Insights – 28(2), July 2015
CrossRef Text and Data Mining Services | Rachael Lammey

CrossRef Text and Data Mining Services

Based on a breakout session presented at the 38th UKSG Annual Conference, Glasgow, March 2015

The discipline of text and data mining (TDM) is growing in visibility in the scholarly research community. In May 2014, CrossRef launched CrossRef Text and Data Mining Services, which aims to alleviate the issues that text mining researchers encounter when trying to collect the corpus of content that they want to mine from the different publishers who have produced it. This article will discuss what the CrossRef service does, the problem that it is trying to solve and the current status of the rollout of the service across CrossRef member publishers.

RACHAEL LAMMEY

Product Manager
CrossRef

Introduction

CrossRef is a not-for-profit membership organization of international scholarly publishers, founded in 2000. CrossRef now has over 5,000 member publishers, representing all disciplines and comprising commercial publishers, academic societies, open access (OA) publishers and university presses. It also has over 80 affiliate members and 2,000 library affiliates who make use of the CrossRef database to look up digital object identifiers (DOIs) and metadata. There are other DOI registration agencies within the scholarly publishing sphere, but CrossRef is the largest DOI registration agency and has assigned over 70 million DOIs to date. The UKSG community has always been on familiar terms with CrossRef, but these numbers serve to emphasize CrossRef's current coverage.

CrossRef was conceived and set up by publishers to provide a service that would enable publishers to link to each other persistently in the age of online publishing. As publishers started to move online, they found that websites would change and content would move so that the links that they had put into their articles to link them to other articles stopped working. The DOI provides a way to prevent this phenomenon, known as 'link rot'. It works as follows: if the web address where a piece of content is hosted changes, a CrossRef member publisher can update the metadata stored with CrossRef to point the link to the content in its new location. Once they have done this, any other source linking to that piece of content via the DOI will get redirected to the content at its current URL.

Membership of CrossRef facilitates the technical aspects of this linking; however, the other element is one of industry collaboration. By joining CrossRef, a publisher is committing to linking their references to other publishers' content using the DOI, understanding that other publishers will do the same. This saves individual publishers having to make their own agreements with other publishers to link out to each other's content, and it also saves each of them having to come up with their own solutions to ensure the links to their content persist over time.

CrossRef Text and Data Mining Services

The fact that CrossRef has existing relationships with over 5,000 publishers is key to the thinking behind it providing a service to try to solve some of the logistical issues relating to text and data mining (TDM).

'a service to try to solve some of the logistical issues relating to TDM'

As noted in PLOS ONE Community blog[1], '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'. It is a way for a computer to try to read and try to find links within the vast volumes of scholarly literature that may not be found otherwise. These links may not be conclusive in themselves, but may generate enough information for researchers to follow up on.

Researchers have developed TDM tools to extract from and try to analyse the scholarly literature, but the issue they often face, and that CrossRef's service is trying to solve, is the one of actually obtaining the full-text content for the purposes of mining. This comes down to the large scale of the literature and the fact that it is often dispersed across a wide range of publishers.

As noted, CrossRef has thousands of members and a researcher may be interested in obtaining full-text content from a large proportion of them for the purposes of mining. It is impractical for that researcher to contact even 100 publishers to ask for the content to be delivered to them for mining. These individual requests also need to be handled by the publisher, and the content delivery arranged. If 100 researchers contact 100 publishers, this will result in 10,000 different transactions. Even on that small scale, it seems evident that some infrastructure is needed to facilitate this process.

Traditionally, there have been a number of ways for researchers to get the full text in bulk. One way is to work with the publisher to have the content delivered to them. This could be sent to them via file transfer protocol (FTP) or via physical media, i.e. on disc. This can cause issues if a researcher wants a feed of the most recent relevant content delivered to them on an ongoing basis, and it is also difficult to scale across multiple publishers and researchers. Some researchers will try to go directly to the publisher's website and screen-scrape the full text from it. Screen-scrapers can impact on the performance of a publisher's website, and if slight changes are made to the website, this can break the tool doing the screen-scraping.

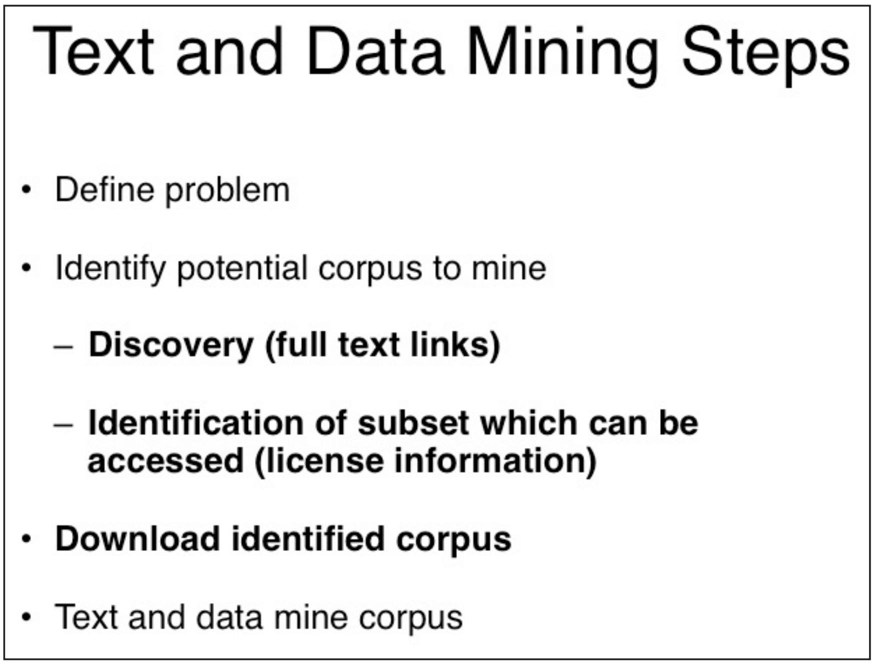
CrossRef Text and Data Mining Services aims to provide an alternative solution that has the capacity to work in a similar way across all publishers. This is done by providing a standard application programming interface (API) and standard data representations to facilitate TDM across a wide range of publishers. This approach will enable computer-to-computer requests for the full text, and can apply across all publisher business models. Even if content is open access, a researcher still needs to be able to easily pull back the full text over a wide range of publishers in as centralized a way as possible.

'an alternative solution that has the capacity to work in a similar way across all publishers'

The service is set up to do two things; firstly, it will give researchers a way to access the full text from the publisher site for OA or subscribed content. This is done via the CrossRef Text and Data Mining API, which gives a common protocol for requesting machine-readable full text from publishers who participate in the CrossRef service. The second issue concerns permissions. Researchers want to know whether text and data mining is allowed on specific content via their current subscription agreements, and if not, get permission to mine the content. CrossRef Text and Data Mining Services makes use of licensing information embedded in article metadata at CrossRef and a registry for supplemental text and data mining terms and conditions for those publishers who make use of them.

There are no content mining tools provided by CrossRef by way of this service, but the API associated with the service can be built into these tools. Figure 1 shows, in bold, the steps that the CrossRef service helps accomplish, and the basic workflow of the service is presented in Figure 2.

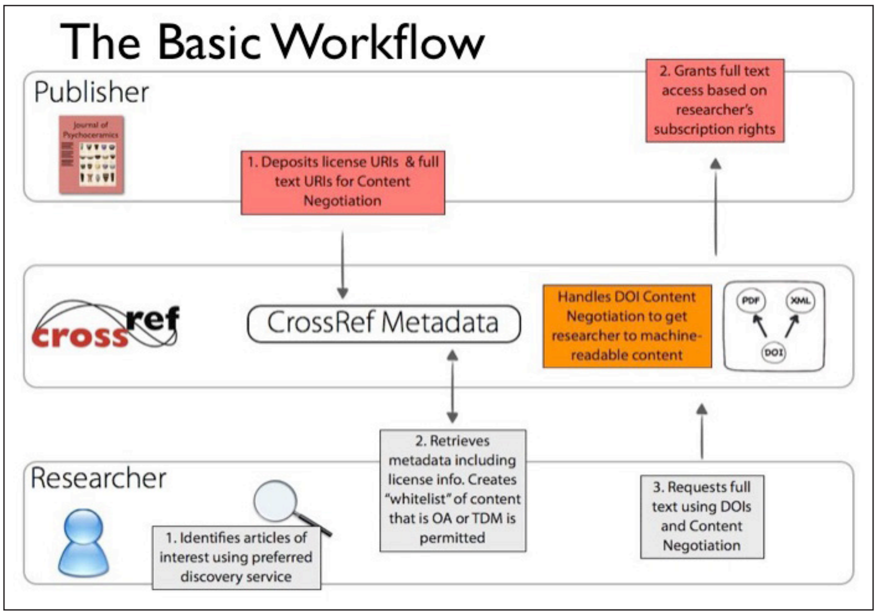
To walk through the process: the first thing a researcher will do is to identify a problem and the corpus of content that they want to mine. They are likely to identify this corpus using one of their preferred discovery tools, e.g. PubMed, Scopus or other search interfaces, rather than come to CrossRef to perform this search (as CrossRef only holds bibliographic metadata for articles and not the full text).



Text and Data Mining Steps

- Define problem
- Identify potential corpus to mine
  - **Discovery (full text links)**
  - **Identification of subset which can be accessed (license information)**
- **Download identified corpus**
- Text and data mine corpus

Figure 1. Steps in the text and data mining workflow



The Basic Workflow

Publisher

1. Deposits license URIs & full text URIs for Content Negotiation

2. Grants full text access based on researcher's subscription rights

CrossRef Metadata

Handles DOI Content Negotiation to get researcher to machine-readable content

Researcher

1. Identifies articles of interest using preferred discovery service

2. Retrieves metadata including license info. Creates "whitelist" of content that is OA or TDM is permitted

3. Requests full text using DOIs and Content Negotiation

Figure 2. The basic CrossRef Text and Data Mining Services workflow

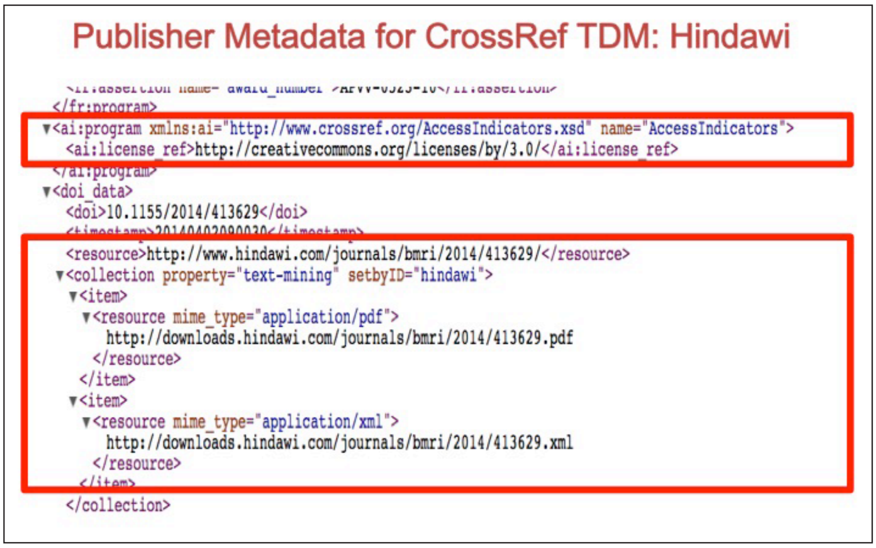
Once the researcher has a list of the content they are interested in, they can identify the DOIs associated with that content and come to CrossRef with that list to start the process of trying to retrieve the full text associated with those DOIs.

From the publisher side, they need to let the researcher know, on a per DOI basis, where that full text can be located on the publisher website and whether the researcher has permission to mine it. Publishers participating in CrossRef Text and Data Mining Services do this by making two additions to their CrossRef metadata: direct links to the full text into the CrossRef metadata for each DOI, and links to the licence under which the DOI is published. The publisher can choose which formats of the full text they link to, and can link to more than one format from the CrossRef metadata, giving the researcher the opportunity to work with the format they prefer. Since April 2014, participating publishers have been providing links to the full text in a mixture of PDF, XML and plain-text formats.

When the researcher comes to CrossRef with the list of DOIs they are interested in mining, they can use the CrossRef Text and Data Mining API to programmatically look at the metadata for each DOI and find out where the full text is for that DOI, and if they have permission to mine it under the licence the metadata shows it is published under. TDM researchers will have a whitelist of licences that they know will allow TDM – including the Creative Commons licences.

If researchers can see that they are allowed to mine the content under the terms of the licence, they can use a programming command to request the full text from the publisher website (shown by the full-text link in the metadata). The publisher can then return the full text from their website when it is requested by the researcher so that the researcher can start to compile the corpus of content that they will run their text mining tools on. (Figure 3 shows a snippet of CrossRef metadata showing full-text links and licence information for a journal article.)

> ‘researchers … can use a programming command to request the full text from the publisher website’



Publisher Metadata for CrossRef TDM: Hindawi

```
    <fr:assertion name="award_number">AFVV-0523-10</fr:assertion>
  </fr:program>
▼<ai:program xmlns:ai="http://www.crossref.org/AccessIndicators.xsd" name="AccessIndicators">
    <ai:license_ref>http://creativecommons.org/licenses/by/3.0/</ai:license_ref>
  </ai:program>
▼<doi_data>
    <doi>10.1155/2014/413629</doi>
    <timestamp>20140402090030</timestamp>
    <resource>http://www.hindawi.com/journals/bmri/2014/413629/</resource>
  ▼<collection property="text-mining" setbyID="hindawi">
    ▼<item>
      ▼<resource mime_type="application/pdf">
          http://downloads.hindawi.com/journals/bmri/2014/413629.pdf
        </resource>
      </item>
    ▼<item>
      ▼<resource mime_type="application/xml">
          http://downloads.hindawi.com/journals/bmri/2014/413629.xml
        </resource>
      </item>
    </collection>
```

Figure 3. A snippet of CrossRef metadata showing full-text links and licence information for a journal article

Because the researcher is going to the publisher site to retrieve the content (albeit in a programmatic way), the publisher is still in control of returning the full text to the researcher. For an OA publisher, they can simply return the full text when requested, and subscription publishers can choose to return the full text to subscribers using their existing access control mechanisms like IP recognition. The service does not offer a way for researchers to get access to subscription content they do not already have access to, but it will enable recognized TDM for subscribed content.

> ‘The service does not offer a way for researchers to get access to subscription content they do not already have access to’

The CrossRef API that researchers use to interact with the service also supports rate-limiting so that publishers can define the rate at which they will return the full text to the researcher in order not to affect the performance of their website for other users.

The click-through service

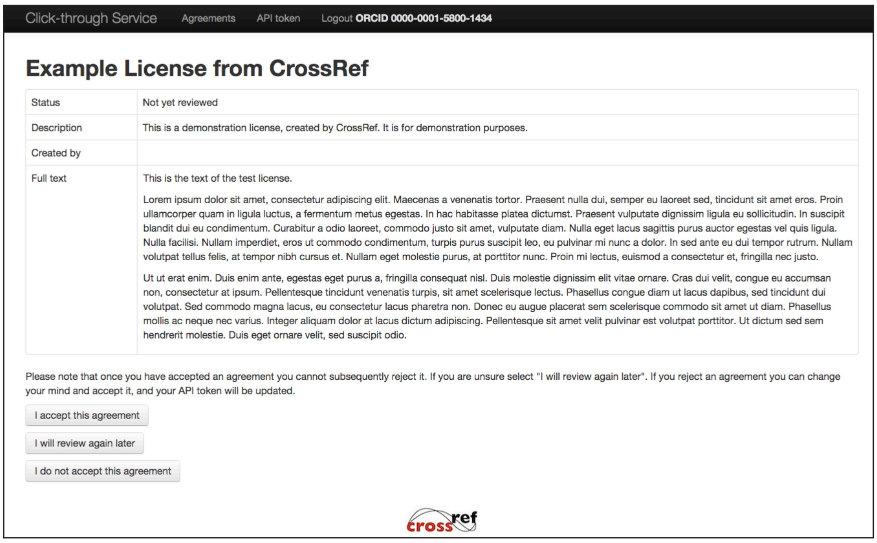
The basic workflow for CrossRef Text and Data Mining Services applies to OA publishers and those publishers who allow content mining as part of their existing subscription agreements. For publishers where this applies, all they need to do to participate in the CrossRef TDM service is to deposit the additional metadata with CrossRef, and optionally implement rate-limiting. CrossRef has provided some tools, including a file-upload option, to help publishers populate their existing DOI metadata with this additional information.

However, some publishers ask researchers who already have access to the content via a subscription to agree to additional terms and conditions (T&Cs) before they are allowed to mine the content. Publishers who fall under these criteria will also need to make use of the click-through service that CrossRef provides. The click-through service is a portal in which publishers can upload and manage these additional T&Cs so that they can be reviewed and accepted by researchers in one central place.

Publishers can communicate these additional conditions to researchers via the licence they link to in the CrossRef metadata. When a researcher follows the licence link, they can view the T&Cs under which they can mine the content. These terms may relate to how the content can be reused, or the rate at which it can be downloaded from the publisher.

‘When a researcher follows the licence link, they can view the T&Cs under which they can mine the content’

These terms and conditions will be displayed within a click-through service on the CrossRef website[2]. To interact with the service, researchers can register for it using their ORCID log-in details. Using their ORCID details means that the researcher can be disambiguated from other potential users of the service, and it means that CrossRef is not holding user ID and password information for them. When they are logged in, they will be able to see additional T&Cs posted by individual publishers. They can click to view the licences from the publishers whose content they are interested in mining, and on the page where the licence is displayed, they can choose to accept or reject the licence, or choose to review it again later. Figure 4 shows an example of this page.



Example License from CrossRef

Status	Not yet reviewed
Description	This is a demonstration license, created by CrossRef. It is for demonstration purposes.
Created by	
Full text	This is the text of the test license. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Maecenas a venenatis tortor. Praesent nulla dui, semper eu laoreet sed, tincidunt sit amet eros. Proin ullamcorper quam in ligula luctus, a fermentum metus egestas. In hac habitasse platea dictumst. Praesent vulputate dignissim ligula eu sollicitudin. In suscipit blandit dui eu condimentum. Curabitur a odio laoreet, commodo justo sit amet, vulputate diam. Nulla eget lacus sagittis purus auctor egestas vel quis ligula. Nulla facilisi. Nullam imperdiet, eros ut commodo condimentum, turpis purus suscipit leo, eu pulvinar mi nunc a dolor. In sed ante eu dui tempor rutrum. Nullam volutpat tellus felis, at tempor nibh cursus et. Nullam eget molestie purus, at porttitor nunc. Proin mi lectus, euismod a consectetur et, fringilla nec justo. Ut ut erat enim. Duis enim ante, egestas eget purus a, fringilla consequat nisl. Duis molestie dignissim elit vitae ornare. Cras dui velit, congue eu accumsan non, consectetur at ipsum. Pellentesque tincidunt venenatis turpis, sit amet scelerisque lectus. Phasellus congue diam ut lacus dapibus, sed tincidunt dui volutpat. Sed commodo magna lacus, eu consectetur lacus pharetra non. Donec eu augue placerat sem scelerisque commodo sit amet ut diam. Phasellus mollis ac neque nec varius. Integer aliquam dolor at lacus dictum adipiscing. Pellentesque sit amet velit pulvinar est volutpat porttitor. Ut dictum sed sem hendrerit molestie. Duis eget ornare velit, sed suscipit odio.

Please note that once you have accepted an agreement you cannot subsequently reject it. If you are unsure select "I will review again later". If you reject an agreement you can change your mind and accept it, and your API token will be updated.

Figure 4. An example licence displayed in the click-through service

The click-through service supports version control. Once a researcher has accepted a licence, the publisher cannot change that licence. They can retire it via the publisher version of the click-through service, and issue a new version of the licence, but they cannot change terms and conditions that a researcher has already agreed to.

The researcher also needs a mechanism to show the publisher if they have accepted their T&Cs when they come to request the full text via the CrossRef service. They do this via an API key that they can access when they register for the click-through service. The API key is an alphanumeric string specific to that researcher that they can add to their programmatic requests to the publisher for the full text.

Publishers using the click-through service to display additional T&Cs also get their own iteration of an API and an API key. When they get requests from a researcher for the full text, they can use their API in combination with their API key to verify that the researcher making the request has accepted the relevant licences.

If the results of the publisher's API query show that the researcher has agreed to the relevant T&Cs, and they have access to the content, then the publisher can return the full text to them.

Conclusion

The benefits of this service are that it streamlines researcher access to the distributed full text for the purposes of mining and enables machine-to-machine automated access for recognized TDM. It is set up to work across any publishers who choose to participate in the CrossRef service, eliminating the need for researchers and publishers to undertake lots of individual, manual interactions. It also enables article-level licensing information to be communicated and an easy mechanism for supplemental terms and conditions for TDM to be reviewed and accepted. CrossRef does not charge publishers any addition to their CrossRef membership fees to participate in the service, and similarly, the service is free for researchers to use.

As of April 2015, over 14 million DOIs have been enabled for use via the service, i.e. they have the full-text links and licence information deposited at CrossRef. Elsevier, the American Physical Society (APS) and Hindawi have registered the majority of these DOIs. However, over 100 publishers and societies have deposited these links in different volumes, including full participation by the Korean Association of Medical Journal Editors (KAMJE). Other publishers, such as Wiley, Springer, PLOS and IOP Publishing, have also supported the service; these publishers (among others) advised on the creation and set-up of the service, are working towards making their content available via the service and now serve on the CrossRef Text and Data Mining Services committee. Further information on the service is available from the CrossRef website and the support pages for researchers and publishers[3].

‘over 14 million DOIs have been enabled for use via the service’

The service has been live for just under 12 months and is available for use by researchers. Because researchers do not have to register to use the CrossRef API, it is difficult to track specific instances of the services being used, although publishers can see usage reflected in the download statistics from their platforms or look at requests for content containing the API key generated from the click-through service (if applicable). CrossRef solicited feedback from researchers during the pilot stage of the service, however, and the API has been used by a number of researchers. One of these researchers, Eric Lease Morgan (University of Notre Dame), has also blogged about his early experiences with it[4]. He notes that the service is ‘a step in the right direction’ but that publishers need to start to populate their metadata with the relevant information in order to participate.

‘publishers need to start to populate their metadata with the relevant information in order to participate’

Now that publishers have started to enable their content for use in this way, however, CrossRef is keen to support researchers using the service and get feedback on how they are integrating it into their workflows. CrossRef is also aiming to continue to promote the service to publishers and provide ways to help them participate to help researchers avail of their content in a streamlined way.

Competing interests: The author has declared no competing interests. As clearly stated, she is Product Manager at CrossRef.

References

1. Madhusoodanan, J, 17 April 2013, Announcing the PLOS Text Mining Collection, PLOS ONE Community Blog: http://blogs.plos.org/everyone/2013/04/17/announcing-the-plos-text-mining-collection/ (accessed 20 February 2014).

2. CrossRef terms and conditions on click-through service [ORCID ID needed to access]: https://apps.crossref.org/clickthrough/researchers/#/login

3. CrossRef website and support pages for researchers and publishers: http://tdmsupport.crossref.org/ (accessed 6 May 2015).

4. Morgan, E.L., 11 June 2014, CrossRef’s Text and Data Mining (TDM) API, Days in the Life of a Librarian: https://blogs.nd.edu/emorgan/2014/06/tdm/ (accessed 5 May 2015).

Rachael Lammey
Product Manager
CrossRef, Oxford, UK
E-mail: rlammey@crossref.org

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

To cite this article:

Lammey, R, CrossRef Text and Data Mining Services, *Insights*, 2015, 28(2), 62-68; DOI: http://dx.doi.org/10.1629/uksg.233

Published by UKSG in association with Ubiquity Press on 07 July 2015