

CrossRef: From cross-publisher reference linking to cross-publisher plagiarism screening in eight short years

by Amy Brand

CrossRef is a not-for-profit publisher association whose general purpose is to promote the cooperative development of innovative technologies for scholarly publishing and research. In addition to operating the first collaborative reference linking service, the association recently announced a new service, CrossCheck, to aid editors in screening submitted manuscripts for plagiarism. How did this initiative come about? Where does it fit within CrossRef's broader mission? What specifically is it intended to do and how does it work? These are the questions this article addresses.

CrossRef was founded in 2000 through the joint efforts of a small group of prestigious scientific, technical, and medical journal publishers. At the time, publishers moving their journals online wanted a way to cross-link journal articles while avoiding the common problem of broken links, or '404 page-not-found' errors. A new infrastructure for permanent identification of online information, the DOI (Digital Object Identifier) system, had just been introduced. Publishers who recognized the DOI system as their opportunity to create a cross-platform network joined together in a non-profit, independent association; CrossRef went live as the first collaborative citation-linking network in June 2000.

CrossRef's mission, formally put, is "to enable easy identification and use of trustworthy electronic content by promoting the cooperative development and application of a sustainable infrastructure." In the eight years since its inauguration, CrossRef has evolved along several dimensions. It interlinks the publications of thousands of information providers and offers a variety of services. It registers 16,000 diverse content items each day, seven days per week. Its members include not only traditional academic publishers and societies, but also institutional repositories that house dissertations, working papers, and datasets; government offices that output technical reports; and web-based reference environments with dynamically aggregated pages. CrossRef has become the first place publishers turn when they want to work together on new initiatives to harness technology for better information navigation and dissemination in the complex world of publishing content.

In 2006, the CrossRef membership identified plagiarism screening as a top priority for the academic publishing

community and decided to develop an aid for publishers to protect the integrity of the published record. In 2007, CrossRef conducted a pilot with six well-known publishers. The pilot's goal was to assess the feasibility of launching a self-sustaining plagiarism detection service with a business model that encouraged industry-wide participation through low barriers to entry. Specifically, we wanted to:

- understand the logistics and costs involved in creating and maintaining a database of our members' full-text content
- allow some of our members to experiment with the iThenticate user interface and to think about how they might use the system within their particular editorial environments
- understand what publishers are likely to need to do in order to integrate the plagiarism detection step into their existing manuscript tracking and editorial tools

The most important aspect of a plagiarism detection system is that its database contains a critical mass of relevant content; otherwise those checking manuscripts against the system will encounter an unacceptable number of false negatives. We therefore selected our pilot participants with the

A human being has to judge if intentional plagiarism has occurred or not

goal of getting a significant sample of content in two specific disciplines—computer science and biomedical research. Following completion of the pilot the CrossCheck service was officially launched on 19 June 2008.

CrossCheck is intended to help academic publishers verify the originality of works submitted for publication. CrossCheck has two parts, a database of scholarly publications and a web-based tool, iThenticate, to check authored works against that database. The result is a form of computer-assisted editing, in which the process of detecting textual overlap between documents—or otherwise verifying the originality of a document in the absence of any overlap—is largely automated. Clearly, the tool cannot, on its own, identify plagiarism. A human being has to examine areas of overlap in context and use judgment to determine if intentional plagiarism has occurred or not.

Screening tools like CrossCheck are only effective if they are checking texts against a relevant, comprehensive database. Although there are several plagiarism detection tools

CrossCheck helps academic publishers verify the originality of works submitted for publication

CrossRef: From cross-publisher reference linking to...

in use, they are not well-suited to filtering academic content simply because they have not had access to the relevant (often proprietary) full-text literature to screen against. CrossCheck has a continuously growing database of archival and current scholarly literature, text-fingerprinted for accurate document comparison.

As of June 2008, the CrossCheck database is already slated to cover over 20 million journal articles from the following publishers: Association for Computing Machinery, American Society of Neuroradiology, BMJ Publishing Group, Elsevier, Institute of Electrical & Electronics Engineers, International Union of Crystallography, Nature Publishing Group, Oxford University Press, Sage, Informa UK (Taylor & Francis), and Wiley Blackwell. With the launch, both publisher participation and the CrossCheck database are expected to grow quickly.

CrossRef's partner in this initiative, iParadigms, is a leading provider of web-based plagiarism detection services. Via CrossCheck, publishers can screen documents against billions of pages of open web content that iThenticate has crawled and indexed, in addition to the CrossCheck database itself. When a document is submitted to the service for checking, it does not become part of the database. Instead, the system creates a digital fingerprint of the document based on a special set of algorithms, and that fingerprint is run against the vast database of pre-indexed content. The output of this process is a 'matching report' that lists sources sharing a significant degree of textual overlap with the submitted text.

It is important to note that the service will only help editors identify cases of verbatim plagiarism, along with cases that may entail simple word substitution or sentence addition. The system cannot detect subtle forms of plagiarism like paraphrasing or idea plagiarism, and cannot detect copying of images and graphs, unless they also plagiarize significant textual elements such as captions. At the same time, the system can produce false positives when a portion of text has been legitimately duplicated; examples include boilerplate text, bibliographic references, and mathematical proofs.

CrossRef's current priority for CrossCheck is to recruit as much published scholarship into the database as possible. Even publishers who decide not to integrate screening into their editorial processes at this time are encouraged to allow their content to be indexed so that others can check against it. A 'CrossCheck Depositor' logo will be used by those contributing to the database, to help increase public awareness of the initiative; 'CrossCheck Deposited' tags will be placed on individual publications that have been indexed in the database to help deter future plagiarism.

It is too early to offer definitive advice to editors on where in the editorial process to add the plagiarism-checking step. For now, participating publishers are providing distributed access to the tool, so that internal and external editorial

staff can use it as they see fit. While some may opt for routine checking of every submitted article, others will only screen submissions that a reviewer or editor flags. With iThenticate's open Application Programmer's Interface (<http://en.wikipedia.org/wiki/API>), which allows publishers to integrate the system with their in-house tools, CrossCheck is currently being integrated with several manuscript tracking services, to better streamline editorial processes around the use of the tool. CrossRef is also

The database covers over 20 million journal article

working with leading community policy organizations to develop best practices to help publishers use CrossCheck effectively and ethically, and is planning a variety of research projects that will help the community better understand the issues and trends surrounding plagiarism.

In closing, CrossCheck is not just a plagiarism detection tool or a database, but rather a multi-pronged initiative to make plagiarism checking feasible for the academic publishing community. CrossCheck was created by publishers, for publishers, and its success will depend on the CrossRef membership joining in to allow their published content to play a part. Community interest in the initiative is high and the future looks promising for CrossCheck, yet another way publishers are working together through CrossRef to ensure the integrity of the published scholarly record.

Amy Brand

*CrossRef (operated by Publishers International Linking Association, Inc)
Lynnfield, Massachusetts, USA
crosscheck_info@crossref.org
www.crossref.org*

Vital signs

Dear TWS

You seem to have a certain theme in every issue. I wondered if this means that if I write an article for the journal it needs to be written around the specific theme or if I have to wait until my topic is relevant, or how it works?

Claire Gillow

Claire.Gillow@perceptive.com

Note from editor

Every issue of *TWS* has a theme but the articles in the issue do not all relate to the theme. Articles on any topic of interest to medical writers are always welcome.