THE PUBLISHERS ASSOCIATION

Enabling Text Mining Of Scholarly Content: Update On Current Developments

Introduction

This paper sets out the practical steps being taken to facilitate and enable text and data mining of published works, especially articles in scientific and scholarly journals.

Publishers are committed to developing licensing solutions designed to facilitate text and data mining, and to continuing to improve the licensing process for researchers and would-be licensees. As this paper demonstrates, digital market-based solutions to some of the emerging issues around text and data mining can be addressed through improvements to licensing and technological standards – and in some cases a combination of these. Such an approach is far superior to the introduction of a copyright exception to allow (in theory but not necessarily in practice) such mining. Whilst an exception may remove the publisher's exclusive right to prevent copying for the purposes of text and data mining, a researcher's ability easily to engage in the activity across different publisher platforms is dependent upon there being an alignment of technological and contractual standards and practices.

The PA and its colleagues in the Federation of European Publishers and the International Association of Scientific, Technical and Medical Publishers (STM) remain firmly committed to the process of identifying practical licensing solutions. Publishers will continue to work closely with the research community to enable and develop text mining services, especially where these have the potential to drive discoveries in medical and scientific disciplines. With this in mind, work continues on the following elements of an industry-led solution.

Licence Standardisation

Many publishers' standard subscription agreements with universities already provide for a certain amount of re-use of subscribed content, including text and data mining for scientific research purposes.

However, given the competitive nature of the market in academic publishing and the fact that different companies have developed policies and practices separately to one another, different licence terms and conditions often apply, depending on the publisher and content in question. These differences in approach have created some concerns for those seeking to mine publishers' content, who require a more simplified licence framework.

For this reason and to enable publishers, both large and small, to simply and clearly adopt licence terms to facilitate mining of subscribed content, STM has developed three sample licence clauses. These clauses, governing usage terms and methods of access, are designed to be inserted or added to subscription agreements between publishers and universities and/or research institutions. This provides publishers with flexible means to grant permissions whilst providing researchers with the security that they are legally entitled to text mine copyright works.

Publishers have been encouraged to consider inserting these sample clauses allowing text and data mining for non-commercial purposes as part of subscription agreements.

Some of the larger publishers have indicated that they are pro-actively inserting TDM clauses into their new subscription agreements, while many consider inserting them also at the time of renewal of existing subscription agreements (many subscription agreements are renewed every three years, while some are renewed annually and others every five years).

Some publishers have a TDM clause ready to be deployed as and when approached by researchers or library representatives for inclusion, but will only deal with TDM on request to gain further experience; to allow for the adaption of their electronic platforms; and to render the published materials more easily minable first.

Irrespective of the way in which TDM is taken care of legally through subscription and content access agreements, many publishers will participate in Crossref's Prospect programme (see below). As part of that programme researchers at subscribing institutions will be able to get ready access to content for TDM non-commercial purposes by way of a so-called click-through licence and by using Crossref's API standard.

Technology Standardisation

Technological specifications differ across publishers' platforms, and this poses problems for miners seeking to access content from different publishers as seamlessly as possible.

Different systems allow or can cope with different load rates and download rates. They also have differing abilities to convert text to machine-readable form. These differences in publisher site capabilities are a direct reflection of the differing technologies which exist to carry out mining: there are a plethora of different robots, spiders and crawlers and automated downloading programs, algorithms and devices. They each have different characteristics in terms of their searching, scraping, extracting, linking and indexing.

Such differences in specification would not be swept away or in any way mitigated by a copyright exception. Researchers would continue to face the difficulty of needing to apply different techniques to different platforms.

Some publishers have therefore developed their own specific Application Programming Interfaces (APIs) to ensure that their platforms can engage with text and data mining technology. For mining to be possible, and so as not to disrupt the integrity of the whole platform and degrade its performance, it is critical that miners adhere to the technical approach specified by the publisher.

However, not all publishers have the resources to develop their own APIs and so are dependent upon developing a common standard, one example of which is CrossRef.

CrossRef Prospect is already in development, leveraging existing CrossRef and publisher infrastructure to establish an automated, centralized, yet distributed, and efficient mechanism to allow researchers and publishers to agree to the terms of a standard text mining licence and to enable a standard cross-publisher mechanism for identifying and retrieving the full text of journal articles for text mining for non-commercial research purposes, from one single portal.

Researchers are able to select the publishers of interest from one single portal (Prospect) and, having accepted the relevant terms and conditions, the researcher receives a unique token (API key) for immediate use. This unique token identifies the researcher and their request(s) and acts as an identifier through which publishers can validate requests and make the content available to researchers for text mining.

This process is facilitated with a “click through” licence approach, which helps to strip the complexity out of the text mining process.

This process also allows mining to take place away from publishers’ own platforms, protecting platform stability and integrity whilst mining is occurring. The researcher’s API key also provides publishers with the certainty that their content is being provided to a bona fide researcher, in line with underlying subscription entitlements.

Crossref will roll its Prospect service out between November 2013 and first quarter 2014. Beta-version access should thus be available this year, facilitating the mining of a significant proportion of STM content and increasing experience of how mining can be streamlined for subscribed content and open access content alike.

Small Scale Usage

Often a researcher may not know which publisher owns the content they wish want to mine or how to go about contacting them to secure permission. Given the number, variety and location of publishers of research material, it can be quite a challenging undertaking for researchers. Publishers are working to resolve this issue, particularly for the ‘long tail’ of rights holders whose content may be less discoverable or well known.

(Advocates of a copyright exception point primarily to this problem as one which an exception would solve since not having to ask permission takes away the problem of obtaining it. However, as with the problems of different technological specifications, a blanket exception would go no way to solving the stated problem.

The Publishers Licensing Society (PLS), a collective management organisation representing publishers, holds an extensive database of publishers’ rights that can potentially be adapted and enhanced as an entry point or Clearing House for researchers to contact appropriate publishers using a common interface. PLS has started work to enable such a service, called *PLS Clear.*

PLS Clear is a digital clearing house which leads researchers through a "permission request form". This gathers basic information about the project and the content to be mined that publishers require in order to assess permission requests. It has been developed with the assistance of a group of leading publishers and researchers. PLS Clear then forwards the form automatically to the appropriate licensing manager for consideration. (PLS maintains an extensive database of publisher information) for consideration. The same form can be used to make requests of multiple publishers at the same time. It is currently being piloted and is expected to be available at the end of 2013.

The service will be open to all researchers, whether or not linked to a subscribing institution, and whether their research is for commercial or non-commercial purposes.

PLS is planning to develop a simple licence that can be used in cases where the researcher is not linked to an institution which already has a TDM clause in its subscription agreement.

PLS is also working on behalf of the publishing industry to reduce complexity in the licensing system, facilitating easier identification of rights owners and the licensing of content, through The Copyright Hub.

Both the Copyright Hub and the PLS Clearing House may in time develop the functionality to facilitate ‘click through’ licensing transactions, making it easier for researchers not just to identify rights holders, but to secure the relevant permission to mine it too.

Other Initiatives

The Copyright Clearance Centre (CCC) has been piloting a new Text and Data Mining service with users and publishers in the US, UK and EU since May 2013. The service makes it easy for commercial researchers to gain quick access to full-text content in a centralized manner with a common interface. CCC is currently piloting the service for the remainder of 2013 and, if successful, plans to commercialize the offering in 2014.

Features of this TDM service include the provision of a single source of XML full text content so that researchers can: download full-text from one place obtained directly from multiple publishers; search across subscribed and unsubscribed content from publishers to obtain the broadest possible set of results; keep track of content spending and integrating with library subscription holdings; obtain a common set of terms and conditions across publishers; and use customer-specific analysis and indexing techniques, which is a necessary functionality given the fact that most users have different requirements and different domain knowledge with respect to TDM.

Publishing Industry TDM Roadmap (as at November 2013)

Objective	Status
Develop STM standard licence clause for pharmaceutical customers	Objective achieved
Develop STM licence clause for non-commercial text and data mining	Objective achieved
Cross Ref Prospect Beta version launched	Objective achieved
Cross Ref final version launched	Work in progress – expected launch date Dec 2013 / Jan 2014
CCC pilot development and user testing phase	Work in progress
25% STM coverage participation rate in Cross Ref Prospect and/or insertion of STM non commercial clause in standard licences	Objective achieved
PLS Clear pilot	Work in progress – expected completion by December 2013
50% participation rate in Cross Ref Prospect and/or insertion of STM non commercial clause in standard licences	Work in progress – expected achievement date end of 2013
Copyright Hub beta phase launch	Objective achieved