Aggregating and Enriching Publications to Improve the Research Process

Commercial use (Y/N): N
Related project (*if any – please specify funder and project name*):
Author: Petr Knoth
Organization: Knowledge Media institute, The Open University, United Kingdom
Date: 19/12/2014

Use case description – *Describe the problem you are trying to address, the research area it affects, and the TDM application used. Please state all participants, whether and how the results were subsequently used.*

COnnecting REpositories (CORE) is a not-for-profit service run by the Knowledge Media institute, Open University. Our research into aggregating and text-mining of research papers, supported by a range of funders including Jisc and the European Commission, has resulted in the creation of a platform with a number of applications built on top of it, providing benefits to a range of stakeholders and the general public.

CORE contains 20m+ open access research papers from worldwide repositories and journals, on any topic and in more than 40 languages. In a single month CORE records 500,000+ visits from 90,000+ unique visitors. By processing both full-text and metadata, CORE serves three communities: (1) developers, text-miners, scientometricians, etc. who need large-scale machine access to research papers, (2) researchers and the general public who need better paywall-free access to research literature and (3) funders and government organisations needing to discover scientific trends and evaluate research impact.

As part of this work, CORE heavily enriches the aggregated research publications with the use of text and data mining (TDM). The goal is to extract useful information and knowledge from the publications. This can sometimes only be done at the level of the aggregation and cannot be effectively realised at the individual data sources. The main applications of TDM include:

- **Information extraction** from full-texts of research papers, including named entity recognition, ranging from basic metadata (authors, titles, institutions, emails, URLs) to more advanced metadata, such as names of scientific methods, names of species, locations or opinion phrases/sentiment.
- **Extraction of citations** from full-texts of research papers and identification of unique identifiers of these publications (such as DOIs). This leads to the creation of an open citation dataset.
- **Content recommendation** of related research papers based on full-texts.
- Matching of papers to patents, funding opportunities, open courses, etc. to support a range of stakeholders.
- **Mining the licence** of research papers to determine if they are compatible with the OA definition.
- **Supporting scientific knowledge discovery**[1] by improving access to research literature.
- **Text-categorisation** of papers to determine the subject class and allow the monitoring of research trends.

[1] http://polaris.gseis.ucla.edu/gleazer/296_readings/swanson_undiscovered.pdf

CORE offers applications for a number of user groups. They include the CORE Portal, which provides a faceted search interface for open access research papers, CORE API & Data dumps, which offer unrestricted machine access to the aggregated and semantically enriched research literature available for TDM, and the CORE Recommendation Plugin which, following the ideas of exploratory search, proactively assists researchers in discovering related materials across repositories and journals. CORE has become the Jisc UK Aggregator of research papers and has been listed among Top 10 Search Engines for Research that Go Beyond Google[2]. It has been used by many researchers as a source of valuable data for text-mining and has also been integrated with journal, repository and library systems, including the The European Library portal.

Content sources – *List the targeted content sources and types of content to used.*

CORE harvests primarily Open Access (OA) research papers from repositories (institutional, subject-based) as well as research papers from publishers' systems. At the moment, CORE makes use of the UK TDM Copyright Exception to, for example, to text-mine articles in order to identify their licence. This is due to the fact that many publishers' systems do not provide this information consistently in the metadata.

The content sources of CORE currently include 600+ repositories worldwide and 10,000+ journals. The sources include large subject-based repositories, such as PubMed and arXiv. Overall, there are over 20 million records stored in the CORE database with full-texts enriched using TDM (provided when full-texts are available from the source). The content includes a wide range of research output types, including accepted research papers, preprints, postprints, monographs as well as master and doctoral thesis.

Targeted users – *Describe end users, their number and expertise.*

CORE serves three communities:
(1) commercial and non-commercial services developers, text-miners, scientometricians, etc. who need large-scale machine access to research papers.
(2) researchers and the general public who need better paywall-free access to research literature
(3) funders and government organisations needing to discover scientific trends and evaluate research impact.

The CORE Portal currently registers 500,000+ visits from 90,000+ unique visitors per month. The CORE API and Data Dumps have been used by over 100 text-miners/developers many of which have developed new services on top of CORE or managed to publish new papers using the data provided by the service. The CORE Content Recommendation Plugin has been used by a number of organisations including UNESCO and the European Library. CORE is well integrated with the UK repository infrastructure (being the national aggregator) and is going to be used to provide access to UK content in collaboration with OpenAIRE2020.

2 http://www.jisc.ac.uk/inform/inform37/SearchingBeyondGoogle.html?utm_medium=email&utm_source=JISC+Executive&utm_campaign=2793129_Jisc+Inform+37&utm_content=searchingbeyondgoogle&dm_i=QHI,1NV6X,9ZS4RV,5U7J3,1#.VJQH_AUt

Impact – *Describe all possible impacts of the use case. Specify any cross-border, societal and economic effects (if possible state any monetary benefits and market advantages) .*

- **Increasing TDM efficiency and reducing costs.** About 90% of text-miners' time is spent in gathering data; CORE enables quick, cost-free access to data for text-mining projects. This can dramatically increase efficiency of text-miners and provide better return on investment in TDM research.
- **Free data access sustainability for TDM services**. CORE constantly updates its database, offering a free data sustainability option for TDM services developed on top of the CORE API.
- **CORE supports businesses relying on TDM**: For example, William Cullerne Bown, the CEO of a London based company ResearchResearch which used CORE to develop a text-classification system for funding opportunities, said: *`The CORE team understand data mining. As an independent company, we had no obvious access to big, diverse scholarly data - a killer in our drive to develop classification algorithms. The CORE repository, available in bulk, was a breakthrough. Now our algorithms outperform even those from huge publishers.'* CORE has also attracted a donation from an Italian company Ciaotech S.r.L. which has used CORE to find prior art for patents as part of a tool to support innovation according to the TRIZ methodology.
- **Enabling reproducibility of large-scale TDM research.** CORE offers unrestricted machine access to OA research publications for text-miners. This means that studies carried out on the CORE dataset can lead to reproducible research (shipping data with the SW), thus providing better return on investment for funders.
- **Enabling business intelligence and decision support using TDM.** CORE enables development of analytical services and decision support systems on top of the data, such as applications to monitor compliance with open science policies or applications to analyse growth and trends in scientific research to direct funding.
- **Enabling new research on impact metrics.** CORE enables research on developing new research impact metrics, such as Semantometrics[3], that are not based on the false premise of linking quality with the number of interactions in the scholarly communication network, such as Bibliometrics, Webometrics and Altmetrics. New research metrics that go beyond citations are needed to make research more efficient, quality oriented, collaborative rather than competitive and enable the full transition to open access.
- **Enabling large-scale TDM research using publications as corpora.** CORE enables research on a large number of TDM applications including exploratory searches, information extraction, link discovery, text-classification, data science and data archiving.

[3] http://www.dlib.org/dlib/november14/knoth/11knoth.html

Constraints – *Describe any legal, technical, economic, societal, organizational, cultural, multi-lingual or other limitations and you have encountered.*

- **Accessibility of Open Access content on publishers' websites –** Some of the major commercial publishers deny machine access to their OA content using the Robots Exclusion Protocol[4]. However, large commercial organisations, such as Google and Microsoft, typically have special conditions on machine access to these OA materials. We are currently investigating whether this approach violates the competition law.
- **Text-mining sometimes needed to determine licence –** In many cases, we are required to text-mine the full-text content in order to determine the licence of the content. The licence is typically not provided as part of the metadata. We are currently depending on the recent UK Copyright Exception for text-mining[5] to do so, but a European-wide approach would be helpful.
- **CORE could extend some of its functionality to non-OA content if adequate exceptions were available.** Our system could provide much wider benefits if a TDM exception for research papers, regardless of their licence, would be granted.

Comments – *Other comments and any recommendations you may have.*

It should be recognised that many commercial organisations have been text-mining research papers for years as they decided to undertake the risk. They include large organisations, such as Google and Microsoft, as well as innovative and succesful start-ups, such as Mendeley and Linguamatics. However, universities and non-commercial organisations are more risk-averse.

We do not support the view that these innovative commercial organisations should be punished for their activities, but advocate for making it clear that TDM of research papers and associated materials does not breach the law so that universities and non-commercial organisations could also participate.

We also strongly suggest that providers of content should not discriminate between services with respect to the right of crawling content from their website. Clear rules on machine access that universally apply to all search engines and aggregators should always be established. No system should receive an unfair advantage, as this limits the ability of new players to enter the market, and decreases the competition.

---

[4] http://core-project.kmi.open.ac.uk/files/oa-metadata-to-oa-content.pdf Section 5.
[5] http://www.legislation.gov.uk/ukdsi/2014/9780111112755