How to Mine Millions of Articles: a three stage solution to successful TDM

What is Text and Data Mining?

Text and data mining (TDM) can help answer specific research questions by uncovering trends and patterns through a process of searching, extracting and analysing large amounts of information, using computers. It's not just simple search: it's like building your own search engine from scratch and using it to discover trends and patterns within the results that would not otherwise be possible.

Why do Text and Data Mining?

TDM has the potential to drive further research and innovation in Europe. It uses advanced software that allows computers to read and digest digital information far more quickly than a human being can. With over 2.5 million articles published each year[1], it has the potential to help researchers accelerate new discoveries.

What do we need to do Text and Data Mining?

Quality content
Publishers support TDM by providing access to quality content and working with researchers to overcome technical challenges.

The right skills
Content mining is a growing trend requiring specific skills to get the most out of it. Uptake depends on education, awareness, tools and infrastructure.

Technical solutions
Software which “crawls” websites to select and download content can affect site performance, so mining access must be managed to make sure that the experience of other users is stable. Further technical knowledge and software is then needed to mine the content for information.

Good user experience
As a result, many publishers have developed APIs which provide direct access to data, providing the best possible experience for all users.

Collaboration
TDM requires publishers, researchers, librarians and the wider community to work together. We all have a role to play in making content mining efficient and effective.

Investment
Successful and seamless TDM requires access to quality content and technical expertise. Investment by publishers helps make this happen and users need to work in line with copyright for its long term sustainability.

[1] The STM Report, March 2015, http://www.stm-assoc.org/2015_02_20_STM_Report_2015.pdf (p 6)

Three stage solution to successful Text and Data Mining

1. Tackle the technical challenge

Uptake of TDM is growing, but still modest, with 15% of publishers receiving requests in 2014[2]. Researchers need to be equipped with the right tools, technology and best practices which will enable TDM to be effective and successful. It's not a question of being allowed to mine, it is about having the technical skills and support to be able to do it.

2. Continue developing the right infrastructure

There are a number of services that have already been developed to enable stable and sustainable TDM. These include:

- **CrossRef Text and Data Mining API** provides a standard API that works across thousands of publishers.
- **Copyright Clearance Center XML for Mining Service** provides centralized access to normalized licensed full text XML from multiple publishers.
- **PLSClear for Text and Data Mining** helps researchers identify rights holders and communicate project requirements efficiently.
- **JSTOR Data for Research** is a free service for researchers wishing to analyze content on JSTOR through a variety of lenses and perspectives

Researchers, librarians and publishers will continue to collaborate to make content mining work.

3. Make it long term and sustainable

Permissions contained within licences provide clear, fast access and rights to content for text and data mining. Publishers already offer a working solution based on APIs. These provide stable options allowing computers to access, download and mine content without disturbing our human users!

Use case Hedge fund managers and algorithmic traders buy licenses from traditional content sources like news agencies and publishers to gain access to breaking stories. Traders then use TDM software to mine those feeds to predict movements of markets for everything from government bonds to commodities. For example: subtle shifts in weather patterns might predict a downturn in the price of wheat, oil or gas.
Use case A text miner needs to copy 15 million articles indexed in PubMed in the last 20 years. How can the content be stored safely? And what reassurances are there that the content will not be redistributed and republished for commercial gain? Using a supported API enables access to content from thousands of publishers. Tools such as *PLSclear* help to identify rights holders to communicate the scope of research.

[2] Publishers Licensing Society, August 2015 http://www.pls.org.uk/news-events/n-tdm-august-15/