CrossRef: the missing link

Ed Pentz

Paper presented at the 24th UKSG Annual Conference, Heriot-Watt, April 2001

References are a core feature of online journals, and links across titles and publishers' sites are essential. CrossRef provides the linking infrastructure by collecting metadata in a standard format and maintaining a DOI Directory for articles. Issues currently being addressed include unsubscribed content, multiple URLs, archiving and expansion beyond journal articles.

Ed Pentz
Executive Director
CrossRef
Burlington, MA USA
e-mail: epentz@crossref.org

Introduction

CrossRef is a collaborative linking service that enables reference links in scholarly journal articles. More precisely, CrossRef enables persistent links to full text scholarly journal articles. CrossRef, just over a year old, is a non-profit membership organisation made up of primary scholarly publishers that sets standards and rules and runs a system to make linking efficient and manageable. The main focus for CrossRef over the last year has been on getting primary publishers to join the organization, with increased effort to inform the scientific and library community about CrossRef since they will benefit from it.

CrossRef develops and manages a linking infrastructure and is committed to the long-term development and management of that infrastructure. Therefore it is very important to have ongoing support and development of the system. Many other initiatives, like eBioSci and Open Archives (http://www.openarchives.org/), are similar to CrossRef, in that they are collaborative ventures developing infrastructure to enable services to be created for scholars and to improve scholarly communication in general.

The CrossRef system

The key components of the CrossRef linking infrastructure are persistent article identifiers, standardized metadata and a resolution system to get from the identifiers to the content itself. For article identifiers, CrossRef uses DOIs, and is the first full-blown practical application of the Digital Object Identifier (DOI) system. The International DOI Foundation (IDF) was set up in 1998 and has laid out a general framework for DOIs and metadata. As the first IDF Registration Agency, CrossRef deposits DOIs, assigned by the publishers, together with the URLs to

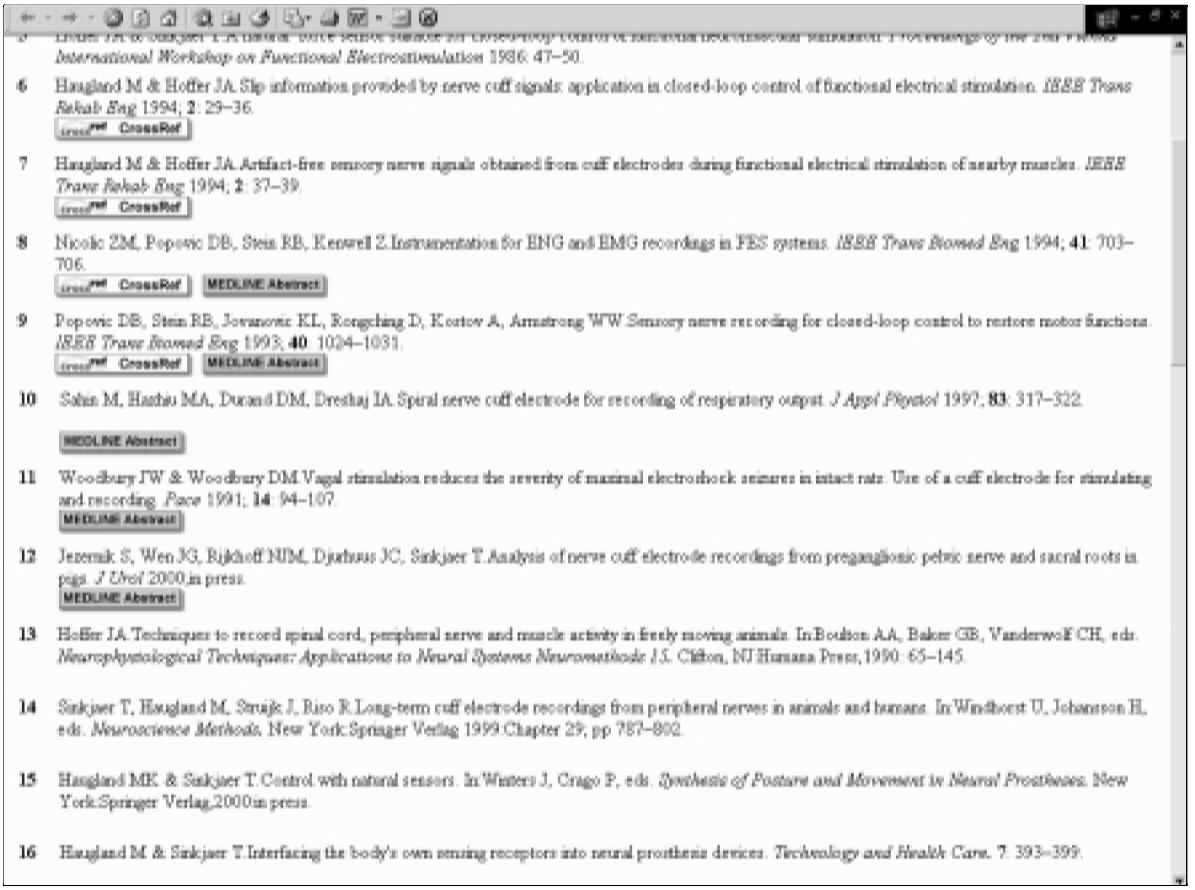

Figure 1

which the DOIs point, into the DOI system on behalf of member publishers.

Another part of the CrossRef service is a repository of article metadata. This is effectively a directory of article identifiers. The metadata is deposited in CrossRef in a standard XML format. Publishers deposit just the bibliographic metadata for an article, not the article or even the abstract. The mandatory information for publishers to send to CrossRef is journal title, volume, issue, article title, page, first author, year, DOI and URL. The DOI and URL are sent to the DOI Directory.

The final piece of the system is the Reference Resolver, which enables publishers to submit bibliographic data to look up the DOIs. Looking up a DOI is comparable to using the telephone directory to lookup a telephone number. If you want to find someone's telephone number you go to the telephone directory and, using some metadata about the person, their name and their address, you get their telephone number.

Once a DOI is obtained for an article, it is placed in a URL (http://dx.doi.org/10.1006/jmbi.2000.4282). The DOI is sent to the DOI Directory, which automatically redirects the user to the URL registered by the publisher.

Reference Linking

A key benefit of online journals for scholars is reference linking. Since references are how authors make explicit the links between their work and prior scholarship, references are a core feature of online journals. Readers expect references to be linked. Reference links are not a nice add-on like online commentary or multimedia content, they are essential and journals that do not have them will be seen as less valuable. Publishers have to add reference links to their journals, and the links will only really benefit scholars, if they work across publications and across various publisher sites.

Figure 1 shows a section of the references from Sinkjaer, Thomas. "Integrating Sensory Nerve Signals Into Neural Prosthesis Devices". *Neuromodulation 3* (1), 34-41 published by Blackwell Science. The article's DOI is 10.1046/j.1525-1403.2000.00035.x. Being a member of CrossRef, Blackwell Science just sends the

bibliographic metadata in the reference to CrossRef and a DOI is returned. In this case there are DOIs for references 6-9. Blackwell Science creates a link using the DOI and a user clicking the link is automatically redirected to the cited publisher's site. Before CrossRef, the process of linking to other publishers was more laborious.

The first step was to figure out the publisher of the journal based on the journal abbreviation. Next, the linking publishers would have to know the format for creating links to all the different publishers cited in the references and would need to determine whether or not the article was available online. A publisher would not want to link to articles not available online. In order to do this, every publisher would have to know and track any changes in the linking format, for all the publishers that they might link to and they would also have to have online holdings information for all the publishers that they would want to link to. As a result of these complications, primary publishers were not linking directly to one another. Large secondary publishers and large primary publishers were signing bilateral linking agreements. These agreements made sure there were no surprises and allowed for organizations to exchange data about holdings and linking formats. If just a few publishers are linking, bilateral agreements are not a problem, but with many publishers, bilateral agreements are scalable. Signing the linking agreements can be time consuming and a drain on resources, especially for smaller publishers. By taking away the need for bilateral linking agreements and providing a central location to lookup unique article identifiers, CrossRef makes broad-based linking efficient and manageable for publishers.

When a reader links to the publisher's site, the publisher's system handles the access control. In most cases publishers check the user's IP address to see if they have access to full text. In most cases users arrive at the abstract page for an article, but the publisher determines access to the full text and abstract. The minimum requirement is that publishers show a full bibliographic citation for the article for all users and provide information on acquiring the article or subscribing to the journal. CrossRef is business model neutral because the full text can be limited to subscribers, available on a pay-per-view basis, or it can be free. It is up to publishers to set the access terms.

CrossRef is also unusual because publishers are not known for co-operating, but the system was up and running very quickly. Linking allows publishers to add value and, since linking is a quid pro quo (I link to you and you link to me), the links add value to the source journal and the target journal. All journals get increased traffic and smaller publishers can easily be part of a linking network with the largest publishers. End users benefit because they have links that span journals and publishers.

CrossRef maximises the power of the web. The web is all about distributed content and CrossRef is acting in a distributed way. The content stays on the publisher's site and a minimal amount of data is stored on central servers to enable the linking.

Adding value

A key issue for information providers and publishers is how to add value. In the Internet world the barriers are lower, so if you do not add value, someone else will. In scholarly publishing there are e-prints, author self-archiving and initiatives like the Public Library of Science, which is calling for all content to be freely available. It is useful to look at the music industry to see parallels.

As with scholarly journals, the music industry is struggling with issues around online digital content and how to add value. Napster has been in the news recently because they quickly got about 50 million users for their service that allows the sharing of digital music files. Napster makes digital files available through a distributed peer-to-peer system and it was convenient. However, Napster was free for users and was found to be violating the copyright of the records labels since they did not have permission to use the files and was not charging any fees or paying any royalties to the record companies. Even those who agree that Napster was violating copyright also agree that the music industry was missing a great opportunity to create new economic models. The main issue is not technology; it is business models.

Recently, a number of joint ventures have been announced between record labels to start online

music services. One example is that AOL Time Warner, REAL Networks, Bertelsmann and EMI are setting up a subscription service called MusicNet (http://www.musicnet.com/). These large record companies are having to collaborate with their competitors to set up a service that will offer users a broad range of digital music in a convenient online service. This is very similar to what happened with CrossRef – the publishers had to get together and collaborate to make reference linking a reality.

It is important to note that Napster is not that convenient and one of the main problems is that there are no identifiers or standardised metadata. The central Napster servers just read the filenames that users have on their computers and these names can be wrong. This makes it very difficult to search for content. When the court ordered Napster to filter copyrighted material, users just changed the names of the files slightly to avoid the filtering, although this made it even more difficult for users to find what they were looking for. What the music industry will be grappling with, over the next year, will be identifiers and metadata because they are crucial components of any online distribution service. The scholarly publishing industry is slightly ahead of the music industry on this issue.

CrossRef: current status

The members of CrossRef are primary publishers of scholarly material, but there is also the category of Affiliates. Any organisation creating links to full text articles, meaning secondary databases, A&I databases and even libraries, can use CrossRef. The initial focus was getting publishers to sign up but we now are trying to get other organisations to use and participate in CrossRef, because the more links there are, the better. There are currently 71 member publishers, of which 60% are non-profit and they cover all areas of scholarly publishing. There are large commercial publishers and small society publishers from all over the world. CrossRef has signed up organisations like Cambridge Scientific Abstracts and EBSCO Publishing as affiliates, who will use CrossRef to lookup DOIs and create links to full text articles at publishers' sites.

CrossRef has metadata deposited for three million articles from about 3800 journals. The system went live in June 2000 and there are now thousands of journals with links. Between 500,000 and 1 million new articles per year are to be added on an on-going basis. Right now, to the end user, CrossRef has not made a large impact. The links, for a variety of reasons, are going to take some time to build up. First of all, publishers tend to deposit their content in the system and that has only really just got going in a big way. Then they wait a few months, while they do work on their own systems to actually add the reference links at a later date. So, whilst about 50 publishers are actually actively depositing content, only about 15 publishers are adding links to their references. It also varies by subject area, but somewhere in the order of 60% of the references in journals are not available online, so there will only be links to them when the content is digitised. The other thing is that the journals themselves vary in quality, so that if a publisher doesn't do copy editing on their journal, there are quite a lot of mistakes in the references and the page number might be wrong, or the volume number might be wrong. A project has just been started to develop new software and one of the key components of the software is going to be fuzzy matching. This means that, if somebody submits a reference and one of the pieces of information is wrong, the system would be able to accommodate that and still make a link. However, there are approximately 400,000 DOIs clicked each month by readers of online journals, so CrossRef has already started to make an impact on scholarly journals.

CrossRef: key developments

A big issue is access to "non-subscribed" content. A user may arrive at the publisher's site and not be able to get the full text if they are not a subscriber to the journal. There will be demand to make getting to full text easier and take into account user's institutional context. Asking users to enter credit cards at each publisher's site to purchase articles is not a very good solution. CrossRef, publishers and librarians will have to work on solutions to this problem (see Localized Linking below).

Another key issue is archive repositories. Organisations like JSTOR (http://www.jstor.org/) and the Astrophysics

Data System (http://adswww.harvard.edu/) have a lot of older digital content. DOIs should be assigned to the articles in those systems – in fact that has already started to happen. The oldest article in the CrossRef system is from 1849 from the Astronomical Journal. Someone could cite an article from 1849 and the user looking at the reference would be able to click on it and go to the full text article. Being able to assign DOIs to an article and get traffic coming to the article is a strong incentive for publishers to digitise their content and many publishers have ambitious plans to digitise their content. The American Physical Society will be digitising their content back to volume number one for all their journals

Multiple Resolution is another feature being developed. Currently, one DOI routes to one URL registered by the publisher. However, articles can exist in many different places. The DOI system is being expanded so that a single DOI can have many URLs associated with it. The idea is that, when a user clicks on a DOI link, a choice of links is displayed for the user. It would be nice to have some kind of automated system to take the user to the appropriate site for that article, but standard web browsers are not able to handle this at the moment. CrossRef will be working with the IDF on implementing this functionality for DOIs assigned to journal articles.

Localised linking

A key issue for libraries is localised linking, also called the appropriate copy problem. This refers to the fact that articles can be available through many different sources. A user may have a local electronic copy, there may be a print copy in the library, the publisher may have a copy on their server, or it could be available through an aggregator like Ovid. The question is how do you create a system so that you can show the user all the various options that are available for an article? CrossRef is working with the Digital Library Federation, CNRI, the IDF, Ex Libris, University of Illinois, Ohio State University and Los Alamos National Laboratory on a prototype that shows a DOI link actually used in conjunction with SFX or another local server. The local linking server presents options to the user based on their affiliation and what their institution licences. The prototype has gone very well. The prototype is working with OpenURL, a proposed NISO standard that is a protocol for transmitting metadata in a URL, and localised linking is also referred to as the OpenURL Framework. [For more information see Van de Sompel, Herbert, and Beit-Arie, Oren. 2001. Open Linking in the Scholarly Information Environment Using the OpenURL Framework. *D-Lib Magazine*. http://www.dlib.org/dlib/march01/vandesompel/03vandesompel.html]. CrossRef is working together with open URL, SFX and other services and is not in competition with them.

Another key development for CrossRef is expanding beyond journal articles. Conference proceedings and reference works will be added to the CrossRef system. Other things like patents and chemical information could also be included. Just as linking across publishers and journals is desirable, references should also link to any content that they cite.

Conclusion

CrossRef is a collaborative linking network. It takes away the need for these bi-lateral linking agreements and we have the infrastructure in place to build on for the future. It is also very important to note that CrossRef has no direct interaction with end users. End users are finding the links and clicking on the links, but they are not directly interacting with CrossRef. Standards are critical to the whole process and to many online systems. CrossRef is only part of the solution to effective linking and we are planning on being interoperable with many different systems. We are only at the beginning of the process of developing content that takes full advantage of the online environment – reference linking is just the first step.