Stakeholders in academic publishing: text and data mining perspective and potential

Maria ESKEVICH[1]

Radboud University, Nijmegen, The Netherlands

Abstract. In this paper we discuss the concept of open access in academic publishing with the focus on the right to mine the data once the right to read is granted. Thus we en- visage the roles and types of the stakeholders in academic publishing from the per- spective of the potential text and data mining (TDM) applications. Further on, we briefly introduce FutureTDM project that aims to improve TDM uptake in Europe.

Keywords. digital libraries, text and data mining, FutureTDM project

1. Introduction

The main incentive for academic publishing is to share the knowledge acquired through experimental and/or empirical observations with an overall aim to promote further scientific development and knowledge distribution. Hence, initially the publishing empowered more researchers and thinkers to improve their expertise and to expand the overall knowledge ontology. The dialogue was set up between the content creators through the medium of written and printed text providers. However, the societal and technological development in combination with population growth over the recent centuries lead to a more complicated framework of agents in the field of scientific knowledge sharing, as the same agents can play different roles.

Th The publication process, while having a target to broaden the access to the knowledge across communities, is a service that is being provided, and thus over the years it converted into a business model which raised a pay wall between the ultimate content consumers, i.e. researchers, general audience, and the content itself. As a vast amount of research is being carried out on public funding, the new frameworks for efficient scientific knowledge transfer are discussed and promoted, with the Open Access (OA) strat- egy being the main focus (De Grandis, Lomazzi, Rettberg). Following the OA principles defined in Budapest and Berlin Declarations, the European Commission (EC) defines Open Access as 'the practice of providing online access to scientific information that is free of charge to the end user and that is reusable', where scientific information can refer to (i) peer-reviewed scientific

[1] Corresponding Author: Maria Eskevich, The Netherlands; E-mail: m.eskevich@let.ru.nl

research articles (published in scholarly journals) or (ii) research data (data underlying publications, curated data and/or raw data) . These definitions describe 'access' in the context of open access as including not only basic elements such as the right to read, download and print, but also the right to copy, distribute, search, link, crawl, and mine2. Mining of both the publication text itself and of the corresponding data sets can be carried out with the help of diverse text and data mining (TDM) tools. Based on the last decades development in this domain, it is evident that TDM mechanisms are present throughout scientific and cultural environments, but not in a systematic or infrastructural way. TDM could help in solving scientific problems, which is why we see it in the heart of the future of Open Science. It is often used in domains that are rather advanced in their open and interoperable practices, e.g. bioinformatics, signifying a change in the modus operandi of performing research already showcasing a shift in approach to organizing Science. However, as reported in the Royal Society 2012 report

"Science as an Open Enterprise" , new text-mining technologies and developments in multidisciplinary research would be empowered if TDM barriers were lowered, and there are global policy and political signals that this is not only scientifically desirable, but ultimately inevitable.

In this paper we outline the field of text and data mining that in our view should be incorporated into the publishing practice framework in order to profit from the state-of- the-art TDM research technologies which can be helpful across all fields of science. Thus we regard the structure of the academic publishing stakeholders from the angle of TDM technologies involvement. The remainder of this paper is structured as follows: Section 2 introduces the open access publication agenda (2.1) and introduces the concept of usefulness of the text and data mining as its next step (2.2); Section 3 describes in details what kind of different potential roles (3.1) the diverse players in publishing, research and overall society (3.2) can take in order to promote further beneficent interaction of TDM technologies and the digital publishing; Section 4 reports the state-of-the-art research that can represent potential implementation use cases; Section 5 introduces the FutureTDM project that analyses current TDM uptake across different fields and outlines its potential; and finally Section 6 gives conclusions and outlines directions for future work.2.

2. Academic publishing: challenging background

The amount of academic publications is steadily growing across different fields of re- search with thousands of papers being produced each year, e.g. Figure 1 illustrates the case of Computer Science publications over the past 20 years. The sheer volume of publications pool and the growing trend impede research community, as well as generally the society members, to track all the trends within one field, and it becomes even more challenging to target multidisciplinary domains or to promote cross-domains methods applications. Having the access to this content, TDM can help researchers to cope with the tripling rate of growth of scientific output (Laren 2010).

2.1 Open Access Publication Agenda

The European Commission has made open access a general principle of Horizon 2020 in order to boost innovation capacity1. 'Open access' publications make scholarly literature freely available on the Internet, so that it can be read, downloaded, copied, distributed, printed, searched, text minded, or used for any other lawful purpose, without financial, legal or technical barriers, subject to proper attribution of authorship. Open access improves the pace, efficiency and efficacy of research. It heightens the visibility of authors and the potential impact of their work.

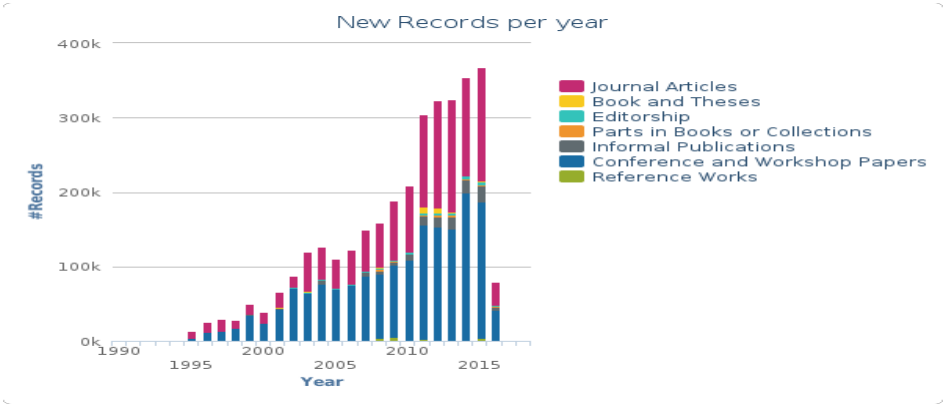

Figure 1: Total number of publications of the different publication types according to DBLP Computer Science Bibliography. [Accessed on 03.2016]

It removes geographical and structural barriers that hinder the free circulation of knowledge. Thereby contributing to increased collaboration, and ultimately strengthening scientific excellence and societal progress. It would seem therefore that open access is a major factor in increasing the uptake of TDM. Yet, it seems that the potential of open access as a means to facilitate data-driven innovation may be undermined by lack of interoperability between licenses and the proliferation of licenses which prohibit the creation of derivatives. This transition requires cooperation of all the stakeholders in the field.

2.2 Is the right to read becoming a right to mine?

The right to mine that is stated as the principle of the open access implies the availability and development of the TDM techniques which in reality requires a framework of data storage, access and processing that may be built with the collaboration between different stakeholders in the field. The TDM research is rapidly growing, but its incorporation into the publishing agenda is still affected by several factors (Hargreaves): economic issues of the market practices changes that it should bring on, the legal issues of copyright (Handke 2015), the lack of awareness among key potential stakeholders, the need for additional training of librarians, researchers, etc.

3. Stakeholders in the field and their different yet overlapping agendas

Stakeholders in the field may be actively engaged in publishing and/or text and data mining directly in their day to day activities, as service providers or developers; or they may have an indirect interest in knowledge discovery, analyze and/or make use of the information gleaned through content mining.

3.1 Stakeholder Roles

We assume a number of general roles that can be taken by different/same stakeholders in the field. Each role is associated with a different step in the circle of knowledge sharing:

Direct work with the TDM process, legal and financial support of this work. Building of a sustainable infrastructure for TDM requires the main stakeholders to undertake the following roles:

- Data Provider: in the framework of academic publishing it implies both papers writing, editing for publication, indexing in the database of publications and as- sociated resources;
- Processing Techniques Developer: the core TDM research is to be implemented based on the state-of-the-art scientific accomplishments in the field;
- Service Providers: once the TDM techniques are developed into a software, the results of the automatic analysis in terms of trends analysis, building solutions based on TDM trends can be released as a service;
- User of TDM techniques and results: the new insights into problems and extended knowledge based on the TDM extracted data and trends that can and should be accessible to use for the research community and general audience.

3.2 Stakeholder Types

The community that can benefit from the new academic publishing framework expanded with the TDM perspective is broader than simply researchers, publishers, and librarians. We foresee the involvement of both public/non-profit and industry sectors. Figure 2 exemplifies how the different stakeholders roles listed in Section 3.1 can be associated with different agents in the community.

Within this framework, content providing means both the provision of the content of the publications, as well as general data that the paper's discussion and experiment sections can be built on. TDM research is being carried out within the research institutions or departments in both public and industrial sectors, and its outcome is available both through the relevant publications or through the services to the general audience.

The algorithms behind TDM research are created and thoroughly investigated within natural language processing (NLP) communities. However, due to the legal restrictions of access, these studies are more often carried out on limited corpora when directly applied to the scientific publications, or otherwise the NLP researchers test their

theories on the other datasets with an assumption of potential further technology transfer across datasets.

4. TDM techniques use cases

In this section we outline current trends in the applied computational linguistics research that are already directly applied to academic publications in order to extract the knowledge or demonstrate potential within the outlined framework. Overall, there are three directions for these applications: information extraction of the content across a set of publications; summarization of the information across a set of publications; use the TDM techniques to reinvent the impact measurements of the scientific publications:

- Information extraction: Each section of a paper can be treated separately when different information is to be extracted for further analysis. It varies from simple detection of the papers published across different venues within the same project and funding scheme to more complicated cases of citations sentiment detection which allows better comprehension of the relevance of the current paper to the work in the field (Hong 2015).

Examples of diverse agents from Public/non-profit and Industry sectors	
Public/non-profit sector	Industry sector
Role: TDM content providers	
Publishers, national and university library organisations; Repositories, open access facilitators, databases, open access publishers; The World Wide Web; Citizens	Private publishers; Industry collecting data on the customers (Energy, Financial, Health, Retail); Journalists, Telecommunication Services
Role: TDM Creators and Developers	
Data scientists in Research institutions	Industrial Research and Development
Role: TDM Service providers: Knowledge aggregators and analysis based on TDM technologies	
Researchers and their associated organisations, Research libraries	Technology experts/data centers, service providers, i.e. telecommunications, software applications, storage providers for data, developers big data analytics providers, data services, journalists, news services, search services
Role: Consumers of TDM	
Research councils, universities, data scientists, Research institutes professional associations, Citizens	Journalists, Retail organisations, governments, public sector bodies etc. Energy, Financial, Health Care, Information Technology, Telecommunication Services,
Role: Funders	
Public funders of TDM initiatives: EU institutions and national Governments Inter-governmental organisations, public sector bodies	Private funding initiatives, Funding of internal research and development
Role: Policy shapers	
EU institutions, public sector bodies, national Governments Inter-governmental organisations, advocacy groups and legal experts	Lobbyists

Figure 2: Examples of diverse agents from Public/non-profit and Industry sectors

- Information summarization Once the separate facts are extracted from the papers, this information can be automatically summarized for further analysis using TDM and NLP techniques.
- Impact measurement Evaluation of the impact of the content, raising the profile of the publications using novel approaches to bibliometrics (Mayr 2015, Athar 2012), as not only the sheer number of citations might be representative of the quality of a certain research paper, but such details as the context of citations can bring better insight into understanding of papers overall value and mutual relevance.

5. Future TDM Project

FutureTDM2 project supports the uptake of TDM across all sectors of economy, considering publishing sector being of high importance, as so much of scientific information is confined within the deluge of publications which can be profitable for both commercial and non-profit use. In this paper, we discuss the scientific publications context, while within this project in general we aim to contact various types of stakeholders from different sectors in order to identify how their progress in the field can be supported when embracing TDM on the large scale, and how the structure of the roles is to be readjusted for each specific sector accordingly.

6. Conclusions

In this paper we have discussed the importance of TDM technologies for the future development of the academic publishing, and introduced the structure of the roles and types of the stakeholders in the field accordingly. This discussion should raise the awareness of the TDM potential and bring better understanding of the interaction structure.

7. Acknowledgments

This work is supported by the European Commission's HORIZON 2020 Programme (H2020) under GARRI-3-2014 – 665940 (FutureTDM).

References

Hargreaves et al. Expert Group Report on standardisation in the area of innovation and technological development, notably in the field of Text and Data Mining, http://ec.europa.eu/research/ innovation-union/pdf/TDM-report from the expert group-042014.pdf [Accessed on 12.2015]

Berlin Declaration of 2003 http://openaccess.mpg.de/67605/berlin declaration engl.pdf [Accessed on 12.2015]

De Grandis G., Neuman Y. Measuring Openness and Evaluating Digital Academic Publishing Models: Not Quite the Same Business. Journal of Electronic Publishing, Volume 17, Is- sue 3: Metrics for Measuring Publishing Value: Alternative and Otherwise, Summer 2014. DOI: http://dx.doi.org/10.3998/3336451.0017.302

Lomazzi L., Chartron G. The implementation of the European Commission recommendation on open access to scientific information: Comparison of national policies. Inf. Services and Use, vol. 34 (3-4), pp. 233–240, 2014.

Rettberg N., Schmidt B., RossA. Infrastructures for Policies: How OpenAIRE Supports the EC's Open Access Requirements. New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science: Scale, Openness and Trust, ELPUB 2015, pp. 185 - 189, IOS Press.

C. Handke, L. Guibault, J.-J.Vallbe´. Is Europe falling behind in data mining? Copyright's impact on data mining in academic research. New Avenues for Electronic Publishin in th eAge of Infinite Collections and Citizen Science. ELPUB 2015, pp. 185 - 189, IOS Press.

P.O. Laren, M. von Ins. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. Scientometrics 84(3): 575603. 2010.

P. Mayr, P.Schaer, A. Scharnhorst, P. Mutschke. Editorial for the Bibliometric-Enhanced Information Retrieval Workshop at ECIR 2014. CoRR abs/1404.7099 (2014). Amsterdam, the Netherlands.

P. Mayr, I. Frommholz P. Mutschke. Editorial for the 2nd Bibliometric-Enhanced Information Retrieval Workshop at ECIR 2015. CEUR Workshop Proceedings, Vol. 1344. Vienna, Austria, March 29th, 2015.

A. Athar, S. Teufel. Detection of Implicit Citations for Sentiment Detection. Proceedings of the Work- shop on Detecting Structure in Scholarly Discourse at ACL '12, Jeju, Republic of Korea, pp. 18–26, 2012.

K. Hong, M. Marcus, and A. Nenkova. System Combination for Multi-document Summarization. EMNLP, page 107-117. The Association for Computational Linguistics, (2015)

.