Use of CrossRef and OAI-PMH to Enrich Bibliographical Databases

Mehmet Ali Abdulhayoglu[1] and Bart Thijs[2]

[1]*Mehmetali.abdulhayoglu@kuleuven.be*
Centre For R&D Monitoring (ECOOM), K.U. Leuven, Waaistraat 6, B-3000 Leuven (Belgium)

[2]*Bart.thijs@kuleuven.be*
Centre For R&D Monitoring (ECOOM), K.U. Leuven, Waaistraat 6, B-3000 Leuven (Belgium)

Introduction

Today prominent and comprehensive databases such as Thomson Reuters' Web of Science (WoS) or Elsevier's Scopus are highly in use for bibliometric research. However, these databases do not index full texts hindering researchers to carry out more detailed analyses. Besides, it is possible that some indexed publications do not have DOI numbers playing an important role to access full texts. This paper focuses on how these above-mentioned deficiencies might be overcome by harnessing the Web sources CrossRef and OAI-PMH. Glenisson, Glänzel, Janssens, & De Moor (2005) and Alexandrov, Gelbukh, & Rosso (2005) stated and showed that full text can have an added value in comparison to abstract and title combination when mapping or clustering disciplines and subfields are in question. Therefore, automatic, rapid and free access to full texts of scientific publications might yield a significant contribution to bibliometric research.

Sources

CrossRef

CrossRef provides, besides its other valuable services, a Text and Data Mining (TDM) service enabling researchers to access full-texts of scientific papers for free (Lammey, 2014). This initiative might be a good alternative when considering the policies of the publishers over TDM hindering or retarding the scientific initiatives (Van Noorden, 2012). In this context, by means of a CrossRef REST API, which is free to be used by the public, the developer can access the metadata that CrossRef assembles from more than 4,400 publishers. Besides the metadata such as title, source (e.g. journal, book chapter etc.) name, co-author names, volume year, volume, issue, subject category, two additional important items might be given. These records are *license* and *links* where *link* gives the related full text link and *license* presents an URL link to the license which must be accepted when a GET request is triggered to access the full text. Figure 1 depicts how to access a full text through CrossRef for a given sample digital object identifier (DOI) and a java GET request. In CrossRef's web site, other methods are given to access full text. Since it is not mentioned in the site, we opt to give a java sample through a snippet.

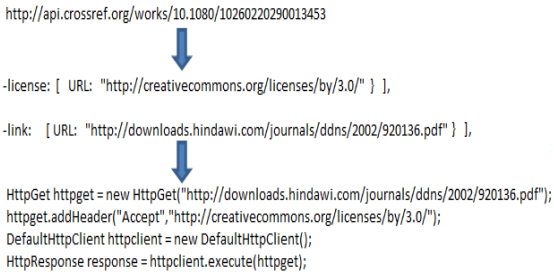

Figure 1. Process of accessing a full text presented by CrossRef by applying *license* and *link* information.

As of 22/12/14, CrossRef has thousands of publications metadata having both full text and license info from the publishers using creative commons license (CC-BY) which encourages the reuse and distribution of content. These publishers are given in Figure 2.

Publisher	CrossRef Number	CrossRef & WoS Number
HINDAWI PUBLISHING CORPORATION	123552	30737
PENSOFT PUBLISHERS	2233	1712
AIP PUBLISHING	273	5
AMERICAN ASSOCIATION OF PHYSICISTS IN MEDICINE (AAPM)	39	11
AMERICAN VACUUM SOCIETY	4	1
ACOUSTICAL SOCIETY OF AMERICA (ASA)	1	0

Figure 2. Number of publications according to publishers using creative commons license (CC-BY) with full text info within CrossRef and within CrossRef-WoS DOI combination.

On the figure's last column, the number of publications, which appear in both CrossRef and WoS, is given for those WoS records only having a DOI. Even though only a few publishers are willing to allow their contents to be mined, we believe that this number will increase over time as also stated by Van Noorden (2014).

Open Archives Initiative – Protocol for Metadata Harvesting (OAI-PMH)

OAI-PMH emerged aiming at enabling e-print archives to be interoperated (Van de Sompel & Lagoze, 2000). The content of the metadata depends on data provider, for example, while BMC

is providing full texts as well as other metadata, most of data providers such as arXiv do not provide full text or they just mention the URL link not guaranteeing that the full text can be freely downloaded. Below, some example links are given from arXiv and BMC which can be applied to harvest data.

http://www.pubmedcentral.nih.gov/oai/oai.cgi?verb=ListRecords&from=2014-01-01&metadataPrefix=pmc&set=bmcbiology (1)

http://export.arxiv.org/oai2?verb=ListRecords&metadataPrefix=arXiv&set=cs (2)

While former link gives the results only for the journal BMC Biology and those recorded in the repository later than 2014/01/01, later link invokes all the data from computer science discipline in arXiv repository without any date limitation. Note that both results will be invoked in accordance with their own XML schema.

Application

Combining WoS - arXiv - CrossRef

Leveraging arXiv repository, we harvested their OAI-PMH compatible data (See (2)) to combine with our WoS database by matching titles through a character N-Gram text matching process (Abdulhayoglu, Thijs, & Jeuris, 2014). In particular, from arXiv we retrieved title and DOI information for only the *computer science(cs)* discipline to deal with a relatively small data set. There were about 60,000 arXiv records while we have, in WoS, more than 35 million records indexed between 1991 and 2014. We searched for arXiv records within WoS and we found around 18,000 matches having a Salton similarity score higher than 0.90.

Besides 10,000 matches having identical titles, there were more than 7,000 matches having both Salton and Kondrak scores higher than 0.90. Finally, there were only about 200 matches having lower similarity Kondrak scores which can be re-checked manually or simply removed.

We examined the matches having very high similarity scores around 0.90-0.99 and saw that the small character corruptions might appear both on the database or repository side. Additionally, some terms might be given as a text string while it might appear as a symbol in the other source for exp. alpha and α. As a result a similarity score higher than 0.90, especially for Kondrak, can be applied for string matches. So, considering the observations just mentioned, we retained about 6,000 matches having both Salton and Kondrak scores higher than 0.90 and DOI information from the arXiv side.

The retrieved DOI numbers were supposed to be used for accessing full texts through CrossRef. However, a few accessed records have a CC-BY license and we could only grab 286 publications and download their full texts in pdf format. We controlled each full text whether they are correct by checking titles. During this optional process we applied a java pdf parser (*itextpdf*) and correctly extract the title information of those 286 publications. Besides *itextpdf*, CrossRef has its own tool named *pdfextract,* however, it is only applied on Linux environment. Lipinski, Yao, Breitinger, Beel, & Gipp (2013) compare some other extractors.

Conclusions and Discussions

Employing CrossRef and OAI-PMH, a process of accessing full texts of scientific publications indexed in WoS database is explained. Computer science articles from arXiv repository are matched with whole WoS database. Despite a high number of matches, the number of publications appearing within CrossRef repository having creative commons license is quite low. Though a small number of publications has creative commons license, CrossRef seems to ease the issue of accessing full texts freely in time (Van Noorden, 2014).

Acknowledgments

Authors would like to thank Rachael Lammey and Karl Ward from CrossRef, Meshna Koren from Elsevier, Mikail Shaikh from Springer and IT admins from arXiv for their valuable guidance and helps for their TDM systems.

References

Abdulhayoglu, M. A., Thijs, B., & Jeuris, W. (2014). Matching bibliographic data from publication lists with large databases using N-Grams. *Available at SSRN 2464065*.

Alexandrov, M., Gelbukh, A., & Rosso, P. (2005). An approach to clustering abstracts. In *Natural Language Processing and Information Systems* (pp. 275-285). Berlin: Springer.

Glenisson, P., Glänzel, W., Janssens, F., & De Moor, B. (2005). Combining full text and bibliometric information in mapping scientific disciplines. *Information Processing & Management*, *41*(6), 1548-1572.

Lammey, R. (2014). CrossRef's Text and Data Mining Services. *Learned Publishing*, 27(4), 245-250.

Lipinski, M., Yao, K., Breitinger, C., Beel, J., & Gipp, B. (2013, July). Evaluation of header metadata extraction approaches and tools for scientific PDF documents. In *Proc. 13th ACM/IEEE-CS Joint Conf. on Digital Libraries* (pp. 385-386). ACM.

Van de Sompel, H., & Lagoze, C. (2000). The Santa Fe convention of the open archives initiative. *D-Lib Magazine*, (2), 2011-10.

Van Noorden, R. (2014). Elsevier opens its papers to text-mining. *Nature*, *506*(7486), 17-17.

Van Noorden, R. (2012). Trouble at the text mine. *Nature*, *483*(7388), 134-135.