FutureTDM
Explore . Analyse . Improve

CHALLENGES

IMPROVING UPTAKE OF TEXT AND DATA MINING IN THE EU

Facts
Project No: 665940
Program: H2020 | CSA | GARRI-3-2014
Duration: 09/2015 - 08/2017

BARRIERS

Academic Researchers Experience

Throughout the FTDM project we have examined a variety of case studies to find out which are the barriers that academic researchers experience while practicing TDM. The researchers which were interviewed were not given a definition of "barriers" beforehand but were let unbiased to highlight all issues hindering their work on TDM. Given the high level of expertise of the participants, their input provided useful insights on problems encountered in the following areas:

1. Technical and Infrastructure

Researchers have indicated that the use of Application Programming Interfaces (API) may help content owners, i.e. publishers, to avoid TDM activity overloading their systems and control over who can access what content, but they create access barriers. Users who have exceeded a number of downloads are sent warning messages and/or might even be blocked. Limitations may be arbitrary. As a result, the lawful researcher still needs to contact and negotiate terms under which TDM can be done, which can be time consuming and costly. In addition when APIs are used, practitioners are unsure of how much of the content was actually accessible or they are aware that full content access is prohibited resulting in low TDM results.

Another issue is not being able to mine across content providers. Again, the absence of a platform or standard API makes TDM very time consuming if not impossible.

The quality of data, has also been highlighted. Some data come in formats which are considered more TDM friendly, XML for example, in contrast to others which are highly problematic (e.g. PDF).

But even if the data quality and formats are appropriate, still there are not enough available TDM tools easy to use and effective while the existing ones are often too expensive or they do not fit for purpose and need to be tailored to become suitable. The tools which have been the outcome of research projects are sometimes non reusable due to project specificities. Moreover even the most potential useful software is provided with limited or unclear documentation so people still don't know how to use it and take advantage of its potential.

2. Legal and content

Reseachers have pointed out that the lack of awareness of what a specific TDM technology can achieve can be a problem. Most researchers use small scale samples and as a result, they may not encounter problems while remaining under a certain threshold. But access

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

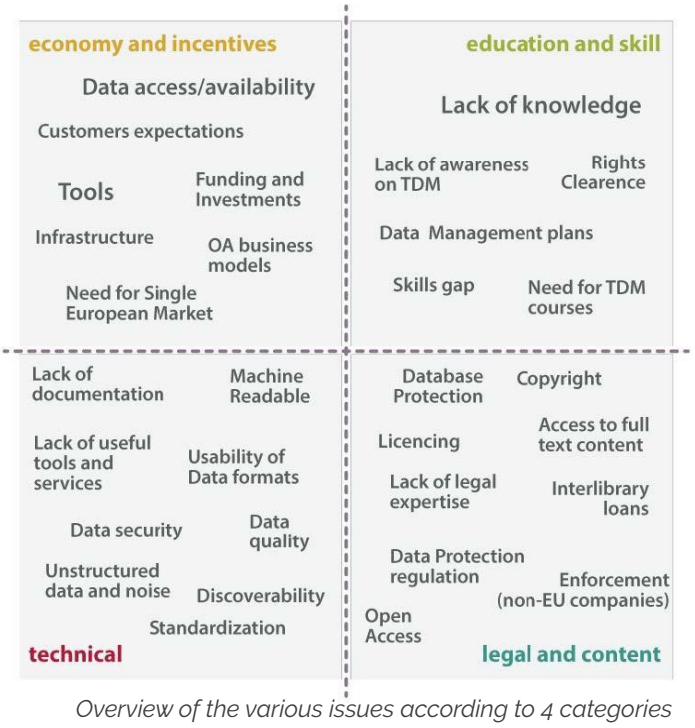
to data is not easy to obtain. Although rightsholders say there is a willingness to provide easy access and permission to use copyright protected materials, in practice the process of obtaining permission proves to be a serious barrier. Researchers usually rely on their institutional affiliations which provide access to data through their subscriptions to publisher content and interlibrary loans. The researchers themselves often do not have the time or resources available to negotiate access rights and only use data freely available without any restrictions or data with no license at all. As a consequence, some topics of research are not covered and research is being biased due to not using all the relevant data. A copyright exception for TDM being limited to the research community is said to be problematic given the ongoing trend of research cooperation between academia and industry making it impossible to distinguish between research for scientific purposes and research for product development. In the same way, the distinction of commercial vs non commercial TDM is not considered a feasible solution, as more and more frequently there are public-private partnerships and commercial spin-offs from academic research.
When researchers want to share their data, on the other hand, they are not sure what license to use; the fear of losing control of their data leads to not sharing it or use restrictive licenses.

3. Education and skill

There is a lack of understanding and awareness amongst researchers about the use and benefits of TDM. Those working in academia and industry agree that there should be a joint effort to raise awareness and to help fill the current demand for TDM practitioners. Industry can help promote and facilitate educational programs by being more involved, providing resources and clarity about career opportunities. Universities are urged to develop courses not only targeted at those who will become TDM practitioners and developers of TDM tools and services, but to include courses on TDM in the general educational curriculum improving general computer science literacy amongst all disciplines.

4. Economy and Incentives

TDM market is still very immature and the products and services available are far from being perfect. The challenge is to develop tools and services that meet the expectations and needs of the different stakeholders. Market access is being hindered by fragmentation due to the EU different regulations and languages. In addition, there is not enough academic funding that should be made available for research to address domain specific barriers, infrastructure and data acquisition



Overview of the various issues according to 4 categories