Training Material

Using the Crossref Metadata API to explore publisher content

Rachael Lammey

Crossref, Oxford, UK

Abstract

Crossref is a not-for-profit membership association for scholarly publishers, founded in 2000. It is the largest digital object identifier (DOI) registration agency and provides publisher members with the capacity to deposit DOIs and their associated metadata to support persistent linking between different types of academic content. In a previous paper in *Science Editing* [1], Crossref mentioned a newly-created interface (http://search.crossref.org) that allowed publishers, libraries and researchers to search across nearly 50 million Crossref metadata records for items such as journal articles, books and conference proceedings. Since that paper was published, the search service has graduated into a live production service covering over 81 million DOIs, and Crossref has documented and made available the application programming interface used to build and support the search interface. This paper will provide information on this Crossref Metadata API, which is being widely adopted and used by many different stakeholders in the scholarly community, and give examples of how it is being employed by these parties.

Introduction

When publishers register Crossref digital object identifiers (DOIs), they do so by depositing, at minimum the bibliographic metadata related to each article: journal/book title, ISSN (International Standard Serial Number), work title, author, publication date (print and online), URL (uniform resource locator) of the content and the DOI itself. By providing this information, the piece of content can be distinguished from other, similar pieces of content in other publications. This enables the article, book or conference proceedings to be linked to and cited distinctly by other researchers, and for publishers to be able to track how widely the work is being used.

Over time, the metadata that Crossref collects from publishers has expanded in scope as the workflows that publishers need to support has grown. Over and above the standard biblio-

graphic information about a piece of content, publishers can also deposit ORCID iDs, funding and license information, full-text links (to enable indexing and text mining), updates to content and abstracts. With such a wealth of information being made available through the publisher metadata, it was becoming increasingly important that this information could be easily and widely disseminated. This way, anyone interested in using it could do so, and use it as a mechanism for finding information on publisher content, linking to it effectively and to build their own services on top of the information.

Before the current Crossref Metadata API was launched, it was possible to receive and interrogate the publisher metadata, but the process would have been as follows. To find out which licenses *Science Editing* uses via the Crossref metadata, an interested party would have to sign up to Crossref Enhanced Metadata Services [2], download around one terabyte of extensible markup language via the OAI-PMH protocol, and then parse and scan that information for *Science Editing* DOIs and the license information associated with those. Many third parties use this route, but many more needed information that would update dynamically as publishers deposited DOIs for new content and updated the information for existing content.

The Crossref Metadata API lets anyone search, filter, facet and sample Crossref metadata related to over 81 million content items with unique. It is free to use, the code is publically available and end-users can do whatever they want with the data. In exposing the authoritative cross-publisher metadata to the community in this way, it becomes more accessible, functional and much simpler to integrate with third party systems and services (from the publisher and the end-user side). This leads to smoother workflows and increased discoverability without changing existing publisher processes.

The history of the Crossref Metadata API

The Crossref Metadata API started life with the Crossref labs team in early 2013. The year before, Crossref had started a pilot in collaboration with publishers and funders to collect funding information in a consistent way in the publisher metadata so that it could then be used by funders to find and report on the outputs of the research they funded.

Crossref funding data [3] launched in May 2013, but to accompany the service there needed to be an efficient mechanism for funders to be able to get this data once it had been provided by publishers. It also needed to update dynamically as publishers added to or changed existing metadata, and funders needed to be able to filter and facet their searches to look for specific subsets of information to report on the KPIs

they were interested in. They also wanted reporting tools to be able to download, review and share this information as simply as possible.

Karl Ward, one of Crossref's Research & Development team worked on a revised, modern version of Crossref's existing application programming interfaces (APIs) to create a REST API that fulfilled the criteria that funders, research institutions and other third parties could use. Crossref also started to use it to build some of it's own tools like a search interface for funding information (http://search.crossref.org/funding) where anyone could come and ask for a list of the content that had been funded by one of the parties in the Open Funder Registry—a taxonomy of over 12,000 standardized funder names.

With the launch of funding data, Crossref started to see the API being used extensively. Coupled with that, the increased breadth of the metadata that publishers could provide Crossref has also been growing - letting it be interrogated and used in lots of interesting ways.

Current use cases for the application programming interface at Crossref

The metadata API is used extensively within Crossref to power various tools and services. As noted, it provides the backbone for Crossref Metadata Search and the linked funding data search interface. Using the full-text links and license links provided by publishers, the API can be leveraged to provide cross-publisher support for text and data mining applications [4].

It can also power reports and reporting. There is top-level information accessible via the API on the metadata Crossref holds (e.g., how many journal DOIs does Crossref have), article level information, or interesting subsets of information e.g., how many publishers are depositing ORCID iDs (and which ones?) longer term, Crossref plans to build publisher participation reports from the API so that members can easily check the completeness of the metadata they are depositing with Crossref.

Use cases by third parties

Third parties can, and do use the API to integrate publisher metadata into their own products and services. Organisations leveraging the metadata to report on funder information and compliance with funder mandates were our first use case, but that has grown to include: (1) searching and placing references dynamically in scientific blog posts e.g., in Coko Foundation's Pubsweet 'science blogger' alpha [5] science blog platform; (2) helping authors find and verify their publications.

Kudos [6] use this to help their authors identify the works they have published; (3) built-in citation search in authoring tools/DOI reference matching like Authorea [7]; (4) helping build databases of specific content types e.g., open access journals; (5) assessing license information as described by Impactstory in their blog (http://blog.impactstory.org/find-and-reward-open-access); and (6) it also has the potential to be used in helping streamline open access workflows within academic institutions. Crossref is working with Jisc in the UK and other interested parties on https://www.jisc.ac.uk/blog/new-publisher-led-initiatives-to-support-reporting-to-funders-21-mar-2016.

Even at this relatively early stage, it is apparent that the API has a wide variety of uses, which will continue to grow over time. Crossref has also been working with developer communities on the service. Scott Chamberlain of rOpenSci has built a set of robust libraries for accessing the Crossref API [8], available in the R, Python and Ruby languages. There's also a javascript library [9] authored by https://github.com/darobin so users can interact with the API in the programming language they prefer to use.

Conclusion

The Crossref Metadata API currently sees around 32 million requests a month, up from 20 million just a few months ago. Crossref doesn't require users to register to use the API, so success is measured by the volume of usage seen, but also in the diversity of use-cases for the API. Crossref plans to provide an optional service level agreement version of the service in order to provide additional functionality and increased reliability to users dependent on it for their own products and services. Crossref will work with them to gather requirements, resource these and provide a service level agreement version of the API. And of course, as publishers deposit more, richer metadata with Crossref, the scope of what the API can do and support will continue to grow in turn, enhancing discovery, linking, citation and collaboration - all of the principles that Crossref was set up to uphold when it was created.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

References

1. Lammey R. CrossRef developments and initiatives: an update on services for the scholarly publishing community from CrossRef. Sci Ed 2014;1:13-8. http://dx.doi.org/10.6087/kcse.2014.1.13
2. Crossref. Crossref enhanced metadata services [Internet]. Oxford: Crossref; 2016 [cited 2016 June 12]. Available from: http://www.crossref.org/cms/index.html
3. Crossref. Crossref funding data [Internet]. Oxford: Crossref; 2016 [cited 2016 June 16]. Available from: http://www.crossref.org/fundingdata/
4. Crossref. Crossref text and data mining services [Internet]. Oxford: Crossref [cited 2016 June 18]. Available from: http://tdmsupport.crossref.org
5. Collaborative Knowledge Foundation. PubSweet 1.0 "Science Blogger" alpha, INK 1.0 alpha releases [Internet]. [place unknown]: Collaborative Knowledge Foundation; 2016 [cited 2016 June 5]. Available from: http://coko.foundation/blog.html
6. Kudos [Internet]. Oxfordshire: Kudos [cited 2016 June 20]. Available from: https://www.growkudos.com/
7. Authorea [Internet]. New York: Authorea [cited 2016 June 20]. Available from: https://www.authorea.com/
8. ROpenSci [Internet]. ropensci/rcrossref. [place unknown]: GitHub; 2016 [cited 2016 June 20]. Available from: https://github.com/ropensci/rcrossref
9. Berjon R. scienceai /crossref [Internet]. [place unknown]: GitHub; 2016 [cited 2016 June 20]. Available from: https://github.com/scienceai/crossref