Experiencia de aplicación del programa Cumulus en la asignatura Fuentes de Información Especializada en Ciencia y tecnología

Manuel BLÁZQUEZ OCHANDO
Universidad Complutense de Madrid
manuel.blazquez@pdi.ucm.es

Recibido: 20/02/2011
Aceptado: 27/03/2011

RESUMEN
El programa informático Cumulus, especializado en la gestión de fuentes de información científica y técnica ha sido utilizado y probado extensivamente por los alumnos de la asignatura Fuentes de Información Especializada en Ciencia y Tecnología del curso 2009/2010, obteniendo resultados cuantitativos y cualitativos sobre una muestra de centros de investigación internacionales, que han sido estudiados de forma exhaustiva y asistida con la aplicación. Algunas de las conclusiones alcanzadas reflejan la importancia de utilizar programas que asistan al documentalista en el análisis y tratamiento de las fuentes de información científica más allá de la referenciación bibliográfica. Por otro lado se obtienen datos que reflejan el dinamismo y continua evolución de las fuentes científicas, su tipología, clasificación temática, así como su relevancia.
Palabras clave: Fuentes de información, Cumulus, documentación científica, automatización, centros de investigación internacional.

Implementation experience of Cumulus program: Focus management of sources for scientific and technical information.

ABSTRACT
Cumulus is an application specifically designed to deal with scientific and technical information sources. It has been extensively tested by the students of the course of "Sources for specialized scientific and technical information" during the academic year 2009/2010, with significant results, both qualitative and quantitative, on a sample of international research centres, that have been intensively studied with the assistance of this application. Some of our conclusions mirror the importance of using applications that assist in the analysis and management of information sources beyond the pure bibliographical reference. On the other side some data are obtained that project the dynamism and continuous evolution of scientific sources, as well as their typology and thematic classification together with its relevance.
Keywords: Information sources, Cumulus, scientific documentation, automation, international research centres.

1. INTRODUCCIÓN

El programa Cumulus constituye una plataforma básica para el desarrollo de aplicaciones especializadas en la gestión de fuentes de información. Desde su presentación (BLÁZQUEZ OCHANDO, M., 2010a) hasta la fecha, dicha herramienta ha sido explotada y probada por un grupo de 18 alumnos de la asignatura *Fuentes de Información Especializada en Ciencia y Tecnología* (BLÁZQUEZ OCHANDO, M., 2010b), impartida en la Facultad de Ciencias de la Documentación de la Universidad Complutense durante el curso 2009/2010. El objetivo del estudio es doble. Por un lado analizar una muestra de los principales centros de investigación científica y determinar cuáles son sus características, tipología y contenidos. Por otro, evaluar la herramienta experimental Cumulus, detectar qué aspectos podrían ser perfeccionados y cuáles resultan positivos para el estudiante y el docente-investigador. Hay que señalar que el programa Cumulus presenta un nuevo enfoque a la hora de tratar fuentes de información en red, basado en la distinción y definición de su tipología según su origen institucional, documental o personal (CHAÍN NAVARRO, C., 1995, pp.81-99). En el caso de los centros de investigación en ciencia y tecnología ello se aplica perfectamente, dada la jerarquía organizativa de las instituciones productoras de documentación científica, sostenidas por grupos de investigación y especialistas. En definitiva una compleja red de relaciones científico-institucionales entre la incoación de los estudios que se desarrollan bajo líneas de investigación y proyectos concretos.

2. METODOLOGÍA

La investigación parte de la inclusión de la herramienta Cumulus en el contexto de la asignatura Fuentes de Información en Ciencia y Tecnología. Se propuso la elaboración de un trabajo de curso consistente en el análisis y catalogación de 4 centros de investigación científica y sus fuentes de información derivadas. Éste se llevaría a cabo entre el 24 de marzo de 2010 hasta el 27 de julio de 2010. En esta prueba participaron 18 alumnos a los que se les proporcionaron accesos a la herramienta y los recursos pertinentes. Estos fueron tomados de la fuente original (Ranking Web of Research Centers, 2010) que elabora todos los años el Laboratorio de Cibermetría (CybermetricsLabs: Observatorio de ciencia y tecnología en internet, 2008) del CSIC. Distribuidos los recursos iniciales, se presentaron las siguientes bases metodológicas:

1. **Análisis principal:** Consiste en la descripción exhaustiva de los 4 centros de investigación asignados a cada alumno, atendiendo a su identificación y control, describiendo su denominación original, dirección URL y canales de sindicación de contenidos disponibles. En segundo lugar la tipificación de la fuente de información atendiendo a la clasificación según el tipo de recurso web, origen, nivel y contenido. A continuación se procede a la clasificación científico-temática de los contenidos partiendo de las directrices generales del sistema de clasificación UNESCO a dos primeros niveles (UNESCO, 1989). Seguidamente se utilizan los campos especiales de anotación en los que se describe el orga-

nigrama del centro de investigación, determinando centros dependientes, laboratorios, departamentos o áreas funcionales del mismo. Por otro lado se incluirá, si procede y existen, la referencia a catálogos bibliográficos y sistemas de consulta OPAC correspondientes a los centros de información y documentación presentes.

2. **Tratamiento de fuentes de información derivadas:** Por cada centro de investigación analizado se pidió la descripción de fuentes de información derivadas, concretamente 3 de tipo institucional (departamentos de investigación derivados, grupos de investigación, investigadores o asociaciones relacionadas) y 5 de tipo documental (bases de datos, revistas, directorios, monografías y demás publicaciones científicas). Hay que señalar que la cifra varió dependiendo de la fuente, por lo que se optó por un seguimiento flexible de cada institución, proporcional a la cantidad de recursos encontrados.
3. **Interrelación de las fuentes de información:** El programa Cumulus permite la relación de todas las fuentes de ingresadas entre sí según subordinación, nivel de equivalencia, alternancia, colección y tipo de servicio. Ello permitiría la construcción de una web semántica completa.

3. RESULTADOS DE LA INVESTIGACIÓN

La investigación ha dado como resultado el análisis de 662 fuentes de información de las que se extrajeron 229 autoridades corporativas, 305 personales, 126 geográficas y 72 editoriales. Tales datos indican hasta 732 puntos de acceso alternativos que posibilitan una mejor recuperación de cada fuente de información, *véase tabla1*. Todos los resultados pueden ser consultados en red a través del directorio del programa Cumulus disponible en: http://mblazquez.es/testbench/cumulus/

Datos generales	Nº total
Fuentes de información analizadas	662
Autoridades corporativas	229
Autoridades personales	305
Autoridades geográficas	126
Autoridades editoriales	72
Alumnos implicados en la investigación	18

Tabla 1. Datos generales de la investigación

La distribución de las fuentes de información según su clasificación temática demuestra una mayor preponderancia de las ciencias tecnológicas (157 fuentes) y de las ciencias del espacio (117 fuentes) frente a otras áreas de investigación como las ciencias

médicas, astronomía y astrofísica, *véase tabla2*. Esta disparidad puede deberse a una alta concentración de fuentes especializadas como por ejemplo informática (22 fuentes), telecomunicaciones (29 fuentes) y energía (16 fuentes) englobadas dentro de la categoría ciencias tecnológicas. Dentro de la categoría ciencias de la tierra y el espacio las principales áreas de investigación son espacio (35 fuentes), meteorología (21 fuentes) y geofísica (18 fuentes).

Tesauro operativo	**Nº total**
Clasificación general	15 categorías
Ciencias tecnológicas	157
Ciencias de la tierra y el espacio	117
Ciencias médicas	57
Organizaciones científicas	51
Ciencias de la vida	50
Astronomía y astrofísica	48
Física	36
Ciencias económicas	28
Humanidades	16
Matemáticas	15
Pedagogía	13
Ciencia de los materiales	7
Química	7
Lógica	1
Otros	59
Clasificación de términos específicos	128 categorías

Tabla 2. Clasificación temática de las fuentes basada en la clasificación UNESCO CTIC

En cuanto a la cobertura idiomática, la investigación desvela que el inglés es el idioma predominante con 469 fuentes analizadas, frente a otros como el francés o el español que apenas alcanza el medio centenar, *véase tabla3*.

Lengua de las fuentes de información	**Nº total**
Inglés	469
Francés	50
Español	49
Otros idiomas	94

Tabla 3. Cobertura idiomática de las fuentes de información analizadas

En cuanto al análisis de las fuentes según su nivel, se advierte un fenómeno de solapamiento, provocado por la ambivalencia de los recursos de información, dicho de

otra forma, una fuente puede proporcionar información de tipo primario y secundario. Los datos obtenidos muestran una importante cantidad de fuentes primarias y secundarias, quedando en menor medida las de tipo terciario o complementario, *véase tabla4*. La tasa de solapamiento entre las fuentes primarias y secundarias es del 35%, mientras que en categorías inferiores esta desciende paulatinamente, hasta el 28% cuando se trata de fuentes secundarias y terciarias.

Tipología de las fuentes según su nivel	**Nº total**
Fuentes primarias	510
Fuentes secundarias	440
Fuentes terciarias	184
Fuentes complementarias	52
Tasas de solapamiento	**Porcentaje**
Entre fuentes primarias y secundarias	35%
Entre fuentes secundarias y terciarias	28%
Entre fuentes primarias, secundarias y terciarias	12%

Tabla 4. Tipología de las fuentes según su nivel

La tipificación según el origen muestra dos clases de fuentes; institucionales y documentales que predominan en mayor medida, *véase tabla5*. Resulta de interés comprobar que la tasa de solapamiento entre fuentes documentales e institucionales apenas supera el 10% lo que indica una distinción clara de los recursos analizados. Por otro lado, al existir 333 fuentes institucionales, se advierte la existencia de un gran conglomerado de entidades, departamentos, etc. relacionados con los centros de investigación estudiados. Esta afirmación es posible al comprobar que existen hasta 120 grupos o proyectos de investigación y más de 229 autoridades corporativos que también fueron analizadas.

Tipología de las fuentes según su origen	**Nº total**
Fuentes documentales	412
Fuentes institucionales	333
Tasas de solapamiento	**Porcentaje**
Entre fuentes documentales e institucionales	11%

Tabla 5. Tipología de las fuentes según su origen

La tipología del recurso de información según su página web, también fue objeto de tipificación, obteniéndose resultados destacables como por ejemplo una mayoría de contenidos soportados con tecnología de la web dinámica, fundamentalmente PHP y MySQL, con más de 400 fuentes tratadas. En menor medida sitios web de tipo estático con 178 fuentes. Resulta reseñable la importancia de la recuperación de información en recursos científicos y la presencia de buscadores, en más de 130 casos. Por último las redes sociales y blogs tienen una discreta aparición o uso en el análisis global. *Véase tabla6*.

Tipología de las fuentes según su web	**Nº total**
Sitio web dinámico	416
Sitio web estático	178
Buscador	132
Social	52
Blog	34
Comercio electrónico	20
Directorio de recursos	17
Otros	23
Tasas de solapamiento	**Porcentaje**
Sitio web dinámico y buscador	13%
Sitio web dinámico y social	9,6%
Sitio web dinámico y blog	7,5%
Sitio web dinámico y directorio	2,8%

Tabla 6. Tipología de las fuentes según su web

La tipificación según contenidos resulta de las más importantes en el estudio, dado que permite conocer la topografía informacional de las fuentes de información en ciencia y tecnología. Los contenidos más abundantes son las publicaciones científicas con más de 580 recursos analizados, de los que destacan artículos y ensayos científicos con 286, seguidos de informes técnicos y reviews con 100 y actas de congresos, ponencias y simposios con más de 70. Otro gran grupo de recursos son los directorios, conformados por 254 recursos de los que 101 corresponden nuevamente a publicaciones científicas, 59 a servicios institucionales y 42 a centros de información como bibliotecas, archivos o centros de documentación. Las bases de datos merecen mención especial al haberse detectado en 186 recursos y al determinarse una tasa de solapamiento del 20% con respecto a los artículos y ensayos científicos, lo que demuestra que una gran parte son especializadas en este tipo de contenidos. Otro dato de interés es el alto número de proyectos y grupos de investigación, concretamente 120 recursos que encaja con el alto índice de fuentes institucionales analizadas. Finalmente existe una importante presencia de catálogos y repertorios bibliográficos especializados, con más de 140 referencias. *Véase tabla7*.

Tipología de las fuentes según sus contenidos	Nº total
Publicaciones científicas	**585**
Artículos y ensayos científicos	286
Monografías científicas	49
Patentes y modelos de utilidad	10
Actas de congresos, ponencias y simposios	73
Revistas científicas	67
Informes técnicos y reviews	100
Otras publicaciones	56
Directorios	**254**
Directorios de autoridades	18
Directorios de publicaciones científicas	101
Directorios de bases de datos	11
Directorios de centros de información y documentación	42
Directorios de editoriales científicas	15
Directorios de servicios institucionales	59
Otros directorios	8
Bases de datos	**186**
Grupos y proyectos de investigación	**120**
Catálogos y repertorios bibliográficos	**143**
Catálogos bibliográficos	136
OPACS	7
Tasas de solapamiento	**Porcentaje**
Entre artículos científicos y bases de datos	20,8%
Entre artículos científicos y directorios	9,6%
Entre artículos científicos y proyectos de investigación	12,6%
Entre artículos científicos y catálogos bibliográficos	4,36%
Entre bases de datos y directorios	7,7%

Tabla 7. Tipología de las fuentes según sus contenidos

El programa Cumulus también calculó el PageRank de todas las fuentes de información ingresadas en el sistema. Se obtiene que los recursos mejor valorados fueron encabezados por el *National Institute of Health*, el *National Weather Service* y el *Centre National de la Recherche Scientifique*. La primera fuente española, el *Consejo Superior de Investigaciones Científicas*, se encuentra en la posición 21 con un rango de 8, encabezando el segmento. *Véase tabla8*.

PageRank	Fuente cabecera de cada segmento	Nº de fuentes
9	NIH National Institute of Health	20
8	CSIC Consejo Superior de Investigaciones Científicas	68
7	National Institute for Health and Clinical Excellence	131
6	Bureau International des Poids et Mesures	153
5	Istituto Nazionale di Geofisica e Vulcanologia	104
4	Revista Española de Documentación Científica	186
Para más información puede consultar el listado completo en: http://www.mblazquez.es/documents/experiencia-cumulus_100pagerank.html		

Tabla 8. Cantidad de fuentes según su PageRank

Como resultado de establecer las relaciones entre las fuentes de información y sus autoridades, el programa Cumulus generó un archivo semántico con todos los contenidos analizados. Dicho archivo puede ser consultado en la siguiente dirección URL, http://www.mblazquez.es/testbench/cumulus/output/sources.rdf.

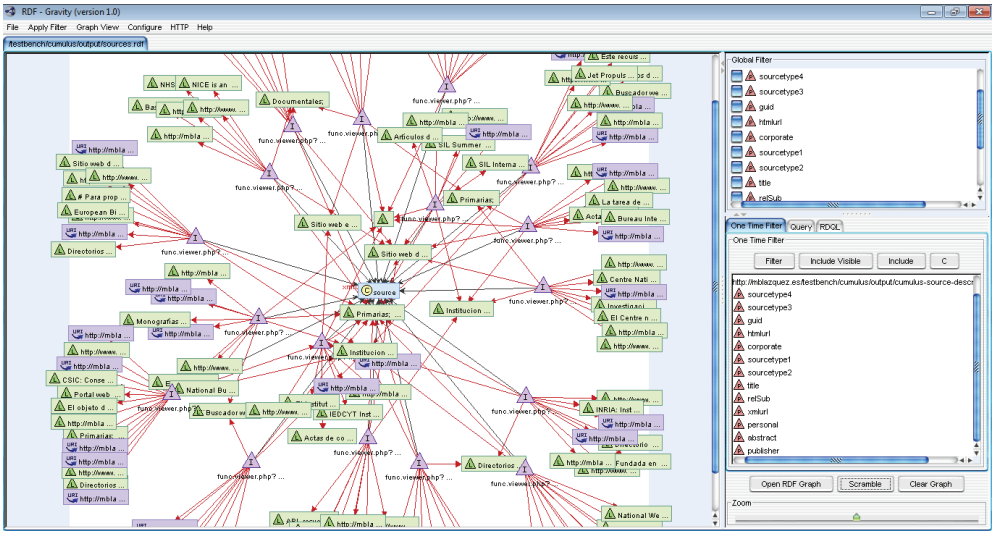

Figura 1. Web semántica de las fuentes de información analizadas con CUMULUS

Para lograr una representación gráfica del mismo se ha empleado la herramienta especializada RDF Gravity (GOYAL, S. y WESTENTHALER, R.) Se obtiene como resultado una perspectiva aproximada de las dimensiones y tejido con que está confeccionada la red, advirtiéndose toda la estructura de triples en un modelo de representación radial, *véase figura1*.

5. FUTURAS LINEAS DE DESARROLLO Y MEJORAS

Pese a todos los resultados obtenidos, no hay que olvidar que el programa Cumulus aún sigue siendo una herramienta experimental en desarrollo. Este hecho justifica en mayor medida la atención a todos los defectos y problemas que pueda causar al usuario, en este caso a los alumnos participantes en el experimento. Al finalizar su trabajo se les pidió evaluar libremente la herramienta exponiendo sus ventajas, inconvenientes y aspectos mejorables. El resultado fue satisfactorio porque en general casi todos estaban de acuerdo en destacar la originalidad y necesidad de un programa informático que les asistiera en la gestión y catalogación de fuentes de información. El método ordenado y coherente de análisis, así como la puesta en práctica de las distintas tipificaciones de las fuentes de información, fueron señaladas como los principales rasgos de calidad y utilidad del programa para el desempeño de su trabajo. Por otro lado se detectaron algunas deficiencias en la programación del mecanismo de relaciones del sistema, que si bien no impedían la labor, la dificultaban por su complejidad. Además se señaló la importancia de mejorar los procesos de gestión de autoridades, con especial énfasis en la recuperación y detección de duplicados. Todas estas consideraciones han sido tenidas en cuenta y serán resueltas en la próxima versión del programa, denominada Cumulus2, con la que se formarán a los futuros cursos de Fuentes de Información.

6. CONCLUSIONES

1. Cumulus es una herramienta ideal para obtener más información sobre la topografía y características de las fuentes de información al obtener datos del tipo de fuentes, contenidos, cobertura idiomática, pagerank, autoridades y categorización temática. Por otro lado, proporciona un directorio automático que puede ser consultado en línea de tal forma que la recuperación pueda llevarse a cabo mediante texto libre, índices y canales de sindicación especializados.
2. Los contenidos predominantes en las fuentes de información especializadas en ciencia y tecnología son los artículos y ensayos científicos por encima de otro tipo de publicaciones como las patentes y monografías. De hecho su nivel de interrelación con las bases de datos y los proyectos de investigación detectados, desvela que es el principal medio de comunicación científica. En menor medida se encuentran los blogs y redes sociales cuya presencia aún sigue siendo bastante reducida.
3. Se detecta una gran cantidad de fuentes de información de tipo institucional, refrendado por el importante número de autoridades corporativas, editoriales así como por el número de grupos de investigación y departamentos analizados. Esto supone que algo menos del 50% de las fuentes de información científicas que han sido analizadas son de tipo institucional y que de ellas depende cerca de un 60% de las grandes fuentes documentales como directorios, bases de datos con acceso a artículos científicos, bibliografías, catálogos etc.

4. Se comprueba una vez más que el idioma conductor de la investigación científica en la mayor parte de los centros y consejos internacionales de investigación es el inglés, quedando en minoría el francés, español y alemán. La mayor parte de las fuentes de información analizadas corresponden al ámbito de ciencias del espacio, tecnología informática, telecomunicaciones, ciencias médicas, ciencias de la vida, astronomía y astrofísica, que suman en total 480 fuentes, un 72% del total.
5. La mayor parte de las fuentes analizadas presenta tecnología dinámica PHP y MySQL lo que implica un amplio nivel de desarrollo, cercano al 62% de todos los contenidos analizados. Pese a este hecho la web estática puede llegar a soportar contenidos de otra índole como especificaciones, informes, artículos o monografías que no requieren de actualizaciones o cambios dinámicos, siguiendo la pauta del archivado digital. Teniendo en cuenta ese caso, la web estática se cifra en torno al 26% del total estudiado.
6. El índice de PageRank obtenido en el estudio, indica que existen 6 centros de investigación españoles entre los 100 mejor valorados, pero no existe aún ninguno con rango 9, lo que indica la necesidad de mejorar no sólo la visibilidad, sino los contenidos que deben publicarse en inglés para obtener una mayor referenciación por parte de entidades e instituciones internacionales.

7. BIBLIOGRAFÍA

BLÁZQUEZ OCHANDO, M. 2010a. Gestión de fuentes de información en ciencia y tecnología: desarrollo del programa CUMULUS. *In*: *VII Seminario Hispano Mexicano de Investigación en Bibliotecología y Documentación*. México: CUIB.

BLÁZQUEZ OCHANDO, M. 2010b. *Fuentes de Información en Ciencia y Tecnología*. [online]. [Accessed 10 Feb 2010]. Disponible en: http://ccdoc-fuentescienciatecnologia.blogspot.com/

CHAÍN NAVARRO, C. 1995. *Introducción a la gestión y análisis de recursos de información en ciencia y tecnología*. Murcia: Servicio de publicaciones, Universidad.

CybermetricsLabs: Observatorio de ciencia y tecnología en internet. 2008. [online]. [Consultado 15 Feb 2010]. Disponible en: http://internetlab.cindoc.csic.es/

GOYAL, S. and R. WESTENTHALER. *RDF Gravity (RDF Graph Visualization Tool)*. [online]. [Consultado 30 Diciembre 2010]. Disponible en: http://semweb.salzburgresearch.at/apps/rdf-gravity/index.html

Ranking Web of Research Centers. 2010. [online]. [Consultado 12 Febrero 2010]. Disponible en: http://research.webometrics.info/

UNESCO. 1989. *Internationa standard nomenclature fields for science and technology*. [online]. [Consultado 02 Mayo 2010]. Disponible en: http://unesdoc.unesco.org/images/0008/000829/082946eb.pdf

AGRADECIMIENTOS

Quisiera agradecer la colaboración de los alumnos de la asignatura de Fuentes de Información en Ciencia y Tecnología del curso 2009/2010 en la Facultad de Ciencias de la Documentación de la Universidad Complutense por sus aportaciones durante el experimento y pruebas realizadas con la herramienta Cumulus.