Releasing 1.8 million open access publications from publisher systems for text and data mining

Text and data mining offers an opportunity to improve the way we access and analyse the outputs of academic research. But the technical infrastructure of the current scholarly communication system is not yet ready to support TDM to its full potential, even for open access outputs. To address this problem, ***Petr Knoth****,* ***Nancy Pontika*** *and* ***Lucas Anastasiou*** *have developed the CORE Publisher Connector, a toolkit service designed to assist text miners in accessing content though a single machine interface. The Connector aims to solve the heterogeneity among publisher APIs and assist text miners with data collection, provide a centralised point of access to all openly available scientific publications, and provide a high-performance, constantly updated access interface.*

Research literature contains some of the world's most vitally important information; such as how to treat diseases, solve difficult engineering problems, and answer many of the global challenges we face today. The entire body of research literature is currently estimated at 100-150 million publications, with an annual increase of around 1.5 million. Systematically reading and analysing the full body of knowledge is now beyond the capacities of any human being.

While text and data mining offers an opportunity to improve the way we access and analyse the outputs of research, the technical infrastructure of the current scholarly communication system is not ready to support TDM to its full potential. Furthermore, the legal framework in most countries imposes limitations on mining research literature, at least for non-research purposes. Little attention has been given, though, to the technical issues preventing large-scale analysis of research articles. Surprisingly, mining even open access research papers is problematic for technical reasons.

Open access and text mining of research papers have one thing in common: both aim to improve access to scientific knowledge for people. As a result, text mining is performed in large corpuses of text. In fact, many of the text mining tasks, such as semantic search, recommender systems, question answering, or content summarisation, are only able to realise their full potential when run on an as large a corpus of publications as possible. This means that text miners must typically invest considerable time, effort, and resources in collecting their corpus of interest. Sometimes, this task may prove impossible due to the technical restrictions and limitations of publisher platforms. According to a 2014 Jisc report, text miners can spend up to 90% of their total investigation time on the data collection.

To eliminate these extra steps and save time and money for text miners we have developed the CORE Publisher Connector, a toolkit service designed to assist text miners on accessing content though a single machine interface.

Development of the Publisher Connector was not easy. Though one might think the machine accessibility of open access scientific outputs from publishers' interfaces would be relatively straightforward, we found it to be quite the opposite. After conducting a survey of the largest open access and hybrid publishers – which included questions on: a) their machine interface; b) how it is possible to access full text; and c) whether they impose any download restrictions – we soon realised the complexity of the issue. Not all publishers support well-adopted standards for aggregation, such as the OAI-PMH, usually retracting to custom proprietary API implementations. We also discovered issues when linking a publication's metadata with the full text, leading to ambiguity when machines attempt to aggregate content and make a metadata and full text match. To further compound the situation, some publishers do not support a machine interface on their websites at all. More specifically, they offer content on their websites only, typically in combination with restricting robots, with only Google and a handful of others permitted to access it. This prohibits any interaction with aggregators, or tasks related to content extraction via computer agents.

Date originally posted: 2018-03-22

Permalink: http://blogs.lse.ac.uk/impactofsocialsciences/2018/03/22/releasing-1-8-million-open-access-publications-from-publisher-systems-for-text-and-data-mining/

Blog homepage: http://blogs.lse.ac.uk/impactofsocialsciences/

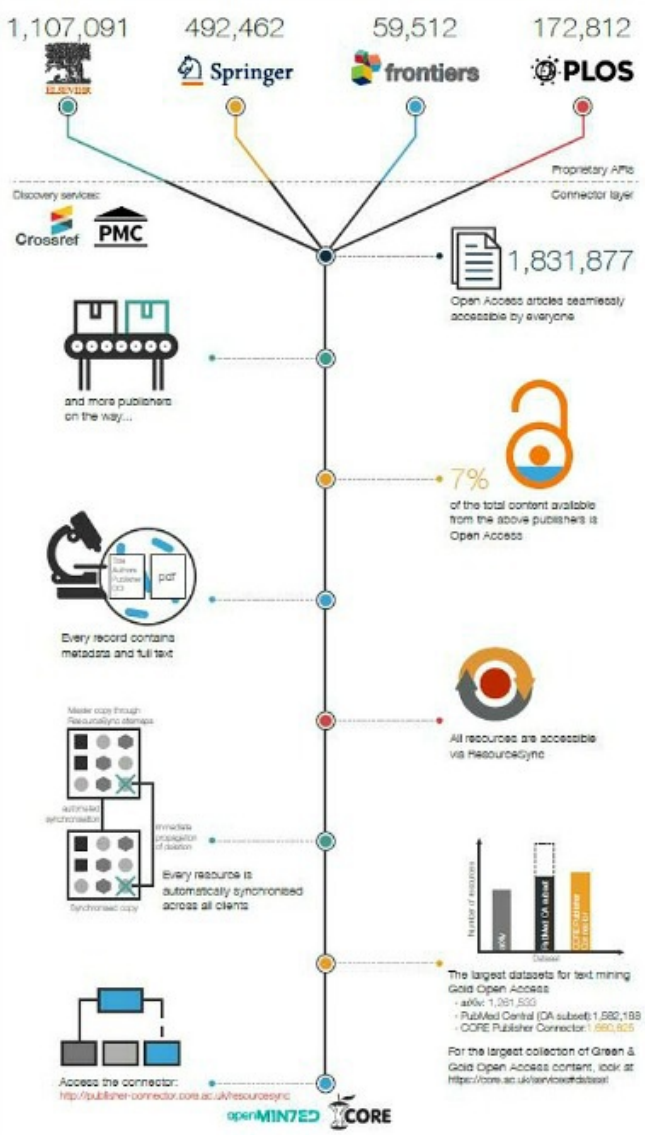

Figure 1: CORE Publisher Connector. A full-size, high-resolution version is available here.

In response to these survey findings, we launched a new infrastructure, the CORE Publisher Connector. The Connector consists of several software modules, with each able to harvest content from a given publisher making it available to the public for download. The synchronisation is conducted using the ResourceSync protocol. When compared to OAI-PMH, which provides interoperability for harvesting metadata only, ResourceSync permits the sharing of any kind of resource – i.e. both metadata and the actual data – and offers an advanced synchronisation mechanism over the web. The resources gathered by the Connector are made available via the ResourceSync protocol here.

We have so far released more than 1.8 million full text, open access articles. To further increase the coverage of the Publisher Connector, we plan to add more software modules from several other publishers.

Date originally posted: 2018-03-22

Permalink: http://blogs.lse.ac.uk/impactofsocialsciences/2018/03/22/releasing-1-8-million-open-access-publications-from-publisher-systems-for-text-and-data-mining/

Blog homepage: http://blogs.lse.ac.uk/impactofsocialsciences/

Publisher	Volume of open access articles
Elsevier	1,107,091
Springer	492,462
Frontiers	59,512
PLoS	172,812
TOTAL	1,831,877

Table 1: Breakdown of open access articles per publisher available in the Publisher Connector as of December 2017.

The aim of the CORE Publisher Connector is to:

- *Create a seamless layer for accessing content across publishers*: the Connector attempts to solve the heterogeneity among publisher APIs and assist text miners with data collection
- *Provide a generic, centralised point of access to all available resources***:** this is a large corpus of millions of open access scientific publications
- *Provide a high-performance, up-to-date access interface:* the corpus will be constantly updated to easily surface open access scientific literature.

In this initial release we have developed connectors to four leading publishers – Elsevier, Springer Nature, Frontiers, and PLoS – and investigated how their APIs work, comparing our findings with the information provided on their respective websites. In an expertise directory, we present our aggregation approach, each publisher's features – including limitations and solutions – and propose improvements to the publishers' existing machine interfaces. The directory also contains the code used for the implementation of the Publisher Connector. Notably, we realised that the technical information we received from some publishers during the survey was not fully consistent with reality. For this reason we are yet to develop connectors for other key publishers. However, and to our surprise, our work has so far been warmly received; with some publishers, including Taylor & Francis and Copernicus, asking to opt in after seeing presentations about the Publisher Connector at conferences.

The release of over 1.8 million of open access articles made available via the CORE Publisher Connector introduces a new era in the open access field. We are now in position to easily estimate the number of gold and hybrid open access publications per publisher and monitor future growth. For example, from those publishers we have investigated so far, 7% of the total published articles are available as gold or hybrid open access. Additionally, we have defined the architecture of an effective workflow of harvesting publishers and a good start is half the battle won. We now offer implementations of generic retrievers, coverage of which will be increased in the future, since the process for implementing new modules for each publisher is now more straightforward. With this work we also aim to motivate publishers to create a more unified and less heterogeneous system; one which would be less time-consuming and perplexing, and remove obstacles to text mining practices. In general, our contribution aims to offer a worry-free interface for text miners, enabling them to concentrate on the content, rather than the peculiarities of how to access it.

This work is innovative in two ways: it constitutes the first systematic aggregation of gold and hybrid open access content from key publishers; content that aggregators like CORE and OpenAIRE have not been harvesting so far. This work liberates nearly two million papers from key publishers for a range of activities, including TDM. It is also the first deployment of ResourceSync as an effective technology for distributing a large corpus of scholarly literature.

This work has been conducted by the team developing the CORE aggregator within the OpenMinTeD project, which aims to create an e-infrastructure for text and data miners across Europe. CORE is a global aggregation service which has harvested over 83 million metadata records from more than 3,600 data sources and tens of thousands of journals. The majority of these records contain links to article full texts. In addition, as CORE has already ingested the content from the Publisher Connector, CORE is now directly hosting more than 10 million open access full texts, making CORE the world's largest full text aggregator. The full texts are available directly from CORE via the CORE API (REST and ResourceSync) and downloadable as a dataset.

Date originally posted: 2018-03-22

Permalink: http://blogs.lse.ac.uk/impactofsocialsciences/2018/03/22/releasing-1-8-million-open-access-publications-from-publisher-systems-for-text-and-data-mining/

Blog homepage: http://blogs.lse.ac.uk/impactofsocialsciences/

Featured image credit: iPhone unlock by Elaine Smith, via Flickr (licensed under a CC0 1.0 license).

Note: This article gives the views of the authors, and not the position of the LSE Impact Blog, nor of the London School of Economics. Please review our comments policy if you have any concerns on posting a comment below.

About the authors

***Petr Knoth** leads an R&D team working in the domains of text mining, digital libraries, and open access/science. He is the founder, product and team leader for CORE, a service that aggregates millions of open access articles from around the world and makes them available for people to search and machines to text-mine. Previously, he worked as a Senior Data Scientist at Mendeley on information extraction and content recommendation for research. He has a deep interest in the use of AI to improve research workflows. He has co-founded Semantometrics.org which aims to go beyond bibliometrics and altmetrics to produce new research evaluation methods that make use of the publication full texts in research assessment. Petr has been involved as a researcher and as a PI in a number of European Commission, national, and international-funded research projects in the areas of text mining, open science, and eLearning. His ORCID iD is 0000-0003-1161-7359.*

***Nancy Pontika** holds a PhD in Open Access from the School of Library and Information Science at Simmons College, Boston. Her main areas of interest are Open science, responsible research and innovation and the role of text and data mining in scientific literature with respect to openly accessible research outputs. She advocates for pure open access, machine access to open access research papers, and the promotion of open science for the advancement of research. For the time being she is an Open Access Aggregation Officer in CORE, a service that aggregates millions of open access research papers, while in the past she worked for the Repositories Support Project (RSP) and as a repository manager at Royal Holloway, University of London. She is currently involved in three European projects, FOSTER, OpenMinTeD, and FIT4RRI. She serves as an Editor at the Open Access Directory, as an External Liaison Officer at the UK Council of Research Repositories, and tags for the Open Access Tracking Project. Her ORCID iD is 0000-0002-2091-0402.*

***Lucas Anastasiou** is an Open Access Publishing Project Officer. He is a graduate of National Technical University of Athens (Electrical and Computer Engineering) and University College London (Information Security). He has been employed by the Knowledge Media Institute where he has been involved as a research assistant in various academic projects; e.g. LTfLL, Stellar, CORE, eCloud, and OpenMinTed. He is a strong advocate and enthusiast of open access and open science. His ORCID is 0000-0002-1587-5104.*

Date originally posted: 2018-03-22

Permalink: http://blogs.lse.ac.uk/impactofsocialsciences/2018/03/22/releasing-1-8-million-open-access-publications-from-publisher-systems-for-text-and-data-mining/

Blog homepage: http://blogs.lse.ac.uk/impactofsocialsciences/