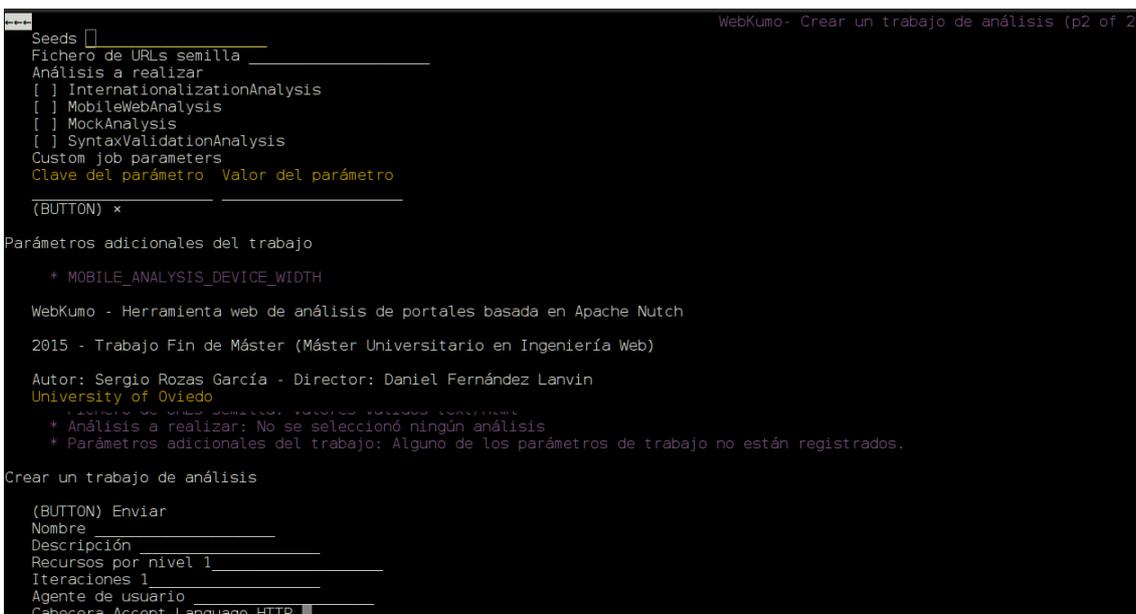


Fig. 98: Página de creación de un trabajo mostrada por un navegador de texto

WebKumo WEBKUMO Panel de control Panel de trabajos Servicios admin ▾

Panel de trabajos / Crear un trabajo de análisis

### Crear un trabajo de análisis Enviar

Nombre	Job 1				
Descripción	...				
Recursos por nivel	1				
Iteraciones	1				
Agente de usuario	WebKumo-0.1				
Cabecera Accept-Language HTTP	es-ES, en-us,en-gb,en;q=0.7,*;q=0.3				
Seeds	http://www.example.org				
Fichero de URLs semilla	<input type="button" value="Examinar..."/> No se ha seleccionado ningún archivo.				
Análisis a realizar	Ninguno ▾				
Custom job parameters <input type="button" value="Q+"/>	<table><thead><tr><th>Clave del parámetro</th><th>Valor del parámetro</th></tr></thead><tbody><tr><td><input type="text"/></td><td><input type="text"/></td></tr></tbody></table>	Clave del parámetro	Valor del parámetro	<input type="text"/>	<input type="text"/>
Clave del parámetro	Valor del parámetro				
<input type="text"/>	<input type="text"/>				

WebKumo - Herramienta web de análisis de portales basada en Apache Nutch University of Oviedo

2015 - Trabajo Fin de Máster (Máster Universitario en Ingeniería Web)

Autor: Sergio Rozas García - Director: Daniel Fernández Lanvin

**Fig. 99: Página de creación de un trabajo renderizada sin imágenes**

WEBKUMO Panel de control Panel de trabajos Servicios admin ▾

Panel de trabajos / Crear un trabajo de análisis

### Crear un trabajo de análisis Enviar

Nombre	Job 1				
Descripción	...				
Recursos por nivel	1				
Iteraciones	1				
Agente de usuario	WebKumo-0.1				
Cabecera Accept-Language HTTP	es-ES, en-us, en-gb, en; q=0.7; *; q=0.3				
Seeds	http://www.example.org				
Fichero de URLs semilla	<span>Examinar...</span> No se ha seleccionado ningún archivo.				
Análisis a realizar	Ninguno ▾				
Custom job parameters <span>Q+</span>	<table><thead><tr><th>Clave del parámetro</th><th>Valor del parámetro</th></tr></thead><tbody><tr><td><input type="text"/></td><td><input type="text"/></td></tr></tbody></table>	Clave del parámetro	Valor del parámetro	<input type="text"/>	<input type="text"/>
Clave del parámetro	Valor del parámetro				
<input type="text"/>	<input type="text"/>				

**WebKumo - Herramienta web de análisis de portales basada en Apache Nutch**  
2015 - Trabajo Fin de Máster (Máster Universitario en Ingeniería Web)  
Autor: Sergio Rozas García - Director: Daniel Fernández Lanvin

**Fig. 100: Página de creación de un trabajo a una resolución típica de dispositivo móvil (720x1080px).**

## Crear un trabajo de análisis

Enviar

Nombre

Descripción

Recursos por nivel

Iteraciones

Agente de usuario

Cabecera Accept-Language HTTP

Seeds

Fichero de URLs semilla  No se ha seleccionado ningún archivo.

Análisis a realizar

- 
- Seleccionar todos
- InternationalizationAnalysis
- MobileWebAnalysis
- MockAnalysis
- SyntaxValidationAnalysis

Custom job parameters

Clave del parámetro	Valor del parámetro
<input type="text"/>	<input type="text"/>

### Parámetros adicionales del trabajo

- MOBILE\_ANALYSIS\_DEVICE\_WIDTH

*Fig. 101: Página de creación de un trabajo sin hojas de estilo aplicadas*

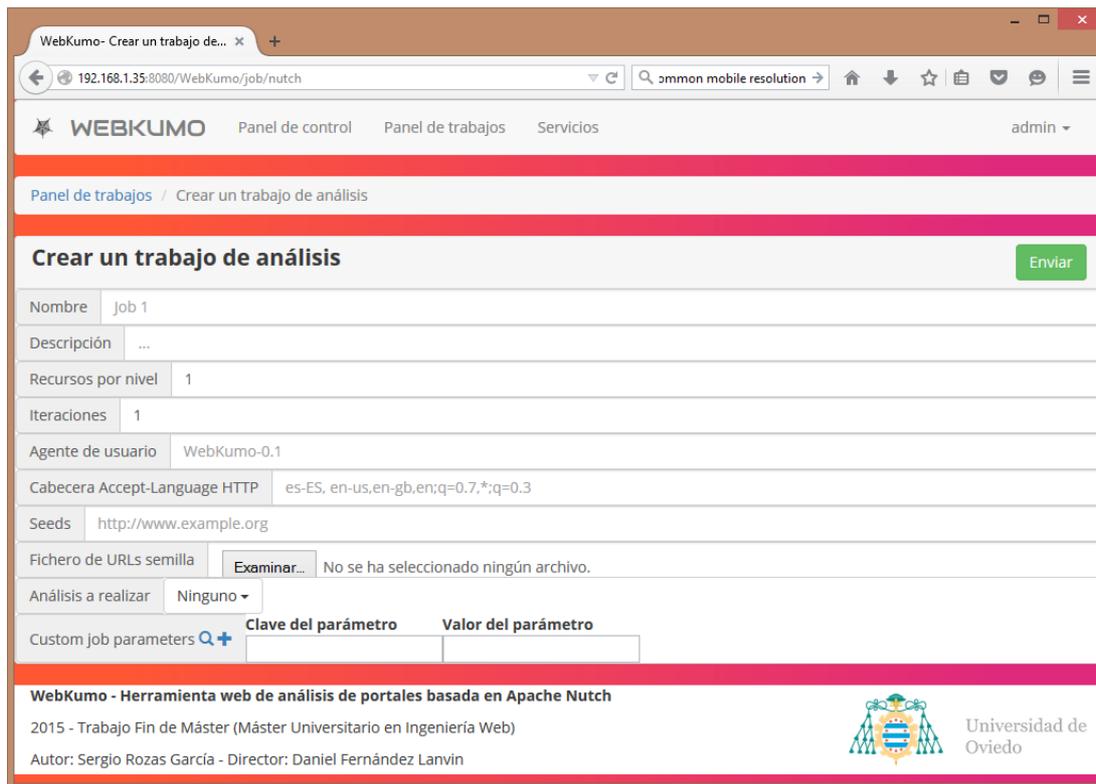


Fig. 102: Página de creación de un trabajo renderizada por Mozilla Firefox

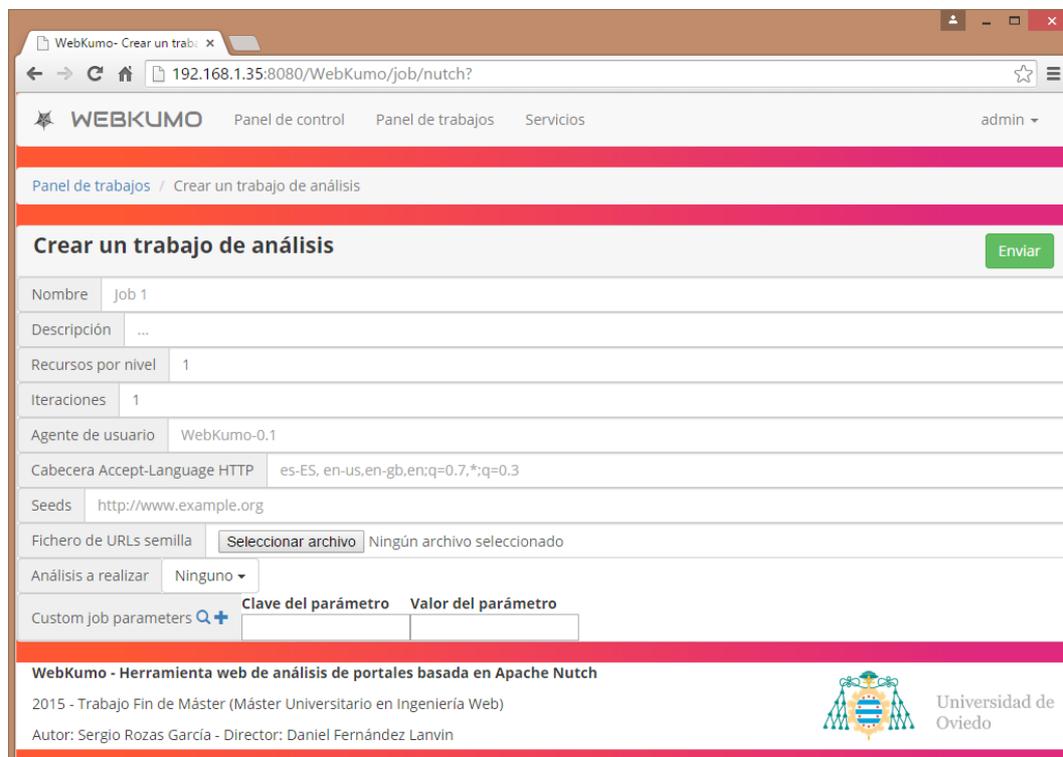


Fig. 103: Página de creación de un trabajo renderizada por Google Chrome

WEBKUMO Panel de control Panel de trabajos Servicios admin

Panel de trabajos / Crear un trabajo de análisis

**Crear un trabajo de análisis** Enviar

Nombre Job 1

Descripción ...

Recursos por nivel 1

Iteraciones 1

Agente de usuario WebKumo-0.1

Cabecera Accept-Language HTTP es-ES, en-us,en-gb,en;q=0.7,\*;q=0.3

Seeds http://www.example.org

Fichero de URLs semilla  Examinar...

Análisis a realizar Ninguno

Clave del parámetro	Valor del parámetro
<input type="text"/>	<input type="text"/>

WebKumo - Herramienta web de análisis de portales basada en Apache Nutch

2015 - Trabajo Fin de Máster (Máster Universitario en Ingeniería Web)

Autor: Sergio Rozas García - Director: Daniel Fernández Lanvin

Universidad de Oviedo

**Fig. 104:** Página de creación de un trabajo renderizada por Microsoft Internet Explorer

### 8.3.2.2.5 Revisión del panel del administrador

```

WebKumo - Gestión de cuentas de usuario
WebKumo
WEBKUMO
* Panel de control
* Panel de trabajos
* Servicios

* admin
+ Mi cuenta
+
+ (BUTTON) Cerrar sesión

Gestión de cuentas de usuario
(BUTTON) Crear un nuevo usuario
(BUTTON) Crear un nuevo grupo

Usuarios

Identificador Nombre de usuario Nombre completo Email Eliminar
(BUTTON) x

Eliminar

¿Está seguro?

La operación solicitada fue llevada a cabo con éxito

La operación solicitada no se pudo completar

Recuerde que un grupo sólo puede eliminarse si no posee miembros y un usuario si no posee trabajos.
(BUTTON) No (BUTTON) Si
(BUTTON) Volver

```

**Fig. 105:** Página del panel del usuario del administrador vista en un navegador textual

WebKumo WEBKUMO Panel de control Panel de trabajos Servicios admin ▾

### Gestión de cuentas de usuario

Crear un nuevo usuario  
Crear un nuevo grupo

#### Usuarios

Mostrar 10 registros Buscar:

Identificador	Nombre de usuario	Nombre completo	Email	Eliminar
#1	admin	Sergio Rozas	uo204054@uniovi.es	✕
#10	user12	John Doe	johndoe12@nowhere.org	✕
#11	user22	John Doe	johndoe22@nowhere.org	✕
#12	user32	John Doe	johndoe32@nowhere.org	✕
#13	user42	John Doe	johndoe42@nowhere.org	✕
#14	user13	John Doe	johndoe13@nowhere.org	✕
#15	user23	John Doe	johndoe23@nowhere.org	✕
#16	user33	John Doe	johndoe33@nowhere.org	✕
#17	user43	John Doe	johndoe43@nowhere.org	✕
#18	user14	John Doe	johndoe14@nowhere.org	✕

Mostrando registros del 1 al 10 de un total de 21 registros

Anterior 1 2 3 Siguiente

**Fig. 106: Página del panel del usuario del administrador sin imágenes**

WEBKUMO Panel de control Panel de trabajos Servicios admin ▾

## Gestión de cuentas de usuario

Crear un nuevo usuario

Crear un nuevo grupo

### Usuarios

Mostrar  registros

Buscar:

Identificador	Nombre de usuario	Nombre completo	Email	Eliminar
#1	admin	Sergio Rozas	uo204054@uniovi.es	✕
#10	user12	John Doe	johndoe12@nowhere.org	✕
#11	user22	John Doe	johndoe22@nowhere.org	✕
#12	user32	John Doe	johndoe32@nowhere.org	✕
#13	user42	John Doe	johndoe42@nowhere.org	✕
#14	user13	John Doe	johndoe13@nowhere.org	✕
#15	user23	John Doe	johndoe23@nowhere.org	✕
#16	user33	John Doe	johndoe33@nowhere.org	✕
#17	user43	John Doe	johndoe43@nowhere.org	✕
#18	user14	John Doe	johndoe14@nowhere.org	✕

Mostrando registros del 1 al 10 de un total de 21 registros

Anterior **1** 2 3 Siguiente

**Fig. 107:** Página del panel del administrador a una resolución típica de dispositivos móviles (720x1080px)

Gestión de cuentas de usuario

Crear un nuevo usuario

Crear un nuevo grupo

**Usuarios**

Mostrar  registros

Buscar:

Identificador	Nombre de usuario	Nombre completo	Email	Eliminar
<a href="#">#1</a>	admin	Sergio Rozas	uo204054@uniovi.es	
<a href="#">#10</a>	user12	John Doe	johndoe12@nowhere.org	
<a href="#">#11</a>	user22	John Doe	johndoe22@nowhere.org	
<a href="#">#12</a>	user32	John Doe	johndoe32@nowhere.org	
<a href="#">#13</a>	user42	John Doe	johndoe42@nowhere.org	
<a href="#">#14</a>	user13	John Doe	johndoe13@nowhere.org	
<a href="#">#15</a>	user23	John Doe	johndoe23@nowhere.org	
<a href="#">#16</a>	user33	John Doe	johndoe33@nowhere.org	
<a href="#">#17</a>	user43	John Doe	johndoe43@nowhere.org	
<a href="#">#18</a>	user14	John Doe	johndoe14@nowhere.org	

Mostrando registros del 1 al 10 de un total de 21 registros

- [Anterior](#)
- [1](#)
- [2](#)
- [3](#)
- [Siguiente](#)

**Eliminar**

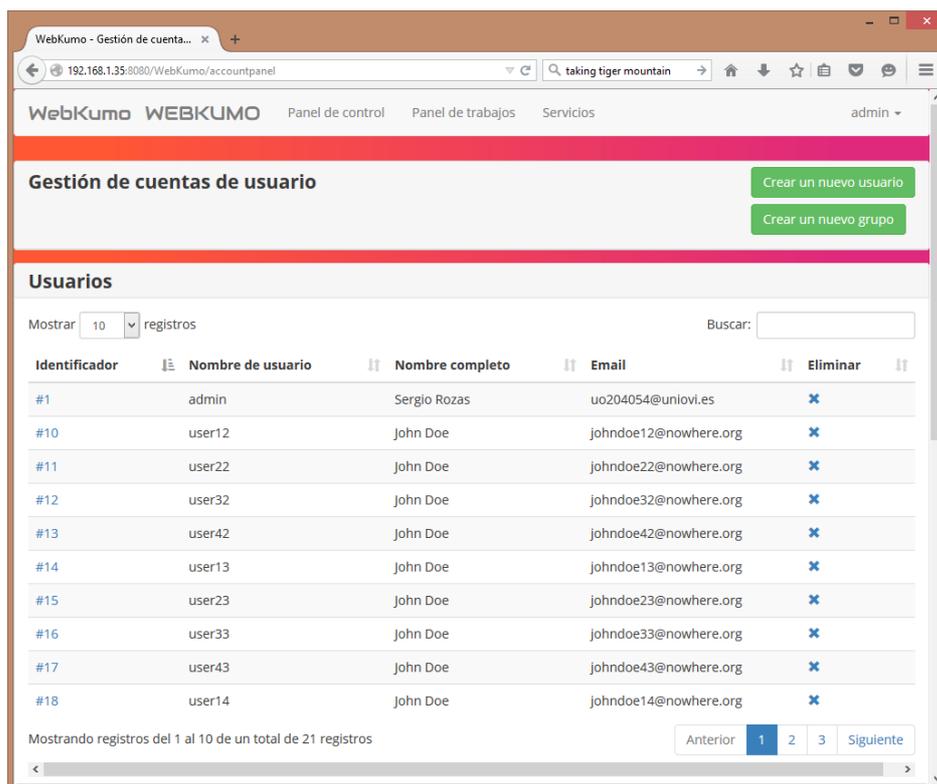
¿Está seguro?

La operación solicitada fue llevada a cabo con éxito

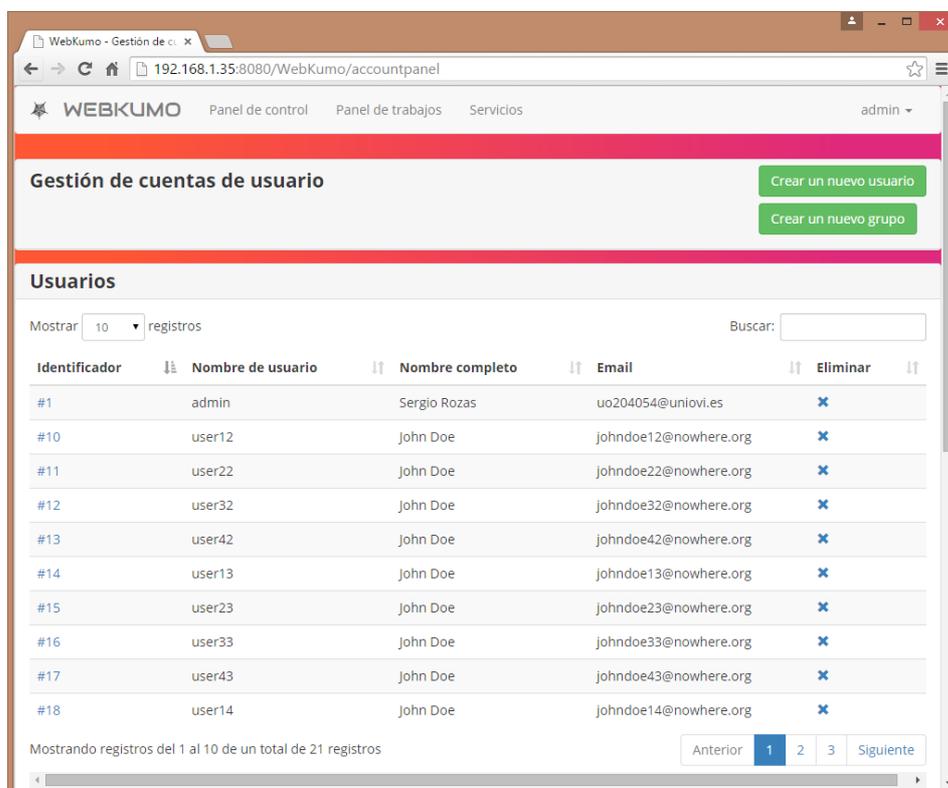
La operación solicitada no se pudo completar

Recuerde que un grupo sólo puede eliminarse si no posee miembros y un usuario si no posee trabajos.

**Fig. 108: Página del panel del administrador sin hojas de estilos aplicadas**



**Fig. 109: Página del panel del administrador renderizada por Mozilla Firefox**



**Fig. 110: Página del panel del administrador renderizada por Google Chrome**

The screenshot shows the WebKumo administrator interface. At the top, there's a navigation bar with 'WEBKUMO', 'Panel de control', 'Panel de trabajos', and 'Servicios'. The main heading is 'Gestión de cuentas de usuario'. Below this, there are two green buttons: 'Crear un nuevo usuario' and 'Crear un nuevo grupo'. A section titled 'Usuarios' contains a search bar and a table of users. The table has columns for 'Identificador', 'Nombre de usuario', 'Nombre completo', 'Email', and 'Eliminar'. The data rows show users with IDs #1 through #18, including an 'admin' user and several 'user' accounts with 'John Doe' as the full name and '@nowhere.org' email addresses. At the bottom, it indicates 'Mostrando registros del 1 al 10 de un total de 21 registros' and has pagination controls for 'Anterior', '1', '2', '3', and 'Siguiete'.

Identificador	Nombre de usuario	Nombre completo	Email	Eliminar
#1	admin	Sergio Rozas	uo204054@uniovi.es	X
#10	user12	John Doe	johndoe12@nowhere.org	X
#11	user22	John Doe	johndoe22@nowhere.org	X
#12	user32	John Doe	johndoe32@nowhere.org	X
#13	user42	John Doe	johndoe42@nowhere.org	X
#14	user13	John Doe	johndoe13@nowhere.org	X
#15	user23	John Doe	johndoe23@nowhere.org	X
#16	user33	John Doe	johndoe33@nowhere.org	X
#17	user43	John Doe	johndoe43@nowhere.org	X
#18	user14	John Doe	johndoe14@nowhere.org	X

**Fig. 111: Página del panel del administrador renderizada por Microsoft Internet Explorer**

### 8.3.2.3 Comentarios sobre la revisión preliminar

Tras realizar el análisis de las pantallas seleccionadas se comprobó que principalmente, derivado del uso de JavaScript, las páginas pierden funcionalidad al ser visitadas usando un navegador de texto. Por ejemplo en el caso de la validación de los formularios (aunque no funcione la validación mediante JavaScript, la validación se realiza igualmente desde el servidor) o el caso de la pantalla de situación general donde la lista de últimos trabajos se genera de forma dinámica

No existe una forma sencilla de eliminar los problemas de accesibilidad derivados del uso de AJAX para carga dinámica de elementos. En ocasiones, como en la pantalla de vista general, se podría sustituir la carga dinámica por procesamiento del lado del servidor; sin embargo, en la carga de los campos de un informe de un

análisis, no se podría sustituir la carga mediante AJAX por problemas de rendimiento (ver apartado 8.4).

Por la naturaleza de la aplicación, cerrada, de carácter interno a cierto personal de la Universidad de Oviedo y con un reducido grupo de usuarios, se considera más importante mantener el uso de AJAX asumiendo los inconvenientes de accesibilidad que introducen.

En cuanto a la presentación para casos especiales que se contemplaron en la revisión, se concluye que no existen problemas de relevancia si se prescinden de hojas de estilo, imágenes o si se utiliza un dispositivo móvil.

Se utilizaron muy pocas imágenes a lo largo de la aplicación (algo lógico tratándose de una herramienta) lo que implica que si se desactivan, la experiencia casi no se ve afectada. En cuanto a las hojas de estilos, el desactivarlas modifica sensiblemente la experiencia de usuario (por ejemplo con mensajes de error que sólo se muestran cuando se produce un error y sin hoja de estilo se muestran) pero por lo general el sitio sigue siendo usable. Pasando al uso de la aplicación en pantallas pequeñas, las pruebas con resolución 720x1080 píxeles (resolución HD típica de dispositivos móviles actuales) mostraron un resultado satisfactorio. Se ve en las capturas como, gracias al uso del framework CSS *Bootstrap*, los elementos se reordenan automáticamente al decrecer el ancho de la pantalla.

La prueba realizada con distintos navegadores web gráficos fue satisfactoria. Los tres navegadores utilizados renderizaron las distintas páginas correctamente sin apenas diferencias entre ellos. No se ha utilizado ningún código específico para ningún tipo de navegador web en el desarrollo de las interfaces.

Se comprobó también que la navegación haciendo uso únicamente del teclado fuera aceptable. Mediante el uso de la tecla tabulador se ha podido navegar por las páginas sin problemas.

## 8.3.2.3.1 Conformidad de la página de situación general

The screenshot shows the W3C Markup Validation Service interface. At the top, it says "This document was successfully checked as HTML5!". Below this, the "Result" is "Passed, 4 warning(s)". The "Source" section displays the HTML code of the document, including the doctype, head, title, meta tags, and link tags. The "Encoding" is set to "utf-8", the "Doctype" is "HTML5", and the "Root Element" is "html".

**W3C** Markup Validation Service  
Check the markup (HTML, XHTML, ...) of Web documents

Jump To: Notes and Potential Issues Validation Output Congratulations · Icons

This document was successfully checked as HTML5!

**Result:** Passed, 4 warning(s)

**Source:**

```
<!doctype HTML>
<html>
  <head>
    <title>WebKumo-
      Panel de Control</title>
    <meta charset="utf-8"/>
    <meta name="description" content="WebKumo - An Apache Nutch based, web analysis tool"/>
    <meta name="keywords" content="Web análisis, Apache Nutch"/>
    <meta name="author" content="Sergio Rozas - MIW - University of Oviedo"/>
    <meta name="viewport" content="width=device-width, initial-scale=1"/>
    <link rel="stylesheet"
      href="https://maxcdn.bootstrapcdn.com/bootstrap/3.3.1/css/bootstrap.min.css">
    <link rel="stylesheet" href="/WebKumo/stvles/base.css" type="text/css"/>
  </head>
  <body>
  </body>
</html>
```

**Encoding:** utf-8 (detect automatically)

**Doctype:** HTML5 (detect automatically)

**Root Element:** html

**W3C VALIDATOR Suite** Try now the W3C Validator Suite™ premium service that checks your entire website and evaluates its conformance with W3C open standards to quickly identify those portions of your website that need your attention.

Fig. 112: Validación sintáctica del código HTML de la página de situación general

WCAG 2.0 NIVEL A					
Criterio	Título	Comentarios	SI	NO	N/A
1.2.1	Contenido no textual	Todas las imágenes tienen alternativa textual. No existen otros elementos que requieran alternativa textual.	X		
1.2.1	Sólo audio y video (grabado)	No existen audios ni videos en la página			X
1.2.2	Subtítulos (grabados)	No existen audios ni videos en la página			X
1.2.3	Audiodescripción o medio alternativo (grabado)	No existen audios ni videos en la página			X
1.3.1	Información y relaciones	Se utiliza <i>role</i> y <i>aria-expanded</i> para el botón desplegable de la barra de navegación.	X		
1.3.2	Secuencia significativa	No hay elementos visuales que alteren el orden natural de lectura ni su contenido	X		
1.3.3	Características	No hay elementos accionables cuyo	X		

	sensoriales	descubrimiento dependa de características sensoriales (forma, tamaño etc.).			
1.4.1	Uso del color	Se utiliza el color para indicar el estado de los trabajos en la parte derecha de la pantalla (color verde si se ejecutaron con éxito, rojo si fallo).	X		
1.4.2	Control del audio	No hay audios en la página			X
2.1.1	Teclado	Se puede navegar por todos los elementos accionables de la página utilizando la tecla tabulador.	X		
2.12	Sin trampas para el foco del teclado	El foco no queda atrapado al navegar con el teclado	X		
2.2.1	Tiempo ajustable	Sin límites temporales			X
2.2.2	Pausar, detener, ocultar	No existen elementos que presenten información móvil			X
2.3.1	Umbral de tres destellos o menos	No hay elementos de este tipo en el sitio			X
2.4.1	Evitar bloques	No existen grandes bloques repetidos en la página que dificulten la navegación.			X
2.4.2	Titulado de páginas	La página posee título indicando el nombre de la aplicación y nombrando la página actual	X		
2.4.3	Orden del foco	La secuencia de navegación utilizando la tecla de tabulación es correcta.	X		
2.4.4	Propósito de los enlaces (en contexto)	Los enlaces presentes se describen por su contexto (barra de navegación o panel derecho de trabajos).	X		
3.1.1	Idioma de la página	No está presente el atributo lang en el elemento html.		X	
3.2.1	Al recibir el foco	Los elementos al recibir el foco no varían automáticamente	X		
3.2.2	Al recibir entradas	No existen entradas en esta página			X
3.3.1	Identificación de errores	No hay formularios en esta página			X
3.3.2	Etiquetas o instrucciones	No hay formularios en esta página			X
4.1.1	Procesamiento	El código HTML es sintácticamente válido	X		

4.1.2	Nombre, función, valor	No existen controles de usuario propios ni campos de formulario			X
-------	------------------------	---	--	--	---

WCAG 2.0 NIVEL AA					
Criterio	Título	Comentarios	SI	NO	N/A
1.2.4	Subtítulos (en directo)	No hay videos en la página			X
1.2.5	Audiodescripción (grabado)	No hay videos en la página			X
1.4.3	Contraste mínimo	La herramienta de análisis utilizada (WCAG Contrast Checker para Firefox) indica de relaciones de contraste insuficientes.		X	
1.4.4	Cambio de tamaño del texto	Al aumentar el tamaño del texto al 200% se reconfigura la página sin perder legibilidad.	X		
1.4.5	Imágenes de texto	No se utilizan imágenes de texto en ningún momento	X		
2.4.5	Múltiples vías	Sólo se proporciona una barra de navegación. Es necesario implementar otra vía.		X	
2.4.6	Encabezados y etiquetas	Se utilizan encabezados para describir las secciones	X		
2.4.7	Foco visible	Se muestra mediante un recuadro alrededor de elemento que tiene el foco.	X		
3.1.2	Idiomas de las partes	Los estados de los trabajos se definen en inglés. No se indica dicho idioma con el atributo lang.		X	
3.2.3	Navegación consistente	Los elementos de navegación se sitúan siempre en los mismos lugares	X		
3.2.4	Identificación coherente	Se utilizan las mismas metáforas visuales que en el resto de páginas	X		
3.3.3	Sugerencias ante errores	No hay formularios en esta página			X
3.3.4	Prevención de errores	No hay formularios en esta página			X

### 8.3.2.3.2 Conformidad de la página de ficha de un trabajo

The screenshot shows the W3C Markup Validation Service interface. At the top, it says "Markup Validation Service" and "Check the markup (HTML, XHTML, ...) of Web documents". Below that, there are navigation links: "Jump To: Notes and Potential Issues", "Validation Output", and "Congratulations · Icons". A green banner states "This document was successfully checked as HTML5!". The main content area shows the validation result: "Result: Passed, 4 warning(s)". Below this, the source code of the HTML document is displayed, including the doctype, head, title, and meta tags. At the bottom, there are dropdown menus for "Encoding" (set to utf-8) and "Doctype" (set to HTML5), and a "Root Element" field set to html. A logo for "W3C VALIDATOR Suite" is also present, along with a promotional text: "Try now the W3C Validator Suite™ premium service that checks your entire website and evaluates its conformance with W3C open standards to quickly identify those portions of your website that need your attention."

**Fig. 113: Validación sintáctica del código HTML de la página de información de un trabajo.**

WCAG 2.0 NIVEL A					
Criterio	Título	Comentarios	SI	NO	N/A
1.2.1	Contenido no textual	Todas las imágenes tienen alternativa textual. No existen otros elementos que requieran alternativa textual.	X		
1.2.1	Sólo audio y video (grabado)	No existen audios ni videos en la página			X
1.2.2	Subtítulos (grabados)	No existen audios ni videos en la página			X
1.2.3	Audiodescripción o medio alternativo (grabado)	No existen audios ni videos en la página			X
1.3.1	Información y relaciones	Se utiliza aria-controls y aria-expanded para el botón desplegable de muestra de campos de informe.	X		
1.3.2	Secuencia significativa	No hay elementos visuales que alteren el orden natural de lectura ni su contenido	X		
1.3.3	Características	No hay elementos accionables cuyo	X		

	sensoriales	descubrimiento dependa de características sensoriales (forma, tamaño etc.).			
1.4.1	Uso del color	Se utiliza el color para indicar la puntuación indicada en el análisis para cada URI (en rojo si la puntuación es menor a 50 puntos, naranja entre 50 y 75 y verde para mas de 75).	X		
1.4.2	Control del audio	No hay audios en la página			X
2.1.1	Teclado	Se puede navegar por todos los elementos de la página excepto por los enlaces generados al cargar los campos de un informe. Falta añadir el atributo tabindex a dichos enlaces.		X	
2.12	Sin trampas para el foco del teclado	El foco no queda atrapado al navegar con el teclado	X		
2.2.1	Tiempo ajustable	Sin límites temporales			X
2.2.2	Pausar, detener, ocultar	No existen elementos que presenten información móvil			X
2.3.1	Umbral de tres destellos o menos	No hay elementos de este tipo en el sitio			X
2.4.1	Evitar bloques	No existen grandes bloques repetidos en la página que dificulten la navegación.	X		
2.4.2	Titulado de páginas	La página posee título indicando el nombre de la aplicación y nombrando la página actual	X		
2.4.3	Orden del foco	La secuencia de navegación utilizando la tecla de tabulación es correcta.	X		
2.4.4	Propósito de los enlaces (en contexto)	Los enlaces presentes se describen por su contexto.	X		
3.1.1	Idioma de la página	No está presente el atributo lang en el elemento html.		X	
3.2.1	Al recibir el foco	Los elementos al recibir el foco no varían automáticamente	X		
3.2.2	Al recibir entradas	No existen entradas en esta página			X
3.3.1	Identificación de errores	No hay formularios en esta página			X
3.3.2	Etiquetas	No hay formularios en esta página			X

4.1.1	Procesamiento	El código HTML es sintácticamente válido	X		
4.1.2	Nombre, función, valor	No existen controles de usuario propios ni campos de formulario	X		

WCAG 2.0 NIVEL AA					
Criterio	Título	Comentarios	SI	NO	N/A
1.2.4	Subtítulos (en directo)	No hay videos en la página			X
1.2.5	Audiodescripción (grabado)	No hay videos en la página			X
1.4.3	Contraste mínimo	La herramienta de análisis utilizada (WCAG Contrast Checker para Firefox) indica de relaciones de contraste insuficientes.		X	
1.4.4	Cambio de tamaño del texto	Al aumentar el tamaño del texto al 200% se reconfigura la página sin perder legibilidad.	X		
1.4.5	Imágenes de texto	No se utilizan imágenes de texto en ningún momento			X
2.4.5	Múltiples vías	Sólo se proporciona una barra de navegación. Es necesario implementar otra vía.		X	
2.4.6	Encabezados y etiquetas	Se utilizan encabezados para describir las secciones	X		
2.4.7	Foco visible	Se muestra mediante un recuadro alrededor de elemento que tiene el foco.	X		
3.1.2	Idiomas de las partes	Los estados de los trabajos se definen en inglés. No se indica dicho idioma con el atributo lang.		X	
3.2.3	Navegación consistente	Los elementos de navegación se sitúan siempre en los mismos lugares	X		
3.2.4	Identificación coherente	Se utilizan las mismas metáforas visuales que en el resto de páginas	X		
3.3.3	Sugerencias ante errores	No hay formularios en esta página			X
3.3.4	Prevención de errores	No hay formularios en esta página			X

## 8.3.2.3 Conformidad de la página del panel de servicios

Markup Validation Service  
Check the markup (HTML, XHTML, ...) of Web documents

---

**Jump To:** [Notes and Potential Issues](#)
[Validation Output](#)
[Congratulations · Icons](#)

This document was successfully checked as HTML5!

<b>Result:</b>	Passed, 4 warning(s)
<b>Source:</b>	<pre>&lt;!doctype HTML&gt; &lt;html&gt;   &lt;head&gt;     &lt;title&gt;WebKumo-       Servicios&lt;/title&gt;     &lt;meta charset="utf-8"/&gt;     &lt;meta name="description" content="WebKumo - An Apache Nutch based, web analysis tool"/&gt;     &lt;meta name="keywords" content="Web analysis, Apache Nutch"/&gt;     &lt;meta name="author" content="Sergio Rozas - MIW - University of Oviedo"/&gt;     &lt;meta name="viewport" content="width=device-width, initial-scale=1"&gt;     &lt;link rel="stylesheet"       href="https://maxcdn.bootstrapcdn.com/bootstrap/3.3.1/css/bootstrap.min.css"&gt;     &lt;link rel="stylesheet" href="/WebKumo/styles/base.css" type="text/css"/&gt;   &lt;/head&gt; &lt;/html&gt;</pre>
<b>Encoding:</b>	utf-8 <small>(detect automatically)</small>
<b>Doctype:</b>	HTML5 <small>(detect automatically)</small>
<b>Root Element:</b>	html

VALIDATOR Suite
Try now the [W3C Validator Suite™](#) premium service that checks your entire website and evaluates its conformance with W3C open standards to quickly identify those portions of your website that need your attention.

Fig. 114: Validación sintáctica del código HTML de la página de panel de servicios

WCAG 2.0 NIVEL A					
Criterio	Título	Comentarios	SI	NO	N/A
1.2.1	Contenido no textual	Todas las imágenes tienen alternativa textual. No existen otros elementos que requieran alternativa textual.	X		
1.2.1	Sólo audio y video (grabado)	No existen audios ni videos en la página			X
1.2.2	Subtítulos (grabados)	No existen audios ni videos en la página			X
1.2.3	Audiodescripción o medio alternativo (grabado)	No existen audios ni videos en la página			X
1.3.1	Información y relaciones	Se utiliza aria-controls y aria-expanded para el botón desplegable que muestra la salida del servicio. Los botones de accionamiento de servicio no están asociados a elementos accionables de la interfaz.	X		

1.3.2	Secuencia significativa	No hay elementos visuales que alteren el orden natural de lectura ni su contenido	X		
1.3.3	Características sensoriales	No hay elementos accionables cuyo descubrimiento dependa de características sensoriales (forma, tamaño etc.).	X		
1.4.1	Uso del color	Se utilizan iconos diferenciados por colores para indicar los estados de los servicios. Además si una operación falla es mostrada mediante un mensaje en rojo mientras que si tiene éxito éste se muestra en verde.	X		
1.4.2	Control del audio	No hay audios en la página			X
2.1.1	Teclado	Se puede navegar por todos los elementos de la página utilizando el tabulador.	X		
2.12	Sin trampas para el foco del teclado	El foco no queda atrapado al navegar con el teclado	X		
2.2.1	Tiempo ajustable	Sin límites temporales			X
2.2.2	Pausar, detener, ocultar	No existen elementos que presenten información móvil			
2.3.1	Umbral de tres destellos o menos	No hay elementos de este tipo en el sitio			X
2.4.1	Evitar bloques	No existen grandes bloques repetidos en la página que dificulten la navegación.			X
2.4.2	Titulado de páginas	La página posee título indicando el nombre de la aplicación y nombrando la página actual	X		
2.4.3	Orden del foco	La secuencia de navegación utilizando la tecla de tabulación es correcta.	X		
2.4.4	Propósito de los enlaces (en contexto)	Los enlaces presentes se describen por su contexto y su descripción.	X		
3.1.1	Idioma de la página	No está presente el atributo lang en el elemento html.		X	
3.2.1	Al recibir el foco	Los elementos al recibir el foco no varían automáticamente	X		
3.2.2	Al recibir entradas	No existen entradas en esta página			X

3.3.1	Identificación de errores	No hay formularios en esta página			X
3.3.2	Etiquetas o instrucciones	No hay formularios en esta página			X
4.1.1	Procesamiento	El código HTML es sintácticamente válido	X		
4.1.2	Nombre, función, valor	No existen controles de usuario propios ni campos de formulario			X

WCAG 2.0 NIVEL AA					
Criterio	Título	Comentarios	SI	NO	N/A
1.2.4	Subtítulos (en directo)	No hay videos en la página			X
1.2.5	Audiodescripción (grabado)	No hay videos en la página			X
1.4.3	Contraste mínimo	La herramienta de análisis utilizada (WCAG Contrast Checker para Firefox) indica de relaciones de contraste insuficientes.		X	
1.4.4	Cambio de tamaño del texto	Al aumentar el tamaño del texto al 200% se reconfigura la página sin perder legibilidad.	X		
1.4.5	Imágenes de texto	No se utilizan imágenes de texto en ningún momento			X
2.4.5	Múltiples vías	Sólo se proporciona una barra de navegación. Es necesario implementar otra vía.		X	
2.4.6	Encabezados y etiquetas	Se utilizan encabezados para describir las secciones	X		
2.4.7	Foco visible	Se muestra mediante un recuadro alrededor de elemento que tiene el foco.	X		
3.1.2	Idiomas de las partes	No existen partes con idiomas distintos al del usuario. La parte de salida del servicio es la generada por el propio software y puede ser considerada como término técnico.			X
3.2.3	Navegación consistente	Los elementos de navegación se sitúan siempre en los mismos lugares	X		

3.2.4	Identificación coherente	Se utilizan las mismas metáforas visuales que en el resto de páginas	X		
3.3.3	Sugerencias ante errores	No hay formularios en esta página			X
3.3.4	Prevención de errores	No hay formularios en esta página			X

### 8.3.2.3.4 Conformidad de la página del creación de un trabajo

**W3C** Markup Validation Service  
Check the markup (HTML, XHTML, ...) of Web documents

Jump To: [Notes and Potential Issues](#) [Validation Output](#) [Congratulations · Icons](#)

This document was successfully checked as HTML5!

**Result:** Passed, 4 warning(s)

**Source:**

```
<!doctype HTML>
<html>
  <head>
    <title>WebKumo -
      Crear un trabajo de análisis</title>
    <meta charset="utf-8"/>
    <meta name="description" content="WebKumo - An Apache Nutch based, web analysis tool"/>
    <meta name="keywords" content="Web analysis, Apache Nutch"/>
    <meta name="author" content="Sergio Rozas - MIW - University of Oviedo"/>
    <meta name="viewport" content="width=device-width, initial-scale=1">
    <link rel="stylesheet"
      href="https://maxcdn.bootstrapcdn.com/bootstrap/3.3.1/css/bootstrap.min.css">
    <link rel="stylesheet" href="/WebKumo/styles/base.css" type="text/css"/>
  </head>
</html>
```

**Encoding:** utf-8 (detect automatically)

**Doctype:** HTML5 (detect automatically)

**Root Element:** html

**W3C VALIDATOR Suite** Try now the **W3C Validator Suite™** premium service that checks your entire website and evaluates its conformance with W3C open standards to quickly identify those portions of your website that need your attention.

Fig. 115: Validación sintáctica del código HTML de la página de creación de un trabajo.

WCAG 2.0 NIVEL A					
Criterio	Título	Comentarios	SI	NO	N/A
1.2.1	Contenido no textual	Todas las imágenes tienen alternativa textual. No existen otros elementos que requieran alternativa textual.	X		
1.2.1	Sólo audio y video (grabado)	No existen audios ni videos en la página			X
1.2.2	Subtítulos (grabados)	No existen audios ni videos en la página			X
1.2.3	Audiodescripción	No existen audios ni videos en la			X

	o medio alternativo (grabado)	página			
1.3.1	Información y relaciones	Se utiliza aria-hidden para indicar que el panel que muestra los parámetros personalizables de análisis están ocultos.	X		
1.3.2	Secuencia significativa	No hay elementos visuales que alteren el orden natural de lectura ni su contenido	X		
1.3.3	Características sensoriales	No hay elementos accionables cuyo descubrimiento dependa de características sensoriales (forma, tamaño etc.).	X		
1.4.1	Uso del color	Se utiliza el color rojo para indicar los campos erróneos del formulario	X		
1.4.2	Control del audio	No hay audios en la página			X
2.1.1	Teclado	Se puede navegar por todos los elementos de la página utilizando el tabulador salvo para los botones que muestran el listado de parámetros de trabajo personalizables y el que añade un campo nuevo para indicar otro parámetro personalizable.		X	
2.12	Sin trampas para el foco del teclado	El foco no queda atrapado al navegar con el teclado	X		
2.2.1	Tiempo ajustable	Sin límites temporales			X
2.2.2	Pausar, detener, ocultar	No existen elementos que presenten información móvil			X
2.3.1	Umbral de tres destellos o menos	No hay elementos de este tipo en el sitio			X
2.4.1	Evitar bloques	No existen grandes bloques repetidos en la página que dificulten la navegación.	X		
2.4.2	Titulado de páginas	La página posee título indicando el nombre de la aplicación y nombrando la página actual	X		
2.4.3	Orden del foco	La secuencia de navegación utilizando la tecla de tabulación es correcta.	X		
2.4.4	Propósito de los enlaces (en	Los enlaces presentes se describen por su contexto y su descripción.	X		

	contexto)				
3.1.1	Idioma de la página	No está presente el atributo lang en el elemento html.		X	
3.2.1	Al recibir el foco	Los elementos al recibir el foco no varían automáticamente	X		
3.2.2	Al recibir entradas	No hay cambios de contexto automáticos en los campos del formulario.	X		
3.3.1	Identificación de errores	Existe validación de los campos del formulario con indicación de errores.	X		
3.3.2	Etiquetas o instrucciones	Se utilizan <i>placeholders</i> para indicar ejemplos de entradas válidas.	X		
4.1.1	Procesamiento	El código HTML es sintácticamente válido	X		
4.1.2	Nombre, función, valor	Los nombres de los campos de formularios están sin etiquetar (están marcados como texto <span>).		X	

WCAG 2.0 NIVEL AA					
Criterio	Título	Comentarios	SI	NO	N/A
1.2.4	Subtítulos (en directo)	No hay videos en la página			X
1.2.5	Audiodescripción (grabado)	No hay videos en la página			X
1.4.3	Contraste mínimo	La herramienta de análisis utilizada (WCAG Contrast Checker para Firefox) indica de relaciones de contraste insuficientes.		X	
1.4.4	Cambio de tamaño del texto	Al aumentar el tamaño del texto al 200% se reconfigura la página sin perder legibilidad.	X		
1.4.5	Imágenes de texto	No se utilizan imágenes de texto en ningún momento			X
2.4.5	Múltiples vías	Sólo se proporciona una barra de navegación. Es necesario implementar otra vía.		X	
2.4.6	Encabezados y etiquetas	Se utilizan encabezados para describir las secciones	X		
2,4.7	Foco visible	Se muestra mediante un recuadro	X		

		alrededor de elemento que tiene el foco.			
3.1.2	Idiomas de las partes	No existen partes con idiomas distintos al del usuario.			X
3.2.3	Navegación consistente	Los elementos de navegación se sitúan siempre en los mismos lugares	X		
3.2.4	Identificación coherente	Se utilizan las mismas metáforas visuales que en el resto de páginas	X		
3.3.3	Sugerencias ante errores	Al introducir un valor inválido en el formulario se le muestra al usuario una indicación con los valores válidos.	X		
3.3.4	Prevención de errores	Existe información ante errores, además se puede modificar la información de un trabajo si se envía de una manera incorrecta	X		

### 8.3.2.3.5 Conformidad de la página del panel del administrador

The screenshot shows the W3C Markup Validation Service interface. At the top, it says "Markup Validation Service" and "Check the markup (HTML, XHTML, ...) of Web documents". Below that, there are navigation links: "Jump To: Notes and Potential Issues", "Validation Output", and "Congratulations - Icons". A green banner states "This document was successfully checked as HTML5!". The main content area shows the validation result: "Result: Passed, 4 warning(s)". Below this, the "Source" code is displayed, including the DOCTYPE, head, title, meta, and link tags. At the bottom, there are dropdown menus for "Encoding" (set to utf-8) and "Doctype" (set to HTML5), and a "Root Element" field set to html. The W3C Validator Suite logo is visible at the bottom left, along with a promotional text: "Try now the W3C Validator Suite™ premium service that checks your entire website and evaluates its conformance with W3C open standards to quickly identify those portions of your website that need your attention."

Fig. 116: Validación sintáctica del código HTML de la página del panel del administrador

WCAG 2.0 NIVEL A					
Criterio	Título	Comentarios	SI	NO	N/A
1.2.1	Contenido no textual	Todas las imágenes tienen alternativa textual. No existen otros elementos que requieran alternativa textual.	X		
1.2.1	Sólo audio y video (grabado)	No existen audios ni videos en la página			X
1.2.2	Subtítulos (grabados)	No existen audios ni videos en la página			X
1.2.3	Audiodescripción o medio alternativo (grabado)	No existen audios ni videos en la página			X
1.3.1	Información y relaciones	Se utiliza aria-hidden para marcar como oculto el diálogo de borrado y aria-label para marcar los botones de eliminación de las tablas.	X		
1.3.2	Secuencia significativa	No hay elementos visuales que alteren el orden natural de lectura ni su contenido	X		
1.3.3	Características sensoriales	No hay elementos accionables cuyo descubrimiento dependa de características sensoriales (forma, tamaño etc.).	X		
1.4.1	Uso del color	Se utiliza el color rojo para indicar la imposibilidad de eliminar un grupo o un usuario. Se utiliza el color verde para indicar éxito en la misma operación.	X		
1.4.2	Control del audio	No hay audios en la página			X
2.1.1	Teclado	Se puede navegar por todos los elementos de la página utilizando el tabulador incluyendo las tablas.	X		
2.1.2	Sin trampas para el foco del teclado	El foco no queda atrapado al navegar con el teclado	X		
2.2.1	Tiempo ajustable	Sin límites temporales			X
2.2.2	Pausar, detener, ocultar	No existen elementos que presenten información móvil			X
2.3.1	Umbral de tres destellos o menos	No hay elementos de este tipo en el sitio			X

2.4.1	Evitar bloques	No existen grandes bloques repetidos en la página que dificulten la navegación.	X		
2.4.2	Titulado de páginas	La página posee título indicando el nombre de la aplicación y nombrando la página actual	X		
2.4.3	Orden del foco	La secuencia de navegación utilizando la tecla de tabulación es correcta.	X		
2.4.4	Propósito de los enlaces (en contexto)	Los enlaces presentes se describen por su contexto.	X		
3.1.1	Idioma de la página	No está presente el atributo lang en el elemento html.		X	
3.2.1	Al recibir el foco	Los elementos al recibir el foco no varían automáticamente	X		
3.2.2	Al recibir entradas	No hay cambios de contexto automáticos en los campos del formulario.	X		
3.3.1	Identificación de errores	No existen formularios en esta página.			X
3.3.2	Etiquetas o instrucciones	No existen formularios en esta página.			X
4.1.1	Procesamiento	El código HTML es sintácticamente válido	X		
4.1.2	Nombre, función, valor	No existen formularios en esta página.			X

WCAG 2.0 NIVEL AA					
Criterio	Título	Comentarios	SI	NO	N/A
1.2.4	Subtítulos (en directo)	No hay videos en la página			X
1.2.5	Audiodescripción (grabado)	No hay videos en la página			X
1.4.3	Contraste mínimo	La herramienta de análisis utilizada (WCAG Contrast Checker para Firefox) indica de relaciones de contraste insuficientes.		X	
1.4.4	Cambio de tamaño del texto	Al aumentar el tamaño del texto al 200% se reconfigura la página sin	X		

		perder legibilidad.			
1.4.5	Imágenes de texto	No se utilizan imágenes de texto en ningún momento			X
2.4.5	Múltiples vías	Sólo se proporciona una barra de navegación. Es necesario implementar otra vía.		X	
2.4.6	Encabezados y etiquetas	Se utilizan encabezados para describir las secciones	X		
2,4.7	Foco visible	Se muestra mediante un recuadro alrededor de elemento que tiene el foco.	X		
3.1.2	Idiomas de las partes	No existen partes con idiomas distintos al del usuario.			X
3.2.3	Navegación consistente	Los elementos de navegación se sitúan siempre en los mismos lugares	X		
3.2.4	Identificación coherente	Se utilizan las mismas metáforas visuales que en el resto de páginas	X		
3.3.3	Sugerencias ante errores	No existen formularios en esta página.			X
3.3.4	Prevención de errores	No existen formularios en esta página.			X

### 8.3.2.4 Corrección de problemas y declaración de conformidad general

Como se vio en el apartado anterior, al realizar la comprobación de criterios de éxito para WCAG 2.0 (niveles A y AA), se revelaron algunos puntos que no son satisfechos por la aplicación. Se procedió a subsanar los problemas encontrados como se indica a continuación.

1. **Criterio 3.1.1 (A):** Se introdujo el idioma en el que se muestran las páginas como atributo lang del elemento html del código de hipertexto.

```

<!DOCTYPE html>
<html lang="es">
<head>
  <title></title>
  <meta charset="utf-8"></meta>
  <meta content="WebKumo - An Apache Nutch based, web analysis tool" name="description"></meta>
  <meta content="Web analysis, Apache Nutch" name="keywords"></meta>
  <meta content="Sergio Rozas - MIW - University of Oviedo" name="author"></meta>
  <meta content="width=device-width, initial-scale=1" name="viewport"></meta>
  <link href="https://maxcdn.bootstrapcdn.com/bootstrap/3.3.1/css/bootstrap.min.css" rel="stylesheet"></link>
  <link type="text/css" href="/WebKumo/styles/base.css" rel="stylesheet"></link>
  <link type="text/css" href="/WebKumo/styles/navbar.css" rel="stylesheet"></link>
  <link type="text/css" href="/WebKumo/styles/footer.css" rel="stylesheet"></link>
  <script src="https://code.jquery.com/jquery-2.1.3.min.js"></script>
  <script src="https://maxcdn.bootstrapcdn.com/bootstrap/3.3.1/js/bootstrap.min.js"></script>
  <link type="text/css" href="styles/dashboard.css" rel="stylesheet"></link>
  <script src="/WebKumo/scripts/spin.min.js"></script>

```

Fig. 117: Criterio WCAG 2.0 3.1.1 corregido.

2. **Criterio 1.4.3 (AA):** Se corrigieron las paletas de colores en las páginas analizadas para conseguir que el ratio de contraste sea 4.5:1. Se muestra a continuación el análisis del contraste (nótese que en todas las páginas analizadas, la herramienta indica un elemento que no cumple la relación de contraste, este elemento corresponde al logotipo de WebKumo al cuál no se le aplica el criterio **[WC143]** y por tanto se puede ignorar).

The image shows the WCAG Contrast Checker tool interface. On the left, a table lists elements with their contrast ratios and corresponding HTML tags. The top entry shows a ratio of 4.22 for the element '4-small: [a,mng,div,span]'. Below the table, there are options for 'Normal Vision' and 'Normal Color Vision', and a 'Normal Vision' section with color pickers for 'Primer plano' and 'Fondo'. A small table at the bottom shows contrast levels for AA and AAA.

Nivel	Luminosidad	small	large
AA	4.22	✗	✓
AAA	4.22	✗	✗

The right side of the image shows a dashboard page titled 'SERVICIOS' with a large orange warning icon and the text 'Existen algunos servicios que requieren atención'. Below this, a table lists services: APACHE HADOOP (Parado), APACHE NUTCH (En ejecución), APACHE SOLR (Parado), and APACHE CASSANDRA (Parado). The last update is noted as 'Última actualización: 26-Jun-2015 12:04:43'.

Fig. 118: Criterio 1.4.3 corregido para la página de situación general

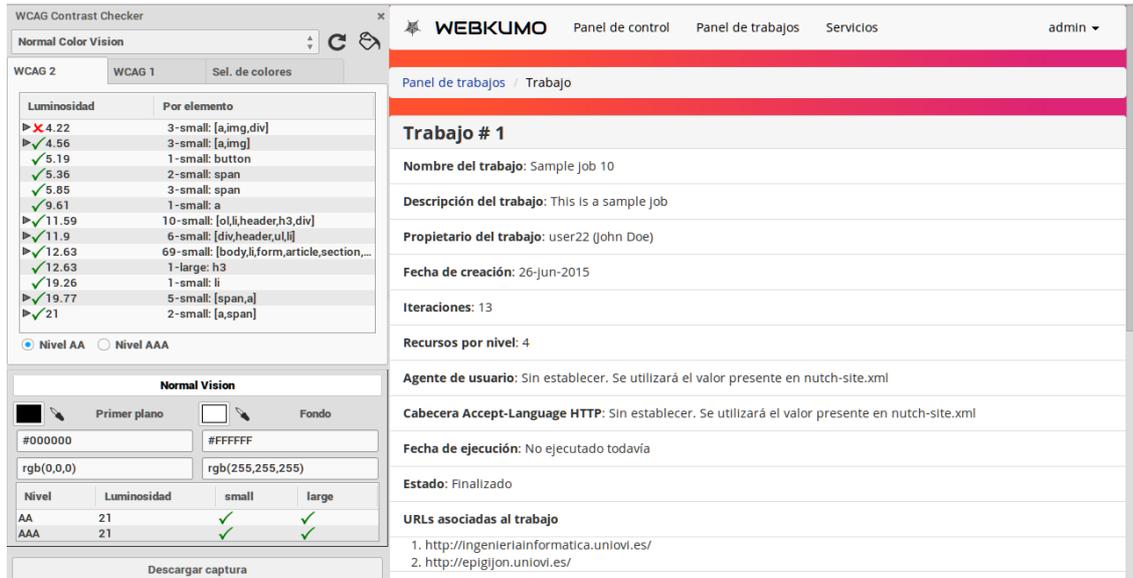


Fig. 119: Criterio 1.4.3 corregido para la página de información de un trabajo

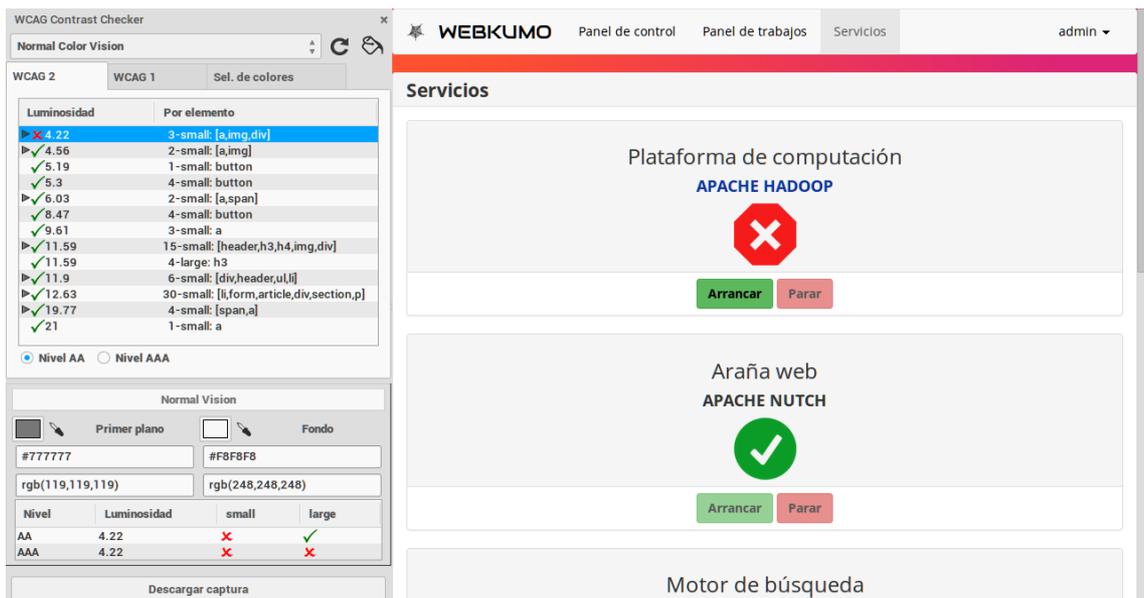


Fig. 120: Criterio 1.4.3 corregido para la página de panel de servicios

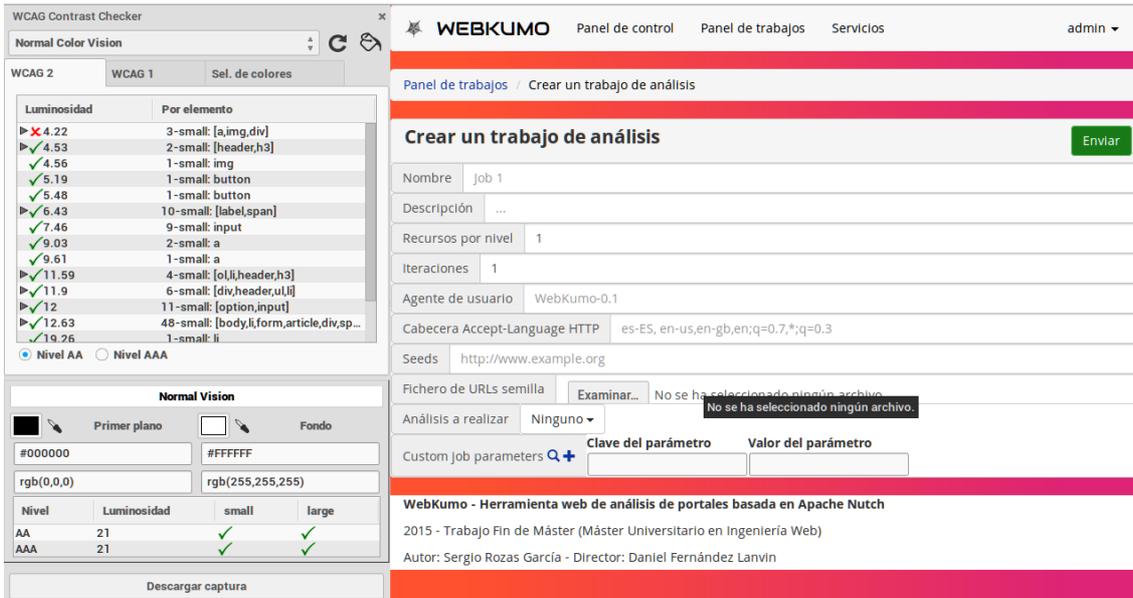


Fig. 121: Criterio 1.4.3 corregido para la página de creación de un trabajo

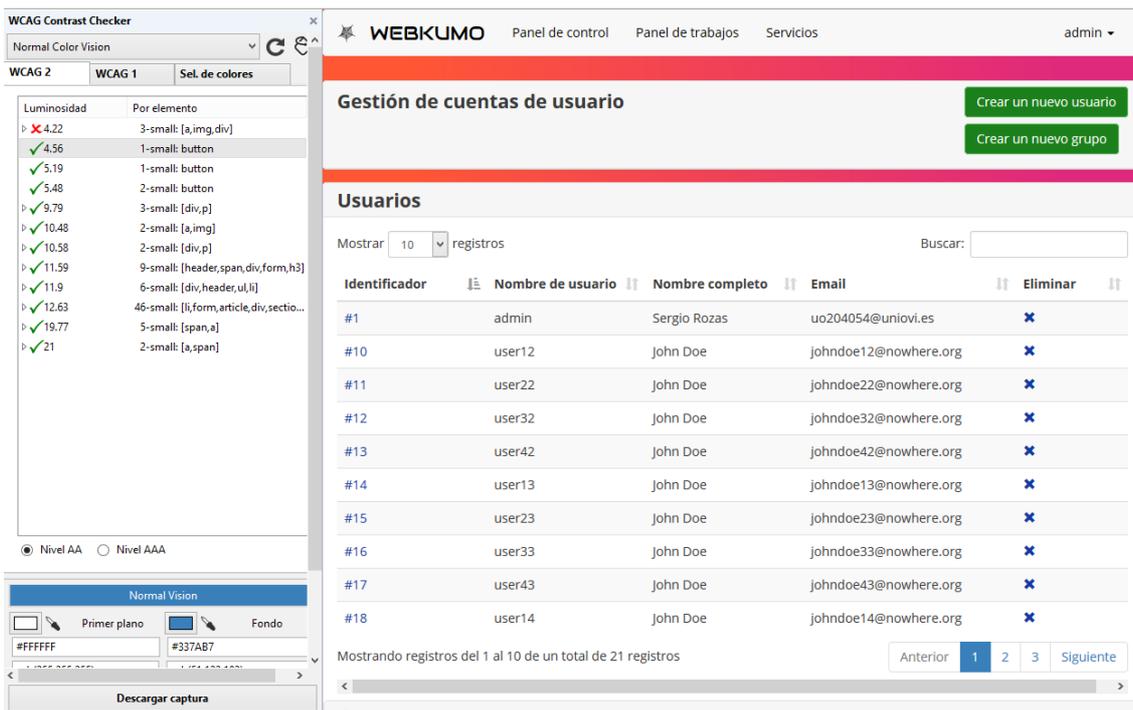
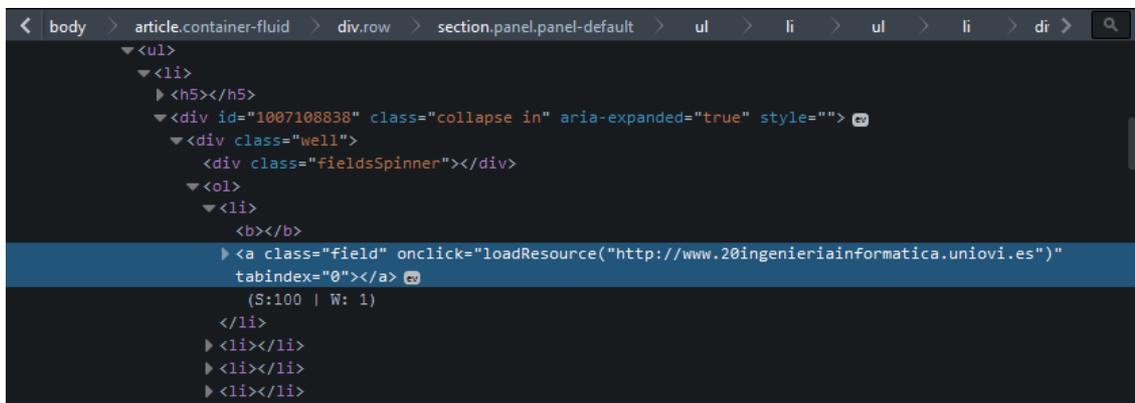


Fig. 122: Criterio 1.4.3 corregido para la página del panel del administrador

- Criterio 2.1.1:** Se añadieron atributos `tabindex` en aquellos elementos de enlace anchor que no los tuvieran.



```
<ul>
  <li>
    <h5></h5>
    <div id="1007108838" class="collapse in" aria-expanded="true" style="">
      <div class="well">
        <div class="fieldsSpinner"></div>
        <ol>
          <li>
            <b></b>
            <a class="field" onclick="loadResource('http://www.20ingenieriainformatica.uniovi.es')"
              tabindex="0"></a>
              (S:100 | W: 1)
          </li>
          <li></li>
          <li></li>
          <li></li>
        </ol>
      </div>
    </li>
  </ul>
```

**Fig. 123:** Criterio 2.2.1 corregido mediante el uso del atributo `tabindex`.

- Criterio 2.4.5:** No se disponen de múltiples vías de navegación, únicamente existe la barra de navegación superior. Como se trata de una aplicación web con un número de documentos relativamente reducido, se opta por adoptar la técnica WCAG 2.0 G185 **[WG185]** añadiendo en el pie de página del sitio principal (la vista de situación general) enlaces a las otras páginas del sitio.



**Fig. 124:** Criterio 2.4.5 corregido mediante el insertado de enlaces a otras páginas del sitio.

5. **Criterio 3.1.2:** Se marcaron los estados de los trabajos con idioma Inglés.

```
▼<li class="list-group-item">
  ▶<a href="/webkumo/job/nutch/1"></a>
  Sample job 10 : 26/6/2015 2:00:00 (user21)
  ▼<span class="pull-right label label-success label-as-badge" lang="en">
    FINISHED
  </span>
</li>
```

*Fig. 125: Criterio 3.1.2 corregido mediante el insertado de marcas de idioma.*

6. **Criterio 4.1.2:** Se etiquetaron los nombres de los campos del formulario utilizando el elemento <label>.

```
▼<div class="input-group">
  ▶<label id="langHeader-text" class="input-group-addon" for="langHeader"></label>
  <input id="langHeader" class="form-control" type="text" aria-describedby="langHea
</div>
▼<div class="input-group">
  ▶<label id="uri-text" class="input-group-addon" for="uri"></label>
  <input id="uri" class="form-control" type="text" aria-describedby="uri-text" plac
</div>
```

*Fig. 126: Criterio 4.1.2 corregido mediante el uso de elementos label.*

Una vez subsanados los problemas descritos y en función de la evaluación de conformidad realizada, se puede afirmar que **el portal web WebKumo cumple los requisitos declarados en las normas de accesibilidad W3C WCAG 2.0 para el nivel AA** (equivalente norma UNE 139803:2012).

## 8.4 Pruebas de rendimiento

Tal y como se explicó en la especificación técnica de las pruebas, se detectó que podrían surgir problemas de rendimiento cuando, ante un trabajo con un tamaño importante, se mostraran los resultados del informe de análisis. Esto se debe a que cada informe está compuesto por campos que son persistidos en la base de datos. Si se tiene un trabajo que haya generado un informe con muchos campos, puede suponer un problema al extraerlos de la base de datos.

A continuación se muestran las pruebas de rendimiento realizadas en el sentido descrito anteriormente. Las condiciones de las pruebas son:

1. Las peticiones son realizadas con la caché del navegador vaciada para forzar la descarga completa de los recursos web. Se utiliza el medidor de tiempos de respuesta que integra el navegador web Mozilla Firefox.
2. Los parámetros de la máquina virtual de Java son los que se proporcionan por defecto (tamaños de memoria, recolector de basura **[GCOL]** etc.).
3. Únicamente se tienen abiertos como procesos de usuario, el navegador web, el servidor de aplicaciones (Apache Tomcat 8 en este caso) y el entorno de desarrollo (IDE).
4. La prueba se ha realizado tanto con la base de datos de desarrollo (H2) como con la base de datos de producción (MySQL).

La máquina de pruebas tiene las siguientes características:

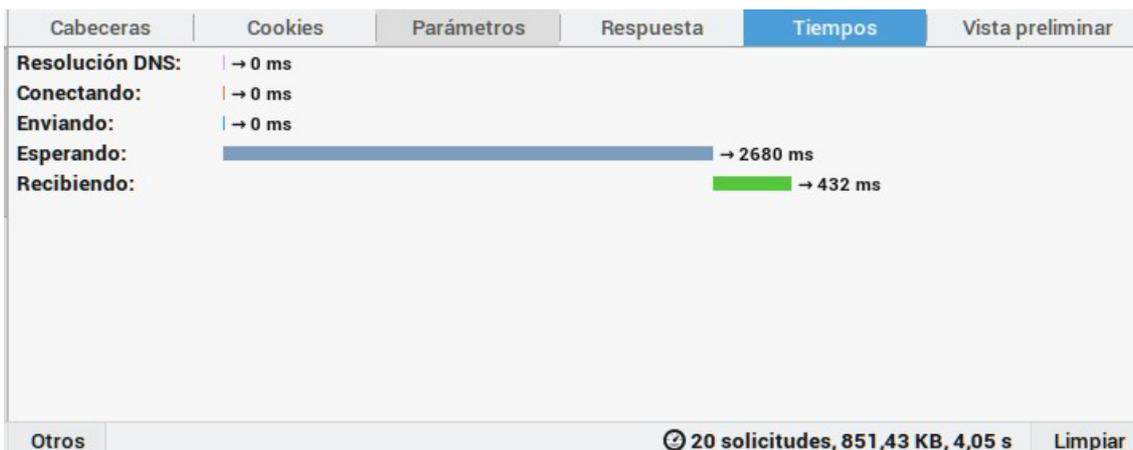
- ◆ **CPU:** Intel Core i5 a 2.3Ghz
- ◆ **Memoria principal:** 4GB 1333Mhz.
- ◆ **Memoria secundaria:** Disco duro magnético 500GB 5400rpm
- ◆ **Sistema Operativo:** GNU/Linux Mint 17.1

Se comenzó probando la carga de un trabajo con 100 campos en sus informes:



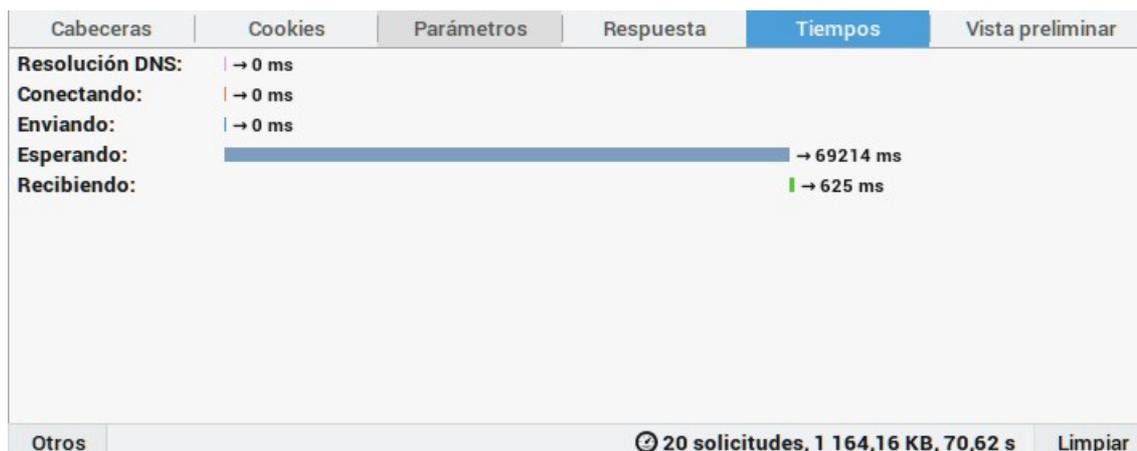
**Fig. 127 : Prueba de rendimiento para 100 campos.**

En la imagen se observa un tiempo de carga razonable de aproximadamente un segundo y medio. Se procede a ejecutar la misma prueba para 1000 registros.



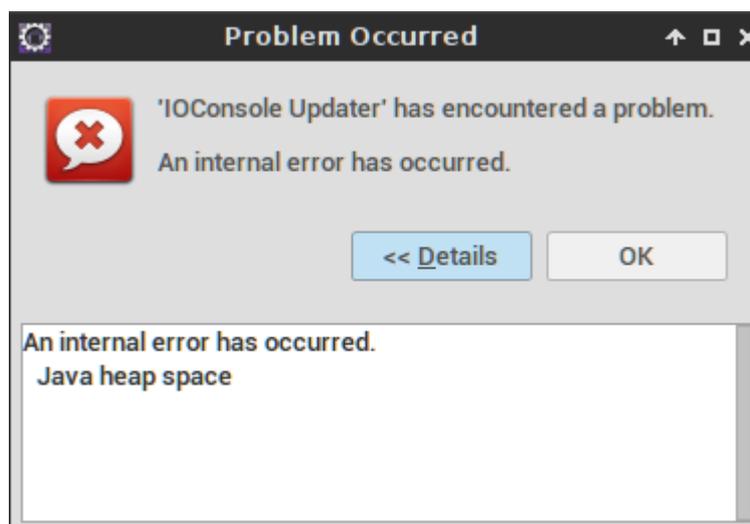
**Fig. 128: Prueba de rendimiento para 1000 campos.**

En este caso el tiempo de respuesta total es de unos cuatro segundos, lo cuál comienza a aproximarse al límite de 10 segundos que se toma como referencia **[NIELS]** a partir de los cuales el usuario deja de esperar y comienza otra tarea. Se continúa con la prueba para 2500 campos:



**Fig. 129: Prueba de rendimiento para 2500 campos.**

Para 2500 campos comienzan a aparecer problemas de memoria lo cuál demora la petición hasta, como se ve en la imagen anterior, los 70 segundos. Este tiempo de respuesta es inaceptable.



**Fig. 130: Problemas de memoria en el IDE al ejecutar la prueba con 2500 campos.**

Puede desprenderse de las pruebas anteriores que, no es práctica la carga de los campos completos de los informes al acceder a los resultados de un trabajo. Si se tiene un trabajo de cierta envergadura es probable que se alcancen estas cantidades de elementos a obtener de la base de datos e incluso superiores. Por lo anterior, se ha decidido implementar la carga dinámica de los campos en la vista de resultados de un trabajo mediante AJAX.

# 9 Manuales

## 9.1 Manual de instalación

A continuación se detallarán los pasos de instalación de la infraestructura que sustenta el sistema. Esta infraestructura está compuesta por los siguientes componentes:

1. **Apache Hadoop:** Software de computación distribuida que se desplegará en modo pseudo-distribuido.
2. **Apache Cassandra:** Software de persistencia de la araña web cuyo despliegue será de una instancia.
3. **Apache Nutch:** Software de araña web. El despliegue se realizará en el mismo nodo maestro que Apache Hadoop.
4. **Apache Solr:** Software de motor de búsqueda desplegado en modo de nodo único.

Se realizará el despliegue en la plataforma recomendada que es una distribución de GNU/Linux (*Debian Jessie* concretamente) mediante interfaz de línea de comandos. Los distintos programas serán instalados en el directorio `/usr/local`. Como pre-requisito **se requiere tener instalado en el sistema el kit de desarrollo de Java (JDK)** al menos en su versión 6. Pude obtenerse mediante el gestor de paquetes de cada distribución concreta.

### 9.1.1 Instalación de Apache Hadoop

Se utilizará la distribución binaria de Apache Hadoop 1.2.0 que en el momento de escribir este manual no es la última versión (es la 1.2.1). Como la versión de Apache Nutch que se utilizará se empaqueta con Apache Hadoop 1.2.0, se recomienda descargar la versión 1.2.0. Se procede a la descarga [**HADO5**] mediante el programa `wget` (también puede utilizarse `curl` o similar), descompresión del fichero `tar.gz` descargado y posterior movimiento a `/usr/local`. Por comodidad se moverá con el nombre de carpeta `hadoop`.

```

sergiorozas@webkumo-master:~$ wget https://archive.apache.org/dist/hadoop/common/hadoop-1.2.0/hadoop-1.2.0-bin.tar.gz
--2015-06-13 09:04:03-- https://archive.apache.org/dist/hadoop/common/hadoop-1.2.0/hadoop-1.2.0-bin.tar.gz
Resolviendo archive.apache.org (archive.apache.org)... 140.211.11.131, 192.87.106.229, 2001:610:1:80bc:192:87:106:29
Conectando con archive.apache.org (archive.apache.org)[140.211.11.131]:443... conectado.
Petición HTTP enviada, esperando respuesta... 200 OK
Longitud: 37650688 (36M) [application/x-gzip]
Grabando a: "hadoop-1.2.0-bin.tar.gz"

hadoop-1.2.0-bin.tar.gz 100%[=====>] 35,91M 6,93MB/s en 10s
2015-06-13 09:04:15 (3,48 MB/s) - "hadoop-1.2.0-bin.tar.gz" guardado [37650688/37650688]

sergiorozas@webkumo-master:~$ tar -xzf hadoop-1.2.0-bin.tar.gz
sergiorozas@webkumo-master:~$ su -c 'mv hadoop-1.2.0 /usr/local/hadoop'
Contraseña:
sergiorozas@webkumo-master:~$ ls /usr/local/
bin etc games hadoop include lib man mondex sbin share src
sergiorozas@webkumo-master:~$

```

**Fig. 131: Descarga de Apache Hadoop.**

Una vez completado este paso, se debe crear un grupo del sistema para el usuario que será encargado de ejecutar Apache Hadoop. El usuario puede ser uno existente o bien uno dedicado. Para generar el nuevo grupo y añadir al usuario al grupo se ejecuta lo siguiente (en el ejemplo se crea el grupo hadoop y se añade el usuario sergiorozas al mismo).

```

sergiorozas@webkumo-master:~$ su -c 'groupadd hadoop'
Contraseña:
sergiorozas@webkumo-master:~$ su -c 'usermod -a -G hadoop sergiorozas'
Contraseña:
sergiorozas@webkumo-master:~$ █

```

**Fig. 132: Usuario y grupo añadido para Hadoop.**

A continuación se debe añadir una variable de entorno que identifica a nivel del sistema operativo el directorio raíz para Hadoop. Esta variable será utilizada internamente por Apache Nutch. La variable es HADOOP\_HOME con el valor que se muestra en las capturas siguientes. Además, si no existía previamente, se debe crear también la variable JAVA\_HOME que apunta al directorio de instalación del JDK.

```

export HADOOP_HOME=/usr/local/hadoop

```

**Fig. 133: Variable de entorno establecida para Hadoop.**

Esta línea debe escribirse en el fichero `.bashrc` del directorio personal del usuario para Apache Hadoop (o similar fichero de configuración de sesión según el intérprete de comandos por defecto que utilice el usuario, en este caso es `bash`). Además es recomendable o bien cerrar e iniciar sesión de nuevo o bien ejecutar la misma línea para hacer efectiva la variable.

El siguiente paso es generar la llave ssh **[SSH]** que Apache Hadoop necesita para la comunicación entre nodos. Nótese que para poder comunicarse, al necesitar comunicación ssh, Hadoop requiere un servidor ssh ejecutándose en la máquina (puede instalarse el paquete `openssh-server` o similar mediante el gestor de paquetes de la distribución en uso).

```
sergiorozas@webkumo-master:~$ ssh-keygen -t rsa -P ""
Generating public/private rsa key pair.
Enter file in which to save the key (/home/sergiorozas/.ssh/id_rsa):
Your identification has been saved in /home/sergiorozas/.ssh/id_rsa.
Your public key has been saved in /home/sergiorozas/.ssh/id_rsa.pub.
The key fingerprint is:
14:36:cc:94:97:dd:4e:15:66:5e:5a:9b:9f:df:67:cf sergiorozas@webkumo-master
The key's randomart image is:
+---[RSA 2048]-----+
|
|  +=. o . ==|
| .+oo . *o+|
| .. o.+ |
| . . o |
| S . . |
| . o |
| = |
| oo |
| E |
|-----+
sergiorozas@webkumo-master:~$
```

**Fig. 134: Generación de la llave RSA para SSH.**

La clave se generará por defecto en el fichero `id_rsa.pub` mostrado anteriormente.

A continuación se debe copiar el contenido de la clave generada al fichero de claves autorizadas mediante el comando mostrado a continuación:

```
sergiorozas@webkumo-master:~$ cat .ssh/id_rsa.pub >> .ssh/authorized_keys
sergiorozas@webkumo-master:~$ cat .ssh/authorized_keys
ssh-rsa AAAAB3NzaC1yc2EAAAADAQABAAQD0tqd91TmG0RrWEM5UXVVu20eTZebHNz07cIvA482o7UvtQJK7uWQM9ubmNwD0E0WtKnFB1hIYken
j6Pjd8AA3hniC0LHG1AAMWFKhvMyVTBTyarKws5g9aPmp0xEfJ60akxKpanVdYaa65Ggf+P8gxohTCnGG141y6+h0LySQyi9yNgPgYbmgGrxU90v2Db
TJh4jo7U9DVWBe/R7bvDC4ybp7rq7o0/n9fU7QHCmpJP06LT2IxHyLnK6kN77F7/3xb3yQ2DFZwxreP1BSCb8p2E0WPwrfk3g2TB348GcivtXuhjGJx
drmu4NEC0MLF9KSpzBbiRbv0AqLuFFfR66j sergiorozas@webkumo-master
sergiorozas@webkumo-master:~$
```

**Fig. 135: Copia de la llave SSH al fichero de llaves autorizadas.**

Si la distribución en la que se está realizando es Debian o derivada (Ubuntu, Mint etc.) se recomienda ejecutar un comando para solventar un problema [HADO6] que aparece con el soporte de Hadoop para el protocolo de internet versión 6 [IPV6]. El siguiente paso consiste en establecer permisos para el usuario que se encargará de ejecutar Apache Hadoop. En la imagen siguiente se ve los comandos a ejecutar para el usuario `sergiorozas` perteneciente al grupo `hadoop` creado anteriormente.

```
root@webkumo-master:~# chown -R sergiorozas:hadoop /usr/local/hadoop/
root@webkumo-master:~#
```

**Fig. 136: Asignación del propietario del directorio de instalación de Apache Hadoop.**

Se debe también crear la jerarquía de directorios que servirá de soporte de almacenamiento para el sistema de archivos distribuido HDFS en el nodo maestro que es el que se está creando en este manual. Se crean los siguientes directorios y se les asignan permisos para el usuario que ejecutará Hadoop (`sergiorozas` del grupo `hadoop`) según se visualiza en la siguiente imagen:

```
root@webkumo-master:~# mkdir /app
root@webkumo-master:~# mkdir /app/tmp
root@webkumo-master:~# mkdir /app/tmp/dir
root@webkumo-master:~# chown -R sergiorozas:hadoop /app
root@webkumo-master:~# chmod -R 750 /app
root@webkumo-master:~#
```

**Fig. 137: Creación de la jerarquía de directorios locales para HDFS.**

Finalmente se procederá a la modificación de los ficheros de configuración de Apache Hadoop. Primeramente se debe modificar el fichero `hadoop-env.sh` del directorio `conf` del directorio de instalación de Hadoop. Se debe configurar la variable `HADOOP_CLASSPATH` para que incluya el directorio `lib` contenido en el directorio de instalación de Hadoop y el directorio `resources` incluido en este proyecto. Adicionalmente se recomienda descomentar la línea `HADOOP_HEAPSIZE` que indica el tamaño del montón de memoria que usará Hadoop en su ejecución en la máquina virtual de Java y también la variable `JAVA_HOME` que indica a Hadoop la instalación de Java que debe utilizar.

```
# The java implementation to use. Required.
export JAVA_HOME=/usr/lib/jvm/java-7-openjdk-amd64

# Extra Java CLASSPATH elements. Optional.
export HADOOP_CLASSPATH=${HADOOP_HOME}/lib:/home/sergiorozas/tfm_webkumo/TFM_WebKumo/src/main/resources

# The maximum amount of heap to use, in MB. Default is 1000.
export HADOOP_HEAPSIZE=2000
```

**Fig. 138: Configuración del fichero `hadoop-env`.**

Otro fichero que es necesario configurar es el fichero `core-site.xml` que es el que contiene las principales propiedades para el sistema de archivos distribuido HDFS. Nótese que `webkumo-master` es el nombre de la máquina (podría ser sustituido por la dirección IP del servidor si esta es fija o bien por nombre de dominio asignado).

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<!-- Put site-specific property overrides in this file. -->

<configuration>
  <property>
    <name>fs.default.name</name>
    <value>hdfs://webkumo-master:9000</value>
  </property>
  <property>
    <name>hadoop.tmp.dir</name>
    <value>/app/tmp/hadoop</value>
  </property>
</configuration>
```

**Fig. 139: Configuración del fichero `core-site`.**

La propiedad `fs.default.name` determina el punto de acceso al sistema de archivos HDFS y la propiedad `hadoop.tmp.dir` se refiere al directorio de almacenamiento interno configurado previamente.

Otro fichero que es necesario configurar es `hdfs-site.xml` en el mismo directorio que `core-site.xml`. Se establece aquí el número de nodos que participan en el replicado de la información. Como en este manual se realiza sobre un único nodo en modo pseudo-distribuido, se debe configurar como sigue:

```
<configuration>
<property>
  <name>dfs.replication</name>
  <value>1</value>
</property>
</configuration>
```

**Fig. 140: Configuración del fichero hdfs-site.**

Una vez llegado a este punto, sólo falta darle formato al namenode (al sistema de archivos HDFS) mediante el uso del comando `hadoop` que se encuentra en la carpeta `bin`.

```
sergiorozas@webkumo-master:/usr/local/hadoop/bin$ ./hadoop namenode -format
Warning: $HADOOP_HOME is deprecated.

15/06/13 13:05:32 INFO namenode.NameNode: STARTUP_MSG:
/*****
STARTUP_MSG: Starting NameNode
STARTUP_MSG: host = webkumo-master/127.0.1.1
STARTUP_MSG: args = [-format]
STARTUP_MSG: version = 1.2.0
STARTUP_MSG: build = https://svn.apache.org/repos/asf/hadoop/common/branches/branch-1.2 -r 1479473; compiled by '
hortonfo' on Mon May 6 06:59:37 UTC 2013
STARTUP_MSG: java = 1.7.0_79
*****/
15/06/13 13:05:33 INFO util.GSet: Computing capacity for map BlocksMap
15/06/13 13:05:33 INFO util.GSet: VM type = 64-bit
15/06/13 13:05:33 INFO util.GSet: 2.0% max memory = 2027290624
15/06/13 13:05:33 INFO util.GSet: capacity = 2^22 = 4194304 entries
15/06/13 13:05:33 INFO util.GSet: recommended=4194304, actual=4194304
15/06/13 13:05:33 INFO namenode.FSNamesystem: fsOwner=sergiorozas
15/06/13 13:05:33 INFO namenode.FSNamesystem: supergroup=supergroup
15/06/13 13:05:33 INFO namenode.FSNamesystem: isPermissionEnabled=true
15/06/13 13:05:33 INFO namenode.FSNamesystem: dfs.block.invalidate.limit=100
15/06/13 13:05:33 INFO namenode.FSNamesystem: isAccessTokenEnabled=false accessKeyUpdateInterval=0 min(s), accessTo
kenLifetime=0 min(s)
15/06/13 13:05:34 INFO namenode.FSEditLog: dfs.namenode.edits.toleration.length = 0
15/06/13 13:05:34 INFO namenode.NameNode: Caching file names occuring more than 10 times
15/06/13 13:05:34 INFO common.Storage: Image file of size 117 saved in 0 seconds.
15/06/13 13:05:34 INFO namenode.FSEditLog: closing edit log: position=4, editlog=/app/tmp/hadoop/dfs/name/current/e
dits
15/06/13 13:05:34 INFO namenode.FSEditLog: close success: truncate to 4, editlog=/app/tmp/hadoop/dfs/name/current/e
dits
15/06/13 13:05:34 INFO common.Storage: Storage directory /app/tmp/hadoop/dfs/name has been successfully formatted.
15/06/13 13:05:34 INFO namenode.NameNode: SHUTDOWN_MSG:
/*****
SHUTDOWN_MSG: Shutting down NameNode at webkumo-master/127.0.1.1
*****/
sergiorozas@webkumo-master:/usr/local/hadoop/bin$
```

**Fig. 141: Formateo del namenode.**

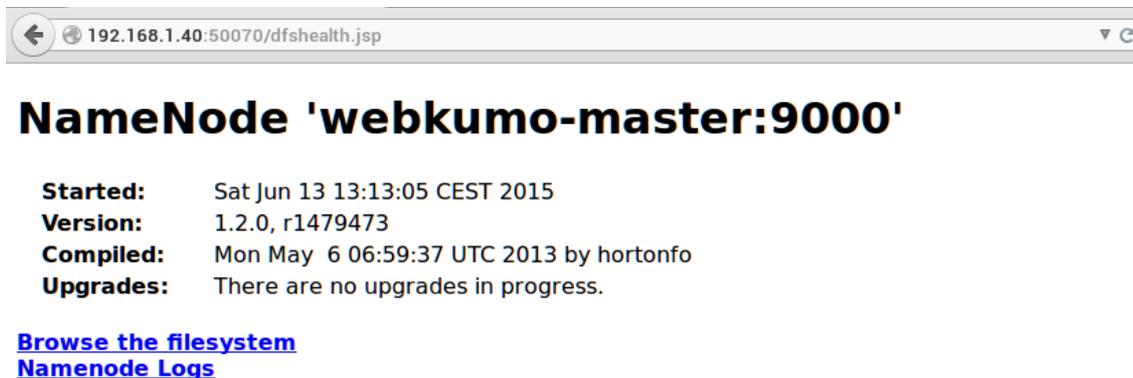
Finalmente se debe arrancar el nodo de Hadoop con el script `start-all.sh` localizado en la carpeta `bin`.

```
sergiorozas@webkumo-master:/usr/local/hadoop/bin$ ./start-all.sh
Warning: $HADOOP_HOME is deprecated.

starting namenode, logging to /usr/local/hadoop/libexec/../logs/hadoop-sergiorozas-namenode-webkumo-master.out
localhost: starting datanode, logging to /usr/local/hadoop/libexec/../logs/hadoop-sergiorozas-datanode-webkumo-master.out
localhost: starting secondarynamenode, logging to /usr/local/hadoop/libexec/../logs/hadoop-sergiorozas-secondarynamenode-webkumo-master.out
starting jobtracker, logging to /usr/local/hadoop/libexec/../logs/hadoop-sergiorozas-jobtracker-webkumo-master.out
localhost: starting tasktracker, logging to /usr/local/hadoop/libexec/../logs/hadoop-sergiorozas-tasktracker-webkumo-master.out
sergiorozas@webkumo-master:/usr/local/hadoop/bin$
```

**Fig. 142: Arranque del nodo Hadoop.**

Para comprobar que el nodo ha arrancado como se espera, se puede mediante un navegador web acceder a la aplicación web del namenode que se encuentra aceptando peticiones en el puerto 50070.



**Fig. 143: Vista de la interfaz web del namenode en ejecución.**

Con esto concluye la instalación del nodo de Apache Hadoop en modo pseudo-distribuido, suficiente para la ejecución del sistema de este proyecto. El despliegue de otros nodos para conformar un clúster es básicamente el mismo recomendándose la guía indicada **[HADO2]** por la Fundación Apache. Únicamente recordar que **si se desea realizar la computación en un clúster, WebKumo y las demás infraestructuras de computación deben de estar localizadas en la misma máquina que alberga el nodo principal** de Hadoop.

## 9.1.2 Instalación de Apache Cassandra

Para sustentar las descargas realizadas por la araña web, se tiene la base de datos NoSQL Apache Cassandra que se debe instalar. Para completar este punto se debe comenzar descargando la versión binaria **[CASSA2]** que proporciona la Fundación Apache (DataStax proporciona una versión Community liberada bajo licencia libre que sirve para el propósito de este proyecto).

Se realizará la descarga de la última versión estable (la 2.1.6 en el momento de escribir este manual) mediante una herramienta tipo curl o wget (o mediante navegador si se dispone de un entorno gráfico). Posteriormente se descomprime el fichero .tar.gz utilizando tar y se mueve a la localización destino en /usr/local.

```
sergiorozas@webkumo-master:~$ wget http://apache.rediris.es/cassandra/2.1.6/apache-cassandra-2.1.6-bin.tar.gz
--2015-06-13 14:01:55-- http://apache.rediris.es/cassandra/2.1.6/apache-cassandra-2.1.6-bin.tar.gz
Resolviendo apache.rediris.es (apache.rediris.es)... 130.206.1.5
Conectando con apache.rediris.es (apache.rediris.es)[130.206.1.5]:80... conectado.
Petición HTTP enviada, esperando respuesta... 200 OK
Longitud: 24561032 (23M) [application/x-gzip]
Grabando a: "apache-cassandra-2.1.6-bin.tar.gz"

apache-cassandra-2.1.6-bin.tar. 100%[=====] 23,42M 4,78MB/s en 21s
2015-06-13 14:02:18 (1,11 MB/s) - "apache-cassandra-2.1.6-bin.tar.gz" guardado [24561032/24561032]

sergiorozas@webkumo-master:~$ tar -xzf apache-cassandra-2.1.6-bin.tar.gz
sergiorozas@webkumo-master:~$ su -c 'mv apache-cassandra-2.1.6 /usr/local/cassandra'
Contraseña:
sergiorozas@webkumo-master:~$ █
```

**Fig. 144: Descarga y descompresión de Apache Cassandra.**

Apache Cassandra no requiere configuración adicional por lo que se procede a arrancar Apache Cassandra mediante el comando `cassandra` del directorio `bin` que se encuentra en el directorio de instalación. Durante el arranque se imprimirá múltiple información sobre el proceso (no existe una versión silenciosa para omitir la salida del arranque).

```
539514210405550827, -7579751717351464778, -7824392738680614470, -78919751906547408, -7895690584307796738, -7954432315323015590
, -7986722066726611744, -7994322090485184680, -8222399236367606456, -8460266841558696081, -8473405694826705797, -8570026643257
678878, -8639234844637632098, -8786129812243691503, -8797409421014124737, -8838721983500565055, -8903970252687634482, -9074279
221892713826, -9127074768534315827, -945117827468801167, 1110115891504761865, 1204245364477304913, 1277294123641458588, 130241
5567343495084, 1321766604385369627, 1348199780952099682, 1364924831207757860, 1407789077904683980, 1442855167511334282, 145561
7890199398393, 1515344483667538068, 1518195827194580127, 1529117781791939856, 154342306284550271, 1591965512150294501, 1640366
907231732921, 1690069136838404634, 1839472724055535845, 1845176145428435446, 1872324400137901792, 1874197418178320000, 1932252
601154117800, 1965748548425910078, 1976712688164448825, 2120436929243859941, 2181771140012545486, 2204006832193826883, 2216545
152775644843, 2353215256079356746, 2458812769520899903, 2549683895494803351, 2591650879612538774, 2652995945728470602, 2942922
148690419029, 2956960460076878240, 3041117769121949314, 3096399394387970470, 3111256184974537200, 3196647834418304622, 3225925
013064854836, 3249405344050366948, 3320026785334793759, 3327632635996067396, 3431010381668921631, 349157657374818957, 3490522
789108811390, 3515852818890102281, 3517733087383665905, 3537236179687065683, 3613127272889518353, 3617113214573346991, 3625432
363519794056, 3648447806158993046, 3681130204971433387, 3753059868369684625, 3798438752177920048, 3810199702684344928, 3927375
967661737351, 4004345995775070187, 4095403558569313486, 427216954374618117, 4347059160826817194, 4397688767037395663, 44325698
06945229588, 44821223625708268, 4497200305543758829, 456433829219104301, 4656118487749990708, 4704416712571674162, 4786785575
216359178, 4864231646212481166, 4955656878891517906, 5020770896900058975, 5044907402392348551, 5097817648901074474, 5110922132
879758756, 5121926728608282621, 513605729191697524, 5178178075983112049, 5236090271644786902, 5255089204956300200, 52630046977
86873454, 5263217207783558955, 5334345875040773798, 5360238047943377339, 5425459499481729867, 542776508367005763, 544142828246
5705820, 5526784604335348180, 5540866562209380532, 5776451443704787062, 5835287001854506727, 5958550776681900994, 618585195090
2988268, 6257315475146347426, 6337211314358027011, 6526238390310100273, 6559008120705117706, 6579324103303855719, 658809931294
9511457, 6698843590294911199, 6809051217313941919, 6979402782279431485, 7017155559972670215, 720665090059036065, 7250968049547
710268, 7259885773586602269, 7290397954327981674, 7293854144101750628, 7344842652817060086, 735538812839964365, 74066980226257
92797, 7534395228911767269, 7584266589932108354, 7682182228312305799, 7910385665727253256, 7964488914878111707, 81739454958798
03271, 8175502882238118401, 8216045082998124838, 8314682429037480820, 8317706627884399937, 8389257658231588984, 87757697247844
83863, 8891084258923089624, 903185997179177249, 9170392737250432027, 966017115497555048]
INFO 12:22:38 Enqueuing flush of local: 716 (0%) on-heap, 0 (0%) off-heap
INFO 12:22:38 Writing Memtable-local@964658829(112 serialized bytes, 4 ops, 0%/0% of on/off-heap limit)
INFO 12:22:38 Completed flushing /usr/local/cassandra/bin/././data/data/system/local-7ad54392bcd35a684174e047860b377/system
-local-ka-18-Data.db (115 bytes) for commitlog position ReplayPosition(segmentId=1434198155149, position=97038)
INFO 12:22:38 Enqueuing flush of local: 51474 (0%) on-heap, 0 (0%) off-heap
INFO 12:22:38 Writing Memtable-local@1115288483(8560 serialized bytes, 259 ops, 0%/0% of on/off-heap limit)
INFO 12:22:38 Compacted 4 sstables to [././data/data/system/local-7ad54392bcd35a684174e047860b377/system-local-ka-17,]. 11
.419 bytes to 5.724 (-50% of original) in 292ms = 0,018695MB/s. 4 total partitions merged to 1. Partition merge counts were
{4:1, }
INFO 12:22:39 Completed flushing /usr/local/cassandra/bin/././data/data/system/local-7ad54392bcd35a684174e047860b377/system
-local-ka-19-Data.db (5264 bytes) for commitlog position ReplayPosition(segmentId=1434198155149, position=109597)
INFO 12:22:39 Node localhost/127.0.0.1 state jump to normal
sergiorozas@webkumo-master:~/usr/local/cassandra/bin$ █
```

**Fig. 145: Proceso de arranque de Apache Cassandra.**

Es recomendable que se prueba la conexión con Apache Cassandra mediante el intérprete de comandos de Cassandra `cqlsh` que se incluye en la carpeta `bin`.

```
sergiorozas@webkumo-master:/usr/local/cassandra/bin$ ./cqlsh
Connected to Test Cluster at 127.0.0.1:9042.
[cqlsh 5.0.1 | Cassandra 2.1.6 | CQL spec 3.2.0 | Native protocol v3]
Use HELP for help.
cqlsh> █
```

**Fig. 146: Intérprete CQLSH en ejecución.**

En este punto ya está listo el nodo de Apache Cassandra listo para su uso. Adicionalmente comentar que Apache Cassandra puede ejecutarse también de forma distribuida mediante el despliegue de varios nodos **[CASSA3]** que se comunicarán vía HTTP. Esto permite replicación asíncrona de datos y mayor capacidad de almacenaje si fuera necesario.

### 9.1.3 Instalación de Apache Solr

Como se ha descrito a lo largo de esta documentación, se tendrá desplegada una instancia del motor de búsqueda Apache Solr con el fin de indexar los recursos web descargados por la araña web para facilitar la búsqueda de dichos recursos una vez descargados. Además Solr contendrá también el código fuente de los recursos.

Se desplegará la versión binaria de Apache Solr 4.10 que es soportada por Nutch 2.2.1 descargándolo **[SOLR4]** desde su sitio web oficial (la última versión de Solr es la 5 pero no está soportada por Apache Gora todavía). La descarga se ejecutará, como en ocasiones anteriores mediante `wget` o `curl` u otra herramienta similar. Tras la descarga, se realizará la descompresión del fichero `.tar.gz` y el posterior copiado al directorio donde se vienen agrupando las instalaciones de la infraestructura.

```

sergiorozas@webkumo-master:~$ wget ftp://ftp.cixug.es/apache/lucene/solr/4.10.3/solr-4.10.3.tgz
--2015-06-14 16:58:34-- ftp://ftp.cixug.es/apache/lucene/solr/4.10.3/solr-4.10.3.tgz
=> "solr-4.10.3.tgz"
Resolviendo ftp.cixug.es (ftp.cixug.es)... 193.144.61.75
Conectando con ftp.cixug.es (ftp.cixug.es)[193.144.61.75]:21... conectado.
Identificándose como anonymous ... ¡Dentro!
==> SYST ... hecho. ==> PWD ... hecho.
==> TYPE I ... hecho. ==> CWD (1) /apache/lucene/solr/4.10.3 ... hecho.
==> SIZE solr-4.10.3.tgz ... 150010621
==> PASV ... hecho. ==> RETR solr-4.10.3.tgz ... hecho.
Longitud: 150010621 (143M) (probablemente)

solr-4.10.3.tgz          100%[=====] 143,06M  12,4MB/s  en 19s
2015-06-14 16:58:53 (7,38 MB/s) - "solr-4.10.3.tgz" guardado [150010621]
sergiorozas@webkumo-master:~$ █

```

**Fig. 147** Descarga y descompresión de Apache Solr.

El siguiente paso es copiar un fichero de configuración que se proporciona con Apache Nutch (ver su proceso descarga mas adelante) para que se pueda indexar sus contenidos. El fichero se denomina `schema-solr4.xml` y se encuentra en la carpeta `conf` dentro del fichero comprimido de Apache Nutch. Se muestra a continuación el proceso de copia teniendo ya descomprimido Apache Nutch en `/usr/local/nutch`:

```

sergiorozas@webkumo-master:/usr/local/solr$ cp /usr/local/nutch/conf/schema-solr4.xml /usr/local/solr/example/solr/collection1/conf
sergiorozas@webkumo-master:/usr/local/solr$ mv /usr/local/solr/example/solr/collection1/conf/schema-solr4.xml /usr/local/solr/example/solr/collection1/conf/schema.xml

```

**Fig. 148:** Copia del fichero `schema` para Apache Nutch.

Una vez copiado el fichero debe, como se ha hecho en la imagen, renombrarse a `schema.xml`. A continuación debe el mismo fichero modificarse para indicarle a Solr que indexe la versión de los recursos que Apache Nutch descargará; este paso es requerido para el funcionamiento de la fase de indexado. Se añade el siguiente elemento como hijo del elemento `fields` en el fichero XML copiado.

```

<!-- Nutch's indexing -->
<field name="_version_" type="long" indexed="true" stored="true"/>█
</fields>

```

**Fig. 149:** Edición del fichero `schema.xml`.

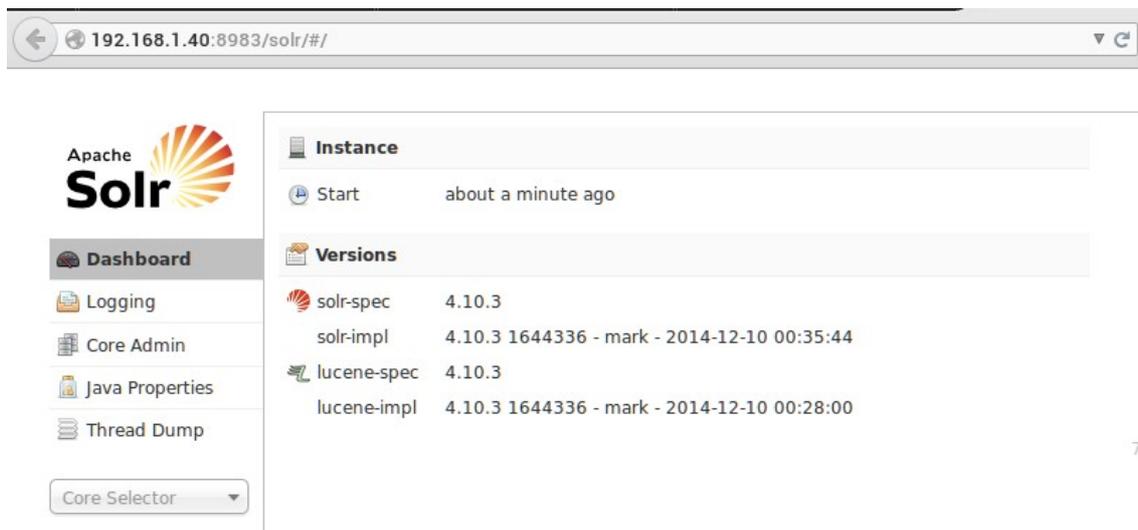
Se puede proceder en este momento al arranque del mismo mediante el ejecutable `solr` que puede encontrarse en el directorio `bin` de la ruta de instalación de Apache Solr. Se debe especificar el parámetro `start` al comando mencionado para efectuar el arranque.

```
sergiorozas@webkumo-master:/usr/local/solr/bin$ ./solr start
Waiting to see Solr listening on port 8983 [\]
Started Solr server on port 8983 (pid=14791). Happy searching!

sergiorozas@webkumo-master:/usr/local/solr/bin$
```

**Fig. 150: Arranque de Apache Solr.**

Puede comprobarse que Apache Solr se ha iniciado correctamente accediendo a su interfaz web que arranca en el puerto 8983 por defecto.



**Fig. 151: Interfaz web de Apache Solr.**

En este punto, Apache Solr estará listo para ser utilizado. Adicionalmente, comentar que al igual que ocurre con Apache Cassandra, Apache Solr puede desplegarse en un clúster de forma distribuida mediante Solr Cloud **[SOLR2]** para el aprovechamiento de las características de replicación de datos y balanceo de carga, entre otras.

## 9.1.4 Instalación de Apache Nutch

Apache Nutch es el software de araña web que será ejecutado como trabajo MapReduce en el software de computación cuya instalación fue cubierta en el punto anterior. Se utilizará la **versión de código fuente** (contiene src en el nombre del archivo) **2.2.1**.

Se comienza con la descarga [**NUTCH6**] del software, al igual que con Apache Hadoop, se realizará mediante herramienta `wget`, `curl` o similar (si la máquina dispone de interfaz gráfico se puede utilizar un navegador web sin ningún problema).

```
sergiorozas@webkumo-master:~$ wget https://archive.apache.org/dist/nutch/2.2.1/apache-nutch-2.2.1-src.tar.gz
--2015-06-13 16:30:11-- https://archive.apache.org/dist/nutch/2.2.1/apache-nutch-2.2.1-src.tar.gz
Resolviendo archive.apache.org (archive.apache.org)... 140.211.11.131, 192.87.106.229, 2001:610:1:80bc:192:87:106:229
Conectando con archive.apache.org (archive.apache.org)[140.211.11.131]:443... conectado.
Petición HTTP enviada, esperando respuesta... 200 OK
Longitud: 3839858 (3,7M) [application/x-gzip]
Grabando a: "apache-nutch-2.2.1-src.tar.gz"

apache-nutch-2.2.1-src.tar.gz 100%[=====] 3,66M 1,31MB/s en 2,8s
2015-06-13 16:30:16 (1,31 MB/s) - "apache-nutch-2.2.1-src.tar.gz" guardado [3839858/3839858]

sergiorozas@webkumo-master:~$ tar -xzf apache-nutch-2.2.1-src.tar.gz
sergiorozas@webkumo-master:~$ su -c 'mv apache-nutch-2.2.1 /usr/local/nutch'
Contraseña:
sergiorozas@webkumo-master:~$
```

**Fig. 152: Descarga y descompresión de Apache Nutch.**

El siguiente paso es, si no se ha realizado durante la instalación de Hadoop, es configurar la variable `JAVA_HOME` como variable de sesión (modificando `.bashrc` o similar según el intérprete, ver la sección de instalación de Hadoop). Se deben realizar las operaciones de configuración mediante la modificación de los ficheros de propiedades de Apache Nutch según se indica en los siguientes pasos.

El primer paso es modificar el fichero de configuración principal `nutch-site.xml` situado en la carpeta `conf` del directorio raíz de instalación de Apache Nutch. Se debe configurar como sigue:

```
<configuration>
  <property>
    <name>http.agent.name</name>
    <value>Mozilla/5.0 (Windows NT 6.3; rv:36.0) Gecko/20100101 Firefox/36.0</value>
  </property>
  <property>
    <name>http.accept.language</name>
    <value>en-us,en-gb,en;q=0.7,*;q=0.3</value>
  </property>
  <property>
    <name>parser.character.encoding.default</name>
    <value>utf-8</value>
  </property>
  <property>
    <name>storage.data.store.class</name>
    <value>org.apache.gora.cassandra.store.CassandraStore</value>
  </property>
  <property>
    <name>plugin.includes</name>
    <value>protocol-httpclient|urlfilter-(regex|domain)|parse-(html|tika)|index-(basic|anchor)|scoring-opic|urlnormalizers
  </property>
  <property>
    <name>plugin.folders</name>
    <value>/usr/local/nutch/runtime/local/plugins</value>
  </property>
  <property>
    <name>http.accept</name>
    <value>
      text/html,application/xhtml+xml,application/javascript,text/css;q=0.9,*/*;q=0.8
    </value>
  </property>
</configuration>
```

**Fig. 153: Configuración del fichero `nutch-site`.**

Se detallan todas las propiedades a continuación en forma de tabla.

NUTCH-SITE.XML		
Nombre	Descripción	Valor
<code>http.agent.name</code>	Agente de usuario que Apache Nutch utilizará a la hora de descargar los usuarios. Por defecto se usará uno de un navegador de escritorio.	Mozilla/5.0 (Windows NT 6.3; rv:36.0) Gecko/20100101 Firefox/36.0
<code>http.accept.language</code>	Valor de la cabecera Accept-Language <b>[ACLAN]</b> de HTTP que Nutch utilizará.	en-us,en-gb,en;q=0.7,*;q=0.3
<code>parser.character.encoding.default</code>	Nombre de la codificación por defecto que se utilizará para procesar los recursos web descargados.	utf-8
<code>storage.data.store.class</code>	<i>Driver</i> de conexión de Apache Gora para ser utilizado por Nutch. Varía en función del <i>backend</i> de persistencia de Nutch (Cassandra en este caso).	org.apache.gora.cassandra.store.CassandraStore
<code>plugin.includes</code>	Lista de plugins de Apache Nutch que serán incluidos durante la ejecución. Esta propiedad acepta una expresión regular. Como puede apreciarse, destacan 'protocol-httpclient' que es necesario para la descarga de recursos bajo HTTPS y 'webkumconnector' que es el nombre del plugin que se ha escrito para comunicar la aplicación web y la araña web.	protocol-httpclient urlfilter-(regex domain) parse-(html tika) index-(basic anchor) scoring-opic urlnormalizer-(pass regex basic) urlmeta webkumconnector
<code>plugin.folders</code>	Directorio donde se encuentra la estructura de directorios de los plugin de Nutch. Se generará una vez se compile.	/usr/local/nutch/runtime/local/plugins
<code>http.accept</code>	Tipos MIME que debe aceptar Apache Nutch durante sus operaciones de descarga.	text/html,application/xhtml+xml,application/javascript,text/css
<code>http.robots.agents</code>	Agente de usuario que Nutch buscará en el fichero robots.txt <b>[ROBOT]</b> durante	*

	la descarga de recursos. Sirve para indicarle a Nutch que respete las indicaciones de robots.txt.	
<code>db.ignore.internal.links</code>	Le da instrucciones a Apache Nutch para que no ignore los enlaces internos (el mismo dominio) al iniciar la descarga de los recursos partiendo de las semillas iniciales.	false
<code>db.ignore.external.links</code>	Le indica a Apache Nutch que ignore los enlaces a recursos externos (otro dominio). Esta propiedad, en conjunción con la anterior hace que Apache Nutch sólo descargue recursos bajo los dominios de las semillas indicadas.	true
<code>db.max.outlinks.per.page</code>	Número máximo de enlaces a analizar por recurso HTML. Por defecto esta propiedad está establecida a 100 enlaces lo cuál puede ser insuficiente en algunos casos.	9999
<code>file.crawl.parent</code>	Esta opción controla que Apache Nutch busque ficheros de semillas en directorios padres a los que se especifiquen. Se establece a falso para asegurar que Apache Nutch use únicamente el fichero de semillas que se le indique y no otros que puedan existir de ejecuciones previas.	false
<code>db.fetch.interval.default</code>	Intervalo por defecto que utilizará a Nutch para determinar cada cuánto tiempo tiene que volver a descargar los recursos.	0
<code>db.fetch.interval.max</code>	Intervalo máximo temporal que usa Nutch para determinar cada cuánto tiempo tiene que descargar los recursos que ya ha descargado en el pasado. Se usa esta propiedad junto con la anterior para indicarle a Nutch que siempre descargue los recursos nuevamente (para que distintos	0

	análisis sobre una misma semilla se ejecuten sobre los mismos recursos).
--	--

Otro aspecto que hay que modificar es el fichero `ivy.xml` del directorio `ivy` localizado en el directorio raíz de la instalación de Apache Nutch. Este fichero gestiona el sistema de dependencias Java de Apache Nutch mediante Apache Ivy [IVY]. Este sistema de gestión de dependencias funciona de forma similar a otros sistemas como Maven, Gradle o Sbt.

Se debe descomentar la línea que se muestra a continuación para permitir que Ivy descargue la versión del *driver* de Apache Gora que Nutch pueda utilizar Apache Cassandra mediante Gora.

```
<!--=====-->
<!-- Gora artifacts -->
<!--=====-->
<dependency org="org.apache.gora" name="gora-core" rev="0.3" conf="*->default"/>
<!-- Uncomment this to use SQL as Gora backend. It should be noted that the
gora-sql 0.1.1-incubating artifact is NOT compatible with gora-core 0.3. Users should
downgrade to gora-core 0.2.1 in order to use SQL as a backend. -->
<!--
<dependency org="org.apache.gora" name="gora-sql" rev="0.1.1-incubating" conf="*->default" />
-->
<!-- Uncomment this to use MySQL as database with SQL as Gora store. -->
<!--
<dependency org="mysql" name="mysql-connector-java" rev="5.1.18" conf="*->default"/>
-->
<!-- Uncomment this to use HBase as Gora backend. -->
<!--
<dependency org="org.apache.gora" name="gora-hbase" rev="0.3" conf="*->default" />
-->
<!-- Uncomment this to use Accumulo as Gora backend. -->
<!--
<dependency org="org.apache.gora" name="gora-accumulo" rev="0.3" conf="*->default" />
-->
<!-- Uncomment this to use Cassandra as Gora backend. -->
<dependency org="org.apache.gora" name="gora-cassandra" rev="0.3" conf="*->default" />
```

**Fig. 154: Modificación del fichero `ivy.xml`**

Otra de las acciones necesarias para que Apache Nutch utilice el *backend* de almacenamiento (Apache Cassandra) es descomentar otra línea en el fichero `gora.properties` del directorio `conf` situado en el directorio de instalación de Nutch. Se debe descomentar la línea que se ve a continuación:

```
#####
# CassandraStore properties #
#####

gora.cassandrastore.servers=localhost:9160
```

**Fig. 155: Configuración de Apache Gora mediante su fichero `properties`.**

Como servidor se debe especificar el mismo que se indicó en la configuración de Apache Cassandra (en este caso la máquina se llama webkumo-master como se pudo observar con anterioridad).

A continuación se realizarán modificaciones necesarias en la configuración para que Apache Nutch incluya el *plugin* WebKumo Connector en el proceso de compilación.

El primer paso es modificar el fichero `default.properties` que se encuentra en el directorio raíz de Apache Nutch para modificar la versión de Java contra la que se realizará la compilación. En este caso concreto en el que se está utilizando la versión 7 del JDK, se establecerá como sigue la propiedad `javac.version`:

```
javac.debug=on
javac.optimize=on
javac.deprecation=on
javac.version= 1.7
```

**Fig. 156: Configuración de la versión del compilador de Java.**

Además de indicar la versión del compilador a utilizar, se debe indicar en el fichero `build.xml` de configuración de la herramienta de compilación Apache Ant **[ANT]**; que se incluya el *plugin* WebKumo Connector desarrollado en el proceso de compilación. Para eso se debe comenzar creando un directorio `webkumo-connector` en `src/plugins` partiendo de la raíz de la instalación de Nutch. En dicho directorio se deben copiar los 3 ficheros XML proporcionados en el paquete `Java org.rz.webkumo.connector`. A continuación se debe crear la jerarquía de paquetes según se ve en la imagen a continuación y copiar los ficheros Java que contienen.

```
sergiorozas@webkumo-master:~/usr/local/nutch/src/plugin$ mkdir webkumo-connector
sergiorozas@webkumo-master:~/usr/local/nutch/src/plugin$ mkdir webkumo-connector/src
sergiorozas@webkumo-master:~/usr/local/nutch/src/plugin$ mkdir webkumo-connector/src/java
sergiorozas@webkumo-master:~/usr/local/nutch/src/plugin$ cp ~/src/main/java/org/rz/webkumo/connector/*.xml webkumo-connector/
sergiorozas@webkumo-master:~/usr/local/nutch/src/plugin$ cp ~/src/main/java/org/webkumo-connector/src/java/
sergiorozas@webkumo-master:~/usr/local/nutch/src/plugin$ rm -R webkumo-connector/src/java/org/rz/webkumo/business/
sergiorozas@webkumo-master:~/usr/local/nutch/src/plugin$ rm -R webkumo-connector/src/java/org/rz/webkumo/conf/spring/
sergiorozas@webkumo-master:~/usr/local/nutch/src/plugin$ rm -R webkumo-connector/src/java/org/rz/webkumo/view/
sergiorozas@webkumo-master:~/usr/local/nutch/src/plugin$ rm -R webkumo-connector/src/java/org/rz/webkumo/persistence/
sergiorozas@webkumo-master:~/usr/local/nutch/src/plugin$
```

**Fig. 157: Copia de los ficheros de código de WebKumo Connector.**

Se eliminan los paquetes `business`, `conf/spring`, `view` y `persistence` por innecesarios para el *plugin* WebKumo Connector.

Además de copiar el código en sí, hay que indicarle a Apache Ant que lo compile junto al resto de *plugins* que se incluyen con Apache Nutch. Se deben añadir las siguientes líneas al fichero `build.xml` del directorio `src/plugin` que se encuentra en el directorio de instalación de Apache Nutch. Para cada elemento `target` del fichero XML se debe añadir un elemento hijo `ant` para WebKumo Connector en el que se haga casar el nombre del `target` como el del elemento hijo. Los elementos a añadir son :

- ◆ `<ant dir="webkumo-connector" target="test"/>`
- ◆ `<ant dir="webkumo-connector" target="test"/>`
- ◆ `<ant dir="webkumo-connector" target="test"/>`

Se muestra a modo de ejemplo uno de los elementos introducidos:

```

<!-- ===== -->
<!-- Build & deploy all the plugin jars. -->
<!-- ===== -->
<target name="deploy">

    <ant dir="webkumo-connector" target="deploy"/>

    <ant dir="creativecommons" target="deploy"/>
    <ant dir="index-anchor" target="deploy"/>
    <ant dir="index-basic" target="deploy"/>

```

**Fig. 158: Configuración del fichero `build.xml` para los `plugins` de Apache Nutch.**

Análogamente a lo realizado con el fichero `build.xml` de `src/plugin`, hay que modificar el fichero `build.xml` que se encuentra en la raíz de Apache Nutch. Dentro de los elementos `target` con nombres `release` y `javadoc` se debe añadir

- ◆ `<packageset dir="${plugins.dir}/webkumo-connector/src/java"/>`

Si se ha instalado Apache Hadoop en un directorio distinto al que se indica en este manual (`/usr/local/hadoop`), debe indicarse la ruta correcta de instalación en el fichero `plugin.xml` bajo el elemento con nombre `webkumo_connector_hadoop_home` así como indicar el nombre de dominio (con el puerto correcto, 9000 por defecto) en el elemento con nombre `webkumo_connector_hdfs` que hace la función de punto de acceso al sistema de archivos distribuido HDFS. Este paso es necesario para el funcionamiento de WebKumo con la implementación de línea de comandos del servicio Apache Nutch.

No se requieren más configuraciones que realizar llegado a este paso. El paso siguiente es realizar la compilación de Apache Nutch usando la herramienta `ant` (que debe ser instalada previamente mediante el gestor de paquetes de la distribución). Se procede ejecutando el comando `ant clean && ant runtime` desde el directorio de instalación de Apache Nutch donde se encuentra el fichero `build.xml` que le indica a `ant` cómo debe construir Apache Nutch. Al finalizar la compilación debe mostrarse un mensaje `BUILD SUCCESSFUL` como sigue:

```
runtime:
[mkdir] Created dir: /usr/local/nutch/runtime
[mkdir] Created dir: /usr/local/nutch/runtime/local
[mkdir] Created dir: /usr/local/nutch/runtime/deploy
[copy] Copying 1 file to /usr/local/nutch/runtime/deploy
[copy] Copying 2 files to /usr/local/nutch/runtime/deploy/bin
[copy] Copying 1 file to /usr/local/nutch/runtime/local/lib
[copy] Copying 1 file to /usr/local/nutch/runtime/local/lib/native
[copy] Copying 26 files to /usr/local/nutch/runtime/local/conf
[copy] Copying 2 files to /usr/local/nutch/runtime/local/bin
[copy] Copying 118 files to /usr/local/nutch/runtime/local/lib
[copy] Copying 179 files to /usr/local/nutch/runtime/local/plugins
[copy] Copied 2 empty directories to 2 empty directories under /usr/local/nutch/runtime/local/test

BUILD SUCCESSFUL
Total time: 58 seconds
sergiorozas@webkumo-master:/usr/local/nutch$
```

**Fig. 159: Compilación exitosa de Apache Nutch.**

Se debe por último copiar las librerías (JAR) descargadas durante el proceso de compilación al directorio de librerías de Apache Hadoop, esto es, copiar los ficheros JAR que se encuentran en el directorio `runtime/local/lib` y `runtime/local/plugins/webkumo` al directorio `lib` localizado en la instalación de Apache Hadoop. Esto sólo es necesario si se va a utilizar la implementación `org.rz.webkumo.model.services.NutchService` que ejecuta Apache Nutch mediante su API Java.

```
sergiorozas@webkumo-master:/usr/local/nutch$ cp runtime/local/lib/*.jar /usr/local/hadoop/lib
sergiorozas@webkumo-master:/usr/local/nutch$ cp runtime/local/plugins/webkumo-connector/*.jar /usr/local/hadoop/lib
sergiorozas@webkumo-master:/usr/local/nutch$
```

**Fig. 160: Copia de las librerías JAR descargadas.**

Aunque ya está Apache Nutch instalado llegado a este punto, se procederá a continuación a describir el proceso de prueba de instalación. Se requiere para esta prueba que todos los sistemas de infraestructura (Apache Hadoop, Apache Solr y Apache Cassandra) estén en ejecución (puede consultarse cómo iniciarlos en los manuales de instalación anteriores).

Se comienza creando un directorio `urls` de que contendrá el fichero de texto con las semillas y añadiendo dicha semilla al filtro de urls. Obsérvese la imagen siguiente:

```
sergiorozas@webkumo-master:/usr/local/nutch/runtime/local$ mkdir urls
sergiorozas@webkumo-master:/usr/local/nutch/runtime/local$ echo "http://www.uniovi.es/" >> urls/seeds.txt
sergiorozas@webkumo-master:/usr/local/nutch/runtime/local$ echo "http://www.uniovi.es/" >> conf/domain-urlfilter.txt
sergiorozas@webkumo-master:/usr/local/nutch/runtime/local$
```

**Fig. 161: Creación del fichero de semillas**

A continuación se debe acceder al directorio `runtime/local/bin` desde la carpeta de instalación de Nutch y se enviarán los comandos que se muestran a continuación que ejecutará una iteración del ciclo de ejecución de Apache Nutch. Lo primero es la fase de inyección que se ejecuta mediante lo siguiente:

```
sergiorozas@webkumo-master:/usr/local/nutch/runtime/local/bin$ ./crawl ../urls/ -depth 3 -topN 5
InjectorJob: starting at 2015-06-14 16:16:09
InjectorJob: Injecting urlDir: ../urls
InjectorJob: Using class org.apache.gora.cassandra.store.CassandraStore as the Gora storage class.
InjectorJob: total number of urls rejected by filters: 0
InjectorJob: total number of urls injected after normalization and filtering: 1
Injector: finished at 2015-06-14 16:16:13, elapsed: 00:00:03
sergiorozas@webkumo-master:/usr/local/nutch/runtime/local/bin$
```

**Fig. 162: Inyección de semillas**

En este momento se debe comprobar que Apache Nutch haya creado el *keyspace* **[KEYSP]** webpage en Apache Cassandra mediante la ejecución del intérprete `cqlsh` del que ya se habló en el manual de instalación de Apache Cassandra.

```
sergiorozas@webkumo-master:/usr/local/nutch/runtime/local/bin$ /usr/local/cassandra/bin/cqlsh
Connected to Test Cluster at 127.0.0.1:9042.
[cqlsh 5.0.1 | Cassandra 2.1.6 | CQL spec 3.2.0 | Native protocol v3]
Use HELP for help.
cqlsh> describe keyspaces;

webpage system_traces system
cqlsh>
```

**Fig. 163: Keyspace webpage presente en Apache Cassandra**

El siguiente paso en la fase de ejecución de Nutch es generar el trabajo MapReduce que se ejecutará en Apache Hadoop. Se realiza la generación mediante el comando que se observa a continuación:

```
sergiorozas@webkumo-master:/usr/local/nutch/runtime/local/bin$ ./nutch generate -topN 5
GeneratorJob: starting at 2015-06-14 16:34:16
GeneratorJob: Selecting best-scoring urls due for fetch.
GeneratorJob: starting
GeneratorJob: filtering: true
GeneratorJob: normalizing: true
GeneratorJob: topN: 5
GeneratorJob: finished at 2015-06-14 16:34:20, time elapsed: 00:00:03
GeneratorJob: generated batch id: 1434292456-1483253251
sergiorozas@webkumo-master:/usr/local/nutch/runtime/local/bin$
```

**Fig. 164: Ejecución de la fase de generación de Nutch.**

Puede verse en la captura que se ha generado un trabajo con número de lote 1434292456-1483253251. El siguiente paso es realizar la descarga (fetch) según se ve a continuación:

```
sergiorozas@webkumo-master:/usr/local/nutch/runtime/local/bin$ ./nutch fetch 1434292456-1483253251
FetcherJob: starting
FetcherJob: batchId: 1434292456-1483253251
Fetcher: Your 'http.agent.name' value should be listed first in 'http.robots.agents' property.
FetcherJob: threads: 10
FetcherJob: parsing: false
FetcherJob: resuming: false
FetcherJob : timelimit set for : -1
Using queue mode : byHost
Fetcher: threads: 10
QueueFeeder finished: total 1 records. Hit by time limit :0
fetching http://www.uniovi.es/ (queue crawl delay=5000ms)
Fetcher: throughput threshold: -1
Fetcher: throughput threshold sequence: 5
-finiishing thread FetcherThread1, activeThreads=8
-finiishing thread FetcherThread2, activeThreads=7
-finiishing thread FetcherThread7, activeThreads=6
-finiishing thread FetcherThread6, activeThreads=5
-finiishing thread FetcherThread5, activeThreads=4
-finiishing thread FetcherThread4, activeThreads=3
-finiishing thread FetcherThread3, activeThreads=2
-finiishing thread FetcherThread8, activeThreads=1
-finiishing thread FetcherThread9, activeThreads=1
-finiishing thread FetcherThread0, activeThreads=0
0/0 spinwaiting/active, 1 pages, 0 errors, 0.2 0 pages/s, 66 65 kb/s, 0 URLs in 0 queues
-activeThreads=0
FetcherJob: done
sergiorozas@webkumo-master:/usr/local/nutch/runtime/local/bin$
```

**Fig. 165: Ejecución de la fase de descarga de Nutch.**

En la imagen anterior se ve que se ha descargado el recurso que se encuentra en la URL de la semilla especificada. La siguiente fase es ejecutar el procesado (*parsing*) de lo descargado. En esta fase es donde WebKumo Connector se ejecutará aunque es irrelevante para esta prueba.

```
sergiorozas@webkumo-master:/usr/local/nutch/runtime/local/bin$ ./nutch parse 1434292456-1483253251
ParserJob: starting
ParserJob: resuming:      false
ParserJob: forced reparse: false
ParserJob: batchId:      1434292456-1483253251
Configuration: nutch-default.xml, nutch-site.xml, /usr/local/hadoop/conf/core-site.xml
Configuration: nutch-default.xml, nutch-site.xml, /usr/local/hadoop/conf/core-site.xml
Parsing http://www.uniovi.es/
ParserJob: success
sergiorozas@webkumo-master:/usr/local/nutch/runtime/local/bin$
```

**Fig. 166: Ejecución de la fase de procesado de Nutch.**

A continuación se debe actualizar Apache Cassandra con lo descargado mediante el comando que inicia la fase updatedb que se ve a continuación:

```
sergiorozas@webkumo-master:/usr/local/nutch/runtime/local/bin$ ./nutch updatedb
DbUpdaterJob: starting
DbUpdaterJob: done
sergiorozas@webkumo-master:/usr/local/nutch/runtime/local/bin$
```

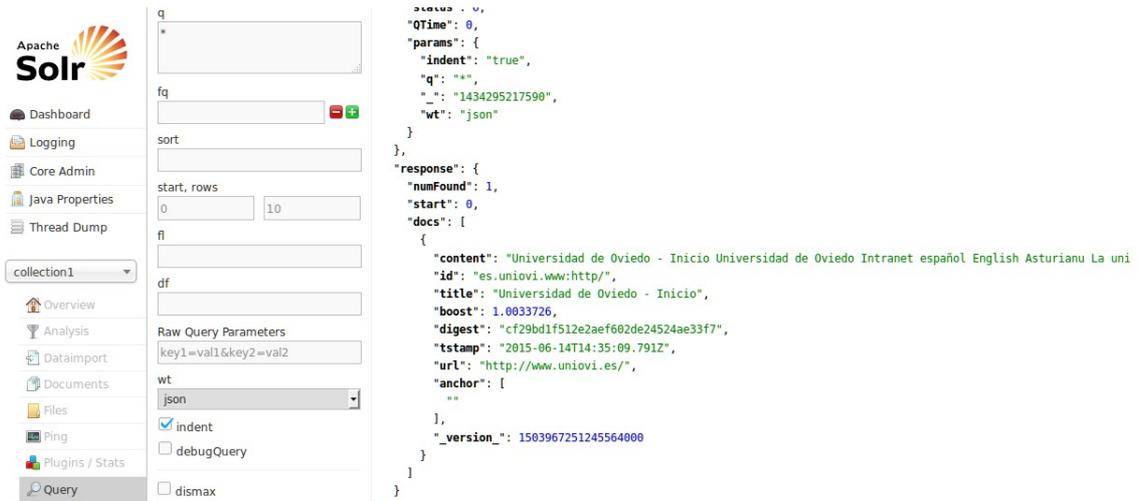
**Fig. 167: Ejecución de la fase de actualización de Apache Cassandra.**

Finalmente se procede a enviar a Apache Solr el contenido que Apache Nutch ha descargado y almacenado en Apache Cassandra. Se debe ejecutar el comando que se muestra en la siguiente imagen:

```
sergiorozas@webkumo-master:/usr/local/nutch/runtime/local/bin$ ./nutch solrindex http://localhost:8983/solr -reindex
SolrIndexerJob: starting
Adding 1 documents
SolrIndexerJob: done.
sergiorozas@webkumo-master:/usr/local/nutch/runtime/local/bin$
```

**Fig. 168: Ejecución de la fase de indexado de Apache Nutch.**

Con esto se habrá finalizado la ejecución manual de Apache Nutch. Para finalizar se comprobará que el recurso descargado (la página web principal de la Universidad de Oviedo) ha sido indexado por Apache Solr, lo que confirmará el éxito del proceso. Para comprobarlo se accede a la interfaz web de Apache Solr (que acepta peticiones HTTP en el puerto 8983), se selecciona a continuación la colección de documentos por defecto `collection1` en el selector core de la parte izquierda de la interfaz. A continuación se seleccionará “Query” en el menú de la colección lo cual cargará un formulario para ejecutar consultas sobre Solr. En el campo “q” del formulario se prueba a escribir “\*” para buscar todos los documentos indexados. Finalmente se ejecutará la consulta mediante el botón del formulario “Execute Query” y se comprueba que se indexó el documento como puede verse en la captura siguiente:



**Fig. 169:** Documento principal del portal web de la Universidad de Oviedo indexado en Solr.

### 9.1.5 Contenedor de aplicaciones J2EE

La aplicación web WebKumo se ejecuta sobre un servidor web de aplicaciones J2EE. Se utilizará Apache Tomcat **[TOMC1]** en su versión 8 por ser el servidor mas popular **[TOMC2]** de aplicaciones web empresariales de Java. El uso de este servidor garantiza un alto nivel de soporte disponible así como la existencia de actualizaciones frecuentes.

Para la instalación de Apache Tomcat, puede utilizarse el empaquetado en el repositorio de paquetes de la distribución GNU/Linux que se esté utilizando. En este caso, para facilitar el proceso, se utilizará la distribución binaria (debido a que sirve para cualquier distribución).

Es un proceso sencillo, sólo hay que descargar el software en su versión core (comprimido en zip o tar .gz) y descomprimirlo como se ve a continuación:

```
sergiorozas@webkumo-master:~$ wget http://ftp.cixug.es/apache/tomcat/tomcat-8/v8.0.23/bin/apache-tomcat-8.0.23.tar.gz
--2015-06-14 18:14:26-- http://ftp.cixug.es/apache/tomcat/tomcat-8/v8.0.23/bin/apache-tomcat-8.0.23.tar.gz
Resolviendo ftp.cixug.es (ftp.cixug.es)... 193.144.61.75
Conectando con ftp.cixug.es (ftp.cixug.es)[193.144.61.75]:80... conectado.
Petición HTTP enviada, esperando respuesta... 200 OK
Longitud: 9076980 (8,7M) [application/x-gzip]
Grabando a: "apache-tomcat-8.0.23.tar.gz"

apache-tomcat-8.0.23.tar.gz 100%[=====] 8,66M 10,2MB/s en 0,8s
2015-06-14 18:14:41 (10,2 MB/s) - "apache-tomcat-8.0.23.tar.gz" guardado [9076980/9076980]
sergiorozas@webkumo-master:~$ tar -xzf apache-tomcat-8.0.23.tar.gz
sergiorozas@webkumo-master:~$
```

**Fig. 170:** Descarga y descompresión de Apache Tomcat

Además de lo anterior se debe configurar un usuario de Apache Tomcat que se utilizará para controlar el servidor desde otras aplicaciones (necesario para lo que se mostrará en el manual de ejecución). Para lo comentado se debe modificar el fichero `tomcat-users.xml` localizado en el directorio `conf` que se encuentra en el directorio creado al descomprimir de Apache Tomcat

```
<user password="contraseña" roles="manager-script,admin" username="usuario"/>
```

**Fig. 171: Configuración del usuario de Apache Tomcat**

Como puede observarse se añadió un elemento `user` como hijo del elemento principal `tomcat-users`. Se debe sustituir `contraseña` y `usuario` por las credenciales que se deseen.

## 9.2 Manual de ejecución

Esta sección contiene las instrucciones necesarias para poner en marcha el sistema, tanto los servicios de infraestructura como la aplicación web WebKumo. Se indicará además como detener los mismos.

Para el arranque y/o detención de los servicios existen dos opciones principalmente, o realizarlo de forma manual mediante el comando apropiado, o bien mediante la interfaz web de WebKumo. Esta última manera se cubrirá en el manual del usuario.

### 9.2.1 Apache Hadoop

Para arrancar Apache Hadoop, como ya se indicó en el manual de instalación se debe ejecutar el script `start-all.sh` que se encuentra en el directorio `bin` localizado en el directorio de instalación de Hadoop (`/usr/local/hadoop` si se ha seguido el manual contenido en esta documentación). Esto debe ejecutarse mediante el usuario que se haya preparado para que ejecute Apache Hadoop (ver manual de instalación para mas aclaraciones):

```
sergiorozas@webkumo-master:/usr/local/hadoop/bin$ ./start-all.sh
Warning: $HADOOP_HOME is deprecated.

starting namenode, logging to /usr/local/hadoop/libexec/../logs/hadoop-sergiorozas-namenode-webkumo-master.out
localhost: starting datanode, logging to /usr/local/hadoop/libexec/../logs/hadoop-sergiorozas-datanode-webkumo-master.out
localhost: starting secondarynamenode, logging to /usr/local/hadoop/libexec/../logs/hadoop-sergiorozas-secondarynamenode-webkumo-master.out
starting jobtracker, logging to /usr/local/hadoop/libexec/../logs/hadoop-sergiorozas-jobtracker-webkumo-master.out
localhost: starting tasktracker, logging to /usr/local/hadoop/libexec/../logs/hadoop-sergiorozas-tasktracker-webkumo-master.out
sergiorozas@webkumo-master:/usr/local/hadoop/bin$
```

**Fig. 172: Ejecución del script de inicio de Apache Hadoop.**

En el caso de que se desee detener Apache Hadoop se debe ejecutar el script `stop-all.sh` localizado en el mismo directorio que el script de arranque.

```
sergiorozas@webkumo-master:/usr/local/hadoop/bin$ ./stop-all.sh
Warning: $HADOOP_HOME is deprecated.

no jobtracker to stop
localhost: no tasktracker to stop
stopping namenode
localhost: no datanode to stop
localhost: stopping secondarynamenode
sergiorozas@webkumo-master:/usr/local/hadoop/bin$
```

**Fig. 173: Ejecución del script de parada de Apache Hadoop.**

## 9.2.2 Apache Solr

En el caso de Apache Solr, se debe ejecutar el script `solr` localizado en el directorio `bin` que se encuentra en el directorio de instalación de Apache Solr (`/usr/local/solr` si se ha seguido el manual de instalación anterior). Al script se le debe indicar el parámetro `start` como se ve a continuación.

```
sergiorozas@webkumo-master:/usr/local/solr/bin$ ./solr start
Waiting to see Solr listening on port 8983 [\]
Started Solr server on port 8983 (pid=14791). Happy searching!

sergiorozas@webkumo-master:/usr/local/solr/bin$ █
```

**Fig. 174: Ejecución del script de inicio de Apache Solr**

Para detener Apache Solr hay que ejecutar el mismo script que para iniciarlo pero con los parámetros `stop -all`.

```
sergiorozas@webkumo-master:/usr/local/solr/bin$ ./solr stop -all
Sending stop command to Solr running on port 8983 ... waiting 5 seconds to allow Jetty process 5145 to stop gracefully.
sergiorozas@webkumo-master:/usr/local/solr/bin$ █
```

**Fig. 175: Ejecución del script de parada de Apache Solr**

## 9.2.3 Apache Cassandra

Para el software de persistencia de la araña web que se ha utilizado en este proyecto, Apache Cassandra, debe ejecutarse el *script* de arranque `cassandra` localizado en la carpeta `bin` de instalación (`/usr/local/cassandra` si se ha seguido la guía del capítulo anterior).

```
539514210405550827, -7579751717351464778, -7824392738680614470, -78919751906547408, -7895690584307796738, -7954432315323015590
, -7986722066726611744, -7994322090485184680, -8222399236367606456, -8460266841558606081, -8473405694826705797, -8570026643257
678878, -8639243844637632098, -8786129812243691503, -8797409421014124737, -8838721983500565505, -89039702562867634482, -9074279
221892713826, -9127074768534315827, -945117827468601167, 1110115891504761865, 12042453644779304913, 1277294123641458588, 130241
5567343495084, 1321766604385369627, 1348199780952099682, 1348199780952099682, 1407789077904683980, 1442855167511334282, 145561
7890199398393, 1515344483667538068, 1518195827194580127, 1529117781791939856, 154342306284550271, 1591965512150294501, 1640366
907231732921, 1690069136838404634, 1839472724055535845, 1845176145428435446, 1872324400137901792, 1874197418178320000, 1932252
601154117800, 1965748548425910078, 1976712688164448825, 2120436929243859041, 2181771140012945486, 2204006832193826883, 2216545
152775644843, 2353215256079356746, 2458812769520899903, 2549683895494803351, 2591650879612538774, 2652995945728470602, 2942922
148690419029, 2956960460076878240, 3041117769121949314, 30969939487970470, 3111256184974537200, 3196647834418304622, 3225925
013064854836, 3249405344050366948, 3320026785334793759, 3327632635996067396, 3431010381669921631, 3491576573743818957, 3499522
789108811390, 3515852818890102281, 3517733087383665905, 3537236179687065683, 3613127272889518353, 3617113214573346991, 3625432
363519794056, 3648447806158993046, 3681130204971433387, 3753059868369684625, 3798438752177920048, 3810199702684344928, 3927375
967661737351, 4004345995775070187, 4095403558569313486, 427216954374618117, 4347059160826817194, 4397688767037395663, 44325698
06945229588, 448212233625708268, 4497200305543758829, 456433829219104301, 4656118487749990708, 4704416712571674162, 4786785575
216359178, 4864231646212481166, 4955656878891517906, 5020770986900058975, 5044907402392348551, 5097817648901074474, 5110922132
879758756, 5121926728608282621, 513605729191697524, 5178178075983112049, 5236090271644786902, 5256089204956300200, 52630046977
86873454, 5263217207783558955, 5334345875040773798, 5360238047943377339, 5425459499481728867, 542776508367005763, 544142828246
5705820, 55267846604335348180, 5540866562209380532, 5776451443704787062, 5835287001854506727, 5958550776681090994, 618585195090
2988268, 6257315475146347426, 6337211314358027011, 6526238390310100273, 6559008120705117706, 6579324103303855719, 658809931294
9511457, 6698843590294911199, 6809051217313941919, 6979402782279431485, 7017155559972670215, 7206650900590360665, 7250968049547
710268, 7259885773586602269, 7290397954327981674, 7293854144101750628, 7344842652817060086, 735538812839964365, 74066980226257
92797, 7534395228911767269, 7584266589932108354, 7682182228312305799, 7910385665727253256, 7964488914878111707, 81739454958798
03271, 8175502882238118401, 8216045082998124838, 8314682429037480020, 8317706627884399937, 8389257658231589884, 87757697247844
83863, 8891084258923089624, 903185997179177249, 9170392737250432027, 96601115407555048]
INFO 12:22:38 Enqueueing flush of local: 716 (0%) on-heap, 0 (0%) off-heap
INFO 12:22:38 Writing Memtable-local@964658829(112 serialized bytes, 4 ops, 0%/0% of on/off-heap limit)
INFO 12:22:38 Completed flushing /usr/local/cassandra/bin/././data/data/system/local-7ad54392bccdd35a684174e047860b377/system
-local-ka-18-Data.db (115 bytes) for commitlog position ReplayPosition(segmentId=1434198155149, position=97038)
INFO 12:22:38 Enqueueing flush of local: 51474 (0%) on-heap, 0 (0%) off-heap
INFO 12:22:38 Writing Memtable-local@1115288483(8560 serialized bytes, 259 ops, 0%/0% of on/off-heap limit)
INFO 12:22:38 Compacted 4 sstables to [././data/data/system/local-7ad54392bccdd35a684174e047860b377/system-local-ka-17,]. 11
.419 bytes to 5.724 (-50% of original) in 292ms = 0,018695MB/s. 4 total partitions merged to 1. Partition merge counts were
{4:1, }
INFO 12:22:39 Completed flushing /usr/local/cassandra/bin/././data/data/system/local-7ad54392bccdd35a684174e047860b377/system
-local-ka-19-Data.db (5264 bytes) for commitlog position ReplayPosition(segmentId=1434198155149, position=109597)
INFO 12:22:39 Node localhost/127.0.0.1 state jump to normal
sergiorozas@webkumo-master:~/usr/local/cassandra/bin$
```

Fig. 176: Ejecución del script de arranque de Apache Cassandra

Por desgracia Apache Cassandra no proporciona ningún script ni otro mecanismo de detención, por lo que para parar Cassandra es necesario enviarle la señal SIGKILL **[SIGKIL]** al proceso de Cassandra que se está ejecutando en el marco del sistema operativo. Como puede verse en la imagen siguiente, se ejecuta ps ax para listar los procesos filtrando su salida mediante grep para facilitar la búsqueda del proceso de Apache Cassandra. Una vez encontrado se le envía la señal mediante kill -9 <PID> donde PID es el identificador de proceso de Cassandra que se aprecia en la imagen:

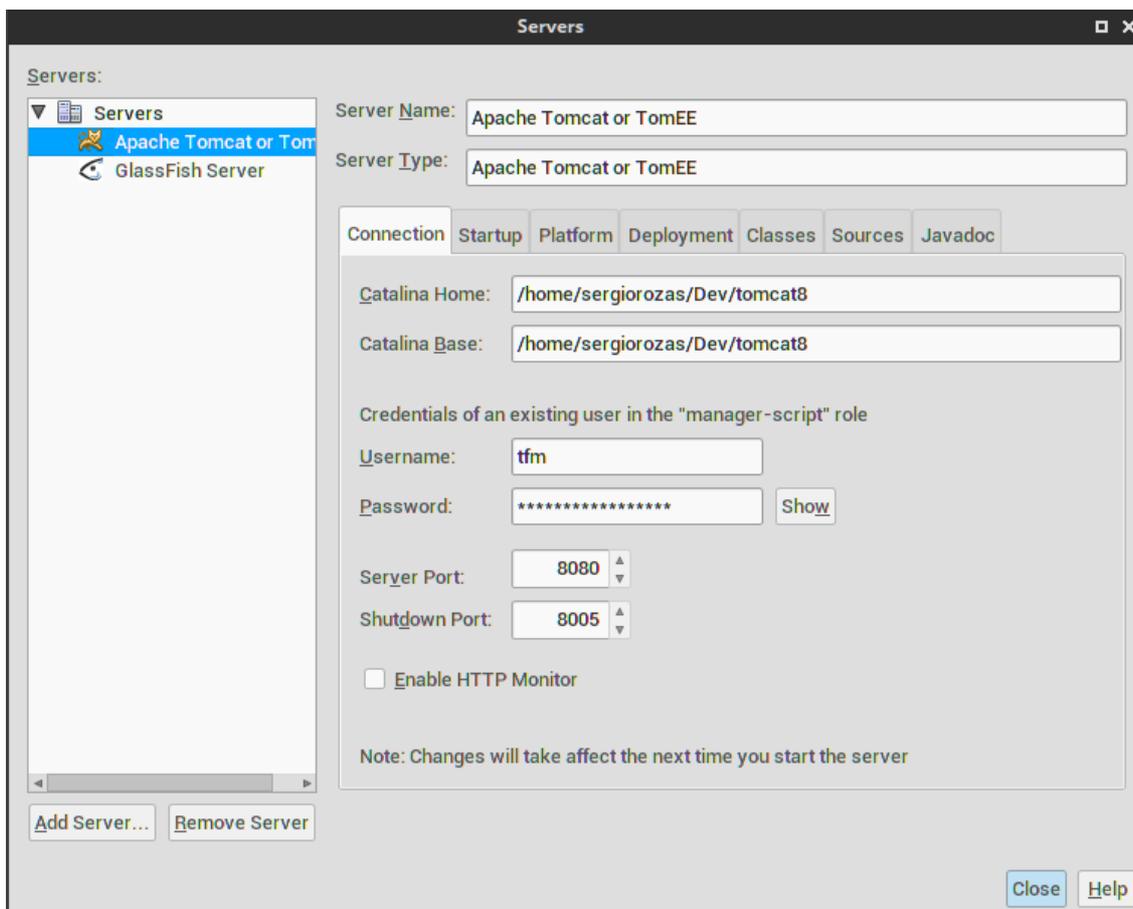
```
sergiorozas@webkumo-master:~$ ps ax | grep cassandra
3223 pts/0 Ssl 1:22 /usr/lib/jvm/java-7-openjdk-amd64/bin/java -ea -javaagent:/usr/local/cassandra/bin/./lib/jamm-0.3.
0.jar -XX:+CMSClassUnloadingEnabled -XX:UseThreadPriorities -XX:ThreadPriorityPolicy=42 -Xms500M -Xmx500M -Xmn100M -XX:+HeapD
umpOnOutOfMemoryError -Xss256k -XX:StringTableSize=1000003 -XX:+UseParNewGC -XX:+UseConcMarkSweepGC -XX:+CMSParallelRemarkEnab
led -XX:SurvivorRatio=8 -XX:MaxTenuringThreshold=1 -XX:CMSInitiatingOccupancyFraction=75 -XX:+UseCMSInitiatingOccupancyOnly -X
X:+UseTLAB -XX:CompileCommandFile=/usr/local/cassandra/bin/./conf/hotspot_compiler -XX:CMSWaitDuration=10000 -XX:+CMSParallel
InitialMarkEnabled -XX:+CMSEdenChunksRecordAlways -XX:CMSWaitDuration=10000 -XX:+UseCondCardMark -Djava.net.preferIPv4Stack=tr
ue -Dcassandra.jmx.local.port=7199 -XX:+DisableExplicitGC -Dlogback.configurationFile=logback.xml -Dcassandra.logdir=/usr/loca
l/cassandra/bin/./logs -Dcassandra.storagedir=/usr/local/cassandra/bin/./data -cp /usr/local/cassandra/bin/./conf:/usr/loca
l/cassandra/bin/./build/classes/main:/usr/local/cassandra/bin/./build/classes/thrift:/usr/local/cassandra/bin/./lib/ST4-4.0
.8.jar:/usr/local/cassandra/bin/./lib/airline-0.6.jar:/usr/local/cassandra/bin/./lib/antlr-runtime-3.5.2.jar:/usr/local/cass
andra/bin/./lib/apache-cassandra-2.1.6.jar:/usr/local/cassandra/bin/./lib/apache-cassandra-clientutil-2.1.6.jar:/usr/local/c
assandra/bin/./lib/apache-cassandra-thrift-2.1.6.jar:/usr/local/cassandra/bin/./lib/commons-cli-1.1.jar:/usr/local/cassandra
/bin/./lib/commons-codec-1.2.jar:/usr/local/cassandra/bin/./lib/commons-lang3-3.1.jar:/usr/local/cassandra/bin/./lib/common
s-math3-3.2.jar:/usr/local/cassandra/bin/./lib/compress-lzf-0.8.4.jar:/usr/local/cassandra/bin/./lib/concurrent-linkedhashmap
-lru-1.4.jar:/usr/local/cassandra/bin/./lib/disruptor-3.0.1.jar:/usr/local/cassandra/bin/./lib/guava-16.0.jar:/usr/local/cas
sandra/bin/./lib/high-scale-lib-1.0.6.jar:/usr/local/cassandra/bin/./lib/jackson-core-asl-1.9.2.jar:/usr/local/cassandra/bin
/./lib/jackson-mapper-asl-1.9.2.jar:/usr/local/cassandra/bin/./lib/jamm-0.3.0.jar:/usr/local/cassandra/bin/./lib/javax.inje
ct.jar:/usr/local/cassandra/bin/./lib/jbcrypt-0.3m.jar:/usr/local/cassandra/bin/./lib/jline-1.0.jar:/usr/local/cassandra/bin
/./lib/jna-4.0.0.jar:/usr/local/cassandra/bin/./lib/json-simple-1.1.jar:/usr/local/cassandra/bin/./lib/libthrift-0.9.2.jar:/
usr/local/cassandra/bin/./lib/logback-classic-1.1.2.jar:/usr/local/cassandra/bin/./lib/logback-core-1.1.2.jar:/usr/local/cas
sandra/bin/./lib/lz4-1.2.0.jar:/usr/local/cassandra/bin/./lib/metrics-core-2.2.0.jar:/usr/local/cassandra/bin/./lib/netty-
all-4.0.23.Final.jar:/usr/local/cassandra/bin/./lib/reporter-config-2.1.0.jar:/usr/local/cassandra/bin/./lib/slf4j-api-1.7.2
.jar:/usr/local/cassandra/bin/./lib/snakeyaml-1.11.jar:/usr/local/cassandra/bin/./lib/snappy-java-1.0.5.2.jar:/usr/local/cas
sandra/bin/./lib/stream-2.5.2.jar:/usr/local/cassandra/bin/./lib/super-csv-2.1.0.jar:/usr/local/cassandra/bin/./lib/thrift-
server-0.3.7.jar org.apache.cassandra.service.CassandraDaemon
7259 pts/0 S+ 0:00 grep cassandra
sergiorozas@webkumo-master:~$ kill -9 3223
sergiorozas@webkumo-master:~$ ps ax | grep cassandra
7262 pts/0 S+ 0:00 grep cassandra
sergiorozas@webkumo-master:~$
```

Fig. 177: Detención de Apache Cassandra mediante la parada de su proceso

## 9.2.4 Aplicación web WebKumo

El último componente cuyo arranque y parada se explicará es la aplicación web WebKumo que realiza el papel de intermediario entre la infraestructura de computación y análisis y el usuario. Se trata de una aplicación web J2EE gestionada mediante Apache Maven por lo que se ejemplificará su arranque usando un entorno de desarrollo (IDE) con soporte para Maven. En concreto se utilizará Netbeans en su última versión (8.0.2). Otros entornos como Eclipse o IntelliJ funcionan de manera similar.

Se comienza añadiendo el servidor Apache Tomcat configurado en el manual de instalación anterior para que sea gestionado por el IDE. Se especificarán las credenciales indicadas en el fichero `tomcat-users.xml` anteriormente así como la ruta de instalación creada.



**Fig. 178: Configuración de Apache Tomcat en Netbeans**

Con esto ya se podrá arrancar la aplicación web pero primero se describirán ciertos parámetros de configuración de la aplicación. La configuración se controla mediante los ficheros *properties* de la carpeta *resources/conf* del proyecto Maven.

Para arrancar la aplicación web, se utilizará la utilidad del IDE que permite tal cosa. Internamente, se ejecutará el *goal* u objetivo Maven correspondiente al arranque que es el *goal install*.

### 9.2.4.1 Initialization.properties

El fichero *initialization.properties* se encarga de definir el perfil **[SPBOO]** en que se ejecutará la aplicación web. Existen dos perfiles cargados:

1. **Development:** Perfil para la fase de desarrollo, incluye el uso de una base de datos empotrada (almacenamiento en memoria) y autenticación en memoria (sin usuarios en la base de datos). Este perfil es idóneo para tareas de desarrollo ya que elimina la necesidad de desplegar una instancia de una base de datos, reduciendo la carga del sistema.
2. **Production:** Este perfil ya está pensado para utilizar una base de datos separadamente cuyos datos no se perderán al detener el servidor web. La autenticación en este perfil se realiza contra la base de datos.

Para arrancar la aplicación en un perfil u otro se debe descomentar la línea adecuada en el fichero *properties*. Se muestran las dos opciones en la siguiente imagen:

```
webkumo.profile=development
#webkumo.profile=production
```

**Fig. 179: Valores válidos de *initialization.properties*.**

### 9.2.4.2 Datasource.properties

Este fichero controla la configuración de la base de datos que WebKumo utilizará para su funcionamiento. Se divide la configuración en dos partes *spring.dev.\** y *spring.production.\** para las respectivas propiedades en función del perfil de la aplicación.

En el caso de las propiedades para el perfil de desarrollo, se tiene únicamente el tipo de base de datos empotrada que se utilizará en dicho perfil (H2 Database **[H2]** por defecto). Los valores válidos para dicha propiedad son H2 o HSQLDB para la base de datos HyperDatabase **[HSQLDB]**. Nótese que dichas bases de datos están empotradas y se incluyen mediante la dependencia Maven correspondiente en el fichero pom.xml que se incluye en el proyecto.

En cuanto a las propiedades del perfil de producción se incluyen la URI al punto de acceso a la base de datos, el nombre de usuario y la contraseña de acceso así como el driver JDBC **[JDBC]** que el software ORM debe utilizar. Para la propiedad spring.production.datasource.type, los valores válidos son MYSQL (MySQL), POSTGRESQL (PostgreSQL), ORACLE (Oracle DB) o SQL\_SERVER (Microsoft SQL Server) en función de la implementación concreta de la base de datos a utilizar.

```
#####
# Development DB.
spring.dev.datasource.type=H2

#####
# Production DB
spring.production.datasource.type=MYSQL
spring.production.datasource.uri=jdbc:mysql://localhost/webkumo
spring.production.datasource.username=user
spring.production.datasource.password=pass
spring.production.datasource.driver-class-name=com.mysql.jdbc.Driver
```

**Fig. 180: Fichero de configuración datasource.properties**

### 9.2.4.3 Infrastructure.properties

La configuración de los diversos servicios de infraestructura se controla mediante este fichero de propiedades que WebKumo utiliza para inicializar los servicios que utiliza internamente.

Se tienen las siguientes opciones de configuración:

INFRASTRUCTURE.PROPERTIES		
Nombre	Descripción	Valor recomendado
jpa.production.database	Tipo de base de datos <b>[SPDATA1]</b> que se utilizará bajo el perfil de producción.	MYSQL

<code>jpa.production.showsql</code>	Determina si las consultas SQL generadas son reflejadas en el log de la aplicación.	false
<code>jpa.production.generateDdl</code>	Establece si el software ORM debe generar las tablas de la base de datos.	true
<code>jpa.dev.database</code>	Tipo de base de datos que se utilizará bajo el perfil de desarrollo.	H2
<code>jpa.production.showsql</code>	Determina si las consultas SQL generadas son reflejadas en el log de la aplicación.	true
<code>jpa.production.generateDdl</code>	Establece si el software ORM debe generar las tablas de la base de datos.	true
<code>hibernate.production.generate_statistics</code>	Controla la generación de estadísticas para el ORM Hibernate el perfil producción	true
<code>hibernate.production.dialect</code>	Dialecto <b>[HIBER1]</b> de base de datos a utilizar por parte de Hibernate bajo el perfil de producción.	org.hibernate.dialect.MySQL5Dialect
<code>hibernate.production.hbm2ddl</code>	Acción a tomar <b>[HIBER2]</b> en caso de que <code>jpa.production.generateDdl</code> valga cierto.	create
<code>hibernate.dev.generate_statistics</code>	Controla la generación de estadísticas para el ORM Hibernate el perfil desarrollo	true
<code>hibernate.dev.dialect</code>	Dialecto de base de datos a utilizar por parte de Hibernate bajo el perfil de desarrollo.	org.hibernate.dialect.H2Dialect
<code>hibernate.dev.hbm2ddl</code>	Acción a tomar en caso de que <code>jpa.dev.generateDdl</code> valga cierto.	create
<code>hadoop.master</code>	Nombre de la máquina (o dirección IP) del nodo maestro de Hadoop.	Dependiente de la máquina
<code>hadoop.master.port</code>	Puerto en el que se ejecuta la interfaz web del <i>namenode</i> .	50070
<code>hadoop.hdfs.port</code>	Puerto de comunicaciones con el sistema de archivos distribuido HDFS.	9000
<code>hadoop.home</code>	Directorio de la máquina en el que está instalada la instancia de Apache Hadoop.	/usr/local/hadoop

<code>hadoop.scripts</code>	Directorio relativo a <code>hadoop.home</code> que contiene los scripts de gestión de Hadoop.	<code>bin</code>
<code>hadoop.start</code>	Nombre del script de arranque de Hadoop.	<code>start-all.sh</code>
<code>hadoop.stop</code>	Nombre del script de parada de Hadoop.	<code>stop-all.sh</code>
<code>hadoop.user</code>	Usuario que se ha configurado para ejecutar Hadoop.	Dependiente de la máquina
<code>solr.host</code>	Nombre de la máquina que alberga el nodo de Apache Solr	Dependiente de la máquina
<code>solr.port</code>	Puerto en el que escucha Solr esperando por conexiones.	8983
<code>solr.home</code>	Directorio de la máquina en el que Apache Solr está instalado.	<code>/usr/local/solr</code>
<code>solr.bin</code>	Directorio donde se encuentra el script de control de Solr. Ruta relativa a <code>solr.home</code>	<code>bin</code>
<code>solr.script</code>	Nombre del script contenido en <code>solr.bin</code> que controla Solr.	<code>solr</code>
<code>solr.start.options</code>	Parámetros que se pasarán al script definido en <code>solr.script</code> para el arranque.	<code>start</code>
<code>solr.stop.options</code>	Parámetros que se pasarán al script definido en <code>solr.script</code> para ella parada de Solr.	<code>stop -all</code>
<code>solr.user</code>	Usuario que ejecutará Solr.	Dependiente de la máquina
<code>cassandra.host</code>	Nombre de la máquina que alberga el nodo de Apache Cassandra	Dependiente de la máquina
<code>cassandra.home</code>	Directorio de la instalación de Apache Cassandra en la máquina.	<code>/usr/local/cassandra</code>
<code>cassandra.script</code>	Nombre del script que arranca la instancia de Cassandra.	<code>cassandra</code>
<code>cassandra.user</code>	Nombre del usuario encargado de ejecutar el proceso de Apache Cassandra	Dependiente de la máquina

<code>cassandra.port</code>	Puerto donde escuchará la instancia de Apache Cassandra	9160
<code>langdetector.profiles</code>	Directorio donde se encuentran los ficheros de perfil de idioma que utiliza la librería de detección de idioma utilizada en el análisis de internacionalización.	Dependiente de la máquina
<code>nutch.cli.home</code>	Ruta absoluta al directorio de instancia local de Apache Nutch	/usr/local/nutch/runtime/local

#### 9.2.4.4 Webmvc.properties

Este fichero de configuración controla los directorios que contienen las distintas partes de interfaz web de la aplicación. Se tienen dos tipos de propiedades, separados por perfiles, para especificar el directorio que contienen las plantillas de Freemarker (`view.(dev|production).templates.folder`), las hojas de estilo (`view.(dev|production).css.folder`), los ficheros JavaScript (`view.(dev|production).js.folder`) y las imágenes (`view.(dev|production).img.folder`). Se recomienda no modificar estos valores.

#### 9.2.4.5 Implementación de los servicios

La selección de las clases de implementación de los servicios se realiza mediante la factoría que es la clase `org.rz.webkumoconf.BackendServicesFactory`. Esta clase controla los servicios que se utilizarán para la infraestructura de procesamiento (Apache Hadoop, Apache Cassandra, Apache Solr y Apache Nutch).

Para la entrega de este proyecto únicamente se proporcionan alternativas de implementación para el servicio de Apache Nutch (mediante su API Java o mediante interfaz de línea de comandos). Esto puede ser controlado antes de arrancar WebKumo mediante la modificación del método que obtiene el servicio para que instancie un tipo de servicio u otro. Se muestra a continuación la función de factoría concreta.

```
/**
 * Factory for the crawling backend service.
 *
 * @return A reference to the backend.
 * @throws java.lang.Exception
 */
public static synchronized GenericCrawler getCrawlerService()
    throws Exception {
    if (BackendServicesFactory.crawlerSvc == null) {
        //BackendServicesFactory.crawlerSvc = NutchService.getInstance();
        BackendServicesFactory.crawlerSvc = NutchCommandLineService.getInstance();
    }
    return BackendServicesFactory.crawlerSvc;
}
```

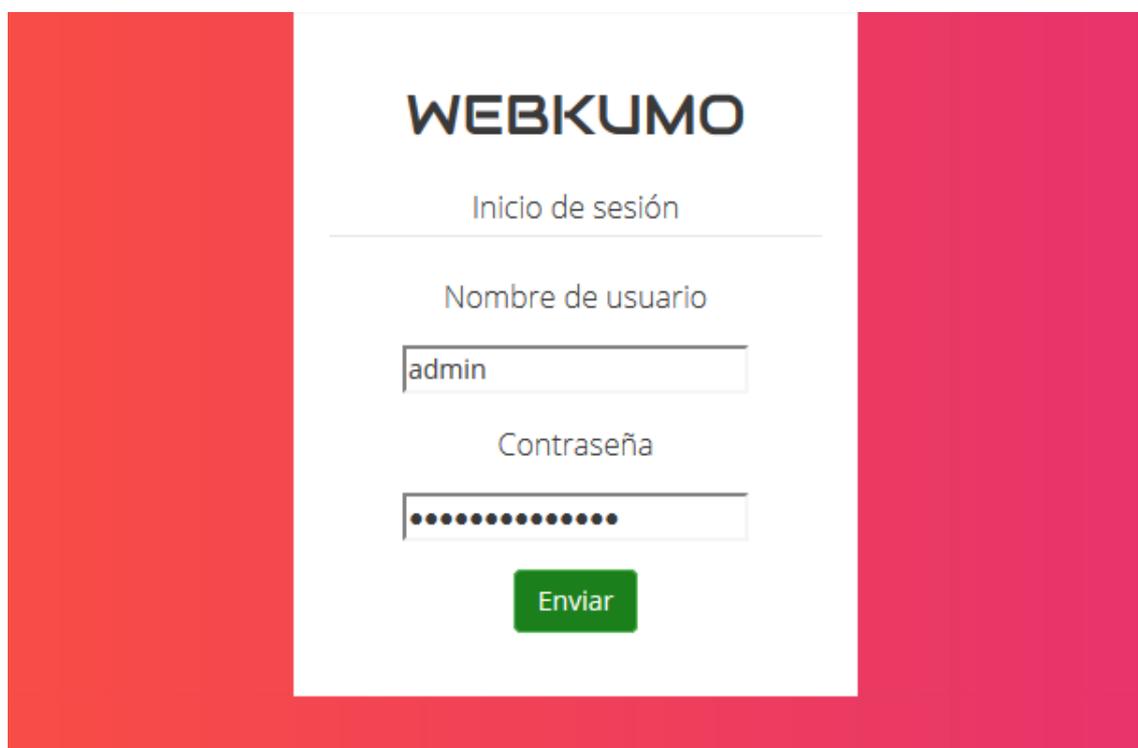
**Fig. 181: Factoría de servicios para Apache Nutch.**

## 9.3 Manual del usuario

Esta es la guía de manejo de la aplicación WebKumo que sirve de intermediario entre el usuario y la funcionalidad de análisis y procesamiento que el sistema que ocupa este proyecto ofrece. Se agruparán las siguientes subsecciones según funcionalidad atendiendo a las operaciones que se pueden realizar en la aplicación.

### 9.3.1 Inicio de sesión

Para utilizar WebKumo es necesario identificarse ante el sistema, para ello se debe introducir las credenciales que el administrador de WebKumo le haya suministrado en la página de inicio de sesión (se muestra por defecto al acceder a la URI en la que esté alojado WebKumo sin haber iniciado sesión) utilizando los campos del formulario de "Nombre de usuario" y "Contraseña" como se muestra a continuación:



The image shows a login form for WebKumo. At the top, the word "WEBKUMO" is displayed in a bold, dark blue font. Below it, the text "Inicio de sesión" is centered. A horizontal line separates the header from the input fields. The first field is labeled "Nombre de usuario" and contains the text "admin". The second field is labeled "Contraseña" and contains a series of 12 black dots. Below the password field is a green button with the text "Enviar" in white.

**Fig. 182: Formulario de inicio de sesión**

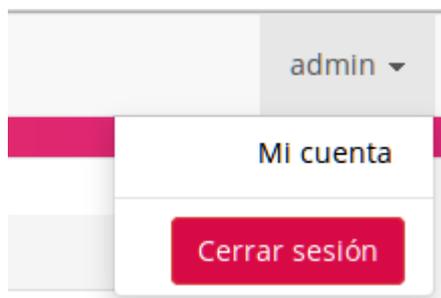
En caso de introducir una combinación incorrecta de usuario y contraseña, aparecerá en pantalla el siguiente mensaje explicativo:



**Fig. 183: Credenciales incorrectas en el inicio de sesión**

### 9.3.2 Cierre de sesión

Cuando se termine de trabajar con WebKumo, es recomendable cerrar la sesión del usuario para evitar usos no autorizados de la cuenta. Para llevar a cabo esto se debe usar el botón que aparece en la parte derecha de la barra de navegación superior como se puede apreciar en la imagen siguiente:



**Fig. 184: Botón de cierre de sesión**

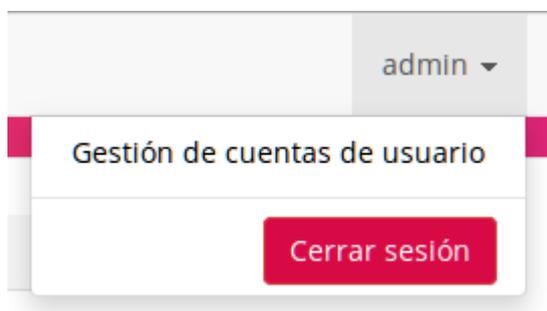
Al finalizar la sesión se dirigirá al usuario a la pantalla de inicio sesión donde verá un mensaje indicando que el cierre de sesión se ha efectuado. Se ve a continuación el mensaje de cierre de sesión:



**Fig. 185: Mensaje de cierre de sesión**

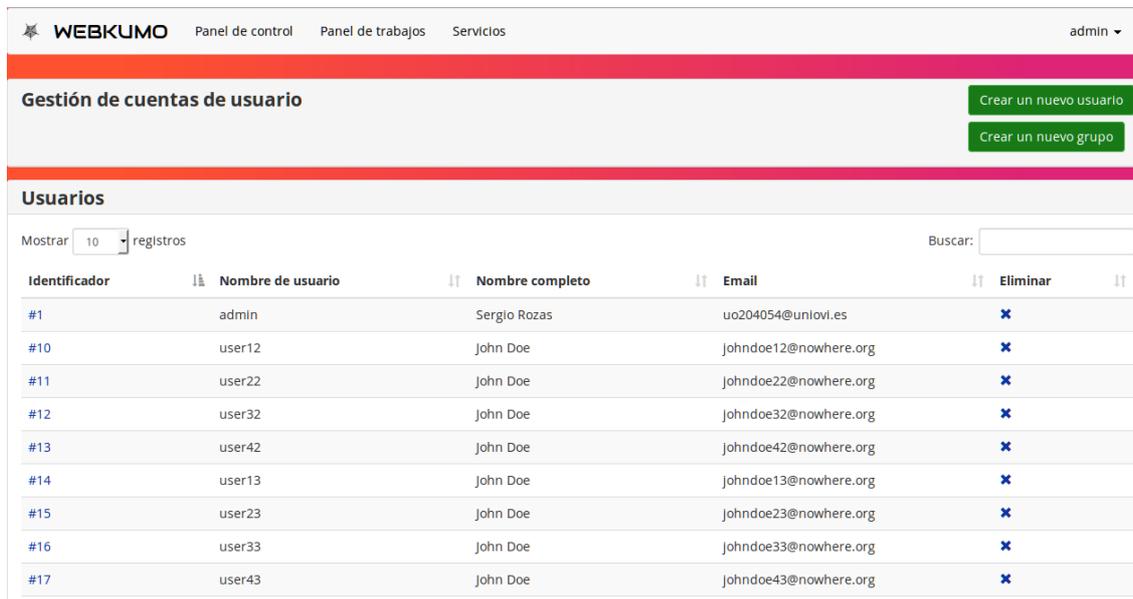
### 9.3.3 Panel del usuario (administrador)

Un usuario administrador puede acceder al panel de gestión de cuentas de usuario y grupos mediante el botón desplegable de la barra de navegación como puede apreciarse en la siguiente imagen:



**Fig. 186: Botón de acceso al panel de administración**

Tras utilizar el botón descrito se accede al panel del administrador que tiene el aspecto siguiente:



**Fig. 187: Vista del panel de administración**

Como administrador se presentan tres partes diferenciadas, una superior que contiene los botones para crear un nuevo usuario o grupo, una intermedia con una tabla con el listado de usuarios y una inferior con una tabla que recoge los grupos dados de alta en el sistema. Para facilitar la visualización, ambas tablas están paginadas:



**Fig. 188: Selector de páginas de tabla**

Las tablas permiten también el filtrado de la información mediante un recuadro de búsqueda:



**Fig. 189: Cuadro de búsqueda en tabla**

También permiten la ordenación (ascendente o descendente) por columnas mediante el uso de los botones de la cabecera de la tabla así como personalizar el número de elementos por página de tabla utilizando el selector de la parte superior izquierda.



**Fig. 190: Botón de ordenación de columnas y selector de registros por página**

### 9.3.4 Creación de usuario (administrador)

Desde el panel de administrador visto anteriormente se puede proceder al alta de un usuario nuevo mediante el botón correspondiente que aparece en la parte superior del panel visto en el apartado del panel de administración.

Hacer click en el botón mencionado enviará al usuario al formulario de creación de un usuario que se muestra a continuación:

**Editar información de usuario**
Enviar

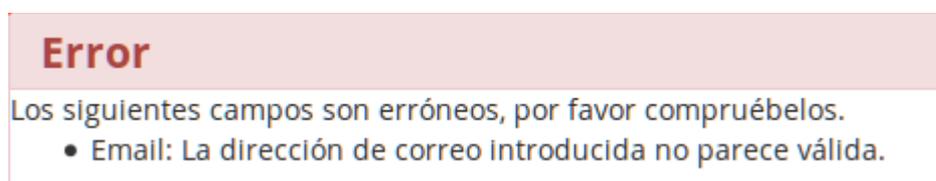
Nombre de usuario	carmen
Nombre	Carmen
Apellidos	Vázquez
Email	cvazquez@correo.dominio
Contraseña	****
Repita contraseña	****
Grupos	#2: group2, #6: group3 ▾
Rol	<input type="radio"/> Administrador <input checked="" type="radio"/> Usuario regular
Permisos	<input checked="" type="checkbox"/> Puedes ver tus propios trabajos <input checked="" type="checkbox"/> Puedes crear nuevos trabajos <input type="checkbox"/> Puedes ver trabajos de otros usuarios que estén en los mismos grupos que tú.

**Fig. 191: Formulario de creación de usuario**

Para proceder a la creación del usuario se deben rellenar todos los campos del formulario de forma adecuada. Se describen a continuación los campos del formulario así como indicaciones para hacerlo correctamente:

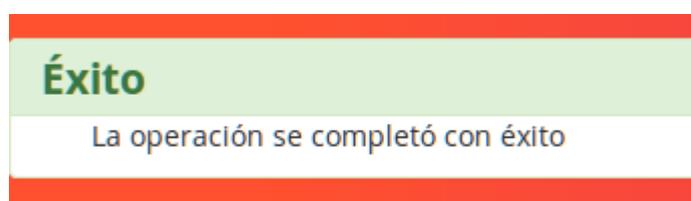
<b>CREAR USUARIO</b>		
<b>Nombre</b>	<b>Descripción</b>	<b>Valores válidos</b>
Nombre de usuario	El nombre que identificará al usuario nuevo ante el sistema	Texto entre 3 y 50 caracteres.
Nombre	Nombre de pila de la persona que utilizará la cuenta de usuario	Texto entre 3 y 50 caracteres.
Apellidos	Apellidos de la persona que utilizará la cuenta de usuario	Texto entre 2 y 50 caracteres.
Email	Dirección de correo electrónico de contacto	usuario@host.dominio
Contraseña	Para evitar accesos no autorizados a la aplicación, las cuentas de usuarios están protegidas por contraseña.	Texto de mínimo 3 caracteres.
Repita contraseña	Confirmación de la contraseña indicada anteriormente (para evitar errores).	Texto de mínimo 3 caracteres igual al de la contraseña.
Grupos	Conjunto de grupos a los que el usuario pertenece (se pueden seleccionar múltiples manteniendo pulsada la tecla control). Irrelevante si el usuario a crear va a tener el rol de administración.	Ningún grupo seleccionado, uno o varios (mediante lista de selección múltiple).
Rol	Tipo de usuario. Puede ser regular o administrador. Prestar especial atención a este punto ya que ser administrador otorga privilegios para gestionar otros usuarios y los servicios.	Administrador o Regular (mediante botón de selección única).
Permisos	Permisos que tendrá el usuario creado. Los permisos pueden ser de lectura (ver tus propios trabajos), de escritura (crear tus propios trabajos) y de lectura a nivel de grupo (ver trabajos de otros usuarios). Irrelevante si el usuario a crear va a tener el rol de administración.	Los indicados en el formulario mediante botón de selección múltiple.

Si alguno de los campos indicados en el formulario es incorrecto, se mostrará un mensaje explicativo al intentar enviar dicho formulario (mediante el botón correspondiente localizado en la parte inferior del formulario). Se ve a continuación un ejemplo de este mensaje de error:



*Fig. 192: Indicaciones de error en el formulario de creación de usuario.*

Indicar que **no se podrá dar de alta un usuario con un nombre de usuario (*login*) que ya esté en uso por parte de otro usuario** (en caso de intentarlo se indicará mediante un mensaje de error). En caso de que todos los campos del formulario sean válidos, se procederá al alta del nuevo usuario, dirigiendo al usuario al panel del administrador y mostrándole un mensaje de confirmación como el que se aprecia a continuación:



*Fig. 193: Mensaje de éxito en la operación.*

### 9.3.5 Edición de usuario (administrador)

Para editar un usuario, el usuario administrador debe acceder desde el panel del administrador a la información del usuario que quiera modificar. Para ello deberá hacer en el identificador de usuario mostrado en la tabla como se aprecia a continuación:

Identificador	 Nombre de usuario	 Nombre completo
#22	carmen	Carmen Vázquez

**Fig. 194: Acceso a la edición de un usuario**

Esto dirigirá al usuario al formulario de modificación del usuario que es idéntico al de creación del usuario con la salvedad que los datos del usuario ya estarán presentes en los campos del formulario. Se editarán entonces los aspectos deseados y se confirmará la operación mediante el botón situado al final del formulario a tal efecto.

Refiérase a las indicaciones dadas anteriormente para la creación de un nuevo usuario para obtener mas información sobre el proceso de edición, el cuál es idéntico a partir de este punto al de creación.

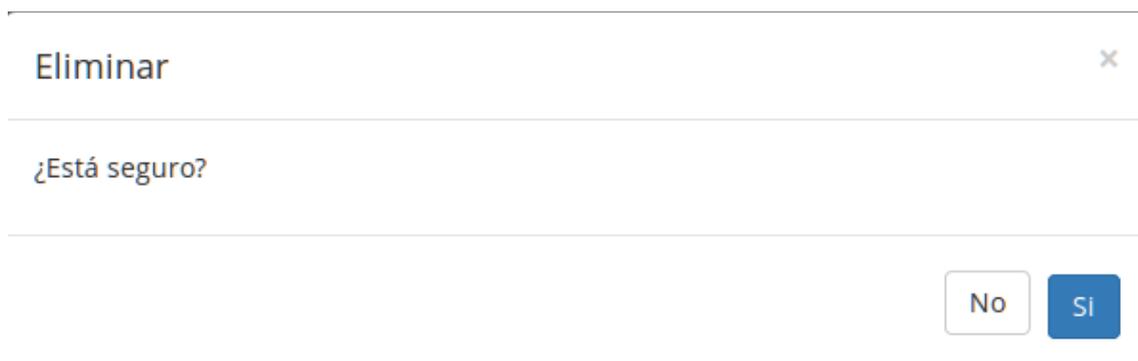
### 9.3.6 Eliminación de un usuario (administrador)

Si un usuario administrador desea eliminar la cuenta de usuario de cualquier otro debe dirigirse primeramente al panel del administrador descrito anteriormente. En la tabla de usuarios debe localizar al usuario que desee y hacer click en el icono de eliminación que puede verse en la parte derecha de la fila de la tabla como se aprecia a continuación:

Carmen Vázquez	cvazquez@correo.com	
----------------	---------------------	---

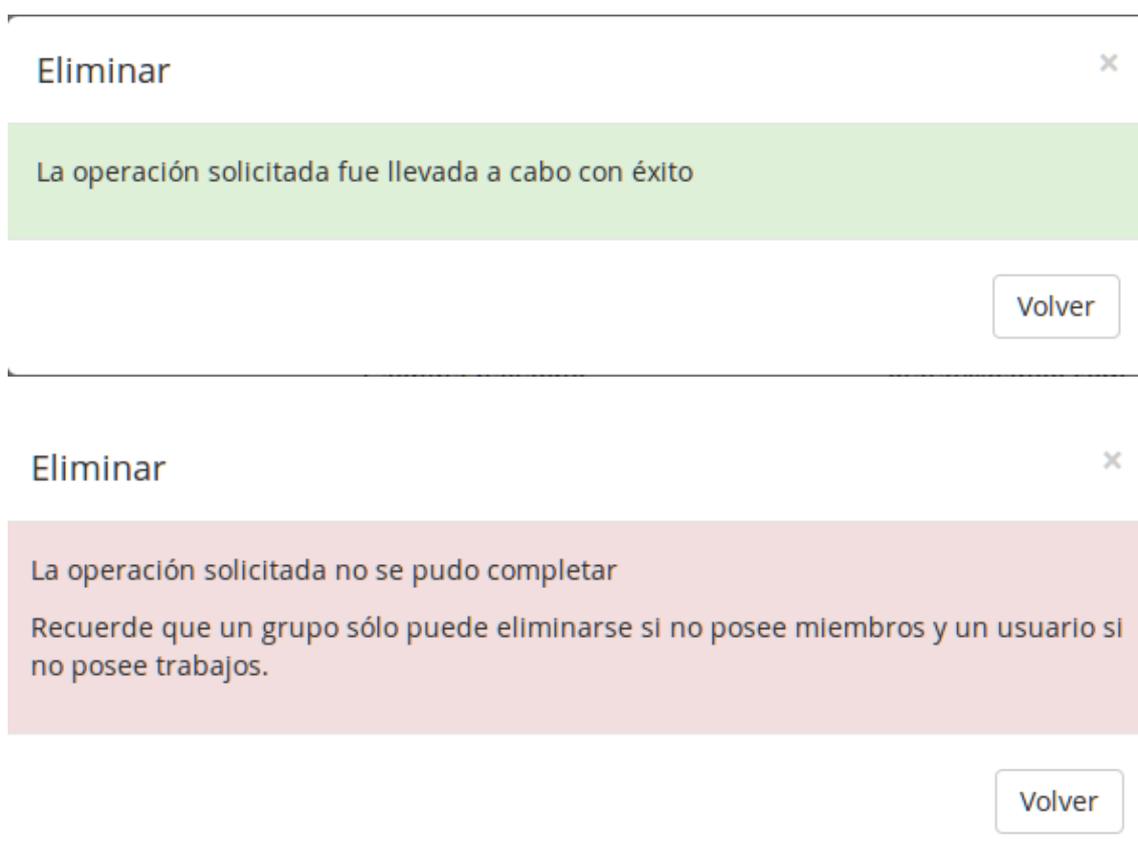
**Fig. 195: Botón de borrado de usuario.**

Tras esto se le mostrará al usuario un diálogo de confirmación para que verifique si desea realmente realizar la operación, en caso de responder negativamente se le devolverá al panel del administrador sin realizar ninguna operación. Si el administrador confirma su intención de eliminar al usuario seleccionado, se le indicará el éxito o fallo de la operación mediante un mensaje como el que sigue:



**Fig. 196: Mensaje de confirmación de borrado del usuario.**

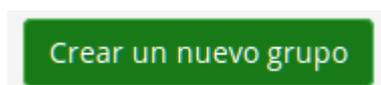
Destacar que **la operación fallará y así se indicará si el usuario que se intenta eliminar posee trabajos de análisis asociados.**



**Fig. 197: Mensaje de éxito y error en el borrado de un usuario.**

### 9.3.7 Creación de un grupo (administrador)

Los usuarios pueden aglutinarse mediante grupos de usuarios para gestionar la visibilidad de los trabajos que se creen en WebKumo. Cuando un administrador desee crear un nuevo grupo, deberá utilizar el botón situado a tal efecto en el panel del administrador como se muestra a continuación:



**Fig. 198: Botón de creación de un grupo**

Al pulsar el mencionado botón se dirigirá al usuario al formulario de creación que tiene el aspecto siguiente:

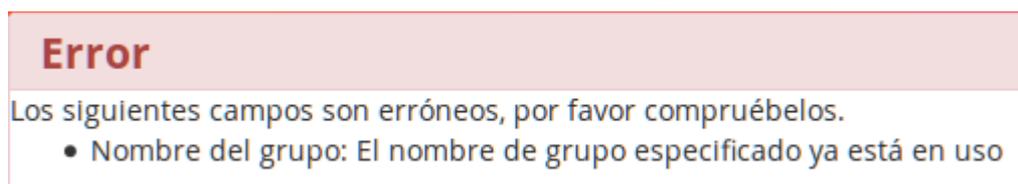
 A screenshot of a web form titled "Editar la información de grupo" with an "Enviar" button. The form contains two input fields: "Nombre del grupo" with the value "Grupo de Investigación" and "Descripción del grupo" with the value "Este grupo está compuesto por miembros del departamento de Informática."

**Fig. 199: Formulario de creación de grupo**

Se tienen en dicho formulario los siguientes campos a rellenar:

CREAR GRUPO		
Nombre	Descripción	Valores válidos
Nombre	El nombre que identificará al nuevo grupo de usuarios.	Texto entre 3 y 50 caracteres.
Descripción	Breve explicación del cometido del grupo	Texto entre 3 y 200 caracteres.

Si alguno de los campos del formulario contiene valores erróneos (ver tabla anterior) **o bien existe ya un grupo con el nombre especificado** se indicará mediante un mensaje de error explicativo como el que sigue:



**Fig. 200: Error en el formulario de creación de grupo.**

En caso opuesto, se procederá a la creación del grupo y se dirigirá al usuario al panel del administrador y de igual modo que ocurría en el caso de creación de un usuario, se le notificará que la operación ha sido llevada a cabo (ver creación de un usuario).

### 9.3.8 Modificación de un grupo (administrador)

Análogamente a lo que ocurría con los usuarios, los grupos se pueden modificar mediante el uso del formulario de edición al cuál se accede haciendo click en el identificador de grupo mostrado en la tabla correspondiente del panel del administrador:

Id	Nombre del grupo
#16	Grupo de Investigación

**Fig. 201: Acceso al formulario de edición de grupo.**

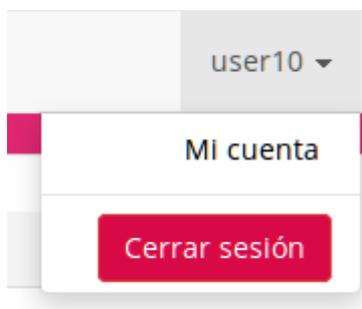
Tras acceder al formulario mencionado se repite el proceso de creación de grupo con la diferencia de que en este caso los campos del formulario estarán ya ocupados con la información del grupo. El proceso a partir de aquí es idéntico al de creación de un grupo (ver apartado anterior).

### 9.3.9 Borrado de un grupo (administrador)

Del mismo modo en el que se explicó cómo dar de baja un usuario, se puede dar de baja un grupo de usuarios **siempre que este no tenga miembros**. Aplica el mismo procedimiento que para eliminar un usuario con la salvedad de que hay que utilizar el botón de eliminado de la fila correspondiente de la tabla de grupos en lugar de la tabla de usuarios del panel del administrador. Consultar las indicaciones para la eliminación de un usuario para mas detalle.

### 9.3.10 Panel del usuario (regular)

Los usuarios no administradores también pueden modificar su propia cuenta de usuario mediante el panel del usuario al cuál se accede (al igual que el administrador accede al panel del administrador) mediante el botón contenido en el desplegable de la parte derecha de la barra superior de navegación.



**Fig. 202: Botón de acceso al panel del usuario.**

El panel del usuario tendrá la siguiente vista:



**Fig. 203: Panel del usuario regular**

Puede apreciarse que en la parte superior aparecen los botones para proceder a modificar la información del usuario regular que esté ejecutando WebKumo y en la parte inferior aparece un pequeño resumen de la información del usuario.

### 9.3.11 Modificación del usuario (regular)

Para que un usuario regular pueda modificar sus datos debe comenzar haciendo click en el botón situado a tal efecto en la parte superior del panel del usuario como se aprecia a continuación:



**Fig. 204: Botón de edición del usuario regular**

Tras utilizar el botón mencionado, se dirigirá al usuario al formulario de edición que se muestra a continuación:

John Doe (johndoe10@nowhere.org) Enviar

Nombre	John
Apellidos	Doe
Email	johndoe10@nowhere.org
Contraseña	****
Repita contraseña	****

**Fig. 205: Formulario de edición de un usuario regular**

Puede apreciarse que se trata de un formulario web estándar que contempla los siguientes campos:

MODIFICAR USUARIO REGULAR		
Nombre	Descripción	Valores válidos
Nombre	El nombre de pila del usuario.	Texto entre 3 y 50 caracteres.
Apellidos	Apellidos del usuario que utiliza la cuenta de usuario.	Texto entre 3 y 200 caracteres.

Email	La dirección de correo electrónico de contacto del usuario.	usuario@host.dominio
Contraseña	Campo para modificar la contraseña que utiliza el usuario para iniciar sesión	Texto de mas de 3 caracteres.
Repita contraseña	Confirmación de la contraseña introducida con anterioridad.	Texto entre 3 y 50 caracteres.

Como ocurría al modificar un usuario por parte de un administrador si se introducen valores inválidos en alguno de los campos se informará mediante el mensaje pertinente (consultar en el apartado de modificación de un usuario por parte de un administrador para ver un ejemplo de mensaje). En caso contrario se procederá a la modificación y el usuario podrá ver sus nuevos datos en el panel del usuario.

### 9.3.12 Consulta de la situación general del sistema

Cuando un usuario inicia sesión, es dirigido a la página de vista del sistema que tiene el aspecto que sigue:

The screenshot displays two main panels: 'Servicios' and 'Panel de trabajos'.

**Servicios:** Features a large orange warning icon and the text 'Existen algunos servicios que requieren atención'. Below this is a table of services:

APACHE HADOOP	Parado
APACHE NUTCH	En ejecución
APACHE SOLR	Parado
APACHE CASSANDRA	Parado

Última actualización: 27-jun-2015 11:31:23

**Panel de trabajos:** Lists six sample jobs with their status:

#1 Sample Job 10 : 27/6/2015 2:00:00 (user11)	FINISHED
#2 Sample Job 20 : 27/6/2015 2:00:00 (user12)	STOPPED
#3 Sample Job 11 : 27/6/2015 2:00:00 (user12)	FINISHED
#4 Sample Job 21 : 27/6/2015 2:00:00 (user11)	STOPPED
#5 Sample Job 12 : 27/6/2015 2:00:00 (user10)	FINISHED
#6 Sample Job 22 : 27/6/2015 2:00:00 (user23)	STOPPED

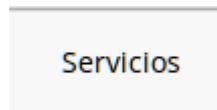
**Fig. 206: Panel de vista general del sistema**

Pueden observarse dos áreas bien diferenciadas a izquierda y a derecha de la pantalla. En la parte izquierda se especifica la situación actual de los servicios de infraestructura mediante un icono grande que identifica la situación general de forma visual. Justamente debajo se indican los estados de cada servicio individual así como la fecha de comprobación de los servicios.

En la parte derecha se le mostrarán al usuario los últimos cinco trabajos (si es que los tiene) que puede visualizar ordenados por fecha de creación. Se puede haciendo click en el identificador de cada trabajo (precedido por un cuadradillo "#") acceder directamente a la ficha del trabajo sin pasar por el panel de trabajos que se describirá posteriormente.

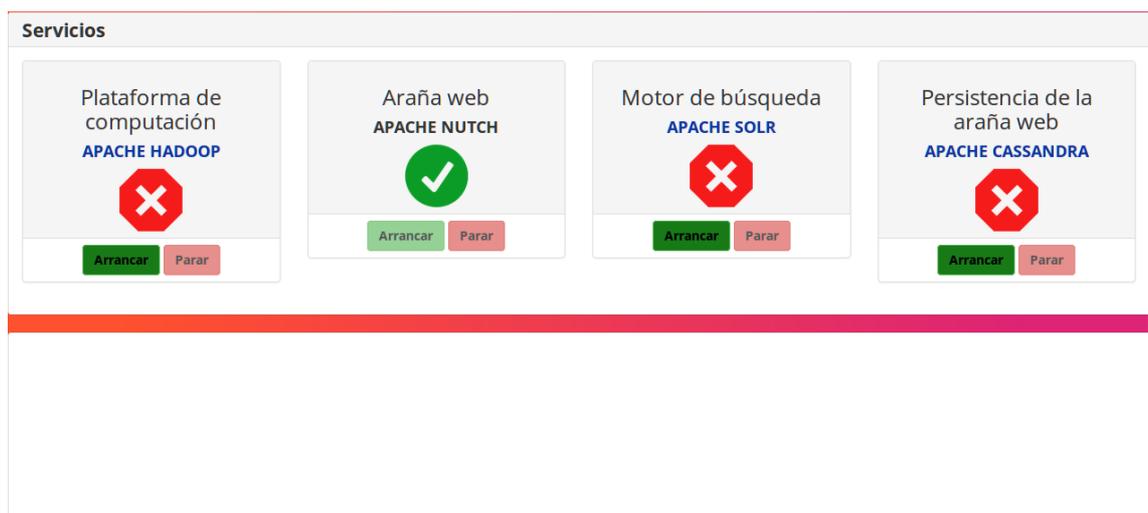
### 9.3.13 Panel de servicios (administrador).

Los usuarios administradores tendrán permitido gestionar los servicios de infraestructura mediante el panel de servicios al que pueden acceder mediante la opción correspondiente que se observa en el menú de navegación superior:



**Fig. 207: Botón de acceso al panel de servicios**

Al acceder al panel, la vista que se le presenta al usuario es la siguiente:



**Fig. 208: Vista del panel de servicios**

En la parte superior se aprecia el listado de servicios con botones accionables para gestionarlos así como su estado actual e información. Si un servicio no se puede detener y/o arrancar aparecerá su botón correspondiente desactivado. En la parte inferior, en blanco, inicialmente aparecerá información sobre el resultado de las operaciones efectuadas sobre cada servicio.

### 9.3.14 Arranque/Parada de un servicio (administrador).

Este punto es aplicable a cualquier servicio que se pueda arrancar o detener. Ambas operaciones se ejecutan de manera similar por lo que se agrupan en este punto único.

Se debe comenzar accediendo al panel de servicios como se indicó con anterioridad y posteriormente localizar en la parte superior de la página el servicio sobre el que se quiere operar. Una vez completado estas tareas se debe hacer click en el botón correspondiente que muestra a continuación:



**Fig. 209: Botones de acciones sobre un servicio**

Se debe hacer click en la opción que corresponda según se quiera arrancar o parar el servicio. Tras lo cuál aparecerá en la parte inferior del panel un indicador de que la operación está en curso como puede verse en la siguiente imagen:



---

**Fig. 210: Indicador de operación en curso sobre un servicio.**

Al finalizar la ejecución de la operación de arranque o parada se indicará el éxito de la misma mediante un mensaje como el que se ve en la imagen siguiente. Además podrá comprobarse la salida capturada del servicio durante su arranque como también se ve en la imagen:

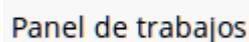


**Fig. 211: Éxito en la operación de arranque de un servicio y su salida.**

Es importante comentar que **la detención un servicio fallará si existen trabajos en el sistema que se encuentren en ejecución**. Si esto no fuera así los trabajos en ejecución fallarían al no disponer de comunicación con los servicios de infraestructura en los que se apoyan.

### 9.3.15 Panel de trabajos

Los usuarios pueden gestionar los trabajos mediante el panel de trabajos al cuál se accede mediante el enlace apropiado de la barra de navegación superior que se ve en la siguiente imagen:



Panel de trabajos

**Fig. 212: Botón de acceso al panel de trabajos.**

El panel de trabajos tiene el aspecto que sigue:

Identificador	Estado	Nombre del trabajo	Usuario propietario	Fecha de creación	Fecha de ejecución	Eliminar trabajo
#1	FINISHED	Sample job 10	user11	27/6/2015 2:00:00	---	✘
#2	STOPPED	Sample job 20	user12	27/6/2015 2:00:00	---	✘
#3	FINISHED	Sample job 11	user12	27/6/2015 2:00:00	---	✘
#4	STOPPED	Sample job 21	user11	27/6/2015 2:00:00	---	✘
#5	FINISHED	Sample job 12	user10	27/6/2015 2:00:00	---	✘
#6	STOPPED	Sample job 22	user23	27/6/2015 2:00:00	---	✘

**Fig. 213: Vista del panel de trabajos.**

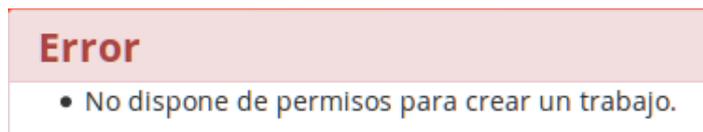
Se aprecia en la parte superior un botón para proceder a la creación de un nuevo trabajo y en la parte inferior una tabla (también paginada para hacer más cómoda la visualización de los trabajos) que contiene el listado de trabajos que el usuario puede visualizar. Cabe destacar que en esta tabla, cuando un trabajo esté en ejecución se mostrará mediante una barra de progreso su porcentaje de ejecución.

Identificador	Estado	Nombre del trabajo
#2	41%	Sample job 20

**Fig. 214: Indicador de progreso de un trabajo en ejecución**

### 9.3.16 Crear un trabajo

Los usuarios podrán dar de alta nuevos trabajos para su posterior ejecución mediante un formulario al cuál se accede desde el panel de trabajos explicado anteriormente. En el panel se debe utilizar el botón de la parte superior. Si el usuario no poseyera permisos de escritura (para crear un nuevo trabajo) se le mostraría el mensaje de error que se ve a continuación:



**Fig. 215: Mensaje de error para un usuario sin permisos de escritura al intentar crear un trabajo**

En caso de que el usuario sí disponga de permisos para crear un trabajo se mostrará el formulario web que se ve a continuación:

Crear un trabajo de análisis		Enviar				
Nombre	Trabajo de ejemplo					
Descripción	Este trabajo analizará algo					
Recursos por nivel	3					
Iteraciones	4					
Agente de usuario	Mozilla/5.0 (X11; Ubuntu; Linux x86_64; rv:38.0) Gecko/20100101 Firefox/38.0					
Cabecera Accept-Language HTTP	es, en-gb;q=0.8, en;q=0.7					
URL semilla	http://www.uniovi.es/					
Fichero de URLs semilla	<input type="button" value="Examinar..."/> No se ha seleccionado ningún archivo.					
Análisis a realizar	MobileWebAnalysis ▾					
Parámetros adicionales del trabajo	<table border="1"> <thead> <tr> <th>Clave del parámetro</th> <th>Valor del parámetro</th> </tr> </thead> <tbody> <tr> <td>MOBILE_ANALYSIS_DEVICE</td> <td>720</td> </tr> </tbody> </table>		Clave del parámetro	Valor del parámetro	MOBILE_ANALYSIS_DEVICE	720
Clave del parámetro	Valor del parámetro					
MOBILE_ANALYSIS_DEVICE	720					

**Fig. 216: Formulario de creación de un trabajo**

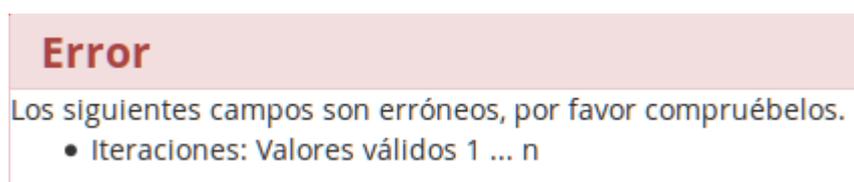
Se trata de un típico formulario web que se debe rellenar con valores válidos para la creación de un trabajo. Se describen a continuación los campos que lo componen:

CREAR UN TRABAJO		
Nombre	Descripción	Valores válidos
Nombre	El nombre del trabajo (opcional).	Texto de máximo 255 caracteres
Descripción	Descripción del trabajo a crear (opcional).	Texto de máximo 255 caracteres
Recursos por nivel	La cantidad de recursos web por cada iteración (ver campo siguiente) que serán descargados y analizados.	De 1 en adelante
Iteraciones	La cantidad de ciclos de ejecución de Apache Nutch que se ejecutarán. A cada ciclo se irán descargando los	De 1 en adelante

	recursos descubiertos a través de hipervínculos en recursos descargados en iteraciones anteriores.	
Agente de usuario	Identificador del agente de usuario que Apache Nutch utilizará para la descarga (útil para obtener un portal web adaptado para cierto dispositivo). Si no se especifica, se utilizará el indicado en el fichero de configuración <code>nutch-site.xml</code> (ver manual de instalación).	Cualquiera.
Cabecera Accept-Language	Valor enviado a los servidores de los que se descargan recursos para, mediante negociación de contenido, especificar el idioma requerido. Si no se especifica, se utilizará el indicado en el fichero de configuración <code>nutch-site.xml</code> (ver manual de instalación).	Cualquiera <b>[ACLAN]</b>
Semilla	URL que hará la tarea de semilla a partir de la cuál se comenzará a descargar y analizar recursos.	<code>http://www.host.dominio</code>
Fichero de semillas	Si se desea, se pueden especificar las semillas en un fichero de texto que contenga en cada línea las URL de las semillas (excluyente con la opción de "Semilla").	Fichero de texto no vacío con las semillas (una en cada línea).
Análisis	Se debe especificar el/los análisis que ejecutará el trabajo mediante una lista de elementos seleccionables (se pueden seleccionar múltiples análisis manteniendo pulsado el botón control).	Alguno de la lista de elementos seleccionables.
Parámetros adicionales	Parámetros extra que se pueden establecer para ser utilizados para dar instrucciones a los análisis sobre cómo de realizar ciertas operaciones. Pueden observarse los parámetros válidos haciendo click en el icono adjunto en forma de lupa. Si se necesitan establecer múltiples puede	Los de los parámetros registrados para el nombre del campo y cualquiera para su valor.

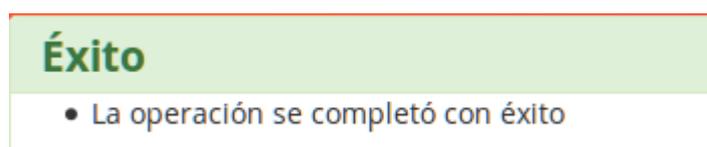
	hacerse mediante el botón en forma de cruz adyacente.	
--	---	--

Si alguno de los campos especificados es incorrecto se indicará mediante un mensaje de error como el siguiente:



**Fig. 217: Mensaje de error en la creación de un trabajo.**

En caso de que el trabajo se cree con éxito, se devolverá al usuario al panel de trabajos donde se mostrará un mensaje de confirmación del alta como se ve:



**Fig. 218: Mensaje de confirmación en el panel de trabajos.**

### 9.3.17 Consulta de un trabajo

Cuando se quiera consultar la información relativa a un trabajo, tarea que se realizará normalmente para visualizar el resultado de los trabajos, se debe acceder al trabajo mediante el panel de trabajos. Se debe hacer click en el identificador del trabajo como se muestra aquí:

Identificador	Estado	Nombre del trabajo
#1	FINISHED	Sample job 10

**Fig. 219: Acceso al trabajo**

Al acceder a la ficha del trabajo se muestra una página como sigue:

Trabajo # 1
<b>Nombre del trabajo:</b> Sample Job 10
<b>Descripción del trabajo:</b> This is a sample job
<b>Propietario del trabajo:</b> user22 (John Doe)
<b>Fecha de creación:</b> 27-Jun-2015
<b>Iteraciones:</b> 3
<b>Recursos por nivel:</b> 13
<b>Agente de usuario:</b> Sin establecer. Se utilizará el valor presente en nutch-site.xml
<b>Cabecera Accept-Language HTTP:</b> Sin establecer. Se utilizará el valor presente en nutch-site.xml
<b>Fecha de ejecución:</b> No ejecutado todavía
<b>Estado:</b> Finalizado
<b>URLs asociadas al trabajo</b>
1. http://ingenieriainformatica.uniovi.es/ 2. http://epigijon.uniovi.es/
<b>Análisis a realizar</b>
1. MockAnalysis (Análisis de prueba que extrae las URL y los nombres de los recursos.)
<b>Parámetros adicionales del trabajo:</b>
1. TEST_CUSTOM_PARAMETER : VALUE

**Fig. 220: Ficha del trabajo**

En la vista del trabajo puede observarse en la parte superior la información registrada del trabajo junto con un botón que permite la edición del trabajo (ver siguiente apartado del manual). En la parte inferior, si el trabajo ha sido ejecutado, aparecerán los resultados del mismo ordenados por análisis y por URI semilla.

El dato más importante que se obtiene es la puntuación global obtenida (sobre 100 puntos) que el análisis ha otorgado al sitio analizado. Si se utiliza el botón en forma de flecha que se encuentra junto a cada semilla, se desplegarán los resultados para cada campo de análisis concreto.

Cada campo está compuesto del nombre del campo (si se ha establecido), la URI que analiza y marcado como campo "S" la puntuación individual para el campo y como "W" el peso relativo dentro del cálculo de la puntuación general.

## Resultados del análisis

- **MockAnalysis**  
Análisis de prueba que extrae las URL y los nombres de los recursos.  
epigijon.uniovi.es  
  - Porcentaje global de cumplimiento: 0% ▼
- ingenieriainformatica.uniovi.es  
  - Porcentaje global de cumplimiento: 0% ▼

**Fig. 221: Panel de resultados de ejecución**

Se puede consultar el código fuente de los recursos descargados haciendo click en la URI del recurso según se ve a continuación:

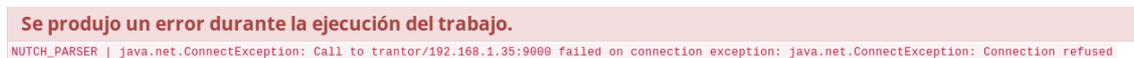
Source code x

```
&lt;!DOCTYPE html> <html xml:lang="es" class="ltr" dir="ltr" lang="es-ES"> <!-- CABECERA --> <head> <!--
Carga de ultima version de IE --> <meta http-equiv="x-ua-compatible" content="IE=edge"/> <!-- TITULO DEL
PORTAL --> <title>Universidad de Oviedo - Ingeniería Química y Tecnología del Medio Ambiente</title> <!--
METAINFORMACION DEL PORTAL --> <meta http-equiv="Content-Type" content="text/html; charset=UTF-8" />
<meta name="author" content="Centro de Innovación" /> <meta name="description" content="Sitio web
institucional de la Universidad de Oviedo" /> <meta name="keywords" content="Universidad de Oviedo,
Uniovi, Asturias, Investigación, Grado, Máster, Doctorado, Erasmus, I+D+i" /> <meta name="geo.placename"
content="Oviedo, Principado de Asturias, Spain" /> <meta name="geo.position" content="43.362 , -5.846" />
<meta name="cinn.nodo" content="webuniovi-02.lan.local" /> <meta name="viewport" content="width=device-
width, initial-scale=1.0"> <!-- FICHEROS JAVASCRIPT Y NOCONFLICT | CSS PRETTYPHOTO--> <script
src="/Uniovi-theme/js/jquery-1.9.1.min.js" type="text/javascript"></script> <script src="/Uniovi-theme/js
/jquery.menu-aim.min.js" type="text/javascript"></script> <script src="/Uniovi-theme/js/cinn.js"
type="text/javascript"></script> <!--[if lt IE 9]> <script src="/Uniovi-theme/js/html5shiv.js"></script>
<![endif]--> <script type="text/javascript"> var $jq = jQuery.noConflict(); </script> <script
src="/Uniovi-theme/js/jquery.prettyPhoto.js" type="text/javascript"></script> <meta content="text/html;
charset=UTF-8" http-equiv="content-type" /> <link href="/Uniovi-theme/images/favicon.ico" rel="Shortcut
Icon" /> <link href="/" rel="canonical" /> <link href="/en_US" hreflang="en-US" rel="alternate" /> <link
href="/ast_ES" hreflang="ast-ES" rel="alternate" /> <link href="/&#x2f;html&#x2f;css&#x2f;main&#x2e;
css&#x3f;browserId&#x3d;other&#x26;themeId&#x3d;Uniovi_WAR_Uniovi-theme&#x26;minifierType&#x3d;css&#x26;
languageId&#x3d;es_ES&#x26;b&#x3d;6101&#x26;t&#x3d;1333360784000" rel="stylesheet" type="text/css" />
<link href="/html/portlet/journal_content/css/main.css?browserId=other&
amp;themeId=Uniovi_WAR_Uniovi-theme&minifierType=css&languageId=es_ES&b=6101&
```

Volver

**Fig. 222: Vista del código fuente de un recurso indexado.**

Si por algún motivo se produce algún error irrecuperable durante la ejecución del trabajo, se indicará el error mediante un mensaje en la parte inferior de la ficha de trabajo.

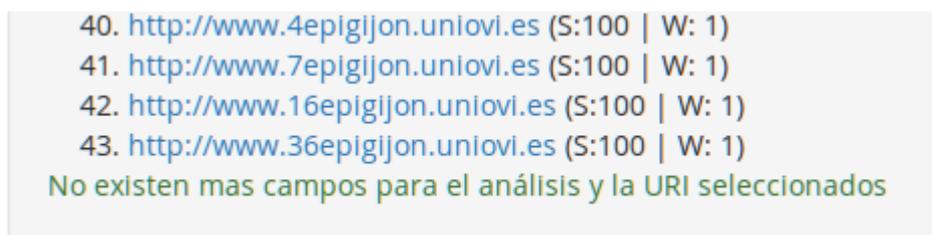


**Fig. 223: Mensaje de fallo de ejecución del trabajo.**

En la parte inferior del desplegable que carga los campos de un informe existe un botón que permite cargar mas campos del informe si los hay. Por defecto sólo se cargan los 100 primeros campos de cada informe.



**Fig. 224: Botón de carga de campos.**



**Fig. 225: Mensaje que indica que ya se han cargado todos los campos.**

### 9.3.18 Modificar un trabajo

Un trabajo, **siempre que no haya sido ejecutado**, puede ser modificado en sus parámetros. Para esto se debe hacer click en el botón de edición del trabajo que aparece en la ficha del trabajo comentada con anterioridad del panel de trabajos como se ve a continuación:



**Fig. 226: Identificador de trabajo clickable para acceder a la edición de un trabajo.**

A partir de este paso el proceso es el mismo que para crear un trabajo con la diferencia que en este caso los datos del trabajo ya aparecerán establecidos en los campos del formulario.

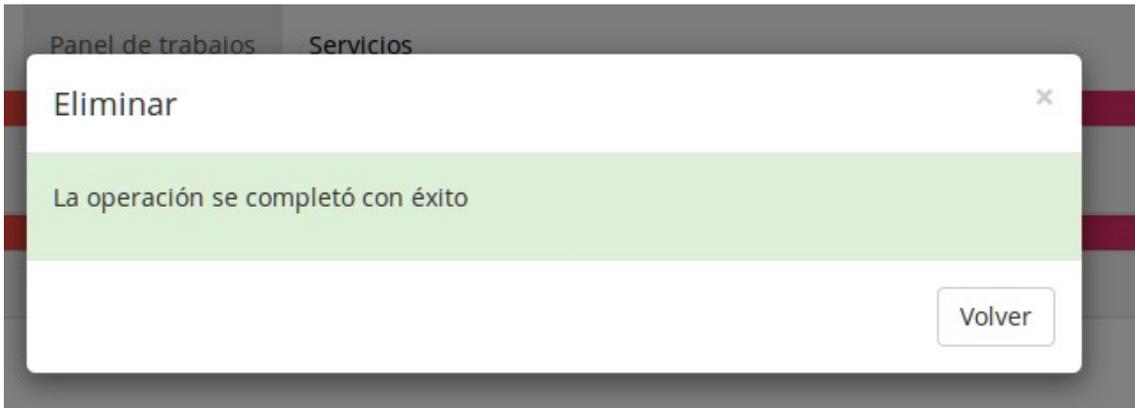
### 9.3.19 Borrar un trabajo

Para eliminar un trabajo **que no haya sido ejecutado**, se debe acceder primeramente al panel de trabajos que se describió en anteriores ocasiones. A continuación se debe localizar el trabajo a eliminar en la tabla de trabajos visibles por el usuario y utilizar el botón de la fila correspondiente de eliminación de trabajo que es el que se muestra:

Fecha de creación	Fecha de ejecución	Eliminar trabajo
27/6/2015 2:00:00	---	

**Fig. 227: Botón de borrado de un trabajo.**

Tras esto se le mostrará al usuario un diálogo de confirmación para que verifique si desea realmente realizar la operación, en caso de responder negativamente se le devolverá al panel de trabajos sin realizar ninguna operación. Si el usuario confirma su intención de eliminar el trabajo seleccionado, se le indicará el éxito o fallo de la operación mediante un mensaje como el que sigue:



**Fig. 228: Mensaje de confirmación de borrado de un trabajo.**

## 9.4 Manual del programador

Esta sección tiene como fin mostrar cómo se puede ampliar la funcionalidad de WebKumo mediante la codificación de nuevos algoritmos de análisis. Se dividirá este manual por pasos incluyendo explicaciones sobre la arquitectura Java de análisis diseñada y la integración de dicho análisis en Apache Nutch.

Se comienza describiendo la idea general a la hora de añadir un nuevo análisis. El nuevo análisis, debe incluirse como clase en el paquete `org.rz.webkumo.model.analysises.algorithms` y requiere para su funcionamiento implementar la clase abstracta `GenericAnalysis`. Un análisis poseerá además una clase de informe que deberá estar localizada en `org.rz.webkumo.model.analysises.reports` y que también deberá implementar una clase abstracta, en este caso es `GenericReport`. Un informe estará compuesto por campos (fields) que serán clases que implementen la clase abstracta `GenericReportField` y que constituirán las unidades que informarán de los aspectos concretos analizados.

Se recomienda tener a mano para el seguimiento de esta clase el diagrama de clases anexo a esta documentación que contiene todas las clases involucradas.

### 9.4.1 Creación de la nueva clase de análisis

Se debe comenzar mediante la creación de la nueva clase utilizando la herramienta que el programador considere oportuna (en este manual, por comodidad, se seguirá el proceso mediante el uso de un entorno de desarrollo, Netbeans 8.0.2 en concreto).

Este proceso se ejemplificará mediante el análisis de prueba `MockAnalysis` (proporcionado con la aplicación) que únicamente lista los recursos descargados.

El siguiente paso a tomar es hacer que la clase creada implemente la clase abstracta `org.rz.webkumo.model.analysises.GenericAnalysis` mediante la palabra reservada `extends`. Entonces se podrán redefinir los métodos de análisis que esta clase abstracta ofrece (que a su vez implementa la interfaz `org.rz.webkumo.model.analysises.Analyzable`). El uso de esta clase abstracta proporciona una serie de atributos e implementaciones de métodos comunes que evitan que el programador tenga que reescribirlos en cada clase de análisis.

Se comienza mediante la codificación de dos constructores obligatorios y mediante la redefinición de los métodos que se ven en el siguiente fragmento de código:

```

@Entity
@DiscriminatorValue(value = "MOCK_ANALYSIS")
@JsonTypeName("mock_analysis")
public class MockAnalysis extends GenericAnalysis {

    @JsonCreator
    public MockAnalysis() {
        this.reportType = MockReport.class;
        this.description = "algorithms.mock";
        this.mimeTypes.add(MimeType.CSS);
        this.mimeTypes.add(MimeType.HTML);
        this.mimeTypes.add(MimeType.XHTML);
        this.mimeTypes.add(MimeType.JAVASCRIPT);
    }

    @JsonCreator
    public MockAnalysis(Map<String, String> customParameters)
        throws InstantiationException,
        IllegalAccessException, IllegalArgumentException,
        InvocationTargetException, NoSuchMethodException {
        // Superclass will initialize this.customParameters
        super(customParameters);
        this.reportType = MockReport.class;
        this.description = "algorithms.mock";
        this.mimeTypes.add(MimeType.CSS);
        this.mimeTypes.add(MimeType.HTML);
        this.mimeTypes.add(MimeType.XHTML);
        this.mimeTypes.add(MimeType.JAVASCRIPT);
    }

    @Override
    public void analyzeScript(ScriptResource resource) {
        String text = "JS: ";
        GenericTextField field = new GenericTextField(text);
        field.setName("JS Resource Field");
        field.setResourceUri(resource.getUri());
        this.reports.get(resource.getUri().getHost()).addField(field);
    }

    @Override
    public void analyzeCss(CSSResource resource) {
        String text = "CSS: ";
        GenericTextField field = new GenericTextField(text);
        field.setName("CSS Resource Field");
        field.setResourceUri(resource.getUri());
        this.reports.get(resource.getUri().getHost()).addField(field);
    }

    @Override
    public void analyzeHtml(HTMLResource resource) {
        String text = "HTML: ";
        GenericTextField field = new GenericTextField(text);
        field.setName("HTML Resource Field");
        field.setResourceUri(resource.getUri());
        this.reports.get(resource.getUri().getHost()).addField(field);
    }
}

```

El constructor por defecto inicializa los distintos parámetros:

1. **reportType:** La clase de informe asociada a este análisis (necesaria para deserialización).
2. **description:** Nombre de la clave que contendrá la descripción del algoritmo en los ficheros properties que contienen los textos internacionalizados (resources/i18n/messages\_XX.properties).
3. **this.mimeTypes.add(MimeType):** Tipos MIME que procesará este análisis. En este ejemplo se aceptarán HTML, XHTML, CSS y JavaScript.

Existe otro constructor obligatorio que es el que WebKumo Connector utilizará para instanciar el análisis dentro del ciclo de ejecución de Apache Nutch. Este constructor es idéntico al sin parámetros con la diferencia que este recibe como parámetro un `Map<String, String>` que contiene los parámetros adicionales (ver creación de un trabajo en el manual del usuario) que se le puedan pasar al análisis (mas adelante se verá cómo se declaran).

Ambos constructores deberán ser anotados con `@JsonCreator` para que la deserialización de los análisis pueda ser llevada a cabo (proceso que forma parte de la comunicación entre WebKumo Connector y WebKumo).

El resto de métodos redefinidos son aquellos que exponen los recursos (ver paquete `org.rz.webkumo.model.sites`) al programador. Estos métodos `analyzeHTML`, `analyzeCSS` y `analyzeScript` serán invocados por el método `analyze` de la clase padre (`GenericAnalysis`, salvo que se redefina) durante la fase de procesado (*parsing*) de Apache Nutch. Esto hace que sea en estos métodos donde se implemente la lógica de procesamiento que se quedará, por tanto, en el modelo de datos de la aplicación. **El código fuente de los recursos puede obtenerse para su análisis mediante el método `getSourceCode()`** como se aprecia en el ejemplo. También puede obtenerse el nombre del mismo (o título HTML) y su codificación

Los métodos de `MockAnalysis` que se utilizar para ejemplificar únicamente generan un campo (*field*) de texto genérico `GenericTextField` (proporcionado con la aplicación) que albergará el título del recurso y la URI en la que se encuentra. Pueden definirse campos en el paquete `reports.fields` para necesidades especiales siempre que implementen la clase abstracta `GenericReportField`.

Un campo por defecto se valorará con la máxima puntuación (100 puntos) y tendrá un peso relativo completo (1). Los campos que contenga un informe de un análisis deben tener **o bien todos un peso de 1** (WebKumo calculará entonces automáticamente la puntuación proporcional de forma igualitaria) **o bien sumar entre todos los pesos 1**. Se puede definir el peso a utilizar, para centralizarlo, en `DefaultFieldWeight` del paquete `reports`. Se debe tener esto en cuenta ya que **la puntuación de un análisis para una semilla concreta se calcula como la media ponderada de los campos de su informe (peso multiplicado por puntuación)**. Además, como se ve en el código anterior, **se deben establecer el nombre y la URI del recurso que valora el campo mediante `field.setName()` y `field.setResourceUri()` obligadamente**. Este nombre debe ser único entre todos los nombres de campos para el análisis.

Para añadir un campo al informe, como se ve en el código anterior, se añade a un `Map<String, GenericReportField>` en el que la clave es el host al cuál se le está aplicando el análisis (obtenible mediante `resource.getUri().getHost()`). **Si no se añade el campo generado al analizar el recurso al informe, el análisis será ignorado.**

Adicionalmente se debe anotar la clase del análisis como sigue:

1. **@Entity**: El análisis y sus entidades asociadas debe persistirse en la base de datos por lo que esta anotación de JPA debe estar presente.
2. **@DiscriminatorValue**: Esta anotación indica al software ORM el valor para discernir entre los análisis de un tipo y otro. El valor de la anotación debe ser único (no debe existir ningún otro análisis con el mismo valor anotado).
3. **@JsonTypeName**: Nombre que se le darán a las instancias serializadas a JSON de este análisis. Es necesario para que el software de deserialización sepa de qué implementación concreta de `GenericAnalysis` se trata. Como contrapartida, se debe incluir también en la clase `GenericAnalysis` la nueva clase de análisis creada en la anotación `@JsonSubtypes` para indicarle al software de serialización las clases que implementan la clase abstracta. El valor de `@JsonTypeName` debe coincidir con el de `name` en `GenericAnalysis` según se ve en el siguiente código.

```
@JsonSubTypes({
    @JsonSubTypes.Type(value = MockAnalysis.class, name = "mock_analysis")
})
```

Recordar que **si se añade algún campo o colección accesoria como atributo de la clase de análisis, debe gestionarse su persistencia siguiendo la API de anotaciones JPA**. Si no es necesario persistir el atributo, basta con anotarlo con `@Transient` o bien, si es una colección, con `@ElementCollection`.

Estas son las directrices básicas para implementar un nuevo algoritmo de análisis.

## 9.4.2 Parámetros personalizados

La arquitectura de análisis se ha diseñado de tal modo que se puedan especificar, a la hora de crear el trabajo de análisis, ciertos parámetros que los trabajos pueden utilizar para realizar su computación. Por ejemplo, en el análisis de directrices móviles del W3C que se entrega con el sistema (`MobileWebAnalysis`), se utiliza este parámetro adicional para que el usuario indique qué ancho de pantalla (en píxeles) debe suponer el análisis a la hora de realizar los cálculos de adaptación a dispositivos móviles. Si no se especifica, el análisis toma por defecto el dispositivo de referencia indicado por la guía del W3C.

Para indicar un nuevo parámetro adicional disponible para su uso, debe declararse en la clase del análisis correspondiente como sigue:

```
@CustomJobParameter
public static final String MOBILE_ANALYSIS_DEVICE_WIDTH =
"mobile_analysis_device_width";
```

Se debe utilizar la anotación `@CustomJobParameter`, de este modo cuando un usuario cree un trabajo y quiera definir este parámetro, lo podrá ver en la lista que se despliega en el formulario de creación de trabajos (ver manual del usuario).

Para acceder durante el análisis al valor de la variable establecido, se debe utilizar el mapa de parámetros adicionales `customParameters`, que forma parte de los atributos heredados de `GenericAnalysis`. De este modo para acceder por ejemplo al parámetro adicional ficticio `PARAMETER`, se usaría el siguiente código:

```
String confParam = this.customParameters.get(PARAMETER);
if ((confParam != null) {
    // El parámetro ha sido establecido por el usuario
}
else {
    // El parámetro no ha sido establecido
}
```

### 9.4.3 Creación de una nueva clase de informe

Aunque no es estrictamente necesario, cada clase de análisis debería contener su propia clase de informe según se indicó anteriormente. Estas clases de informe implementan la clase abstracta `GenericReport` del paquete `org.rz.webkumo.model.analysis.reports`.

Se muestra a continuación como ejemplo, la clase de informe para el análisis de internacionalización `InternationalizationAnalysis`.

```

@Entity
@DiscriminatorValue("I18N_REPORT")
@JsonTypeName("i18n_report")
public class InternationalizationReport extends GenericReport {

    public InternationalizationReport() {
    }

    @Override
    public void prepareReport() {
        for (GenericReportField field : this.getFields()) {
            if (field instanceof LanguageDetectionField) {
                this.checkLanguageDetectionResults((LanguageDetectionField)
field);
            }
            if (field instanceof LanguageTagsField) {
                this.checkLanguageTags((LanguageTagsField) field);
            }
            if (field instanceof DatesDetectionField) {
                this.checkDatesFormat((DatesDetectionField) field);
            }
            if (field instanceof NumberFormattingField) {
                this.checkNumberFormat((NumberFormattingField) field);
            }
        }
    }

    private void checkLanguageDetectionResults(LanguageDetectionField
field) {
        // Comprueba que el idioma del texto del HTML case con el formato
esperado
        // para el Locale para el que se ejecutó el trabajo (definido en el
formulario de creación
        // del trabajo).
    }

    private void checkLanguageTags(LanguageTagsField field) {
        / Comprueba que el idioma declarado mediante etiquetas HTML case con
el formato esperado
        // para el Locale para el que se ejecutó el trabajo (definido en el
formulario de creación
        // del trabajo).
    }

    private void checkDatesFormat(DatesDetectionField field) {
        // Comprueba que el formato de las fechas encontradas case con el

```

```

formato esperado
    // para el Locale para el que se ejecutó el trabajo (definido en el
formulario de creación
    // del trabajo).
    }

    private void checkNumberFormat(NumberFormattingField field) {
        // Comprueba que los números encontrados en el recurso HTML sigan el
formato
        // del Locale para el que se ejecutó el trabajo (definido en el
formulario de creación
        // del trabajo).
    }

```

Se observa que, al igual que ocurría en la definición de un nuevo análisis, el informe debe anotarse con `@Entity`, `@DiscriminatorValue` y `@JsonTypeName` y por tanto el nombre debe incluirse en la declaración de subtipos `@JsonSubTypes` en la clase `GenericReport` (ver apartado anterior de creación de un análisis para explicación detallada).

**El método que mas relevancia tiene es `prepareReport()` que será invocado una vez se haya completado el análisis y cuyo fin es realizar tareas posteriores al análisis si fueran necesarias.** En el ejemplo del análisis de internacionalización este método se utiliza para recalculer la puntuación de los campos de análisis, ya que la puntuación del análisis será determinada en función de si el sitio web procesado ha respondido con su versión en el idioma solicitado (mediante negociación de contenido, ver manual del usuario, creación de un trabajo) o no. Si el idioma detectado por los distintas comprobaciones del algoritmo de análisis (idioma del texto, formato de las fechas etc.) coincide con el solicitado, se le asignará una puntuación alta y en caso contrario una puntuación baja.

Recordar que **si se añade algún campo o colección accesoria como atributo de la clase de informe, debe gestionarse su persistencia siguiendo la API de anotaciones JPA.** Si no es necesario persistir el atributo, basta con anotarlo con `@Transient` o bien, si es una colección, con `@ElementCollection`.

## 9.4.4 Creación de una nueva clase de campo de informe

En la mayoría de ocasiones, basta con utilizar el campo de texto genérico que se proporciona `GenericTextField` (paquete `analyses.reports.fields`) que alberga una cadena de texto que luego será impresa en el informe para informar al usuario. Si se requiriera una forma de expresar un campo de formulario distinta

a un campo de texto, puede crearse un nuevo campo mediante la implementación de la clase abstracta `GenericReportField`. Se muestra a continuación un ejemplo mediante el campo que alberga el formato de los números encontrados en el texto de un recurso HTML (utilizado en el análisis de internacionalización) `NumberFormattingField`.

```
@Entity
@JsonTypeName("number_format_field")
@DiscriminatorValue(value = "NUMBER_FORMAT_FIELD")
public class NumberFormattingField extends GenericReportField {

    @ElementCollection
    @JsonView(AnalysisJSONView.WebKumoConnector.class)
    private List<String> numbers;

    public NumberFormattingField() {
        this.name = "NumberFormattingField";
        this.weight = DefaultFieldWeight.NUMBER_FORMAT_FIELD;
        this.numbers = new ArrayList<>();
    }

    public List<String> getNumbers() {
        return this.numbers;
    }

    public void addNumber(String number) {
        this.numbers.add(number);
    }
}
```

Como se aprecia en el código anterior, debe de nuevo anotar el campo con las ya explicadas anotaciones `@Entity`, `@DiscriminatorValue` y `@JsonTypeName` y por tanto se debe añadir el campo al vector de anotaciones `@JsonSubTypes` de `GenericReportField` (ver apartado anterior de creación de nueva clase de análisis para mas información).

Este campo de informe contiene una colección de cadenas de caracteres conteniendo los formatos encontrados durante el análisis del recurso web HTML correspondiente. Obsérvese la anotación `@ElementCollection` para indicarle al software ORM que ignore el campo a la hora de persistirlo (ya que no es necesario) y `@JsonView` **[JACVIE]** que determina conjuntos de atributos a serializar. **Todos aquellos atributos que se requieran almacenar como parte del análisis para su posterior procesado en `prepareReport()`** (ver apartado anterior de creación de un informe) **deben anotarse así** para que se incluya dicha propiedad en la serialización.

Se debe remarcar también, que la presentación del campo como parte de un informe al usuario, está compuesta por el nombre del campo, su peso relativo y su puntuación. Por lo que **se debe definir el nombre del campo** (mediante el nombre de su método por ejemplo) **estableciendo el valor del atributo name** según sea pertinente.

## 9.4.5 Gestión de dependencias

Durante el desarrollo de un algoritmo de análisis nuevo, es muy probable que se necesite utilizar una librería determinada para llevar a cabo la computación (un *parser* concreto, una librería de detección de idioma etc.). Para conseguir esto, se deben completar dos pasos.

Primeramente y debido a que WebKumo utiliza Apache Maven **[MVN]** como gestor de dependencias, se debe añadir la mencionada dependencia al fichero `pom.xml` del proyecto Maven para que esta esté en disposición de ser utilizada.

Se recomienda separar la versión de la dependencia de su declaración, para ello se recomienda seguir el esquema proporcionado en el fichero. Se ejemplifica con la dependencia de Jsoup **[JSOUP]**, que es el *parser* utilizado en los análisis proporcionados para HTML:

```
<version.jsoup>1.8.2</version.jsoup>
<!-- jsoup HTML parser library -->
<dependency>
  <groupId>org.jsoup</groupId>
  <artifactId>jsoup</artifactId>
  <version>${version.jsoup}</version>
</dependency>
```

**Fig. 228: Ejemplo de dependencia Maven para WebKumo.**

La versión de la dependencia debe añadirse junto al resto como hija del elemento `properties` del `pom.xml`. El elemento `dependency` y sus hijos deben declararse como hijos del elemento `dependencies`.

De esta forma se pueden obtener una gran cantidad de dependencias Java del repositorio central **[MVNRE]** de Maven. También es posible añadir repositorios de terceros e incluso instalar artefactos (JAR) de manera manual. Para un uso avanzado de Maven se recomienda recurrir a las guías **[MVNGUI]** de la Fundación Apache.

Además de indicarle a la aplicación WebKumo las dependencias requeridas para los análisis que se creen mediante Apache Maven, se deben indicar también dichas librerías a Apache Nutch. El procedimiento es bastante similar, editar un fichero XML de configuración, pero en este caso es un fichero de Apache Ivy, otro sistema de gestión de dependencias de la Fundación Apache, que es el que utiliza Nutch.

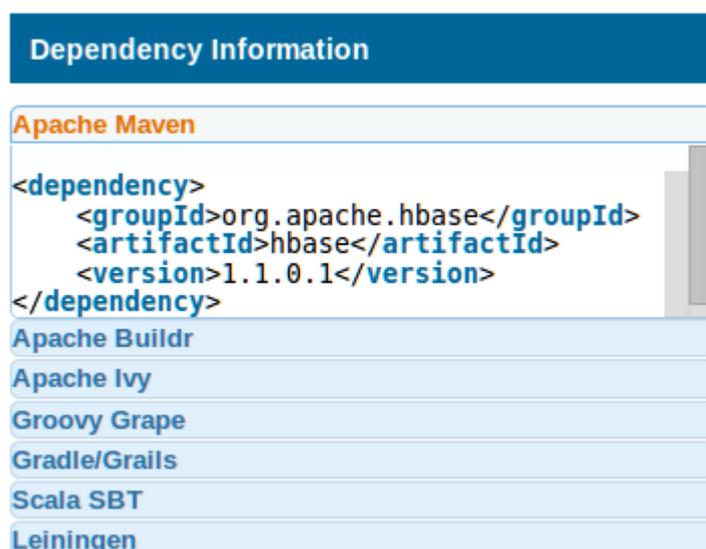
Se debe comenzar editando el fichero `ivy.xml` contenido en la carpeta `src/plugins/webkumo-connector` situada en el directorio de instalación de Apache Nutch (`/usr/local/nutch` si se siguió el manual de instalación que se proporciona en el apartado 9.1). Las dependencias en Apache Ivy se descargan también del repositorio central de Maven y se declaran como sigue:

```
<dependencies>
  <dependency org="net.sf.jung" name="jung2" rev="2.0.1" conf="*->*" />
  <dependency org="org.ccil.cowan.tagsoup" name="tagsoup" rev="1.2" conf="*->*" />
  <dependency org="org.hibernate" name="hibernate-core" rev="4.3.7.Final" conf="*->*" />
  <dependency org="org.hibernate" name="hibernate-entitymanager" rev="4.3.7.Final" conf="*->*" />
  <dependency org="org.hibernate" name="hibernate-validator" rev="5.1.3.Final" conf="*->*" />
  <dependency org="com.fasterxml.jackson.core" name="jackson-core" rev="2.5.3" conf="*->*" />
  <dependency org="com.fasterxml.jackson.core" name="jackson-databind" rev="2.5.3" conf="*->*" />
  <dependency org="com.fasterxml.jackson.core" name="jackson-annotations" rev="2.5.3" conf="*->*" />
  <dependency org="nu.validator" name="validator" rev="15.3.10" conf="*->*" />
  <dependency org="com.phloc" name="phloc-css" rev="3.7.7" conf="*->*" />
  <dependency org="org.apache.httpcomponents" name="fluent-hc" rev="4.4.1" conf="*->*" />
  <dependency org="org.apache.httpcomponents" name="httpClient" rev="4.4.1" conf="*->*" />
  <dependency org="com.datastax.cassandra" name="cassandra-driver-core" rev="2.1.5" conf="*->*" />
  <dependency org="org.jsoup" name="jsoup" rev="1.8.2" conf="*->*" />
  <dependency org="com.norconex.language" name="langdetect" rev="1.3.0" conf="*->*" />
  <dependency org="com.joestelmach" name="natty" rev="0.10.1" conf="*->*" />
  <dependency org="org.mozilla" name="rhino" rev="1.7.6" conf="*->*" />
</dependencies>
```

**Fig. 229: Declaración de dependencias para WebKumo Connector mediante Apache Ivy.**

Se puede observar en la imagen que se declara la dependencia como elemento `dependency` (hijo del elemento `dependencies`) indicando el desarrollador bajo el que está publicado como atributo `org`, el nombre de la dependencia como `name` y la versión como `rev` (en Maven esto se declaraba con los elementos `groupId`, `artifactId` y `version` respectivamente). El atributo `conf` controla bajo qué circunstancias y cómo se deben resolver las dependencias, debe establecerse a `"*->*"` al igual que en la imagen.

Puede resultar de utilidad el buscador que incorpora el sitio web del repositorio central de Maven ya que, para una dependencia concreta muestra para distintos gestores cómo se declararía. Se muestra un ejemplo a continuación:



**Fig. 230:** Muestra de una dependencia en el repositorio central de Maven en sus distintas formas

## 9.4.6 Integración de nuevos análisis en Apache Nutch

Para que el análisis sea efectivamente ejecutado por Apache Nutch durante su ciclo de ejecución, se hace necesario recompilar dicho software.

Primeramente se debe asegurar que las dependencias que requiere la nueva clase de análisis hayan sido añadidas al fichero `pom.xml` de WebKumo (Apache Maven) y al fichero `ivy.xml` de Apache Nutch (Apache Ivy) como se indicó en el apartado anterior.

Además, como es lógico, se deben copiar las nuevas clases codificadas (tanto de análisis como de informe y de campos de informe) a la jerarquía de paquetes localizada en el directorio `src/plugins/webkumo-connector/src/java` (partiendo desde el directorio de instalación de Nutch). Debe respetarse la misma jerarquía de paquetes que se tiene para WebKumo, esto es:

- ◆ **Análisis (extends `GenericAnalysis`):** `src/plugins/webkumo-connector/src/java/org/rz/webkumo/model/analyses/algorithms`
- ◆ **Informes (extends `GenericReport`):** `src/plugins/webkumo-connector/src/java/org/rz/webkumo/model/analyses/reports`

- ◆ **Campos de informe (extends GenericReportField):** `src/plugins/webkumo-connector/src/java/org/rz/webkumo/model/analyses/reports/fields`

A continuación se procederá al recompilado de Apache Nutch mediante la herramienta Apache Ant usando el comando (ejecutado desde el directorio raíz de instalación de Apache Nutch) `ant clean && ant runtime`. Tras lo cuál deben copiarse los ficheros JAR nuevos a la carpeta `lib` del directorio de instalación de Apache Hadoop (ver parte final del manual de instalación de Apache Nutch en la sección 9.1.4).

Como añadidura, comentar que **este proceso debe realizarse con el sistema detenido** y que cada vez que se modifique el código fuente de un análisis, informe o campo de informe debe detenerse el sistema y procederse al recompilado de las fuentes.

# 10 Conclusiones y ampliaciones

## 10.1 Conclusiones

A la finalización de este proyecto, como conclusión principal, se extrae que se han cumplido los objetivos marcados para la realización del proyecto. El objetivo principal era en resumen, crear una arquitectura software extensible sustentada por una infraestructura de procesamiento de alta capacidad para que, en caso de ser necesario, se aproveche la fundamentación en dicha estructura para ejecutar grandes cargas de trabajo.

Adicionalmente se ha llegado a otras conclusiones adicionales relacionadas con la gestión y el desarrollo de un proyecto de cierta envergadura. Sobre todo destaca la gestión de imprevistos que provoquen cambios en la planificación y cómo estos no se contemplaron en la previsión inicial de riesgos realizada en la asignatura de Gestión y Dirección de Proyectos. Esto evidencia la necesidad de adquirir experiencia en proyectos reales en los que estas situaciones están al orden del día para poder ampliar las miras y dejar atrás la planificación orientada cuasi estrictamente al ámbito técnico del proyecto.

También se pudo advertir en la aplicación de este proyecto la complejidad de los sistemas distribuidos y de otro software relacionado con el Big Data con el cuál el proyectante no poseía experiencia previa. Además se pudo apreciar de manera importante el valor de la documentación de un proyecto ya que se trabajó con un software como Apache Nutch que está muy pobremente documentado.

Como conclusión adicional, se pone en valor la importancia del desarrollo de herramientas automatizadas de análisis y computación que faciliten y mejoren los procesos diversos en los que se pueden aplicar este proyecto. Sin este tipo de herramientas la evaluación manual de portales web y en general, diversas tareas de investigación y desarrollo serían mucho mas complicadas de llevar a cabo.

## 10.2 Ampliaciones

Este proyecto es susceptible de implementar gran cantidad de extensiones debido a su propia naturaleza que contempla la extensión de los criterios de modificación por parte del cliente. Sin embargo, al margen de lo anterior, también se contemplan las mejoras que se indican a continuación.

1. Integración en profundidad de la herramienta con un sistema de indexado de los sitios web descargados. Sería interesante explorar esta posibilidad con Apache Solr ya que dicha integración se facilita en gran medida (Apache Nutch incorpora funcionalidad nativa para esta tarea).
2. Preparar soporte para otros software de persistencia (en especial Apache Hbase y Apache Accumulo).
3. Ofrecer una API mediante un servicio web para controlar la ejecución de la herramienta (comandar los trabajos de descarga, análisis y descarga de de los resultados en algún formato comprensible por máquinas).
4. Ofrecer otras plataformas alternativas (como *Crawl4J* o *Heritrix*) de araña web y permitir su coexistencia junto con Apache Nutch dentro de la aplicación. Además en este caso se debería permitir la elección de la araña a utilizar para cada tarea.
5. Realizar la gestión de la plataforma de computación distribuida (Apache Hadoop) mediante su API Java.
6. Integrar un sistema completo de búsqueda en el índice de Apache Solr desde la propia aplicación WebKumo.
7. Dar soporte a tipos MIME no textuales (PDF, ODF, imágenes etc.) mediante Apache Tika **[TIKA]** o similar.

# 11 Presupuesto

Se describirá en esta sección el presupuesto entregado al cliente así como el desglose de su contenido incluyendo mano de obra y materiales.

La elaboración de este presupuesto se ha basado en la planificación del proyecto entregada como anexo, donde se indican la cantidad de horas y recursos asignadas a cada tarea que se ha requerido llevar a cabo.

<b>CONCEPTOS DEL PRESUPUESTO</b>			
<b>Mano de Obra</b>			
<b>Concepto</b>	<b>Cantidad</b>	<b>Precio unitario</b>	<b>Total (€)</b>
Tareas de análisis y diseño	206h	23€/h	4738
Tareas de dirección	245h	35€/h	8575
Tareas de pruebas de software	32h	22€/h	704
Tareas de programación	1929h	18€/h	34722
Tareas de administración de sistemas	154h	24€/h	3696
Tareas de formación	64h	21€/h	1344
<b>Hardware</b>			
<b>Concepto</b>	<b>Cantidad</b>	<b>Precio unitario</b>	<b>Total (€)</b>
PC de Desarrollo	2355	0,01€/h	23,55
Servidor de despliegue	5250h	0,15€/h	787
<b>Otros</b>			
<b>Concepto</b>	<b>Cantidad</b>	<b>Precio unitario</b>	<b>Total (€)</b>
Desplazamientos	7	6,70€	46,9
Licencias software	-	0,0€	0

RESUMEN DEL PRESUPUESTO	
Concepto	Coste (€)
Mano de obra	53779
Hardware	810,55
Otros	46,9
<b>SUMA</b>	<b>54636,45</b>
<b>Beneficio industrial (15%)</b>	<b>8195,47</b>
<b>Subtotal</b>	<b>62831,92</b>
<b>IVA (21%)</b>	<b>13194,70</b>
<b>TOTAL</b>	<b>76026,62</b>

Los precios de cada concepto se han calculado según se indica a continuación:

- ◆ **Mano de obra:** Obtenido del salario medio para cada puesto en función su rol (administrador de sistemas, analista de sistemas etc.). Datos obtenidos del portal web PayScale.com **[PAYSC]**.
- ◆ **PC de desarrollo:** Coste imputado por horas de uso para un computador portátil con un consumo eléctrico de 0.06 Vatios/hora. Se tomó el precio medio de 0.124985€ **[ELEPR]** por Kilovatio y hora según la tarifa regulada en España en 2014.
- ◆ **Servidor de despliegue:** Horas estimadas desde primer uso hasta cierre del proyecto para una máquina disponible 24x7. Coste obtenido como media de distintos servicios *Platform As A Service* **[PAAS]** en sus modalidades intermedias.
- ◆ **Desplazamientos:** Coste de billete de ida y vuelta entre Gijón y Oviedo en tren de cercanías.

# 12 Referencias bibliográficas

## 12.1 Referencias en internet

**WINDEX:** The Web Index Report, 2015, <http://thewebindex.org/report/>

**GOOB:** We knew the web was big, 2008, <http://googleblog.blogspot.com.es/2008/07/we-knew-web-was-big.html>

**PMBOK:** Project Management Book Guide, 2014, Extraído de PMBok Guide 5th Edition

**WCAG:** Web Content Accessibility Guidelines, 2015, <http://www.w3.org/WAI/WCAG20/quickref/>

**WAI:** Web Accessibility Initiative, 2015, <http://www.w3.org/WAI/>

**W3C:** World Wide Web Consortium, 2015, <http://w3.org>

**UWEM:** Unified Web Evaluation Methodology, 2015, [http://www.wabcluster.org/uwem1\\_2/](http://www.wabcluster.org/uwem1_2/)

**TAW:** Tawdis, 2015, <http://www.tawdis.net/>

**ACHECK:** A Checker, 2015, <http://achecker.ca/>

**WAVE:** WAVE Web Aim, 2015, <http://wave.webaim.org/>

**AKEYS:** Access Color, 2015, <http://www.accesskeys.org/tools>

**HERA:** HERA: Sidar Foundation, 2015, <http://www.sidar.org/hera/>

**GOOMOV:** Google Mobile Friendly, 2015, <http://google.com/webmasters/tools/mobile-friendly/>

**SEO:** Search Engine Optimization, 2015, [http://en.wikipedia.org/wiki/Search\\_engine\\_optimization](http://en.wikipedia.org/wiki/Search_engine_optimization)

**WOORNK:** Seo Analyzer, 2015, <http://www.seoworkers.com/tools/analyzer.html>

**SAUDI:** Site Auditor, 2015, <http://raventools.com/seo-website-auditor/>

**STOOL:** Seobook Tool, 2015, <http://tools.seobook.com/>

**W3HTMLV:** W3C HTML Validator, 2015, <https://validator.w3.org/>

**W3CSSV:** W3C CSS Validator, 2015, <http://jigsaw.w3.org/css-validator/>

**NUVAL:** NU Validator, 2015, <https://validator.nu/>

**W3RDFV:** W3C RDF Validator, 2015, <http://www.w3.org/RDF/Validator/>

**SCANSVR:** Scan My Server, 2015, <https://www.scanmyserver.com/>

**SUCURI:** Sucuri Sitecheck, 2015, <https://sitecheck.sucuri.net/>

**VIRTOTAL:** Virus Total, 2015, <https://www.virustotal.com/>

**ACUNETIX:** Acunetix Scanner, 2015, <http://www.acunetix.com/vulnerability-scanner>

**MINIFIC:** Minimización de código, 2015, [http://en.wikipedia.org/wiki/Minification\\_%28programming%29](http://en.wikipedia.org/wiki/Minification_%28programming%29)

**COC:** Convention Over Configuration, 2015, [http://es.wikipedia.org/wiki/Convención\\_sobre\\_Configuración](http://es.wikipedia.org/wiki/Convención_sobre_Configuración)

**JSERVLET:** Java Servlet API, 2015, <http://docs.oracle.com/javaee/7/api/javax/servlet/package-summary.html>

**ORM:** Object Relational Mapping, 2015, [http://es.wikipedia.org/wiki/Mapeo\\_objeto-relacional](http://es.wikipedia.org/wiki/Mapeo_objeto-relacional)

**JPA:** Java Persistence API, 2015, <http://docs.oracle.com/javaee/7/api/javax/persistence/package-summary.html>

**JAAZ:** Java Authorization & Authentication API, 2015,

<http://docs.oracle.com/javase/7/docs/api/java/security/package-summary.html>

**MVC:** Model View Controller, 2015,

<http://martinfowler.com/eaDev/uiArchs.html#ModelViewController>

**IOC:** Inversion Of Control, 2015, [http://es.wikipedia.org/wiki/Inversi%C3%B3n\\_de\\_control](http://es.wikipedia.org/wiki/Inversi%C3%B3n_de_control)

**JNDI:** Java Naming & Directory Interface, 2015, <http://es.wikipedia.org/wiki/JNDI>

**STRU:** Apache Struts, 2015, <http://struts.apache.org/>

**JSP:** Java Server Pages, 2015, [http://es.wikipedia.org/wiki/JavaServer\\_Pages](http://es.wikipedia.org/wiki/JavaServer_Pages)

**JSON:** Javascript Oriented Notation, 2015, <http://es.wikipedia.org/wiki/JSON>

**XML:** Extensible Markup Language, 2015,

[http://es.wikipedia.org/wiki/Extensible\\_Markup\\_Language](http://es.wikipedia.org/wiki/Extensible_Markup_Language)

**SGBD:** Sistema de Gestión de Bases de Datos, 2015,

[http://es.wikipedia.org/wiki/Sistema\\_de\\_gesti%C3%B3n\\_de\\_bases\\_de\\_datos](http://es.wikipedia.org/wiki/Sistema_de_gesti%C3%B3n_de_bases_de_datos)

**CRUD:** Create, Read, Update, Delete, 2015, <http://es.wikipedia.org/wiki/CRUD>

**HQL:** Hibernate Query Language, 2015, <https://docs.jboss.org/hibernate/orm/5.0/manual/en-US/html/ch16.html>

**CRITQ:** Criteria Queries, 2015, <https://docs.oracle.com/javaee/6/tutorial/doc/gjrij.html>

**HDFS:** Hadoop Distributed File System, 2015, [http://en.wikipedia.org/wiki/Apache\\_Hadoop#HDFS](http://en.wikipedia.org/wiki/Apache_Hadoop#HDFS)

**MRED:** Map Reduce, 2015, <http://es.wikipedia.org/wiki/MapReduce>

**BSP:** Bulk Synchronous Parallel, 2015, [http://en.wikipedia.org/wiki/Bulk\\_synchronous\\_parallel](http://en.wikipedia.org/wiki/Bulk_synchronous_parallel)

**YAH01:** Hadoop Tutorial - Yahoo Dev Network, 2015,

<https://developer.yahoo.com/hadoop/tutorial/module1.html#data>

**HAD01:** Hadoop Tutorial - Single Node, 2015,

[http://hadoop.apache.org/docs/r1.2.1/single\\_node\\_setup.html#Local](http://hadoop.apache.org/docs/r1.2.1/single_node_setup.html#Local)

**HAD02:** Hadoop Tutorial - Multinode cluster, 2015,

[http://hadoop.apache.org/docs/r1.2.1/cluster\\_setup.html](http://hadoop.apache.org/docs/r1.2.1/cluster_setup.html)

**JAR:** Java Archiver, 2015, [http://es.wikipedia.org/wiki/Java\\_Archive](http://es.wikipedia.org/wiki/Java_Archive)

**HTCOM:** Apache HTTP Components, 2015, <https://hc.apache.org/>

**NUTCH1:** Apache Nutch - Plugins, 2015, <https://wiki.apache.org/nutch/PluginCentral>

**HBASE:** Apache HBase, 2015, <http://hbase.apache.org/>

**CASSA:** Apache Cassandra, 2015, <http://cassandra.apache.org/>

**ACCUM:** Apache Accumulo, 2015, <http://accumulo.apache.org/>

**SOLR1:** Apache Solr, 2015, <http://lucene.apache.org/solr/>

**ESEAR:** ElasticSearch, 2015, <https://www.elastic.co/products/elasticsearch>

**GORA:** Apache Gora, 2015, <http://gora.apache.org/>

**NUTCH2:** Apache Nutch API, 2015, <http://nutch.apache.org/apidocs/apidocs-2.2.1/index.html>

**TIKA:** Apache Tika, 2015, <https://tika.apache.org/>

**NOSQL1:** SGBD NoSQL, 2015, <http://en.wikipedia.org/wiki/NoSQL>

**NOSQL2:** SGBD NoSQL - Tipos, 2015,

[http://en.wikipedia.org/wiki/NoSQL#Types\\_of\\_NoSQL\\_databases](http://en.wikipedia.org/wiki/NoSQL#Types_of_NoSQL_databases)

**SHARD:** Escalabilidad Horizontal, 2015, [http://en.wikipedia.org/wiki/Shard\\_%28database\\_architecture%29](http://en.wikipedia.org/wiki/Shard_%28database_architecture%29)

**CAP:** Consistencia, Disponibilidad, Tolerancia a Particiones, 2015,

[http://es.wikipedia.org/wiki/Teorema\\_CAP](http://es.wikipedia.org/wiki/Teorema_CAP)

**CASBEN:** Comparación de SGBD NoSQL, 2015, <http://planetcassandra.org/nosql-performance-benchmarks/>

**SOLR2:** Apache Solr Cloud, 2015, <https://cwiki.apache.org/confluence/display/solr/SolrCloud>

**SQLIN:** Inyección SQL, 2015, [http://es.wikipedia.org/wiki/Inyecci%C3%B3n\\_SQL](http://es.wikipedia.org/wiki/Inyecci%C3%B3n_SQL)

**CSRF:** Cross Site Request Forgery, 2015, [http://es.wikipedia.org/wiki/Cross\\_Site\\_Request\\_Forgery](http://es.wikipedia.org/wiki/Cross_Site_Request_Forgery)

**JPAC:** Paquete Java, 2015, [http://es.wikipedia.org/wiki/Paquete\\_Java](http://es.wikipedia.org/wiki/Paquete_Java)

**HAD003:** Apache Hadoop Guide - Starting a single node cluster , 2015,

[http://wiki.apache.org/hadoop/GettingStartedWithHadoop#Starting\\_a\\_Single\\_node\\_cluster](http://wiki.apache.org/hadoop/GettingStartedWithHadoop#Starting_a_Single_node_cluster)  
**NUTCH3**: Apache Nutch CLI, 2015, <https://wiki.apache.org/nutch/CommandLineOptions>  
**CASSA2**: Descarga de Apache Cassandra, 2015, <http://cassandra.apache.org/download/>  
**SOLR3**: Apache Solr - REST API, 2015,  
<https://cwiki.apache.org/confluence/display/solr/Schema+API>  
**HADO3**: Apache Hadoop - Toolrunner, 2015,  
<https://hadoop.apache.org/docs/current/api/org/apache/hadoop/util/ToolRunner.html>  
**ROBU**: Diagramas de robustez, 2015,  
<http://www.agilemodeling.com/artifacts/robustnessDiagram.htm>  
**ANEM**: Modelos de datos anémicos, 2015, [http://en.wikipedia.org/wiki/Anemic\\_domain\\_model](http://en.wikipedia.org/wiki/Anemic_domain_model)  
**AJAX**: Asynchronous JavaScript and XML, 2015, <http://es.wikipedia.org/wiki/AJAX>  
**BDREL1**: Base de datos relacional, 2015, [http://es.wikipedia.org/wiki/Base\\_de\\_datos\\_relacional](http://es.wikipedia.org/wiki/Base_de_datos_relacional)  
**GPL**: Licencia Pública de GNU, 2015, [http://es.wikipedia.org/wiki/GNU\\_General\\_Public\\_License](http://es.wikipedia.org/wiki/GNU_General_Public_License)  
**BDREL2**: Cuotas de mercado de bases de datos relacionales, 2015,  
[http://en.wikipedia.org/wiki/Relational\\_database\\_management\\_system#Market\\_share](http://en.wikipedia.org/wiki/Relational_database_management_system#Market_share)  
**ACID**: Atomicidad, Consistencia, Aislamiento y Durabilidad, 2015,  
<http://es.wikipedia.org/wiki/ACID>  
**QUCAC**: Caché de consultas, 2015, <https://dev.mysql.com/doc/refman/5.1/en/query-cache.html>  
**SPECI**: Spring Data Specifications, 2015, <http://spring.io/blog/2011/04/26/advanced-spring-data-jpa-specifications-and-querydsl/>  
**CRITE**: , 2015, <https://docs.oracle.com/javaee/6/tutorial/doc/gjiv.html>  
**TESNG**: TestNG, 2015, <http://testng.org/doc/index.html>  
**MVN**: Apache Maven, 2015, <https://maven.apache.org/>  
**JDOC**: Java Document, 2015, <http://www.oracle.com/technetwork/articles/java/index-jsp-135444.html>  
**HTML5**: HTML5 , 2015, <http://www.w3.org/TR/html5/>  
**LINK**: Hiperenlace, 2015, <https://es.wikipedia.org/wiki/Hiperenlace>  
**CSS**: Hojas de estilo en cascada, 2015, [https://es.wikipedia.org/wiki/Hoja\\_de\\_estilos\\_en\\_cascada](https://es.wikipedia.org/wiki/Hoja_de_estilos_en_cascada)  
**DOM**: Árbol DOM, 2015, [https://es.wikipedia.org/wiki/Document\\_Object\\_Model](https://es.wikipedia.org/wiki/Document_Object_Model)  
**OJDK**: Open JDK , 2015, <http://en.wikipedia.org/wiki/OpenJDK>  
**JQUE**: JQuery, 2015, <http://en.wikipedia.org/wiki/JQuery>  
**BOOTS**: Bootstrap, 2015, [https://en.wikipedia.org/wiki/Bootstrap\\_%28front-end\\_framework%29](https://en.wikipedia.org/wiki/Bootstrap_%28front-end_framework%29)  
**GNLIN**: GNU/Linux, 2015, <http://es.wikipedia.org/wiki/GNU/Linux>  
**LMINT**: Linux Mint, 2015, <http://www.linuxmint.com/>  
**NBEAN**: Netbeans IDE, 2015, <https://netbeans.org/>  
**MZFIR**: Mozilla Firefox, 2015, [https://es.wikipedia.org/wiki/Mozilla\\_Firefox](https://es.wikipedia.org/wiki/Mozilla_Firefox)  
**UML**: Lenguaje de modelado unificado, 2015,  
[http://es.wikipedia.org/wiki/Lenguaje\\_unificado\\_de\\_modelado](http://es.wikipedia.org/wiki/Lenguaje_unificado_de_modelado)  
**DIA**: GNU Dia, 2015, <https://wiki.gnome.org/Apps/Dia>  
**MSQWO**: MySQL Workbench, 2015C, <https://www.mysql.com/products/workbench/>  
**LOFFI**: Suite Ofimática Libre Office, 2015, <https://es.libreoffice.org/>  
**TOMC1**: Apache Tomcat, 2015, <http://tomcat.apache.org/>  
**NATTY**: Detector de fechas en lenguaje natural, 2015, <http://natty.joestelmach.com/>  
**DTAB**: DataTables, 2015, <http://www.datatables.net/>  
**LANDE**: Java Language Detector, 2015, <https://github.com/shuyo/language-detection>  
**NUPAR**: Parser HTML NU, 2015, <https://about.validator.nu/htmlparser/>  
**PHLOC**: Parser CSS Phloc, 2015, <https://code.google.com/p/phloc-css/>  
**JSOUP**: Parser HTML Jsoup, 2015, <http://jsoup.org/>  
**NUTCH4**: Apache Nutch Mailing List, 2015, [http://mail-archives.apache.org/mod\\_mbox/nutch-user/201504.mbox/%3C07F72DA45B6F8846B15E715EBE92E3D948CDEFC5@exmbx05-](http://mail-archives.apache.org/mod_mbox/nutch-user/201504.mbox/%3C07F72DA45B6F8846B15E715EBE92E3D948CDEFC5@exmbx05-)

cdc.nexus.csiro.au%3E

**DHELL:** Infierno de dependencias, 2015, [http://en.wikipedia.org/wiki/Dependency\\_hell](http://en.wikipedia.org/wiki/Dependency_hell)

**HADO4:** Apache Hadoop ISSUE 10562, 2015, <https://issues.apache.org/jira/browse/HADOOP-10562>

**NUTCH5:** Apache Nutch ISSUE 1791, 2015, <https://issues.apache.org/jira/browse/NUTCH-1791>

**JACONF1:** Migración de configuración XML a Java Config para Spring, 2015, <http://www.robinhowlett.com/blog/2013/02/13/spring-app-migration-from-xml-to-java-based-config/>

**JACONF2:** Ejemplo de uso de Java Config en Spring MVC, 2015, <https://github.com/arey/spring-javaconfig-sample>

**EXSTA:** Códigos de salida, 2015, [https://en.wikipedia.org/wiki/Exit\\_status](https://en.wikipedia.org/wiki/Exit_status)

**HASS:** Guía de evaluación heurística, Hassan Y. Martín F.J, 2003, <http://www.nosolousabilidad.com/articulos/heuristica.htm>

**LYNX:** Navegador textual Lynx, 2015, [https://en.wikipedia.org/wiki/Lynx\\_%28web\\_browser%29](https://en.wikipedia.org/wiki/Lynx_%28web_browser%29)

**BROSH:** Cuota de mercado de navegadores, 2015, [https://en.wikipedia.org/wiki/Usage\\_share\\_of\\_web\\_browsers#Summary\\_tables](https://en.wikipedia.org/wiki/Usage_share_of_web_browsers#Summary_tables)

**MSIEX:** Microsoft Internet Explorer, 2015, [https://en.wikipedia.org/wiki/Internet\\_Explorer](https://en.wikipedia.org/wiki/Internet_Explorer)

**GOCHR:** Google Chrome, 2015, [https://es.wikipedia.org/wiki/Google\\_Chrome](https://es.wikipedia.org/wiki/Google_Chrome)

**WC143:** Criterio de éxtio 1.4.3 WCAG 2.0, 2015, <http://www.w3.org/TR/2008/REC-WCAG20-20081211/#visual-audio-contrast-contrast>

**WG185:** Técnica G185 WCAG 2.0, 2015, <http://www.w3.org/TR/WCAG20-TECHS/G185.html>

**GCOL:** Recolector de basura, 2015, [https://es.wikipedia.org/wiki/Recolector\\_de\\_basura](https://es.wikipedia.org/wiki/Recolector_de_basura)

**NIELS:** Tiempos de respuesta límite, 2015, <http://www.nngroup.com/articles/response-times-3-important-limits/>

**HADO5:** Descarga de Apache Hadoop, 2015, <https://archive.apache.org/dist/hadoop/common/hadoop-1.2.0/hadoop-1.2.0-bin.tar.gz>

**SSH:** Secure Shell, 2015, [http://es.wikipedia.org/wiki/Secure\\_Shell](http://es.wikipedia.org/wiki/Secure_Shell)

**HADO6:** Apache Hadoop IPv6, 2015, <http://wiki.apache.org/hadoop/HadoopIPv6>

**IPV6:** Protocolo de Internet versión 6, 2015, <http://es.wikipedia.org/wiki/IPv6>

**CASSA3:** Apache Cassandra configuración Multinodo, 2015, [https://wiki.apache.org/cassandra/GettingStarted#Configuring\\_Multinode\\_Clusters](https://wiki.apache.org/cassandra/GettingStarted#Configuring_Multinode_Clusters)

**SOLR4:** Descarga de Apache Solr, 2015, <ftp://ftp.cixug.es/apache/lucene/solr/4.10.3/solr-4.10.3.tgz>

**NUTCH6:** Descarga de Apache Nutch, 2015, <https://archive.apache.org/dist/nutch/2.2.1/apache-nutch-2.2.1-src.tar.gz>

**ACLAN:** Cabecera HTTP 'Accept-Language', 2015, [https://developer.mozilla.org/en-US/docs/Web/HTTP/Content\\_negotiation#The\\_Accept-Language.3a\\_header](https://developer.mozilla.org/en-US/docs/Web/HTTP/Content_negotiation#The_Accept-Language.3a_header)

**ROBOT:** Robots.txt, 2015, [http://es.wikipedia.org/wiki/Est%C3%A1ndar\\_de\\_exclusi%C3%B3n\\_de\\_robots](http://es.wikipedia.org/wiki/Est%C3%A1ndar_de_exclusi%C3%B3n_de_robots)

**IVY:** Apache Ivy, 2015, <http://ant.apache.org/ivy/>

**ANT:** Apache Ant, 2015, <http://ant.apache.org/>

**KEYSP:** Espacio de claves, 2015, [https://en.wikipedia.org/wiki/Keyspace\\_%28distributed\\_data\\_store%29](https://en.wikipedia.org/wiki/Keyspace_%28distributed_data_store%29)

**TOMC2:** Cuota de mercado de Apache Tomcat, 2015, <http://zeroturnaround.com/rebellabs/java-tools-and-technologies-landscape-for-2014/8/>

**SIGKI:** Señal POSIX SIGKILL, 2015, <http://es.wikipedia.org/wiki/SIGKILL>

**SPBOO:** Spring Profiles, 2015, <http://docs.spring.io/spring-boot/docs/current/reference/html/boot-features-profiles.html>

**H2:** Base de datos en memoria H2, 2015, <http://www.h2database.com/html/main.html>

**HSQldb:** HyperSQL Database, 2015, <http://hsqldb.org/>

- JDBC:** Java Database Connectivity, 2015, [http://es.wikipedia.org/wiki/Java\\_Database\\_Connectivity](http://es.wikipedia.org/wiki/Java_Database_Connectivity)
- SPDATA1:** Spring Data DB Vendors, 2015, <http://docs.spring.io/spring/docs/current/javadoc-api/org/springframework/orm/jpa/vendor/Database.html>
- HIBER1:** Dialectos para ORM Hibernate, 2015, <http://docs.jboss.org/hibernate/core/3.6/reference/en-US/html/session-configuration.html#configuration-optional-dialects>
- HIBER2:** Opciones de configuración adicionales de Hibernate, 2015, <https://docs.jboss.org/hibernate/orm/3.3/reference/es-ES/html/session-configuration.html#configuration-misc-properties>
- JACVIE:** Jackson JSON Views, 2015, <http://wiki.fasterxml.com/JacksonJsonViews>
- MVNRE:** Repositorio central de Maven, 2015, <http://search.maven.org/>
- MVNGUI:** Guia avanzada de Maven, 2015, <http://maven.apache.org/guides/index.html>
- PAYSC:** PayScale.com, 2015, <http://www.payscale.com/research/ES/Country=Spain/Salary>
- ELEPR:** Informe 2014 precio de la electricidad en Europa - FACUA, 2015, [http://facua.org/es/documentos/study\\_electricity\\_europe\\_en.pdf](http://facua.org/es/documentos/study_electricity_europe_en.pdf)
- PAAS:** Plataforma como servicio, 2015, [https://en.wikipedia.org/wiki/Platform\\_as\\_a\\_service](https://en.wikipedia.org/wiki/Platform_as_a_service)

## 12.2 Libros

1. **Design Patterns: Elements of Reusable Object-Oriented Software:** 1994, Gamma E, Helm R., Johnson R., Vlissides J.
2. **Head First Design Patterns:** 2004; Freeman E., Sierra K., Bates B.
3. **Hadoop: The Definitive Guide:** 2011, White T.
4. **Head First Ajax:** 2008, Riordan R.
5. **Eloquent JavaScript:** 2014, Haverbeke M.
6. **Head First HTML5:** 2011, Freeman E., Robson E.

# 13 Glosario

- ◆ **Araña web:** Software que realiza la descarga de portales web.
- ◆ **Computación distribuida:** Modelo de computación en la que el procesamiento se distribuye entre varios computadores.
- ◆ **Motor de búsqueda:** Software que indexa recursos de naturaleza variada y permite la consulta rápida de los mismos.
- ◆ **Persistencia de la araña web:** Sistema de gestión de bases de datos que la araña web utiliza para almacenar los recursos descargados.
- ◆ **WebKumo:** Nombre de la aplicación web que hace de intermediaria entre el usuario y la funcionalidad del sistema.
- ◆ **WebKumo Connector:** Plugin de Apache Nutch para integrar la funcionalidad de análisis de WebKumo en la fase de ejecución de la araña web Apache Nutch.
- ◆ **Análisis:** Evaluación de algún aspecto técnico de un portal web mediante es escrutinio de sus recursos web HTML, CSS y/o JavaScript.
- ◆ **Campo:** Elemento básico de un análisis que reporta un resultado sobre un aspecto evaluado.
- ◆ **Informe:** Conjunto de campos relacionados con un dominio de internet para un análisis concreto. Produce una puntuación sobre 100 para cada par análisis / dominio.
- ◆ **Trabajo:** Ejecución de uno o varios análisis sobre uno o varios dominios concretos. Los trabajos poseen diversas características configurables.
- ◆ **Servicio:** Infraestructura software que sustenta la funcionalidad del sistema.
- ◆ **Semilla:** URI de dominio a partir de la cuál se ejecutarán las funciones de un trabajo.

- ◆ **Iteraciones (rounds):** Número de iteraciones que la araña web ejecutará en el proceso de descarga. Cada iteración descubrirá nuevos enlaces a recursos del portal.
- ◆ **Recursos por nivel (topN):** Número de recursos que la araña web descargará por cada iteración.
- ◆ **Parámetro personalizado:** Pares de clave y valor que pueden utilizarse para comandar los análisis.
- ◆ **Salida de un servicio:** Texto explicativo que ofrece el servicio cuando se efectúa una operación sobre el mismo.

# 14 Contenido entregado

Se indica a continuación en forma de tabla los directorios que componen el contenido del proyecto entregado.

DIRECTORIO	DESCRIPCIÓN
./	Directorio raíz. Contiene un documento <code>readme.pdf</code> con esta tabla y el resto de directorios.
./src	Contiene el código fuente del proyecto así como un fichero de texto con la licencia.
./src/TFM_WebKumo	Directorio principal del proyecto. Contiene el fichero <code>pom.xml</code> que define el proyecto Apache Maven y sus dependencias.
./src/TFM_WebKumo/src	Directorio que contiene la estructura global de paquetes Java.
./src/TFM_WebKumo/doc	Contiene la API de documentación generada mediante JavaDoc.
./documentacion	Directorio que contiene la documentación del proyecto.
./documentacion/plan	Contiene los anexos de planificación del proyecto
./documentacion/diagramas	Alberga los diagramas realizados de clases y demás con el software libre de diagramas Dia.

El directorio `./src/TFM_WebKumo/src` puede desglosarse en los siguientes elementos:

DIRECTORIO	DESCRIPCIÓN
<code>./main</code>	Directorio de paquetes Java.
<code>./test</code>	Contiene la estructura de paquetes de las clases de test.
<code>./main/java</code>	Directorio que alberga los paquetes Java de WebKumo y WebKumo Connector
<code>./main/resources</code>	Alberga los recursos necesarios del classpath. Contiene subcarpetas y los ficheros de configuración que requiere Apache Hadoop y Apache Nutch cuando es ejecutado mediante su API Java.
<code>./main/resources/conf</code>	Contiene los ficheros de configuración de WebKumo
<code>./main/resources/i18n</code>	Este directorio contiene las cadenas de texto en Castellano e Inglés de WebKumo.
<code>./main/resources/lang</code>	Contiene los perfiles del detector de idioma requeridos por el análisis de internacionalización.
<code>./main/webapp</code>	Parte de la estructura de directorios que almacena los ficheros relacionados con la interfaz web.
<code>./main/webapp/META-INF</code>	Directorio de empaquetado (sin uso, pero requerido por Apache Maven).
<code>./main/webapp/WEB-INF</code>	Contiene en subcarpetas los ficheros que componen la interfaz.
<code>./main/webapp/WEB-INF/img</code>	Almacena las imágenes utilizadas en las páginas de la aplicación
<code>./main/webapp/WEB-INF/scripts</code>	Directorio que contiene los ficheros JavaScript utilizados en WebKumo.
<code>./main/webapp/WEB-INF/styles</code>	Contiene las hojas de estilo aplicadas en la interfaz de WebKumo.
<code>./main/webapp/WEB-INF/templates</code>	Almacena las plantillas de Freemarker que generan los documentos HTML de la interfaz web.