How TDM can unlock a goldmine of information

From September 6th- September 8th, over 200 people with an interest in open science came together in Athens for the Open Science Fair. OpenMinTeD was one of the co-organisers, and also organised a workshop on text and data mining. The first part of the workshop showcased successful TDM initiatives. The second part was focused on content providers and was more OpenMinTeD specific.

PART 1: TDM showcases

This part of the workshop highilighted four showcases, related to life sciences, agroculture, social sciences and entrepeneurship.

The Big Mechanism: from text to experiments using their text mining

The first showcase focused on text mining related to cancer biology. Sophia Ananiadou of NaCTeM, University of Manchester, showed how you can mine large volumes of scientific papers, in order to find information about cancer related processes inside the body. One of the aspects they focused on is detecting uncertainty. For example: if one protein **might** interact with another protein and it is written in a text like that, you can detect it. By combining 'machine learning' and 'automatic rule induction' they were able to find these kinds of relations better. Her research group also worked on an interactive visual analytics tool, LitPathExolorer, that allows them to visualize corroborating and new discoveries linked with pathways by including the user in the loop. These tools can contribute to a better understanding of cancer mechanisms and hence support experiments and drug discovery

- View the presentation on Slideshare

Left: Sophia Ananiadou presenting. Right: audience member asking a question

What can semantic text mining do for food quality improvement?

The second showcase was presented by Robert Bossy, of the French National Institute for Agricultural Research INRA. He started by announcing that he would talk about cheese. Are you wondering what text mining has to do with cheese? Well, in order to design better tasting and healthier cheese, researchers need to understand which microorganisms are found in different kinds of cheese and where they come from. They for instance designed metagenomics experiments that screen the whole diversity of microorganisms in French and Italian cheese. Textmining can help, because it can analyse thousands of papers and for example find out where these microorganisms were found in nature. INRA worked on a textmining infrastructure that can do this. Now they are trying to make it interoperable and to integrate it with OpenMinTeD. This will allow them to mine more full-text papers in a secure environment with good capabilities and a unified interface. This will also improve reproducibility and reusability of software components for other projects.

- View the presentation on Slideshare

Data Analytics meets Socials Sciences

The workshop moved on from medical and life sciences to social sciences. Haris Papageorgiou of the Institute for Language and Speech Processing (ILSP) of the Athena Research Centre introduced the audience to the concept of computational social sciences. It basically means that computers are used to model, simulate, and analyze social phenomena. You can, for example, predict the outcome of elections, examine trends in society, predict conflict regions and explain collective behavior. In order to do this you need three ingredients: a data-driven methodology, architecture and infrastructure. Haris presented two examples of TDM in social sciences. One project used text and data mining to map, document and analyse the dynamics of social movements. The other project analysed large volumes of different kinds of data in order to learn more about xenophobia in Greece during the economic crisis and migration crisis.

- View the presentation on Slideshare

Haris Papageorgiou

The science-industry relationship driven by competitive intelligence: how to surf emerging technologies

Open data and open science are not only valuable for science, but also for industry. Entrepreneur Manual Noya, CEO of Linknovate took the stage to elaborate on this. There is a need for non-manual tools that gather intelligence and translate it to business insights. For companies it is important to understand what is going on around them in terms of competitors, newcomers and opportunities in emerging markets. This is mostly enabled by emerging technologies such as IoT, cybersecurity, biometrics and artificial intelligence. Linknovate helps with industrial and academic analysis to generate technology maps, and analyse key companies (how they are investing, partnering, patenting, and other signals). Linknovate uses machine learning algorithms to stitch together hundreds of 'fresh' data sources, distills insights from this and provides the capability to contact experts directly and get unique insights from this (primary) field research.

- View the presentation on Slideshare

Manuel Noya

PART 2: OpenMinTeD and content providers

The second part of the workshop was more focused on OpenMinTeD and content providers.

From Open Access to Open Science: making sense of scientific content

Stelios Piperidis kicked off the second part of the workshop by presenting the need for text and data mining and the background of the OpenMinTeD project. On average a new scientific paper is published every twelve seconds. It would be impossible to read all this for a human, so even to keep track of one particular field, you need text and data mining. The project OpenMinTeD works on an infrastructure that makes text and data mining resources, tools and services interoperable.

- View the presentation on Slideshare

SOME FACTS ABOUT SCIENTIFIC LITERATURE
ATHENS, SEPTEMBER 6-8, 2017
OPEN SCIENCE FAIR

Stelios Piperidis presenting some facts about scientific literature

Machine accessibility of open access scientific publications

Open access publications in publisher systems are supposed to be open to everyone, but how open are they for machines to read? Petr Knoth of the Open University explained the work that he has been doing for OpenMinTeD in this context. First they asked different large publishers "how do you want us to do TDM on your content?" and "Can we use your APIs?". After that they tried if it really worked. Since part of it didn't work, they decided to make workarounds. Finally, the publications will be connected to the OpenMinTeD platform via CORE.

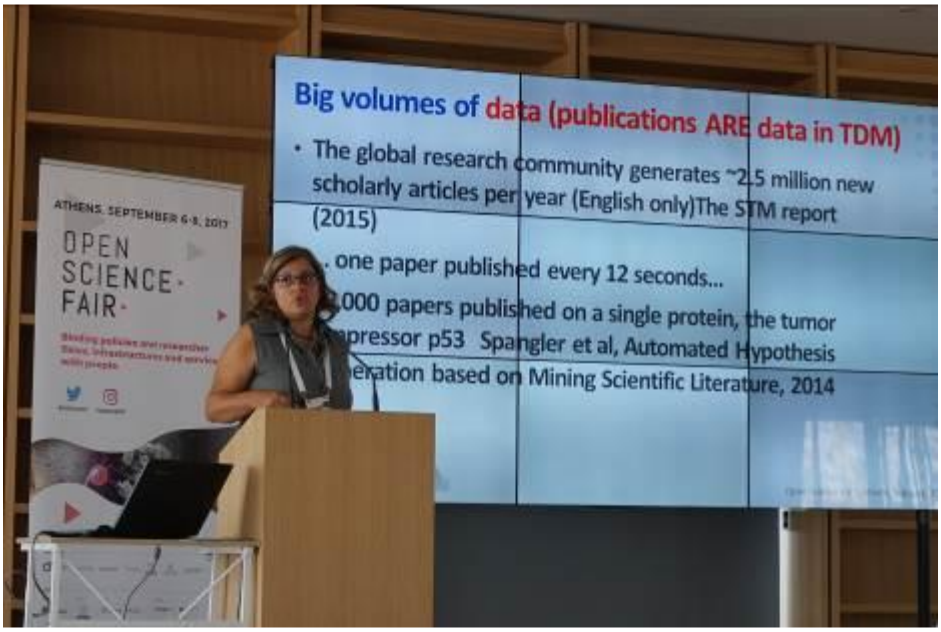
- View the presentation on Slideshare

Petr Knoth

How we explore, model, analyze and visualize systematic research in OpenAIRE utilizing Topic Modeling Services

The next presenter was Natalia Manola of Athena Research and Innovation Centre. She explained that OpenAIRE started mining scholarly literature, because they wanted to have numbers on the percentage of open access articles. The result of mining scholarly publications is linked information, for example on the relation between topics and funders, where authors are based, references and citations. With topic modeling you can find out if a topic is trendy and what the top most related publications are. After finding this information, visualization is also important for communication of the results. If all these steps are taken, the new information can help funders and institutions to make informed decisions.

- View the presentation on slideshare


Big volumes of data (publications ARE data in TDM)
• The global research community generates ~2.5 million new scholarly articles per year (English only)The STM report (2015)
. one paper published every 12 seconds...
000 papers published on a single protein, the tumor
pressor p53 Spangler et al, Automated Hypothesis
eration based on Mining Scientific Literature, 2014
ATHENS, SEPTEMBER 6-8, 2017
OPEN SCIENCE FAIR

Natalia Manola

Open Science (legal) check list for repositories and publishers

There are a number of legal shortcomings when it comes to text and data mining. Thomas Margoni of the University of Glasgow presented the opportunities that repositories and publishers interested in Open Science can seize. He summarized this in 5 steps:

1. Apply the right license to your repository (CC BY 4.0)
2. Don’t forget the metadata
3. Content should also be licensed
4. Papers, articles, songs, photographs, films etc. Should be licensed under a CC BY 4.0 . Why CC BY? Because otherwise you create license incompatibility
5. In case of data and databases: please don’t use CC-BY but CC0 (no attribution). It is a delicate topic, but remember that scientists can also be acknowledged for producing data in other ways.
6. Require that uploaders choose a licence. No licence means ALL RIGHTS RESERVED . Statements as ‘open access’ are not license and introduce legal uncertainty. Don’t do this.
7. Indicate the preferred open science licence.

- [View the presentation on slideshare]()

ATHENS, SEPTEMBER 4-5, 2017
OPEN SCIENCE FAIR
4) Papers, articles, songs, photographs, films, etc