Aggregating Research Papers from Publishers' Systems to Support Text and Data Mining: Deliberate Lack of Interoperability or Not?

Petr Knoth, Nancy Pontika
The Open University
Walton Drive, Milton Keynes, United Kingdom
petr.knoth@open.ac.uk, nancy.pontika@open.ac.uk

Abstract

In the current technology dominated world, interoperability of systems managed by different organisations is an essential property enabling the provision of services at a global scale. In the Text and Data Mining field (TDM), interoperability of systems offering access to text corpora offers the opportunity of increasing the uptake and impact of TDM applications. The global corpus of all research papers, i.e. the collection of human knowledge so large no one can ever read in their lifetime, represents one of the most exciting opportunities for TDM. Although the Open Access movement, which has been advocating for free availability and reuse rights to TDM from research papers, has achieved some major successes on the legal front, the technical interoperability of systems offering free access to research papers continues to be a challenge. COnnecting REpositories (CORE) (Knoth and Zdrahal, 2012) aggregates the world's open access full-text scientific manuscripts from repositories, journals and publisher systems. One of the main goals of CORE is to harmonise and pre-process these data to lower the barrier for TDM. In this paper, we report on the preliminary results of an interoperability survey of systems provided by journal publishers, both open access and toll access. This helps us to assess the current level of systems' interoperability and suggest ways forward.

Keywords: Interoperability, publishers, standardisation

1. Context

Each year approximately 1.5 million research papers are being published and only 4% of these are available via an open access journal (Björk and Lauri, 2009). Even though the availability of this high volume of scientific papers brings new opportunities for content discoverability, enables the advancement of the disciplines through the practice of TDM, and constitutes an important financial asset, there are still threats that do not allow its application. Mainly, there are two types of challenges, legal and technical, which have been discussed extensively in the European Union reports (Science Europe, 2015; European Commission, 2015). In this paper, we will explore the technical challenges and, more specifically, we will focus on the interoperability of publisher systems and whether the aggregation of their content is feasible. Furthermore, we would like to advocate for clear interoperable annotation resources regardless of their license and format. The initial idea of conducting the machine accessibility survey follows one of the outcomes of the technical prototyping work of the Open Mirror feasibility study (Knoth and Russell, 2014), commissioned by a non-departmental funding body, Jisc, which highlighted the technical difficulty in aggregating open access content from the systems offered by the major publishers.

A study into the Value and Benefits of Text Mining authorised by Jisc in 2012 (McDonald and Kelly, 2015) concluded that text-mining of research outputs offers the potential to provide significant benefits to the economy and the society in the form of increased research efficiency, by unlocking hidden and developing new knowledge and improving the research process and its evidence base. These benefits will result in significant cost savings and productivity gains, innovative new service developments, new business models, new medical treatments, etc. In order to realise these benefits, we need a harmonised access to research content for TDM.

CORE is a global aggregator service, collecting metadata and full-text of the open access scientific papers from repositories and journals from around the world. CORE is collecting the metadata of resources using the Open Archives Initiative Metadata Harvesting Protocol (OAI-PMH), which is one of the most popular standards (Horwood and Garner, 2004). The metadata are typically formatted using the Dublin Core schema[1], but we also need to be able to consume other protocols, such as METS[2] or RIOXX[3]. While these protocols appear as standardised solutions, the way metadata is expressed by different systems claiming to conform to them is highly inconsistent.

As there is no widely adopted standard for full-text harvesting[4], CORE uses a range of approaches to harvest the content. For instance, we have developed approaches that:

- recognise links to full-texts in metadata,
- apply a focused crawling approach starting from a particular web resource with the goal to discover a specific paper,
- are completely custom-built for a particular provider. Managing such an infrastructure, which lacks technically, is challenging as one cannot rely on it.

At CORE, we have a great interest in the increased interoperability of publishers' systems as it enables us to con-

[1] http://dublincore.org/

[2] https://www.loc.gov/standards/mets/

[3] http://rioxx.net/

[4] We do not consider ResourceSync http://www.openarchives.org/rs/1.0/resourcesync as a widely adopted standard at this stage as also revealed by our survey.

centrate on helping the TDM community rather than dealing with problems of aggregating content on a provider by provider basis. At the moment, we are required to have a detailed understanding of the technical details of hundreds of systems providing machine access to research papers. As we are now interested in enriching the CORE collection by gaining access to open access articles published by commercial (toll access) publishers, we have conducted a survey of the machine accessibility of open access articles stored in publishers' systems.

2. Survey

The survey was initially sent to sixty publishers by email. However, the response rate was extremely low. Surprisingly only Elsevier originally responded. This could indicate that publishers were originally not ready to respond due to their lack of knowledge of the TDM needs or due to them being unable/not ready to direct the survey to the appropriate person within their organisation. As a second step, we started calling publishers asking for a conversation with the person(s) responsible for policy decision making issues and/or technology related issues in their organisation. This route proved to be problematic as well; a number of publishers have a no-name policy, while, those who provided us with a contact name and a phone number, were not reachable when we tried to contact them. This lead to our third attempt, which proved to be the most successful. We asked a UK funding organisation, which deals often with publishers, to share their contacts with us in order to be able to proceed with the survey. The organisation shared with us 16 publisher contact information and we received a response from 11 of them.

The survey was composed of 10 questions, both closed and open ended, where the publishers were asked to provide information on the following themes: open access publishing activities; machine interface availability; type of machine interface; identification of open access papers; access to full-text of open access papers; restrictions on accessing full-text; licenses used for open access articles; open access machine interface; and planned machine interface.

2.1. Publishers profiles

The publishers who responded to our survey were a mix of both subscription based or toll access and open access publishers (Table 1). Even thought the survey's response rate was relatively low, nonetheless we were satisfied that we received responses from international publishing houses, such as Elsevier and Palgrave Macmillan, which are subscription based publishers, and eLife Sciences and PeerJ, both open access publishers.

In an effort to collect as much information as possible from the publishers and to be able to address the current state of the interoperability requirements more accurately, we included in our survey the question of approximately how many open access articles each one of them has published so far (Table 2). We discovered that indeed a large number of open access journals has already been published, the content of which could be used for TDM purposes with potential great benefits for the various subject fields and the advancement of the society.

Toll Access	Open Access
Elsevier	eLife Sciences
Palgrave Macmillan	PeerJ
Cambridge University Press	Frontiers
IOP Publishing	
Royal Society of Chemistry	
HighWire Press	
Dove Medical Press	
Publishing Technology Plc	

Table 1: Publishers' publication models

Publishers	Open Access Articles
Elsevier	No Response
Palgrave Macmillan	18,500
Cambridge University Press	1,409
IOP Publishing	5,800
Royal Society of Chemistry	2,000
HighWire Press	150,000
Dove Medical Press	5,000
Publishing Technology Plc	1
eLife Sciences	1,600
PeerJ	1,600
Frontiers	1,600

Table 2: Publishers' number of open access publications

In addition, we asked the publishers to provide us with an estimation of the forthcoming year's open access publications (Table 3).

Publishers	Open Access Articles
Elsevier	No Response
Palgrave Macmillan	15,000
Cambridge University Press	500
IOP Publishing	10,00
Royal Society of Chemistry	2,000
HighWire Press	15,000
Dove Medical Press	5,800
Publishing Technology Plc	10,000
eLife Sciences	900
PeerJ	1,000
Frontiers	14,000

Table 3: Publishers' estimation of open access publications for the forthcoming year

Based on their responses, one can conclude that the number of open access articles is steadily growing, something that could be attributed to the continuous growing of funders' open access policies[5]. The current situation presents a large opportunity for the development of TDM that cannot be overseen, but acquiring methods and ensuring access to this content must be further investigated.

[5] http://roarmap.eprints.org/

3. Preliminary Survey Results on Interoperability

Even though this is still work in progress, we thought that it would be a good opportunity to use this workshop to present some of the preliminary survey results, discuss the findings and address the challenges relating to the interoperability of publishers systems and whether these allow and enable the aggregation of the open access content.

Based on the responses collected, we discovered that the biggest proportion (N=11, *n*=7, 63.6%) of the publishers who responded provide a machine interface to the metadata of papers published on their websites.

With regards to the standards used that enable the machine accessibility we saw that there was approximately an equal number of publishers that are using the international standard OAI-PMH and have their own API (Table 4). We received only one response regarding the use of the Z39.50 protocol, which can be explained based on the fact that it is an old protocol and not widely used lately.

Standard	No. of Publishers
OAI-PMH	6
Own API	5
Z39.50	1
ResourceSync	0
Other	0

Table 4: Standards followed by publishers

On the question on whether the article's full-text is referenced in the article metadata we received again a mixture of responses (Figure 1). Not providing a direct link to the full-text significantly complicates content harvesting causing a situation in which a metadata record is often not unambiguously linked to the item it describes. Such approaches have been repeatedly discouraged[6] (Knoth, 2013). Unfortunately, providing only a DOI cannot be seen as a good practice on its own as DOIs often do not resolve to the full-text but only to a article "splash page". Two publishers declared that their interface supports the transfer of the full text document, which is a good approach, and only one mentioned that they provide the link to a "splash page". Four publishers did not provide an answer to this question.

In the end we asked the publishers if there are any restrictions on programmable accessing the full-text of the articles. Eight publishers responded to this question and the most popular answer (*n*=7) was that that they offer this content through their website, four mentioned that they release it through an API, while there was one publisher using the FTP functionality. From these publishers, three of them offer both a website and an API functionality. However, offering full-text content only through a web interface is completely insufficient for TDM purposes where the aggregator needs to quickly transfer and process large quantities of content. This is a particularly important issue due to the fact that many publishers completely disallow or significantly limit the access to robots on their website with Googlebot being usually the only exception.

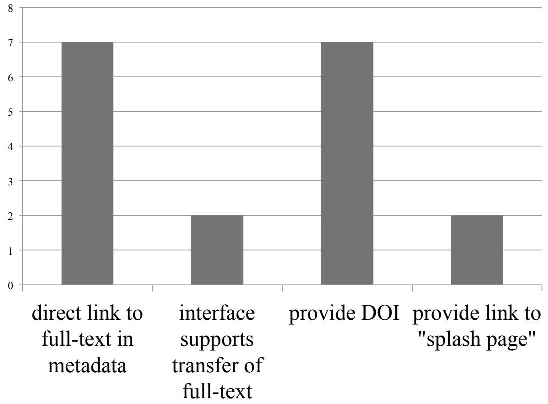

Figure 1: Reference of article's full-text in the metadata.

3.1. Significance of the investigation and the results

The purpose of this work is to explore an issue that has not been investigated in the past, the machine access interoperability of publishers' systems. This topic is of great importance not only to those interested to engage in TDM activities, but also to sponsors of publicly funded research and consequently to the society (McDonald and Kelly, 2015). We started our research with a list of sixty publishers, but we received only one response. Our second attempt, contacting the publishers by phone, was not successful as well. We perceive, thought, that our third attempt provided us with a very high response rate. From the 16 publishing houses we contacted, 11 publishers responded to our survey, a response rate of 68.7%.

4. Future Work

Our next steps are to analyse the results we have received in depth. In addition, we will investigate TDM information provided on publishers' websites, especially those who did not respond to our research. In the past, we have seen that there is often a substantial discrepancy between the standardisation level declared as supported by the system providers and the level actually provided. Consequently, we plan to validate the declared results by actively harvesting open access content from these systems, measuring their response time, success rate and other parameters. We plan to make these results openly available. We aim to provide these results as a feedback to the content providers and research funders as we believe this could lead to an improved situation.

5. Conclusion

Enabling harmonised access to all research papers for TDM purposes continues to be a technically challenging problem. In a recent study Sompel and Nelson (Van de Sompel and Nelson, 2015) recommend the creation of interoperable systems to enable a "thriving web-based scholarly ecology". The results of our survey show that there is a pressing need to improve not just the adoption of standards on the content provider's side, but also the application of

[6] http://www.rioxx.net/

good practices of their use, such as the principles for direct linking to full-text. The CORE project is putting effort in monitoring the size of problem, harmonising the access to research papers and encouraging content providers to adopt relevant standards and good practices.

6. Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation programme (H2020-EINFRA-2014-2) under grant agreement No. 654021. It reflects only the author's views and the EU is not liable for any use that may be made of the information contained therein.

7. Bibliographical References

Björk, R. and Lauri, M. (2009). Scientific Journal Publishing: Yearly Volume and Open Access Availability. *Information Research*, 14(1).

European Commission. (2015). Standardisation in the Area of Innovation and Technological Development, Notably in the Field of Text and Data Mining: Report from the Expert Group and the Need for a Science-friendly EU Copyright Reform.

Horwood, S. Sullivan, E. Y. and Garner, J. (2004). OAI Compliant Institutional Repositories and the Role of the Library Staff. *Library Management*, 25(4/5).

Knoth, A. R. and Russell, R. (2014). Open Mirror Feasibility Study: Appendix A: Technical Prototyping Report.

Knoth, P. and Zdrahal, Z. (2012). CORE: Three Access Levels to Underpin Open Access. *D-Lib Magazine*, 18(11/12).

Knoth, P. (2013). From Open Access Metadata to Open Access Content: Two Principles for Increased Visibility of Open Access Content. In *Proceedings of the Open Repositories Conference 2013 (OR2013)*, Charlottetown, Canada, july.

McDonald, D. and Kelly, U. (2015). Value and Benefits of Text Mining.

Science Europe. (2015). Text and Data Mining and the Need for a Science-friendly EU Copyright Reform.

Van de Sompel, H. and Nelson, M. (2015). Reminiscing About 15 Years of Interoperability Efforts. *D-Lib Magazine*, 21(11/12).