FutureTDM
Explore . Analyse . Improve

ORGANISATIONS

IMPROVING UPTAKE OF TEXT AND DATA MINING IN THE EU

Facts

Project No: 665940

Program: H2020 | CSA | GARRI

Duration: 09/2015 - 08/2017

TDM and ContentMine

Barriers and Enablers of TDM

ContentMine

What is ContentMine?

ContentMine is a non-profit organisation founded by Dr Peter Murray-Rust, a chemist, molecular informatician and advocate for open science. Murray-Rust faced barriers throughout his career in trying to apply his TDM technologies to the scientific literature.

In 2014 the south African philanthropic funder Shuttleworth Foundation supported him with a two year fellowship to set up the ContentMine project, which initially sought to liberate 100 million 'facts', mostly named entities, from scientific literature. The project also ran TDM training workshops for researchers to promote the usefulness of TDM to researchers facing overwhelming levels of content, reaching around 300 researchers at over 20 workshops. Talks by Murray-Rust and the ContentMine team reached an estimated audience of 2000, promoting the concept of content mining (as a more inclusive term than TDM) and its utility across a wide variety of disciplines.

Murray-Rust was heavily engaged in advocacy for the idea that 'the right to read is the right to mine', a phrase that was later picked up by organisations such as the Wellcome Trust and LIBER in their advocacy and policy work around TDM aiming to give subscribers to scientific articles the right to read them using a machine without seeking additional permissions.

What is the aim of the project?

The major aim of the project and resulting non--profit was to set up a daily feed of 'facts' by accessing a high proportion of the full-text scientific literature via publisher and content providers' application programming interfaces (APIs) and by scraping from websites where necessary. Initial efforts focused on the Open Access literature but the introduction of a UK copyright exception for TDM for noncommercial research in 2014 reduced some legal barriers to use of the closed access literature and the project is now planning to implement a daily pipeline of open data in the form of species names, word frequency data, human genes and other facets in collaboration with librarians at the University of Cambridge.

This project has received funding from the European Union's Horizon 2020 (H2020) Research and Innovation Programme.

10,000
new academic papers
1

6 researchers
as ContentMine Fellows
2

>100 researchers
trained in TDM
3

>100 scrapers and stylesheets
to mine 90% of the literature
4

1 *10,000 - the average number of new academic papers that the ContentMine TDM pipeline will process on a daily basis.*
2 *6 early career researchers using TDM have been supported as ContentMine Fellows*
3 *Over 100 researchers have been trained in TDM through ContentMine workshops*
4 *Mining 90% of the literature would require producing >100 scrapers and stylesheets to obtain and normalise the content*

"The days of manually searching through thousands of academic papers are now gone"

ContentMine

What are the barriers and enablers of TDM for ContentMine?

Technical and Infrastructure

Technically, the barriers reported by ContentMine are related to the heterogeneity of publisher XML and HTML, even when it conforms to a technical standard such as NISO JATS. In order to produce a normalised corpus of articles for easier semantic tagging, custom web scrapers and XML style sheets must be constructed on a publisher by publisher basis, a challenging task for an individual researcher or group. Members of the team have also found multiple instances of publisher barriers such as captchas to prevent bulk downloads and 'traps' such as fake DOIs, which alert the publisher to mining activity or in some cases automatically cut off access from the relevant IP range.

Legal and Content

The legality and ability of researchers to challenge the technical measures is unclear even under the UK exception and represents another barrier beyond statutory legal barriers. Nonetheless, a copyright exception that cannot be overwritten by contract was viewed by ContentMine as a major enabling factor.

Economy and Incentives

Although the organisation is based in a country where a statutory law allows non-commercial use and it is a mission-driven non-profit, finding income streams to remain sustainable and develop software without relying on public funding is challenging without an allowance for commercial use.

Education and Skill

A lack of awareness of TDM was a barrier to the work of organisation. It was clear from discussions between the training team and workshop participants that many researchers lack the skill base to work with highly technical or command line tools and there is a gulf between the types of techniques and protocols they are used to applying and the approaches typically taken by academic TDM groups, who get academic credit for the quality of the mining rather than user interface design. Many groups were doing large scale literature reviews entirely manually at the great expense and effort and the learning curve was substantial barrier regardless of legal status.

ContentMine exemplifies the types of research activities that have been positively enabled by removal of legal barriers but are still impeded by non-legal factors and threatened by lack of sustainable funding models even in a non-profit context.