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

José Manuel Barrueco *

13 de diciembre de 2001

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 del trabajo de grupos de investigadores que ponen en común sus trabajos 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 documentos no son más que la presentación en un soporte de los hallazgos y descubrientos 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 documentos formado así la inmensa red que es la literatura científica.

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

De lo dicho se puede intuir la importancia de las listas de referencias. Autores como (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. Este autor describe un sistema en el cual 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.

*Universitat de València, Servei d'Informació Bibliogràfica barrueco@uv.es

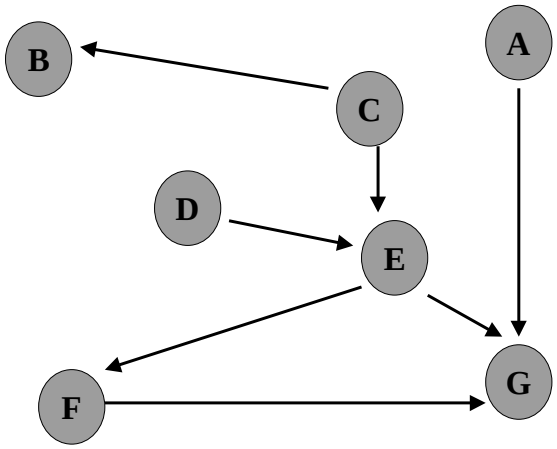

Figura 1: Enlaces de documentos a través de citas

Sin embargo su propuesta requiere que los autores o instituciones proporcionen información sobre las citas 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 propuesta 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 Internet. 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[1] en Física, CogPrints en Ciencias del Conocimiento o RePEc en Economía. Para estudiar la interopetabilidad de dichos archivos (6) se ha creado recientemente la Open Archives Initiative (OAi)[2].

Con la existencia de los documentos en formato electrónico ha surgido la posibilidad de tratar los textos automaticamente 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 cúal es el título, los autores, la fuente donde ha sido publicado y cúal es 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.

La posibilidad de enlazar una referencia con el texto completo del documento al que representa es lo que los anglosajones llaman **reference linking** o enlace de referencias.

Esta idea está siendo objeto de multiples estudios en la actualidad. El interés procede de dos areas claramente diferenciadas. En primer lugar de la indústria editorial y de servicios de resúmenes, quienes han visto el enorme potencial que representa el interconectar los miles de documentos que están poniendo en formato electrónico. El principal exponente de este área es CrossRef, proyecto que se describe más adelante. Por su carácter los resultados están limitados por innumerables barreras económicas. Lamentablemente sólo podrán tener acceso a ellos los investigadores de paises desarrollados. Por el contrario el segundo área de investigación procede de los archivos de documentos electrónicos accesibles gratuitamente en internet. Aquí tenemos iniciativas como OpCit, CiteSeer o CERN que son descritas más adelante. 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.

[1] `http://www.arXiv.org`
[2] `http://www.openarchives.org`

2 CrossRef

Las editoriales comerciales se dieron cuenta desde muy pronto de las ventajas que les aportaría el tener enlazados todos los documentos que publican. Como respuesta a esta necesidad surgió en 1999 el proyecto CrossRef[3] puesto en marcha por la Publishers International Linking Association (PILA) y descrito en (7). Esta es una empresa que engloba a más de ochenta editores y proveedores de servicios de resúmenes de todo el mundo.

El funcionamiento de CrossRef 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 requiere información básica como título de la revista donde se ha publicado, volumen, número, etc. Pueden aportar además otra información adicional según su elección. Además a cada artículo se le asigna un DOI[4] (Digital Object Identifier) (10) y un URL asociado para recuperar el texto completo de los documentos. El intercambio de información entre editoriales y CrossRef se realiza utilizando el formato XML. Es obligación de cada editorial mantener actualizada la correspondencia entre DOIs y URLs de cada documento.

La base de datos contiene en estos momentos casi cuatro 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 encará de devolver el DOI del documento citado si existe. Este código 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 del documento. 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.

El resultado de una interrogación a CrossRef solamente será un código que se resolverá en un url que apunta al texto del documento. Será competencia de cada editorial vigilar si el usuario que está intentando acceder al documento dispone de una subscripción 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. Igualmente no tiene nada que ver con los documentos en si mismos ya que lo único que almacena son sus metadados.

[3] `http://www.crossref.org`
[4] `http://www.doi.org`

3 CiteSeer

CiteSeer ha sido desarrollado en el departamento de investigación de la empresa 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.

Aunque ha sido pensado para trabajar con documentos en el área de informática podría ser aplicado a cualquier disciplina. Se puede ver **CiteSeer** en funcionamiento en la página `http://csindex.org`. Se trata de una base de datos con más de 200.000 documentos indizados, todos ellos accesibles gratuitamente en internet. De ellos se han extraido más de dos 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*.Entre las muchas funciones que aporta este software se podrían destacar las siguientes (9):

- Localización de documentos científicos en la web. Para ello actua 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 búsqueda a documentos de carácter científico (por ejem. 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 automaticamente datos como título y autores de los documentos que indiza.
- Identificación de la lista de referencias. Identifica la sección que contiene la bibliografía y es capaz de aislar cada una de las referencias. Además es capaz de determinar si dos referencias 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.
- Identificación 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 documento citado.
- Identificación de 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 identificar cuales son las autoridades en una materia, entendiendo por tales aquellos documentos que reciben mayor número de citas. Igualmente es capaz de determinar las revisiones, que serían aquellos trabajos que contienen un elevado número de referencias.
- Puede identificar las autocitas, comparando el autor del documento con los autores de las referencias.
- Permite a los usuarios corregir errores en la base de datos, convirtiendose 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 del propio web, nuevos documentos que se ajusten a dichos perfiles. Estos perfiles pueden ser actualizados directamente por el usuario o bien por el propio sistema mediante técnicas de inteligencia artificial.

4 OpCit: Open Citation Project

OpCit es un proyecto financiado conjuntamente por el Join Information Systems Committee del Reino Unido y la National Science Foundation de Estados Unidos. Por parte inglesa participa el 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.

4.1 Cornell Digital Libraries Research Group

El trabajo desarrollado por Cornell está descrito en informes como (2; 3; 1). Según éstos han desarrollado una arquitectura para el enlace de referencias que se compone de dos niveles: el *nivel 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 de tales aplicaciones es la creación de interfaces de usuario para navegar por la red de referencias.

Dada una determinada referencia, el nivel de enlace debe:

1. Buscar su correspondiente contexto en el documento.

2. Hacer un análisis de la referencia para determinar qué obra es, si el documento que representa se encuentra en formato electrónico, si es posible establecer un enlace con él, etc.

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

Por su parte las aplicaciones de enlaces de referencias deben:

1. Convertir la referencia en el texto del documento en un enlace que apunte al texto completo. Por ejemplo en HTML o PDF convirtiendola en un ancla hipertextual

El trabajo del grupo se ha centrado hasta ahora en el primer apartado, lo que sería 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**[5].

Su objetivo era alcanzar un 80% de precisión en el proceso de extracción de información. Este es el mínimo aceptable para que se puedan crear servicios de valor añadido en el nivel de enlaces. Para evaluar los resultados han creado dos indicadores en función de los dos tipos de errores posibles. El primero es un error en la extracción de la información bibliográfica del documento que está siendo analizado. Está representado por el **item accuracy** que es el número de elementos analizados correctamente, dividido por el número total de elementos en el item. Entre los elementos 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. Está representado por el **reference accuracy**. Es el porcentaje de los elementos de una referencia que son analizados correctamente. Estos elementos 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.

4.2 Intelligence, Agents, Multimedia Research Group

El socio inglés de OpCit está trabajando en establecer enlaces entre los documentos disponibles en el archivo **arXiv**. ArXiv es el archivo de prepublicaciones más antigüo que existe en internet. Fue diseñado por Paul Ginsparg en Los Alamos National Laboratory (USA) a comienzos de los años 90. En la

[5] `http://www.dlib.org`

actualidad almacena casi la mitad de la literatura que se genera en Física de Altas Energias. 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 la importancia del archivo la da el hecho de que más de 35.000 personas lo consultan diariamente.

El trabajo consiste en enlazar internamente las referencias de los documentos depositados en arXiv. En este caso, el disponer de los documentos originales, da la posibilidad de manipularlos para insertar determinada información que luego servirá para crear enlaces. 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. Puede verse en `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. No obstante está siendo muy poco utilizado por los usuarios, quienes 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 (Open Archives Initiative)[6] y un nuevo formato para la descripción bibliográfica de documentos científicos llamado AMF, descrito en (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 estudiando las prácticas de los usuarios al utilizar el sistema.

Finalmente el otro punto que están investigado 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 fueran los que mayor número de veces hayan sido citados.

5 CERN

En la misma línea de trabajo el CERN Document Server ha anunciado 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 ello supone más de dos millones de enlaces desde sus documentos a

[6] `http://www.openarchives.org`

revistas electrónicas u otras prepublicaciones. Una descripción del proyecto apareció en (5).

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 éste haya recibido. Todo ello en un sistema perfectamente integrado con lo que los usuarios venian utilizando hasta ahora.

Al igual que OpCit, esta iniciativa se beneficia de la particular forma de construir las referencias que existe en Física. Es un formato muy estandarizado en el que es habitual referenciar 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. Todos ellos están en formato PDF. Los autores pueden enviar los trabajos en cualquier formato incluyendo MS Word, LaTeX, etc. Luego son convertidos a PDF. Esto proporciona una gran uniformidad en el resultado, pues todos los ficheros habrán sido creados con el mismo procedimiento. La extracción de los datos se ha realizado sobre estos ficheros en PDF. Se ha convertido cada documento a formato ASCII y después se ha analizado para extraer las referencias. La tasa de efectividad de este proceso se situa en el 91%, incluyendo en los errores tanto documentos que no están en inglés como aquellos que no contienen una sección de referencias.

De los documentos analizados se han extraido casi tres millones de referencias, el 80% de las cuales se han podido analizar correctamente. Por análisis correcto se entiende que el sistema ha sido capaz de identificar la revista donde ha sido publicado el documento. 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 enlaces a aquellas revistas disponibles en la biblioteca del CERN o a otras prepublicaciones almacenadas en el propio servidor.

Finalmente resaltar que se ha realizado una integración total de los datos extraidos 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 MARC. 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.

Referencias

[1] Arms, W., Bergmark, D., and Lagoze, C. An architecture for reference linking. Tech. Rep. CSTR 2000-1820, Cornell Digital Library Research Group, 2000.

[2] Bergmark, D. Automatic extraction of reference linking information from onlin documents. Tech. Rep. CSTR 2000-1821, Cornell Digital Library Research Group, 2000.

[3] Bergmark, D., and Lagoze, C. Reference linking the web's scholarly papers. Tech. rep., Cornell Digital Library Research Group, 2000.

[4] Cameron, R. D. A universal citation database as a catalyst for reform in scholarly communication. *First Monday 2*, 4 (1997).

[5] Claivaz, J.-B., Meur, J.-Y. L., and Robinson, N. From fulltext documents to structured citations: Cern's automated solution. *High Energy Physics Libraries Webzine*, 5 (2001). http://library.cern.ch/HEPLW/5/papers/2.

[6] de Sompel, H. V., and Lagoze, C. The santa fe convention of the open archives initiative. *D-Lib Magazine 6*, 2 (2000).

[7] Keefer, A. Gestión de enlaces entre artículos electrónicos: el sistema crossref. *El Profesional de la Información 10*, 4 (2001), 32–33.

[8] Krichel, T., and Warner, S. Design of a metadata framework to support scholarly communication. In *International Conference on Dublin Core and Metadata Applications* (October 24-26 2001). http://openlib.org/home/krichel/papers/kanda.a4.pdf.

[9] Lawrence, S., Bollacker, K., and Giles, C. L. Indexing and retrieval of scientific literature. In *Eighth International Conference on Information and Knowledge Management, CIKM99* (1999), pp. 139–146.

[10] Testal, C. G. Digital object identifier. *El Profesional de la Información 10*, 7-8 (2001), 26–31.