The Cross-Publisher Originality Screening Initiative

Rachael Lammey
Product Manager, CrossRef
October/November 2014
I'm going to show you our CrossCheck service, but before I do that I'd like to give you a little bit of background on CrossRef - what CrossRef is, and why we are running an originality screening service. CrossRef was founded fourteen years ago to solve the problem of broken links. The web is all about links, but links break. This is annoying if you’re browsing the web and want to follow an interesting link, but in the context of scholarly publishing it becomes more than annoying - if you can’t follow a citation from one paper to another you’re being hampered in your research. Citation linking is one of the greatest benefits of online publishing, but it really does need to be reliable. ...and publishers were finding that web sites changed, content moved, and links that they had put into their articles stopped working. So they started a multi-publisher initiative to solve this problem of broken links.
This is done using the DOI - the Digital Object Identifier, which I’m sure many of you are familiar with. A CrossRef DOI is simply a unique identifier for a piece of content. Once assigned, it doesn’t change. It is to all intents and purposes a meaningless number, but it allows that piece of content to be located on the web.

10.1098/rstl.1665.0001
And it works like this: publishers use CrossRef DOIs to link to content, usually from the references at the end of articles. Users click on those DOI-based links and are referred via the CrossRef database to the cited article at its correct location on the web. If content moves the publisher only has to update the CrossRef database once, and all of the publishers that are linking to their content using CrossRef DOIs will be redirected to the content in its new location.
A few numbers for you to give some idea of how CrossRef has grown in the fourteen years since its launch...
Books are the fastest growing at the moment - most publishers have assigned DOIs to their journals and journal archives, but more and more are now starting to assign them to their books, and to register their book metadata with CrossRef. Publishers are also registering components - 274,000 so far.
Every month there are around 90 million clicks on CrossRef DOI links, so 90 million citations resolved to content.
And a little about us as an organisation.
* UK and Boston, MA
* Publishers of all sizes, subjects, and nationalities - you can see the list on our website if you're interested.
* One of the reasons we achieve so much with such a small staff is that we're lucky to be supported by a network of working groups and committees.

- 24 staff (7 UK, 17 USA)
- 16 person Board of Directors, cross-section of publishers
- Working groups and committees
So let’s look at how much of a problem plagiarism is in scholarly publishing. It’s certainly not a new problem, but is it one that’s getting worse? It has never been easier to search across vast amounts of content in online publications and databases. And with more content being produced than ever before it’s much harder for reviewers to have thoroughly read everything in their field.
And finally a study that has been reported and commented on in a number of places. A team led by Harold Garner and now based at the Virginia Bioinformatics Institute analysed the contents of Medline using a piece of text comparison software called eTBLAST. By comparing texts against each other computationally they identified pairs of articles with high levels of similarity.
Where the software threw up matches they went on to manually compare the documents - as you can see here the areas of this article that appear in an earlier article have been highlighted. And they have written up their findings in each case in the deja vu database.
You can access the database with their findings at the URL I've listed here. At the time that this article appeared in Science in May of 2009 year there had been at least 48 retractions of suspicious papers that were flagged to editors as a result of this project.
The team went on to contact the authors and journals involved and documented their responses - some of them were written up in Science in March 2009.

Deja Vu

- 238 pairs of articles with high similarity and no shared authors
- 1602 pairs with high similarity and one or more shared authors
- 588 pairs that eTBLAST flagged that are in fact legitimate

August 2009
"I fully endorse your proposal that publishers everywhere use text recognition software. This will be a useful at-source deterrent, as you point out... biomedical journals should at least have the standards widely used in today's high schools!"

Author of plagiarised article found by Garner et al. using eTBLAST

This is a quote from one of the authors whose work was plagiarised. READ SLOWLY.
He is referring to plagiarism detection software used in schools, which I'll come back to a little later.
So we started to develop CrossCheck. We ran a pilot towards the end of 2007 and the start of 2008 with 7 major publishers and a technology partner iParadigms, and on successful conclusion of the pilot we launched the live CrossCheck service in June last year. So it’s been running now for around 7 years.
The first thing that I always say when I talk about CrossCheck is that although we call it a plagiarism detection service, it doesn't actually detect plagiarism.
A machine cannot detect plagiarism. A machine can look at written text and tell you where it matches other written text, but it cannot tell you why that text matches, and this is critical. It takes a human being with a certain amount of domain expertise to analyse the results of any computer programme in order to determine the intent of the author or authors.
There are legitimate reasons why text might be the same in two documents. Here's a bibliography section which will almost-certainly be repeated in numerous places. A mathematical proof might be repeated in order to be extended, and so on... a human can spot some of these examples very easily whereas a piece of software cannot.
So let’s take a look at how CrossCheck actually works by starting with a simple overview. You have a manuscript or document that you want to check for originality. You submit it to the CrossCheck system, where it is broken down and analysed, then compared against a database of other scholarly publications. Wherever overlapping or similar passages of text are found, they are highlighted in a report for an editor to take a look at.
- Software that analyses and compares text
- Database of content to check text against
To effectively screen research material you need to compare it with other research material, and most of that is in publications that are on many different publisher platforms and often behind access control. So even if you find a match using Google Scholar you will still need to go to the publisher’s website to see the abstract, which may or may not contain the matching text. If it doesn’t, you need to get access to the full text, which may or may not involve paying, and so on and so forth.
This is where CrossRef is able to bring something new to a service such as this. We already have working relationships with over 4000 publishers and societies, and can circumvent the need for so many bilateral agreements to create a comprehensive database of content.
And this is precisely what we’ve done - we’ve facilitated the indexing of full text content from CrossRef publishers who join CrossCheck, and with iParadigms have put this content into a database to screen documents against. Just to talk about our technology partner for this project for a moment. Several technology providers were considered when we were planning CrossCheck. The one that we decided to work with is an organisation called iParadigms, who are based in Northern California. Their proven technology is probably best known as powering the Turnitin plagiarism screening tools for higher education. Turnitin is used widely in the UK and USA and I believe in several other European countries. Since 2003 it has also been available in the form of a commercial product called iThenticate. And it’s the iThenticate system that is made available to CrossCheck participants.
So to look at the process in a little more detail: you submit your manuscript to the iThenticate system, and it is by default checked against three databases of content. It is checked against web content - iThenticate indexes web pages in much the same way as a search engine, but with the added advantage that they keep an archive of web pages going back eight years. The manuscript is checked against the CrossCheck database, which contains the content from all of the participating CrossCheck publishers. And it’s also checked against a growing repository of online and offline content that iThenticate is gathering and indexing, including databases from Gale and Ebsco, and sites such as PubMed and Arxiv.org. And as before, matches retrieved by comparison with these databases are pulled into a report for an editor to examine in more detail.
## IThenticate Database Content

**STM Publishers:**
- American Chemical Society
- American Institute of Physics
- American Physical Society
- Elsevier
- IEEE
- Institute of Physics
- John Wiley & Sons
- Nature Publishing
- Ovid
- Oxford University Press
- Sage Publications
- Springer
- Taylor & Francis
- Wiley Blackwell

**Content Partners:**
- ABC-CLIO
- Language Learning
- EBSCOhost - 1.5m periodicals, biographies, yearbooks, encyclopedias, magazines, journals, books, and abstracts
- Eumed Journals
- Gale 90m articles
- Harcourt, McGraw-Hill and Wiley:
  - 2,000 academic textbooks
- ProQuest:
  - 360,000+ theses and dissertations
- PubMed/Medline:
  - 1.8m abstracts and citations: medical resources
- SAGE Reference:
  - 166v encyclopedias

**Internet:**
- 45 Billion

IThenticate’s proprietary internet crawler is comparable to major search engines. Archived back nearly a decade, IThenticate currently crawls 1.5m+ sites per day.

[http://www.ithenticate.com/content](http://www.ithenticate.com/content)
Why would something not be in the database?

- Not all publishers are CrossRef members
- Not all CrossRef members participate in CrossCheck: http://www.crossref.org/crosscheck_members.html
- Indexing needs to be enabled
- Timing – usually a few weeks between publication and indexing
- Needs a DOI!
This is the screen that you see when you’ve uploaded one or more manuscripts to iThenticate. You can see the article titles on the left, author and date processed on the right. The Report column with the square buttons beneath tells you what percentage of text within the manuscript has been found to match text in other documents.

Explain highlighting and thresholds.
And you get to this, which is the first of four different report manipulations available - this one is called the Similarity Report: Manuscript on left, matches on right from highest to lowest. Scroll up and down to compare. URLs (plus date) or citation depending on database. Links. Ability to exclude a match if you know it’s not relevant. Click on the left to see side by side report.

Point out print icon and filter and exclude items. Show link to Document Viewer and touch on report view.
Here you can see the two matching pieces of content side by side:
Glance feature, full source view
Importantly - entire article or piece of content on the right.
Scroll up and down and have opportunity to establish the context.
This is another benefit of using CrossCheck. We negotiated with
iThenticate to allow users to see the complete article where there
is matching text. If you use the commercial ithenticate product
directly you only see snippets of text either side of the match. But
we feel that it’s important for publishers and editors to be able to
see more than that.
You might have spotted in the previous examples that the technology isn’t just looking for word for word matches. The way that it breaks the text down allows it to spot passages of text with word substitutions, so it is looking for similar as well as identical text. In this example you can see that some of the words have been very subtly substituted or moved but the technology still picks them up.
Some additional features that you should be aware of: it’s possible to exclude certain things to help reduce background noise. You can opt to exclude anything that’s included in quotation marks. You can exclude the reference section, and you can choose not to be shown any matches below a certain number of words - so perhaps strings of fewer than 25 words would not be shown.

You should be aware that the first two of these features work on fairly hard and fast rules, so there need to be opening and closing quotation marks for a quote to be spotted and excluded, and the exclude bibliography feature relies on there being a recognisable section heading for
the references to identify it at the end of a document. So some documents will slip through these filters.
Limitations:
photos or images
graphs and tables
formulae
Text only
Accepted Formats

We currently support file upload for the following document types:

- Word
- Word XML
- Text
- PostScript
- PDF
- HTML
- WordPerfect WPO
- RTF

Also copy and paste
Three obvious places where you might want to do plagiarism screening
1) On submission  2) At some defined point in the review and editorial process - - and obviously this is a massive over-simplification not in the least because this process varies widely from publisher to publisher
3) Just prior to acceptance
And we have CrossCheck members taking each of these approaches - trend moving towards on submission.
615 publishers
Over 41 million content items indexed
  121,000+ titles
  170,000+ manuscripts checked each month

The progress of CrossCheck to date.
Very comprehensive database - can see list of titles on our website.
This is a partial list of member publishers - now that we're over 500 it's hard to fit them on one page and still make them readable but you can see the complete list on the CrossRef website so don't strain your eyes trying to read this one. Just to emphasise really that they are publishers of all sizes, of multiple nationalities and representing many disciplines.
And we are also building something of a CrossCheck community. We have a users email listserv that members are invited to use to ask questions or share experiences. In addition to the excellent support available from the iThenticate team, there are CrossRef staff such as myself available to help with setting up and running CC. And with the supporting CC committee we’re looking to create guidelines and best practices for use of CC, and template plagiarism policies for those publishers that may not have one of their own.

- Webinars
- CrossRef support
- Formation of guidelines, best practices etc.
- See: http://www.crossref.org/crosscheck_plagiarism_resources.html
- User group meetings
Advanced CrossCheck training from ME Support team

• Scheduled for November and December.

• 3 webinars:
  – Interpreting results and using the flowchart in the plagiarism policy (looking at examples).
  – Exploring different CrossCheck settings.
  – Types of CrossCheck reports.
Dealing with Plagiarism: Cochrane’s plagiarism policy

- Available in the Cochrane Editorial and Publishing Policy Resource
- Launched September 2014

www.cochrane.org/editorial-and-publishing-policy-resource/plagiarism

If you’ve got a case of suspected plagiarism on your hands, it’s important that it is dealt with sensitively and carefully. Cochrane offers a guide based on this. Cochrane has their own policy which is COPE-approved.
And flowcharts adapted from the COPE/Committee on Publication Ethics – a resource widely used in the publishing world. These walk you through the steps needed and the appropriate parties to contact at each point in the process to ensure the correct steps are followed.
And that’s exactly why the CrossCheck service has been created - to help publishers screen for originality before it gets published. Once plagiarised content is published and out there it becomes quite a messy problem to fix. With CrossCheck we’re providing tools that will hopefully mean any problems are found and sorted out much earlier on in the process. And ultimately it’s about maintaining the integrity of your publications, and adding value through the publishing process.
How to access CrossCheck

• All Cochrane Review Groups can receive a login to CrossCheck.
• Contact Gavin Stewart at Wiley (gstewart@cochrane.org).
Thank You
www.crossref.org/crosscheck.html

Remind about Cochrane follow-up training.