

Sistemas de información

Reference linking: un nuevo concepto para facilitar el acceso a la literatura científica

Por José Manuel Barrueco

1. El concepto de *reference linking*

LA CIENCIA MODERNA es una actividad social en la que no existe el trabajo individualista de una sola persona. Los avances y descubrimientos son fruto de la actividad de grupos de investigadores que ponen en común sus estudios por encima de limitaciones geográficas o temporales. Igual que no existe el investigador individual, tampoco puede existir un documento científico aislado.

Si este tipo de material no es más que la presentación en un soporte de los hallazgos y descubrimientos que los autores han realizado, es lógico que éstos hagan referencia explícita a cuáles han sido sus puntos de referencia, sus colaboradores, los trabajos que le han precedido, etc. Dichas menciones se llevan a cabo a través de las citas y las referencias bibliográficas. De esta forma, cada documento nos habla de otros, formando así la inmensa red que es la literatura científica.

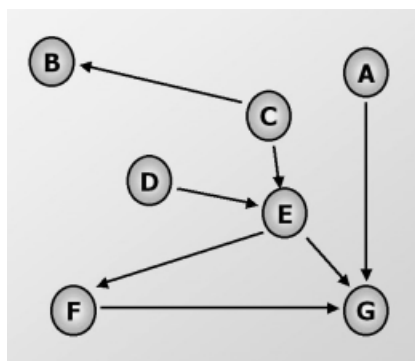


Figura 1



José Manuel Barrueco

Gráficamente se la podría representar como se ve en las figuras 1 y 2. Cada nodo representa un documento. Las flechas, a su vez, se identifican con las relaciones marcadas por las referencias bibliográficas. Las que parten de los nodos son las que ha realizado su autor a otros documentos publicados con anterioridad. Aquellas que llegan a ellos serían citas que tales documentos han recibido. Así por ejemplo, E referencia dos trabajos: F y G, mientras que a su vez es citado por otros dos C y D.

«Cada documento nos habla de otros, formando así la inmensa red que es la literatura científica»

De lo dicho se puede intuir la importancia de las listas de referencias. Autores como **Cameron**⁴ han imaginado una base de datos universal de citas que permitiera unir cualquier trabajo científico escrito en la historia con los documentos a los que hace referencia.

Describe un sistema en donde cualquier documento estaría disponible y podría ser localizado a través de internet. La base de datos incluiría información sobre citas y sería exhaustiva y actualizada.

Sin embargo, su propuesta requiere que los autores o instituciones proporcionen esta información en un formato específico. Esto es muy costoso de hacer en el futuro y muy difícil de lograr sobre el material del pasado, por lo que no es sorprendente que su proyecto nunca se haya hecho realidad. Al menos una de las condiciones que plantea **Cameron** sí que se está cumpliendo: cada vez son más los trabajos que están disponibles en formato electrónico a través de la Red. El número de investigadores que están colocando publicaciones en sus páginas web o archivos de documentos está creciendo cada día. Es el caso por ejemplo de *arXiv* en física, *CogPrints* en ciencias del conocimiento o *RePEc* en economía. Para estudiar la interoperatividad de dichos archivos⁶ se

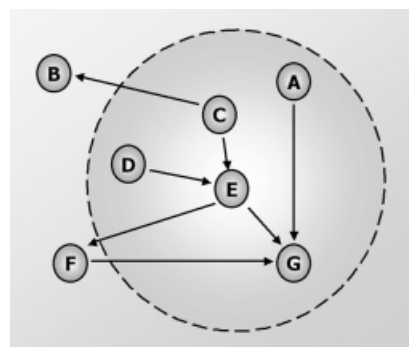


Figura 2

ha creado recientemente la *Open Archives Initiative (Oai)*.

Con la existencia de los documentos en formato electrónico ha surgido la posibilidad de tratar los textos automáticamente para extraer las referencias sin intervención humana; y también sin añadir un trabajo adicional a los autores. De esta forma, si un sistema informático es capaz de determinar dónde se encuentra la sección de referencias en un documento, individualizar cada una de ellas y diferenciar cuál es el título, los autores, la fuente donde ha sido publicado y su dirección electrónica, cabría la posibilidad de crear un enlace entre ambos documentos que nos permitiera ir de uno a otro a través del hiperespacio electrónico.

«CrossRef no tiene que ver nada con control de accesos a los documentos»

La posibilidad de unir una referencia con el texto completo del documento al que representa es lo que los anglosajones llaman *reference linking* o enlace de referencias y que está siendo objeto de múltiples estudios en la actualidad. El interés procede de dos áreas claramente diferenciadas: en primer lugar de la industria editorial y de servicios de resúmenes, quienes han visto el enorme potencial que representa la interconexión de los miles de documentos que se están distribuyendo en formato electrónico. El principal exponente de esta área es *CrossRef*, proyecto que se describirá con posterioridad.

Por su carácter, los resultados están limitados por innumerables barreras económicas y lamentablemente sólo podrán tener acceso a ellos los investigadores de países desarrollados. Por el contrario el segundo campo de investigación procede de los archivos de documentos electrónicos accesibles

gratuitamente en internet. Aquí tenemos iniciativas como *OpCit*, *CiteSeer* o *Cern* que también serán descritas en este texto. Desde el punto de vista del profesional de la información son mucho más interesantes pues se están desarrollando sistemas basados en inteligencia artificial que son capaces de realizar el enlace de referencias de forma automática.

2. CrossRef

Las editoriales comerciales se dieron cuenta desde muy pronto de las ventajas que les aportaría tener enlazados todos los documentos que publican. Como respuesta a esta necesidad surgió en 1999 el proyecto *CrossRef* puesto en marcha por la *Publishers International Linking Association (Pila)* y descrito en **Keefer**⁷. Ésta es una empresa que engloba a más de ochenta editores y proveedores de servicios de resúmenes de todo el mundo y su funcionamiento es bastante sencillo. No tiene nada que ver con las iniciativas que veremos, dado que aquí las referencias ya están disponibles como parte de la edición del texto; además se dispone de abundante metadata por lo que no es necesario realizar un análisis del documento para encontrarla.

Los editores participantes aportan información bibliográfica sobre los documentos que publican a una base de datos común. Por cada artículo se requieren datos básicos como el título de la revista donde se ha publicado, volumen, número, etc., aunque pueden aportar otra información adicional según su elección. Además a cada artículo se le asigna un doi (*digital object identifier*)¹⁰ y una url asociada para recuperar el texto completo de los documentos. El intercambio de información entre editoriales y *CrossRef* se realiza utilizando el formato xml, siendo obligación de cada una mantener actualizada

Algunas direcciones de interés

<http://www.arxiv.org>

<http://www.openarchives.org>

<http://www.crossref.org>

<http://www.doi.org>

<http://www.dlib.org>

la correspondencia entre dois y urls de cada documento.

La base de datos contiene en estos momentos casi 4 millones de documentos. Cuando una editorial va a publicar un nuevo artículo debe, por un lado, remitir sus datos bibliográficos a *CrossRef* y por otro tomar la lista de referencias e interrogar la base de datos de *CrossRef* para ver si están disponibles en formato electrónico. Para ello, las referencias son enviadas a un *reference resolver* que se encargará de devolver el doi del documento citado si existe y que se incluirá en el artículo publicado. En el futuro, cuando alguien quiera acceder al documento citado será el sistema doi quien se encargue de convertir ese código en la dirección de internet donde se almacene el texto completo. Adicionalmente este proceso se podría realizar también de forma dinámica, en el momento en que un usuario seleccionara una referencia de la bibliografía del artículo una vez publicado.

«CiteSeer es un verdadero índice de citas construido de forma automática»

El resultado de una interrogación a *CrossRef* solamente será un código que se resolverá en una url que apunte al texto del documento. Será competencia de cada editorial vigilar si quien está intentando acceder a él dispone de una suscripción o, en su caso, establecer un sistema de pagar por ver para cada usuario individual. En definitiva, *CrossRef* no tiene que ver nada con

control de accesos a los documentos, e igualmente no tiene relación con ellos en sí mismos ya que lo único que almacena son sus metadatos.

3. *CiteSeer*

Ha sido desarrollado en el departamento de investigación de *NEC* por **Steve Lawrence**, **Kurt Bollacker** y **C. Lee Giles**. Es el punto de referencia obligado para cualquier trabajo sobre identificación de citas en documentos electrónicos e incluso puede verse su funcionamiento en la url indicada. Aunque ha sido pensado para trabajar con materiales en el área de informática podría ser aplicado a cualquier disciplina. Se trata de una base de datos con más de 200.000 documentos indizados, todos accesibles gratuitamente en internet, de los cuales se han extraído más de 2 millones de referencias. *CiteSeer* no es exactamente un software para enlazar referencias sino que tiene un objetivo mucho más amplio. Es un verdadero índice de citas construido de forma automática, lo que los autores denominan un *autonomous citation index*.

<http://csindex.org>

Entre las muchas funciones que aporta este software se podrían destacar las siguientes⁹:

—Localización de documentos científicos en la web. Para ello actúa como un metabuscador utilizando múltiples buscadores como *Altavista* o *Google*, a los cuales envía palabras clave junto con términos específicos para limitar el ámbito de la consulta a documentos de carácter científico (por ejemplo: papers, conference, proceedings, etc.).

—Indización del texto completo de los documentos encontrados que estén en formato pdf, postscript o html.

—Extracción de información bibliográfica de los documentos. Incluye algoritmos y técnicas de aprendizaje para detectar automáticamente datos como título y autores de los documentos que indiza.

—Identificación de la lista de referencias. Establece la sección que contiene la bibliografía, siendo capaz de aislar cada una de las referencias. Además puede determinar si dos de ellas, con formatos diferentes, se refieren al mismo documento.

—Identificación de los distintos datos que contiene una referencia, entre otros: año de publicación, título y autores.

—Localización del contexto donde se ha producido la cita, es decir, la frase o frases que ha empleado el autor en el cuerpo del documento para referirse al aquel que cita.

—Puede determinar los documentos relacionados utilizando la información sobre citas disponible. Ante cada documento el sistema sugiere al usuario una serie de trabajos alternativos en función del número de referencias que tengan en común, de los documentos que lo han citado, etc.

—Es capaz de analizar las redes de literatura científica con objeto de determinar cuales son las autoridades en una materia, entendiendo por tales los documentos que reciben mayor número de citas. Igualmente es capaz de establecer las revisiones, que serían aquellos trabajos que contienen un elevado número de referencias.

—Puede identificar las autocitas, comparando el autor del documento con los de las referencias.

—Permite a los usuarios corregir errores en la base de datos, convirtiéndose así en un sistema interactivo.

—Tiene la posibilidad de mantener un perfil de los usuarios, de

tal forma que puede recomendar, bien por correo electrónico o a través de la propia web, nuevos documentos que se ajusten a dichos perfiles que pueden ser actualizados directamente por el usuario o bien por el propio sistema mediante técnicas de inteligencia artificial.

4. *OpCit (Open Citation Project)*

Es un proyecto financiado conjuntamente por el *Join Information Systems Committee* del Reino Unido y la *National Science Foundation* de EUA. Por parte inglesa participa *Intelligence, Agents, Multimedia Research Group* de la *Universidad de Southampton*, mientras el socio americano es el *Cornell Digital Libraries Research Group* de la *Universidad de Cornell*.

1. *Cornell Digital Libraries Research Group*

El trabajo que desarrollan está descrito en varios informes^{1,2 y 3}, según los cuales han creado una arquitectura para el enlace de referencias que se compone de dos niveles: el de enlace de referencias sobre la web, que se encarga de proporcionar suficientes datos para una variedad de servicios de valor añadido, o como ellos los definen aplicaciones de enlaces de referencias. Un ejemplo es la creación de interfaces de usuario para navegar por la red de referencias; dada una en concreto el nivel de enlace debe:

—Buscar su correspondiente contexto en el documento.—Realizar un análisis sobre ella para determinar que obra es, si el documento que representa se encuentra en formato electrónico, si es posible establecer un enlace con él, etc.

—Proporcionar acceso a todos estos datos para que puedan ser usados por las aplicaciones.

Por su parte las aplicaciones de enlaces de referencias deben:

—Convertirla, dentro del documento, en un enlace que apunte al texto completo. Por ejemplo en html o pdf, transformándola en un ancla hipertextual.

El trabajo del grupo se ha centrado hasta ahora en el primer apartado, lo que serían las fases de análisis, acceso y distribución de los datos. La aplicación práctica de esta base teórica la han llevado a cabo con la revista *D-lib*. Su objetivo era alcanzar un 80% de precisión en el proceso de extracción de información. Éste es el mínimo aceptable para que se puedan crear servicios de valor añadido en el nivel de enlace. Para evaluar los resultados han creado dos indicadores en función de los dos tipos de errores posibles. El primero se produce en la extracción de la información bibliográfica del documento que está siendo analizado. Está representado por el *item accuracy*: número de elementos analizados correctamente, dividido por la cantidad total de elementos en el *item*. Entre los que se intentan identificar están: el título del documento, cada uno de los autores, el año de publicación y los contextos de las referencias.

El segundo son los errores en el análisis de las cadenas de referencias que contiene el documento y está representado por el *reference accuracy*: el porcentaje de los elementos de una referencia que son analizados correctamente y los cuales incluyen: título, cada autor, año, contextos y url si existe. Los resultados que han obtenido para estos indicadores se acercan bastante a los objetivos previstos, con un 75% en el caso del *item accuracy* y un 70% de *reference accuracy* sobre un conjunto representativo de referencias.

2. Intelligence, Agents, Multimedia Research Group.

El socio inglés de *OpCit* está trabajando para establecer enlaces entre los documentos disponibles

en el archivo *arXiv* (el archivo de prepublicaciones más antiguo que existe en internet). Fue diseñado por **Paul Ginsparg** en *Los Alamos National Laboratory* (EUA) a comienzos de la década de los 90. En la actualidad almacena casi la mitad de la literatura que se genera en física de altas energías. Contiene más de 150.000 documentos a texto completo que pueden ser descargados gratuitamente desde cualquiera de los más de 15 mirrors que existen por todo el mundo. Una idea de su importancia la da el hecho de que más de 35.000 personas lo consultan diariamente.

«El Cern document server anunció el 1 de noviembre de 2001 que han llevado a cabo el enlace de todos los documentos que distribuyen»

El trabajo consiste en enlazar internamente las referencias de los documentos depositados en *arXiv*. En este caso, disponer de los originales da la posibilidad de manipularlos para insertar determinada información que luego servirá para crear vínculos. Es decir, pueden llevarse a cabo todos los niveles de la arquitectura descrita en el punto anterior. En estos momentos ya existe una versión interconectada de todo el archivo y se ha diseñado un interfaz que permite navegar por los documentos usando las referencias.

http://arabica.ecs.soton.ac.uk/cgi-bin/search_tj

Una vez interconectado todo el archivo, *OpCit* cuenta con un servicio duplicado de *arXiv* con nuevas facilidades, aunque está siendo muy poco utilizado por los usuarios, que prefieren acudir al sitio tradicional. El reto que se plantea ahora es devolver la información generada por *OpCit* para que pueda ser integrada a su vez en *arXiv*.

Para ello se ha pensado utilizar las facilidades que aporta la *OAI* y un nuevo formato para la descripción bibliográfica de documentos científicos llamado *AMF*, descrito en **Krichel y Warner**⁸.

Una vez que el proyecto dispone de una importante base de datos de citas, el siguiente paso que han emprendido es la realización de estudios bibliométricos sobre la disciplina. Están investigando la aplicación de indicadores como el factor de impacto a archivos de documentos y están analizando las prácticas de los usuarios al utilizar el sistema.

Finalmente el otro punto de interés es la aplicación de las citas para ordenar los resultados de las búsquedas realizadas sobre el servicio. Es decir, que ante una pregunta los documentos devueltos en primer lugar sean los que mayor número de veces hayan sido citados.

5. Cern

En la misma línea de trabajo, el *Cern document server* anunció el 1 de noviembre de 2001 que han llevado a cabo el enlace de todos los documentos que distribuyen, principalmente prepublicaciones en el campo de la física. En total supone más de 2 millones de vínculos desde sus documentos a revistas electrónicas u otras prepublicaciones. Una descripción del proyecto apareció en **Claivaz, Meur y Robinson**⁹. La versión completamente operativa del sistema puede verse en:

<http://weblib.cern.ch/>

Por el momento ofrecen información tanto de las referencias de un documento como de las citas que haya recibido. Todo ello en un sistema perfectamente integrado con lo que los usuarios venían utilizando hasta ahora. Al igual que *OpCit*, esta iniciativa se beneficia de la particular forma de construir las referencias que existe en física:

un formato muy estandarizado en el que es habitual indicar solamente los autores y la publicación donde ha aparecido el trabajo en formato abreviado.

El *Cern* almacena más de 170.000 documentos a texto completo y todos ellos están en pdf. Los autores pueden enviar los trabajos en cualquier formato incluyendo *Word*, *LaTeX*, etc., pues con posterioridad son convertidos a ficheros pdf (sobre los cuales se efectúa la extracción de los datos). Esto proporciona una gran uniformidad en el resultado, ya que todos los ficheros habrán sido creados con el mismo procedimiento. Cada documento se ha convertido a ascii y después se ha analizado para extraer las referencias. La tasa de efectividad de este proceso se sitúa en el 91%, incluyendo en los errores tanto documentos que no están en inglés como aquellos que no contienen una sección de referencias. De todos los analizados se han extraído casi 3 millones de referencias, el 80% de las cuales se han podido analizar correctamente, entendiéndose por tal que el siste-

ma ha sido capaz de identificar la revista donde ha sido publicado.

En 1.937.162 referencias se ha podido establecer un enlace entre la referencia y el documento electrónico al que representa. En este caso, se han hecho vínculos a las revistas disponibles en la biblioteca del centro o a otras prepublicaciones almacenadas en el propio servidor. Finalmente resaltar que se ha realizado una integración total de los datos extraídos en el registro bibliográfico del documento analizado elaborado por la biblioteca. Para ello se ha utilizado el formato marc: por cada referencia encontrada se ha añadido un campo 909 al correspondiente registro. La dirección electrónica, si existe, se ha colocado en el subcampo \$x. Esta es una característica que la hace diferente del resto de iniciativas vistas.

Notas

1. **Arms, W.; Bergmark, D.; Lagoze, C.** *An architecture for reference linking*. Technical report Cstr 2000-1820, Cornell Digital Library Research Group, 2000.

2. **Bergmark, D.** *Automatic extraction of reference linking information from online documents*. Technical report Cstr 2000-1821, Cornell Digital Library Research Group, 2000.

3. **Bergmark, D.; Lagoze, C.** *Reference linking the web's scholarly papers*. Technical report, Cornell Digital Library Research Group, 2000.

4. **Cameron, R. D.** "A universal citation database as a catalyst for reform in scholarly communication". En: *First monday*, 1997, v. 2, n. 4.

5. **Claivaz, J. B.; Meur, J. Y. L.; Robinson, N.** "From fulltext documents to structured citations: Cern's automated solution". En: *High energy physics libraries webzine*, 2001, n. 5. <http://library.cern.ch/HEPLW5/papers/2>

6. **De Sompel, H. V.; Lagoze, C.** "The Santa Fe convention of the open archives initiative". En: *D-lib magazine*, 2000, v. 6, n. 2.

7. **Keefer, A.** "Gestión de enlaces entre artículos electrónicos: el sistema Crossref". En: *El profesional de la información*, 2001, marzo, v. 10, n. 4, pp. 32-33.

8. **Krichel, T.; Warner, S.** "Design of a metadata framework to support scholarly communication". En: *International conference on Dublin core and metadata applications*, 2001. <http://openlib.org/home/krichel/papers/kanda.a4.pdf>

9. **Lawrence, S.; Bollacker, K.; Giles, C. L.** "Indexing and retrieval of scientific literature". En: *8th International conference on information and knowledge management, Cikm99*, 1999, pp. 139-146.

10. **García Testal, C.** "Digital object identifier". En: *El profesional de la información*, 2001, julio-agosto, v. 10, n. 7-8, pp. 26-31.

José Manuel Barrueco, *Universitat de València, Servei d'Informació Bibliogràfica.*

barrueco@uv.es