science editing

pISSN 2288-8063
eISSN 2288-7474

Sci Ed 2015;2(1):22-27
http://dx.doi.org/10.6087/kcse.32

Training Material

CrossRef text and data mining services

Rachael Lammey

CrossRef, Oxford, United Kingdom

Abstract

CrossRef is an association of scholarly publishers that develops shared infrastructure to support more effective scholarly communications. It is a registration agency for the digital object identifier (DOI), and has built additional services for CrossRef members around the DOI and the bibliographic metadata that publishers deposit in order to register DOIs for their publications. Among these services are CrossCheck, powered by iThenticate, which helps publishers screen for plagiarism in submitted manuscripts and FundRef, which gives publishers standard way to report funding sources for published scholarly research. To add to these services, CrossRef launched CrossRef text and data mining services in May 2014. This article will explain the thinking behind CrossRef launching this new service, what it offers to publishers and researchers alike, how publishers can participate in it, and the uptake of the service so far.

Keywords

Application programming interface; CrossRef; Digital object identifier; Metadata; Text and data mining

Introduction

Due to its position as a membership organisation for publishers, CrossRef has existing relationships with over 4,000 publishers and societies. These cover all subjects, all business models, and between all of these member publishers they have assigned nearly 70 million digital object identifiers (DOIs) to journal articles, books, conference proceedings and other types of content such as data. CrossRef does not hold the full-text of the content, but it does hold the bibliographic metadata for that content, and the links to the content on the publisher-maintained websites, which makes CrossRef well positioned to provide services that relate to text and data mining (TDM).

Over the past number of years, the issue of TDM has become very important and, because of the registry of unique identifiers and metadata for scholarly content that constitutes CrossRef's infrastructure, CrossRef is in a unique position to extend that infrastructure to make TDM easier for researchers and their institutions and publishers.

Received: November 7, 2014
Accepted: November 17, 2014

Correspondence to Rachael Lammey
rlammey@crossref.org

ORCID
Rachael Lammey
http://orcid.org/0000-0001-5800-1434

What is Text and Data Mining?

Before going into detail on the CrossRef TDM Services and the problem they are trying to solve, it is important to define what TDM consists of. To use a clear a description from PLOS blogs: Text mining is an interdisciplinary field combining techniques from linguistics, computer science and statistics to build tools that can efficiently retrieve and extract information from digital text [1].

In the same way as a person can read an article, in the field of text mining a computer program is reading the literature in order to find links or patterns within it. This may involve reading thousands of papers, which a computer program can do, but would take years for a researcher to do, and even if they could, they may never notice the links between the papers that a more automated process could pick up.

The text mining briefing paper from Joint Information Systems Committee gives a good example of text mining at work. It cites an example from the *Journal of the American Medical Informatics Association*, where a researcher called Marc Weeber used text mining tools to look at potential uses for the drug thalidomide. The example notes the efficacy of text mining tools in order to define a more refined and therefore useful corpus of content: "Type in thalidomide and you get 2,000 to 3,000 hits. Type in disease and you get 40,000 hits. With automated text mining tools we only had to read 100 to 200 abstracts and 20 or 30 full papers. We've created hypotheses for others to follow up," says Weeber et al. [2].

It's important to note that Weeber says that the work from his group has 'created hypothesis for others to follow up.' Some people have pointed to TDM as being a method that could provide new cures for many diseases, but it is not a magic bullet. The results resulting from this type of exploration still need to be analysed and built upon by researchers to test the hypotheses they raise. And of course the corpus of content used by the text mining tools needs to be the most comprehensive and best available.

Why has CrossRef Launched CrossRef Text and Data Mining Services?

CrossRef has launched this new service to try to help facilitate access to the relevant corpus of content for researchers who are interested in mining academic publications produced by CrossRef members. Currently, some issues exist for researchers trying to get the full text in order to mine it.

The first issue is that researchers find it impractical to negotiate multiple differing agreements with subscription-based publishers in order to get authorisation to text and data mine subscribed content i.e. content the researcher would already have access to via an institutional license or personal subscription. Because they may want to mine thousands of papers that are published by many different publishers, a researcher may need to contact many of these publishers to get access to the text which is a time-consuming, manual process.

From the publisher side, subscription-based publishers find it impractical to negotiate multiple bilateral agreements with researchers and institutions in order to authorise TDM of subscribed content. Again, they need to handle the transactions from researchers on a case-by-case basis, which is not an efficient process. As such, the CrossRef TDM services aim to give all parties standard application programming interfaces (APIs) and data representations that they can use to enable more automated TDM transactions across all publishers, regardless of their business model. The service is free for researchers to use.

Also, because it is a CrossRef service, it uses the DOI. This isn't an unnecessary layer of complication, but rather provides several benefits. It provides an easy way to de-duplicate documents that may be found on several sites, as processing the same document on multiple sites could skew text and data mining results. It also provides provenance information for the piece of content i.e., a researcher can see it comes from the publisher of the work who will maintain and steward it, and update the DOI to point to the content in its current location if it ever moves.

CrossRef Text and Data Mining Services: In Detail

The main aspect of CrossRef TDM services is the CrossRef TDM API. The API is designed to allow researchers to easily harvest full text documents from all participating publishers regardless of their business model (e.g., open access, subscription). It makes use of CrossRef DOI content negotiation, which will be explained later in this article, to provide researchers with links to the full text of content located on the publisher's site. As CrossRef does not hold the full-text, the publisher remains responsible for actually delivering the full text of the content requested. Thus, open access publishers can simply deliver the requested content to the researcher, while subscription based publishers can use their existing access control systems to give access to researchers with subscriptions access to the full text.

To explain what is meant by content negotiation, this feature allows a researcher to request a resource in their preferred format. DOI resolvers already use content negotiation to provide different representations of metadata associated with DOIs. A content negotiated request to a DOI resolver is much like a standard hypertext transfer protocol (HTTP) request, except server-driven negotiation will take place based on the

list of acceptable content types a client provides. So a researcher who prefers to work with the content in extensible markup language (XML), can use the API to request that the XML version of the content be returned to them by the publisher, or if they prefer portable document format (PDF) they can use the API to request that. However, this does depend on what formats of the content the publisher can provide. For example, some publishers only have XML for their more recent content, so may only be able to provide back content in PDF format.

As well as content negotiation, the API also supports rate limiting. Rate limiting is a method used to control the rate of traffic sent or received by a website. The API used with CrossRef TDM services employs a set of standard HTTP headers that can be used by servers to convey rate-limiting information to automated TDM tools. Text mining tools can look for these headers when they query publisher sites in order to understand how to adjust their behaviour so as not to affect the performance of the site. The headers allow a publisher to define a "rate limit window"—which is basically a time span (e.g., a minute, an hour, a day) in which they will return a certain number of full-text documents.

The CrossRef TDM HTTP headers are as follows:

- CR-TDM-rate-limit: 1,500 (the rate limit ceiling per window on requests)
- CR-TDM-rate-limit-remaining: 1,387 (number of requests left for the current window)
- CR-TDM-rate-limit-reset: 1,378,072,800 (the remaining time in to be replaced with Coordinated Universal Time epoch seconds before the rate limit resets and a new window is started)

Note that the values given are example values—each publisher can determine their own values based on the needs of their publishing platform, if they choose to use these headers—use of them is optional. This rate-limiting technique is already used by many APIs, including the Twitter API.

In order for researchers to use the CrossRef API, publishers need to add two new pieces of metadata to their CrossRef DOI deposits. They need to deposit a full-text link in the metadata for each DOI so researchers can follow it to get the full-text at the uniform resource identifier (URI) stated. They should also deposit a license URI in the metadata for each DOI so researchers can use this to find out if they have permission to mine the piece of content, and under what conditions they can do so. There is no charge for publishers to deposit this additional metadata with CrossRef.

The section of a CrossRef XML deposit containing these extra pieces of TDM information is shown below.

```
<crossmark>
<crossmark_policy>10.6087/ crossmark_policy</crossmark_policy>
<crossmark_domains>
<crossmark_domain>
<domain>escienceediting.org</domain>
</crossmark_domain>
</crossmark_domains>
<crossmark_domain_exclusive>true</crossmark_domain_exclusive>
<custom_metadata>
<assertion name="published" label="Published" group_name="publication_history" group_label="Publication History" order="0">2014-08-18</assertion>
<ai:program xmlns:ai="http://www.crossref.org/AccessIndicators.xsd" name="AccessIndicators">
<ai:license_ref applies_to="tdm" star_date="2014-08-18">http://creativecommons.org/licenses/by-nc/3.0/</ai:
license_ref>
</ai:program>
</custom_metadata>
</crossmark>
<doi_data>
<doi>10.6087/kcse.2014.1.51</doi>
<resource>
http://escienceediting.org/journal/view.php?doi=10.6087/kcse.2014.1.51
</resource>
<collection property="text-mining" setbyID="kcse">
<item>
<resource content_version="tdm">
http://www.esienceediting.org/upload/se-1-2-51.pdf
</resource>
```

Specifically, the following section relates to the license information for the article:

```
<ai:program xmlns:ai="http://www.crossref.org/AccessIndicators.xsd" name="AccessIndicators">
<ai:license_ref applies_to="tdm" start_date="2014-08-18">http://creativecommons.org/licenses/by-nc/3.0/</ai:
license_ref>
</ai:program>
```

This access indicators program means that publishers can provide a link to the license the article is published under, in the case of *Science Editing*, this is a Creative Commons CC-BY license which allows the content to be mined. By using the start_date information, a publisher can also represent embargo information—so if a paper is published under one license for a certain period of time, then changes to another, the start date can be used to show the date from which the new license will apply.

The following section relates to the full-text links that publishers should add to point the researcher to the full-text of

the content:

```
<collection property="text-mining" setbyID="kcse">
<item>
<resource content_version="tdm">
http://www.escienceediting.org/upload/se-1-2-51.pdf
</resource>
```

Note that by using the collection_property element, a publisher can define the specific purpose of the full-text link (in this case for TDM). CrossRef members may also deposit full text links for use by crawlers like Google, and the iParadigms crawler (for CrossCheck indexing). Because the API supports content negotiation, publishers can deposit full-text links to more than one version of the full-text, as shown in the example below:

```
</collection>
<collection property="text-mining">
<item>
<resource mime_type="application/pdf">
http://annalsofpsychoceramics.labs.crossref.org/fulltext/
10.5555/515151.pdf
</resource>
</item>
<item>
<resource mime_type="application/xml">
http://annalsofpsychoceramics.labs.crossref.org/fulltext/
10.5555/515151.xml
</resource>
</item>
</collection>
```

In this instance, the researcher can choose to use either the PDF or XML version of the article, and request either via content negotiation. More detailed information on formatting the XML relating to the license information is available on the CrossRef TDM support site [3].

If a publisher is an open access publisher or if they allow TDM as part of their standard subscription agreements, then that is all a publisher needs to do to enable their content for TDM via CrossRef. They can deposit those two additional pieces of metadata and optionally implement rate-limiting. They can then point researchers interested in mining their content with the CrossRef TDM services to the relevant page on the CrossRef TDM support site [4] so they can find information on how to use the API, and the commands they can use to call the full-text content. The entire workflow, showing

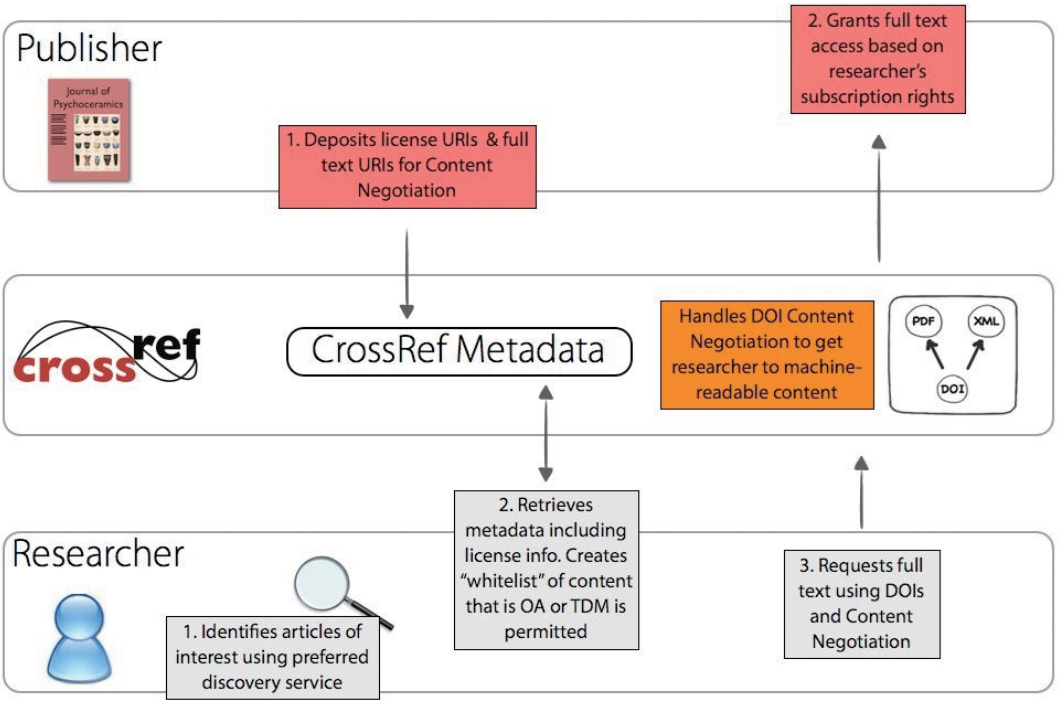

Fig. 1. The CrossRef TDM workflow. URI, uniform resource identifier; PDF, portable document format; XML, extensible markup language; DOI, digital object identifier; OA, open access; TDM, text and data mining.

how CrossRef, publishers and researchers combine to offer and make use of this service is shown in Fig. 1.

The Click-through Service

Publishers who require researchers to agree to a specific set of terms and conditions (T&Cs) before they are allowed to text and data mine content that they otherwise have access to (e.g., through an existing subscription) will need to make use of CrossRef's click-through service, which handles that additional transaction. The use of the click-through service extends the CrossRef TDM services workflow, and the additional aspects are shown in Fig. 2.

The Click-through Service for Publishers

The click-through service allows publishers to upload and manage click-through TDM agreements for their content. It also allows publishers to verify with the service that a researcher has accepted one or more relevant registered T&Cs (via an API token). Publishers can go to the service at: https://apps.crossref.org/clickthrough/publishers/#/login and log in using their existing CrossRef credentials.

Once logged-in, publishers can upload and mange their T&Cs. Every agreement registered must have a unique name, a short description of the T&Cs, a unique URI which points to a copy of these terms on the publisher's site and the full text of the T&Cs. When these are uploaded, the CrossRef member can 'publish' the agreement, thus making it available in the click-through service for researchers to be able to review these terms and decide whether to agree to them.

For the purposes of version control, once an agreement has been published and accepted by even one researcher, it can't be edited or deleted, i.e., the T&Cs cannot change after someone has already agreed them. However, they can be disabled, and a new version then published to reflect any updated terms for researchers.

The Click-through Service for Researchers

If a researcher is interested in mining content from a publisher who requires them to sign an additional license to use their content for mining purposes, then they can review these licenses in the click-through service.

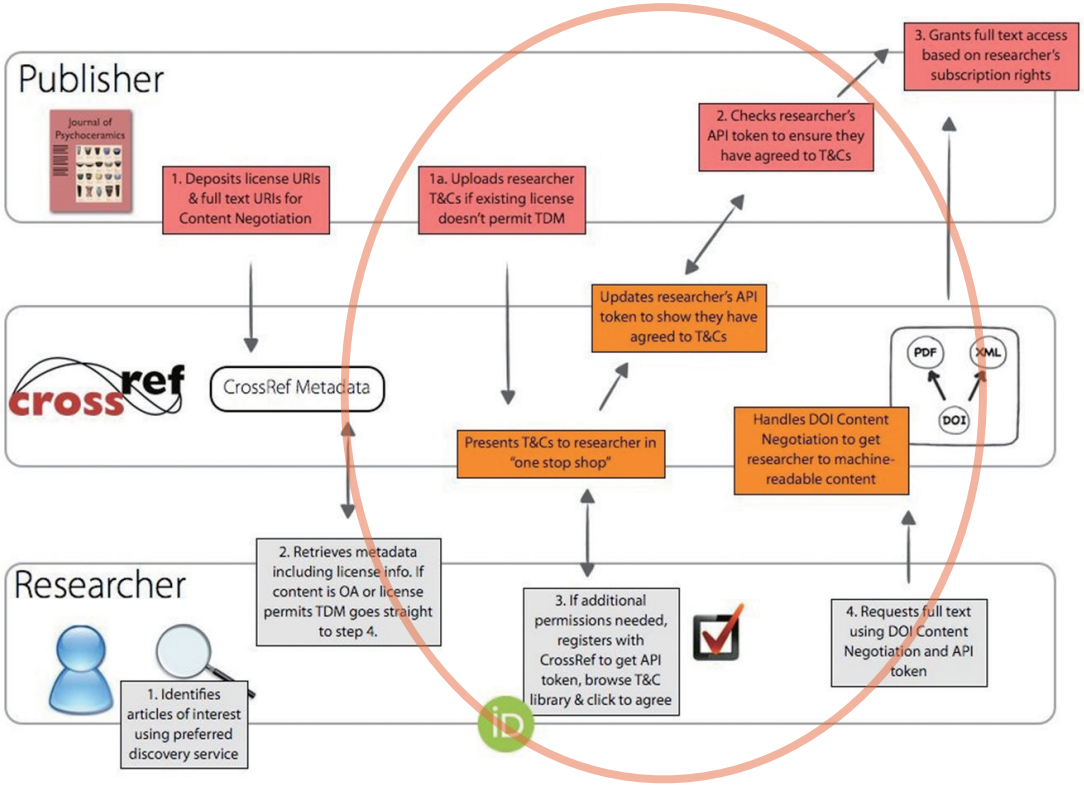

Fig. 2. The extended CrossRef TDM services workflow. URI, uniform resource identifier; T&Cs, terms and conditions; TDM, and data mining; DOI, digital object identifier; PDF, portable document format; XML, extensible markup language; API, application programming interface.

A researcher can log into the click-through service here: https://apps.crossref.org/clickthrough/researchers/#/login, using their Open Researcher and Contributor ID (ORCID) credentials. If they do not have an ORCID or profile, they will need to register for one here: https://orcid.org/register before using the service. The use of the ORCID means that the researcher can be disambiguated from other researchers with the same/similar names, they can use one set of log-in information for multiple services and it means that CrossRef is not storing this sensitive information.

Once registered, a researcher can click on the licenses that apply to the publishers they are interested in to view the specific T&Cs registered. They can then accept or reject the license, or choose to review it again later.

Once a researcher has accepted any license via the click-through service, they are issued with an API token. They can then use this API token in their TDM tools when they request the full-text of the article from the publisher to identify themselves to publishers. It is worth noting that a researcher can regenerate their personal token at any time if they feel that it has been compromised.

This researcher API token combines with a publisher API token that a publisher is given through their version of the click-through service. Using the publisher API token, the CrossRef member can check to see which licenses have been accepted by a particular researcher using an HTTP request. Examples of the request that a publisher can make are available on the click-through service support site [5]. CrossRef doesn't advise publishers to query the API with every single request they get, but rather to do so every certain number of requests, or once every set length of time to avoid overloading the API. The aim of the click-through service section of CrossRef TDM services is to help provide machine-to-machine automated access for recognised mining, by enabling an easy mechanism for the use of supplemental licences for TDM.

Progress of CrossRef Text and Data Mining Services

As of November 2014, CrossRef has seen 17 publishers sign up to CrossRef TDM services, and publishers can register their interest in participation or let CrossRef know when they expect to start depositing the TDM-specific metadata with CrossRef via a web-based contact form [6].

Some publishers have already started to deposit this metadata with CrossRef—Elsevier has populated over 11 million DOIs with the license information and full-text links necessary for CrossRef TDM services, and Hindawi has added this information to over 120,000 of their journal articles. As further publishers join the service in late 2014 and early 2015, CrossRef expects this number to grow quickly. There is no cost for publishers to participate in CrossRef TDM services in 2014, and a decision will soon be communicated to CrossRef members regarding any proposed charges for 2015. As mentioned earlier in this article, there is no charge to researchers for using the service.

Conclusion

This article has aimed to explain the thinking behind CrossRef's TDM services and the technical aspects involved in implementing the service for CrossRef members. It is hoped that this service will see healthy publisher participation and therefore become a useful resource for the TDM community, reducing the time and effort involved for all parties interested in supporting this process.

References

1. Bergman CM, Hunter LE, Rzhetsky A. Announcing the PLOS text mining collection[Internet]. [place unknown]: PLOS One Community Blog; 2013 [cited 2014 Nov 5]. Available from: http://blogs.plos.org/everyone/2013/04/17/announcing-the-plos-text-mining-collection/
2. Weeber M, Vos R, Klein H, De Jong-Van Den Berg LT, Aronson AR, Molema G. Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide. J Am Med Inform Assoc 2003;10:252-9. http://dx.doi.org/10.1197/jamia.M1158
3. CrossRef. Content mining with CrossRef text and data mining services [Internet]. Lynnfield: CrossRef [cited 2014 Nov 2]. Available from: http://tdmsupport.crossref.org/researchers/
4. CrossRef. Text and data mining for publishers [Internet]. Lynnfield: CrossRef [cited 2014 Nov 2]. Available from: http://tdmsupport.crossref.org/publishers/
5. CrossRef. CrossRef click-through service [Internet]. Lynnfield: CrossRef [cited 2014 Nov 6]. Available from: http://clickthroughsupport.crossref.org/publishers/
6. CrossRef. CrossRef text and data mining contact form [Internet]. Lynnfield: CrossRef [cited 2014 Nov 6]. Available from: http://www.crossref.org/tdm/contact_form.html/