D.6. El ecosistema de la información científica: estructura y niveles de agregación

Por **Ernest Abadal** y **Lluís Codina**

14 febrero 2011

Abadal, Ernest; **Codina, Lluís**. "El ecosistema de la información científica: estructura y niveles de agregación". *Anuario ThinkEPI*, 2011, v. 5, pp. 128-131.

Resumen: *Se presentan los principales tipos de productos para el acceso a la información científica junto con sus rasgos diferenciales en cuanto a los contenidos analizados, la técnica utilizada (asignación y recolección de metadatos, indización o búsqueda federada) y los resultados ofrecidos. Los productos considerados son los siguientes: bases de datos bibliográficas, portales de revistas, repositorios, motores de búsqueda académicos, recolectores, metabuscadores académicos y metabuscadores de biblioteca.*

Palabras clave: *Información científica, Agregadores, Motores de búsqueda académicos, Repositorios, Portales de revista, Metabuscadores académicos, Bases de datos bibliográficas, Metabuscadores de bibliotecas.*

Title: ***The scientific information ecosystem: structure and aggregation levels***

Abstract: *The structure and characteristics of the main products for accessing scientific information are described. The different types of content, the technology used (assignment and harvesting of metadata, full text indexing, or federated search) and their results are analysed. The products are: bibliographic databases, academic journals portals, repositories, academic search engines, academic metasearch engines, and library metasearch engines.*

Keywords: *Scientific information, Academic search engines, Repositories, Academic journals portals, Academic metasearch engines, Bibliographic databases, Libraries metasearch engines.*

La información académica

LA INFORMACIÓN ACADÉMICA o científica difunde los resultados de la investigación a través de artículos de revista, contribuciones a congresos, tesis, patentes, etc.

Constituye un sector económico específico que dispone de una industria editorial –con *Reed-Elsevier* y *Thomson Reuters* a la cabeza– que se ha visto afectada en los últimos años por los procesos de digitalización y por la irrupción del acceso abierto.

El número de contenidos generados es altísimo[1] y explica que se hayan creado diversos productos y servicios pensados específicamente para ayudar a los científicos a localizar y consultar documentos de su interés. Durante muchos años –desde finales de los 60– las bases de datos bibliográficas fueron los únicos instrumentos que facilitaban a los investigadores la localización de referencias científicas. A principios de 2000 aparecieron los motores de búsqueda académicos, que incluyen toda clase de documentos publicados en sitios web relacionados con la actividad investigadora (con *Scirus* y *Google Scholar* al frente); y a partir de aquí, otros productos y servicios han hecho acto de presencia.

"Durante años, las bases de datos bibliográficas fueron los únicos instrumentos que facilitaban a los investigadores la localización de referencias científicas"

Nuestro objetivo es presentar una tipología del conjunto de sistemas de acceso a la información científica que actualmente forman un ecosistema con nichos bien separados, pero también con elementos en competencia que se solapan.

Tipos de productos para acceder a la información científica

En la tabla 1 presentamos una propuesta de caracterización de los distintos productos de acceso a la información científica existentes, junto con una descripción de sus rasgos esenciales.

Contenidos analizados

Una primera diferenciación de los productos de la tabla la podríamos establecer en función de los contenidos analizados. De esta manera se pueden distinguir dos niveles de agregación, según se almacene y se indice directamente la fuente original de la información científica (artículos de revista, congresos, tesis, etc.) o se llegue a ellos de forma indirecta, a través de algún otro producto agregador como los portales de revista o los repositorios.

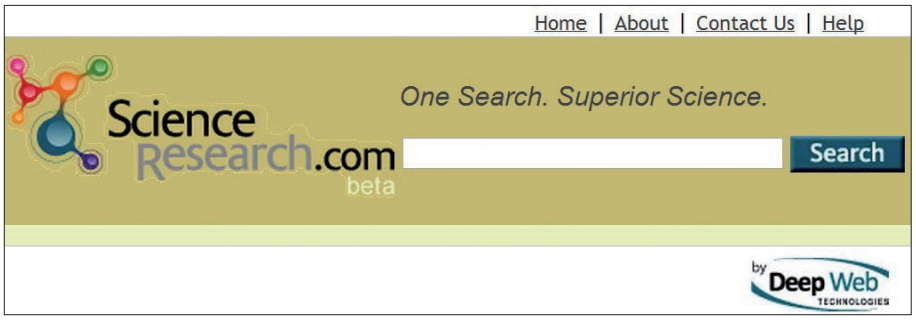
En el primer nivel de agregación encontramos los sistemas de recuperación que se nutren directamente de la fuente original de la información científica, es decir, que toman como referencia los artículos de revista, las contribuciones a congresos, las tesis, etc., independientemente de que almacenen los textos completos o no.

Estos productos son las bases de datos bibliográficas, los portales de revistas (ya sean comerciales o de acceso abierto), y los repositorios.

En el segundo nivel se encuentran aquellos servicios que se nutren del primer nivel, es decir, que incluyen contenidos procedentes de portales de revistas y de repositorios. Estos sistemas no van a buscar las fuentes (las revistas, las tesis o los congresos) en su lugar de origen, sino que llegan a ellas indirectamente por medio de los

Producto	Contenidos analizados	Tecnología	Resultados	Inicios	Coste	Ejemplos	Nivel de agregación
Bases de datos bibliográficas	Fuentes primarias: artículos de revista, congresos, etc.	– Asignación de metadatos (registros bibliográficos)	Registros bibliográficos + acceso a un sistema de resolución de enlaces	Finales de 1960	Comerciales	*Scopus*, *ISI WoS*, *Chemical Abstracts*, *Eric*, *Dialog*	1
Portales de revistas	Fuentes primarias: artículos de revistas	– Asignación de metadatos (registros bibliográficos) – Indización del texto completo	Registros bibliográficos + acceso al documento original	Finales de 1990	Comerciales y gratuitos	*Emerald*, *Scielo*, *ScienceDirect*, *Recyt*	1
Repositorios	Fuentes primarias: artículos de revista, tesis, congresos, etc.	– Asignación de metadatos (registros bibliográficos) – Indización del texto completo	Registros bibliográficos + acceso al documento original	Finales de 1990	Gratuitos	*E-LIS*, *DDD* (*UAB*), *MIT DSpace*, *Repositorium*	1
Motores de búsqueda académicos	– Portales de revistas – Repositorios – Sedes web académicas	Indización del texto completo	Lista de enlaces	2000	Gratuitos	*Google Scholar*, *Scirus*	2
Recolectores	– Portales de revistas – Repositorios	– Recolección de metadatos	Registros bibliográficos + acceso al documento original	Mediados de 2000	Gratuitos	*OAIster*, *Recolecta*, *Hispana*, *Arrow*	2
Metabuscadores académicos	– Portales de revistas – Repositorios – Motores de búsqueda académicos	– Búsqueda federada	Lista de enlaces	Principios de 2000	Gratuitos y comerciales	*ScienceResearch*, *Biznar*	2
Metabuscadores de bibliotecas	– Repositorios – Portales de revistas suscritas – Catálogo de la biblioteca	– Búsqueda federada	Registros bibliográficos + acceso al documento original	Principios de 2000	Comerciales	*MetaLib*, *Encore*	2

Tabla 1. Productos principales para el acceso a la información científica



http://www.scienceresearch.com

– Búsqueda federada: consiste en enviar la misma consulta a cientos de fuentes (agregadores de primer nivel que indizan las fuentes primarias), en lugar de volverlas a indizar directamente. Como en el caso anterior, el usuario también recibe una lista única de resultados.

agregadores de primer nivel. De esta forma les basta con acudir a unos pocos miles de sedes web para hacerse con millones de contenidos.

Aquí están los motores de búsqueda académicos, los recolectores, los metabuscadores académicos y los metabuscadores de biblioteca.

"El mayor número de consultas a los repositorios procede de agregadores de segundo nivel y no tanto de consultas directas"

Tecnología

Los fundamentos técnicos utilizados por los productos analizados son cuatro:

– Asignación de metadatos (catalogación e indización): proceso intelectual (no automático) que consiste en elaborar un registro bibliográfico para cada una de las fuentes originales analizadas. Puede ser realizado por el mismo autor que crea los contenidos (artículos de revista, etc.) o por el analista de la base de datos, portal de revistas o repositorio.

– Indización automática del texto completo: consiste en extraer (todos) los términos de los contenidos seleccionados, que pueden estar más o menos dispersos en servidores, y generar un índice global como resultado.

– Recolección de metadatos: se crea un índice común recolectando (sólo) metadatos de los repositorios que cumplen un protocolo de etiquetado común (OAI-PMH). El usuario recibe una lista única de resultados.

Lista de resultados

Las páginas de resultados de estos productos pueden ser de tres tipos:

a) Registros bibliográficos + acceso al documento original (portales de revista, repositorios, recolectores).

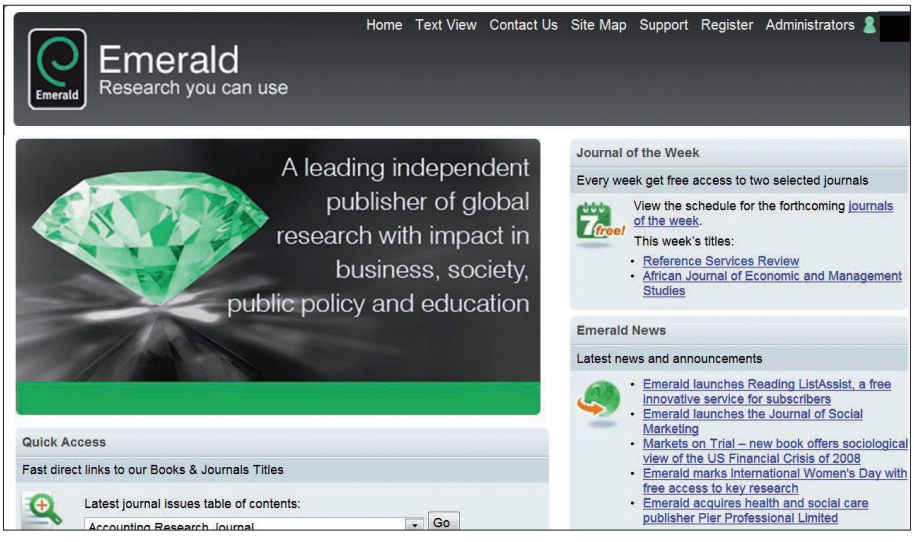
b) Registros bibliográficos + acceso a un sistema de resolución de enlaces (bases de datos bibliográficas).

c) Lista de enlaces (motores académicos, metabuscadores).

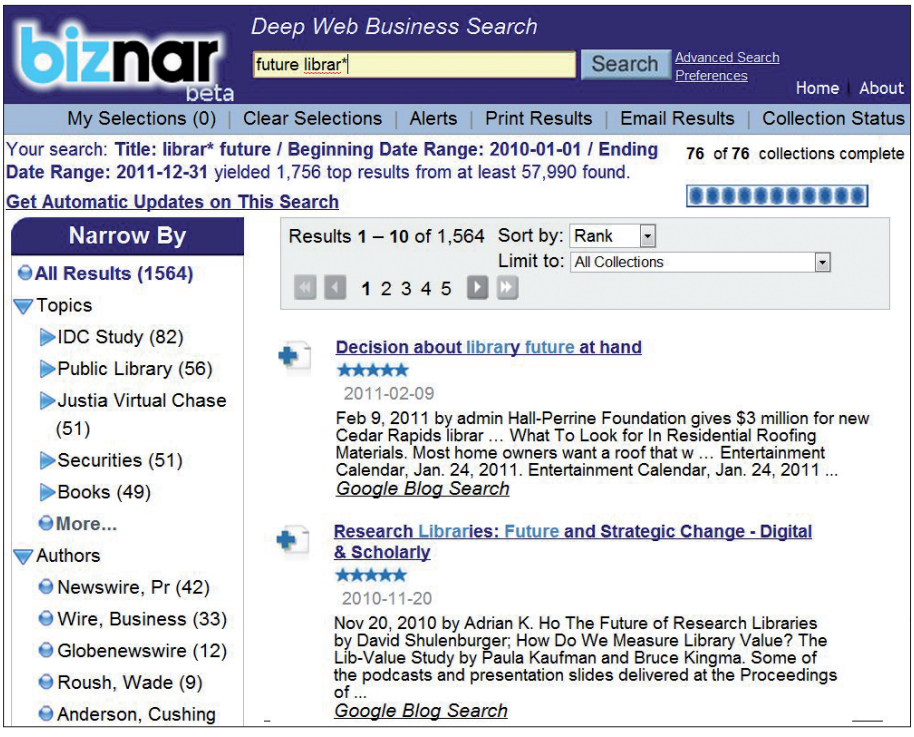
En el primer caso (a), el usuario tiene solucionado siempre de forma directa el paso siguiente a cualquier búsqueda: la obtención del documento.

En el segundo caso (b), lo tiene solucionado de forma parcial, es decir, en ocasiones el sistema de resolución de enlaces lo conducirá al documento completo, pero a veces no, y además deberá hacerlo en varios pasos.

En el tercer caso (c), se ofrece una lista de enlaces a otros sitios web de modo que el acceso en principio es directo, pero lleno de casuística: enlaces rotos, servidores que requieren suscripción, etc.



http://www.emeraldinsight.com



http://biznar.com

Consideraciones finales

De la estructura y niveles de agregación antes descritos se desprenden diversos comentarios para algunos de los principales agentes de la comunicación científica:

Para los científicos como usuarios de información

Es frecuente que los investigadores estén suscritos a las alertas de las revistas de su máximo interés. De todas formas, para las búsquedas sistemáticas y exhaustivas acostumbran a utilizar mayoritariamente, y de forma intensiva, los recursos de segundo nivel (especialmente los motores de búsqueda académicos) y también las bases de datos, que les aseguran el acceso a un mayor número de fuentes primarias. Esto es lógico ya que no es práctico tener que ir recorriendo los centenares de portales de revistas o de repositorios.

Para los editores de las revistas

Dado que los científicos consultan fundamentalmente agregadores de segundo nivel, es importante para una revista estar presente en portales de revistas o en repositorios, ya que son el paso esencial e imprescindible para poder ser incluidas en motores de búsqueda y metabuscadores. Es muy difícil estar en el segundo nivel sin pasar por el primero.

Para los repositorios

Los contenidos incluidos en repositorios tienen asegurada la presencia en el segundo nivel. Esto es muy importante para los contenidos que están depositados en ellos. El mayor número de consultas a los repositorios procede de agregadores de segundo nivel y no tanto de consultas directas.

Para los científicos en tanto que autores

Si quieren asegurar una máxima difusión a sus obras tienen que publicar en revistas incluidas en portales y, si no son de acceso abierto, depositar sus textos en repositorios. De esta forma tienen asegurada su inclusión en motores de búsqueda y metabuscadores académicos y una fácil localización por parte de sus colegas.

Notas

1. Para tener una referencia: tan sólo las revistas académicas activas son unas 78.000 (según *Ulrich's*).

Comentario

La clasificación presentada en este artículo se ha hecho atendiendo al análisis de los contenidos. Además existen empresas que no indizan ni recopilan metadatos, sólo son distribuidores:

– De bases de datos (hosts). Reciben las bases de datos bibliográficas de los productores y las cargan en su ordenador junto a otras, ofreciéndolas al público con un mismo software de consulta para todas. Ejemplos: *Dialog*, *Questel-Orbit*, *EbscoHost*.

– De portales de revistas. Muchas editoriales no quieren instalar equipos informáticos propios para distribuir sus revistas, y subcontratan el servicio a esas empresas. Los distribuidores alojan los pdfs y los ofrecen a través de webs (portales) con la imagen y logotipos de la editorial. Ejemplos: *MetaPress*, *HighWire Press, IngentaConnect*.

– Agentes de suscripciones. Actúan como agregadores de portales de revistas. Ejemplo: *Swets* (*SwetsWise*).