FutureTDM

Explore . Analyse . Improve

REDUCING BARRIERS AND INCREASING UPTAKE OF TEXT AND DATA MINING FOR RESEARCH ENVIRONMENTS USING A COLLABORATIVE KNOWLEDGE AND OPEN INFORMATION APPROACH

Deliverable D4.3

Compendium of Best Practices and Methodologies

Project

Acronym:	**FutureTDM**
Title:	Reducing Barriers and Increasing Uptake of Text and Data Mining for Research Environments using a Collaborative Knowledge and Open Information Approach
Coordinator:	SYNYO GmbH
Reference:	665940
Type:	Collaborative project
Programme:	HORIZON 2020
Theme:	GARRI-3-2014 - Scientific Information in the Digital Age: Text and Data Mining (TDM)
Start:	01. September, 2015
Duration:	24 months
Website:	http://www.futuretdm.eu/
E-Mail:	office@futuretdm.eu

Consortium:

SYNYO GmbH, Research & Development Department, Austria, (SYNYO)

Stichting LIBER, The Netherlands, (LIBER)

Open Knowledge, UK, (OK/CM)

Radboud University, Centre for Language Studies The Netherlands, (RU)

The British Library Board, UK, (BL)

Universiteit van Amsterdam, Inst. for Information Law, The Netherlands, (UVA)

Athena Research and Innovation Centre in Information, Communication and Knowledge Technologies, Inst. for Language and Speech Processing**,** Greece**,** (ARC)

Ubiquity Press Limited, UK, (UP)

Fundacja Projekt: Polska, Poland, (FPP)

Deliverable

Number:	**D4.3**
Title:	**Compendium of best practices and methodologies**
Lead beneficiary:	**OK/CM**
Work package:	WP4: Analyse: Fields of application, projects, best practices and resources
Dissemination level:	Public (PU)
Nature:	Report (RE)
Due date:	30.06.2016
Submission date:	30.06.2016

Authors: **Freyja van den Boom**, OK/CM

Contributors: **Ben White,** BL
Maria Eskevich, RU
Antal van den Bosch, RU
Jenny Molloy, OK/CM
Burcu Akinci, SYNYO
Blaz Triglav, Mediately
Donat Agnosti, Plazi
Malcolm Macleod, Edinburgh University

Review: **Alessio Bertone**, SYNYO
Maria Eskevich, RU

Acknowledgement: This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 665940.

Table of Contents

LIST OF FIGURES

EXECUTIVE SUMMARY

The purpose of the report is to find, challenge and/or provide evidence for what are considered to be barriers and enablers for text and data mining (TDM) in Europe.

We use the general term text and data mining in this report, although the activity can be also referred to as two separate and partly overlapping text mining and data mining processes. Text Mining is the analysis of textual data, as well as all other forms of data converted to text, while Data Mining started from mining databases and evolved to encompass mining of all forms through which information can be transmitted.[1]

For this report, we examine some of these different TDM practices carried out by scientific researchers and small scale companies working in different sectors of economy. Building upon previous deliverables and underlying research done within the FutureTDM project, this deliverable provides a first set of seven Text and Data Mining (TDM) case studies that will further be developed to demonstrate the different text and/or data mining practices. The case studies are set up in such a way that will highlight the apparent barriers to improving the uptake of TDM in Europe.

Chapter 1 starts with the introduction followed by a description of the methodology in Chapter 2.

Chapter 3 provides an overview of the main insights gained from the interviews with respect to the barriers. These have been grouped together under the headings

- 'Technical and Infrastructure'
- 'Legal and content'
- 'Economy and Incentives' and
- 'Education and Skill'.

Chapter 4 describes the conceptual framework and case study selection, Chapter 5 contains the seven different case studies and Chapter 6 concludes this deliverable with the main findings and steps for further research.

The interviews and use cases analysis resulted in the following conclusions on the TDM state-of-the-art uptake and growing potential:

- First, the case studies indicate that not every TDM project or practice is the same. This complexity may make it difficult to develop a 'one fits all' solution. It shows that in practice researchers are facing a number of issues that may not be resolved by any single solution. Therefore recommendations need to take into account that a combination of enablers may be needed.
- Second, some of the aforementioned case studies involve cross-discipline collaborations, private-public partnerships and not only EU but also international collaborations. This

[1]http://project.futuretdm.eu/wp-content/uploads/2016/06/FutureTDM_D4.1-European-Landscape-of-TDM-Applications-Report.pdf

demonstrates that from a policy perspective, the benefits of TDM would best be achieved through different stakeholders working collaboratively.

Having identified the main barriers and gained insight into the issues from different stakeholders perspectives, the follow-up deliverable D4.5 will focus on best practices and recommendations that will improve the uptake of TDM in Europe.

1 INTRODUCTION

This report showcases case studies of TDM practices, focusing on the main barriers to the uptake of TDM from the perspective of different stakeholders. The first part of the report will the provide background on which the case studies in the second part of the report have been selected, namely the stakeholder interviews.

The main purpose of the interviews was to get more insight into the practice of TDM from the perspective of those who work with the actual tools and data. We contacted expert practitioners from different communities and economic sectors, in order to have a representation of different TDM involvement levels and working practices. This helps us to get a better understanding of evidence that is crucial for relevant policy making.

Limitations

The interviews provide insights into the practice of TDM from the main stakeholder communities involved in TDM as identified in Deliverable D2.2.[2] The limited number of interviews planned restricted the choice of participants to a selected group of TDM practitioners. The results may therefore not cover a full representation of the entire TDM community nor can this be considered to cover the wide range of TDM practices that exist in the different fields and EU member states. The results nevertheless represent the most important issues. Given the expertise of the participants, the input provides useful insights that are indicative of the barriers to TDM uptake in a more general sense. Making this deliverable public and presenting it to the community will provide the feedback necessary for the update revision of this deliverable under D4.5.[3]

2http://project.futuretdm.eu/wp-content/uploads/2015/12/FutureTDM_665940_D2.2-Stakeholder-Involvement-Roadmap-And-Engagement-Strategy.pdf

3The focus of the FutureTDM Deliverable 4.3 is on gaining insight into the barriers that the different stakeholders experience in practice. The FutureTDM Deliverable 4.5 will provide an update of the case studies focused in this deliverable and present best practices and recommendations for the different stakeholders on how to improve the uptake of TDM in Europe.

2 METHODOLOGY

2.1 The Interviews procedure

2.1.1 Interviews

A semi-structured method was chosen to benefit from having a common structure for all interviews while at the same time flexibility for the interviewer to ask for clarification or to allow the interviewee to elaborate on specific topics of expertise. The interviews took place between March and June 2016 and were recorded and later manually transcribed. Figure 1 illustrates the location of interviews and KCs.

The participants were well informed and consent was freely given for the recordings and information to be used for the purpose of the FutureTDM project deliverable. After a first set of 3 interviews (10% of the total amount of interviews) the questions were reviewed and adjusted to better cover the research questions, and to keep the discussions within the 45 minutes time frame.

2.1.2 Selection of participants

The 30 interview participants were selected using the internal project stakeholders' directory, the FutureTDM network and recommendations from all partners. One or two participants were chosen from each stakeholder groups to give a sample representation of their specific field of expertise.

2.1.3 Questionnaire

Initial set of questions for the Questionnaire was developed based on the issues that FutureTDM has identified in previous research, meetings and the FutureTDM Knowledge Cafés.

These questions were grouped around four identified themes:

- *Economic and Incentives* theme consisted of six main questions that aim to understand the TDM market and the barriers to enter the market, and how to incentivise different stakeholders to contribute to and improve TDM uptake in Europe.
- *Legal and content* theme consisted of nine questions and an additional set of sub-questions. These were developed to see whether an interviewee was aware of any legal issues and whether they were aware of and experienced any legal barriers, if so, how did this affect their TDM practices and/or research. It also focused specifically on whether there is a need for an exception to copyright and what the requirements would have to be for such an exception to effectively improve research and innovation.
- *Technical and Infrastructure* theme consisted of four main questions with several sub-questions to focus on data management plans in practice, data quality and availability of useful tools for TDM. It also asked about currently available European infrastructure, and necessary steps to improve and/or establish it.
- *Education and skill* theme consisted of eight main questions that addressed the need for qualified people to develop and work on TDM. It also tried to get a better understanding of the current status of TDM education, availability of a skilled workforce and what is considered necessary to improve this.

Figure 1: Geographic Map with the locations of the interviews and knowledge cafes

After an internal review, the questionnaire for the semi-structured interviews was approved and used for a first set of interviews aiming at 10% of the total amount of interviews. After this set was carried out, the questionnaire was reviewed again and adjusted based on the responses. To get clearer information from the participants some questions were removed from the questionnaire because they did not provide a useful response and some other questions and sub-questions were added to get more clarity about a certain topic.[4]

2.2 The case study format

The format of the case studies was chosen to represent the barriers that TDM practitioners face with respect to TDM practices in Europe. The outcome of the interviews led to selection of appropriate case studies to cover these barriers. The case studies aggregate the insights of the FutureTDM project gained from the interviews, knowledge cafes and other FutureTDM dissemination and research activities.

[4] The Questionnaire is available in the Annex 1

Given the complexity of the issue and the specificities of different stakeholders, projects and issues the case studies do not intend to cover all issues that may arise when doing TDM that we encountered during the interviews.

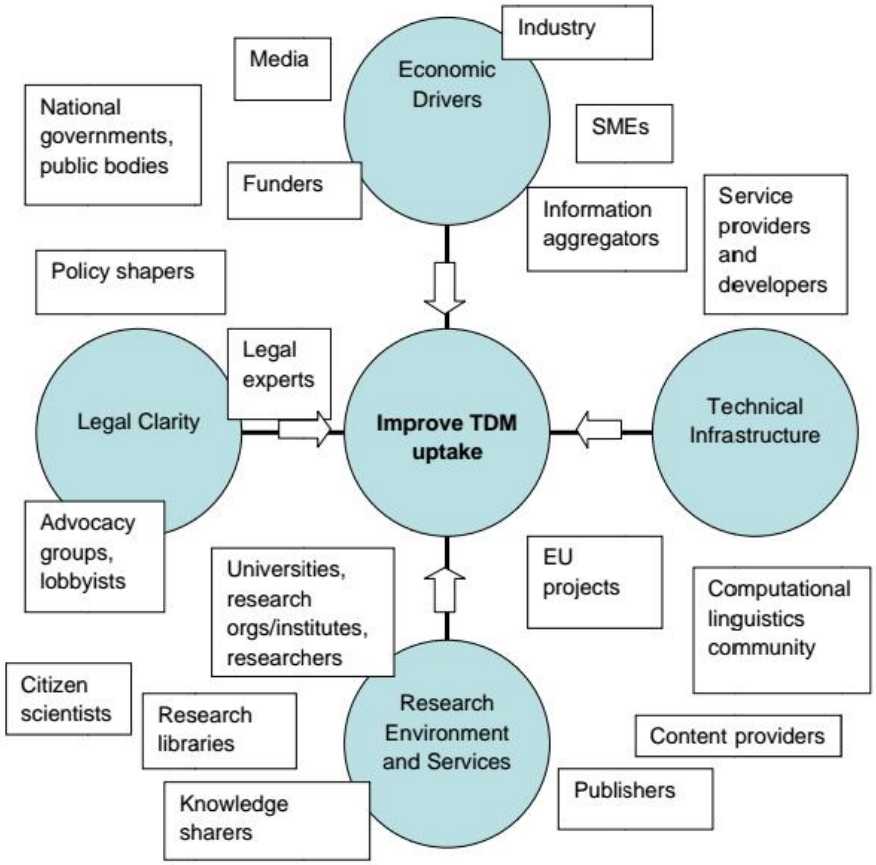

Figure 2: Main four themes identified in FutureTDM Deliverable 2.2[5]

[5] These were identified for the Knowledge Cafe's and subsequently used throughout the FutureTDM project as the main categories of barriers see Future TDM D2.2 Stakeholder and engagement involvement strategy

3 INSIGHTS FROM THE INTERVIEWS

This section provides a summary of the main barriers mentioned by the stakeholders, which formed the basis of the case study selection.

3.1 Introduction

The interviews have been coded, and grouped together under the four headings that are used throughout the FutureTDM project to identify the barriers namely:

- Technical and Infrastructure'
- 'Legal and content'
- 'Economy and Incentives' and
- 'Education and Skill'.

In the following sections we give an overview of the main issues that were raised by different stakeholders and the topics that will be covered by the case studies. You will see that we have included quotes which are all taken from the interviews. They have not been edited to keep the original wording and provide a personal voice from the different stakeholders in the TDM community.

3.2 Technical and Infrastructure

When asked about the technical and interoperability aspects of TDM and whether this may cause barriers to the uptake the following issues were mentioned:

- **Lack of Documentation, tools and services**
- **Data security**
- **Standardization**
- **Machine Readable**
- **Usability of Data formats**
- **Data quality**
- **Unstructured data and noise in datasets**

Below we summarize the main topics that appeared in the discussions about the technical aspects of TDM.[6]

'The need for TDM is clear; however in practice there are technical barriers that hinder the use of TDM and its development.'

[6] The full list of topics that were mentioned and discussed during the interviews can be found in Annex 1 the Questionnaire.

3.2.1 Access barrier: API

There seems to be no agreement over whether the use of an API provided by the publishers is sufficient for TDM. Having an API helps the platform owners, i.e. publishers, to avoid TDM activity overloading their systems. It also gives them control over who can access their content, by what means and for what purposes so that only those who have lawful access are able to do TDM. [7]However it becomes a barrier when not all of the publishers provide an API, when they all provide a different API and when APIs are reportedly insufficient for TDM requirements.

'Research would be easier if the publishers API was Open Access and we could do this in our system.'

The barriers mentioned include not being given access at all, blocking users and/or sending warnings when the use exceeds a limited amount of downloads possible through the API. This amount may not be known beforehand and the limitations may be arbitrary not taking into account requirements for TDM. As a result the researcher may have to enter negotiations which can be time consuming and costly.

Although the publishers say that using the API gives the same results as TDM applied directly to their platforms this may not always be the case in practice. People have reported being blocked when downloading a certain amount of publications without knowing what the maximum amounts for downloads are. Having a limit also impedes on being able to bulk download across various websites. Lack of a standard API across all platforms makes TDM very time consuming for researchers to try and gain lawful access in a quick and reliable way.

3.2.2 Use barrier: quality of data

For some fields the information is not yet in digital format. When data is digitized or 'born digital' it may not yet be in the right format for TDM. This may be because most publications are in a format that is not compatible with TDM practices for example the use of PDF is overall considered to be problematic whereas XML files would be more TDM friendly.

'We need more general evidence of benefits for use of standards: does that research have more impact, is it more widely used, does it have more citations or being used in subsequent research.'

There is an agreement that having a standard format regardless of what format is chosen will greatly benefit all TDM practices. There is concern however about how to incentivise people to adopt existing standards, instead of developing new standards. This also relates to the need for more awareness about the necessity of data management. For example, it is better to apply the standard during the data collection stage, and not to post-process and adapt the data to a certain standard afterwards.

[7] See for example Elsevier's policy https://www.elsevier.com/about/company-information/policies/text-and-data-mining

3.2.3 Use barrier: *tools and infrastructure for TDM*

'TDM is getting better but the accuracy must be high enough so that scientists can rely on it.'

TDM users agree that there are too few easy to use tools that do what users want. This could, however, also be because tools are hard to find, or they exist, but are too expensive. As a result most of the researchers and companies have developed their own tools to fit their specific needs which are seen as a barrier for people who do not have the skills to develop tools or funds available to buy or have them developed.

The specificities of each single research project may not make it possible to have standardised reusable tools available. Available TDM tools may therefore need to be tailored to become suitable. Most people find that documentation of existing software is often missing or lacks clarity so even if potentially useful software is available, people still don't know how to use it because there are no instructions.

From the developer's perspective, technical challenges exist but can be overcome. Their concerns have more to do with the expectations of customers about what TDM can do and how they can use it within their practice. To address the issue of 'no tool fits all' what is often proposed is to have a modular system and a community, such as the open source community, to work together on developing software, making tools interoperable, and improving the tools to address each specific need.

'Ideally what you want is a web based modular system '

3.3 Legal and content

In FutureTDM Deliverable 3.3[8] the following classification was made to identify the legal barriers in the areas of copyright law, database law and data protection law.

- *Restrictiveness:* concerns the legal rules and criteria that in themselves restrict (parts of) mining activities, or only restrict or permit TDM under certain conditions.
- *Fragmentation:* relates to differences and anomalies among national laws and interpretations, but also those between legal regimes or concepts within the laws.
- *Uncertainty:* refers to rules, criteria or concepts in the laws that are not clear.[9]

When asked about the legal and policy aspects of TDM and whether this may cause barriers to the uptake the following issues were mentioned:

- ***Copyright regulation,***
- ***Database protection,***
- ***Access,***
- ***Licensing (attribution, negotiation, fees),***

8http://project.futuretdm.eu/wp-content/uploads/2016/06/FutureTDM_D3.3-Baseline-Report-of-Policies-and-Barriers-of-TDM-in-Europe.pdf
9http://project.futuretdm.eu/wp-content/uploads/2016/06/FutureTDM_D3.3-Baseline-Report-of-Policies-and-Barriers-of-TDM-in-Europe.pdf

- ***Legal expertise (clarity),***
- ***Harmonization,***
- ***European single market,***
- ***Data Protection regulation,***
- ***Enforcement (non-EU companies),***
- ***Commercial and Academic research***
- ***Interlibrary loans,***
- ***Open access publishing.***

3.3.1 Uncertainty, Restrictiveness and Fragmentation

As a result of the lack of legal clarity there is no consensus on the extent to which copyright is a barrier for TDM. Rightsholders say there is a willingness to provide easy access and permission to use copyright protected materials. However in practice the process of obtaining permission from the rightsholders for each individual use proves to be a serious barrier for researchers. They often do not have the time or resources available to negotiate access rights. As a result they refrain from using copyright protected work in their research but instead only use data which is freely available without any restrictions or license or available under an Open Access license.

There are topics which are not covered by researchers because they do not have access to information. Other consequences of the current situation are that people find that research may be biased due to not using all the relevant data.

'We [researcher] would always prefer freely available over data without licence strings attached even if data with licence strings attached was better.'

Many researchers rely on having an institutional affiliation to be able to conduct their research. Institutions provide access to data through their subscriptions to publisher content and through interlibrary loans, although the latter can be a costly and time consuming to get access to necessary materials.

3.3.2 Awareness

When testing the effectiveness of a specific TDM tool most researchers use small scale samples. As a result the researcher may not have any problems getting access to data simply because his TDM remain under a certain threshold. They will therefore not become aware of the problems practitioners face when scaling up to using the TDM tools on real-world big data sets.

'The argument is that if you are researcher on TDM you can solve the question on 500 papers to show it works. You do not need to do it on a million of papers. But these researchers work on the academic level and not in practice so they may not know what the actual problems are because they don't do this on a large scale. '

In an academic project it can be difficult to say beforehand or guarantee what will be done with the data. As a result it can be hard to explain to the rightsholder to obtain permission for the use of the data as there may be no specific purpose.

3.3.3 Access to data

The Companies included in the interviews did not report on copyright as being a legal issue that needed to be solved. They accept that they have to pay for data and access and often rely on their legal advisors to help them gaining permission. Many companies also report taking Google as an example to see what practices are allowed.

'If google has it indexed it is accepted that that is the norm'.

3.3.4 Copyright exception

A proposed solution to the copyright barrier which is discussed at the moment at the EU, is to have an exception for TDM. There was however no agreement amongst the interviewees whether this solution will improve the uptake of TDM. Most publishers do not agree with the proposal for various reasons. One fear among subscription access publishers is that this may lead to a lack of control over who has access. As a result they may not be able to exclude those who do not have lawful access and those who use the works for unwanted and/or illegal purposes.

Research purposes & non-commercial use

With respect to any copyright exceptions for TDM being limited to the research community, this is said to be problematic given the ongoing trend of research cooperation between academia and industry. With multidisciplinary research that involves both academic research and businesses it is not possible to distinguish between research for scientific purposes and research for product development. And it seems to be contradictory to the emphasis policymakers are putting on marketing the outcome of public funded research. This also limits research performed by citizen scientists, who play an increasing role in areas like environmental conservation that rely heavily on observations from the public.

There is also no consensus within the TDM community whether there should be a distinction between commercial and non-commercial purposes. Those in favour of having the exception solely for non-commercial see no problem in making the distinction but most believe that in practice it will become increasingly difficult to separate what use falls under commercial use given the trend towards more public-private partnerships and spin-out companies forming at universities.

'The fear of how someone else is selling the data and taking away your market.'

3.3.5 Licenses

There is uncertainty about what license to use for the data being made available as there is a fear of lack of control over the data. For example because of the uncertainty concerning sharing of data people tend to stay on the safe side choosing restrictive licenses.

3.3.6 Personal data

When it comes to working with personal data, people who are involved in research responded that there was uncertainty about the scope and how to comply with the data protection regulations. Because of this most take the highest level of precaution and only allow anonymized data to be stored for example.

'Copyright exception presumes there is an issue around access to content. Researchers who have lawful access to content are able to text mine with publishers. '

3.4 Education and skill

When asked about skill and awareness aspects of TDM and whether this may cause barriers to the uptake the following issues were mentioned:

- ***Lack of knowledge,***
- ***Skills gap between what industry needs vs what universities provide both in numbers as well as courses,***
- ***Lack of awareness of TDM,***
- ***Need for Data Management plans and incentives to use these,***
- ***Need to understand Rights Clearance.***

'How ethical is it if we do not make research data available wildly to everyone who needs to use it?'

3.4.1 Awareness

Many mention that there is still a lack of understanding and awareness amongst researchers how TDM could be used to improve their research results or help to reduce the time and money spent on finding relevant data. 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. To be able to supply to the increasing demand for TDM practitioners the industry is urging universities to develop courses not only targeted at those who will become text and data miners and developers, but to include courses on TDM in the general educational curriculum.

'If we want to move towards a highly technological and sophisticated society a lot more investment in education and research is needed in general.'

Teaching with open data and open source tools is mentioned as a good practice. Students get positive reinforcement using real datasets and afterwards can continue to apply what they've learned

3.5 Economy and Incentives

In the public consultation on scientific information in the digital age, undertaken by the European Commission, a variety of stakeholders mentioned lack of funding to develop and maintain the necessary infrastructures (80 %) as well as the lack of incentives for researchers (76.4%) as some of the main barriers. When our participants were asked about the economic aspects and incentives for different stakeholders to do TDM and whether this may cause barriers to its uptake, the following issues were mentioned:

- **Data access/availability,**
- **Customers' expectations,**
- **Tools,**
- **Funding and Investments,**
- **Infrastructure,**
- **EU Single market (fragmentation, barriers to entry),**
- **OA business models.**

3.5.1 Market Access

The market for TDM is still very immature, the products and services available are far from being perfect. This is considered a market difficulty by businesses but more so they see it as an opportunity. The challenge is to develop tools and services that meet the expectations and needs of the different stakeholders better than the competition.

3.5.2 Single European market

It has also been mentioned that the idea that very few companies are doing TDM is a misconception. Companies who invest heavily in artificial intelligence (AI) and machine learning (ML) may not specialize in technology but for example Spotify and Zalando are growing European companies whose business models largely rely on data mining.

What companies mention as being a serious barrier is the fragmentation of the EU Market which makes it harder for companies in the EU to grow. There are different regulations, languages and national markets.

'It's possible to do business online, but if you want to develop a network and market presence you still need to open an office in every EU country, which is an investment.'

Some consider that it may be easier if there was one European set of rules but others disagree as the EU legislation tends to be more restrictive than national. It might also introduce additional barriers for using data from web sources.

3.5.3 Access and availability of TDM tools and services

The companies we spoke to say that the availability of TDM tools and the quality of data is not a problem as long as one is willing to invest in the development of these tools and pay for access to high quality data. What the corporate sector is more concerned about is how to deal with data and confidentiality.

'Compared to the academic sector, the corporate sector is willing to pay for solutions '

3.5.4 Academic funding

At the moment there is not enough funding available for academic research on TDM. Researchers have been denied funding for applied TDM research on the basis that their research was not academic research, not focusing on finding new applications. However they all signal a clear need for funding to address domain specific barriers as well as being able to spend parts of their research funding on infrastructure and acquiring data.

'In chemistry and biology for example, research groups often do combined applied and academic research. How much money from your project do you want to dedicate to infrastructure?'

3.5.5 Data access

Access to data for research gets harder when there is a potential to make money.

"They prevent others to exploit their data or at least in a way that undercuts what they want to do themselves. '

Researchers report the pushback they get from publishers but also from public bodies who are now often tasked with making money. In finding a solution often the open access model is mentioned however there is a need for evidence to show how this model can be beneficial and profitable.

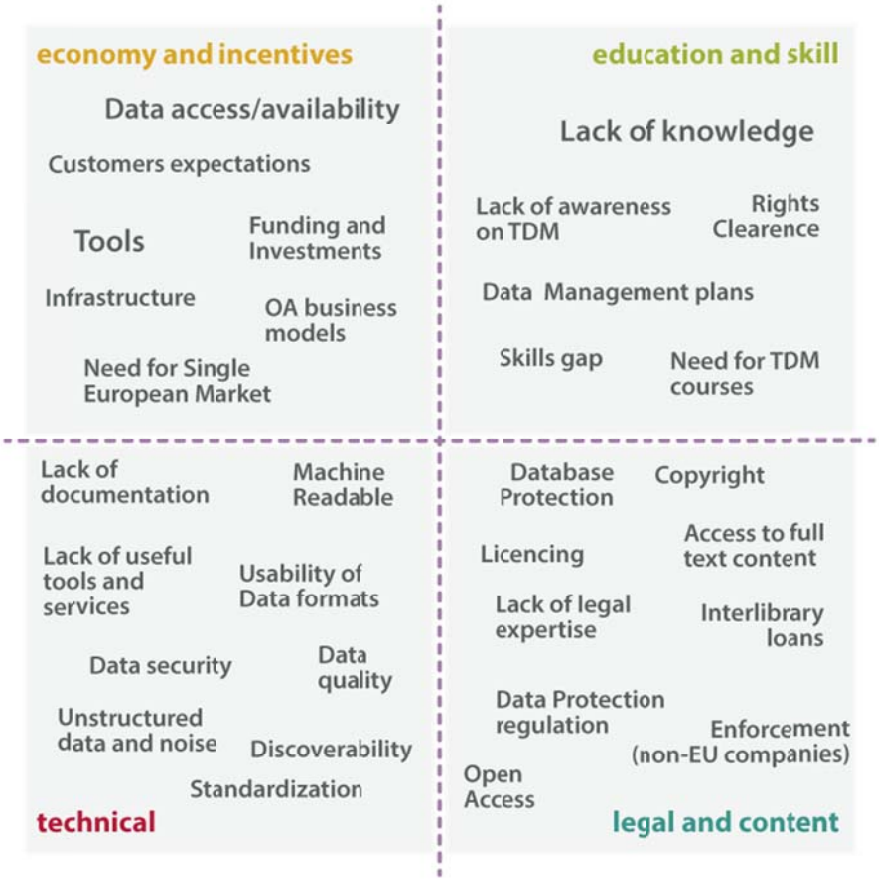

Figure 3: Main barriers derived from the interviews

3.6 Summary and selection criteria

The interviews have provided evidence of the barriers to TDM that different expert practitioners and stakeholders experience in practice. Figure 3 depicts the main barriers that were derived from the interviews.

4 CONCEPTUAL FRAMEWORK AND SELECTION CRITERIA

The aim of the case studies is to provide a better understanding of the issues different stakeholders face that may be a barrier to further uptake of TDM in Europe. Using the interviews as a starting point we have developed seven case studies that help to give a better understanding of some of the specific issues the main stakeholders deal with in practice. In the follow up deliverable we will continue with the analysis taking into account the developments and insights from both the FutureTDM project as well as new developments considering TDM being an emerging technology.

As mentioned in Deliverable D3.1 Research Report on TDM Landscape in Europe, TDM is not defined through a single scientific domain or theory.[10] These case studies therefore also do not intend to cover all of the issues that arise with respect to TDM but are meant to cover the most prominent issues that have been identified throughout the FutureTDM project.

Figure 4: The most frequent terms adopted during the interviews as word cloud

To help improve the knowledge utilization it is proposed to develop case studies of a series of TDM practices (or stakeholders)[11]. The FutureTDM research thus far has employed surveys, content analyses, and other quantitative and qualitative approaches to provide useful understanding. The purpose of this work package is to also take *context* into account and to do so systematically.[12]

Our plan for the case studies involves investigation into the practice of TDM as experienced by the different stakeholders. We will use the interviews as a first starting point and combine this with further collection of data through documents and website review and if necessary a communication follow up with relevant parties involved in the case study practice.

4.1 Selection criteria

For the selection of the case studies we developed criteria based upon insights from the other Work Packages within FutureTDM, in particular from the research deliverables and knowledge cafes carried

[10] http://project.futuretdm.eu/wp-content/uploads/2016/05/D3.1-Research-Report-on-TDM-Landscape-in-Europe-FutureTDM.pdf
[11] Also referred to as KU it aims at improving the use of evaluation evidence in the making of policy
[12] Weiss, Carol H. *Evaluation Research*. Englewood Cliffs, N.J.: Prentice-Hall, 1972.

out with and amongst stakeholders and from the results from the interviews mentioned in the first part of this deliverable (Figure 4 illustrates the most frequent terms adopted in the interviews).

The following criteria were considered relevant:

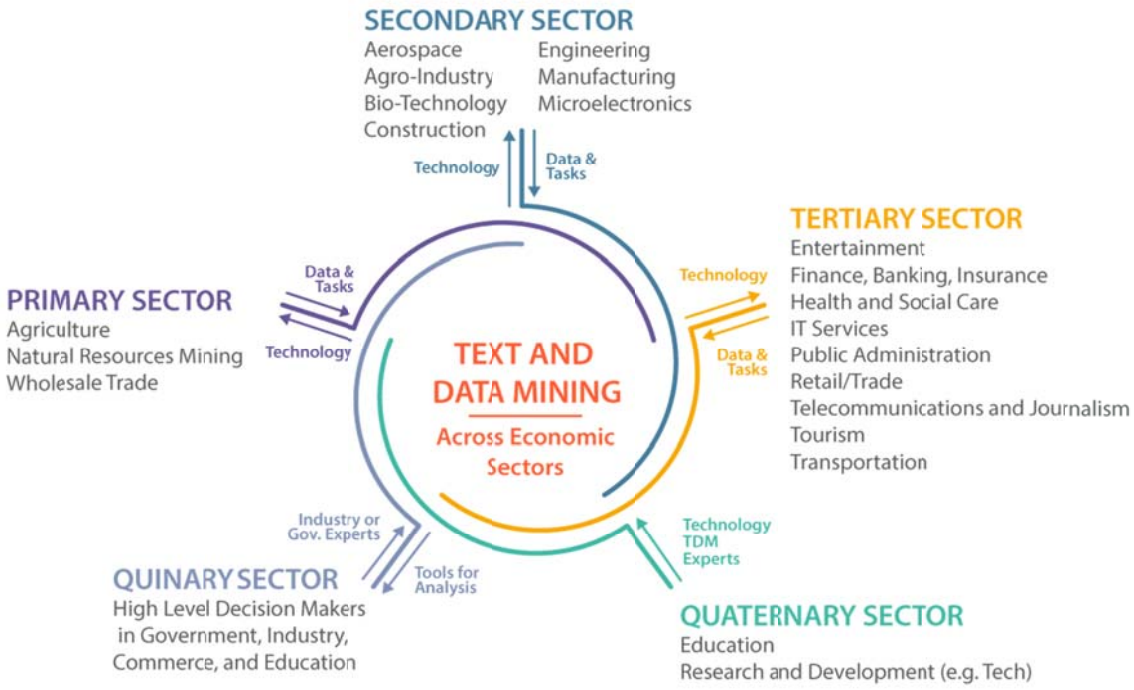

Figure 5: General economic structure and connections between TDM and economic sectors

4.1.1 Sector selection

Figure 5 shows the knowledge-based economy structure where TDM is used across all sectors. Originally, the scope of the FutureTDM project was on the improvement of the uptake of TDM for research environments only. However, analysis of the economy structure has shown that TDM implementation and related research are crucial for all sectors. Therefore, this deliverable focuses on the quaternary sector as it encompasses research, development, and information services that further on affect the growth of other sectors.[13]

4.1.2 Stakeholder selection

The case studies need to be relevant for the different stakeholder communities; therefore they should be representative for the main Stakeholder communities we identified in Deliverable 2.2. Namely:

- Research community,
- TDM content providers,
- Consumers of TDM,
- Funders,
- Policy shapers,
- Service providers,

[13] Busch, Peter. *Tacit Knowledge In Organizational Learning*. Hershey, PA: IGI Pub., 2008.

- Information aggregators and analysts,
- Citizens.

4.1.3 Member State selection

The results from the interviews show there is a differentiation between TDM practices and barriers in different member states. As the project aims to provide recommendations for the increase of uptake in the EU, it is necessary to consider national differences. The interviews gave insight into the existence of potential national barriers and the lack of a single European market. The case studies therefore need to represent this diversity by covering practices in a number of different Member States. The European Union currently has 28 member countries.[14] The aim was not to have specific case studies for all of these but to select cases that illustrate some of the difficulties for TDM in a variety of different member states.

The interviews also included participants from non-European countries. As was often mentioned research does not stop at the European border and the research takes place within international collaborations between the EU and the US for example. We have included a specific non-European case study

4.1.4 TDM process selection

To cover the barriers to the uptake of TDM practices each case study should represent at least one of the following four stages in the TDM process.[15]

- **Crawling and scraping**: this is where the miner searches for the relevant contents they seek to mine and retrieves the information, e.g. by copying it to their own server or terminal equipment.
- **Dataset creation**: The retrieved contents may have to be modified. These contents are extracted to a new (target) dataset that can be used for analysis in the subsequent stage.
- **Analysis;** the dataset is analysed by means of a computer using mining software, according to an algorithm developed or chosen by the miner.
- **Publication:** the TDM user may want to publish the findings from the TDM research. Depending on the purpose of and the context in which TDM is carried out this could include scientific research papers or online journal publication. It could also be circulated only within the closed circle of a company in order to inform decisions.

4.1.5 Selection

Each selected case study covers at least one of the barrier categories. They are presented from one of the perspective of the main stakeholders involved with TDM in at least one of the EU member states. Together these case studies cover the fields that fall under the scope of the FutureTDM project and each of the different steps in the TDM process. Figure 6 and Table 1 represent respectively the geographical distribution and an overview of the case studies taken into account.

[14] http://europa.eu/about-eu/countries/index_en.htm
[15] We refer to D 3.3Baseline report of policies and barriers of TDM in Europe for more information

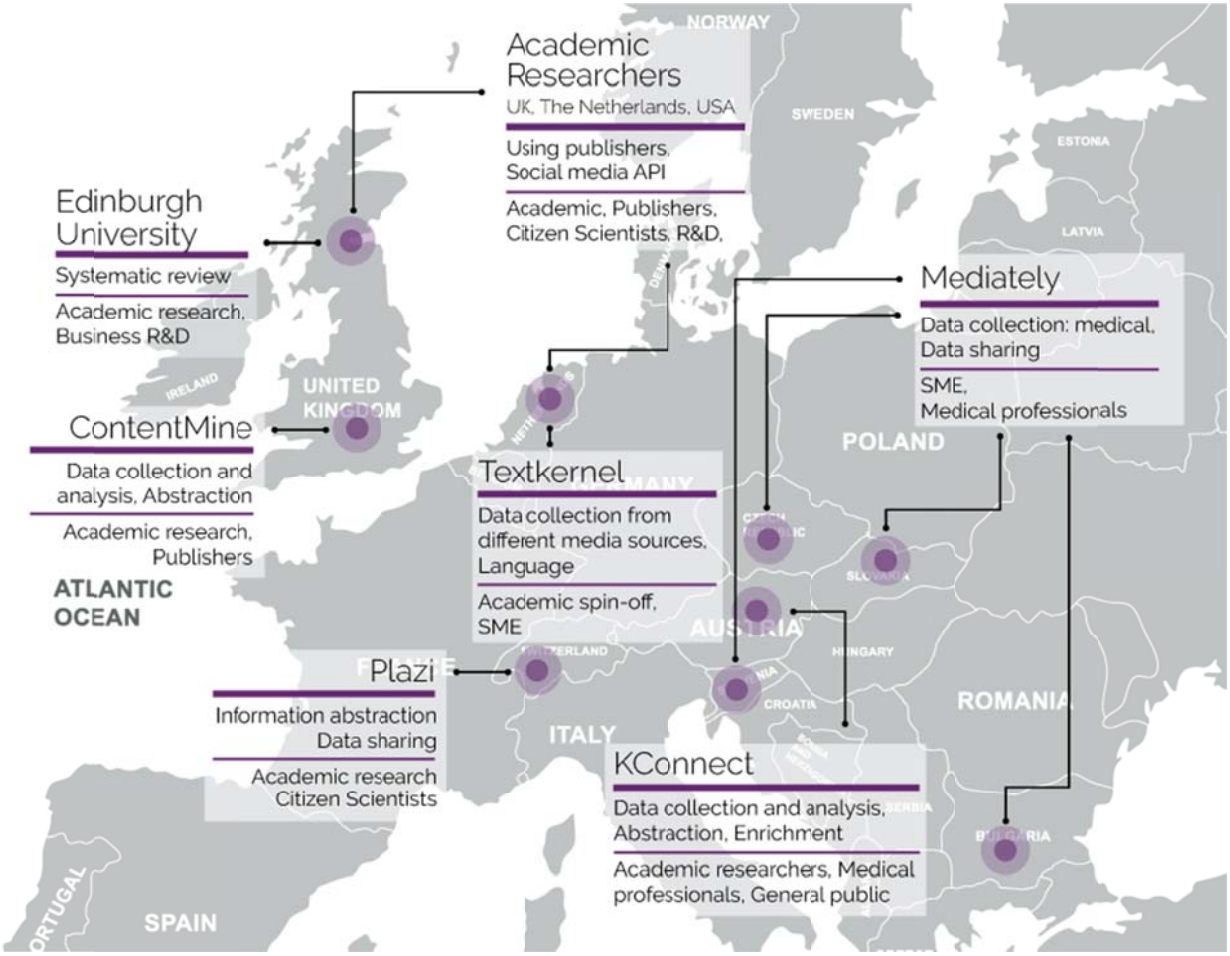

Figure 6: Map of selected Case studies

Table 1: Case studies

	Represented stakeholder	Country	What is the TDM practice	Barriers
5.1	Academic research, Business R&D	Scotland, UK	development and application of systematic review and meta-analysis to the analysis of data from animal studies modelling neurological diseases.	Access to data, data quality, tools, awareness and education
5.2	Academic research,	Switzerland	association supporting and promoting the development of persistent and openly accessible digital taxonomic literature	Access to data, funding for digitization, data sharing, education, awareness, standards
5.3	Academic research, Publishers	United Kingdom	Open-Source cross platform tool for textual analysis. Extracting facts from the academic literature	Access, sharing, re-use, awareness , education
5.4	Academic researchers, Medical professionals, general public	Austria	medical-specific multilingual text processing services, consisting of semantic annotation, semantic search, search log analysis, document classification and machine translation.	Access to data, Data protection, Medical data, confidentiality
5.5	SME, Medical professionals	Bulgaria, Slovakia, Slovenia, Czech republic	medical and health mobile development company	Language, personal data protection, Commercial use, market entry
5.6	Academic spin-off, SME	The Netherlands	software company that specialises in information extraction, document understanding, web mining and semantic searching & matching in the Human Resources sector	Processing textual content in different languages, Single European digital market, competition, standards
5.7	Academic, R&D, citizen scientists	United Kingdom, Netherlands, USA	automatic analysis and extraction of information from large numbers of documents.	Access, data quality, tools, education, awareness, standards

5 CASE STUDIES

The following section provides an analysis of seven different case studies that cover one or more relevant TDM practices, barriers and enablers that different stakeholders have experienced.

Each case study starts with a brief introduction into the activity description (business model or scientific research); followed by the main issues being presented for further analysis of the barriers that were present in this specific case. At a later stage Deliverable 4.5 will also include possible ways to deal with the barriers through providing best practices and methodologies.

5.1 TDM and systematic review

The following case study focuses on the issue of academic research from the perspective of a research consortium looking at the use of TDM to improve systematic reviews of the scientific and medical literature.

5.1.1 Research at Edinburgh University

In healthcare a huge amount of research is produced each year. There are 1.3m new publications published a year in biomedical science alone. It is simply not possible for humans to understand and aggregate all the information there without machine learning intelligence to go through all of it.

The results of different studies often have conflicting findings. This could be the result of study differences, flaws or chance (sampling variation). When between study differences exist it is not always clear which results are most reliable and should be used as the basis for practice and policy decisions. Using systematic reviews it is possible to address these issues by identifying, critically evaluating and integrating the findings of all relevant, high-quality individual studies that cover one or more research questions.

'Using systematic review, we can identify all publications and find relevance to research and research questions.'

A research group led by a Professor of Neurology and Translational Neuroscience at Edinburgh University[16] has shown that much of the research published is at substantial risk of bias. As a consequence the effects observed in animals of particular research may be substantially overstated.[17]

This is a problem because future research – further laboratory work or taking new treatments to clinical trial – is then based on a false premise and is less likely to succeed. For laboratory research, this is a waste of money, time, and animal lives. For clinical trials, human subjects may be put at risk.

[16] http://www.ed.ac.uk/clinical-brain-sciences/people/principal-investigators/professor-malcolm-macleod

[17] The university is also part of CAMARADES an international collaboration which aims to provide a central focus for data sharing. It uses SyRF as NC3Rs *in vivo* systematic review and meta-analysis facility and it aims to provide an easily accessible source of methodological support, mentoring, guidance, educational materials and practical assistance to those wishing to embark on systematic review and meta-analysis of data from in vivo studies

The research group is now looking at developing tools to provide unbiased summaries of what is currently known. And to develop tools that can assess whether indeed the effects in animals are overstated, by comparing results with existing research. Their aim is to then use this information to help guide better design of clinical trials testing treatments in humans.[18]

5.1.2 The application of TDM

If we look at the amount of new research emerging, every week around 3500 new pieces of research involving animal research are published and it seems almost impossible for anyone to stay up to date.

To illustrate the scale of the problem, a specific query can give 80.0000 hits from which only 4000 may turn out to be relevant for the research.

To find these relevant publications, a researcher has to first go through all these hits. However a physical screening of all the material is almost impossible because to screen all the hits by hand it takes a year or two at least by which time your results will already be out of date. Text mining and machine learning can be used to help find publications that include experiments with potential relevance.

The next step is to establish the actual relevance of these publications. However at the moment the tools to establish relevance are not good enough. The research group is currently testing what is available on the market of TDM tools and services. At the moment they have been unsuccessful in finding a company that can actually provide the tools they need to be able to provide what they need. Companies are offering TDM solutions but the outcome of their services is not sufficient. When trying to discuss the development of tools, there seems to be a reluctance to share code and/or solutions or to work together to improve results.

The final step is to actually try to extract the outcome information from the experiment. This is proving to be difficult in practice, for example it is challenging to abstract information from tables and images when these are used instead of text.

'We can get reasonable performance on one dataset but when validating this on another dataset the results are not great.'

There are well established TDM approaches for enriching search results which reduce the amount of screening by 50%. However the aim of the research is to achieve a 90% reduction.

5.1.3 Barriers

Technical and Infrastructure

The group expects that full text access together with a deep learning approach will get substantially better results from TDM. At the moment the only way to identify relevant publications is by going through abstracts, but what is needed are raw PDFs with a title, abstract and the full text of the publication. The issue is that to get the full text PDF and extract outcome data in an unsupervised

[18] Examples of trials they have helped design include EuroHYP-1 - a trial of brain cooling in stroke - and MS-SMART, a trial in secondary progressive multiple sclerosis.

way is impossible. Getting them in a supervised way is sometimes possible, but the technology is not at that stage yet.

Another barrier is the lack of tools available that produce reliable results. At the moment they are screening companies who provide these services but the results are not great. Having an open source modular system would help.

Legal and content

There is an argument for the explicit purpose of systematic review that journals could make their content available for review even if it's not available for reading. The review system has a hierarchy: you can use the abstract which is free or you can buy to read the full paper and/or you can get a TDM license. The argument is that with a TDM license or copyright exception for this specific purpose publishers could make the item available provided that the user is mining the PDF rather than retrieve the PDF.

'Having a copyright exception, things would be easier. We could persuade journals to do this because if we get together all the available data, we can develop a better system, one that is not biased.'

Being based in the UK the research group relies on the exception to copyright for non-commercial TDM practices. The exception, which became effective on 1 June 2014, allows for 'computational analysis' to be carried out legally on material under copyright. This, means that you can do TDM if you have lawful access to the source material and the analysis is undertaken for the purposes of non-commercial research.

The research group reports not having a problem getting access since they can use their university library subscriptions to get the content. It may even be a benefit as many other member states do not have such an exception and are interested to work together with UK researchers.

There is also an issue that not all publications are available through their university subscriptions so they must use interlibrary loans. If a publication is not available they will have to purchase it through interlibrary loans which will cost around £4 per publication. In addition to the costs and time it takes to manually put in the request, the time it takes to receive the publication is long. While this is a hindrance in institutions which enjoy subscriptions to a wide range of journals, for smaller institutions it is a major barrier, and stands in the way of the democratization of science.

Another consequence of the current legal framework is that they cannot share the full results with anyone outside of the institution who does not have the same access subscriptions.

Education and Skills

Using TDM to compile a review of the available literature will provide researchers with more informative and thus better knowledge of the field. However, the implementation of TDM practices for the specific field requires proper understanding of TDM potential and limitations in terms of text processing and data mining, as well as proficiency in the field in question. This combination of skills is rare, and requires both additional educational investments at the University level and personally from the scientists within the project.

Economy and Incentives

The economic barriers that were mentioned had to do with having to rely on companies' willingness to share the working of their systems. At the moment the research has to pay for commercial TDM services without knowing whether they can provide the right solutions. It would be more beneficial to be able to work on developing systems together but there is reluctance from the commercial sector to do so. This could be explained by the competitiveness of the market in providing solutions.

With respect to getting research funding the impression is that if you sell the project well there is funding available.

5.1.4 Conclusion and further study

The purpose of using TDM for systematic review is to make the review more trustworthy and less time consuming. An additional outcome of better systematic review and coverage of all relevant publications is that for researchers and authors in general, they will their data cited more often.

The problem this case study illustrates is that not having access to the full text is a barrier. Ideally there would be a copyright exception or a separate license for TDM for the purpose of systematic review.

Another proposed solution to promote machine learning and text mining as a method is to make it freely available for researchers and SMEs in Europe and on subscription to other companies in other countries.

5.2 TDM and Biodiversity conservation

5.2.1 Information abstraction from biodiversity literature

Global biological diversity is increasingly threatened, making precise and detailed data on biodiversity necessary in order for numerous organisations to provide convincing arguments for conservation and biodiversity management.[19] A large part of our knowledge on the world's species is recorded in the corpus of biodiversity literature with well over hundred five million pages. 17,000 new species are described per year, in many cases based on the 2 – 3 billion reference specimens stored in thousands of natural history institutions. This body of knowledge is almost entirely in paper-print form and though it is increasingly available online, it is rarely in a semantically structured form, rendering access cumbersome and inadequate from the perspective of researchers.

For example, finding relevant literature on a given species is often extremely difficult. This is mainly because there is no comprehensive, global bibliographic database of the publications and no index to the specific taxonomic treatment of species, despite the maturity and ubiquity of a global scientific naming and classification system for species. Searches for a particular name therefore tend to result in a huge array of irrelevant data (e.g. mere citations, or other references to topics that are not relevant for the understanding of the description). Only for a few groups of species is there a

[19] See CBD the Convention on Biological Diversity. [http://biodiv.org] and Target 2010. [http://www.countdown2010.org]

complete species catalogue and access to digital versions of the related literature available, but for the majority of species, including well known groups such as birds and fish this is not the case .[20]

5.2.2 Plazi.org

History: Plazi is an independent not for profit organization.[21]

Aim: The goal of Plazi is to produce open access, semantically enhanced, linked taxonomic documents whose content can be harvested by machines, taxonomic treatments and observation records that can be cited and provided in various formats from HTML to Linked Open Data (RDF), and with this contributing to the Global Biodiversity Knowledge Graph. The motivation is to bridge the gap from a scientific name to what has been published about it.

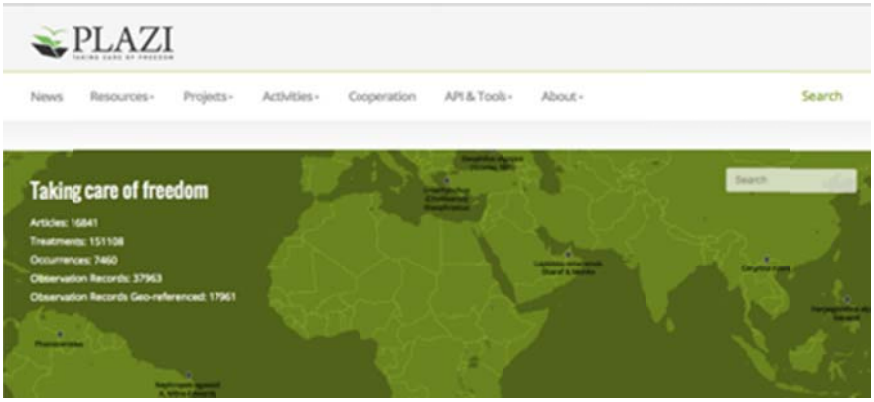

Figure 7: Plazi Home Page

TDM practices: The workflow of Plazi relevant for our research on TDM is as follows:

Their workflow (Figure 8) begins by discovering documents that have not yet been included in their system or that are part of a body of publications to be mined. For a select number of journals this is a fully automated process from scraping the WWW to mine and expose the treatments and facts therein. For those journals, especially those where a born digital PDF is available (that is an idiosyncrasy in taxonomic publishing whereby the PDF is a prerequisite to create available names for taxa new to science), the bibliographic metadata is extracted from the publication semi-automatically and added as (Metadata Object Description Schema) MODS header into the interim XML document

[20] See for example the Antbase.org. , [http://antbase.org] for ants.
[21] Plazi. [http://plazi.org]

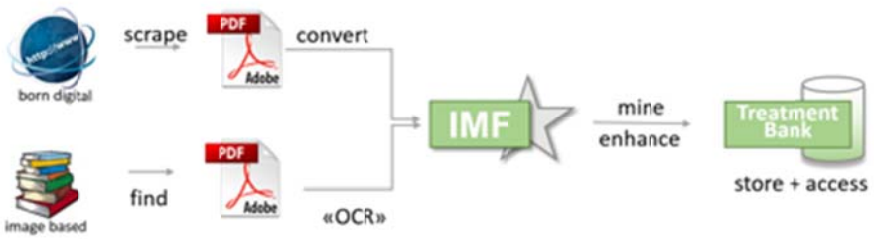

Figure 8: Plazi workflow

The workflow begins either with born digital PDFs or PDFs that are based on scanned page images. Once the documents are converted the files are stored as IMFs. The facts are served from a database.

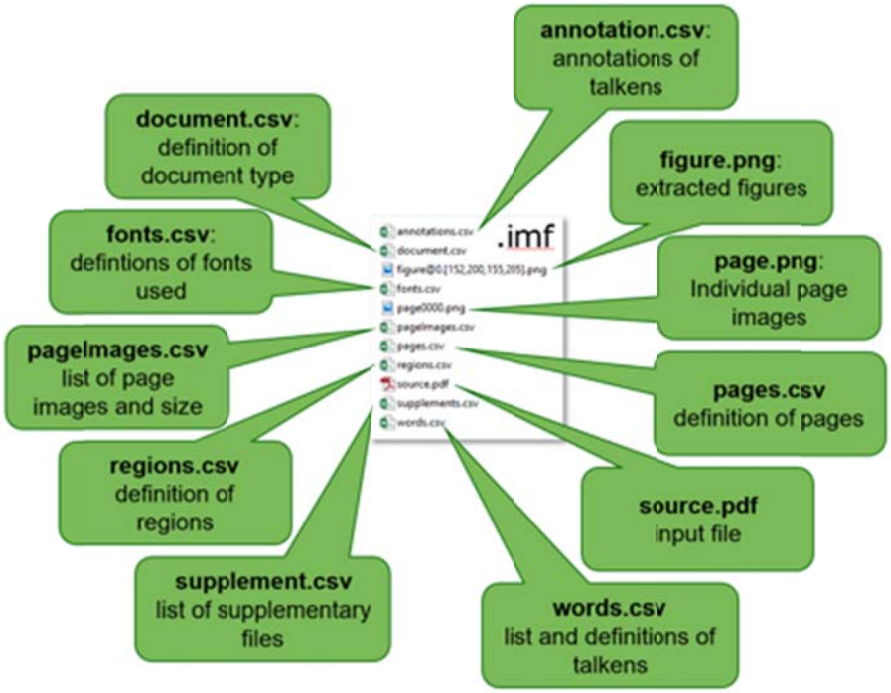

Figure 9: Image Markup File (IMF)

The IMF (Image Markup File, Figure 9) is a container that includes the original file, page images as bases for further linking text to the respective bounding boxes of the tokens to allow editing on the page image, and all the annotations, including links to external resources. The IMF can be read by GoldenGATE Imagine, and the data can be exported from there in various formats (e.g. XML).

The original PDF is uploaded to the Biodiversity Literature Repository (http://biolitrepo.org) at Zenodo/CERN, and a DOI is created in case none is available to cite the article. This way, all the facts can be linked to a digital copy of the cited bibliographic reference

After removing all OCR- and printing artifacts as an initial step in both pathways, the bibliographic references are detected, and marked-up, as well as the citations of bibliographic references in the text linked to the bibliographic references. All the bibliographic references are exported to RefBank, a bucket to collect bibliographic references (http://refbank.org) now including over 600.000 references. Similarly all the tables and images are detected and exported, the captions are marked

and table and figure citations are linked to the captions that will be enhanced with a link to a digital representation. In a next step, all the taxonomic names are tagged and enhanced with their related higher taxa using the Catalogue of Life (http://www.catalogueoflife.org/) and the Global Biodiversity Information Facility (http://gbif.org). Afterwards, all the taxonomic treatments, a dedicated section of an article that includes facts about a particular taxon, are identified. Treatments can then be subdivided into semantic elements. These steps can be highly customized allowing a fully automatic processing from scraping the Web to expose the facts on TreatmentBank. The daily input of new taxa, mainly based on this system is available here (http://tb.plazi.org/GgServer/static/newTodayTax.html) and represents ca 4,800 taxa or 30% of new discovered taxa per annum, and in total > 20,000 treatments. Converting an entire journal run (Zootaxa, 18,000 born digital documents) had a yield of 71% of fully automatic conversion, resulting in 90,000 treatments, 130,000 extracted images, and 200,000 bibliographic references. Using "pluggable architecture", allows Plazi and collaborators to continually improve the automation by the development of software plug-ins written to a published Application Programming Interface (API).

After the mark-up, the documents are uploaded to TreatmentBank. All the marked-up data elements will be saved in respective fields, including the metadata of the publication, to guarantee the provenance of each element.

The markup process is based on GoldenGate's internal XML that can be exported in various flavours such as XML (Figure 10) or RDF, for which a complementary vocabulary is developed for elements that are specific for taxonomic treatments (https://github.com/plazi/TreatmentOntologies). For the remainder, existing vocabularies are used.

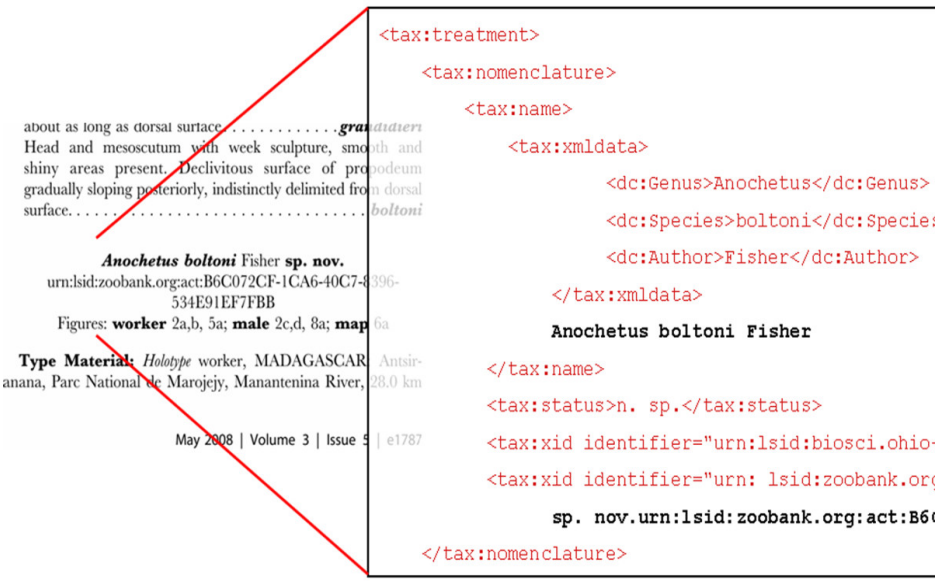

Figure 10: Sample markup page. Left: sample of an original, published taxonomic treatment. Right: Same treatment marked-up in TaxonX XML schema and enhanced with external identifiers.

All application programs used by Plazi are open source except for the commercial ABBYY Finereader. This includes both for those supporting their internet services[22] as well as those created by Plazi

[22] e.g. DSpace, Postgres, Simile and eXist

themselves (GoldenGATE and its plug-ins, SRS), which are licensed under the Berkeley Software Distribution license[23].

5.2.3 Research challenges

As described in the previous sections, the Plazi workflow aims at transforming printed text into semantically enabled documents from which taxonomic information can be extracted.

Their content comes from scientific taxonomic publications, particularly the taxonomic treatments and single materials citations in these publications and from external databases like taxonomic name servers, specimen databases, and bibliographic services. Species names, treatments and other data as well as bibliographic identifiers are then assembled in a publicly accessible repository. For all those elements the source is cited, including the actual page number and if possible sufficient machine-readable data to allow software to locate the original, or at least a digital copy, of the publication.[24]

5.2.4 Legal barriers

The legal barriers Plazi encounters is whether their process of abstracting the data is compatible with existing copyright rules. Is it possible to extract species names and descriptions from protected material without infringing copyright? Secondly, whether they are allowed to make the assembled data available to the interested public?

Data Sources

Copyright protects scientific information such as books and articles when it qualifies as a "literary and artistic work" in the sense of copyright law (art. 1 Berne Convention).[25] Although there is no legal clarity about the scope of protection, Plazi considers that the information in the Plazi's Search and Retrieval Server (SRS) namely the taxonomic treatments as well as the metadata of the publications are not copyright protectable but part of the 'public domain' (Agosti and Egloff 2009).[26]

Taxonomic treatments are formulated in a highly standardized language following highly standardized criteria. They adhere to rules and predefined logic. They are not "individual", nor "original" in the sense of copyright law. The same applies to biological nomenclature which follows standards established by various Commissions installed by the biological community.[27] Text written in accordance with these nomenclatural systems is not individual and cannot qualify as work.

[23] Sautter G, Böhm K, Agosti D: A quantitative comparison of XML schemas for taxonomic publications. Biodiversity Informatics. 2007, 4: 1-13.View ArticleGoogle Scholar

[24] The act of publishing is one of the key criteria required by the Codes governing biological nomenclature to complete a valid 'nomenclatural act', i.e. to create a valid scientific name for a new discovered species.

[25] Berne Convention. [http://www.wipo.int/treaties/en/ip/berne/trtdocs_wo001.html]

[26] Agosti D, Egloff W 2009. Taxonomic information exchange and copyright: the Plazi approach. BMC Research Notes2009,[2:53].DOI:10.1186/1756-0500-2-53,accessed http://bmcresnotes.biomedcentral.com/articles/10.1186/1756-0500-2-53#CR6

[27] including the International Commissions for Zoological Nomenclature ICZN for Botanical Nomenclature (ICBN and for Fungal Nomenclature (Index fungorum). All these aim to preserve logical schemes and structures that are predefined by the scientific community according to pre-established objective criteria.

Data extraction

Plazi creates its database from taxonomic literature that may be copyright protected. The main copyright question with respect to Plazi is therefore quite simple: Is Plazi permitted to extract data from a protected work?

As mentioned before the Plazi workflow includes the reproduction of documents. Works are scanned, they are semi-automatically marked-up and they are processed by algorithms in order to make extraction of names, treatments and finer grained information possible. Texts or pictures will repeatedly be reproduced during this process. For Plazi to be fully effective, it must be able to operate against the full body of taxonomic literature. It is not technically practicable to seek individual permissions on a case-by-case basis. The process concerns millions of documents. Neither can the extraction process be limited, say, to documents published under a copyright waiver. The only feasible solution is to work on the basis of legal licences.

International Collaborations

Plazi is possible because of the exception in Swiss copyright law which allows temporary acts of reproduction, when the copies are transient or incidental, and are an integral and essential part of a technological process, as far as the purpose is to enable a lawful use of the work and for non-commercial purposes,[28] or for art. Swiss Author's Rights Law allows one to download and to reproduce protected works for internal use in administrations, public and private bodies and other institutions.

The Plazi workflow is conceived following these Swiss copyright rules: works are copied several times during the markup and the extraction process, but the copies are only transient. As a result of this process, Plazi presents scientific data and metadata from original sources, including published scientific illustrations, which they do not consider to be work in a legal sense, but not the works themselves. Literary and artistic works such as scientific publications remain restricted to internal use as long as they are stored only for the markup and extraction process. No further use is made of the transient copies used for the extraction process. Therefore, the Plazi workflow is covered by the aforementioned legal exceptions to copyright.

If Plazi was based in any EU country it would have been impossible. By having their base in Switzerland they can also avoid database right protection. This legal instrument, laid down in the Directive 96/9/EC of 11 March 1996 on the legal protection of databases protects databases, "which show that there has been qualitatively and/or quantitatively a substantial investment in either the obtaining, verification or presentation of the contents" through a so called "sui-generis-right".[29] This right allows preventing extraction and/or re-utilization of the whole or of a substantial part of the contents of that database.

This European Database Directive is therefore a serious obstacle to scientific information exchange. That's why Plazi organizes its work in a way that excludes the application of European database

[28] art. 24a Swiss Author's Rights Law implementing art. 5 (1) of the European Directive 2001/29/EC of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society.

[29] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases

protection. The whole workflow, as well as the storage of documents, is based on Swiss law, which does not provide such particular database protection.

Access

Plazi has encountered legal issues concerning the sharing of data.[30]

All the data used by Plazi is published data with publications going back to the year 1756 as the beginning of taxonomy. Anything published after that which is scientific and follows their code becomes part of the system.[31]

With respect to making the assembled data available, Plazi does not make the protected works from which the material may be extracted available. Instead, they present scientific data which is not under copyright and properly cite the containing material.

That copyright can still have a negative impact is clear in the case of the Biodiversity Heritage Library (BHL), a large scale effort to digitize all the biodiversity literature stored in the large US and UK natural history institutions.[32] BHL policy is not to scan and include anything that is presumed to fall under copyright and for which the rights have not been cleared. As a result most information in BHL is outdated as it does not hold any publications younger than 65 years. The more recent publications in biodiversity literature – about 20,000 descriptions of new species each year and an estimated fivefold that number of re-descriptions – are only available to a privileged group of subscribers.[33]

5.2.5 Technical barriers

If we compare TDM in other fields the uptake is low because of a lack of structured data, the tools to mine the data and a lack of shared ontologies. With few exceptions, none of the taxonomic data finds its way into PubMed, the main source for TDM in the biomedical field. There is a huge amount of data which is on people's desk like copies of articles which are not online. This is a problem because this data is then neither accessible nor citable.

'The problem of TDM is that it does not follow the way science works. Our (biodiversity) literature is not made for it - it's almost impossible to get a machine to read it. '

An alternative to the full text search is to embed domain specific markup, such as elements delimiting and identifying scientific names, individual treatments, or materials citations, essentially modeling the logical content. This is available and implemented through a collaboration with the US National Library of Medicine, the Bulgarian Publisher Pensoft and Plazi. The 12 journals by Pensoft use a biodiversity domain specific Journal and Archival Tag Suites version (TaxPub: Catapano et al.

[30] Willi Egloff & Donat Agosti Plazi, Bern (http://plazi.org) Globis-B Workshop Leipzig, 29.2./2.3.2016 Data Sharing Principles and Legal Interoperability

[31] *Plazi has an agreement with ZENODO to make everything up to the year 2000 accessible. This data is chosen somewhat arbitrary but up to 2000 nobody was asked for a cease of rights so nobody could complain and also the data is old enough for it not to be interesting commercially.*

[32] Biodiversity Heritage Library. [http://www.biodiversitylibrary.org]

[33] Polaszek A, 25 co-authors: A universal register for animal names. Nature. 2005, 437: 477-10.1038/437477a.View ArticlePubMedGoogle Scholar

2012)[34]. However, marking-up the literature which is already been published can be expensive and time consuming, not least because of the complexity of PDFs and even more so the uncontrolled scanning of the hundred millions of pages of legacy literature in various languages, fonts, paper quality.

What is needed are new models for publishing taxonomic treatments in order to maximize interoperability with other relevant cyberinfrastructure components (e.g., name servers, biodiversity resources, etc...)

'*We are a data broker: We take unstructured data and make it structured and accessible. '*

5.2.6 Economic and Incentives

Currently, there is no market and thus no business model that allows building a company that provides this conversion service. Scientists depend on abstracting services such as the Index of Organism Names by Thomson Reuters (http://www.organismnames.com/) which are neither complete nor timely - but better than anything else. Catalogue of Life (http://www.catalogueoflife.org/) is neither complete nor provides a near time service for new species. Traditionally, there are taxonomic group specific services, such as the World Spider Catalogue or Hymenoptera Online, and they cater for a very specific community, not to a global "market", nor is their focus on facts in the cited articles and treatments, but rather metadata.

The economic barriers have to do with the funding of digitization projects and the missing vision and drive to build a global name service that provides immediate access to all the new published scientific facts about species.[35]

'*We cannot get the data unless there is funding for it'*

5.2.7 Education and skill

The Plazi members are also involved in advocacy. Mentioned during the interviews were a lack of awareness of researchers and students for the need of open access and the need for data management by researchers making their data available for re-use.

For example there is still a huge amount of data which is on people's desk like copies of articles which are not online. This is a problem because this data is then neither accessible nor citable.

'Our problem is not copyright our problem is attribution, scientists want to make sure they get attributed.'

5.2.8 Conclusion

TDM is used to abstract the data out of publications and make it available for research and innovation. The technical barriers described in this case are not having structured data. The extracted

[34] Penev L, Catapano T, Agosti D, et al. Implementation of TaxPub, an NLM DTD extension for domain-specific markup in taxonomy, from the experience of a biodiversity publisher. In: Journal Article Tag Suite Conference (JATS-Con) Proceedings 2012 [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2012.

[35] 'Horizon 2020 overestimated the the capacity and status of the data and those delivering data. Instead of focusing on the use of data they should be focusing still in making data accessible' Personal communication May 2016

data can be shared but not the full text with markup. Also it is difficult to collaborate on the extraction process outside of Switzerland and the UK due to the lack of harmonization of copyright exceptions.

5.2.9 Project Recommendations

Being dedicated to making the sharing of research data possible, Plazi is working on finding solutions. They advocate and educate the community on maintaining free and open access to scientific discourse and data. What they consider to be of vital importance[36]. Based on their experiences they proposed the following principles: [37]

- Essential Biodiversity Variables should be shared as Open Data, making them available as part of Data-CORE without charge or restrictions on reuse
- Exceptions should be limited to sensitive data: Data whose free accessibility could endanger certain aspects of biodiversity conservation; Data that are qualified as confidential by the competent authority
- Right holder(s) of research data (if any) should dedicate them to the public domain (by CC0-waiver, CC-BY-License or any similar instrument)
- Data, products and metadata should be made available with minimum time delay

5.3 TDM and CONTENTMINE

5.3.1 Introduction

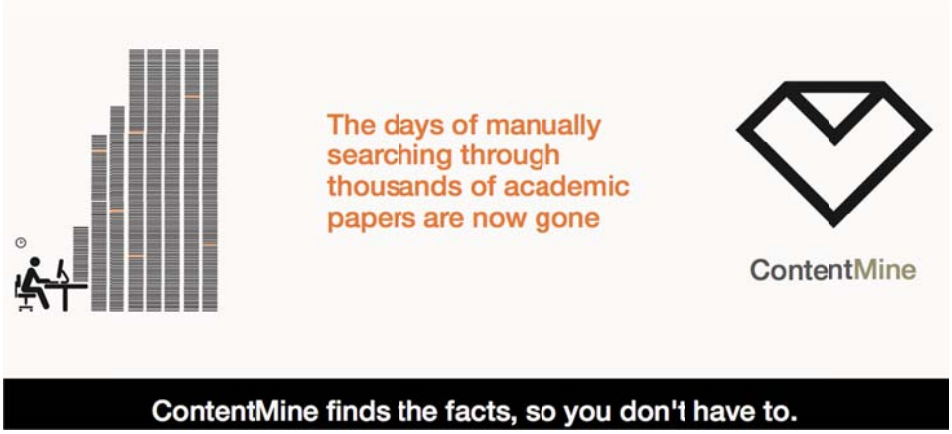

Figure 11: ContentMine website screenshot. Copyright ContentMine Ltd, licensed under CC-BY 4.0.

ContentMine (Figure 11) is a UK non-profit organisation founded by Dr Peter Murray-Rust, a chemist, molecular informatician and advocate for open science. Murray-Rust faced barriers throughout his career in trying to apply his TDM technologies to the scientific literature.

In 2014 the South African philanthropic funder Shuttleworth Foundation supported him with a two-year Fellowship to set up the ContentMine project, which initially sought to liberate 100 million 'facts', mostly named entities, from the scientific literature. The project also ran TDM training

[36] Agosti D, Egloff W (2009). "Taxonomic information exchange and copyright: the Plazi approach" (PDF). BMC Research Notes 2:53: 53.
[37] Agosti D, Egloff W (2009)

workshops for researchers to promote the usefulness of TDM to researchers facing overwhelming levels of content, reaching around 300 researchers at over 20 workshops. Talks by Murray-Rust and the ContentMine team reached an estimated audience of 2000, promoting the concept of content mining (as a more inclusive term than TDM) and its utility across a wide variety of disciplines. Murray-Rust was heavily engaged in advocacy for the idea that 'the right to read is the right to mine', a phrase that was later picked up by organisations such as the Wellcome Trust and LIBER in their advocacy and policy work around TDM aiming to give subscribers to scientific articles the right to read them using a machine without seeking additional permissions.

'The days of manually searching through thousands of academic papers are now gone'

5.3.2 Aim of the project

The major aim of the project and resulting non-profit was to set up a daily feed of 'facts' by accessing a high proportion of the full-text scientific literature via publisher and content providers' application programming interfaces (APIs) and by scraping from websites where necessary. Initial efforts focused on the Open Access literature but the introduction of a UK copyright exception for TDM for non-commercial research in 2014 reduced some legal barriers to use of the closed access literature and the project is now planning to implement a daily pipeline of open data in the form of species names, word frequency data, human genes and other facets in collaboration with librarians at the University of Cambridge.

5.3.3 Technical and Infrastructure

Technically, the barriers reported by ContentMine are related to the heterogeneity of publisher XML and HTML, even when it conforms to a technical standard such as NISO JATS. In order to produce a normalised corpus of articles for easier semantic tagging, custom web scrapers and XML style sheets must be constructed on a publisher by publisher basis, a challenging task for an individual researcher or group. Members of the team have also found multiple instances of publisher barriers such as captchas to prevent bulk downloads and 'traps' such as fake DOIs, which alert the publisher to mining activity or in some cases automatically cut off access from the relevant IP range.

5.3.4 Legal and content

The legality and ability of researchers to challenge the technical measures is unclear even under the UK exception and represents another barrier beyond statutory legal barriers. Nonetheless, a copyright exception that cannot be overwritten by contract was viewed by ContentMine as a major enabling factor.

5.3.5 Economy and Incentives

Although the organisation is based in a country where a statutory law allows non-commercial use and it is a mission-driven non-profit, finding income streams to remain sustainable and develop software without relying on public funding is challenging without an allowance for commercial use. Without approaching publishers one by one to negotiate permissions, which would be challenging as a lean organisation, ContentMine cannot charge researchers for access to its stream of facts as this would likely be viewed as commercial use. It also cannot produce useful insights that it could sell to organisations for money and is therefore restricted to leading or collaborating on grant-funded research projects or offering consultancy services. While these are valid business options, the ContentMine view based on extensive organisational brainstorming about potential routes to

sustainability and impact delivery was that restricting allowable business models limits delivery of innovative ideas and economic impact.

5.3.6 Education and Skill

A lack of awareness of TDM was a barrier to the work of the organisation. It was clear from discussions between the training team and workshop participants that many researchers lack the skill base to work with highly technical or command line tools and there is a gulf between the types of techniques and protocols they are used to applying and the approaches typically taken by academic TDM groups, who get academic credit for the quality of the mining rather than user interface design. Many groups were doing large scale literature reviews entirely manually at great expense and effort and the learning curve was a substantial barrier regardless of legal status.

5.3.7 Conclusion

This case study exemplifies the types of research activities that have been positively enabled by removal of legal barriers but are still impeded by non-legal factors and threatened by lack of sustainable funding models even in a non-profit context.

5.4 TDM and Search technologies for medical information

'Radiologists are drowning in images. At larger hospitals over 100GB (over 100'000 images) are produced per day.'

5.4.1 The Khresmoi project

The main goal of the Khresmoi project is to limit the information overload[38] of radiologists and other clinicians caused by an increasing number of images and an increasing complexity of radiological protocols. For this they developed a multilingual multimodal search and access system for biomedical information and documents.[39]

The idea is to explain the data viewed in a better way, including:

- the use of past cases and recent publications;
- indexing databases of medical images;
- understanding problems of real life patient data in terms of data quality, anonymization, and pre-treatment;
- data reduction when storing 3D and 4D datasets and their visual features through concentrating features on regions different from healthy models.

They were able to achieve this through

- Automated information extraction from biomedical documents, including improvements using manual annotation and active learning, and automated estimation of the level of trust and target user expertise

[38] This problem was identified through a large scale survey. Online health information search: what struggles and empowers the users? Results of an online survey. Natalia Pletneva, Alejandro Vargas, Kostantina Kalogianni and Célia Boyer Stud Health Technol Inform., 2012

[39] This project was supported by the European Commission under the Information and Communication Technologies (ICT) Theme of the 7th Framework Programme for Research and Technological Development

- Automated analysis and indexing for medical images in 2D (X-Rays) and 3D (MRI, CT)
- Linking information extracted from unstructured or semi-structured biomedical texts and images to structured information in knowledge bases
- Support of cross-language search, including multilingual queries, and returning machine-translated pertinent excerpts
- Adaptive user interfaces to assist in formulating queries and interacting with search results

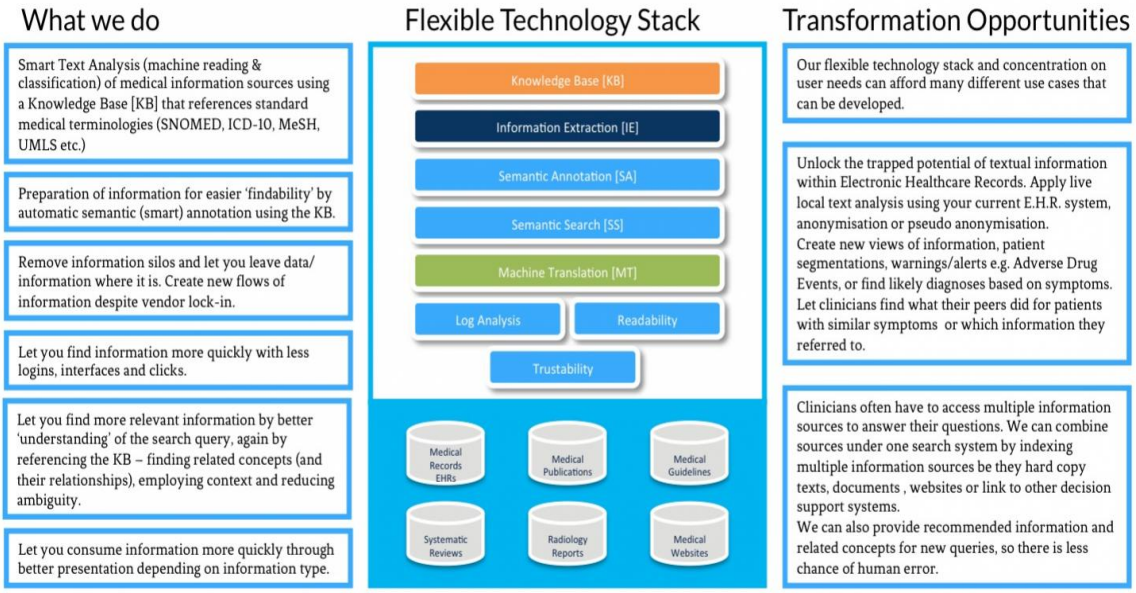

Figure 12: KConnect workflow

5.4.2 KConnect

Khresmoi continued in 2015 as *KConnect,* to bring the developed medical text analysis and search technologies to the market. Figure 12 shows the projects developments including a flexible technology stack that can handle a variety of medical information resources including EHRs, medical publications, best practices and treatment guidelines, systematic reviews, indexed web pages etc.

5.4.3 Text mining and analysis

The ability to search over a number of medical information sources/systems means information is no longer held in silos but people can have access to the most relevant and up-to-date medical information. KConnect provides Medical Text Analysis, Semantic Annotation and Semantic Search services aimed at healthcare professionals, researchers in the biomedical industry and the public.

Text analysis, classification and semantic annotation with the aid of a large medical Knowledge Base allows for improved search (semantic search) results. Text analysis can also add value to textual information that is normally locked, for example inside EHRs (Electronic Healthcare Records).

Figure 13: Screenshot KConnect http://www.kconnect.eu/

Further analysis (post anonymisation or pseudo-anonymisation) of patient EHRs can provide opportunities around: symptoms-diagnoses; patient segmentation; adverse drug events/reaction warnings, increase treatment efficiencies; letting clinicians know how similar patients were treated or simply provide the query basis for further search regarding a patient in other medical information sources.

5.4.4 Specific Challenges to the project

The relevance of text mining for medicine, can be illustrated with the following example; exposure to a potential drug–drug interaction (PDDI) occurs when a patient is prescribed or administered two or more drugs that can interact. Even if no harm ensues, such events are a significant source of preventable drug-related harm.[40] Text mining can help to avoid this from happening by providing more information. There is a pressing need for informatics research on how to best organize both existing and emerging PDDI information for search and retrieval. To overcome the disagreements the following has been proposed: [41] There is a need for

- a more standard way to assess the evidence that a drug combination can actually result in an interaction,
- agreement about how to assess if an interaction applies to a single drug or all drugs in its class,
- guidance on how a drug information source should handle PDDIs listed in product labeling [42].

40 'Toward a complete dataset of drug–drug interaction information from publicly available sources, Serkan Ayvaz,John Horn,Oktie Hassanzadeh,Qian Zhu,Johann Stan,Nicholas P. Tatonetti,Santiago Vilar,Mathias Brochhausen,Matthias Samwald,Majid Rastegar-Mojarad,Michel Dumontier,Richard D. Boyce, Journal of Biomedical Informatics, Elsevier, June 2015 http://www.sciencedirect.com/science/article/pii/S1532046415000738

41 L.E. Hines, D.C. Malone, J.E. Murphy Recommendations for generating, evaluating, and implementing drug–drug interaction evidence, Pharmacother. J. Hum. Pharmacol. Drug Ther., 32 (4) (2012), pp. 304–313

42 http://www.sciencedirect.com/science/article/pii/S1532046415000738#b0010

- there is currently no interoperable standard for representing PDDIs and associated evidence in a computable form (i.e., as assertions linked to evidence).
- Since evidence for PDDIs is distributed across several resources (e.g., product labeling, the scientific literature, case reports, social media), editors of drug information resources (public or proprietary) must resort to ad hoc information retrieval methods that can yield different sets of evidence to assess.

A recommended best practice is that systems that provide access to the lists (for example through API's), should provide results using a interoperable common data model for PDDIs.[43] Furthermore they should inform users that the lists may be incomplete with respect to PDDIs so that clinicians are aware of this.

5.4.5 Legal challenges

One of the main legal challenges for KConnect involves working with medical records. As a result they have big issues of confidentiality and data protection.

One of the challenges to overcome in developing their service that allows users to search through Electronic health records are the data protection regulation requirements. In many countries, there are specific regulations about accessing medical data. In case of personal data or sensitive personal data when the data for example holds medical information about a person the use is strictly limited for other purposes than the one for which the persona data was collected.

As a solution KConnect has developed a unique capacity through the Clinical Record Interactive Search (CRIS) application which allows research use of the *anonymised* mental health electronic records data.[44]

The Dementia Clinical Record Interactive Search (D-CRIS)[45] is a resource that enables large patient datasets to be pooled so that dementia research can be conducted at scale, providing researchers with access to one million patient records and enabling them to identify trends in the data and investigate why treatments work for some patients and are not as effective for others.

The KConnect project will provide semantic annotation and semantic search capability across the complete record with integrated biomedical information extracted from the literature knowledgebase. This capability is believed to transform the way clinicians and researchers use the ECH.[46]

Much of the information within the record will still be hidden from the clinician and researcher but it does allow a set of natural language processing information-extraction applications covering a range of hitherto-unrealised constructs such as symptomatology, interventions and outcomes (e.g. adverse drug reactions).

[43] L. Peters, O. Bodenreider, N. Bahr, Evaluating drug–drug interaction information in NDF-RT and DrugBank, in: Proceedings of the Workshop on Vaccines and Drug Ontology Studies (VDOS-2014), Houston, Texas, 2014.
[44] This was developed At the NIHR Biomedical Research Centre for Mental Health and Unit for Dementia at the Institute of Psychiatry, Psychology and Neuroscience (IOPPN),
[45] http://www.slam.nhs.uk/research/d-cris
[46] Case study Kings College http://www.kconnect.eu/kings-college-london

Applications to access CRIS and the analyses carried out using CRIS are closely reviewed, monitored and audited by a CRIS Oversight Committee, which carries representation from the SLaM Caldicottt Guardian and is chaired by a service user. The Committee is there to ensure that all applications comply with the ethical and legal guidelines.[47] CRIS was developed with extensive service user involvement and adheres to strict governance frameworks. It has passed a robust ethics approval process acutely attentive to the use of patient data. The data is used in an entirely pseudonymised and data-secure format. All patients have the choice to opt-out.[48]

5.4.6 Education and Skill challenges

The project has reported having problem recruiting skilled people.

5.4.7 Technical challenges in developing a secure system for TDM

Medical professionals frequently use general-purpose search engines such as Google, medical research databases and even Wikipedia to answer medical questions online[49]. A potential problem with these resources is that most of them either return large amounts of clinically irrelevant or untrustworthy content (e.g., Google), or that they are mainly focused on primary scientific literature that makes selection of clinically relevant publications very time-consuming (e.g., PubMed).[50]

Another issue is with the quality of data. For example relevant for KConnect service is how to handle text in Electronic Health Records, which often includes is misspellings, neologisms, organisation-specific acronyms, and heavy use of negation and hedging.[51]

5.4.8 Conclusion

The KConnect case study aims to provide more insight in the issue of using data that may include personal data. The project will provide insight in how to comply with the data protection regulations while still hold relevance for further research. As KConnect develops further in bringing the technology to market they will be able to provide possible best practices in the economic aspects of TDM service development. We will continue to analyse the KConnect developments and include an updated version of this case study in the next Deliverable 4.5

5.5 Mediately – A Slovenian Technology Start-Up

5.5.1 Background

Mediately is a Ljubljana based start-up, focusing on improving patient care, by providing health care professionals with a range of treatment related information.[52] It currently offers services to medical

[47] The Clinical Record Interactive Search (CRIS) system has been developed for use by the NIHR Mental Health Biomedical Research Centre and Dementia Unit (BRC and BRU) at the South London and Maudsley (SLaM) NHS Foundation Trust.
[48] See access at june 2016 http://www.slam.nhs.uk/research/d-cris
[49] Kritz M, Gschwandtner M, Stefanov V, Hanbury A, Samwald M. (2013) Utilization and Perceived Problems of Online Medical Resources and Search Tools Among Different Groups of European Physicians. J Med Internet
[50] Samwald, M. & Hanbury, A. (2014). An open-source, mobile-friendly search engine for public medical knowledge. *Proc. Medical Informatics Europe 2014*
[51] The hospital electronic health record (EHR), implemented in 2007, contains records for 250,000 patients in a mixture of structured and over 18 million free text fields.
[52] www.mediately.co

professionals in Slovenian, Serbian, Croatian and Czech via its website but also via downloadable apps available from Apple's App Store or Google Play. It began trading in 2013.

Services:

Mediately analyses patient care related information that it can get access to via the internet, and presents the information in a synthesised form for doctors, nurses as well as other health care professionals. This supports medical decision making and also saves health care professionals time as information is more centralised for them.

Data Sources

The main source of information currently aggregated and translated by Mediately in their online services is publicly available information from the European Medical Agency and various European countries' medical authorities. The information provided by the medical agencies mainly relates to prescription medicines. This information often includes guidance on dosage, how often to take the medicine, what to take the medicine for, known and possible side effects, when to discontinue use, interactions with other medicines etc. Other information collected and made available to healthcare professionals includes officially registered medicines for use in a particular country, official price, manufacturer, whether the cost of the medicine to patients is reimbursed by national insurance schemes etc.

Employees

Currently Mediately employs 8 persons.

Competitors

eMC (UK), epocrates (US), Vademecum (ES), Vidalfrance (Fr).

5.5.2 Technical Access Issues to Data.

There are a number of challenges that Mediately face that relate to the technical access to data. A lot of the data held by medical authorities is presented on websites as PDFs. However it is believed that in some instances that the information received by the medical authorities from the pharmaceutical companies is provided in CSV or other open formats. (CSV formats are easier for technologists to work with as they represent more structured data, than for example a PDF which is a free flow of text.)

The high prevalence of PDFs have required a lot of investment in order that Mediately can be in a position where they can extract the required information automatically that is "buried" in the free flowing text. The company estimates that 70% of the investment required in entering a new language marketplace is employed in normalising and creating structured data. Not only is the financial and time investment high in getting to a position where they can automatically extract the various types of information held within a PDF, but the legal and medical risks of not getting this process right means the company also has to invest a significant amount of their time in building processes to validate and verify the extracted data. This is to ensure the absolute accuracy and correctness of the information they provide in their services, as it is central to patient care.

The time, effort and costs involved in turning the free flowing text held within the PDFs into something computer readable represents a double-edged sword for the company. This is because once the investment has been made in normalising the data, and turning it into something a computer can read, they have a significant first-mover advantage over other companies. Any competitor wanting to enter into the same marketplace would have to replicate this investment.

Mediately also report ongoing costs when organisations that host medical information redesign their websites, as the algorithms and software that Mediately have created in order to create structured data from the material hosted online on that website have to be re-engineered. Mediately estimate that their back-end engineers spend approximately 50% of their time re-engineering their algorithms because of changes to providers' website layouts and changes to other documentation hosted by European Medical Authorities.

5.5.3 Economic challenges: Comparisons to the United States

Mediately are conscious that one of their main competitors in the US, epocrates launched 10 years before Mediately did. The US had a big technological head start compared to Central Eastern Europe, where most of Mediately's markets are located. Availability of data in digital form, and low barriers of access to that data have stimulated a more competitive landscape in the US than Europe. This meant that companies had the ability to put out comparable products much earlier than tech companies based in Central Eastern Europe. Another reason for this is the fact that epocrates only operates in English, whereas Mediately is currently operating in Czech, Croatian, Slovenian and Serbian meaning that all services and activities need to be replicated in these four languages.

For technology companies operating in this space the US is also arguably an easier environment to operate in. Mediately highlights particularly two areas where they feel US competitors have an advantage:

Firstly the US Medical Authority – the Food and Drug Administration – generally release all its information relating to medicines under the most open terms and conditions possible (a CC0 waiver), which waives any intellectual property that may exist in relation to the information held by the FDA. The FDA Terms of Service states:

' Unless otherwise noted, the content, data, documentation, code, and related materials on openFDA is public domain and made available with a Creative Commons CC0 1.0 Universal dedication. In short, FDA waives all rights to the work worldwide under copyright law, including all related and neighboring rights, to the extent allowed by law. You can copy, modify, distribute, and perform the work, even for commercial purposes, all without asking permission.[53]'

This gives companies who are using this information full legal certainty that they can reuse the information they have access to how they want, therefore supporting the business and reducing legal barriers.

Secondly, the FDA launched "Open FDA" in 2014 which through its open API allows developers to freely search, mine and analyse over 3 million reports produced by the FDA between 2004 and

[53] http://www.fda.gov/ Accessed on june 2016 and https://open.fda.gov/terms/

2013.[54] This open API facilitates enormously access to the vasts amount of information that the FDA holds, something which does not exist in Europe.

5.5.4 Legal Issues

Licensing and Copyright

Operating currently in four countries, but soon to expand into a further two European territories, one of the issues that the company has faced is getting permission to use the documentation made available by European government medical authorities. While most countries make this information freely available, Germany charges and different countries have different terms and conditions relating to the use of the information. In addition to differing terms and conditions of use, Mediately reports that frequently they do not get a reply in response to a request to use the medical information made available on the health authority's website. Whether this is because there is no one really in charge of licensing this information in the medical authority, or responding with a clear answer is legally too complex, or whether a request from a small startup company is not seen to be important is open to conjecture.

For a small start-up, Mediately are aware that the legal issues that they face ranging from copyright and licensing, through to liabilities that relate to mistakes in their own systems, that could have a life threatening impact on patients means that they need to be legally aware as a company. Currently Mediately is able to rely on a small group of friends with a legal background who help them, but they predict in the next year they will have to employ a full-time lawyer.

Open Access

Guidelines and international best practice on how to administer drugs most effectively or how to do pre-surgery checks on patients are often published in journals. Currently as Mediately work with pharmaceutical companies, often they rely on the pharmaceutical company to ask researchers for permission to use the article where best practice guidelines are written up. This is time consuming and ad hoc in terms of results. Mediately report little benefit to them as a business in the European-wide investment in gold open access because of the difficulty of easily discovering and establishing reliably whether an articles is available under a CC BY licence or not. Theoretically Mediately could benefit from the UK's investment in Gold Open Access but a lack of a services / portals with correct licensing information and metadata means the company is reluctant to take the risk of using articles that could turn out to be the copyright of a third party.

Investors

As with any startup Mediately rely on investors to fund their business. As outlined above the lack of responses from European Medical Authorities, and general reliance on third party copyrighted information taken and mined from the open web (as well as ad hoc use of published articles) means that Mediately feels it operates sometimes in more legal uncertainty than is needed.

[54] https://open.fda.gov/

5.5.5 Conclusion

As a startup Mediately's business depends on acquiring the medical information it needs, and hard decisions have to be made when no answers are forthcoming from a medical agency. This contrasts strongly with US based competitors who can operate in a climate of legal certainty in regards to information published by the FDA who generally waive all copyright and other intellectual property rights in their publications. The mining of accessible data is also viewed as being allowable and lawful under the US doctrine of fair use, following a number of relevant US court cases.

Mediately reports that it is often asked by potential investors if its scraping and analysis of material from the internet is legal. Given the licensing uncertainty that operates in this area, exacerbated by an occasional lack of response, they report that some potential investors are put off by the complex and legalistic question marks that exist over its data analysis business.

5.6 TDM and Textkernel

5.6.1. Introduction

Successful economic development is helped at a fundamental level when its members efficiently manage to find the jobs suitable to their skills and potential, and when companies and institutions succeed in finding the most suitable talents to carry out the required tasks. This process of successfully matching employees and employers requires TDM on a large scale basis, considering the size of the growing European and international jobs market and overall population sizes. Companies optimize diverse aspects of the recruitment work on the international level, i.e. LinkedIn[55], where clients do much of the knowledge aggregation themselves, as well as focusing on the regional markets and automatization through high-precision TDM, a case in point being the Dutch company Textkernel[56].

TDM for the recruitment process consists of automatic information extraction (skills, education level, experience) from curriculum vitae, as well as the automatic extraction of the same information fields from job advertisements. With current developments in technologies that lead to the emergence of new types of jobs every decade, and constant changes in the vocabulary of job titles and descriptions, smart text mining techniques are required.

Textkernel considers candidate experience as one of the most important aspects of the recruitment process, affecting both the speed and success of the matching and the interest of the candidates in the company in question. With this in mind they develop technologies that simplify this experience by turning it into an on-line process. The candidates should be able to simply upload their CV, or any other earlier prepared documents supporting their qualification, and avoid as much as possible any manual form filling. Text mining helps to parse the incoming documents, extract the information and align it with the most suitable positions in the database. As Textkernel operates in the European market, their TDM technology also has to support multilinguality. Attaining high precision information extraction and adapting the same basic technologies to multiple languages are the two key technical challenges for the company.

[55] https://www.linkedin.com
[56] http://www.textkernel.com

5.6.2. Legal challenges

As long as the content provided by the employers and potential employees stays within the Textkernel facilities, there is no issue of privacy and legal access to the data. Both the companies using the service and the candidates that are looking for a position are interested in the data usage and give consent to its usage.

5.6.3. Education and Skill challenges

When exposing users to advanced technologies, the usual challenge for each innovative company is to balance user expectations and the technical capabilities. Textkernel uses state-of-the-art machine learning and artificial intelligence techniques in order to shape the service they provide within a context that is familiar to their users, i.e. combining core TDM techniques with Internet crawling and advanced matching and searching.

5.6.4. Technical barriers

As for many other companies in the field, Textkernel has limited access to the output of research that is done within the academic communities not published in open access repositories. This can potentially slow down the testing and ingestion of state-of-the-art techniques.

Focusing on the European Union requires the development of tools and solutions that process multilingual content. The bias of existing TDM tools towards the English language[57] implies costly adaptation to other languages than English. Semantic technologies that are multilingual can boost workforce mobility around the continent. Textkernel runs its own research department to overcome these language barriers.

5.6.5. Conclusion

Overall, TDM companies such as Textkernel successfully manage to build their business model by overcoming potential issues of the absence of one unified European market. This is possible due to the fact that they have access to the freely available content on the web and to the one provided by their customers directly. Their main issue lies in the diversity of languages being used in the EU market and the need to adapt the tools accordingly. These problems are dealt with via internal research effort and monitoring of the academic progress in the field.

5.7 Academic Research

A growing number of stakeholders understand the value and importance of allowing researchers worldwide to use TDM as part of the research process. This includes using TDM to find relevant topics for research, doing a systematic review of the literature and applying TDM to be able to generate and analyse data for results[58].

'The real richness is in text and data together. We need to look at mining both.'

[57] See D4.1. for more details on the language tools availability

[58] For example Crossref which enables researchers to mine content across a wide range of publishers, has extended its TDM rights for non-commercial research purposes to researchers at subscribing institutions. See http://tdmsupport.crossref.org/

This case study focuses on these restrictions and other barriers that researchers experience when doing academic research.

5.7.1 Background

Researchers in all disciplines are confronted with an increasing amount of data to process for literature reviews or research analyses. For example for the biomedical sciences, PubMed alone has 21 million citations for abstracts or full articles and this is increasing at a rate of two per minute. In the humanities researchers are tapping into an increasing stream of data from social media accounts such as Facebook and Twitter as a source for their research. In the environmental sciences user generated content such as citizen-reported plant observations supersede the necessarily limited scale of academic observations.

This case study is an account of a few different research practices and the barriers researchers faced when it comes to TDM for academic research. The fictitious examples are based on the interviews with a small number of individual researchers from different disciplines who wish to remain anonymous.

5.7.2 The use of TDM

Discovering relevant research is a key application of TDM and basic search and information retrieval is indispensable to most researchers, while others are exploring more sophisticated TDM techniques. Those we spoke to were primarily interested in performing metanalyses and extracting information from full text publications. This was usually in the form of free text, sometimes focused on a particular section such as the methods, but there was also interest in data extraction from diagrams and tables. It is therefore often the case that the abstract which is made freely available does not hold all the information which is necessary to determine whether or not the article is relevant for research, while also clearly being insufficient for undertaking the research itself.

In order to use TDM the target data needs to be discoverable and accessible for machines. Many researchers noted an access problem in getting the information they need. The data may not yet be available in digital format (see Plazi case study in section 5.2). This is for example still the case in environmental field where much of the information relevant for researchers is still being held in books in libraries which are not yet digitised. Or the data is digitized and indexed but they cannot access it properly because the data is placed behind a (pay) access wall and/or spread diffusely over different repositories or databases owned by various different rights holders such as institutions, repositories and publishers. If researchers want to be systematic and know everything there is to know in the academic literature on a specific topic e.g. a human gene, there are large associated transaction costs in terms of time and potentially funding.

As results, all these researchers raised numerous issues and barriers they have encountered.

5.7.3 Legal barriers

Unless a researcher is gathering his own data, the data will be owned by someone else and/or stored in a database. It is often the case that the use must be cleared by the rightsholder before the researcher can access and make use of the data for his or her research.

Not having access is seen by the different stakeholders as the main barrier for researchers to do TDM. This barrier is often classified as a legal barrier because the rightsholder is the one who can

control the access to the publications. For example publishers can decide whether or not to allow TDM by researchers and under what restrictions. Although publishers aim to facilitate research and provide controlled access to their databases the use of API's is not without problems. For the technical issues see section 5.7.4.

Access restrictions

One of the problems reported by researchers was publishers' use of their APIs or web-based tracking to control and subsequently restrict access.

The main point for discussion between researchers and publishers is when a researcher who has lawful access, for example through his institutional subscription, gets blocked because he is using TDM to access and download a vast amount of publications. This is often as a result of download limits or hidden links that alert the publisher to mining activity and either trigger warning emails or immediately block the triggering IP domain. This is problematic for researchers because being blocked causes delays in their research and may put an additional financial burden on institutions. The researchers expressed frustration that if they had done this manually there would have been no block but because it can be done in a fraction of the time by a machine publishers consider that this no longer falls under the normal use and a new 'right' must be negotiated. However, the researchers we interviewed disagreed with some saying that the 'right to read is the right to mine' so no additional permission should be necessary and a block is not justified.

The researchers feel that such publisher actions are based on unclear terms and conditions such as unspecified download limits They also feel uncertain that they have the same access to the database on both the API and site so are often using website scraping preferentially to avoid reduced access and requirements to sign agreement which restrict the use of TDM results downstream. The uncertainty about the ownership of data and copyright has led to many researchers only using publicly available and open access data. They are willing to work with sometimes inferior data sets to avoid having to ask for permission which they consider a lengthy and tedious process which they do not have time nor negotiating skills for.

Finally, there is some worry about privacy concerning what information on research activities publishers are collecting through their APIs.

(Inter)national data sharing

Another important aspect mentioned by researchers are the legal implications of working with international partners. They have uncertainty about their ability to share research across borders. The UK in this respect has the advantage of being sought out as a strategic partner because of their copyright exception. As a result several research projects we encountered locate the TDM part of the research with an academic partner in the UK. However, the results or practice may not be shared if publishers place conditions on downsteam use of data through their contractual agreements.

This is frustrating for a researcher who after having spent time and effort building a repository finds himself unable to share his work with others who may not have lawful access to the same content. As a result his research may not be validated or cannot be used for further research. The consensus of the researchers we interviewed was that having a copyright exception in the EU would provide the

legal clarity necessary to improve TDM for academic research and that to acknowledge the growing practice of private public partnership the exception should not be limited to academic research or non-commercial use only. Many researchers believe that this would limit their chances to work together with industry on projects that involve both academic and applied research.

Personal data

People are able to track and collect all sorts of interesting data about themselves consciously via exercise watches or more unconsciously by uploading a geotagged photograph to social media or even by submitting a piece of text that has privacy data of the person operating the device submitting the data attached to it. If people for example contribute to a database by posting pictures, their location over time may be collected. The people contributing data to these repositories may not consent to this data to be used for any other purpose. However the status of this data and the need for consent for example metadata is unclear to those who are aware of this issue.

Personal data both in the datasets but also the data that is attached (the metadata) can be very useful for research but researchers report there is too much uncertainty about the data protection regulations. As a result most researchers refrain from using personal data in their research or only when it is anonymized.

5.7.4 Economic barriers

Researchers have a limited amount of resources available to spend on getting the data they need. So they will have to be able to quickly identify where the relevant publications are stored and whether they are available for TDM to access and to extract only the data they may need for their specific research.

One of the barriers described in finding funding for projects using TDM is having to explain its benefits for this specific project. Often it is seen by funders not as academic research but as applied research. As a result, a lot of projects either cannot get funding when they want to use TDM because of misunderstanding or lack of awareness about what TDM is and can do. Or they have been funded without proposing to use TDM so when a researcher wants to use TDM they cannot because it was not written into the grant proposal as a research method.

5.7.5 Technical Barriers

There is a lack of tools available for researchers who would like to do TDM themselves. This includes researchers who do not know how to develop their own tools, but even those who can report difficulty finding effective and easy to use tools, as well as a clear need for documentation on how to use them.

One of the reasons proposed it that there is not enough testing on realistically large datasets during academic projects to develop TDM tools. Often a sample training dataset is used but this could be just a few hundred papers and the tools subsequently don’t work well in practice using real life datasets of tens of thousands of papers or much larger databases. One reason could be that developers do not have access to large datasets and therefore only use openly available datasets. As a result they may also never encounter any problems such as legal issues when trying to apply TDM.

Another technical issue is the reliability of the results. At the moment the success rate is not high enough for researchers to rely on the outcomes of TDM and they do not trust an entirely automated approach, but equally recognise that current practices do not make optimum use of the advantages of machines versus human cognition.

"I don't mean to imply that humans should not be involved in data curation, what I meant was the processing part can be done through machine tools. Humans are good at using [the tools] and deciding what is valid and what is not and what is high quality"

The results do not only depend on the quality of the analysis tools but also of the input data as this drastically impacts the necessary role of the tools in pre-processing datasets. Data quality was reported as being an issue by many of the researchers. If a researcher is looking for text and data he will find that most data first needs to be cleaned up and structured before it can become useful for research. This is a frustrating process and even then an analysis may not lead to satisfying results or match any hypothesis you had about the data.

'It is a problem when data is not in a TDM friendly format.'

There was optimism about the rate of improvement in tools, but some concerns over how to change norms and practices for researchers themselves. The interviewees mentioned that many results of research are still not properly managed and may be stored on a personal computer or on a USB stick instead of being put in a repository. When the research is not indexed it cannot be found and according to these researchers that may include a large amount of research today. As a result a lot of research will never be discovered, papers that could provide insights are never read and the researchers not cited.

A proposed solution for some of the technical issues is to look at Open Source software. The Open Source approach allows different tools to be linked together more easily and there is a strong support from the OS community for science, particularly in certain areas like bioinformatics.

5.7.6 Education and Skill

Not having the chance to learn about TDM and how to use or develop TDM tools has been identified as a problem by many of the researchers. Many were self-taught, having gone online to find information and courses to learn how to use a specific tool or TDM service. They also point out that the more senior generation of researchers and principal investigators, often do not really know about the benefits and value TDM can bring to a project and therefore are not encouraged to include TDM in their research proposals. There is also a gap in knowledge when it comes to funders and institution librarians in recognising the importance of and fostering skills development in this area.

'We are still getting skilled graduates but their skillset isn't a very good match with TDM'

There are some that propose to have more general courses to be introduced for every discipline while others think it should be limited to only those where it is deemed the most useful and instead to promote collaboration between the different disciplines to make use of each other's strengths.

5.7.7 Conclusion

One of the most striking features in developing a case study for researchers was that apart from the lack of awareness and uncertainty about many aspects of TDM; they were strongly reluctant to talk about their TDM practices without anonymity. They tend to be very cautious to share their research practices and reportedly said they were not sure if what they did was allowed or not. Consequently they also did not know where to go with their questions to get legal certainty. The negative impact of uncertainty on practitioners and their research is important to stress in determining actions and best practices to overcome that barrier.

6 CONCLUSIONS

6.1 Main findings

The interviews on which the seven case studies were based aimed to challenge and provide evidence for barriers that limit the uptake of TDM in Europe for scientific research and innovation. With the first set of case studies this report aims to provide insight into specific issues facing stakeholders using or developing TDM practice. What has become clear from both the interviews and the case studies is that there are some important steps the FutureTDM project needs to take moving forward.

6.1.1 Education and skill

The main insights that this deliverable has provided is the need for more education on the benefits and practical use of TDM for researchers: working together with industry, publishing community and academia to develop effective courses aimed at different levels depending on the discipline and type of research that is likely to use TDM. This also includes developing profiles for use in developing educational materials.

6.1.2 Legal and policy

With respect to the proposed exception, there is a need for legal clarity for all stakeholders involved on what is allowed using TDM. At the moment there is no consensus amongst the different communities about the legal status of TDM practices and use of results that are gained through using TDM. This provides a serious barrier in the way of time spend on getting consent and working out how to provide attribution. Another serious effect is that it seems to influence what research is being done and what data is used for research. Researchers mentioned avoiding licensed or copyright protected material, which may lead to a bias in results and chosen topics for research.

Another concern repeatedly mentioned both by researchers and companies is the uncertainty about data protection and data sharing. There is a need for more clarity about how to comply with the data protection regulations.

We refer to FutureTDM Deliverable D3.2 and D3.3 for more insight into the issue on copyright protection and database protection.

6.1.3 Technical and infrastructure

Looking at the technical aspects of TDM, again there is no consensus on whether there is a need for more education for all researchers on how to develop tools or whether the sole focus should be on investing in the development of tools and services. Most stakeholders agree that there should be a combination of directed efforts on making tools easier to use while at the same time teaching various stakeholders not only where to find them, how to use them but also how to provide their data in such a way TDM can be done.

Most of the services and tool providers do not mention technical issues as a problem but seem confident that in a matter of time the tools will improve and better applications will become available. Their main concern is the quality of available data, but again there is no consensus that this is a barrier as there are service providers who take this as a business opportunity.

6.1.4 Economy and Incentives

Barriers that are mentioned are the lack of a single European market, the problems of having multiple languages and a lack of enforcement for US companies.

The usefulness of standards as an enabler is a topic that was been discussed as a technical barrier and is an important topic for further FutureTDM research. Developing standards for data quality is seen as a useful but most likely impossible solution given the diversity in projects and requirements, which would make standards too complex for compliance. In those areas with existing standards there is already an issue in having people comply to them, but some propose a solution would be making them mandatory e.g. through funding requirements, or strongly incentived through mechanisms such as rankings.

6.2 Further research

The interviews and the case studies have provided evidence of and insight into the barriers that exist to TDM in Europe. To what extent these barriers can be solved given the different interests of the stakeholders involved remains a topic for further research within the FutureTDM project. For the follow-up deliverable D4.5 we will provide an updated analysis of the present and possible new case studies to include best practices and methodologies for improving the uptake of TDM in Europe, focused on addressing the barriers presented by the main stakeholders here.

7 REFERENCES

Agosti D, Egloff W 2009. Taxonomic information exchange and copyright: the Plazi approach. BMC Research Notes2009,[2:53].DOI:10.1186/1756-0500-2-53,accessed on 5 June 2016 http://bmcresnotes.biomedcentral.com/articles/10.1186/1756-0500-2-53#CR6

Article 29 Data Protection Working Party, 2013. Opinion 06/2013 on open data and public sector information ('PSI') reuse, 5 June 2013,

Brook, M, Murray-Rust, P, Oppenheim, C. The Social, Political and Legal Aspects of Text and Data Mining (TDM) City, Northampton and Robert Gordon Universities doi:10.1045/november14-brook

Caspers, m, Guibault, L (2016) FutureTDM Deliverable 3.3 Baseline report of policies and barriers of TDM in Europe.

Cocoru D and Boehm M 2016 *An Analytical Review of Text and Data Mining Practices and Approaches in Europe* (London: Open Forum Europe)

Éanna Kelly, "Researchers to Take on Publishers over New EU Copyright Laws," *Science|Business*, 07 May 2015.

European Commission, *Report from the Expert Group on Standardisation in the Area of Innovation and Technological Development, Notably in the Field of Text and Data Mining* (Brussels: European Commission, 2014).

Eskevich, M. & Bosch, A. van den, 2016. Deliverable D3.1: Research Report on TDM Landscape inEurope, Available at: http://project.futuretdm.eu/wp-content/uploads/2016/05/D3.1-Research-Report-on-TDM-Landscape-in-Europe-FutureTDM.pdf.

Eskevich, M., van den Bosch, A., Caspers, M., Guibault, L., Bertone, A., Reilly, S.,Munteanu, C., Leitner, P., Piperidis, St., 2016. FutureTDM Deliverable D3.1 Research Report on TDM Landscape, Available at: http://project.futuretdm.eu/wp- content/uploads/2016/05/D3.1-Research-Report-on-TDM-Landscape-in-Europe-FutureTDM.pdf

Frew,h, White, B, Bertone, A, 2016 FutureTDM Deliverable 2.2 Stakeholder Involvement Roadmap and Engagement Strategy

Filippov, S. 2014. Mapping Text and Data Mining in Academic and Research Communities in Europe. Brussels: Lisbon Council. Available at: http://www.lisboncouncil.net/component/publication/publication/109-mapping-text-and-data-mining-in-academic-and-research-communities-in-europe.html

Handke, C., Guibault, L. & Vallbé, J.-J., 2015. Is Europe Falling Behind in Data Mining? Copyright's Impact on Data Mining in Academic Research, Available at:http://dx.doi.org/10.2139/ssrn.2608513.

Hargreaves, I 2011, "Digital Opportunity: A Review of Intellectual Property and Growth,"

Intellectual Property Office, 2014.Exceptions tocopyright:Research, Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/375954/ReseArch.pdf.

Kaplan, A. (1964). *The conduct of inquiry: Methodology for behavioral science.* San Francisco: Chandler.

Kell,D (2009) "Iron behaving badly: inappropriate iron chelation as a major contributor to the aetiology of vascular and other progressive inflammatory and degenerative diseases," BMC medical genomics, vol. 2, p. 2,

OpenMinTeD Deliverable 5.1: Interoperability Landscaping Report, 28 December 2015.

Plume, A and van Weijen, D (2014) "Publish or Perish? The Rise of the Fractional Author...," *Research Trends*, 38, September 2014.

Simpson, M. S., & Demner-Fushman, D. 2012. Biomedical text mining: A survey of recent progress. In Mining Text Data (pp. 465–517). Springer.

STM, Text and data mining: STM statement and sample licence, 2014.

Tsai, H.-H. 2013. 'Knowledge management vs. data mining: Research trend, forecast and citation approach'. Expert Systems with Applications 40; 3160-3173.

Triaille, J.P., de Meeûs d'Argenteuil, J. & de Francquen, A. 2014. Study on the legal framework of Text and Data mining (TDM), European Commission. Available at:http://ec.europa.eu/internal_market/copyright/docs/studies/1403_study2_en.pdf

Universities UK and UK Higher Education International Unit, *European Commission's Stakeholder Dialogue 'Licences for Europe' and Text and Data Mining* (London: Universities UK, 2013).

Ware, M and Mabe,M 2009 "The stm report: An overview of scientific and scholarly journal publishing,"

Weiss, C. H. (1998). *Evaluation: Methods for studying programs and policies* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

8 ANNEX 1 QUESTIONNAIRE

- What is your experience working with TDM?
- How widespread is the use of text and data mining in your community.
- Are there specific fields where TDM is trending or highly used?

The next section asks about Legal and policy aspects

- What type of data and tools do you use *or provide access* to?
- What do you do with the data that you mine
- Have you experienced legal issues that have prevented you from doing the above?
- have you worked out a means of working around these issues? If so, what is the workaround? If not what are the consequences of these legal barriers
- Could you tell something about the rights clearance process and contractual practice regarding the ability to mine a database or collection, owned by publishers, libraries or other information providers?
- Do you think there is legal clarity around TDM in Europe?
- What legal solutions would increase the uptake of TDM?
- Would a legal exception for copyright and/or database law for TDM improve uptake, and if so what are the requirements considering:
    - the works to be used
    - the beneficiaries of the exception
    - the purpose of the use
    - the character of the use
    - the need to compensate right owners for the harm caused by the use;
    - other requirements

The next section focuses on economic aspects of TDM landscape.

- What companies and what sectors do you see building value on TDM practices,
- Open source and open data? building a business on availability of data (uncertainty)
- Why do you think those are particularly prone to TDM applications?
- What would you say were the economic barriers that prevent uptake of TDM
- Also if there is not much commercial uptake of TDM in a specific field, why do you think this is?
- Could you reflect upon: what is the current economic climate when it comes to
    - The use of TDM
    - The development of new TDM tools.

The next section asks about the technical issues to TDM uptake.

- What do you think about the current quality of data and what is the effect on the uptake of TDM? In your question please refer to
    - Content that undergo mining
    - Resources that are relevant within the mining process itself

- How would you rate the infrastructure for TDM in Europe?
    - (technically) access to content and data to be mined
    - availability of mining tools for certain language, domain of interest, type of data
    - availability of resources and other reference data for certain mining task (e.g. a specific ontology does not exist in certain language)
    - capability to combine different tools in a chain to accomplish certain goal
    - availability of computing resources (hardware/cloud)
    - availability of skillful personnel

- How do you rate the availability and effectiveness of "standards for representing content, data as well as metadata for describing them?"?
- What technical infrastructure would you like to see put in place?

The final section asks about the research environment, education and skills

- Do you think the demand for TDM from researchers is sufficient?
- Are the following actors sufficiently aware of TDM and TDM possibilities : Funders, researchers, librarians, managers, administrators
- Do you think the different actors have sufficient skills? And how could we improve this?
- Is the use of TDM for research properly incentivised? How could funders, institutions etc. incentivise TDM?
- Are there incentives for publishers to support TDM?
- Are there any other aspects in the research environment that you have encountered which provides a barrier to TDM that we have not discussed?
- If you would like to mention anything else that relates to barriers to TDM, or if you know of any other good practice, please mention it here.

9 ANNEX 2 INTERVIEWS

The following are quotes and comments taken from the interviews. They have not been edited to allow the issues people face to be expressed in their own words. They have been colour coded to provide keywords that have been used throughout the deliverable

Technical and Infrastructure

- Referring to the quality of data and datasets ' they are not out of the oven and ready to eat'
- It can be frustrating to spend much time cleaning a large dataset and not being able to find any interesting information that can be used to test your hypothesis.
- Access through crawling is overloading our publisher platform.
- Doing TDM is getting easier because computer are getting faster but interesting enough the computers on the publisher's side they seem to be getting slower. They seem worried about their system being overloaded.
- API is about tracking, we (publishers) want to manage platform access
- Centralised infrastructure development is really important
- Not having tools is not the issue, difficulty is getting the pdf after that it's easy
- As an example: Pubmed central is adopted. We will get XML eventually
- Publishers systems are old and publishers not always the best in IT
- Most publishers have automated download capabilities. They allow that to real text miners like commercial to pull papers. So there is a system to do bulk download, it works well but they [publishers] don't give academic people access to that.
- They use the excuse their systems get overloaded well then give us another way to get these papers but they don't want to do that either.
- I [researcher] think limited downloads of 1000 per day is fine but won't scale to other publishers. There are around 500 in biomedical domain and I can't have 500 different api's.
- crossref api not widely supported for other publishers.
- Publishers have outsourced to silverchair and other I[researcher} need then to contact these third parties to tell them my IP address.
- It is hard to get papers out of pubmed if you [academic library] want an archive of what you have access too. For technical reasons they [publishers] block you.
- it can be technically quite complex how you [researcher] are supposed to give attribution.
- TM and Machine Learning must be used because to screen by hand it takes a year or two at least by which time your [researcher] results are out of date
- Ideally what you want is a web based modular system
- my own sense is the big companies that are in this space are able to deliver more quickly by horizontal integration what they already do in different area's
- documentation and training around these tools are not sufficient.
- a big task is to develop tools for tdm that are going to be beneficial for the research community.
- from a publisher's perspective our role is to have a platform enables everyone to read articles, we also need technical perspective that enables miners to download vast amount of

materials. So sciencedirect platform and separate infrastructure api for miners to access the same content but different structure that allows vast scale downloading

- if you allow people to come to the platform and download massive amounts of content, apart from the security issues, there are technical things need to be considered. Something publishers are resolving and helping to resolve through API.
- api can help to distinguish between legitimate text miners and those who want to abuse.
- TDM is got better but accuracy must be high enough so that scientist can rely on it.
- funders say the same thing they have data management plans that they request. But researchers don't know how to fill it in and don't know how to get support.
- standards for data: problems is a lot of standards around but not clear if they are. often quite extreme the big ones, the researchers who want to move on get frustrated if it's detailed it's difficult to encourage.
- On interoperability: the open source approach allows you to lock different tools together and there is support from the os community.
- standards have a big part to play and in trying to structure and order what sometimes is an unordered environment
- We need more generally evidence of benefits for use of standards: does that research have more impact, is it more widely used, does it have more citations or being used in subsequent research

Proposed solutions

- Learn from sci-hub, they have really good infrastructure: one database and well-structured is what publishers should have done.
- Promote text mining as a method so why do we not make it freely available for researchers and SME in Europe and on subscription to other companies in other companies.

Education and Skill

- The argument is that if you are researcher on TDM you can solve the question on 500 papers to show it works. You do not need to do it on a million of papers. But these researchers work on the academic level and not in practice so they may not know what the actual problems are because they don't do this on a large scale.
- research approach ; frustration is that people use specific data : go to pubmed get a set which is fine to demonstrate the usability of the approach but we need all relevant data not just what is available open access but also stuff behind paywalls etc.
- consequences for research : examples are using oa repositories so they [researchers] are sampling, so that is a biased sample: the importance is in making the data available for all
- at the moment we [researcher] get publications through institutional access, if however it is not available then we purchase them through interlibrary loan is 4p for publication but the time we [researchers] get them is long.
- Using published content: it takes longer; you [researcher] need someone to do these interlibrary loans.
- Research would be easier if the publishers API was Open Access and we could do this in our system.

- On ethics of research: how ethical is it if we do not make the info available wildly to everyone who needs to use it. There is not a great deal of sympathy for the publishers in this situation. They are seen as creating a barrier and one that we know is going to fall sooner or later but they deliberately maintain a barrier until they can figure out a business model that allows them to continue to make money.
- We need more people who have a combination of skills
- TDM is not relevant to everybody's research, where it is relevant they may not know much about it or lack the technical capabilities to code and apply this to their content.
- The profile of a data miner; must understand policy and have the relevant skillset to address the increasing demand for TDM practitioners.
- They [students] don't need to know exactly what TDM is and does but more the concept of TDM and how the tools work.
- Industry can also help promote and facilitate these programs by being more involved, providing more resources and clarity about career opportunities.
- TDM is a complex technical field so education must include general education qualities.
- People are discouraged to study subjects that are considered harder to study.
- A lot of software developers are finding their own way and expertise through online courses, instead of conventional education.
- There is a disconnection between academia and industry in data science. People do not see the applications for example the bonus cards and discounts is huge data mining and science behind it.
- Not so visible to academia that industry is growing and transforming
- They [researchers] aren't aware and don't care about TDM or OA. They care about their next paper and grant.
- Primary responsibility is with national budgets for education. if we want to move towards a highly technological and sophisticated society a lot more investment in education and research is needed in general.
- In my view biggest barrier to progress in the field is education of experts on a large scale. On a much larger scale of what is currently the output of institutions.
- Primary responsibility is with national budgets for education. if move towards a highly technological and sophisticated society a lot more investment in edu and research is needed in general.
- People get grants to fund money for research to be done. The principal investigator needs to be aware and often they are not if they write the grant proposal in a way without TDM they have to proceed that way. It's in the grand they do it this way and is too late to change. Major problem and just lack of awareness generally.
- Researchers do not always know who the right person is to go to within their institution.
- Publishers are committed to support researchers who want to do TDM, The STM declaration where publishers signed up to is committed to this and publishers have done this to their own policies and integrating crossref.
- TDM is not relevant to everybody research, where it is relevant they may not know much about it or have the technical capabilities to code and apply this to their content.
- There is not a lot of off the shelf tools and researchers don't necessary know where to go for support with that.

- The challenge to really change is that you need all stakeholders involved to shift and move at the same time.
- Maybe we need to look at undergraduate courses in more detail if we want provide skills necessary for that type analysis. It might mean dropping more traditional module in favour of module on data analysis.

Possible solutions

- Easier tools and need for educations. People need to learn how to store data in a sensible day. Not just stick it on a disk in bottom draw. Why share how share. Prioritise between all of these things. researchers need to be trained and need to emphasis on easier to use tools and use cases and examples to highlight where people have used it and be beneficial so researchers might see how it
- Teaching with open data and open source tools. Using for example Eurostat because this is real world data and not 'pretend' data which provides students with a real life dataset and they get positive reinforcement: The students can continue to do it themselves.
- The 'geo *for all consortium* 'is a worldwide consortium using opensource tools in teaching environment.

Legal and Content

- I [publisher] don't think TDM is a copyright issue. It is more an infrastructure issue
- For a single researcher it takes a lot of time to contact publishers to request publications. I [researcher] stopped my research as a result.
- Requests for papers are positive from those publishers who are close to the open movement as results my [researcher] papers are biased also to certain domains.
- There is always an issue about licensing of data. Too much administrative hassle so trying to work with CC0 if possible so no restrictions what you can do with it. We [researcher] would always prefer freely available over data without licence strings attached even if data with licence strings attached was better.
- Problems: if you have a project with deadlines you only have so much time for negotiating or finding out.
- Copyright exception presumes there is an issue around access to content. Researchers who have lawful access to content are able to text mine with publishers.
- Having an exception will not solve the issue of skills, education and support for researchers for transforming files into xml.
- The exception is solving a problem that does not exist. [Publisher] There is no access problem.
- An exception will have unintended consequences, it will disrupt the system that is working well and already in place and expose publishers content and undermine content that we invest in for the research community.
- We [researcher] typically did not have any project funds allowed for paying license fees. this would put us off so if we could avoid it.

- In an academic project it's difficult to guarantee what you [researcher] are going to do with the data. You're going to publish it in some way and going to manipulate it in various ways so you don't want to tie yourself down if you can avoid it
- If more detailed attribution it can be technically quite complex how you are supposed to give attribution.
- We [researcher] simply back off if the data is in copyright. you could say that is a problem.
- We [researcher] did get permission from a large publisher to work within a sandbox.[59]
- We [SME] have legal people who know the rights
- Tempting to make money itself [Cultural heritage institution] out of exploiting its own collections
- A public funded body wanted to promote its material but it would be putting it out in the wild and never be able to change its mind. There is caution about what type of licenses to choose.
- We [public body] were not directly selling it but were selling value out of projects around image data. There was worry that if you allow data it will affect all the rest.
- Someone else is selling the data and taking away your market.
- The exception only for non-commercial is not problematic: we [publisher] look at who wants to do the mining, if it's researcher or institution this is non-commercial if its industry its commercial.
- We as publishers community have not communicated well enough what is 'commercial.'
- We [publishers] have no problem with researchers doing research and publishing etc . We do mind the output. The researchers' copyright material that underlines the research, we want to avoid the commercialisation of the underlying material.
- Certainly as an academic researcher it's very frustrating you want all the data to be open. Our attitude was we would do useful stuff with it
- If only there was a solution for people to be able to use your data but then you'd be allowed to change your mind if you don't like the use.
- if you put non-commercial on it it's not open
- UK as an example; we [publisher]have not seen much uptake so it's not enough to change the law you also need to do the infrastructure
- Harmonizing EU exception is fine
- There needs to be a degree for infrastructural development, that everyone is happy with
- A consequence of copyright is that getting permission takes so much time. As a professor you would avoid a topic because a PhD student would just waste a year on this.
- Publishers don't want you to download everything they are afraid that all pdf will be to Russian website. I hoped this argument would disappear because yes you can illegal download all papers from psihuv.
- With the copyright exception publishers can stop worrying and do something else.
- On legal clarity: it will never be tested as most libraries and centres are good customers; why would a publisher sue their customer and they don't want to set precedence and it's expensive.

[59] A sandbox in this case meant researchers worked in an 'isolated' environment that allowed them to only access parts of the repository necessary for their research.

- Access is rarely a problem for rich countries but not in countries like Bulgaria and other countries.
- If you get blocked in the UK do you go to the government to tell them let them [publishers] stop block me?
- Individual contracts say I [researcher] can use the data in context of my work.
- Typical TDM is difficult to explain
- Using institutional library accounts limits the number of publications used within the project. Still human work needed to request library loans etc.
- we [researchers] take the view and supported by UK legislation that if we have got the right to read a pdf through a license to download a pdf that also covers the right to TDM the same pdf
- On web crawling if google has it indexed it is accepted that that is the norm'.
- We [researchers] are building an online tool, but if people use TDM tools this might not be covered by the licensing agreement.
- As Machine Learning gets more complex and efficient than scientists whose work is not available for TDM and ML will see their work is not used.
- To do effective search I[researcher] downloaded the full content in order to locally indexing it myself. Instead of relying on third party such as web of index and google scholar.
- The regulator landscape is so completely unclear that if I[researcher] get approval or a lawyer within the institution to say it's fine we are waiting forever.
- UK exception: we find it being a positive for us [researchers] in that there are international groups to which we belong are keen to partner with us because in the UK TDM will be covered and possible.
- If we TDM 1000 papers and 30% reported this or that, that is aggregated data which is allowable.
- The scientific life cycle is disrupted if data is only held in a few hands and if in more hands we can develop treatments to new diseases.
- It is unlikely to harm their [publishers] model if TDM allowed.
- If i [researcher] loose affiliations to the university I'm dead!
- Best is just be open. It's difficult to say what is commercial or not or what is science or not. The best contribution to science is mostly citizen scientist maybe 2/3 records of observation is from citizen scientists that is not science anymore.
- For scientist is more interesting to be in the West because access to all this data and literature. if you open this up it does not matter where it sits it's accessible to all.
- We [publishers] need to help researchers but is the exception the right way? It will not solve the access issue. The question is HOW can I get lawful access!
- You often get threatening emails; you've been downloading too many papers even when you institution has legitimate subscription agreement with that publisher you get emails identifying your ip address and that your ip is being blocked.
- Content used in responsible way. responsible users such as resources and libraries but also need systems in place to those that do not have legitimate access or use it for illegitimate purposes, to copy and distribute the content.
- If blunt instrument such as an exception would give anyone access to go to publisher's platform and download content, we then do not have any idea if the user is legit.

- Authors disputes about data. That's my data in collaboration and I created the data. Every publisher has author disputes. *Data is the new issue.*
- Follow the guidelines: go to institutions to work it out with their authors and whose data is it. It's not that obvious and collaboration where the lab is who does it etc. is a muddy area. And an area that will only grow to raise questions.
- When the core of the article is the data then technically there is no 'author 'but legally…. A CC0 license solves problem of citation but problematic authors want and should be attributed it counts on cultural norms.

Economic and Incentives

- Compared to the academic sector, the corporate sector is willing to pay for solutions
- The corporate sector is concerned about confidentiality
- The market for TDM is immature.
- One size does not fit all, the TDM market is very diverse and has different needs
- Funders replied when asking for a grant to use TDM: this is not research, this is infrastructure this is downloading files and looking for something
- Response from funders on refusing a proposal: we want research to focus on finding new applications not apply existing applications to more papers
- Money should be available to do applied TDM research. There is the argument why are not librarians doing that? But they are not domain experts
- There should be money available to tackle specific domains instead of for example a national centre for text mining
- The problems in chemistry and biology are that they have groups who work in-between applied research and academic research. How much money from your project do you [researcher] want to dedicate to infrastructure?
- Problem for start-ups : University does research and forks out into small company but they then find themselves without access to publications after leaving
- On permission: When going for something that is being sold you [researcher] get more pushback from publishers. There is potential to make money
- On permission: I [researcher] usually don't know my research doesn't lead anywhere. It is hard to explain what you will use the data for.
- Many public bodies are tasked with making money. They prevent others to exploit their data or at least in a way that undercuts what they want to do themselves.
- As a company we are prepared to pay for access. We want high value data and accept to pay for that.
- The energy industry clearly thinks there is market for data to mine.
- we got a lot of funders who believe in our data model,
- There a lot of start-ups interested in getting high quality and up to data and extracting knowledge and business knowledge faster than competitors can. There is a business advantage through analysing data better.
- AI and data science products and - services are not perfect. This is a difficult market and the challenge is to get the expectations right. Its also a big motivator to meet those customer expectations that drive us [SME] forward.

- Everybody {SME} is working with data and language technology to improve for example their websites and search engines because without you will not be able to survive.
- [service provider] It is a misconception that there is a lack of companies doing TDM. Companies, who are doing this, are perhaps not specialized in technology but for example Spotify and Zalando, these are growing European companies who do use data mining. They have large teams but it's not their core business (music and fashion).
- They [European companies not specialised in TDM] also do invest in developing the technology needed for TDM such as AI a lot.
- In general there is a lot of investments done in these types of economies
- One of the EU problems is that we do not have a large single European market to develop these kind of companies.
- The fragmentation of the EU market makes it harder for EU companies. There is different regulation, language and national markets which each ask for a different marketing approach.
- The EU needs to be stronger in taking momentum of putting real policies that help companies.
- It is a complicated topic but most obvious is language which cannot be removed by policies.
- There are a lot of details like employment policies that do make it kind of hard to move easily across Europe
- It's possible to do business online, but if you want to develop a network and market presence you still need to open an office in every EU country, which is an investment.
- We [publishers] are here to support the research ecosystem. If TDM is increasingly becoming part than we want to make sure tools are available for them to use.
- Publishers make good money from subscriptions and shareholders. We [OA publishers] need to demonstrate that OA gold is a beneficial model and profitable
- It would be easier if there would be one European set of rules but on the other hand the EU legislation tends to be more restrictive than national.
- Harmonization is a mixed blessing for companies because may introduce additional barriers for using data from web sources.