Ligue des Bibliothèques Européennes de Recherche
Association of European Research Libraries

Response to Elsevier's text and data mining policy: a LIBER discussion paper

Over the last twelve months LIBER has devoted a considerable amount of effort to making the case for the need for changes to copyright legislation in order to allow researchers to employ digital research methods to extract facts and data from content.[1] We believe that this will exponentially speed up scientific progress and innovation in Europe. Having explored[2] the issue of TDM with our members and other stakeholders in the research community we have come to the conclusion that licensing will never bridge the gap in the current copyright framework as it is unscalable and resource intensive.

In the current vacuum left by a legal framework that is unfit for the digital age, and with the ensuing lack of legal clarity, it is unavoidable that libraries or researchers will have to agree to further licences for the mining of content to which they already have access. The terms of such licences, however, should be such that they reinforce the position that the right to read is the right to mine, and not impose restrictions on how researchers apply research methods or disseminate their research.

UK members should exercise particular caution when considering TDM licence terms, since an exception in UK law for text and data mining is imminent[3] and, dependent on the wording in this new exception, TDM licence terms may undermine what researchers will be permitted to do under this update to UK copyright law. Ireland is also considering such an exception.

This paper has been released in response to the recent launch of the new Elsevier text and data mining policy and API.[4] It is understood that Science Direct licences will be amended to include language around access for TDM. Many libraries may be considering signing, or have even already signed up to the terms and conditions laid out under this new licence.

Other publishers may also be considering following in the footsteps of Elsevier by introducing similar terms for the licensing of text and data mining activities into their licence agreements. LIBER is concerned that some of the licence's terms and conditions relating to content mining

[1] http://libereurope.eu/news/tdm
[2] http://libereurope.eu/news/workshop-report-from-the-perfect-swell-defining-the-ideal-conditions-for-the-growth-of-text-and
[3] http://www.ipo.gov.uk/hargreaves-copyright
[4] http://www.elsevier.com/connect/elsevier-updates-text-mining-policy-to-improve-access-for-researchers

may be unnecessarily restrictive and that systematic and widespread adoption of such terms and conditions will severely hamper the progress and dissemination of data-driven research.

Below, we analyse the elements of the Elsevier policy and the summarised version of the terms and conditions of its TDM licence, as well as some of the terms contained in the click-through licence for researchers.

The institutional licence agreement for text and data mining

In order for a researcher within a subscribing institution to gain access to Elsevier content for the purpose of mining, it is necessary for the institution to update their licence agreement to allow text mining access. Note that within this agreement "text mining access" does *not* mean access to the content on the Elsevier Website that universities subscribe to. Access to content for the purpose of mining is limited to access via an API. The licence explicitly prohibits the use of robots, spiders, crawlers or other automated programs, or algorithms to download content from the website itself, which are the most common ways of performing content mining. Although the new Elsevier policy claims that it "enshrines text- and data-mining rights" in subscription agreements, in reality, under these terms, it compels institutions to agree to very restrictive conditions in order to gain very narrowly defined "access" to content for the purpose of mining.

Access via an API

An application program interface (API) is a set of programming instructions and standards for accessing a web-based software application. In the case of the API offered by Elsevier, the API provides full-text content in XML and plain-text formats. The use of APIs for the mining of metadata is not uncommon. However, article content is much richer, potentially containing images, figures, interactive content, and videos. For researchers in many different disciplines there is as much value in the images and figures contained in the article as there is in the text. In fact, for researchers in disciplines such as the humanities, genetics, chemistry, these may be the most valuable content elements. The Elsevier API allows access to the **text only*.*** And the access limit is an arbitrary and proportionally tiny 10,000 articles per week.

Crucially, researchers develop their own tools for handling and exploiting this rich and diverse variety of content and formats. In order for students and academics to be able to perform research freely, in the way that makes sense for their own studies, they must have the freedom to interrogate, query and structure content in ways that fit with their own needs, technologies and requirements. The requirement to use pre-defined publisher technologies hampers academic freedom, learning, and data driven innovation.

Even for those researchers for whom the API is sufficient, the licence does not guarantee sustained access to the API, as the following clause indicates:

3.4 Elsevier reserves the right to block, change, suspend, remove or disable access to the APIs and any of its services at any time.

Use of robots

The Elsevier policy expressly forbids the use of robots for content mining on the grounds that it would place too much strain on their infrastructure. Open access publishers, whose infrastructure is exposed to all web users on the open web, have reported that the demand placed on their infrastructure by robots for content mining is negligible and any increase in demand will be easy to manage.[5] For subscription services such as those provided by Elsevier, the demand placed on their infrastructure should be even less, as only users registered at subscribing institutions will have access.

Control of outputs

Under the terms and conditions of the updated licence agreement the outputs are controlled in the following ways:

1. Outputs can contain "snippets" of up to 200 characters of the original text

   This is an arbitrary limit. Because this is essentially a limit on the amount of text that can be quoted from the original source, it could potentially result in misquotation or, at the very least, an inaccurate representation of the original research.

2. Licensed as CC-BY-NC

   In signing up to the Elsevier licence agreement, researchers are asked to agree to make their output available under a CC-BY-NC licence. The outputs of TDM are very often facts and data, which are not subject to copyright; however, the Elsevier licence agreement stipulates that this non-copyright information should be put under a licence for copyright works.

   In addition, the definition of "non-commercial" is highly ambiguous and open to interpretation. In effect, a CC-BY-NC licence prevents downstream use of the results and may also put researchers who are performing research under a grant agreement that mandates that data be openly available in a difficult position. Universities are also increasingly engaging in, and being encouraged by governments to enter into business partnerships with, private business. This is known as the "knowledge transfer agenda". We recommend that universities and researchers decide before signing the Elsevier licence whether there is a possibility that the outputs of the research they wish to undertake are commercial. As facts and data are not copyrightable, LIBER's position is that they should be made available under a CC0 licence.[6]

[5] http://blogs.plos.org/opens/2014/03/09/best-practice-enabling-content-mining/

[6] This licence is recommended so that reuse is not prevented under the sui-generis Database Directive.

Registration and click-through licences

In order for an individual researcher to gain access to the Elsevier content that their institution subscribes to, he/she must register directly with the Elsevier developers portal, provide details about the research they wish to undertake, and agree to the terms of a click-through licence. LIBER is particularly concerned about making such demands of researchers for the following reasons:

1. We want to protect the privacy of our users.

   Libraries have a strong track record of putting measures in place to protect the personal details and reading habits of our patrons. By requiring researchers to register individually and to provide details of their research project, Elsevier is circumventing the protections that libraries have put in place. The reason given by Elsevier for this requirement is that the publisher needs to check the credentials of the individual accessing the content. However, in authenticating individual user accounts the institution has already established the bona fide nature of the researcher. Further verification should not be necessary. We object to data about the research being performed by our users in our institutions being collected by an external third party. It is not the job of a publisher to control, monitor and vet what research takes place at a university.

2. We want to protect our researchers from undue liability.

   Many institutions employ full time experts to negotiate the terms and condition of licence agreements on their behalf. This process can take months, and yet, a researcher is expected to agree to the Elsevier click-through licence in a matter of seconds. The terms of this click-through licence[7] are extremely complex, in many places unclear[8] and could have **serious down-stream implications** for the outputs of the research. We also note that there is no cap on liabilities for a researcher:

   2.3 The User will be solely responsible for all costs, expenses, losses and liabilities incurred, and activities undertaken by the User in connection with TDM Service.

   What is more, Elsevier retain the right to amend the terms, without notice and the changes will be deemed accepted by the researcher immediately. This is unacceptable.

   Many of the responsibilities that are placed on the researcher by the click-through licence will be **difficult to implement in practice e.g. the licence states that copyright notices may not be changed from how they appear in the dataset**. This means that in a dataset derived from 10,000 articles there may be at least 10,000 appearances of the word "copyright". A normal way of dealing with this "noise" would be to remove these irrelevant data from the dataset, but this would contravene the terms of the licence.

[7] http://www.developers.elsevier.com/cms/content/text-and-data-mining-service-agreement

[8] Terms used in the licence such as "recognition" and "classification" (2.1.1) are unclear. Another crucial, term "integration" (3.3) has been left undefined.

The click-through licence also makes it impossible to ensure the **transparency and reproducibility** of research results as the researcher may not share the dataset used for the research project and must delete it after use. The researcher is also expressly prohibited from depositing this dataset in their institutional repository.

Lastly, the licence is silent on post-termination use of the results of content mining. The licence will be terminated if the subscribing university *“does not maintain a subscription to the book and journal content in the ScienceDirect® database”.* If a researcher has mined thousands of articles, how do they check that each and every one is being subscribed to? If one or many are cancelled, what does this mean for the results, categorisations and hypotheses contained in data they have invested time and effort to produce?

Outlook

We estimate that European universities spend in the region of €2 billion a year on Scientific Technical and Medical published content, the vast majority of which is on e-journal subscriptions. The new Elsevier licence terms and added requirement of an additional licence for each and every researcher who wishes to mine the content raises questions about what institutions are actually purchasing when subscribing to digital information. The implication of the Elsevier TDM policy is that institutions only purchase the right to cache, look at, print out, and do a word-search on a PDF. We believe that universities should be able to employ computers to read and analyse content they have purchased and to which they have legal access. An e-subscription fee is paid so that universities can appropriately and proportionately use the content they subscribe to. For what other purpose is a university buying access to information?

Research and innovation is best encouraged in a free-thinking and enabling environment where researchers can fully exploit the content they have access to through their library. Going forward, it is important that libraries can ensure that the scientific freedom of their researchers is not eroded, and the impact of their scientific outputs undermined, by limits imposed through licences.

28/03/14