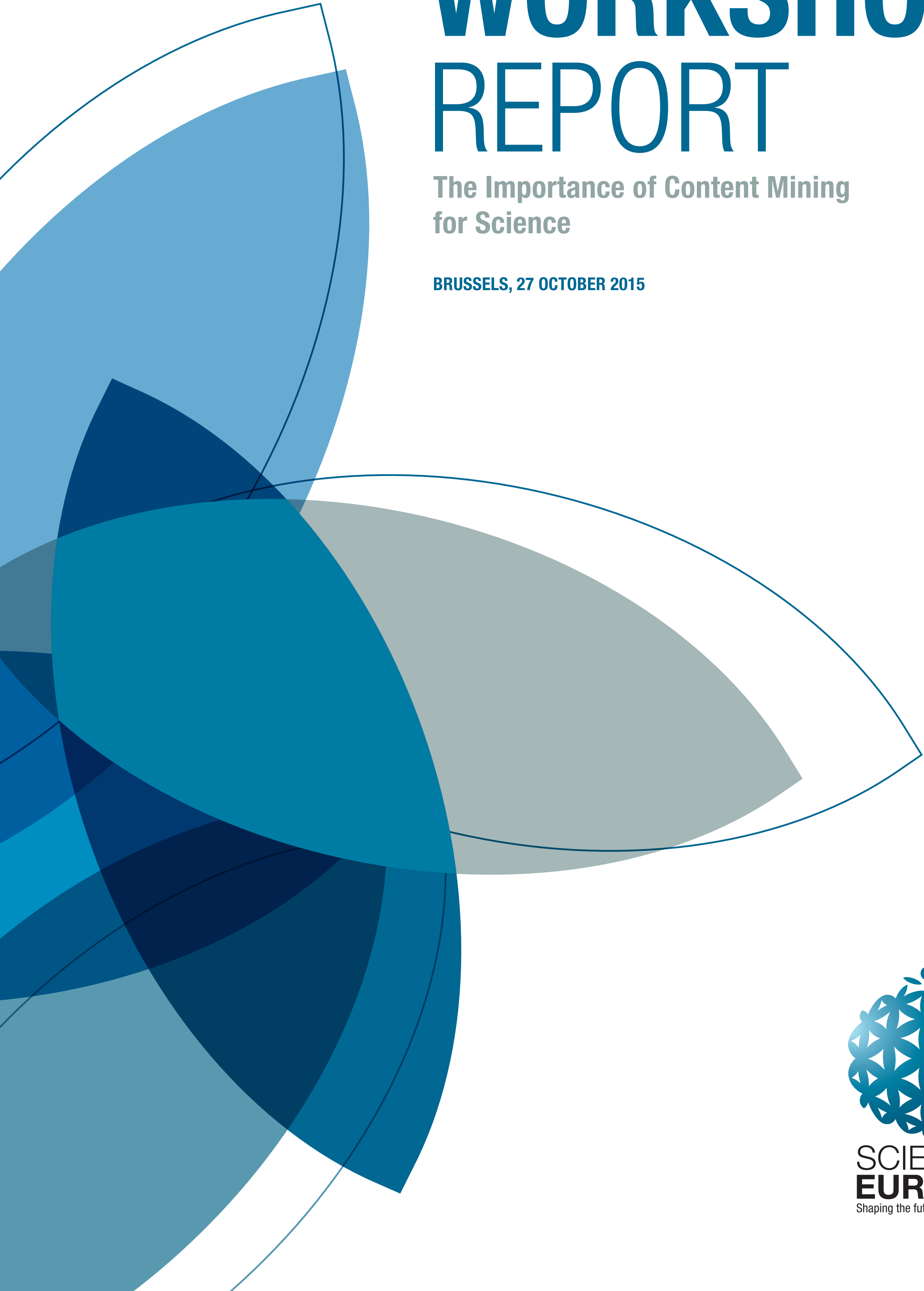
WORKSHOP REPORT

The Importance of Content Mining for Science

BRUSSELS, 27 OCTOBER 2015



SCIENCE
EUROPE
Shaping the future of research

July 2016

‘Workshop Report: The Importance of Content Mining for Science’:
D/2016/13.324/7

Author: Science Europe

Editor: Christoph Bruch (Helmholtz Association, Germany) – Member of the Science Europe Working Group on Research Data (2013–2015)

The Importance of Content Mining for Science

WORKSHOP ORGANISED BY THE SCIENCE EUROPE WORKING GROUP ON RESEARCH DATA

Executive Summary

In 2016, the European Commission is expected to propose changes to EU copyright legislation. From a research perspective, one of the pressing issues that these changes will address relates to the current legal uncertainties surrounding text and data mining (TDM).

Many researchers use TDM technologies. These are methods that can organise and search large amounts of text or data in a way that a person would not be able to do manually. However, the current legal 'grey zone' surrounding TDM in Europe makes aspects of its use uncertain. Copyright laws vary between EU Member States, and currently only UK national copyright law contains a regulation specifically allowing TDM use for research carried out for non-commercial purposes.

Under the current legal framework in Europe, rights holders can prevent researchers from using TDM on copyrighted material. The existing EU Directive on Copyright in the Information Society (Infosoc Directive) contains an optional exception to the reproduction right that allows scientists to reproduce copyrighted material where it is used for a non-commercial purpose; this exception needs to be transposed into national law in Member States, however. In its plans for a Digital Single Market, announced in May 2015, the European Commission highlighted the lack of legal clarity surrounding TDM for commercial and non-commercial use as one factor preventing scientists from using the technology to benefit their research.

Scientists are urging the European Commission to update the legal framework and allow TDM for commercial and non-commercial means as part of the upcoming EU copyright reform. Clarifying the legal position surrounding TDM will encourage better scientific research and could benefit greater society.

In recognition of the importance of this issue, Science Europe organised a workshop on text and data mining for scientific research in October 2015 in Brussels. Delegates heard how the technology can benefit science, with presentations from specialists in the technological methods underpinning TDM and from researchers in the various scientific fields that rely on TDM for their research.

Science Europe hopes that the outputs from this workshop will make an informed contribution to this debate and underline the importance of TDM to research in Europe.

Introduction

Technological developments in text and data mining (TDM) have opened up a wealth of new possibilities for researchers, enabling them to analyse information in ways that were not previously feasible. TDM can be used to extract and display information in a structured, machine-readable way that makes it easier to process and compare with other sources of data. TDM helps make searches more efficient, effectively distilling large quantities of data from publications and research data sets.

These advances enable researchers to use information technology to obtain new insights and develop novel concepts from large collections of data and text. While some analyses carried out with TDM technology could conceivably be performed manually, in practice the size and the diversity of document and data collections make manual analysis prohibitively labour-intensive and costly. Researchers in life sciences, humanities, social sciences and other fields are already using TDM techniques, and TDM technology is evolving rapidly.

However, a lack of legal clarity is hindering scientists from using these techniques to their full extent. This places European researchers at a disadvantage compared to their colleagues in other countries, such as the US. These uncertainties also constrain the development of TDM use as a component in larger applications, such as the OpenMinTeD H2020 project for European text-mining infrastructure.[1]

In autumn 2013, Science Europe's Working Group on Research Data initiated a task group on 'Legal Aspects', in order to develop an internal briefing for Science Europe Member Organisations and to help them to contribute to a European Commission consultation on copyright reform. This briefing was disseminated internally in March 2014. Based on this work, the Briefing Paper 'Text and Data Mining and the Need for a Science-friendly EU Copyright Reform' was published in April 2015.[2]

When the task group on Legal Aspects was initiated, there was still a debate within the European Commission about whether copyright reform was necessary. By 2014, the Commission had decided that it was; in May 2015, it published its communication 'A Digital Single Market Strategy for Europe.[3] This mentions the possibility of introducing an exception for text and data mining as part of a reform of the EU copyright directive:

> "Innovation in research for both non-commercial and commercial purposes, based on the use of text and data mining (e.g. copying of text and datasets in search of significant correlations or occurrences) may be hampered because of an unclear legal framework and divergent approaches at national level.
>
> The need for greater legal certainty to enable researchers and educational institutions to make wider use of copyright-protected material, including across borders, so that they can benefit from the potential of these technologies and from cross-border collaboration will be assessed, as with all parts of the copyright proposals in the light of its impact on all interested parties."

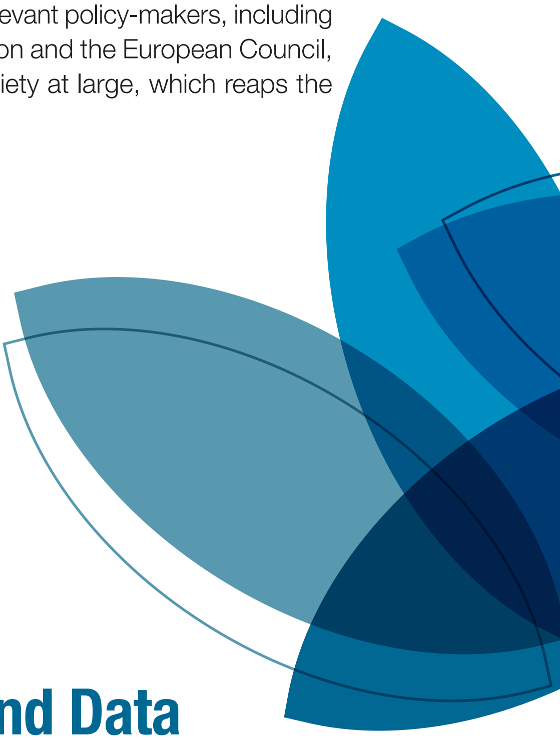
In this very communication, the Commission specifically referred to the fact that legal uncertainty was preventing researchers from using TDM for commercial and non-commercial purposes. Currently, national laws vary between Member States, presenting a challenge to the flow of scientific research across Europe's borders. Any copyright reform should reflect the needs and interests of researchers in exchanging knowledge across borders as well as allowing them to make greater use of emerging TDM technologies. Details relating to TDM are due to be presented in 2016.

Workshop Aims

The aim of the Science Europe TDM workshop was to demonstrate to relevant policy-makers, including Members of the European Parliament (MEPs), the European Commission and the European Council, the importance of text and data mining for scientific research and society at large, which reaps the benefits of research results..

The workshop covered four principal elements:

1. Introduction to TDM
2. Perspectives from Specific Research Fields
3. Benefits for Economy and Society
4. Policy and Legal Perspectives



Session 1: Introduction to Text and Data (Content) Mining

Overview

Sophia Ananiadou, Director of the National Centre for Text Mining, United Kingdom, and professor of computer science at the School of Computer Science, University of Manchester, illustrated how TDM is a crucial research method that can radically change scientific work. Professor Ananiadou stressed the importance of clarifying the legal issues surrounding TDM through copyright reform, and elaborated on some of the other problems confronting researchers who use these techniques, in particular the major issues of licensing restrictions and access to data.

Professor Ananiadou described the legal situation in the UK, the only EU Member State where the national copyright law explicitly states that TDM does not infringe copyright, albeit for non-commercial research purposes only. A number of researchers, funders, and research libraries lobbied for this legislation for several years before its introduction in 2011.

Drawing on her experience as a researcher, Professor Ananiadou explained the technology that underpins TDM. She provided examples of a number of online text mining tools and services used in a variety of types of research. She explained that, at its core, TDM is a way of enriching textual data with semantics. In many cases, textual data is available, but in an unstructured form that is not suitable for machine reading. By indexing data according to semantic markers, TDM imposes structure on textual data, opening it up to machine reading. These annotations help by integrating information from different sources, thus helping researchers find information and generate hypotheses.

There are a number of TDM tools that can add semantic annotations to unstructured textual data. These annotations can support a number of applications including semantic search, summarisation, question-answering, classification, hypothesis generation, and biological network construction. Professor Ananiadou emphasised that the value of adding semantics to unstructured data allows a transition “from big data to big semantics”.

She presented examples of the services offered by the UK National Centre for Text Mining. These include an automatic term and concept extractor, TerMine,[4] and a faceted semantic search system, Kleio,[5] based on named entities suitable for biology. Kleio incorporates features such as species disambiguation and acronym recognition. A more advanced application is EvidenceFinder,[6] part of Europe PubMedCentral, which extracts named entities and large quantities of facts (snippets of information) from over three million full texts. Questions are automatically recommended to users based on the initial query. This means that users are shown questions with known answers. Another application, FACTA (Finding Associated Concepts with Text Analysis),[7] automatically extracts direct and indirect associations between concepts that support literature-based discovery and hypothesis generation from MEDLINE.

According to Professor Ananiadou, a further complication in using unstructured data is that publishers have different rules and licenses for TDM, making data collection challenging for researchers outside the UK, who do not enjoy the copyright exception. If a corpus is to be created using a broad range of works, including works protected by copyright for which different licenses govern how they may be used, then its accessibility and usability will be determined by the most restrictive license applicable..

Algorithm Development for Retrieval, Mining and Visualisation in the Context of Big Data

Matthias Hagen, head of the research group Big Data Analytics at the Bauhaus-Universität in Weimar, Germany, elaborated on Professor Ananiadou's theme. He demonstrated a number of technological developments involving research questions that can be answered if there are known frequent occurrences of web phrases.

He began by recalling IBM's Watson computer that competed on — and won — the American trivia show 'Jeopardy'. This example demonstrated how technology such as the Watson system could successfully locate information on dedicated questions posed to it.

Professor Hagen then demonstrated the Netspeak web service[8] for writing assistance run by the Bauhaus-Universität Weimar. To write in a foreign language is a difficult task, even for an experienced author. Challenges include identifying the correct word or preposition in a given context, finding commonly used expression and avoiding the use of grammatical formulations that reflect the author's native language. Netspeak assists writers in overcoming these issues by using the web as a source of everyday language. The service can be queried with short text phrases to determine how common they are on the web. Wildcard characters can be added to the search, allowing for variations and synonyms of the query phrase, which is returned as a ranked list reflecting the phrases' frequency of occurrence on the web.

According to Professor Hagen, the programme is an example of how TDM can tackle one of the main challenges researchers face when working with big data such as web frequencies: knowing what questions to ask and what problems to solve with the data. As well as Netspeak, he also demonstrated other research questions, including automatic text paraphrasing that uses the web as a source of knowledge and for understanding web search queries, again with the help of web frequencies. Using text and data mining to answer questions is one of the most interesting trends in research and it is connected to the field of machine intelligence assisting humans.

Computer Linguistics as Basic Research for Text Mining

Feiyu Xu, who leads the text analytics research group in the language technology lab at the German Research Center for Artificial Intelligence (DFKI), gave a demonstration of technology to extract valuable information from vast quantities of data. Professor Xu described the project LUcKY,[9] which is being developed by her research centre. As part of this project, the centre created a technology application that helped resolve ambiguities in the meaning of terms used in texts by querying them against freely accessible databases such as Wikipedia.

She emphasised that there were no ‘general purpose’ TDM technologies. Rather, these tools need to be flexible, adaptable to the user’s needs and capable of modifying systems with domain-specific knowledge. This adaptability explains why TDM is important for all strands of research. The effective use of TDM techniques not only demands a legal framework that is conducive to their use but also the expertise and knowledge to develop TDM technologies that will meet the needs of European researchers and businesses.

Session 2: Perspectives from Specific Research Fields

Professor Xu’s assertion that TDM technologies and the underlying linguistic structures are highly versatile and can be adapted to the needs of diverse research topics was taken up in the second session of the workshop. Here, those presenting their research represented a broad range of expertise, including social sciences, humanities and life sciences. They were invited to illustrate how they deployed TDM in their particular area of interest.

Content Mining in Social Sciences

Alexander O’Connor, a lecturer at the ADAPT Centre at Dublin City University, began by describing his experiences when applying TDM to social science research. Social science research relies on tracing human behaviour. Some methods are unobtrusive and record the permanent or temporary evidence of what people are saying or doing. Other methods are obtrusive and rely on direct questioning or measurement.[10] It is the ability to examine both data types that is increasingly important in social science research analysis. Dr O’Connor emphasised the importance of open data for research, since access to the source information allows experiments to be replicated, validated and verified. Without this ability to verify that the appropriate technique was applied correctly to representative data, any results or conclusions from such work are less robust. There is an inherent tension between the rights of the providers, the creators of the data and the aims of the researchers. Licences that are intended to defend the rights of users and providers sometimes render reproducible science much more difficult. Similarly, as pointed out by Professor Ananiadou, annotations cannot be reused if the underlying content cannot be preserved and made available for further research. TDM must have full provenance to be truly open.

Dr O’Connor offered examples of social science studies that used data mining, drawing on sources available freely on the internet: social media, government websites, and emails that were published online as the result of a court case. His examples illustrated the increasing amount of data now becoming available in digital formats, which are often not originally intended for research purposes but which yield significant insights.

The failed American energy company Enron provided an unlikely example of information that could be analysed for social science research, when emails from its employees were made public during a court case. Dr O’Connor referenced a paper[11] by Eric Gilbert, a professor of interactive computing at the Georgia Institute of Technology, which analysed these emails and ranked them according to staff hierarchy within the company. Gilbert categorised phrases that came up frequently in the internal emails, using them to determine whether an employee and the email recipient were of higher or lower ranking. Dr O'Connor said that this social study provided an example of the value of making information publicly available.

Another example was a project called 'Talk of Europe'.[12] This used publicly available information to structure debates from the European Parliament and information about Members of the European Parliament (MEPs). The platform could be consulted to answer specific questions about MEPs using a variety of categories, including their countries of origin, their native languages and their policy areas of interest. This project not only exemplifies a potential use of TDM but also illustrates how TDM technology can provide results of direct relevance for the broader public.

Dr O'Connor told participants that questions of data use may need to be considered from an ethical standpoint, since using TDM on certain personal data could potentially reveal information that the people providing the data would prefer to keep private. This raised a number of other privacy issues; however, data protection was not a focus of this workshop.

The conclusion of the social science TDM discussion focused on the important issue of whether available data can be organised in such a way that researchers can study them. Without access to open, licensed, shareable data, there is a significant risk that conclusions could be distorted by relying on small, regional-specific or otherwise unrepresentative data. It is vital that researchers have certainty that they can safely gather and share their data as well as publish their methodology and findings.

Content Mining in the Humanities

Heike Zinsmeister, professor for German linguistics at Hamburg University, addressed TDM use for researchers in the humanities. She explained that her university's humanities department had recently established an ethics committee to determine how personal data should be used – or not – for research purposes.

The classical approach to humanities research involves analysing texts to uncover symbolism, comparing sources and checking specific hypotheses against texts. Technology is increasingly important in undertaking humanities research. For example, researchers in the field of art history will use TDM tools to identify similarities between works of art, compare images and analyse colour spectra by aggregating image patterns. Tools such as HathiTrust Digital Library[13] make it possible to search millions of texts by automatically analysing a query and returning results from its database that match the words and argument structure of the query. This, she explained, enables researchers to formulate specific questions and obtain precise results that they would not otherwise be able to access. A broad corpus of knowledge is needed to develop and improve this technology.

An example of the increasingly pivotal role of TDM in humanities research is Franco Moretti's 'distant reading' project. Moretti, a professor of literature at Stanford University, uses computer search tools to analyse literary texts without personally reading the texts. This extracts key document details, such as publication date, author, main characters, subjects, and places. This 'distant reading' approach exposes elements of literature that researchers may not always identify when manually searching a vast library of books. These include patterns of literary terms and vocabulary that come up in texts from a particular period or author. Professor Moretti aims to understand literature not by studying particular texts, but by aggregating and analysing massive amounts of data.[14]

Professor Zinsmeister argued that copyright restrictions prevent researchers in the humanities from fully exploiting the potential of TDM. Many researchers seeking information may realise that there may be a copyright limitation on one of their references only after they have cited a particular publication. Citations in academic publications present a legal grey zone, as copyright exceptions concerning citations are not harmonised internationally. Most copyright laws do not allow citation of entire works;

therefore it is often not possible in practice to reproduce results from empirical literature analysis such as ‘distant reading’. Research results may also not be able to cite the source they used for TDM because of potential copyright issues.

Content Mining in Life Sciences

Claire Nédellec, research director at the laboratory for ’Applied Mathematics and Computer Science, from Genomes to the Environment’ (MaIAGE) at the French National Institute for Agricultural Research, explained how TDM is used in life sciences research. Recent developments in molecular research technologies have generated a flood of information from new experiments in the fields of genetics, physiology, proteomics and metabolomics. TDM techniques have helped build databases on microbial biodiversity and to advance knowledge in system biology. According to Professor Nédellec, without the help of technological TDM tools it would be impossible for researchers to systematically review all the newly published information that has become available. TDM not only significantly advances researchers’ abilities to analyse information, it also makes it possible to link the semantic content of academic publications and related research data in bioinformatics applications. For example, a publication may include references to specific molecules. As a first step, these different elements of the publication can be identified. Then, TDM and a database can be used to harmonise these names, despite spelling variations in how the molecule is labelled. Databases with information concerning this molecule can be searched. This search strategy may automatically link the underlying research behind the original publication with information about other research that may not have beendiscovered without the use of TDM. Over and above specific molecules, TDM can extract relations between information.

An example is the collection of properties of a given variety of wheat that are known to be resistant, or not, to given diseases. Such information is only available in thousands of papers, but can be automatically extracted using TDM technologies such as the Alvis Suite. At the same time, breeders and researchers spend time and money to repeat experiments because they lack access rights to TDM results. More generally, the public availability of phenotype information linked to living organisms – not only for wheat or other plants, but also animals, micro-organisms, and humans – is a major issue in life sciences and is highly relevant in agriculture, health and food research. According to Professor Nédellec, one of the greatest hindrances today to the development of such applications is the lack of access rights to the TDM results that draw on the scientific literature where this phenotype information is published.

Session 3: Benefits for Economy and Society

An SME Perspective

Stefan Geißler, from the software company Temis, offered a different perspective. Temis[15] is an SME with offices in Germany, France, the US and Italy, specialising in creating text and data mining software. He described how private businesses use TDM to analyse text and data. The value of the Temis software is that it can quickly convert raw scientific content drawn from thousands of documents into a structured overview of information.

Mr Geißler said that Temis’ clients often find themselves unsure whether they can legally use the software they have purchased. He is hopeful that the European Commission’s announcement of its Digital Single Market plans will clarify the legal situation. He argued that an exception limited to only non-commercial use would not help many of TEMIS’ clients, who are engaged in commercial activities.

According to Mr Geißler, Temis "plays for both teams", by selling its software both to publishers and to private companies that want to analyse large quantities of scientific information. Unsurprisingly, this means that Temis' clients are divided in their opinions on copyright restrictions for TDM use.

Mr Geißler argued that people should be legally able to use Temis' data mining software regardless of whether they have the rights to read a text for commercial or non-commercial use.

Many of Temis' clients are from the pharmaceutical industry and use TDM to extract information from scientific and clinical publications. Often, their industry clients run into difficulties, with scientific publishers claiming that their licence fee does not extend to the use of TDM. However, clients understandably do not want to have this distinction and want to use their subscriptions for reading both manually and with the TDM applications. Companies also argue that TDM helps them determine – in advance – whether articles may be helpful for their research. There should be no reason that they need to pay to access 10,000 articles separately when perhaps only two or three out of the group may be of interest. There has been little progress towards clarifying this in recent years.

Potential Benefits for Knowledge, Economy and Society

Luis Magalhães, professor at the Instituto Superior Técnico, University of Lisbon and chair of the OECD Working Party on Indicators for the Information Society, offered some predictions on the potential economic benefits of TDM use. He began by explaining the various factors that have contributed to the surge in data volume. These include the emergence of new devices, the fall in the cost of data storage and powerful data analytics, all of which enable data mining.

Professor Magalhães believes that if TDM is used to extract and analyse this growing volume of data, it will drive advances in a wide range of research fields. He said that although he was unaware of any comprehensive study detailing the overall economic benefits resulting from TDM use, there have been studies on the benefits to specific sectors, including ICT and healthcare.

Professor Magalhães explained that through the analysis of public sector information (content produced by public sector bodies) it is estimated that hundreds of billions of euros in savings per year could be generated in OECD member states. By using data analytics extensively, public sector costs could be cut by 15–20% per year. Meanwhile, the extensive use of data analytics in the healthcare sector in the US has been estimated to have cut costs by 8%. He also noted that data-driven innovation and analytics is likely to increase efficiency in a number of service sectors, including water, energy, waste disposal and transport. Using TDM will provide public officials with access to better information that will in turn lead to better-informed, more effective policymaking. Other sectors likely to benefit from TDM use include education, healthcare and environmental management.

He identified several factors that may create obstacles to realising the potential benefits of TDM. These include legal barriers, not only copyright restrictions but also data protection and privacy laws that limit how data can be used. Europe lags behind other parts of the world in terms of the number of scientific articles that are mined, a trend he attributed to outdated copyright laws. In the US, 47% of scientific articles are mined, compared with 13% in China and 11% in the UK (although he did not show figures for the rest of Europe, it is possible that the UK may be ahead of continental Europe because of the exception allowing TDM to be introduced into UK copyright law).

Session 4: Policy and Legal Perspectives

The European Commission Perspective

Jean-François Dechamp, policy officer at the Directorate General for Research and Innovation at the European Commission, gave participants an overview of the European Commission's perspective on TDM use. He said the Commission recognised that there are legal hurdles that prevent widespread TDM use. In 2014, the Commission convened an expert group on TDM, the outcome of which was a study that called for strategic reform. In May 2015, the new European Commission presented its Digital Single Market plans. These included a reference to current restrictions on TDM resulting from application of EU copyright law.

Mr Dechamp did not provide details of the Commission's upcoming copyright proposal, or how it is likely to affect TDM use. However, he did say that the Commission recognised the need to take into account the role of TDM in promoting research-driven innovation. He also touched briefly on a report[16] by German MEP Julia Reda[17] and the 2015 Review of the EU copyright framework by the European Parliament Research Service.[18] Both of these state that new legislation should address legal gaps in TDM use. Mr Dechamp said that the European Commissioner for Research, Science and Innovation, Carlos Moedas, wants the new EU copyright law to include an exception allowing TDM use for scientific research.

Is Europe Falling Behind in Data Mining? Copyright's Impact on Data Mining in Academic Research

Lucie Guibault, professor of information law at the University of Amsterdam, pointed out that TDM is not directly mentioned in the Directive on Copyright in the Information Society (Infosoc Directive). Nevertheless, the technology is hindered by the directive because TDM regularly involves making copies of the works to be mined. This infringes the reproduction right that is broadly protected by Article 2 of the Infosoc Directive. Copying protected works either needs an appropriate license or an exception within the law.

Professor Guibault focused particularly on Article 5 of the Infosoc Directive, which lists exceptions to the reproduction right and the right of communication to the public. Article 5(1) of the Infosoc Directive allows for broad "transient and incidental reproductions" of copyrighted works. However, transient copies for TDM purposes are only allowed if they are an integral and essential part of a technological process whose sole purpose is to enable (a) a transmission in a network between third parties by an intermediary, or (b) a lawful use. However, a use can be 'lawful' if it is authorised by either the rights owner or by law. Since there is no specific provision in the Directive authorising TDM, this means that the article does not provide a guarantee of the right to carry out TDM without the consent of rights holders.

Professor Guibault reminded participants that Article 5 of the Infosoc Directive also contains an optional exception to allow reproductions for the purpose of scientific research, as long as it has a "non-commercial purpose." Both exceptions have been implemented in different ways by the EU Member States, creating further legal uncertainty for researchers.

TDM is frequently conducted by extracting information from an online database. However, the 1996 Database Directive allows for only small fractions of a database to be extracted without consent. This creates yet another hurdle for researchers' use of TDM on top of the reproduction right mentioned above.

One recent case that complicated how legal scholars understand the Database Directive stems from the January 2015 European Court of Justice case involving Ryanair and PR Aviation, a company that operates price comparison websites. The court concluded that the information database freely available on Ryanair's website was not protected by copyright law because it was not original. The database was also not protected under the Database Directive because it did not meet the criterion of substantial investment. However, the court sided with Ryanair's argument that PR Aviation had agreed to its terms and conditions, which prohibited it from mining the airline's database.

Comparing the situation in the EU to that in other countries, Professor Guibault referred to Japan's copyright law, which includes a specific exception for TDM. The copyright laws of Israel, South Korea and Singapore probably allow TDM under the 'fair use' doctrine. In the US in October 2015, following years of legal uncertainty, Google Books won a court case that declared the reproduction of parts of published books for the purpose of TDM to be 'fair use'. Guibault argued that EU copyright law needs to allow TDM for scientific research and suggested the Database Directive be repealed or reformed to accommodate researchers.

Conclusions

In her concluding remarks, Science Europe Director **Amanda Crowfoot** emphasised that TDM is an important pillar in helping achieve the European scientific community’s broader aims. Supporting legal approaches to TDM could help researchers achieve Science Europe’s goals of supporting borderless science and improving the scientific environment. Creating legal clarity for TDM use would also help Science Europe to facilitate and communicate science by enabling cross-border research collaboration, create a clear copyright framework and improve research through increased availability of information.

TDM is already a vital tool for many researchers in Europe. As the amount of information that can be read by TDM software and technologies increases, its importance will only continue to grow.

Legal confusion surrounding TDM use in Europe exists for a number of reasons: EU copyright law is difficult to understand, the Database Directive further complicates questions of legal access to content, and there are differences between copyright laws in individual EU Member States. However, since the Google Books ‘fair use’ ruling in the US, European researchers now find themselves at a clear disadvantage when compared to their American counterparts, who enjoy much greater legal certainty when using TDM.

Developments that will influence TDM are imminent. The European Commission will present its proposals for a new EU-wide copyright law within the coming months. It is important that scientific researchers’ needs and concerns are heard, as they are both creators of their own copyrighted work and users of others’ works.

An improved EU copyright law that allows TDM for scientific research will support the broader goals of Science Europe by encouraging the free flow of knowledge across borders. Clear laws that enable TDM will also facilitate better research conditions and help ensure that researchers can access and share each others’ work.

Such a law will not only be important for improving the quality of research within scientific fields. It will also encourage researchers to access each others’ work more freely and to communicate and collaborate without the restraints imposed by cumbersome access restrictions or differences in national copyright laws.

Notes and References

1. http://openminted.eu
2. http://scieur.org/tdm
3. http://europa.eu/rapid/attachment/IP-15-4919/en/DSM_communication.pdf
4. http://www.nactem.ac.uk/software/termine/
5. http://www.nactem.ac.uk/Kleio/
6. http://labs.europepmc.org/evf
7. http://www.nactem.ac.uk/facta/
8. http://www.netspeak.org/
9. http://googleresearch.blogspot.de/2013/07/natural-language-understanding-focused.html
10. Strohmaier, Markus, and Christoph Wagner. "Computational social science for the world wide web." Intelligent Systems, IEEE 29.5 (2014): 84-88.
11. Gilbert, Eric: Phrases that signal workplace hierarchy, in CSCW '12 Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, New York, NY, ACM, 2012, pp 1037–1046
12. http://dhbenelux.org/wp-content/uploads/2014/06/poster-kemman.pdf
13. https://www.hathitrust.org/
14. Moretti, Franco: Distant Reading, Verso, Brooklyn, NY, 2013
15. http://www.temis.com
16. European Parliament; Committee on Legal Affairs (Rapporteur Julia Reda): REPORT on the implementation of Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society (2014/2256(INI)), A8-0209/2015, Brussels, European Parliament, 24.06.2015, http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML+REPORT+A8-2015-0209+0+DOC+PDF+V0//EN
17. http://www.europarl.europa.eu/meps/en/124816/JULIA_REDA_home.html
18. http://www.europarl.europa.eu/RegData/etudes/STUD/2015/558762/EPRS_STU(2015)558762_EN.pdf

Annex

Programme

Tuesday 27 October 2015 // Marriott Renaissance Hotel, Brussels

09.00–09.15 **Opening**

Welcome Messages
Mr Stephan Kuster, Head of Policy Affairs of Science Europe
Dr Hans Pfeiffenberger, Chair of the Science Europe Working Group on Research Data

Workshop Objectives
Dr Christoph Bruch, Helmholtz Association, Germany

09.15–10.45 **Introduction to Text and Data (Content) Mining**
Moderator **Mr Stephan Kuster**

Keynote Speech
Professor Sophia Ananiadou, National Centre for Text Mining, UK

Algorithm Development for Retrieval, Mining and Visualisation in the Context of Big Data
Junior Professor Dr Matthias Hagen, Bauhaus University Weimar, Germany

Computer Linguistics as Basic Research for Text Mining
PD Dr habil Feiyu Xu, German Research Center for Artificial Intelligence, Germany

11.15–12.45 **Perspectives from Specific Research Fields**
Moderator **Professor habil Dr Rūta Petrauskaitė**, Research Council of Lithuania

Content Mining in Social Sciences
Dr Alexander O’Connor, ADAPT Centre, Dublin City University, Ireland

Content Mining in the Humanities
Professor Dr Heike Zinsmeister, Hamburg University, Germany

Content Mining in Life Sciences
Dr Claire Nédellec, French National Institute for Agricultural Research, France

13.45–15.00 **Benefits for Economy and Society**
Moderator **Ms Ana Cristina Neves**, Foundation for Science and Technology, Portugal

An SME Perspective
Mr Stefan Geißler, TEMIS, France

Potential Benefits for Knowledge, Economy and Society
Professor Luis Magalhães, Instituto Superior Técnico, University of Lisbon, Portugal

15.30–16.30 **Policy and Legal Perspectives**
Moderator **Dr Christoph Bruch**

The European Commission Perspective
Mr Jean-François Dechamp, DG Research and Innovation

Is Europe Falling Behind in Data Mining? Copyright’s Impact on Data Mining in Academic Research
Dr Lucie Guibault, Institute for Information Law, University of Amsterdam, the Netherlandsl

16.30–17.00 **Closing Remarks**
Ms Amanda Crowfoot, Director of Science Europe

Science Europe is a non-profit organisation based in Brussels representing major Research Funding and Research Performing Organisations across Europe.

More information on its mission and activities is provided at www.scienceeurope.org.

To contact Science Europe, e-mail office@scienceeurope.org.

Science Europe
Rue de la Science 14
B-1040 Brussels
Belgium

Tel +32 (0)2 226 03 00
Fax +32 (0)2 226 03 01
office@scienceeurope.org
www.scienceeurope.org