TEXT to go with TDM Prezi - https://prezi.com/4h9mbv8hda2o/text-and-data-mining-tdm/

This Prezi presentation provides an overview of what Text and Data Mining (TDM) entails and how STM publishers support it.

What is TDM ?

TDM is a computerised tool that can analyse vast amounts of content – bigger volumes than any human eye or brain could process. Like a bulldozer it crunches through documents, texts and information, deconstructing it into data and reconstructing it into new patterns and new relationships. It can find new knowledge that was previously hidden in single articles, and that could only be discovered by mining thousands or tens of thousands of articles and combining their contents.

TDM is only as good as its content: “garbage in is garbage out”. So in the STM world it is very important that content is of high quality; ideally content that is formatted properly, normalised and harmonised and with metadata added. Without high quality content, TDM outcomes can be very unreliable.

TDM is different from search. Search is used to find and select information. But TDM uses and processes the information –deconstructing and reconstructing it by crunching the information into data from which new patterns can be discovered. In order to do that, a miner must select hundreds or possibly thousands of articles, copy them onto their machine (or save to the Cloud) and run their TDM software through the content. Although technologies vary, copies are made frequently.

In this way new patterns and new relationships can be found and research accelerated.

We all benefit from TDM. Research benefits directly, because computers can process information that is too big in volume for the human eye or brain. Society benefits, because research develops quicker and provides more direct results. TDM can also find increasing civil applications.

High Quality Content necessary

We need to stress the point that for high level TDM, like researchers, we need high quality content. Compare it to a fancy race car – see the Ferrari; if you do not put high quality fuel in it, it will not run efficiently. STM publishers are the content providers for TDM researchers, we provide high quality fuel to make that car (machine) run. STM publishers have made considerable investments to ensure that their content can be properly mined, and for non-commercial users, have committed to including the rights to mine as part of their licenses with their customers at no additional fee.

What else do STM publishers do to support TDM? We provide TDM access to researchers at no extra costs. TDM is included in their licenses.

How researchers use TDM: TDM is only one small step in a much longer research workflow. It sits between literature search, desk research, experiments and lab tests and further analysis. It is one of the many tools that researchers use when undertaking their research.

TDM has a powerful potential across many different disciplines. Originally and still most frequently applied in biomedical research, drug research, and banking it is moving into chemistry, physics, and even law and social sciences.

TDM is a new and promising technique – but, in many cases it is still under development. To apply TDM well, it often needs many rounds of perfecting and tweaking. An important part of the software is often based on so-called machine-learning and the content often needs thorough normalization and harmonization before the software can run properly. The machine can be as sensitive as a Formula 1 race car.

If applied wrongly and without these enhancements, TDM can give very unreliable results. If the content that was used was not high quality, not curated, not cleaned up and preprocessed in the right way, the outcomes can be untrustworthy. Think of the predictions that went wrong for the US elections, when pollsters used incomplete data, data harvested from Internet, Facebook and Twitter. Suddenly everyone was wrong, the polls, the press and the politicians.

Scholarly Content perfect to Mine

STM publishers provide high quality content that is very well suited for TDM. It is curated, digitally available, well structured, has ontologies and taxonomies, is diverse in the sense of covering all disciplines and is well archived, often going back more than 50 or sometimes even 100 years.

Many initiatives have been launched by publishers to make TDM easier. Publishers also invest in cross-industry initiatives, so that miners can retrieve contents from many different STM sources in an easier fashion. CrossRef, for example, offers a TDM facility, via an open API that also checks licenses. It covers over 200 publishers concurrently. Many publishers also offer their own open API, which makes it easy for miners to download content via their own application interfaces. RightFind® XML for mining offers a commercial solution for corporations, large and small, including startups and SMEs to get high quality content from a large number of publishers in a one stop shop.

TDM worldwide; limited exceptions

In the US, content for TDM is generally obtained under licences. The Fair Use clause in US copyright law does NOT apply to TDM automatically. Some believe that Fair Use means that TDM is always allowed in the US, even without a licence. This is not the case. TDM itself is not a "use" that could be delared fair or infringing on a blanket basis. TDM is a tool like a photocopier. Applying the Fair Use rules is done case-by-case. Commercial use is a factor in fair use, bur market substitution and harm to existing markets is more important. Use of scientific content for scientific research will generally require a license.

In Europe two countries have introduced a TDM exception: in the UK and in France these exceptions have been limited to what is extremely important for STM publishers. **The exception is limited to:**

- **non commercial use**
- **content for which lawful access exists and**
- **the right to apply technical protection against piracy .**

The French draft legislation also foresees copies deleted once they are no longer required also for the TDM exercise.
In Japan a general TDM exception exists, but it excludes Databases. This is important for publishers, as their content is held in databases. It means their content can still be subject to TDM licensing.

Tools not Rules
Surveys among researchers have shown that for the TDM-miners, access to content of STM publishers is not their biggest obstacle. They need high quality content and that is what STM publishers offer them, without extra costs involved. They need tools much more than rules; better and easier TDM software.

For STM publishers it is important that a TDM exception is limited to:

- **non-commercial use**
- **content to which the TDM user has lawful access and in addition,**
- **a protection against piracy and theft.**

For commercial use no two users are identical, and the needs vary. STM publishers need to have the continued ability to work with commercial entities and negotiate the specific services that best meet the specific needs of the specific client. These arrangements will almost certainly require investment from the publishers. A broad TDM exception covering commercial use will not only harm existing arrangements, but will certainly discourage publishers from making further investments in high-quality, enriched TDM services.