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 D4.5

Compendium of Best Practices and Methodologies (Update)

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, 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:	**D4.5**
Title:	**Compendium of best practices and methodologies**
Lead beneficiary:	**Open Knowledge International**
Work package:	WP4: Analyse: Fields of application, projects, best practices and resources
Dissemination level:	Public (PU)
Nature:	Report (RE)
Due date:	30.04.2017
Submission date:	30.04.2017
Author(s):	**Freyja van den Boom**, Open Knowledge International
Main Contributors:	**Maria Eskevich**, RU **Antal van den Bosch**, RU **Jenny Molloy,** OK/CM **Peter Murray Rust**, OK/CM **Ben White,** BL **Marco Caspers**, UvA **Stelios Piperidis,** Arc **Burcu Akinci,** SYNYO **Donat Agnosti,** Plazi **Malcolm Macleod,** Edinburgh University **Blaz Triglav,** Mediately **Andre Gaul,** Paperhive **Hilke Reckman**, UNSILO **Frederico Nanni**, Universität Mannheim **Michael Larrobino**, CopyrightClearanceCenter **Jakub Zavrel**, Textkernel **Petr Knoth**, Open Academy **Sara Tonelli,** Foundation Bruno Kessler **Allan Hanbury**, KConnect
Review:	**Burcu Akinci**, 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

Table of Figures

Table of Tables

EXECUTIVE SUMMARY

The aim of this report was to find, challenge and provide evidence for what are considered to be barriers and enablers for text and data mining (TDM) in Europe.

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] We use the general term *text and data mining (TDM)* in this report, although the activity can be also referred to as two separate and partly overlapping text mining and data mining processes.

For this report, we have examined different TDM practices carried out by scientific researchers and small scale companies working in different economic sectors. Building upon the research done within the FutureTDM project, this report provides TDM studies to demonstrate the issues different stakeholders face within their text and/or data mining practices. The case studies are set up in such a way that will highlight the apparent barriers and showcase a compendium of practices and methodologies that may help improve the uptake of TDM in Europe.

Chapter 1 starts with the introduction followed by a description of the methodology used for this report in Chapter 2. Chapter 3 provides an overview of the main insights gained from the interviews. These have been grouped together under the following headings:

- 3.2 'Technical and Infrastructure'
- 3.3 'Legal and content'
- 3.4 'Economy and Incentives'
- 3.5 'Education and Skill'

Chapter 4 describes the conceptual framework and case study selection, Chapter 5 showcases the case studies and Chapter 6 concludes this deliverable with the main findings.

[1]FutureTDM_D4.1-European-Landscape-of-TDM-Applications-Report available on the FutureTDM website

1 INTRODUCTION

This report showcases case studies of potential TDM practices, with focus on drivers and barriers of TDM uptake.

For the stakeholder consultations, we decidedly did not provide a specific definition of what are 'barriers' to TDM to allow participants to share what is a barrier for them. Whether or not a specific issue was mentioned also provided us with insight into the level of awareness of and attitudes towards different aspects of TDM.

With respect to collecting best practices to overcome barriers, we found that respondents were hesitant to talk about 'best' practices dealing with barriers. They often reported they were not aware of standards or code of conducts and did not know whether their practices could be considered as the best. Due to this reported absence of acknowledged 'best practices' we opted to discuss what participants considered to be good and potentially best practices. As of this the recommendations should be taken in this context.

The case studies have been developed based on two sets of semi structured interviews. The main purpose of these stakeholder 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 based on consultations and recommendations from different communities and economic sectors, in order to have a representation of different TDM involvement levels and working practices.

Limitations

The interviews provide insights into the practice of TDM from the main stakeholder communities involved in TDM as identified in previous FutureTDM research.[2] The results of this report may 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 and practices for discussion. Given the high level of expertise of the participants, the input provides useful insights that are indicative of the barriers to TDM uptake in a more general sense. With respect to best practices and methodologies these take into account the different stakeholder perspectives and must be considered in addition to the FutureTDM policy guidelines and practitioners recommendations.

[2]FutureTDM D2.2 Stakeholder Involvement Roadmap and Engagement Strategy

2 METHODOLOGY

2.1 The Interviews procedure

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 first round of interviews took place between March and June 2016
The second round of interviews took place between January and April 2017

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 deliverables. After a first set of 3 interviews (10% of the total number 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.

Selection of participants

The interview participants were selected using the internal project stakeholders' directory, the FutureTDM network and recommendations from all partners.[3] Participants were chosen from each stakeholder groups to give a sample representation of their specific field of expertise, experience and interests.

Questionnaire

There were two rounds of interviews the first focussing on barriers and the second more specific on enablers to help improve TDM. In both rounds the initial set of questions for the questionnaire was developed based on the issues that FutureTDM identified in previous research as well as issues that came up during meetings and the FutureTDM Knowledge Cafés.

These questions were grouped around the four identified themes covering the following issues:

- *Economic and Incentives*
  - to help 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*
  - awareness of legal issues and experiences with legal barriers, if so, how did this affect their TDM practices/research.
  - to see 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.

[3] FutureTDM D4.2 and FutureTDM website knowledge base

- *Technical and Infrastructure*
  - to understand whether the currently available European infrastructure for TDM is sufficient and what would be necessary steps to improve and/or establish it.
  - on data management plans in practice, data quality and availability of useful tools for TDM.
- *Education and skill*
  - to get a better understanding of the current status of TDM education in relation to the need and availability of a skilled workforce, and
  - to gain insight into what is considered necessary to improve TDM skills

Figure 1: Geographic Map with the locations of the interview participants and events

After internal review, the questionnaires for the semi-structured interviews were adopted and used for a first set of interviews. After this set was carried out, the questionnaire was reviewed again and adjusted based on the responses. Some questions were removed from the questionnaire because they did not provide a useful response and some questions were added to get more clarity about a certain topic.

2.1.4 The case study format

The case studies aggregate the insights of the FutureTDM project gained from the interviews, knowledge cafés and other FutureTDM dissemination, consultations and research activities.

The method of case study analysis was chosen to showcase different TDM practices from the main stakeholders perspectives. Given the topic of this report to provide a collection of best practices and methodologies, we were particularly interested in cases that had either experienced or confronted specific TDM barriers and found ways to deal with them successfully and have the potential to be adopted as a best practice.

The case studies present different stakeholders and a variety of practices that are relevant for improving TDM uptake. Included are:

- The TDM researchers/practitioners perspectives,
- TDM content and service providers perspectives, and
- Startups that have entered the European market providing a service or tools for TDM

Figure 2: Main four themes identified in FutureTDM[4]

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

3 INSIGHTS FROM THE INTERVIEWS

This section provides a summary of the barriers and practices and methodologies mentioned by the interview participants. The case studies in section four illustrate what is being identified here as barriers and proposed solutions either through a service, tool or practice.

3.1 Introduction

This report follows the four headings that are used throughout the FutureTDM project to identify the main categories of barriers. The interviews have been coded, and grouped together under these headings:

- 3.2 Technical and Infrastructure
- 3.3 Legal and Content
- 3.4 Economy and Incentives
- 3.5 Education and Skill

In the following sections we give an overview of the main issues (Table 1) that were raised and solutions and practices proposed by different stakeholders as well as the topics that will be covered by the case studies. The quotations used throughout the report are taken directly from the interviews and included here to provide a platform for these different voices in the TDM community to be shared.

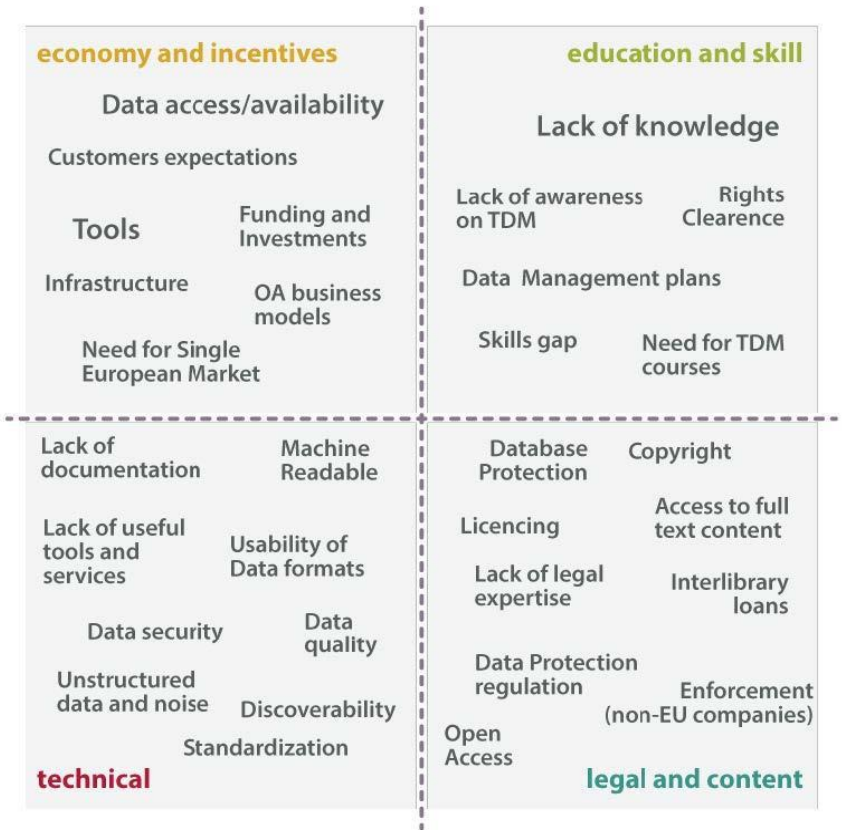

Figure 3: Overview of the various issues reported grouped according to the four main categorie

3.2 Technical and Infrastructure

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

The main issues from the discussions about the technical aspects of TDM include the following:

Access barriers

The use of an Application Programming Interface (API) may help the platform owners, i.e. publishers, to avoid TDM activity overloading their systems. It also gives them control over who can access what content, by what means and for what purposes so that only those who have lawful access are able to do TDM. [5]

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

Researchers however do not recommend restricting access to content via API because of the following reported issues:

- No access, blocking users and/or sending warnings when the use exceeds a limited number of downloads possible through the API. Limitations may be arbitrary. As a result, the lawful researcher still needs to contact and negotiate terms under which TDM can be done, which can be time consuming and costly.
- Not having unrestricted access to the full content. Although several publishers say that using their API gives the same results as TDM applied directly to their platforms this may not always be the case in practice. Various researchers have reported not having access to the full content and 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.
- Not being able to mine across content providers. The absence of a platform or standard API makes TDM very time consuming for researchers who have to try and gain lawful access in a quick and reliable way.

The use of an API can become a barrier when the API is not conform a standard and not interoperable. Other reported issues are on the completeness of the content that is made available through API's. Some practitioners say they did not get access to the full content or they were unsure of how much of the content was actually accessible through the API because there was no notice.[6]

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

[6] All case studies refer to this issue.

Use barrier: *quality of data*

When data is digitized or 'born digital' it may not be in a useful format for TDM. For example, the use of the PDF format 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 consensus amongst stakeholders that having a standard format in which data should be made available, regardless of what format is chosen, will greatly benefit all TDM practices.[7] There is concern however about how to incentivise people to adopt already existing standards instead of developing a new one and to comply with standards as early as possible. It is considered good data management to comply with a given standard during the data collection stage, instead of having to post-process data to a certain standard afterwards.

Use barrier: *tools 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 not enough, easy to use and effective tools available. Tools that are available however are often too expensive, not fit for purpose or they simply cannot be found. Almost all of the practitioners we spoke with have the skills to develop their own tools to fit their specific needs. But others who do not have the knowledge or resources to invest in tools will be left behind.[8]

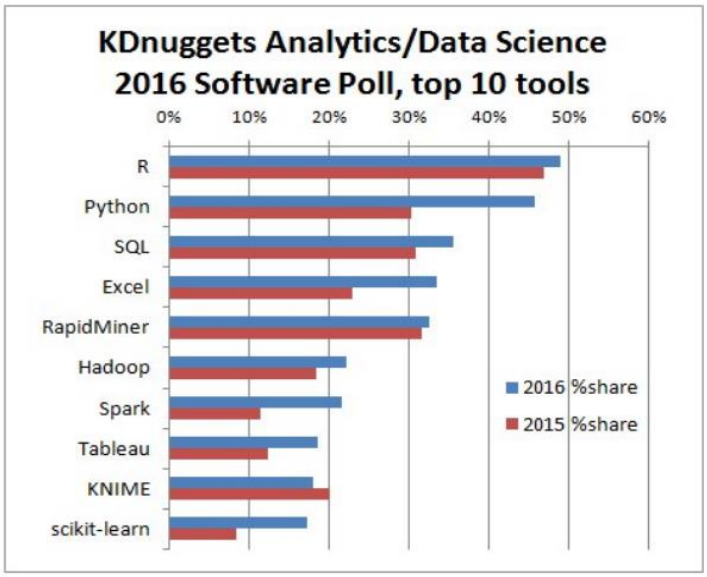

Figure 4: KDNugget screenshot 2016 Software Poll results show R is the most used software tool for mining[9]

[7] See case study 5.1, 5.2, 5.7, 5.8, 5.9, 5.12, 5.13.
[8] See case study 5.1, 5.3, 5.7, 5.11.
[9] Accessed online at http://www.kdnuggets.com/2016/06/r-python-top-analytics-data-mining-data-science-software.html

The development of better and easier to use tools will increase the use of TDM. However, users warn that the specificities of research projects may not make it possible to have standardised reusable tools available for everyone. TDM tools need to be tailored to become suitable. People report 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.

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 it can be useful within their practice.

3.3 Legal and content

The main issues from the discussions about legal and content aspects of TDM.

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 practices remain under a certain threshold. They will therefore not be made 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 at the academic level and not in practice so they say 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 even to guarantee what will be done with the data. As a result, it can be hard to explain the proposed project to the rights holder to obtain permission for the use of the data as there may be no specific purpose.

Access to data

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.

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.

Researchers 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 such as most Open Access licenses or find data with no license at all. Reported consequences due to lack of access include:

- There are topics of research which are not covered by researchers because they do not have access to information.
- Research being biased due to not using all the relevant data.[10]

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

Industry representatives 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 gain permission. Many companies also report taking the search engine Google as an example to see what practices are allowed.

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

Copyright exception

A proposed solution to the copyright barriers, 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. The 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 people who do not have lawful access and/or who use the works for unwanted and/or illegal purposes.

'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. '

Research purposes & non-commercial use

A copyright exceptions for TDM being limited to the research community 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.[11]

Those in favour of having the exception solely for non-commercial see no problem in making a distinction between commercial and non-commercial purposes. But most of the people we spoke to say that it will become increasingly difficult to determine when TDM is for commercial purposes. Especially given the trend towards more public-private partnerships and commercial spin-offs from academic research.

[10] See case study 5.1 and 5.7.
[11] See on this issue use case 5.2, 5.3 and 5.4.

Licenses

People are not sure what license to use for making data available. There is a fear of losing control over the data and uncertainty about what is allowed when it comes to data sharing. Content providers tend to stay on the safe side and chose not to make their data available or when they do they will use restrictive licenses that may restrict the potential of TDM projects.[12]

Personal data

When it comes to working with personal data people are uncertain about the actual scope of the data protection regulations and how to comply. Uncertainty and a lack of legal clarity results in high levels of precaution being taken. As most projects are limited to the use of anonymized data only.[13]

3.4 Education and skill

When asked about the level of TDM skill and education the following issues were mentioned:

Level of awareness

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

There is a lack of understanding and awareness amongst researchers about the use and benefits of TDM. Those working in academia and industry agree that there should be a joint effort to raise awareness and to help fill the current demand for TDM practitioners.

Industry can help promote and facilitate educational programs by being more involved, providing resources and clarity about career opportunities. Universities are urged to develop courses not only targeted at those who will become TDM practitioners and developers of TDM tools and services, but to include courses on TDM in the general educational curriculum improving general computer science literacy amongst all disciplines.

'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 without the need for expensive tools.[14]

3.5 Economy and Incentives

When asked about economic aspects of TDM and incentives for stakeholders, the following issues were mentioned:

[12] See case study 5.2, 5.3 and 5.9 on this issue also we refer to FutureTDM recommendations on legal aspects the FutureTDM D3.3 Baseline Report of Policies and Barriers of TDM in Europe for more information.
[13] See use case 5.4 for proposed practices to deal with this issue.
[14] See use case 5.11 for more information about teaching practices.

Market Access

The market for TDM is still very immature and the products and services available are far from being perfect. This is considered a market difficulty but mostly companies see this as an opportunity. Their challenge is to develop tools and services that meet the expectations and needs of the different stakeholders better than their competition.[15]

What companies mention as being a serious barrier is the fragmentation of the EU Market with its different regulations, languages and national markets making it hard to grow.[16]

'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 EU legislation tends to be more restrictive compared to national regulations. It might also introduce additional barriers that do not exist right now.

Availability and Access of TDM tools and services

The companies we spoke to say that the availability of quality TDM tools and 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.'

3.5.3 Academic funding

There is not enough funding available for academic research on TDM and applied TDM research. Funding should be made available for research to address domain specific barriers, infrastructure and data acquisition.

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

3.5.4 Data access

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

"They {content providers} 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 having to make money. The open access model is mentioned as a possible solution to the problem of lack of access and availability of data.

[15] See use case 5.10
[16] See use cases 5.4, 5.5, 5.6, 5.13 that discuss these issues.

4 CASE STUDY: CONCEPTUAL FRAMEWORK AND SELECTION CRITERIA

The interviews have provided evidence for the legal and content, economic and incentive, education and skill and technical barriers that are hindering the uptake of TDM.

The aim of this chapter is to present a compendium of barrier and best practice case studies to serve as examples for the barriers and enablers for TDM.

The case studies help develop a better understanding of the issues different stakeholders face when it comes to TDM. The case studies selected for this compendium also show examples of practices that proved successful in their specific setting and have the potential to be adopted as 'potential best practices'.

Using the interviews as a starting point the following case studies have been selected to illustrate the barriers and enablers.

Case study 1-7 have been selected for the insight they provide on specific issues main stakeholders face in practice.

Case study 7 - 14 provide more insight on possible practices services and tools that may help overcome barriers and increase the use of TDM.

It became obvious from the discussions on what could be considered best practices that there are no 'one size fits all' solutions.[17] Based on our stakeholder consultations we must conclude that best practices and tools are either not very well developed yet or they simply are not widely shared amongst the different communities since our participants were unaware of them. Because of a lack of best practices accepted as such we approached the question of 'what are the best practices and enabling methodologies for TDM' as being an adaptive learning process. It is therefore important to continue to facilitate discussion between and within the TDM stakeholder communities.

4.1 Selection criteria

For the selection of the case studies we developed criteria based upon insights from the ongoing research within the FutureTDM project. These are *Sector, Stakeholder, Member State, TDM process,* explained in the following sections.

Sector selection

Figure 4 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 affect the growth of other sectors.[18]

[17] D3.1 Research Report on TDM Landscape in Europe FutureTDM.

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

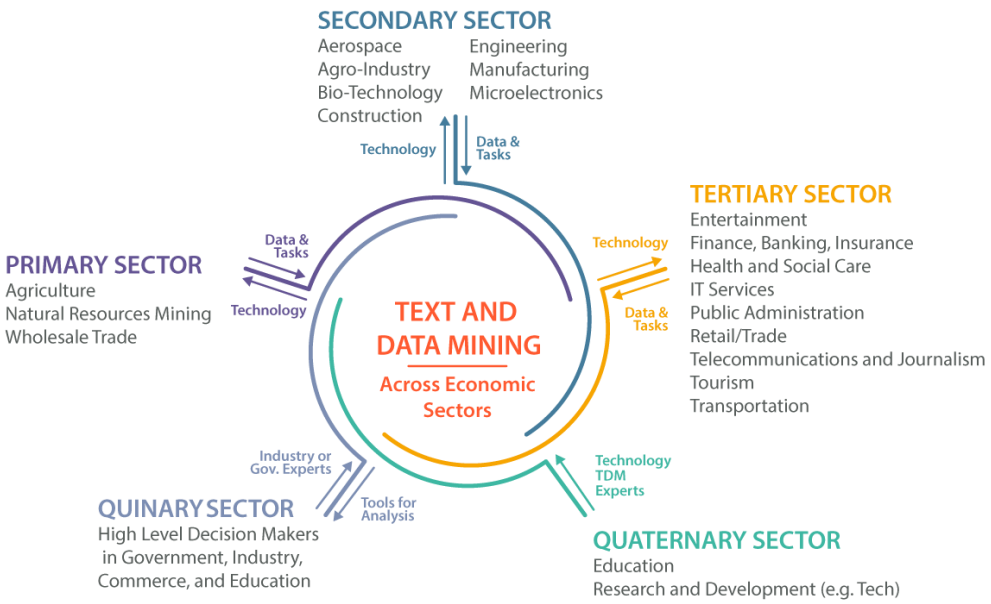

Figure 5: General economic structure and connections between TDM and economic sectors

Stakeholder selection

This compendium with proposed best practices and methodologies should be read as complementary to the FutureTDM recommendations and guidelines proposed for the different stakeholder groups. This is to ensure that issues are raised and addressed with the most adequate stakeholder and right level of involvement.[19]

The case studies represent the main stakeholders as identified by FutureTDM focussing mostly on the following stakeholder communities:

- Researchers,
- TDM content providers,
- Service and tool providers.

EU Member State selection

The stakeholder consultations confirmed national differences with respect to TDM practices and (level of) barriers in the different EU member states. The case studies represent these differences by covering practices in a number of different Member States.

The stakeholder consultations also focussed on participants from outside the EU to provide an international perspective. We have spoken to researchers who conduct TDM across borders as well as international content and service providers.

TDM process selection

Each case study represents at least one of the following four stages in the TDM process.[20]

[19] FutureTDM D5.4 Roadmap for increasing uptake of TDM.

[20] We refer to D 3.3 Baseline report of policies and barriers of TDM in Europe for more information

- **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**
  Contents is extracted to a new (target) dataset that can be used for analysis in the subsequent stage. The retrieved contents may have to be modified.
- **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 within only a closed circle in order to inform decisions.

4.2 Selection of the case studies

Each selected case study covers at least one of the barriers mentioned in chapter 3 and proposes or illustrates a potential best practice. Figure 5 and Table 2 represent respectively the geographical distribution and an overview of the case studies

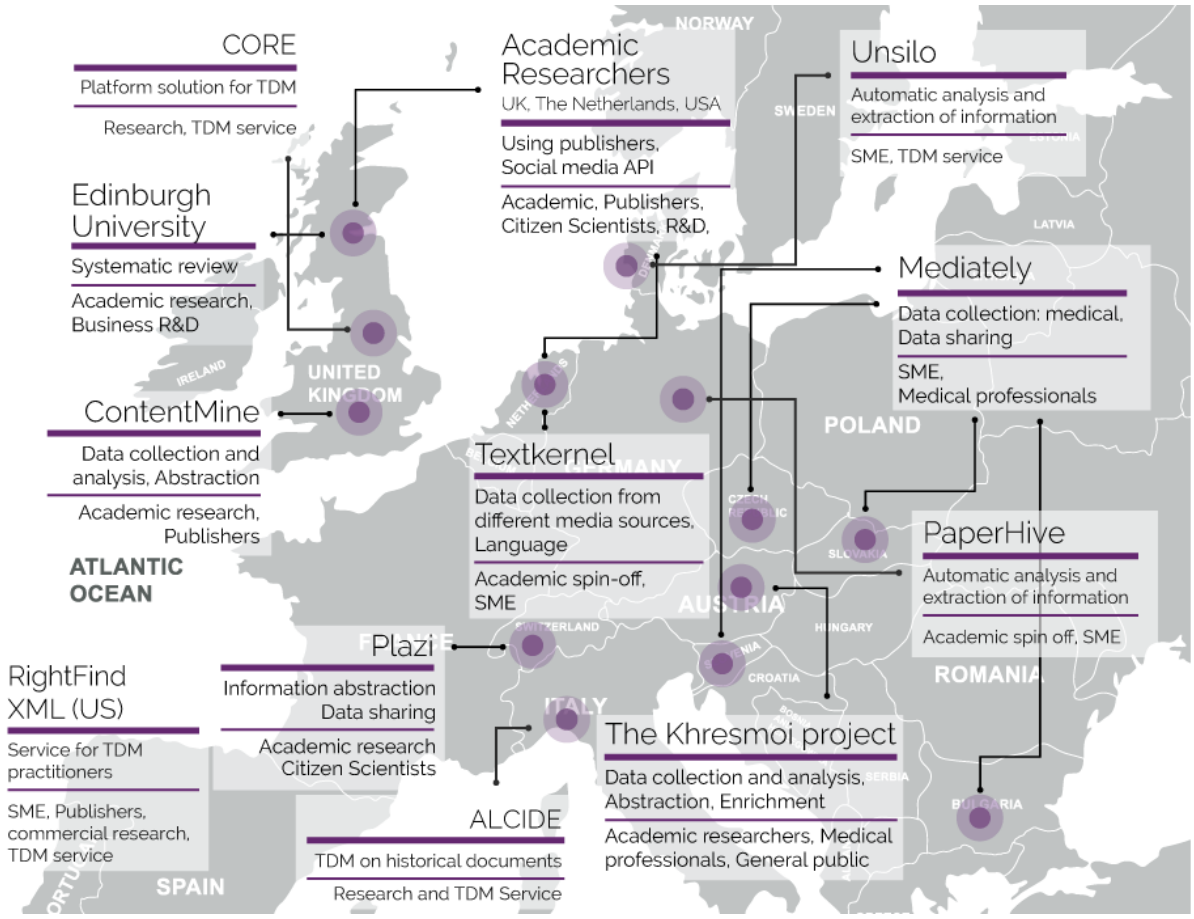

Figure 6: Map of selected case studies

No	Stakeholder	Member State	TDM practice	Main Barriers/Enablers
5.1	Systematic review, Academic research, Industry 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	PLAZI, Biodiversity conservation 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	Contentmine, Academic research, content 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	KConnect, 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	Mediately, Industry SME, Medical professionals	Bulgaria, Slovakia, Slovenia, Czech republic	medical and health mobile development company	Language, personal data protection, Commercial use, market entry
5.6	Textkernel, Academic, commercial spin-off	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 Researchers, R&D, citizen scientists	UK, The Netherlands, USA	automatic analysis and extraction of information from large numbers of documents.	Access, data quality, tools, education, awareness, standards
5.8	ALCIDE, Research, TDM Service	Italy, EU consortium	TDM on Historical documents, Platform development	Platform solution for TDM, Education and skill development
5.9	RightFind XML for Mining, Publishers, commercial research, TDM service	US, Worldwide	Service for TDM practitioners, cross publishers and harmonized contentprovider	Text mining workflow solution, eliminating the manual work that researchers would otherwise need to perform prior to mining content
5.10	UNSILO, Industry SME, TDM service	Denmark	automatic analysis and extraction of information from large numbers of documents. Document DNA	Absence of digital single market, Tool solution: complex natural language parsing and corpus-wide semantic analysis
5.11	Academic research, Education, skill	Germany, Italy, Norway	Awareness and skill, Tool criticism	Awareness, Education and skill, tool evaluation
5.12	CORE, Academic Research, TDM service	UK, EU consortium	platform solution for TDM	Cross content platform solution and service development
5.13	Paperhive, Academic research, startup and spin off	Germany	automatic analysis and extraction of information from large numbers of documents.	Cross content platform solution and tools

Table 1: List of case studies

5 CASE STUDIES

Each case study starts with a brief introduction into the TDM activity description (business model or research); followed by the main issues for further analysis of the barriers that were present in this specific case and/or what best practices and methodologies have been used or proposed for TDM. The case studies conclude with main insights into the barriers and enablers for the uptake of TDM.

5.1 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.

Research at Edinburgh University

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

Looking at the results of studies in healthcare, different studies often have conflicting findings. This could be the result of study differences, flaws or chance (sampling variation). When these 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 make it 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[21] has shown that much of the research published is at substantial risk of bias. As a consequence, the effects observed may be substantially overstated.[22]

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.

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

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

[22] The university is also part of CAMARADES an international collaboration which aims to provide a central focus for data sharing. 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.

overstated, by comparing results with existing research. Their aim is then to use this information to help guide better design of clinical trials testing treatments in humans.[23]

The application of TDM

To illustrate the scale of the problem, in animal research, every week around 3500 new pieces of research are published, making it almost impossible for anyone to stay up to date.

A specific query for publications can give 800.000 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. 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. They have yet to find a company that can actually provide the tools they need for their research. Companies are offering TDM solutions but the outcome of their services are not reliable or sufficient. In their experience, there seems to be a reluctance with companies to share code and/or solutions or to work together to improve results.

The final step in the process is 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.'

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 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. 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. 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.

[23] 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.

Legal and content

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. There is an argument for the explicit purpose of systematic review. 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. Journals could make their content available for review even if it's not available for reading.

'Having a copyright exception, things would be easier. We (researchers) could persuade journals to do this because if we get together all the available data, we can develop a better review 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 an issue that not all publications are available through their university subscriptions. If a publication is not available they will have to purchase it through interlibrary loans which will cost around £4 per publication. On top of the costs and time it takes to manually put in the request, the time it takes to receive the actual publication is often too 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 the research group cannot share the full results with anyone outside of the institution who do 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 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 are able to give a good presentation of the project there is funding available.

Conclusion and proposed best practices

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 PLAZI: Biodiversity conservation

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.[24] 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 species, there is a 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.[25]

Plazi.org

Plazi is an independent not for profit organization.[26] 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

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

[25] See for example the Antbase.org. , [http://antbase.org] for ants.

[26] Plazi. [http://plazi.org]

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.

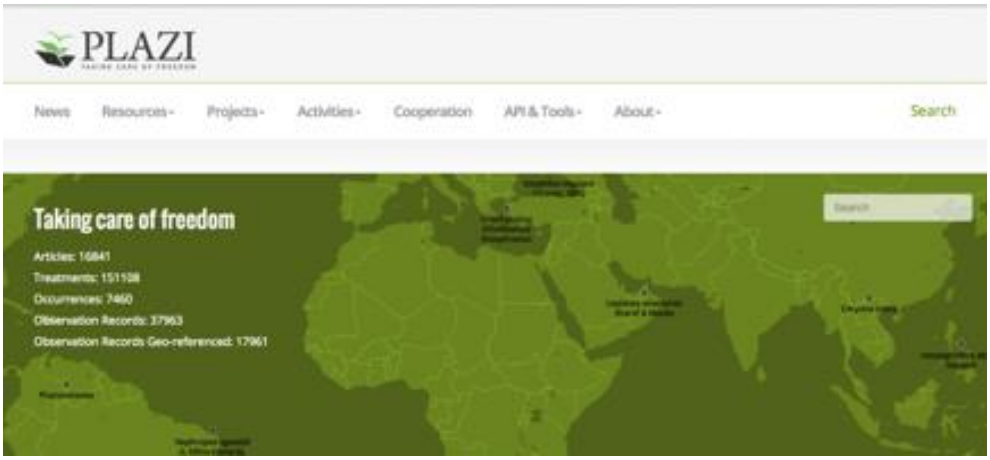

Figure 7: Plazi home page

TDM practice

The Plazi workflow (Figure 7) 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

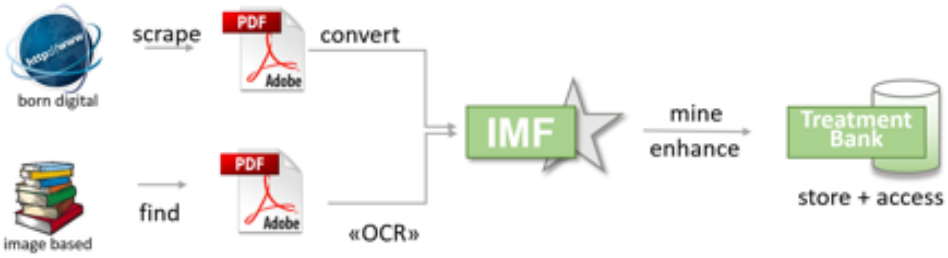

Figure 8: Plazi workflow

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.

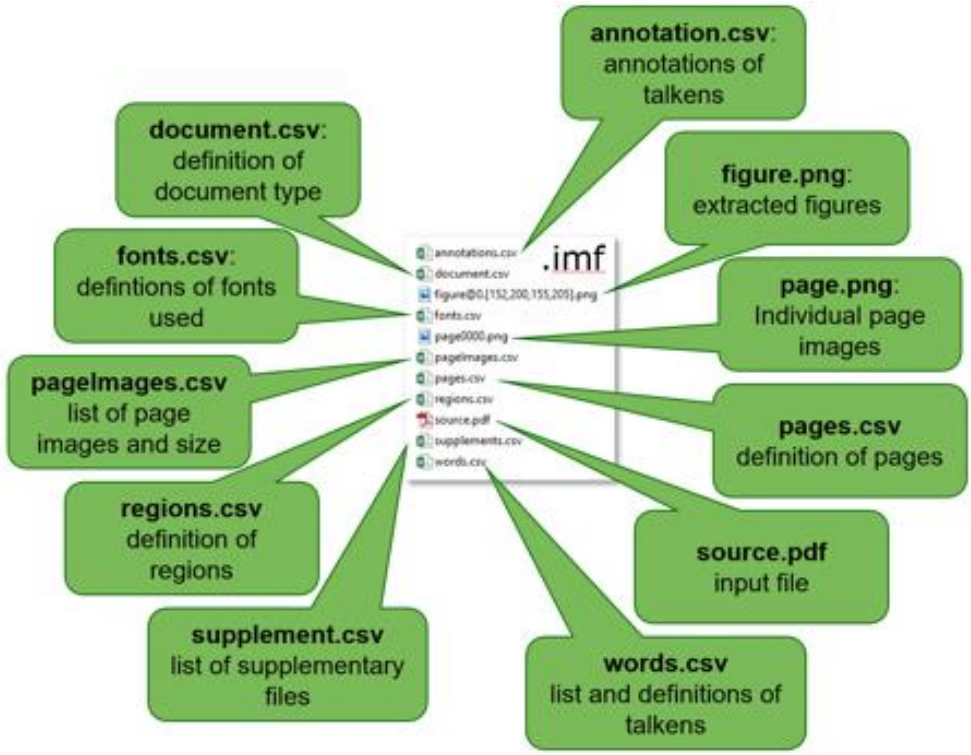

Figure 9: Image Markup File (IMF)

The IMF (Image Markup File, Figure 8) 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 at Zenodo/CERN, and a DOI is created in case none is available to cite the article.[27] This way, all the facts can be linked to a digital copy of the cited bibliographic reference.

After removing all OCR- and printing artefacts 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 now including over 600.000 references.[28] 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 and the Global Biodiversity Information Facility.[29] 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

[27] http://biolitrepo.org
[28] http://refbank.org
[29] http://www.catalogueoflife.org/ and http://gbif.org

automatic processing from scraping the Web to expose the facts on TreatmentBank.[30] 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 9) or RDF, for which a complementary vocabulary is developed for elements that are specific for taxonomic treatments.[31] For the remainder, existing vocabularies are used.

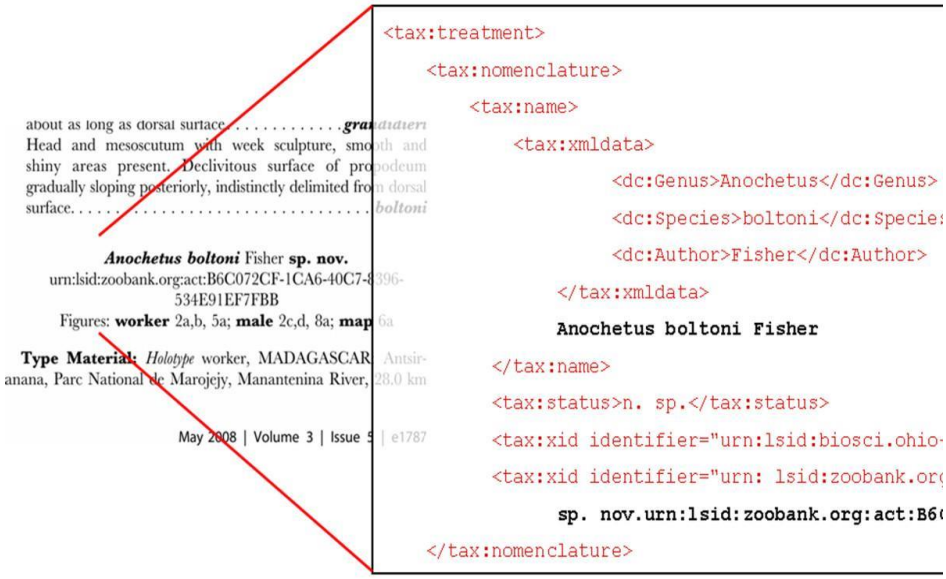

Figure 10: Sample markup page. Left: sample of an original, published taxonomic treatment. Right: Same treatment marked-up in TaxonX XML schema and enhanced with external identifiers

All application programs used by Plazi are open source except for the commercial ABBYY Finereader. This includes both for those supporting their internet services[32] as well as those created by Plazi themselves (GoldenGATE and its plug-ins, SRS), which are licensed under the Berkeley Software Distribution license[33].

[30] The daily input of new taxa, mainly based on this system is available online at: (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.

[31] https://github.com/plazi/TreatmentOntologies

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

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

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.[34]

Legal barriers

The legal barriers Plazi encounters is whether their process of abstracting the data is compatible with existing copyright rules. The question is whether they can 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

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).[35]

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.[36] Text written in accordance with these nomenclatural systems is not individual and cannot qualify as work.

Data extraction

Plazi creates its database from taxonomic literature that may be copyright protected. The main copyright question with respect to Plazi is whether they are 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

[34] 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.

[35] 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). 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

[36] 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.

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,[37] or for works of 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".[38] 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 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.[39]

[37] 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.

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

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

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.[40]

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.[41] 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.[42]

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. 2012)[43]. 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.

[40] 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.

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

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

[43] 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.

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.'

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 which are neither complete nor timely - but better than anything else.[44] Catalogue of Life is neither complete nor provides a near time service for new species.[45] 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.[46]

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

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.'

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 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.

Best practice Recommendations: Licensing proposal

Plazi is working on finding solutions to make sharing of research data possible. They advocate and educate the community on maintaining free and open access to scientific discourse and data. What

[44] http://www.organismnames.com/

45 (http://www.catalogueoflife.org/

[46] 'Horizon 2020 overestimated 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

they consider to be of vital importance[47]. Based on the Plazi experiences the following are considered good practices: [48]

1. 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)
2. 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
3. Essential Biodiversity Variables should be shared as Open Data, making them available as part of Data-CORE without charge or restrictions on reuse
4. Data, products and metadata should be made available with minimum time delay

5.3 ContentMine

5.3.1 Introduction

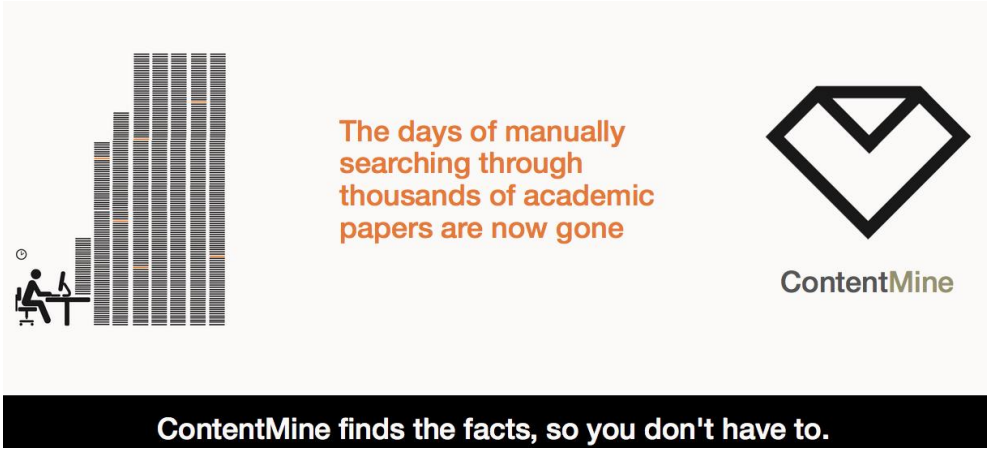

Figure 11: ContentMine website screenshot[49]

ContentMine is a UK non-profit organisation founded by Dr Peter Murray-Rust, a chemist, molecular informatician and advocate for open science.[50] 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 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

[47] Agosti D, Egloff W (2009). "Taxonomic information exchange and copyright: the Plazi approach" (PDF). BMC Research Notes 2:53: 53.
[48] Agosti D, Egloff W (2009)
[49] ContentMine, CC-BY 4.0.
[50] http://contentmine.org/

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.'

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.

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.

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.

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.

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.

Conclusion and recommendations

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.

To help improve the uptake of TDM a group of scholars associated with ContentMine have developed and propose the following principles based on the notion that *'The right to read is the right to mine'*[51]

Principle 1: *Right of Legitimate Accessors to Mine*

We assert that there is no legal, ethical or moral reason to refuse to allow legitimate accessors of research content (OA or otherwise) to use machines to analyse the published output of the research community. Researchers expect to access and process the full content of the research literature with their computer programs and should be able to use their machines as they use their eyes. The right to read is the right to mine

Principle 2: *Lightweight Processing Terms and Conditions*

Mining by legitimate subscribers should not be prohibited by contractual or other legal barriers. Publishers should add clarifying language in their subscription agreements that content is available for information mining by download or by remote access. Where access is through researcher-provided tools, no further cost should be required. Publishers should always explain to subscribers in countries that have implemented an exception to copyright for text and data mining that this exception exists, and should also ensure that in those countries they will not attempt to side-step the exception by adding terms or conditions, or technical barriers, to restrict what subscribers are entitled to do under the law. Users and providers should encourage machine- processing.

Principle 3: *Technical restrictions*

Bona fide content mining should not be restricted by unreasonable or unjustified technical restrictions imposed by publisher servers.

Principle 4: *Agree what is commercial and what is non-commercial*

Publishers should make clear by means of use cases linked to their licences what sorts of downstream activities they reasonably consider to be commercial, and what activities they consider to be non-commercial.

[51] In addition to these principles they have proposed a code for responsible content mining. The code and example workflow incorporating rights clearance can be found in Annex 1.

Principle 5: *Use of Mining Results*

Researchers can and will publish facts and excerpts which they discover by reading and processing documents. They expect to disseminate and aggregate statistical results as facts and context text as fair use excerpts, openly and with no restrictions other than attribution. Publisher efforts to claim rights in the results of mining further retard the advancement of science by making those results less available to the research community.; Such claims should be prohibited. Facts don't belong to anyone[52]

Alongside these principles, the authors proposed a code for responsible content mining (Annex 1). The following is an exemplar workflow for a TDM project detailing the rights required at each stage and the type of activity undertaken (Figure 11).

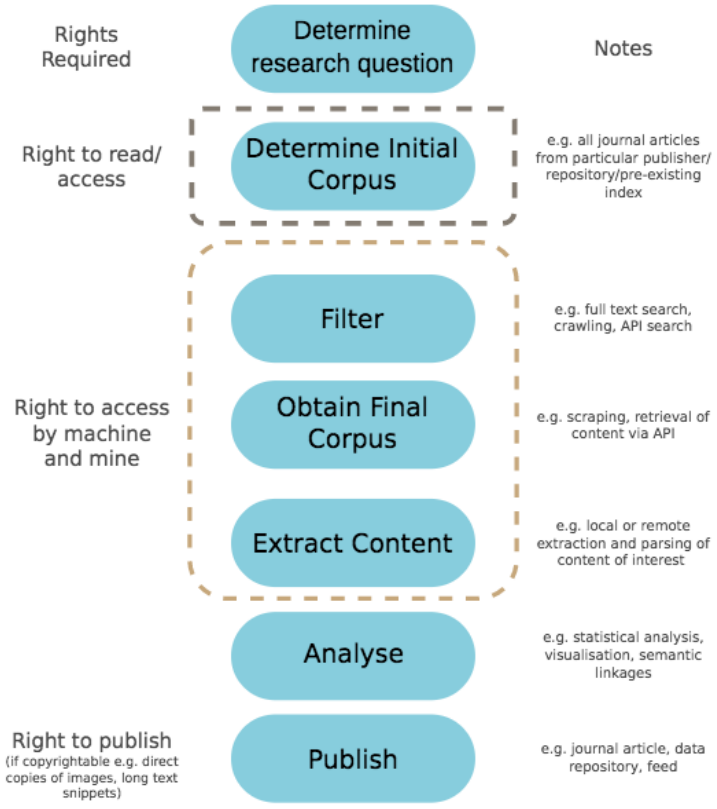

Figure 12: Exemplar workflow for a TDM project

5.4 KConnect: Search technologies for medical information

'Radiologists are drowning in images. At larger hospitals over 100GB (over 100.000 images) are produced per day.'

[52] Responsible Content Mining, Haeussler M, Molloy J, Murray-Rust P and Oppenheim C, June 16, 2015 accessed online at https://contentmining.files.wordpress.com/2015/06/responsible-content-mining-1.pdf

The Khresmoi project

The main goal of the Khresmoi project is to limit the information overload[53] 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.[54]

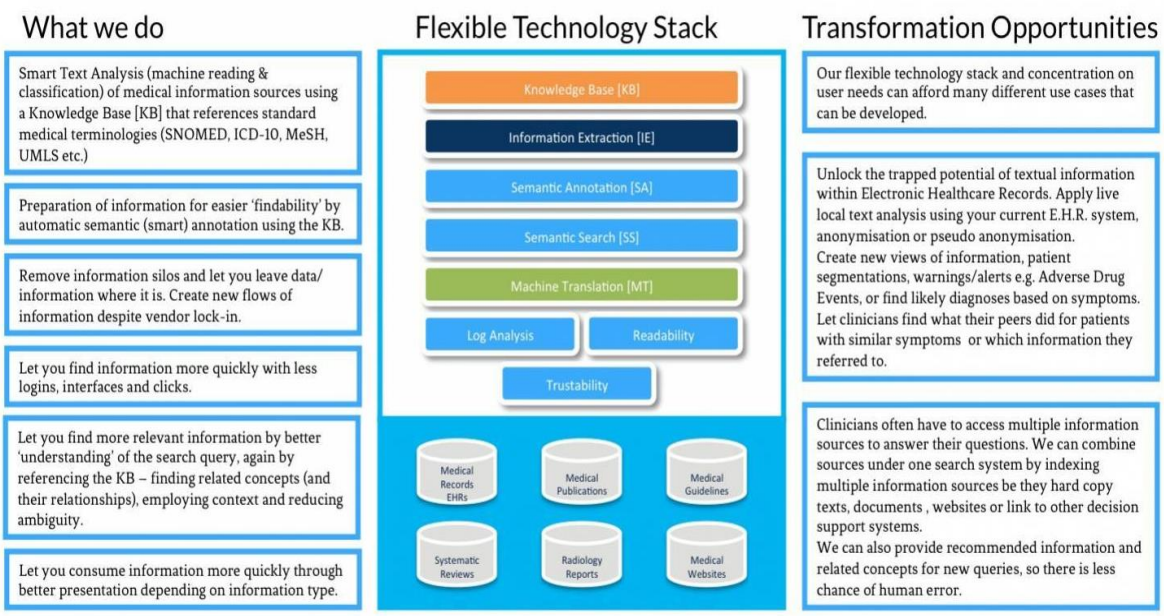

Figure 13: KConnect workflow

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
- 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

[53] 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

[54] 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

- 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

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.

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.

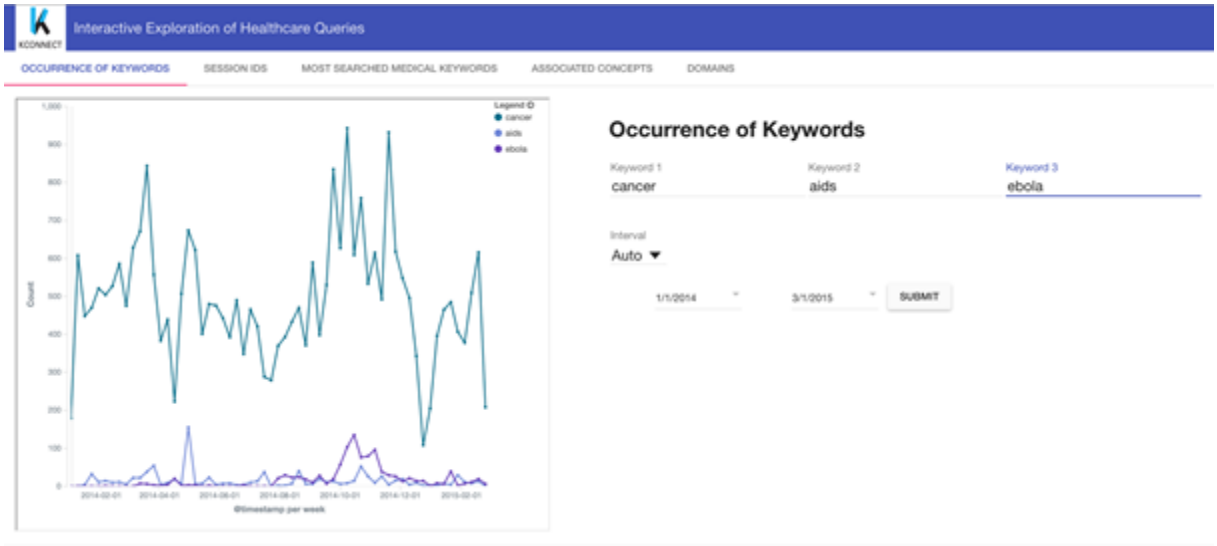

Figure 14: Example of interactive exploration the user is provided with

Text analysis, classification and semantic annotation with the aid of a large medical Knowledge Base allows for improved search (semantic search) results (Figure 14). Text analysis can also add value to textual information that is normally locked, for example inside EHRs (Electronic Healthcare Records).

Figure 15: 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.

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.[55] 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: [56] 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 [57].

[55] '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

[56] 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

[57] 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 an interoperable common data model for PDDIs.[58] Furthermore they should inform users that the lists may be incomplete with respect to PDDIs so that clinicians are aware of this.

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.[59]

The Dementia Clinical Record Interactive Search (D-CRIS)[60] 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.[61]

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).

[58] 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.
[59] 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),
[60] http://www.slam.nhs.uk/research/d-cris
[61] 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 Caldicott Guardian and is chaired by a service user. The Committee is there to ensure that all applications comply with the ethical and legal guidelines.[62] 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.[63]

A second major challenge is license interoperability. The project analyses medical literature cross publishers and makes use of many different types of datasets, vocabularies and ontologies which often have different, restrictive licenses. In practice this makes it complicated to know how these sources can be used and how much you have to pay for using them.

Another serious complication is how to comply with different national systems when providing cloud services. For example, the project makes use of SNOMED CT[64] but because of different licensing models used providing access to researchers from different countries which may or may not have free access to the service, the project would have to implement geoblocking to the cloud service which only adds to an already complicated service. A lack of unified and harmonized licensing provides serious obstacles for the development and commercialisation of products that provide TDM solutions.

Education and Skill challenges

The project has reported having problem recruiting skilled people.

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[65]. 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).[66]

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.[67]

[62] 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.

[63] See access at june 2016 http://www.slam.nhs.uk/research/d-cris

[64] SNOMED CT is the most comprehensive and precise clinical health terminology product in the world http://www.snomed.org/snomed-ct

[65] 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

[66] Samwald, M. & Hanbury, A. (2014). An open-source, mobile-friendly search engine for public medical knowledge. *Proc. Medical Informatics Europe 2014*

[67] 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.

With respect to data quality issues, proposed best practices are:

- Raise awareness and incentivise medical professionals to provide better quality health records and report in a more structured way.
- Reach an agreement on what community standard for abbreviations to use.
- Invest in better technologies to improve the quality of data. Machine Learning can help but there needs to be investment in annotating data and making this data available for machine learning.

Conclusion

The KConnect case study provides insight in the issue of using data that may include personal data. As current technology is not yet 100% reliable in anonymizing data and the consequences of noncompliance are severe many companies will refrain from developing services and tools that would help improve for example medical healthcare. The project however has developed some solutions on how to comply with the data protection regulations while making sure data still holds relevance for further research. As KConnect develops further in bringing the technology to market they will be able to provide more insights into possible best practices dealing with economic and legal barriers for commercialising TDM research projects.[68]

5.5 Mediately

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.[69] It currently offers services to medical 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,

[68] We will continue to monitor and report on the projects developments.

[69] www.mediately.co

manufacturer, whether the cost of the medicine to patients is reimbursed by national insurance schemes etc.

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 has 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.

Economic challenges: Comparisons to the United States

Mediately is conscious that one of their main competitors in the US, epocrates launched 10 years before Mediately did.[70] 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.

[70] https://www.epocrates.com/

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.' [71]***

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 2013.[72] This open API facilitates enormously access to the vast amount of information that the FDA holds, something which does not exist in Europe.

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 is 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

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

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 article 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.

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.

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.[73]

[73] Mediately reports that potential investors are often worried about whether scraping and analysis of material from the internet is legal. For more on the legal aspects of TDM see the FutureTDM D3.3 Baseline report of policies and barriers of TDM in Europe. And for a UK researchers perspective and summary of the legal aspects including scraping of websites: https://blogs.ch.cam.ac.uk/pmr/2016/05/06/sci-hub-and-legal-aspects-of-contentmining/

5.6 Textkernel

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[74], 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[75].

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.

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.

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.

[74] https://www.linkedin.com
[75] http://www.textkernel.com

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[76] 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.

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[77].

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

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

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.[78] 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 science user generated

[76] See FutureTDM Deliverable 4.1. for more details on the language tools availability Online at http://project.futuretdm.eu/wp-content/uploads/2016/07/FutureTDM_D4.1-European-Landscape-of-TDM-Applications-Report.pdf

[77] 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/

[78] https://www.ncbi.nlm.nih.gov/pubmed/advanced

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.

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 meta analysis 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.

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.

International and 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 downstream 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.

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.

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 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.

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 more senior 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.

Conclusion

What became clear from working on the above case study for researchers was that apart from the lack of awareness and uncertainty about many aspects of TDM; researchers were reluctant to talk about their TDM practices. They mentioned not being sure if what they did was allowed or not under the current legal framework. Consequently, they also did not know where to go with their legal questions. The impact of not having legal certainty when it comes to text and data mining is an important factor to be taken into consideration when proposing recommendations.

The following case studies were developed in the second phase of the project and based on the interviews, selected to provide insight not only into the barriers but also focus on what tools and services will help enabler the uptake of TDM.

5.8 ALCIDE

Introduction

ALCIDE (*Analysis of Language and Content In a Digital Environment*) is a web-based platform designed to assist humanities scholars in analysing large quantities of data such as historical sources and literary works.

The first system prototype of the system was developed in 2014 and later redesigned with extended functionalities together with the Italian-German Historical Institute (ISIG) [79]

How it works

The platform gives access to 3 corpora, two in English and one in Italian:

- Nixon's speeches uttered during the U.S. presidential campaign in 1960 (282 documents - 830,000 tokens) [80]
- Kennedy's speeches uttered during the U.S. presidential campaign in 1960 (598 documents - 815,000 tokens) [81]
- All manifestos written by Marinetti between 1909 and 1921 (42 documents - 64,000 tokens) [82]

The system combines a flexible suite of tools to browse through the content of document collections and analyse them along different dimensions, including the lexical, the semantic, the geographical and the temporal level.

The original documents in digital format are converted into XML and then a pipeline of NLP modules processes them to extract a set of relevant information. The project relies mainly on Tint, an open-source NLP suite.[83] All extracted information is stored in a MySql DataBase Management System. In ALCIDE Highcharts are used to present the most common chart types (i.e. bar and line charts), while the most interactive and custom data-driven visualisations (i.e. co-occurrences and networks) are displayed using d3.js. The display of interactive maps is implemented using the Leaflet library.

[79] http://isig.fbk.eu/

[80] Downloaded from the American Presidency Project available online http://www.presidency.ucsb.edu/

[81] Downloaded from the American Presidency Project; http://www.presidency.ucsb.edu/

[82] Digitized by Selena Daly.

[83] Developed at Fondazione Bruno Kessler and based on Stanford CoreNLP that includes modules for tokenisation, sentence splitting, morphological analysis, Part-of-Speech (PoS) tagging, lemmatisation, multiword recognition, keywords extraction, chunking and named entity recognition http://tint.fbk.eu

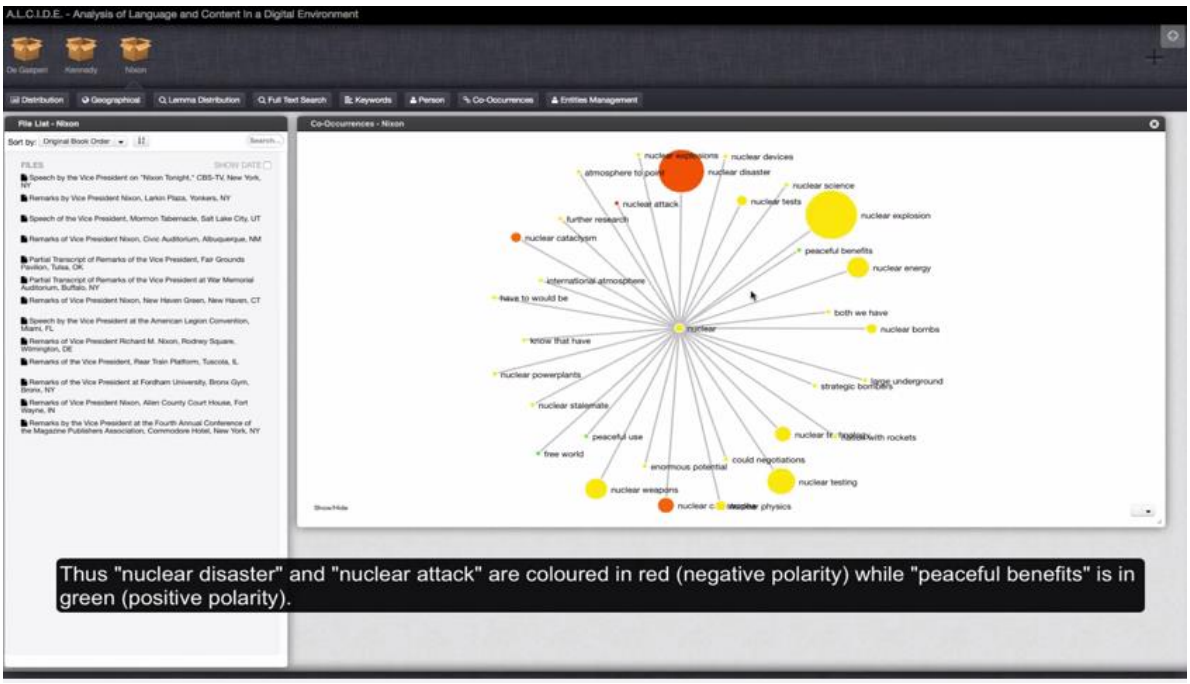

Figure 16: Screenshot ALCIDE Demo video [84]

The goal of the joint project is to enhance historical research, in particular the analysis of political discourse, with state-of-the-art IT technologies. In the first stage of the collaboration the project focused on the writings and speeches of Alcide De Gasperi. The platform is currently extended with new corpora, e.g. presidential speeches.

Legal and content

In digital humanities, legal aspects are a main concern because issues with copyright can stop a project. More and more consortia now have legal experts involved already in the beginning of a project to help negotiate who owns copyright on what and what are people allowed to do.

A given example by the researchers of where issues with copyright impacted a project was a project with a local newspaper whose collection was used as a basis for developing a platform and tools and services. However, when the project ended the newspaper did not agree for the collection to be made available online. As a result, the system and tools the research had developed had to be taken offline after the project had finished.

We don't have any results to show for this project because it's no longer online.'

Copyright

Also within the ALCIDE project there were some challenges. Because the texts from Alcide De Gasperi are still under copyright the project partners had to have permission from the copyright owners to use the corpus. Luckily the rights to the corpus were in the hands of only one publisher who had recently published the texts and the project consortium was able to negotiate the necessary rights for their research. What they did not discuss in the beginning of the project however was whether this also

[84] Available online https://www.youtube.com/watch?v=PhkuOfIod1A&feature=youtu.be

included the right to make the corpus available for download online. As a result, there was some uncertainty whether the platform with De Gasperi's documents would be made available online. Aftera period of uncertainty the project reported that they have reached an agreement with the copyright owner and by the end of 2017 everyone will have access to the texts. What is still under negotiation however is whether the corpus can be downloaded and reused.

Licensing

In practice they choose not to use training data that is licensed and does not allow them to distribute a model trained on this data. The group reports they have had problems with this in the past not being able to share models trained on proprietary datasets. For this reason, they recently put some effort in developing Tint, a NLP suite based on Stanford CoreNLP, which is open source and can be easily shared and modified by contributors, given that many researchers in the NLP community are already familiar with CoreNLP code and analyses.

Research reproducibility

Because for the moment the researchers do not have permission to share the underlying corpus this has had a negative impact on their ability to publish the results in academic journals. They report that an article was refused because they did not have the rights to make the data available alongside the research article.

In addition to having the right to use and share data for TDM the researchers advise to also try and negotiate the right to keep copies of the data for future research.

'You never know what kind of analysis you may want to do now and in the future.'

A similar experience with the barriers for reproducibility when the data cannot be shared is the use of social media data. For example, with Twitter you can show links to tweets but these tweets may disappear over time and when people cannot retrieve them anymore they cannot reproduce the research results.

'Year after year the amount of tweets gets smaller and smaller. For social media researchers this is really an issue because you are not sure that what you do is comparable with what others have done before.'

Open Source Software

The research group is dedicated to Open Source and incorporates this into their research and projects. Their motivation is scientific and based on the US example where it works well. Research groups in the US in the NLP domain have had significant impact because they make their tools available for others for free which has also increased visibility of the work done by the researchers.

Economic issues

The research group want to build a named entity recognizer for Italian language. But all available datasets to train the model are proprietary and would not allow for the model to be released under an open source license. Which is why the group is planning to manually annotate their own corpus, which is very time consuming.

At the moment, EU projects generally do not fund directly this kind of annotation, so the group has to find additional and alternative ways to make this possible.

Skill and education

The group has experience working with social scientists and humanities scholars, whose research benefits from using TDM. In their experience, many social scientists do not know how to gather large amounts of data. They don't know how to get access to data and when they do, they don't have the basic NLP skills to analyze the data.

Many of the social scientists use the same tools which are often old, expensive or too complex. By learning to use the terminal for example and mastering a few commands, it would give them a great advantage and it would be much less expensive.

Awareness

In the NLP community, there is discussion about the need for awareness about the quality of the data collection and the quality of analysis. Researchers must be aware that you cannot draw conclusions about mankind if you have collected your data in a sloppy way. The same goes for AI systems that rely on data processing and can give questionable outcomes if the data that is being fed in the system is questionable.

It is important that researchers learn about the risk of bias in research data and there is a clear need for more diversity in research communities also in the TDM community to avoid research biases.

'People of color and women would not necessarily assume that using social media blogs of a set of similar people equally presents common sense.'

Conclusion

The experiences of the researchers within the ALCIDE project is an example of the importance to have legal agreements about who can do what with the data both during and after the project.

Other good practices and methodologies are;

- to Include rights management in the proposal and have clear agreements on IPR from the beginning to avoid disagreement or miscommunications during and after the project has ended.
- For (social science/humanities) researchers to master a set of relevant tools and techniques that will allow them to know how to gather data and how to analyze large data sets using more up to date tools.
- To promote ethics and have more diversity within the TDM community

5.9 RightFind XML for Mining

Empowering Text Mining with Full-Text XML Articles[85]

[85] RightFind XML product sheet available online at https://www.copyright.com/wp-content/uploads/2015/10/Product-sheet-XML-for-Mining_Life-sciences.pdf

Introduction

A growing number of life science companies use text mining to gather important insights from vast amounts of published information. The results of mining projects inform a wide range of business activities including drug discovery, drug interactions, clinical trial development, drug safety monitoring and competitive intelligence.[86]

RightFind® XML for Mining is a text mining workflow solution developed by Copyright Clearance Center (CCC) with the goal of eliminating the manual work that researchers would otherwise need to perform prior to mining content. [87]

Presently, more than 50 participating publishers have contributed nearly 8 million full text XML articles for mining by users.

How it works

XML for Mining is cloud-based and can be accessed either via a user interface or by API.

A user submits a query and results are returned. Researchers can search across the full text of subscribed, unsubscribed and open access articles. Queries can reference article metadata, the content of entire full text articles, or specific article sections, such as materials and methods, abstracts, conclusions, and/or citations.

Article excerpts enable users to confirm the validity of their search. Results can be filtered by publication year, subscription or open access status. Users can purchase the full text XML of unsubscribed articles directly through CCC's interface, or can choose to simply download abstracts and metadata of unsubscribed articles. Even when they choose the latter, the results are superior to queries run simply across abstracts and metadata, because the initial query was applied to the full text of the articles. Meanwhile, users can, without additional fees, download the full text of subscribed and open access articles.

The user can then load the normalized XML content into their preferred text mining software, such as Linguamatics I2E or IBM's Watson.

After the results are refined and downloaded by the user, they can select to receive weekly, monthly or quarterly updates as new content is published or otherwise added to the database.

The three core benefits XML for Mining aims to provide are;

- It improves the results of text mining because users can search, download and mine full-text articles in XML format from both company subscriptions as well as unsubscribed published material.
- It reduces the time and costs associated with article conversions, content management, and negotiations with publishers, saving time and money for businesses including startups.

86 CCC white paper 3 Reasons to Consider Text Mining available online at http://go.copyright.com/l/37852/2015-08-25/2g9y9m

87 CCC provides services for publishers, businesses and academic institutions on content workflow, document delivery, text and data mining and rights licensing technology.

- By offering consistent licensing terms, it allows users to acquire content for commercial text mining purposes while remaining compliant with copyright law.

'It is not just content that we work to normalize and provide access to, but also the right to mine them it so that companies and researchers feel comfortable from a risk management perspective.'

Barriers Legal and content

What CCC discovered through its consultations with publishers is that they are happy to provide TDM rights to their customers to make the content more available but there are several challenges. The main challenges that XML for Mining addresses are

- content access,
- content normalization,
- consistency of metadata formats,
- limitations on the utility of scraping and converting human readable content into mining-appropriate formats, and
- aggregation.

If for example 10 companies wanted to mine content from 10 publishers, that would require 100 data feeds, 100 exercises in content normalization and 100 agreements dealing with issues such as rights, content, security and output. This is a bigger challenge than simply creating a license.

XML for Mining provides a single standard license across multiple publications, giving users the right to mine the full-text article content for commercial text mining purposes. This solves the problem for users of having to negotiate different licenses when they want to mine across publishers, and also reduces confusion and lack of awareness regarding what researchers can do with the content

Technical and infrastructure

For publishers, XML for Mining solves the issue of not having the bandwidth or technology themselves to provide services to deliver machine-readable content and grant commercial text mining rights on an individual basis. Many publishers have only recently started to publish content in XML format instead of PDF. A lot of data is therefore not yet available. Part of the work CCC is doing is to work with publishers to get the available content in quality formats but also to help them make their backlog available.

XML for Mining provides content to its users in machine-ready XML format. This is more beneficial than when text mining researchers obtain human-readable PDF- formatted content. When a PDF is converted to machine readable XML, it loses metadata, has lower fidelity, and introduces significant noise TDM results. This problem is especially prevalent when content is scraped from a website.

Having access to the full text XML content includes access to tabular data or methods that were not in the abstract. Researchers are now also asking for access to supplemental materials.

Economy and incentives

The XML for Mining business model is based on usage of the service. The more content the user wants to mine across, the more value they will get. Without the service, researchers would have to do this on their own, duplicating CCC's efforts in reaching out to publishers, normalizing data feeds, and so on.

'The pricing takes the value of the service into consideration because it considers the work involved [...], and the technology necessary to provide good precision and recall with full-text search.'

Education and skill

The market is still getting to know some of the applications of full text mining.

Many users are familiar with mining abstracts and metadata but have only recently moved to full text mining. They are beginning to understand the benefits of having access to full text, including the different types of facts and information which can only be found in the full text, such as experimental protocols and methods, and tabular data.

Part of the work is therefore educating both publishers and users on the differences between mining abstracts and mining full text articles.

'Publishers want to be where the need is.'
Michael Larrobino, Product Manager for RightFind XML for Mining (CCC)

Once publishers understand the end user needs and how the platform works, they are interested to know how this will affect the existing relationships they have with their subscribers. And, just like users, they need to be shown how full text mining can accelerate scientific research.

Community feedback and recommendations

The feedback on the service has been positive. Users say it is accelerating project timelines. There are also some recommendations for publishers from users. For optimal text mining, publishers need to provide good quality metadata and a more consistent markup of tabular information.

'We see that TDM is growing as a method and approach to analyzing content'

For corporate end users, it is important to measure the many different ways in which content is used by an organization. A user synthesizing knowledge by reading an article is one aspect; organizations should also consider how articles consumed during the text mining process contribute to business initiatives. This data can be instructive to both end users and publishers in defining where their focus should be.

Best practices for publishers have to do with data are;

- Providing good quality metadata
- To provide a more consistent markup of tabular info
- Get more consistency amongst publishers on how mining can be done across content.

5.10 UNSILO

Introduction

Unsilo is a collaborative search and discovery platform that helps users see patterns across science and innovation.[88] They have developed a sophisticated semantic-based search engine that breaks down the silos of information and makes it easy and fast to find relevant knowledge across different content sources hidden in domain-specific terminology.

[88] https://site.unsilo.com/site/

About UNSILO

When a doctor or researchers is looking for a specific topic, for example “insulin insensitivity in obese children” he may also be interested in documents that have the same meaning but do not use these exact wordings for example research on “overweight girls with reduced hormone responses”.

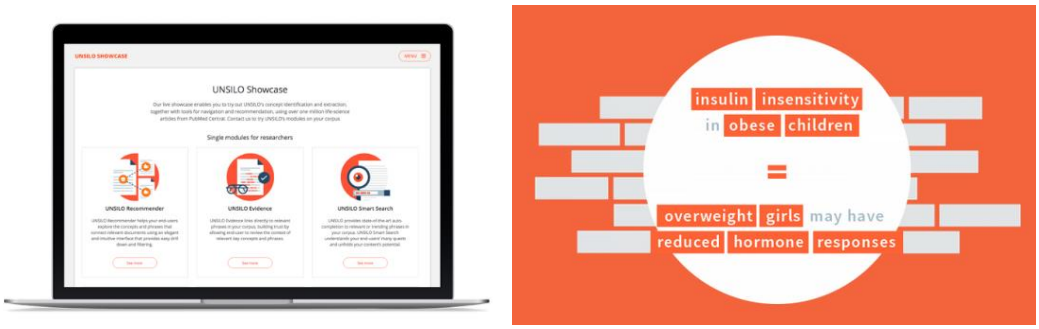

Figure 17: Screenshot UNSILO

UNSILO is developing software that helps researchers find relevant articles and make sense of the scientific literature. This is what UNSILO does, it extracts information from text automatically and matches all documents about for example obese children, even when authors use different words to describe ‘obese’ like heavy, pudgy, or corpulent but also ‘children’ as referred to as youngsters, adolescents or teens. UNSILO captures this multitude of underlying ideas and connections and shows the user what the relevant documents are and why they have been selected as relevant.

UNSILO works with scientific publishers to enrich their content and improve discoverability across domains and disciplines. The UNSILO discovery tools not only capture trending ideas and novel concepts as they emerge, they also help researchers find articles that describe parallel research of similar ideas across different domains and disciplines. Publishers for example want researchers to not only come to their websites to get an article but also to stay and find other interesting articles that are relevant for them too.

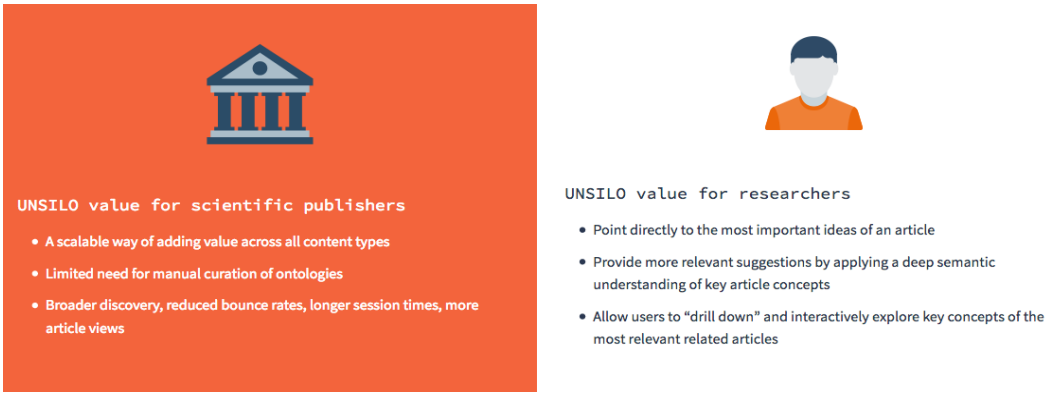

Figure 18: Screenshot UNSILO

The UNSILO analytics engine uses pervasive semantics and machine learning to extract the substance of every document which is examined. By abstracting phrases that represent important concepts and filtering them out they can then figure out which are the most important. Differently from traditional search tools the user is not just presented with an endless list of links but instead the search results are clustered into groups of documents, which share a similar approach or solution strategy. This facilitates the discovery of relevant insights from unexpected sources.

The results are visualized based personal workflow preferences. The process can also easily be shared with colleagues and peers, to facilitate knowledge generation and improve the speed of innovation.

How it works in practice

A common complaint about machine-learning tools is that they are black boxes, operating in ways that cannot be explained. In the case of UNSILO, the automated concept extraction engine has a very precise and well-documented activity. Following is an example how the UNSILO's engine works with related concepts rather than just with strings.

Take, for example, the phrase "secondary brain injury", a medical concept used widely in academic text. A search using the UNSILO showcase (see Figure 18), which comprises around a million medical articles reveals hundreds of hits for this concept, such as the article below.

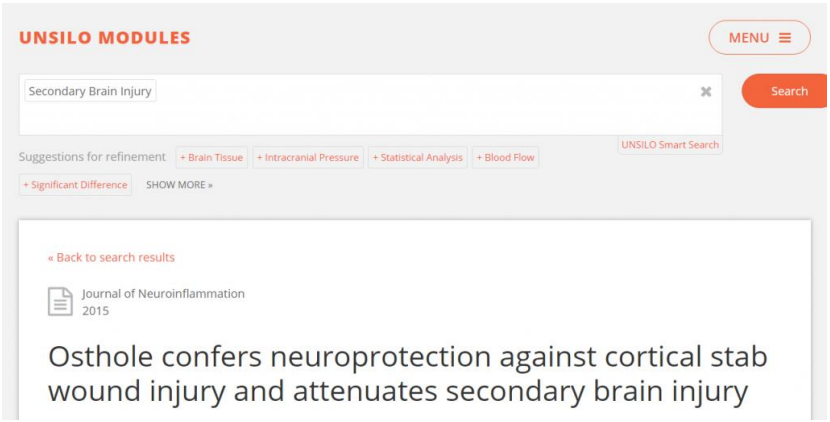

Figure 19: UNSILO module screenshot

Any term or phrase entered in the search box is automatically matched to the closest concept in the index (see Figure 19). However, the English language being what it is, many researchers use the similar concept "secondary brain damage", and a search for this phrase in the Showcase reveals a separate set of hits.

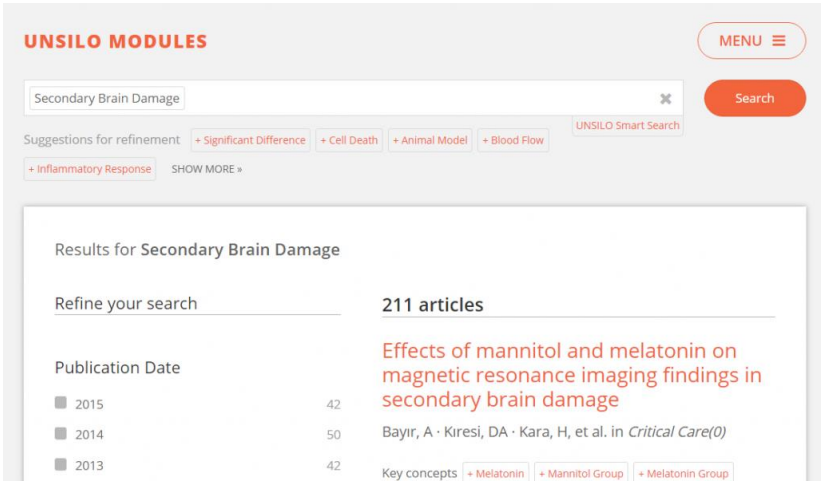

Figure 20: UNSILO module screenshot

Most search tools operate on a string basis, which is not quite what is required here, in the context of academic search. For example, most search engines, including Google, will automatically expand a search syntactically, and so will search for "brains" as well as for "brain", and "damaged" when the user keys in "damage".

This query expansion can cause confusion rather than clarification: for a medical researcher, "secondary brain damage" is the concept they are interested in, not many syntactical variations on the constituent terms. The plural form "brains" will never appear in this context. However, a researcher is very interested indeed in related phrases, including terms that may have little or no syntactical relationship with the original, but a clear meaning relationship. How can we tell there is a meaning relationship? UNSILO's machine-learning capabilities, using statistical methods, identify that many researchers use the phrase "secondary brain injury" as largely synonymous with "secondary brain damage" in articles with a similar context. A really useful indexing tool would identify both phrases as equivalent.

UNSILO does indeed do this: a glance at the concepts extracted by the UNSILO engine shows what is actually taking place. Behind the scenes, the concept extraction engine identifies synonyms and similar expressions in this context, and expands the query by concepts, rather than just by syntax tools. If you look at the concepts identified by the engine, you can see on the second line of the concept results, the engine has expanded the query to include not only "secondary brain injury" but also "secondary brain damage" (see Figure 20).

Figure 21: UNSILO module screenshot

To identify related concepts, rather than just similar strings, is the achievement of the UNSILO automated concept extraction tool, making searches more precise, and related content more rapidly discoverable.

UNSILO makes sure all ideas are tracked from their inception. Each time a new document is published, it is automatically imported and connected to the full corpus. This ensures that the relationships between all documents in a corpus are constantly updated when new research occurs.

The issues for UNSILO: Technical and infrastructure

Discovery is Limited by Manual Tagging

When a researcher publishes a paper, keywords that describe the paper's topics are manually added. But these keywords often cover only some of the actual topics in a paper. When another researcher searches for similar, but not the exact same keywords, it makes it nearly impossible to find the paper again.

Using Triple stores for concept extraction

One major benefit of triple stores is the ability to make inferences from a content repository. But there are many difficulties in creating and maintaining metadata in triple-store form, and using triple stores at present is not a common practice with most publishers.

SPARQL is the most widely used query language for triple stores, but it is not simple to use for non-expert users. Writing queries using SPARQL is too difficult for average users, so a natural-language interface needs to be added but most users are used to keyword search, which can be difficult or even impossible to translate into a meaningful SPARQL query.

Scale

There is currently no standard full-text SPARQL interface to do simple keyword search. There is a problem with scalability: a system being able to manage billions of queries as well terabytes or petabytes of data. Triple-store solutions are complex to scale, and therefore most solutions rely on large replicated nodes that each hold all data in memory to perform well, but this limits the amount of data that can be indexed without sacrificing query performance.

UNSILO has developed software solutions that use machine learning to extract concept and triple data directly from the full text of a paper, which reduces the amount of manual effort required. This type of approach will reduce the cost of creating and maintaining metadata in triple-store form, and likely to help innovation in the fields of storing, scaling, and querying rich knowledge repositories, whether in triple-store form, or in other types of databases.

Tools

There are out of the box tools but in Humanities researchers often want to do difficult things. So, in practice, your research needs may not be served by the tools available. Furthermore, Humanities researchers may not have the technical background to coordinate a push for the development of the appropriate technologies.

UNSILO is developing solutions that automate the content enrichment processes, which at present is the most costly and time consuming part of text mining. The UNSILO Document Enrichment API offers a unified and standardized metadata layer that allows publishers, research institutions, as well as other technology startups to build advanced text analytics and knowledge management applications directly on top of large repositories of natural language documents.

The ability to directly track and compare knowledge across large text collections may unlock new research and innovation appliances.

Education and skill

The implementation of new technology in publishing has often resulted in an increase in the headcount, to manage the complexities of the software being introduced that is supposed to save labor. Managing a triple store and using it via SPARQL queries requires skilled staff as well as considerable computing power.

Legal and content

Open Source and Intellectual Property

UNSILO has experienced some issues in not being able to get commercial licenses to use specific

software and data sources.[89] For example, some open source TDM tools developed by non-profits in the US are offered free for non-commercial use, and are not available under a commercial license. This means that academics might use them for research, but TDM startups cannot use them to build products. It is important that legislation and public funding supports Open Source software development, but unless such software can be used to build commercial products, startups will have to spend time and resources recreating functionality that already exist in the academic realm. UNSILO recommends that Open Source software should be validated in commercial applications before further public funding and support can be awarded.

Access to content

Although in general Open Access will be good because it provides easy legal access to a constantly growing corpus of high quality content, the downside is that the data large variation in the formats used to store the content may not always be good for easy processing of the data. UNSILO is collaborating with CORE[90] which transforms and stores content in a standardized and uniform fashion.

Another issue UNSILO faces is difficulties in obtaining licenses to text mine commercially available bibliographic databases such as the SCOPUS database. This could be evidence of a strategic decision of the part of the database owners to exclude potential competitors, but regardless of the cause, this makes it difficult for startups to enter the commercial market for products aimed directly at researchers.

Uncertainty about the rights to specific data sets is also a barrier. For example, accessing the project Gutenberg website to download a bulk of articles resulted in being blocked. However, it was not clear why this was or what was the 'right' way to access the information because the website did not provide clear information about the rights on its content.

'There are all these different players who own pieces of the puzzle and it's difficult for startups to put the puzzle together.'

Economy and Incentives

As a startup, you have to choose how to use your often limited resources and it may not be the best choice to redo the tools that are already out there.

The market for TDM does not feel overcrowded. There are indeed a lot of companies that do something similar to UNSILO but there is also a lot of work to do and different ways to carve out a specific niche. Especially in the more high-end, there are only a limited amount of companies that have the expertise.

It helps when there is a supportive environment, for example in Denmark. It is easy to get support and there is very little bureaucracy to start a company. As soon as the company grows you will have a lawyer to consult

Funding for TDM projects

European projects on digitizing data are popular but the next step is also to do something with the data and that is currently a bottleneck.

[89] For example, open source tools such as those developed by the Alan Turing institute for AI are only made available for researchers and for non-commercial use.

[90] The world's largest free repository for Open Access content

Traditionally there is not much funding and then there are other issues such as

- In interdisciplinary projects and areas where TD could be applied there are often not enough people involved with technical background. So for people having the technical as well as the subject knowledge is rare
- If you need computer resources in traditional funding models this is not included, for example in literature it is difficult to fund time on computing.

Conclusions and recommended practices

Startups that apply TDM to scientific literature face the same challenge: Access to high quality data, which in the case of the most comprehensive abstract and metadata databases is presently either locked inside commercial products owned by a few entrenched monopolistic players, and, like the most popular research articles, not Open Access at all.

Every year, more articles become Open Access, but it's taking time, and in order to provide high quality services to researchers, startups need access to complete and comprehensive metadata for all of science, not just the articles that are Open Access. Therefore, it is recommended that abstracts and metadata is free and publically available for anyone to build services on top of.

Robust Open source software for text processing and data analysis in the cloud is a great driver of software innovation. From the industry perspective, legislative focus should be on fostering competition and providing funding and support only to initiatives that can be applied in commercial projects, to ensure that academic software development supports the technology startup community.

Creating highly structured data for the semantic web needs a lot of human involvement. It is more efficient to extract concepts and relations automatically without relying solely on human-created ontologies.

- Focus on challenges that presently have huge costs and huge potential savings associated with them. UNSILO combines machine learning with NLP tools to do complex natural language parsing and corpus-wide semantic analysis. UNSILO for example can make accurate and precise links between content objects, which means a more effective recommender engine, or a peer reviewer system that matches authors and reviewers more precisely than a human can and in far less time.
- Cloud computing: Using Amazon cloud instead of buying and maintaining a room full of servers allows startups to process large corpora and compete directly with large software companies.[91]
- Use Open Source components and well-documented formats.[92]

'Improving existing tools may become interesting in the future but there needs to be a business case for it otherwise it's too much effort.'

- Respect the policy of the website whose content you are looking to mine. Some websites like wikipedia allow a dump of their content you can just download the data version which is convenient for everybody.
- Data Quality: Make data available in well-documented formats such as XML, JSON, or LATEX.

[91] UNSILO rents computing power based on the needs of the company to keep the company agile.
[92] Instead of rebuilding what others did UNSILO focuses on building value on top of what is already out there.

Machines already surpass or equal humans in many type of tasks within the areas of text, sound and image analysis, and machine-based analytics will increasingly outpace human capabilities in the years to come.

5.11 Tool evaluation in the Digital Humanities

Introduction

Digital humanities is a diverse field of study that combines a number of different interactions between humanities disciplines and the use of the computer. From the edition of manuscripts in digital form to the use of geographical information system in historical research, from man-computer interactions in media studies to the development of digital libraries, this field of study has gradually attracted more attention.[93] As more and more digital sources such as born digital scientific publications available in repositories of academic institutions researchers becoming available there is also a growing awareness for the need for computational tools to help with data analyze.

When discussing the research on the usefulness and necessity of advanced text mining approaches in the digital humanities the researchers encountered the following barriers.

Legal and content barriers

Access

A barrier mentioned for Digital Humanities is getting the data for research. For the web archive community, people work with large archive of the web. National archives often don't release data for text mining.[94] You only have access through a search engine and are for example not able to scrape the data. As a result, the data cannot be used. There are legal and computational reasons why the data is not shared as people often don't know how to share these often-large datasets.

For the purpose of the workshops it was possible to get permission and access to use only small parts of the archives.[95]

What is confusing is when archives are 'open' but do not provide an easy way to access their data for example via an API.[96] You then have to write your own code to be able to download the data which is what most researchers will end up doing. But this requires programming skills.

[93] See research at the International Centre for the History of Universities and Science (University of Bologna) and overview of projects at University of Mannheim http://dws.informatik.uni-mannheim.de/en/projects/current-projects

[94] Hockx-Yu, H. 2014, Access and Scholarly Use of Web Archives Alexandria, Vol 25, Issue 1-2, pp. 113 – 127, , 10.7227/ALX.0023

[95] http://archivesunleashed.com/

[96] On pro and cons of twitter API: Morstatter, F *Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose* Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13

'When an archive provides an API, you see that this is the data most people will end up using.[97]'

Another issue researchers face is not knowing whether the data made available via API is complete. This can for example be hidden somewhere on the website or in a footnote. It should be made clearer in an obvious way. Also, it is not clear why the data cannot be made available or shared.

This has a negative impact on research when for example a political scientist proposes something but his analysis and the data cannot be reproduced. What people often do is share the code they used to gather the data. If you want to find the data you can then use the same methods they used. However, because this is often 'research' code it can be messy code and not easy to use for others.

A best practice in this regard is to document all the parameters you have tried and the way you have tuned them to get results so others can reproduce your data but also may correct you or give feedback on what you are doing. It is not easy to provide proper documentation because there are a lot of small steps you are taking while you are processing the text.

Learning coding is actually the easy part, understanding what you are actually doing is more difficult and it helps as a social scientist to discuss with people from different fields about the possibilities and risks involved.

Technical and infrastructure

People have different ways of archiving and using data so there is no common practice. Some share their data in CSV files that other can just download and use where others use a database where they may provide a complicated access system you need to learn to understand how to get the data.

'If people want to use your data, they will find out how they can use it.'

Another problem is the storage and availability of data. Often it is the case that the data is available for as long as the researcher is affiliated with the institution. [98] After they leave the dedicated website may disappear or the URL changes and the links no longer direct to the data. This is a big problem and could perhaps be solved by collaboration with University Libraries.[99] They have the experience and knowledge to deal with data. For example, it is important to think about the future of the dataset and where it will be stored and for how long.

'Open source tools provide access to the code and see how they work but often it is not so trivial to understand what goes on.'

[97] And it's the same thing with tools, people use topic modeling with Mallet (http://mallet.cs.umass.edu/topics.php) because it's easy to use – http://programminghistorian.org/lessons/topic-modeling-and-mallet

[98] This is in general a huge problem in academia: http://www.nature.com/news/the-trouble-with-reference-rot-1.17465

[99] This is already often the case – libraries offer the storage, but maybe what misses is communication between researchers and library people.

Education and skill

To do good text mining you have to be an expert in the things you are mining as well as knowing how to properly use the methods. At the moment, we don't see many people who this kind of profile but the next generation of students may be able to do this.

Most of the people who are doing text and data mining in the humanities and social sciences are self-taught. The downside of people who are self-taught in computer science using books and online courses such as coursera is that they may not have learned the right practices. Using the right terminology to describe things so people can understand but also to clean up your code and provide documentation so others can use what you have developed. For example, using text there are many problems with encoding characters. You can share your research but without knowing how you have coded the language others cannot use it.

Another issue when people lack proper skills is that they may take conclusions based on a tool they do not understand. This can be problematic when they use tools that have certain assumptions in them such as topic modelling. This is a method to detect topics in a text. If you don't know that the model you use relies on assumptions because you lack proper training in how these models work you will also not be able to detect when something is wrong.[100]

An important aspect is to learn and understand how to work with the limitations of a tool. There will always be a level of error. When you prepare a proper evaluation of the different tools you will be able to detect when a tool gives a wrong output. If a tool is consistent in this and you become aware of it you can still use the tool to get good results as long as you know how to correct this in your analysis.

It is important to emphasize also that there is a different approach when teaching different disciplines. For example, when teaching social scientists, the topics cover how to deal and evaluate with the tools and how to for example code in Python. When teaching computer scientists, they often have trouble seeing it less as a task that you have to solve but to understand what the research question is.

Recommendations

There are different text mining tools. Some are very simple some are very complex such as convolutional neural networks but you don't know in advance which one will work for what you want to do based on time and resources available. Instead of choosing the most complex and newest tools, it is important to find the one that solves the problem in the smartest way possible.[101]

Improving TDM skills tool evaluation helps researchers to;

- Better define what they want to do (methods, i.e., "how exactly you did something in your research" are not usually discussed in the humanities.[102]
- Better understand that text mining methods are not black boxes or black magic. They are based on specific assumptions and employ statistics and probability theory and – because of

[100] For an introduction to topic models: https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/

[101] On evaluation of topic models see DH2016 https://hal.archives-ouvertes.fr/hal-01483336

[102] The so called "historical method" does not include how sources are identified, analysed and selected some sources and not others.

that – they of course make mistakes. If you understand the assumptions and follow how these assumptions have been embodied into an algorithm, then you can also try to understand why this approach is making some specific mistakes. And if you do that – you can also try to deal with that in research.[103]

- Have a more clear vision of concepts such as qualitative and quantitative research and the potential and limits of both when computational methods are involved.

5.12 CORE

Figure 22: Screenshot CORE website

5.12.1 Introduction

CORE is a global large-scale Open Access aggregation platform that offers access to a large volume of free and open access content.[104] It offers approximately 70 million of bibliographic metadata records and over 6 million of full-text research outputs. The content originates from open access journals and repositories, both institutional and disciplinary.

Background

The last decades have seen a massive increase in the amount of Open Access publications in journals and institutional repositories. Having large volumes of state-of-the-art knowledge freely available online provides benefits in many fields. It helps for example to reduce time and money spend on getting access to and gathering of these publications for research.

Mission

CORE's mission is to aggregate all open access research outputs from repositories and journals worldwide and make them available to the public. In this way, CORE facilitates free unrestricted access to research for all.

CORE:

- supports the right of citizens and general public to access the results of research towards which they contributed by paying taxes,
- facilitates access to open access content for all by offering services to general public, academic

[103] For an interesting example of this regarding the entity "I_Need_To_Know" see Lauscher, A., Nanni, F., Ruiz Fabo, P. and Ponzetto, S.P., 2016. Entities as topic labels: combining entity linking and labeled LDA to improve topic interpretability and evaluability. IJCol-Italian journal of computational linguistics, 2(2), pp.67-88.
[104] https://core.ac.uk/

institutions, libraries, software developers, researchers, etc.,
- provides support to both content consumers and content providers by working with digital libraries, institutional and subject repositories and journals,
- enriches the research content using state-of-the-art technology and provides access to it through a set of services including search, API and analytical tools,
- contributes to a cultural change by promoting open access, a fast growing movement.

CORE harvests openly accessible content available according to the open access definition as was stated in the Budapest Open Access Initiative.

By 'open access' to this literature, we mean its free availability on the public internet, [..] without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited.

How CORE works

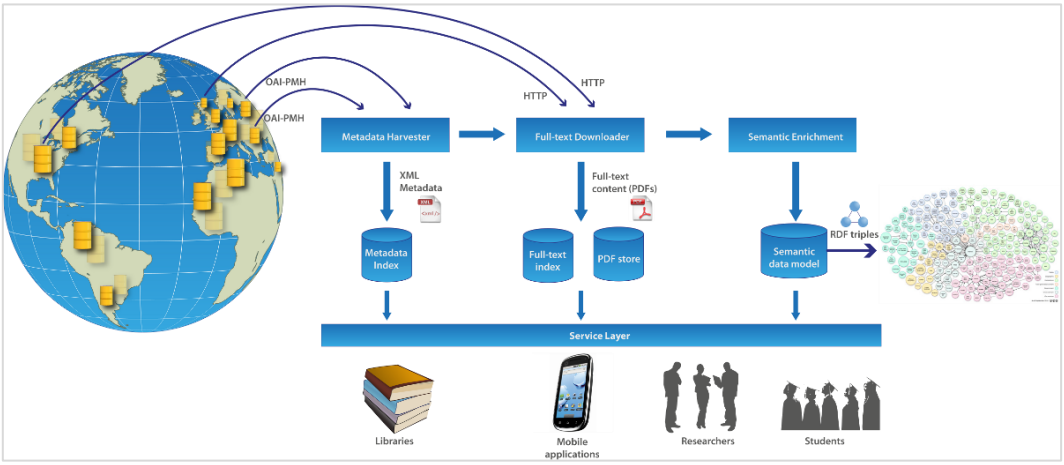

Figure 23: The CORE processes [105]

The Metadata and Content Components

In the metadata and full-text content aggregation phase, the CORE system harvests metadata records and the associated full-text content from Open Access repositories and journals listed in CORE. The harvesting of the metadata is performed using OAI-PMH requests sent to the repositories.[5] Successful requests return an XML document containing information about the papers stored in a repository. A good practice in repositories is to provide as part of the metadata the links to the full-text documents[6].

The CORE system extracts these links and uses them to download full-texts from repositories. The system then carries out format conversions, such as the extraction of plain text.

The CORE system supports the harvesting and downloading of content from multiple repositories at the same time and has been optimised to utilise architectures with multiple processors. The harvesting component in CORE can be controlled using a web interface accessible to the system administrator.

Information Processing and Semantic Enrichment

The goal of the information processing and semantic enrichment is to harmonise and enrich the

[105] http://www.dlib.org/dlib/november12/knoth/CORE-updated-diagram-scaled.png

metadata using both the harvested metadata as well as the full-text content.

In addition to what has become the standard in aggregation systems namely to provide metadata harmonisation and cleaning, the CORE system has additional ways to utilise the full-text.

After running a standard text preprocessing pipeline including tokenisation, filtering, stemming and indexing of the metadata and text. A number of text mining tasks is then performed.

- Discovery of semantically related content — information about the semantic relatedness of content can be used for a number of purposes, such as recommendation, navigation, duplicates or plagiarism detection.
- Metadata extraction
- Extraction of citations and citation resolution — CORE extracts citation information from the publications full-text. This information is used in turn to check if the (cited) target documents are also present in the CORE aggregation to create a link between the cited publications.

Information Exposure

In the information exposure phase, the system provides a range of services for accessing and exposing the aggregated data.

Relevant for text and data mining of the scientific literature is that CORE provides access in various ways which also helps to enable the development of new artificial intelligence-based applications for scientists. At the moment, the services are delivered through the following applications:

- A web-based portal for searching, exploring and accessing the aggregated content.

Because CORE ensures the availability of information specified in the metadata, all search results produced by the system as a response to a user's query will contain links to openly accessible full-texts (unless explicitly requested otherwise by the user), for the purposes of availability and reliability cached on the CORE server. In addition to search, the CORE Portal offers other services on top of the aggregated Open Access collection utilising the information provided by the lower layers, including content recommendation and navigation, duplicates filtering, citation extraction, etc.

- CORE Recommender[106]

This is a platform- and browser-independent plugin for digital libraries & institutional repositories that provides information about related documents. The plugin recommends semantically related papers to the document currently being visited and the recommendations are based on either full-text or metadata.

- An API enabling external systems and services to interact with the CORE system.

The REST API supports tasks such as searching for content using various criteria, downloading documents in PDF or plain text and getting information about related documents.

- CORE Dataset

Users can download all aggregated and enriched metadata and textual content from CORE.

- CORE Repositories Dashboard

The dashboard improves the quality and transparency of the harvesting process of the open access content and provides a two way collaboration between CORE and the content providers. The purpose of the Dashboard is to improve the control with other systems, the harvesting process and broaden

[106] https://core.ac.uk/services#recommender

the discoverability and dissemination of the open access content.

Challenges

Technical and Infrastructure

There is a need for a technical infrastructure for Open Access (OA) research papers which should not only provide search functionality, it should provide support at different access levels addressing the needs of different user groups. One of the most important user groups is comprised of researchers and developers who need access to raw data so that they can analyse, mine and develop new applications. Such an infrastructure does not exist, and we claim that its nonexistence is hindering the positive impact of OA. The CORE system attempts to fill this gap, providing support at different access levels.

Methods and tools are necessary to provide analytical information, including trends, about the OA content. This will strengthen the argument for both academics and publishers to adopt Open Access as a default policy.

Legal and content

Copyright law and other barriers are limiting the use of semantic enrichment technologies, namely text-mining. This creates a strong argument for the wide adoption of OA in research. If semantic enrichment technologies are applied as part of an OA technical infrastructure in a way that provides significant benefits to users, users will prefer OA resources and this will create pressure on commercial publishers.

Education and skill

To fully exploit the OA reuse potential, it is important to better inform the OA community about both the benefits and commitments resulting from OA publishing. In particular, publishers should be aware of the fact that the content they publish might be processed and enriched by machines and the results further distributed. Similarly, the academic community should be better informed about the benefits of increased exposure and reuse potential of their research outputs due to these technologies.

Apart from a few successful OA journals, such as those maintained by PLoS or BioMed central, it is still believed that OA journals today typically do not compare in terms of impact factor with their commercial counterparts. A completely transparent technical infrastructure can help establish new measures of scientific importance.

Conclusions and recommendations

In order to achieve the full potential of having knowledge available, it is necessary to develop systems that:

- make it easy for users to discover and access this knowledge at the level of individual resources,
- explore and analyse this knowledge at the level of collections of resources and
- provide infrastructure and access to raw data in order to lower the barriers to the research and development of systems and services on top of this knowledge.

CORE addresses these needs by providing a system that helps institutional repositories, individuals, researchers, developers, funding bodies and governments.

Furthermore what OA needs is a technical infrastructure demonstrating the advantages of OA policy over traditional publishing models.

5.13 PaperHive

Figure 24: PaperHive Logo

Introduction

PaperHive was created in 2016 by Dr. André Gaul together with Alexander Naydenov.[107]

To help solve the problem of having to spend enormous amounts of time reading during the research process, PaperHive is a web-platform for collaborative reading. It is a tool that allows researchers to engage in collaborative reading which makes reading more effective and efficient. Through Paperhive researchers can easily discover, share and annotate content from different content providers. After starting with arXiv, PaperHive continuously integrates further content on its platform. For example, last year example the collaboration with Elsevier now allows users access the over 12 million articles on ScienceDirect.[108]

PaperHive is part of the startup incubator of the Centre for Entrepreneurship at TU Berlin.[109]

What PaperHive does

PaperHive introduces a platform that allows for seamless discussion of research papers directly in the browser. The platform enables researchers to attach questions, corrections, formulas, figures, further literature, code, or data directly to the original text where everyone can benefit from it. Having in-text discussions about research introduces a deeper level of engagement and understanding of what is being read.

PaperHive has found researchers are benefitting from their software in the following ways;

More productive reading.

By having a platform that enables research to be understood more easily through comments and discussions, researchers are able to devote more time to other academic literature of interest or other research activities.

More learning opportunities.

As people document their questions, thoughts, and ideas within a text, future readers have the opportunity to learn from these documented insights. The value and impact of articles is increased. There are multiple applications in university lectures and seminars.

Increased visibility for both authors and readers.

In addition to authors gaining more exposure when their articles are commented on, readers that respond to an article have the opportunity to share and raise awareness about their own work when

[107] https://paperhive.org/
[108] http://www.sciencedirect.com/
[109] http://www.entrepreneurship.tu-berlin.de/

relevant. The interactions taking place through collaborative reading are also opportunities for sharing and networking (see Figure 25).

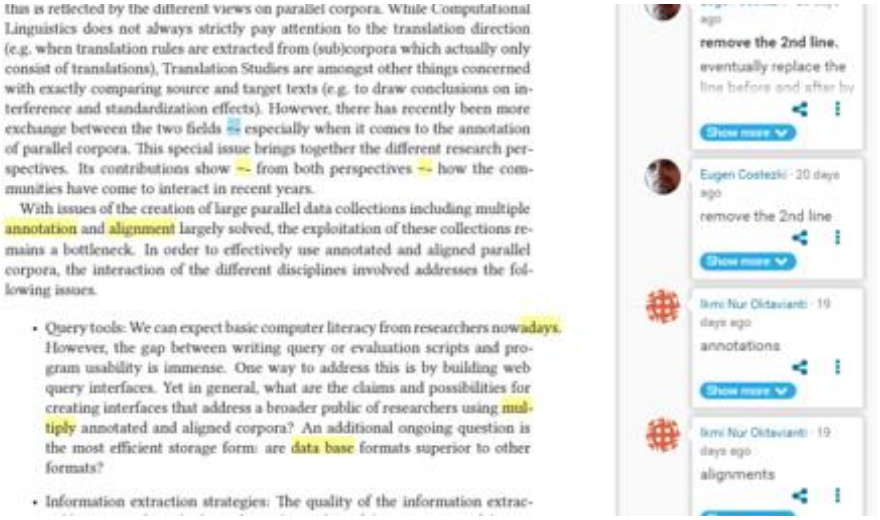

Figure 25: Paperhive screenshot community proofreading

When publishers integrate their journals and books with PaperHive, readers are able to interact on pages and sections of a text by commenting. Readers may leave initial comments, or they can respond to comments others have made. If a person replies to a comment, others will receive a notification and can continue the conversation. With the implementation of the software, users can comment directly into the digital version of the text which can then be read by any other person who accesses the article (see Figure 26). Documents of interest can also be followed, and users will receive a notification if there are updates, new insights, or questions in the future. It gives researchers a reason to return to the content and discover new enrichments they can benefit from.

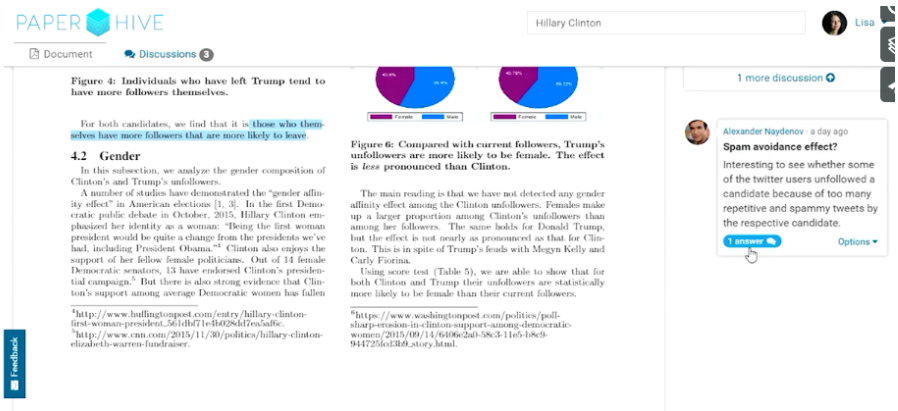

Figure 26: PaperHive example of annotation

The license chosen for the contributions by the public is the CC-BY-4.0[110] allowing reuse while requiring proper attribution of the author.

[110] https://creativecommons.org/licenses/by/4.0/

Ongoing developments

In addition to integrating further publishers' content the team is working on extending PaperHive's feature set for readers, including having private channels for discussion to allow private groups to collaborate together for example as part of a course or specific research project.

With respect to challenges for a TDM start-up PaperHive reported the following issues.

Legal and content issues

'Just putting some content on the web does not mean people know what they can do with the content.'

PaperHive learned from their experiences with ArXive.org which has a 'home brew' license that allows users to choose their own license that the preferred license is one that is as permissive as possible and clear. Unfortunately, it still is the case publishers and repositories do not have clear licenses that explain who has access and for what use. Often it states the right to 'read' the content but not whether this includes TDM or what can be done with the results for example for commercial use. Contracts with publishers may allow mining but are unclear about what can be done with the results of the TDM project.

This uncertainty about the legal aspects of TDM may be a reason why there is a lack of start-ups in the field of text and datamining.

PaperHive extends the concept of a living document and offers an innovative way of displaying content without hosting it. All article traffic goes directly to the publisher's server. The academic documents are dynamically pulled from the publisher's servers. Only users with the proper access rights can view subscription-based content covered under their institutional licence. Paperhive would like to be able to do more with the content including text and datamining services on full content instead of only abstracts. This at the moment is a problem as the current legal framework is unclear.

The Paperhive API includes the metadata for around 16 billion documents. To improve the search results, it would be useful to have access to not only the metadata which may or may not include abstracts but also to the full text articles. This at the moment is not the case for many of the publishers content due to uncertainty about the legal regulatory framework.

Paperhive would like to develop additional services such as recommendations based on TDM but this can only be done when there is more legal clarity about the access and what can be done with the content.

Another aspect is the use of different licenses or the absence of clear communications about what can be done with the content across publishers and repositories.

Technical and infrastructure issues

Data formats and quality

The platform is agnostic with respect to formats but to support as much content as possible PaperHive is currently using the PDF format. Acknowledging the problems with PDF's for text-and datamining

they had to make a tradeoff because PDF is the most widespread data format in use. By the end of the year additional formats will be included making it possible to use the content in more meaningful ways.

When metadata does not include important information such as author and article title, this is an issue because PaperHive does not include articles and books with incomplete metadata because it impairs the user experience.

Data Standards

As standards are important PaperHive follows W3C Web Annotation standards and is a part of the Annotating All Knowledge Coalition.[111] Furthermore all discussions that take place on the platform are safely archived with trusted preservation service providers.

Conclusions and recommendation

Recommended practices from the PaperHive case study include the following points:

Data quality

The quality of the existing metadata should be improved significantly. Valuable information to be included in the metadata are:

- Information such as author or title
- Links to full texts in the metadata of all articles and books
- Information about the format of the full text (e.g., PDF, EPUB, HTML)

Licensing

What is needed for start-ups is to have licensing that states clearly what can be done with the data. It would be helpful if there was one kind of license being adopted as a standard. Publishers should ideally clearly indicate what can be done with the content and not create their own rules. Licenses are already available they just have to be used by the community and not have additional exceptions. It is impossible to respect all these different rules on one platform, as this would mean having to deal 7000 different rules and licenses.

Using the CC-BY license for content is a good development.[112] The Creative Commons licenses are clear and people in general know what they stand for.

[111] https://hypothes.is/annotating-all-knowledge/

[112] The Creative Commons Attribution license allows re-distribution and re-use of a licensed work on the condition that the creator is appropriately credited. https://creativecommons.org/licenses/by/4.0/legalcode

6 CONCLUSIONS AND RECOMMENDATIONS

6.1 Main findings on barriers, practices and methodologies

The interviews on which the case studies were based aimed to challenge and provide evidence for the barriers and enablers for the uptake of TDM in Europe. With the case studies, this report provides insights on selected issues that stakeholders are facing, and the practices they have employed to overcome them. What has become clear is that although there may be a lack of general consensus between the different stakeholders on what constitutes as best practices, practices are being developed which have the potential to help overcome barriers.

The following concludes this report with main findings and practical recommendations gathered from the case studies and interviews. These and others will be made available online through the FutureTDM Knowledge Hub[113].

Education and Skill

The main insights that this deliverable has provided is the existing lack of awareness about text and data mining in general. To help solve this education aimed at improving awareness and skills is required. To stimulate community building and uptake more attention should be given to presenting the benefits of TDM.

With respect to education: collaboration between academia, industry, content providers and publishers must be encouraged to develop educational content covering different levels, disciplines and types of research.

Course development must take into consideration, and provide for the different levels of skill needed within disciplines and projects. It depends on the requirements of each project but researchers should be able to develop at least a basic understanding of

- what tools are most effective,
- how to gather data and data management, and
- how to comply with applicable regulations and codes of conduct.

The importance of research data sharing and publishing of research results as well as underlying data in compliance with regulatory framework must be emphasized. This can be done through education and by providing more examples to help illustrate best practices.

In order to improve the lack of skilled personnel within industry, it is not only the responsibility of academia to provided courses and qualified people, industry and institutions can invest in upskilling personnel with basic TDM skills by providing trainings.

[113] www.futuretdm.eu

Best practice/methodologies for content producers and content providers

Improve access

- To improve availability and access of data: universities and research institutions may consider setting up their own open access journals for publishing research data.
- Data presented in scientific papers should be made available for others for reproducibility purposes. Not publishing supplemental data can cause a negative spiral effect on the willingness of others to also share their data.
- Publishers have an important role to actively make sure supplemental research data is available for TDM and take action if it's not.[114]

Improve clarity and use

- Make it clear how to differentiate between academic and commercial research.
- If data providers require that the copied data must be deleted by the researcher after the TDM project has ended, they should provide a guarantee that it will still be possible to reproduce the results by keeping the data available.
- Raise awareness and take part in educating your community on the importance of data sharing in science. The Human genome project is a good example of the benefits of data sharing for science.
- Tenured researchers have an ideal position to promote TDM by taking a principled stand and promote open access and data sharing.

Best practice/methodologies for practitioners

Educate yourself

- Academia and industry are advised to work together to raise awareness and fill the current demand for TDM practitioners. There is a need for development of courses on basic TDM skills as well as more advanced courses.
- For TDM practitioners: educate yourself and learn at least a few basic commands from the terminal. These are often sufficient for basic TDM projects and a less expensive alternative to available more advanced tools and services. Recommended options for self learning include online courses and MOOCs.
- Be able to identify and critically assess whether tools and services provided are effective for your specific project.
- For researchers, it is advised not to use 'black box' proprietary solutions that do not allow you to look at what goes on in the TDM process and what data is being used and how. A better alternative is to choose open source alternatives. The benefits for researchers who contribute to the development of open source is that it gives more visibility as a researcher which can lead to collaborations and acknowledgement of your work.

[114] At the moment, it is not stimulated without negative repercussions. Authors may become reviewers of your next paper or grant proposal so there are conflicting interests which is why there was a request for an independent authority. COPE is a forum for editors and publishers of peer reviewed journals to discuss all aspects of publication ethics. It also advises editors on how to handle cases of research and publication misconduct.

- Obtain the right skillset to understand the data and to become familiar with the data so you know when something has gone wrong and don't make wrongful assumptions. Understand the importance of the quality of data collection. Analysis based on sloppy data should be avoided.
- Practice constraint in data collection. Having a clear project description and being able to communicate about the possible outcome of the TDM activities will help to identify what data is needed and whether it can be limited to only the data relevant for the project.

Educate others

- Be able to communicate TDM results in an effective manner. Visualization skills are necessary to be able to communicate and transfer the knowledge for the audience to understand.
- Help Increase the level of knowledge of those in decisionmaking positions by sharing and promoting good examples and TDM use cases. For management, directors and trustees to provide more support for TDM projects they need to have a better understanding of TDM.
- Show how data driven processes will help institutions and companies become more sustainable, increase revenue and reach their target groups.

Best practice/methodologies for tool and service providers

Educate others

- Academia and industry should work together to raise awareness and to help fill the current demand for TDM practitioners. Industry to promote and facilitate educational programs by being more involved, providing resources and clarity about career opportunities.
- Avoid the use of black box solutions that do not allow you to look at what goes on and what data is being used.
- Although the skill gap is a barrier this can often be solved by upskilling existing personal instead of hiring new staff. There are resources available to increase knowledge to a level of basic understanding sufficient to do TDM.

Legal and Content

With respect to the proposed Copyright exception, stakeholders report a need for more legal clarity. There is however no consensus amongst our participants on legal classification of TDM practices and the use of the results of mining practices.[115] The amount of time and resources spent on getting access and consent and how to provide attribution are considered important barriers experienced by practitioners. As the choice, what research and what data to use depends on access barriers that limit access must be overcome. For example, researchers mentioned avoiding licensed or copyright protected material, which some say has led to biased and unreliable outcomes.

Another concern repeatedly mentioned both by researchers and companies is uncertainty with respect to privacy, data protection and data sharing. There is a clear need for both industry and academic researchers to have more clarity on how to comply with data protection regulation.

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

Best practice/methodologies for content producers and content providers

- Avoid limitations on access and (re)-use of content for non-commercial use. Researchers are often involved in projects that include commercial partners making it difficult to determine the commercial character of the project. Also, excluding non-commercial use limits the potential of project results to be used and further developed commercially.
- Content made available as Open Access has the least restrictions for those who want to mine publishers content
- Choosing 'open' licenses such as CC-BY or CC0 are recommended because these provide the least restrictions and barriers for TDM.
- When implementing a license, use those that already exist and proven usefull instead of developing one to ensure legal interoperability.

Best practice/methodologies for practitioners

- Already in the proposal stage, think about how the corpus data and research outputs will be managed, whether you may want to have a copy of the corpus data for future research purposes. Define what actions will be taken with respect to the data provided, gathered and developed not only throughout the project but also after the TDM activities have ended. Consult for legal advice about your intended TDM project whether the content, practices and intended use comply with the regional regulatory framework. Make sure you have all necessary permissions and rights clearance in place as soon as possible but preferable before undertaking the research. In collaborative projects make sure all partners understand and agree on the project and TDM involved including the use, sharing and archiving of the results and corpus after the project has finished.
- After the project has ended and the corpus data is no longer needed delete the data. Make sure to have sustainable links to where the underlying corpus can be accessed for reproducibility. Copies made of content need to be stored securely and deleted after the project has ended unless the data would be lost otherwise.
- Comply with legal requirements and community code for data collection and storage. This includes not to collect more data than needed (data limitation) when there are privacy concerns.
- When datasets are not available without permission, develop a proof of concept based on data that is available and use this to negotiate access to unavailable datasets
- Comply with legal requirements and community standards for quotation and contribution.
- Do not use the results of text mining for illegal and/or unethical purposes.
- Subscription agreements to content should adopt a TDM clause that will make it clear that TDM is allowed under the agreement. For existing/running agreements these should be renegotiated. If there is a clause that allows TDM or the rights owner has signed up to the STM agreement there should be a way to police this.

Best practice/methodologies for tool and service providers

- The use of open licenses with the least restrictions possible for software, services and tools is recommended unless there are valid (commercial) reasons not to. It is recommended over proprietary or 'black box' solutions for practitioners, because it allows them to understand the TDM process and improve or make adaptations to better serve the project's requirements

Technical and Infrastructure

Interoperability should be improved because mining is not enabled when content comes from mixed sources where the data is not adequately structured and there is no clarity over the licenses used. Therefore, having a standard API across all platforms would make TDM less time consuming for researchers and industry to gain lawful access in a quick and reliable way.

Most stakeholders agree that there should be a combination of directed efforts on making tools easier to use while at the same time help people find and use the tools and services available. Similarly, more effort should go into improving data management skills to ensure data is made available and ready for TDM.

Most of the service providers and tool developers are confident that in a matter of time the tools will improve and better applications will become available. Their main concern is with the availability of good quality data. Others however take this as a business opportunity.

The following is a compendium of proposed solutions to overcome to the aforementioned barriers

Best practice/methodologies for content producers and content providers

- Adopt a standard API across all platforms.
- Structure the data from the beginning to avoid errors when having to extract data.
- It is better to comply with standards in the data collection stage, and not to post-process and adapt the data to a certain standard afterwards.
- When data is digitized or 'born digital' it should be in a TDM friendly format such as XML files.
- Make content available for TDM in an agreed upon format. Make sure that the necessary metadata is included and then delivered to Crossref or an alternative portal where practitioners can access and download the data through a standard API. Again, for content providers make sure the content is well prepared including consistency in licensing and homogenous structure following community wide adopted standards on version formats. If not all the data is made available for mining but only a subset make this very clear to the users.
- Promote data sharing as this helps move science forward. Examples of best practices in making large data sets available include DNA sequencing and the string database. When sharing data, make the complete set of data available. Only sharing a parsed dataset or minimum amount obstructs others who want to use the data either for their own research or to validate the data and check for errors. Another best practice to help increase the availability of data for mining is to comply with existing best practices for data archiving to make sure the data is useful for others.[116] Providing the data to various repositories helps increase findability.
- Comply with Gold Open Access to make content discoverable, interoperable and as useable as possible
- Store data in approved data repositories such as Uniprot. This is also important for small projects with novel datasets so the data is made available for continuity.
- Help develop and adopt solutions for large file data storage and awareness on how and where to make this available. For example, microscopic images.

[116] To address this issue the COS have developed specific guidelines on Transparency and Openness Promotion (TOP)

- Help develop and adopt a protocol for continuity of datasets being accessible when the data producers have for example changed their affiliation[117]. There is a need for solution where the data underlying research can be stored. At the moment journals do not host supplemental data so people are forced to store this elsewhere which is a problem for archiving and continuity. A proposed best practice for funding agencies is to mandate data storage but also facilitate it. It is not considered to be the role of publishers to maintain servers that can hold all the research accompanied data, however they can provide a data report describing where the data is available and a description of the metadata and use-case.

Best practice/methodologies for TDM practitioners

- Invest in access to high quality data. For those that can pay for services there are companies developing solutions to make TDM easier for example Rightfind XML for mining makes cross publisher journal content available in XML format.
- Keep a copy of the data to use it locally especially for long term projects and reproducibility of the results you need to have control over the dataset as well as know what has gone into the API and comes out. Good academic practice is to safely store copies of the data when in use and delete the data after the project has ended.[118]
- Invest in the development of tools. Best practice for mining is to use your own code and developed API unless there is an API available that allows you to understand what you are doing. [119]
- Open source solutions are good if you have the skillset because it's not a black box and its flexible.[120] The R programming language is one of the most used and recommended software tools it is open source and provides reliable results but you can make mistakes when you do not have enough knowledge.[121]
- For those with no programming skills tools like SPS are recommended, but preferably people should master basic commands and for more advanced mining learn to code.
- It is important to have knowledge but also to get the right advice about TDM tools and services taking into consideration the context. Knowledge about and recommendations on technical solutions are necessary to be able to evaluate what will work in specific situations. Knowledge transfer within the community to get advice on what tools and services are good, when tending for solutions that take into considerations individual environments.

[117] Universities will often delete the dedicated webpage and data along with it.

[118] This would alleviate the fear amongst content providers that their data may be shared without permission. See Transparency and Openness Promotion (TOP) Guidelines.

[119] When you have developed your own tools, you have more control over how it works and don't have to learn to understand how the provided API functions. It also gives you the freedom to change it when necessary which is.

[120] You need to know what is going on and be able to see things in the data which proprietary software does not let you do.

[121] R is the top listed on TDM website KDNuggets. On who uses R including Disney company see https://www.fastcompany.com/3030063/why-the-r-programming-language-is-good-for-business

Best practice/methodologies TDM service and tool developers

- Aim to improve reliability of tools but as is mentioned there will always be a level of error, so at the same time provide expectation management of what can be achieved. Develop tools and services that meet the expectations and needs of the different stakeholders. Solutions should be findable, accessible, interoperable, and reusable. Especially the documentation on how to use the software is very important and often lacking at the moment. Provide easy to find and understand documentation of software.
- Provide the right advice about TDM tools and services taking into consideration the context. Important to have knowledge to understand what legacy system and stacks are used and to share within the community review and rankings. Provide Open source solutions because it's not a black box and its flexible. You need to know what is going on and be able to see things in the data which proprietary software does not let you do.
- Contribute to community building, such as the open source community, to work together on developing software, making tools interoperable, and improving the tools to address each specific need.

Economy and Incentives

Barriers under this heading are the lack of a single European market, the problems of having multiple languages and a lack of enforcement of non-EU companies.

Developing standards for data quality is seen as useful in theory but in practice given the diversity in projects and requirements, standards are likely to become too complex for compliance. In those areas with existing standards the main issue is improving compliance. It could help to make compliance with standards mandatory e.g. through funding requirements, or strongly incentivized through mechanisms such as rankings.

The following is a compendium of proposed solutions from the stakeholder consultations.

Best practice/methodologies for content producers and content providers

- Make more funding available for infrastructure and data acquisition
- Make sharing data common practice to guarantee reproducibility and avoid waste of resources having to replicate data that already exists but people have no access to.
- Accept research publications as a reward for sharing data.
- Do not withhold data longer than what is considered reasonable in your community. Institutions, funders and governments should push for data sharing and not only mandate on sharing of the research papers based on the data.
- Help alleviate fear from content providers who think that sharing data in a standardized format will interfere with their business models. Sharing in a standardized way does not prevent them from selling content.
- Move away from subscription based publishing to Open Access. The Hybrid model for publishing is not considered sustainable on the long term and should be avoided.

Best practice/methodologies for practitioners

- Drive science forward by publishing data and accept publications and citations as your reward.

- Learn from and share examples of researchers who have contributed their research data and as a result helped others. [122]
- Foster a community of sharing and understand others are also strapped for money and time.
- Help raise awareness on policies on data sharing.

Best practice/methodologies for tool and service providers

- Help make funding available for academic research on TDM to address domain specific barriers.
- Help improve knowledge and awareness to avoid the expensive spiral of tech. Many companies will push for the silver bullet and offer solutions that are not good enough.
- Provide realistic expectations about tools and services. Failure to meet the needs from users may result in TDM receiving a bad reputation and reduce adoption.
- Contribute to research by sharing data and be transparent about the use of data. This may help alleviate concerns about use of TDM by industry. And the felt discrepancy when commercial users can have access to the available research data but do not share any of their data in return.

6.2 To Conclude

The purpose of this report was to gain a better understanding of the lack of uptake of TDM in Europe through stakeholder consultations and case study analysis. The questions we aimed to cover were; *Which barriers and practices exist? What works well in these practices and what challenges still need to be addressed?*

We have succeeded in identifying at least the main barriers with respect to Technology and Infrastructure, Legal and Content, Economy and Incentives and Education and Skill. Identifying what are the solutions focussing on best practices and methodologies proved more of a challenge. What became clear from conducting the interviews is that mostly people did not acknowledge or agree on any of the practices as best practices. This can partly be explained by the identified lack of TDM awareness, knowledge and skill. We will take this as one of the focus points in our roadmap for TDM uptake.[123]

Although we have not been able to identify best practices as such our findings do show that there are levels of agreement about potential solutions and practices which could further developed to improve responsible text and datamining.[124] Moving forward it is important to also recognise the willingness amongst the different stakeholder communities to come together to discuss and develop solutions that will help overcome the barriers and take into consideration their differences in interest.

[122] Good examples include the Yeast strains research led by Cadillac and the Holstege group. http://www.princessmaximacenter.com/research/research/our-research-groups/holstege-group/

[123] The roadmap will be developed and published online as FutureTDM D5.4 Roadmap for increasing uptake of TDM

[124] We have included a proposal of what constitutes as responsible TDM for discussion in Annex 1

About the use of the compendium

We have compiled this compendium of practices and methodologies based on self identified successful practices and present them for discussion. When considering these practices for adoption we advise users to first understand the context in which these practices have been developed and will take place. This includes paying attention to available resources, feasibility and implementation that are specific to each project and environment and may affect whether the practice can and should be adopted in this particular situation.

We present this document and the FutureTDM online Hub which will include these and a growing list of best practices and methodologies as a discussion document and invite the TDM community to respond and contribute practices to help overcome barriers and improve uptake of TDM.

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,

Altman, Micah, and Gary King. "A proposed standard for the scholarly citation of quantitative data." D-lib 13, no. 3 (2007):

Borgman, C.L., 2015. Big data, little data, no data: Scholarship in the networked world. Mit Press.

Borgman, C.L., 2010. Scholarship in the digital age: Information, infrastructure, and the Internet. MIT press.

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)

Coalition for a Digital Economy report A GLOBAL BRITAIN: From local startups to international markets Tech and digital policy for skills, investment & trade available http://coadec.com/Coadec-Report-A-Global-Britain.pdf

Dimitrova, V., Open Research Data in Economics. Issues in Open Research Data, p.141.2014.

Digital Science White Paper: A New 'Research Data Mechanics' 10th August 2016

ECORYS UK, DIGITAL SKILLS for the UK ECONOMY, JANUARY 2016 https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/492889/DCMSDigitalSkillsReportJan2016.pdf

https://talent.balderton.com/European_Tech_Talent_Landscape.pdf

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 in Europe, Available at: http://project.futuretdm.eu

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

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

Green, Ben, Gabe Cunningham, Ariel Ekblaw, Paul Kominers, Andrew Linzer, and Susan Crawford. Open Data Privacy (2017). Berkman Klein Center Research Publication. Available at DASH: http://nrs.harvard.edu/urn-3:HUL.InstRepos:30340010.

Griffith University Sam Searle, Best practice guidelines for researchers: Managing research data and primary materials, eResearch Services, 18 September 2014 (version 1.4)

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.

Katz, D., 2014. Transitive credit as a means to address social and technological concerns stemming from citation and attribution of digital products. Journal of Open Research Software, 2(1).

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,

Manyika, Chui, Brown, Bughin, Dobbs, Roxburgh, Hung Byers. Big data: The next frontier for innovation, competition, and productivity, Report - McKinsey Global Institute - May 2011

Nosek, B.A., Spies, J.R. and Motyl, M., 2012. Scientific utopia II. Restructuring incentives and practices to promote truth over publishability. Perspectives on Psychological Science, 7(6), pp.615-631.

Outsell Inc., market report, *"Text and Data Mining: Technologies Under Construction."*

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

Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, Introduction to Data Mining, Person Addison Wesley, 2006.

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

Secker, Morrison, Stewart & Horton, To boldly go… the librarian's role in text and data mining, 19 September 2016 - https://www.cilip.org.uk/blog/boldly-go-librarians-role-text-data-mining

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.

STM principles for article sharing, available online http://www.stm-assoc.org/2015_06_08_Voluntary_principles_for_article_sharing_on_scholarly_collaboration_networks.pdf

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.

ANNEX 1 DISCUSSION: RESPONSIBLE CONTENT MINING CODE[125]

1.Don't break the law

(a) Honour copyright as you understand it and consult about current interpretations in your jurisdiction.

(b) If there is no copyright exemption for content mining in your country, consult your institutional librarians for the terms of your licensing contracts with publishers that do not explicitly permit mining.

(c) Be aware of additional legal permissions required for mining with intended commercial use of results.

2. Don't break servers or services

(a) Set acceptable delays between each crawl.

(b) Try not to recrawl and use public repositories of crawled or submitted materials where they exist and allow this.

(c) Avoid corrupting content in the crawling process.

3.Be visible and polite

(a) Use a defined user-agent string in all HTTP requests that clearly identifies you as a crawler and provide contact details.

(b) If you are using any subscription material inform your library and the publisher of your proposed crawling.

4. Work with other content miners

(a) Consult publicly online about current good practice before starting.

(b) Use de facto standard tools (only write your own if there's a gap).

5. Give credit where credit is due

(a) Credit original producers of mined research outputs whenever possible, as per community norms for the reuse of scholarship

[125] Responsible Content Mining, Haeussler M, Molloy J, Murray-Rust P and Oppenheim C, June 16, 2015 accessed online at https://contentmining.files.wordpress.com/2015/06/responsible-content-mining-1.pdf

ANNEX 2 INTERVIEWS

The following are quotes and comments taken from the interviews. They have been colour coded on keywords and personal data has been removed but no more editing has been done to keep the original voice of the message.

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. We need this 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 education 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 works.
- 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.[126]
- 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.

[126] 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.

- 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.
- 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 kinds 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.