Deliverable D3.3+ Baseline report of policies and barriers of TDM in Europe

(extended version of D3.3)

Caspers, M.; Guibault, L.; McNeice, K.; Piperidis, S.; Pouli, K.; Eskevisch, M.; Gavriilidou, M.

Publication date
2017

Document Version
Final published version

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 D3.3+

Baseline report of policies and barriers of TDM in Europe

(extended version of D3.3)

Project

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

Consortium:

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

Stichting LIBER, The Netherlands, (LIBER)

Open Knowledge, UK, (OK/CM)

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

The British Library Board, UK, (BL)

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

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

Ubiquity Press Limited, UK, (UP)

Fundacja Projekt: Polska, Poland, (FPP)

Deliverable

Number:	D3.3+
Title:	**Baseline report of policies and barriers of TDM in Europe (extended version)**
Lead beneficiary:	**UVA**
Work package:	WP3: ASSESS: Studies, Publications, Legal Regulations, Policies and Barriers
Dissemination level:	Public (PU)
Nature:	Report (RE)
Submission date:	04.04.2017

Authors: **Marco Caspers**, UVA
Lucie Guibault, UVA
Kiera McNeice, BL
Stelios Piperidis, ARC
Kanella Pouli, ARC
Maria Eskevich, RU
Maria Gavriilidou, ARC

Contributors: **Jan Strycharz**, FPP
Freyja van den Boom, OK/CM

Review: **Kiera McNeice**, BL
Ben White, BL
Helen Frew, LIBER

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

Preface

We want to express our special gratitude to the national respondents, who have filled out our legal questionnaire on an entirely voluntary basis. Their input has been essential and very valuable for our comparative analysis in the context of identifying legal barriers to TDM in Europe. The following authors contributed to the project, indicated with the respective country or countries they covered in their reports:

- Teodora Tsenova (*Bulgaria*)
- Tatiana Synodinou (*Cyprus* and *Greece*)
- Matěj Myška, Michal Koščík, Jakub Harašta, Lucie Straková, Jakub Míšek, and Pavel Loutocký (*Czech Republic*),
- Aleksei Kelli (*Estonia*),
- Christophe Geiger and Oleksandr Bulayenko (*France*),
- John Weitzmann (*Germany*),
- Anikó Grad-Gyenge and Andrea Klára Soós (*Hungary*),
- Thomas Margoni, Giulia Dore (*Italy*)
- Rán Tryggvadóttir (*Iceland*),
- Astri Margaret Lund (*Norway*),
- Katarzyna Strycharz and Adam Karpiński (*Poland*),
- Martin Husovec and Linda Lubyová (*Slovakia*),
- Maja Lubarda and Andrej Tomšič (*Slovenia*),
- Stavroula Karapapa (*United Kingdom*).

Table of Contents

LIST OF FIGURES

LIST OF TABLES

1 AIM

The overall aim of the FutureTDM project is to improve the uptake of text and data mining (TDM) in the European Union. It is essential to map the barriers that limit the uptake of TDM, in order to determine what actions stakeholders in Europe need to undertake to overcome these barriers and contribute to an environment that promotes TDM. Therefore, this deliverable identifies the barriers to TDM in Europe, covering many aspects and dimensions in which TDM is hindered. Two basic categories of barriers are distinguished in this regard:

- *Legal and policy barriers*: Barriers related to (legal) regulation and stakeholder policies.
- *Practical barriers*: Barriers relating to (lack of) skills, education, technical issues, funding or business environment.

The findings regarding the first category are the result of an extensive research into legal regulation, and in particular intellectual property regimes and data protection law, as well as stakeholder policies dealing with legal rights and obligations ensuing from these legal areas. The results mainly flow from the research tasks of WP3. In addition, the findings from the practical barriers reflect the activities across all work packages, consisting of stakeholder engagement through workshops and interviews, other informal meetings with stakeholders, research into the technical and application landscape of TDM, and economic research.

This deliverable is an extended version of Deliverable D3.3, which only covered legal and policy barriers. Now that we have been able to identify all sorts of barriers throughout all tasks within the FutureTDM project, we have updated it to provide a full overview of barriers that hinder the uptake of TDM in the EU.

2 STRUCTURE

This report is structured in the following way. It consists of four parts, of which Part I familiarises the reader with our understanding of the concept of TDM, as well as the benchmark that we use to group barriers according to their nature. Parts II and III constitute the core of the report, mapping and explaining the different barriers identified throughout the project. Part II addresses the legal and policy barriers, while Part III addresses the practical barriers; both parts conclude with a short overview to summarise the barriers. Part IV provides an overarching overview and conclusion, based on all the findings in the report.

PART I CONCEPT AND BENCHMARK

3 CONCEPT OF TEXT AND DATA MINING

Text and data mining is a very broad concept that is used as an "umbrella term encompass[ing] diverse techniques that allow interpretation of content of any type ranging from raw data, e.g. sensor data, text, images and multimedia, to processed content, e.g. diagrams, charts, tables, references, maps, formulas, chemical structures, and metadata from semi-structured sources, on a large scale through the identification of patterns".[1] TDM can also be referred to as simply text mining or data mining (possibly as subcategories of TDM), or more popular terms such as big data analytics. In our concept of TDM, all these terms are covered, where we generally refer to any activity where computer technology is used to index, analyse, and evaluate and interpret mass quantities of content and data.

To identify the barriers to the uptake of TDM activities, it is important to have an overview of what the whole process of TDM looks like. Note that this may look different for text or data mining, while it may also depend on the purpose for which TDM is carried out. Generally, a distinction can be made between four stages in the TDM process:

- *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. This stage is not necessarily relevant in every TDM activity. For example, a TDM user may also have access to the content through an application programming interface (API), thereby lacking the need (or potential) to build their own database of contents to be mined.
- *Create target dataset*: the miner may need to transform or modify the retrieved content, for example, to another format for standardisation purposes, he might enrich the subject-matter with metadata, or he may only select a part of the content necessary for the analysis. This content is extracted to a new (target) dataset that can be used for analysis in the subsequent stage.
- *Analysis*: the dataset is analysed by means of a computer using mining software, according to an algorithm developed or chosen by the miner. Choice of algorithm may be based on earlier analyses conducted in this stage.
- *Publication*: the TDM user may want to publish the findings from the TDM research. This may for example be in the form of an online or paper report, research paper in a journal, newspaper article or weblog, but it could also be circulated only within the closed circle of a company in order to base decisions on. Also, a TDM user may want to publish their data(sets) aggregated from the mined contents or data It all depends on the purpose of and the context in which TDM is carried out, as well as common practice within the field of research or industry.

[1] FutureTDM Deliverable D3.1, p.6.

The TDM process is graphically represented, possibly somewhat simplistically, in Figure 1. Based on this model, an assessment can be made of what acts, if any, in the TDM model are unlawful or inhibited by practical barriers and to what extent.

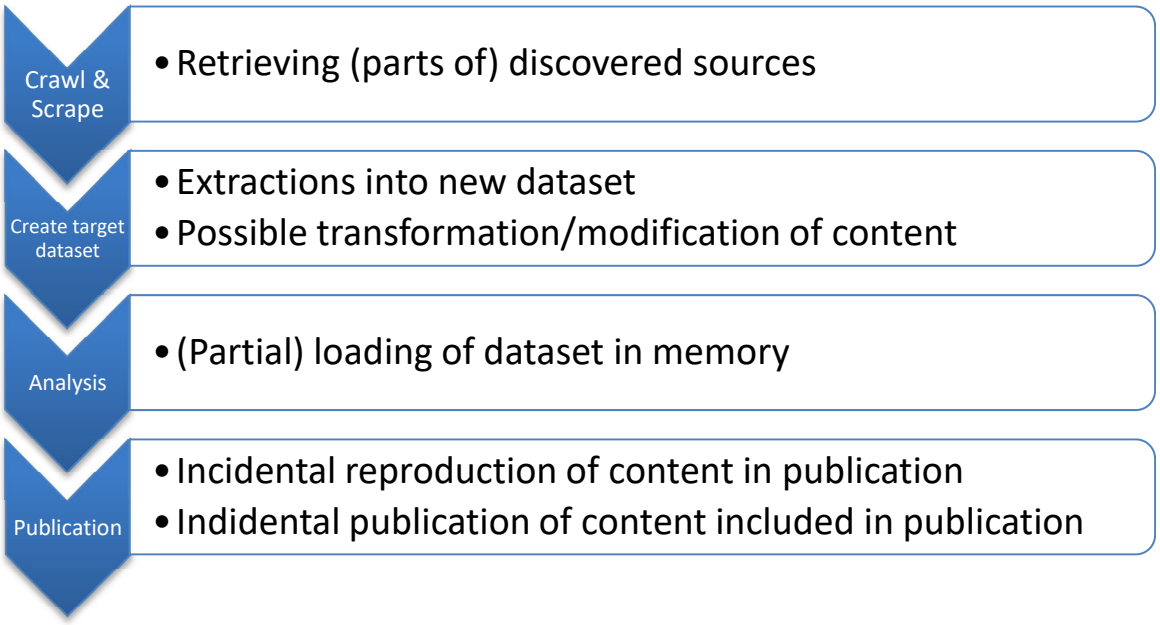

Figure 1 – Model of acts carried out in the TDM process

4 BENCHMARK

To better understand the nature, form and extent of the barriers affecting TDM activities in Europe, we have developed a benchmark test that allows us to qualify our findings in a manner that will lead, at a later stage, to clearer policy recommendations. These criteria allow us to compare the information gathered in a systematic way and to highlight the largest obstacles to TDM uptake. They are represented in Figure 2.

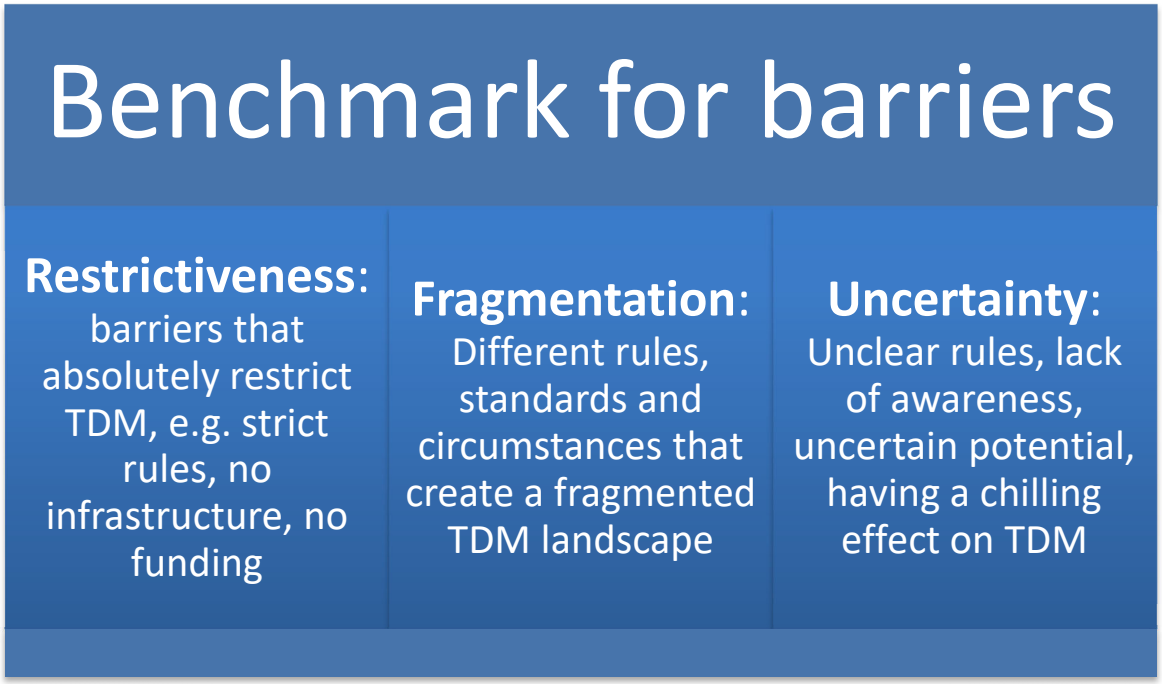

Figure 2 – Categories of barriers

The identified categories can be explained as follows:

- *Restrictiveness*: concerns the barriers that restrict TDM in an absolute way. For example, legal rules and criteria might in themselves restrict (parts of) mining activities, or only permit TDM under certain conditions. Another example would be the lack of funding or availability of data analysis tools.
- *Fragmentation*: relates to any differences and anomalies observable, which create a fragmented environment for TDM, preventing, inter alia, cross-border or cross-domain collaboration or sharing of data and tools among academic or industry sectors. For example, such differences may exist between
    - legal rules and their interpretations in different countries,
    - different policies, applied by different stakeholders,
    - gaps between TDM skills required by industry and skills taught in (higher) education,
    - standards used to format data, or
    - interoperability of infrastructure.

- *Uncertainty*: refers to uncertainty and lack of clarity and awareness, which has a chilling effect on potential TDM users' ability to actually carry out TDM. For example, this category may include uncertainty about applicable legal rules, lack of clear documentation and manuals for TDM tools, or awareness of the actual potential benefits of TDM for a research or a business.

This benchmark is used to group all identified barriers in this report into these three categories. This serves to distinguish and acknowledge the different nature of the barriers, as well as to find commonalities in nature of barriers across the different domains in which barriers are identified. This serves to support the development of the policy framework in deliverable D5.1, which provides high-level principles and recommendations to overcome the most important barriers identified in this report.[2] Therefore, at the end of each section of **Parts II** and **III**, an overview and summary is provided of the barriers identified in the concerned domain, where they are grouped in accordance with our benchmark.

[2] FutureTDM Deliverable D5.1.

PART II LEGAL AND POLICY BARRIERS

5 INTRODUCTION

The purpose of Part II is to identify the barriers to text and data mining (TDM) in Europe that relate to regulations that render TDM activities unlawful and policies that restrict these activities. As is illustrated by the following pyramid in Figure 3, barriers may exist on several layers.

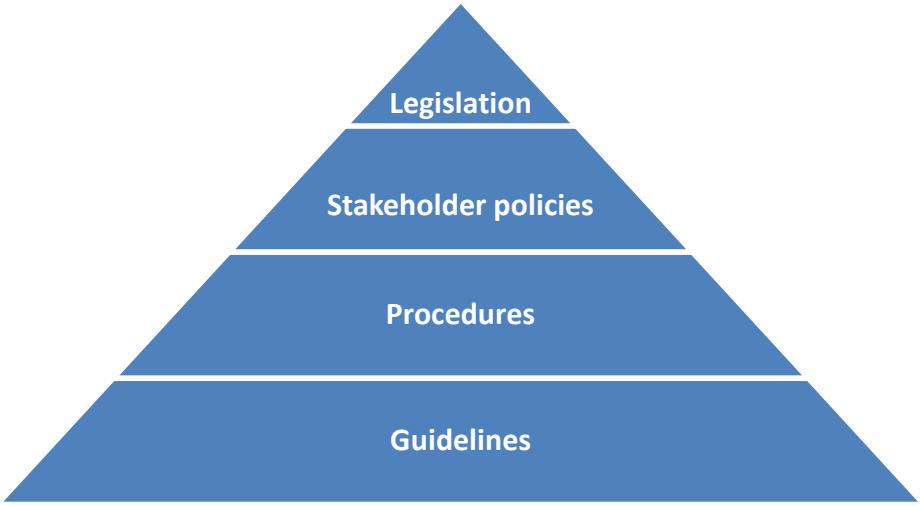

Figure 3 - Pyramid of TDM barriers

Legislation refers to the laws that affect the uptake of TDM, for example by prohibiting certain acts carried out in TDM or due to uncertainty on their permissiveness towards TDM. *Stakeholder policies* may either restrict or permit TDM activities; restrictions may relate to legislation, e.g. when a content provider holds the copyrights in its contents or when individuals do not consent to the processing of their personal data, but they can also be merely based on contracts or unilateral actions; therefore, stakeholder policies are below legislation in the pyramid, covering a broader range of acts. Such policies may also encourage TDM by stimulating the lawfulness of TDM, such as associations of stakeholders pushing the for the authorisation of TDM acts or funders of research requiring that output is made available without restrictions to re-use. It is these first two types of barriers that are addressed in this deliverable.

Procedures, which relate more to the practical aspects such as interoperability, skills and education, may also either restrict or promote TDM; they are below the stakeholder policies in the pyramid, as they may cover a range of aspects that can be broader than legislation and stakeholder policies. *Guidelines*, which are rather concerned with providing clarity and certainty on how to carry out TDM and what to take into consideration (such as legal and technical aspects), may be used to instruct e.g. users and other stakeholders, on how to overcome barriers of any kind; they are the lowest in the pyramid, as they may cover issues that go beyond, but may include, those addressed in legislation, policies and procedures.

To address the top two levels of the pyramid, this part consists of a *legal* part and a *policy* part. Legal barriers are identified both on the European level as well as on the national level. We use the benchmark test developed in Section 4 to assess, in a systematic and methodical way, how the regulations and policies affect TDM activities. In Sections 0, 8 and 9, we identify the barriers in the context of copyright law, database law and data protection law, while we look at recently adopted or proposed TDM exceptions in the EU in Section 10, after which Section 11 provides an overview of TDM regulation outside the EU. Subsequently, Section 12 concludes part II with an assessment of stakeholder policies as to the extent they stimulate or hamper TDM activities.

6 METHOD FOR STYDING LEGAL BARRIERS IN EUROPE

To identify legal barriers, our preliminary research was guided by the question: *Which regulations in the EU and its Member States serve as legal barriers to TDM activities and which allow or promote TDM?* This has led us to identify the regimes of copyright law, database law and data protection law, as the acts identified in our model of TDM are likely to trigger the application of those regimes. In particular, the - permanent and temporary - reproductions of content made in the process or as a result of TDM activities are likely to affect these areas of law, as well as the publication of the research results, where (parts of) the mined data may be included.

6.1 Questionnaire

This report covers legal barriers in both EU law and national laws. To gather input on the national legal barriers, a questionnaire was designed, based on the findings of our preliminary research, and sent to legal experts in all 28 Member States of the EU and in the three additional countries of the European Economic Area (EEA). Input was received from sixteen countries, of which the reports for the Netherlands and Poland were filled out by partners in the FutureTDM project.[3]

The design of the questionnaire was two-fold. First, we sought information on legal *barriers* in relation to the legal regimes identified as potentially restricting TDM. We formulated questions as to how national law restrict TDM or certain acts in the TDM process, more specifically, how the European rules of copyright law[4] and database law[5] are implemented at national level and how they relate to the acts carried out in the course of TDM. We also enquired about the national rules on

[3] Bulgaria (Teodora Tsenova), Cyprus & Greece (Tatiana Synodinou), Czech Republic (Matěj Myška, Michal Koščík, Jakub Harašta, Lucie Straková, Jakub Míšek, and Pavel Loutocký), Estonia (Aleksei Kelli), France (Christophe Geiger and Oleksandr Bulayenko), Germany (John Weitzmann), Hungary (Anikó Grad-Gyenge and Andrea Klára Soós), Iceland (Rán Tryggvadóttir), Italy (Thomas Margoni, Giulia Dore), Norway (Astri Margaret Lund), Poland (Katarzyna Strycharz and Adam Karpiński), Slovakia (Martin Husovec and Linda Lubyová), Slovenia (Maja Lubarda and Andrej Tomšič), United Kingdom (Stavroula Karapapa).

[4] Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society
OJ L 167, 22.6.2001 (*Copyright Directive*), p. 10–19.

[5] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases OJ L 77, 27.3.1996, p. 20–28.

data protection and their application to TDM activities, rules which have been recently further harmonized by the General Data Protection Regulation 2016/679.[6]

In relation to the policy barriers, the questionnaire also included questions as to contractual practice, industry codes of conduct and self-regulation, mostly connected with the legal regimes of copyright, database and data protection law. We also asked for examples of government policies and examples regarding the re-use of public sector information (PSI). In this context, the PSI framework was not in itself regarded as a regulation potentially serving as a legal barrier; it was rather the practical implementation and use by national governments that was considered to potentially impede TDM: providing access and enabling re-use under certain (licence) conditions is rather a matter of policy.

At the outset, it is important to note that, due to lack of budget, the national correspondents were asked to respond to the questionnaires on a voluntary basis. Therefore, they may have invested less time and effort on this task than they would otherwise have done under paid circumstances.[7] Second, a response was not received for all Member States, despite the search for other experts in certain countries and despite multiple reminders. Nevertheless, a sample of sixteen countries is sufficient for an exploration of national legal barriers to TDM and to indicate the most problematic aspects in this context. Third, as only one respondent in each country provided answers, certain aspects of the law that are controversial or uncertain could have been reported differently by another expert. Moreover, some respondents may not necessarily be expert in all fields of law covered by the questionnaire. This bias is partly overcome by the fact that for some national reports, experts from the respective fields have collaborated with others to accurately report on the law in their country.

6.2 Benchmark for legal barriers

To better understand the nature, form and extent of the legal barriers affecting TDM activities in Europe, we use the benchmark of Section 4 to qualify our findings in a manner that will lead, at a later stage, to clearer policy recommendations. The information contained in the national answers to the questionnaire is therefore analysed in the light of the three main criteria so as to provide an indication of the lawfulness of TDM activities not only in each jurisdiction, but also across the European Union. These criteria allow us to compare the information gathered in a systematic way and to highlight the biggest legal obstacles to TDM.

In the context of legal barriers, the benchmark can be explained as follows:

- *Restrictiveness*: concerns the legal rules and criteria that in themselves restrict (parts of) mining activities, or that only restrict or permit TDM under certain conditions.

[6] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance), *OJ L 119, 4.5.2016* (*GDPR*)*, p. 1–88*

[7] Note that some reports were of an exceptionally high standard that one would not necessarily expect in this context.

- *Fragmentation*: relates to differences and anomalies observable between national laws and interpretations regarding specific legal concepts, as well as existing differences and anomalies between legal regimes, e.g. between copyright and database law.

- *Uncertainty*: refers to rules, criteria or concepts in the laws that are not clear, for example, because the concept is vaguely defined and has never been interpreted by the legislator or the courts, or because there is no consensus as to the interpretation in the field.

For each field, we examine the general framework as to how it generally affects the activities that are generally carried out in the course of TDM, followed by a discussion on several exceptions or concepts that deviate from the general rules, thereby potentially permitting TDM under certain conditions.

6.3 TDM in foreign regulation

As data-driven innovation and research, which are enabled through TDM activities, are global endeavours, the section's findings on the legal barriers to TDM activities must be put in a broader context. We therefore examine how Europe's main trading partners deal with TDM issues in their respective intellectual property regimes. To this end, we take a brief look at the copyright laws of the United States, China, Australia, Canada, and Japan to see whether TDM activities are allowed to take place and if so, on what grounds and under what conditions. It is important to note, that none of the countries examined have enacted an intellectual property regime that is comparable to the European Database Directive. Among the countries studied here, only Japan offers extra protection against the misappropriation of databases by competitors. The legal regime relevant for TDM activities outside Europe is copyright law.

7 COPYRIGHT AND *SUI GENERIS* DATABASE LAW

The mining of collections has the potential to affect two intellectual property regimes: copyright law and sui generis database law. Copyright may exist in the individual works that are part of mined collections, but the collection as a whole may also constitute a *work* protected under copyright law. The latter exists independently of any copyright in the contents of the collection. Sui generis database rights can only exist in the collections, regardless of any copyright in their contents or any copyright protection on the collections themselves. In this section, we will analyse protected subject-matters under these regimes, as well as the exclusive rights granted to the right holders and the exceptions to those rights. This will be done from both a European perspective, discussing the harmonised aspects of copyright that are relevant to TDM, and from a national perspective, examining the national exceptions in the context of their European counterparts.

This section first covers the protected subject-matters under copyright and sui generis database law. We subsequently elaborate on the exclusive rights that are granted to the beneficiaries of those intellectual property regimes and the extent to which acts carried in TDM fall within their scope. This will be followed by an analysis of the exceptions in both copyright and database law, as provided for by EU law and as implemented in national laws. The national reports form the basis as regards the barriers in national laws. It must be noted that the overview is in no way exhaustive as not all reports have covered all aspects. The mapping of legal barriers must therefore be regarded as being explorative, which is useful in order to indicate the most problematic bottlenecks that impede the uptake of TDM activities.

7.1 Protected subject matter

7.1.1 Copyright law: *works*

7.1.1.1 Works in general

Copyright law gives the *author* of a *work* several rights to, inter alia, enable him to exploit the work. Which rights are provided to the author and what is to be considered a work are traditionally highly national matters. However, many harmonisation efforts have been made both on the international level and at EU level to provide a minimum level of protection across borders. *What* is protected by copyright, is minimally harmonised by the Berne Convention (BC), which currently has 171 contracting parties:[8] "literary and artistic works".[9] It is considered to include a broad range of forms of expression, among which are books, pamphlets, choreographic works, dramatico-musical works, musical compositions, cinematographic works, paintings, architecture, sculpture, maps, plans, works of applied arts, and so forth.[10] Even a collection of those works may constitute an intellectual creation "by reason of the selection and arrangement of [its] contents" and therefore be protected as a work under copyright.[11] Similar wording is used by the Agreement on Trade-Related Aspects of

[8] http://www.wipo.int/treaties/en/ShowResults.jsp?lang=en&treaty_id=15 (accessed on 31 May 2016).
[9] Article 1(1) of the Berne Convention for the Protection of Literary and Artistic Works (*BC*).
[10] Article 2(1) of the BC.
[11] Article 2(5) of the BC.

Intellectual Property Rights (TRIPS) and the WIPO Copyright Treaty (WCT) – both of which all EU Member States, including the EU itself, are contracting parties to – where they provide that "[c]ompilations of data or other material" in whatever form shall be protected if they "by reason of the selection or arrangement of their contents constitute intellectual creations".[12]

The harmonisation of protected subject-matter in the European Union is characterised by a rather fragmented approach of the European legislator and a highly active role of the Court of Justice of the European Union (CJEU). Initially, the EU copyright framework only harmonised protection for the subject-matters of computer programs, photographs and databases.[13] As a shared concept of originality, they are protected if they constitute "the author's own intellectual creation". Similar provisions in the directives affecting copyright law are lacking for other types of works, even in the comprehensive Copyright Directive (2001/29/EC).[14] Nevertheless, the CJEU has consistently applied the criterion of *author's own intellectual creation* to other works (than software, photographs and databases) in its case law since 2009. In its *Infopaq* decision, the Court ruled that an extraction from a newspaper, consisting of eleven words, falls within the scope of the author's exclusive right of Article 2 to authorise or prohibit the making of (partial) reproductions, if that part constitutes the author's own intellectual creation.[15] Likewise, the CJEU ruled that graphical user interfaces (GUIs), programming languages and data file formats, as well as sound and graphic elements in a videogame are protected subject-matter under the *acquis communautaire*, provided that they constitute the author's own intellectual creation.[16]

What is to be understood by the concept of *author's own intellectual creation* is not defined precisely by the Court, although it has provided some points of reference. The concept of *creativity*, *creative freedom* and *free and creative choices* seems to be crucial in this originality assessment. According to the CJEU, this is found in the "choice, sequence and combination" of words in newspaper articles,[17] the "specific arrangement or configuration" of a GUI's components,[18] or in the case of photographic works this might lie in the choice of - inter alia - background, subject's pose, framing, angle of view or developing technique.[19] The Court has not stated anything about the amount of creativity required for a subject-matter to become an author's own intellectual creation, but it has defined the boundaries of the requisite originality in a negative way: where choices are dictated by technical

[12] See Article 10(2) of the Agreement on Trade-Related Aspects of Intellectual Property Rights (*TRIPS*), and Article 5 of the WIPO Copyright Treaty (*WCT*), respectively.
[13] See, respectively, Article 1(3) of Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs (*Software Directive*), Article 6 of Directive 2006/116/EC of the European Parliament and of the Council of 12 December 2006 on the term of protection of copyright and certain related rights (*Term Directive*), and Article 3(1) of Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases (*Database Directive*).
[14] In this report, Copyright Directive refers to *Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society*.
[15] CJEU 16 July 2009, C-5/08 (*Infopaq*), §§34-36.
[16] See respectively: CJEU 22 December 2010, C-393/09 (*BSA*), §§43-46; CJEU 2 May 2012, C-406/10 (*SAS Institute*), §45; CJEU 23 January 2014, C-355/12 (*Nintendo*), §23.
[17] CJEU Infopaq, §§44-45.
[18] CJEU BSA, §48.
[19] CJEU 1 December 2011, C-145/10 (*Painer*), §91.

function,[20] rules or constraints,[21] the author is not able to "[express] his creative ability in an original manner by making free and creative choices".

The effect of the CJEU's decision does not only imply that there is an autonomous originality test under EU copyright law - where the standard seems rather low, since eleven word extracts are potentially covered by this concept, as appears from the Court's Infopaq decision; it also implies an open-ended work concept that is able to cover any sort of subject-matter that can be considered to constitute the author's own intellectual creation.[22] Consequently, an exhaustive list of types of subject-matter that are protected under a national copyright law would not be consistent with EU copyright law.

In the context of TDM, it is useful to make a distinction between *high-level* and *low-level* data when considering the likelihood of mined contents being protected under copyright law. High-level data would consist of 'rich' contents, such as newspaper or journal articles, books, music, photographs or cinematographic works. Low-level data would consist of more 'raw' data, such as genetic information, measurement data in any field of science, name and address data, GPS coordinates, phone numbers or financial data.

Subject-matter that falls within the category of high-level data is highly likely to constitute an author's own intellectual creation and therefore be subject to copyright law. As a result, any act falling within the scope of the economic rights granted by a copyright law triggers its protection regime. The same applies where the collection as such, "by reason of the selection or arrangement of their contents",[23] constitutes an author's own intellectual creation and is protected as a work of copyright. For those cases, the acts carried out on and with the protected subject-matter in the TDM process need to be assessed to determine whether they fall within the ambit of the author's exclusive right to authorise and prohibit them.

On the contrary, contents that fall within the category of low-level data are highly unlikely to be protected under copyright law. As a principle of copyright law, copyright protection does not exist in isolated facts and, considering the CJEU's case law, not in data which is created from mere technical or functional considerations.

7.1.1.2 Databases as copyrightable works

Moreover, copyright may exist in the collection of either low-level or high-level data, independent of any copyright existing in the contents themselves. The Database Directive (96/9/EC) defines a database as "a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means". It requires a case-by-case analysis to assess whether a certain collection falls within this definition. Triaille et al. argue that, for example, whole text corpora may not be "'arranged in a systematic or methodical way' nor 'individually accessible by electronic or other means'" and therefore may not be protected under copyright law. Whether the contents of a database are regarded as *independent*, depends on whether they retain an autonomous *informative value* after they are extracted from the collection.

[20] CJEU BSA, §48-49.
[21] CJEU 1 March 2012, C-604/10 (*Football Dataco*), §39.
[22] Eechoud (2012), p.68; Rosati (2013), p.65.
[23] Article 3 of the Database Directive.

The CJEU has ruled that this "must be assessed in the light of the value of the information not for a typical user of the collection concerned, but for each third party interested by the extracted material".[24] Following this reasoning, it concluded that geographical information extracted from a map to produce and market another map retained "sufficient informative value" to qualify as "independent materials".[25]

If a collection qualifies as a *database* under the Database Directive, it is only protected when it, "by reason of the selection or arrangement of [its] contents, constitute[s] the author's own intellectual creation".[26] Following the above discussion on the CJEU's interpretation of this criterion, this excludes databases where selection and arrangement are solely based on functional or technical considerations.

7.1.2 Database law: *databases*

In addition to any copyrights, a sui generis database right may exist in a collection or database. Such is only vested in the maker of a database "which shows that there has been qualitatively and/or quantitatively a substantial investment".[27] The investment must be made in the obtaining, verification or presentation of its contents. *Verification* may refer to ensuring that the information in the database is reliable and the monitoring of the accuracy of the collected materials when the database was created, but also during its operation.[28] Investments in the *obtaining* of materials means the use of resources to "seek out existing independent materials and collect them in the database".[29] The investment must be made independently of the resources used to create the materials.[30]

Before the ECJ's rulings on what costs are taken into account to establish whether a substantial investment has been made in a database, a *spin-off theory* had developed in case law in several Member States, but most prominently in The Netherlands.[31] The spin-off theory concerns databases that spun off another activity. For example, a company may produce data from its main activities, where the inclusion thereof in a database would be a by-product.[32] It implies that if a database is used for various purposes, a substantial investment has to be made separately for each purpose. However, in 2002, the Dutch Supreme Court found that neither the Database Directive nor the Dutch Database Act (DDA) provides a basis for such an argument,[33] although it is interpreted as only a rejection of the theory in the given circumstances of that underlying case. Advocate General Stix-Hackl also rejected this theory and stated that the "objective pursued in obtaining the contents of

[24] CJEU 29 October 2015, C-490/14 (*Freistaat Bayern*), §27.
[25] CJEU Freistaat Bayern, §29.
[26] Article 3(1) of the Database Directive.
[27] Article 7(1) of the Database Directive.
[28] ECJ 9 November 2004, C-203/02 (*BHB v. William Hill*), §34.
[29] ECJ 9 November 2004, C-444/02 (*Fixtures Marketing v. OPAP*), §40; ECJ 9 November 2004, C-46/02 (*Fixtures Marketing v. Veikkaus Oy*), §34; ECJ 9 November 2004, C-338/02 (*Fixtures Marketing v. Svenska Spel*), §37.
[30] *BHB v. William Hill*, §35; *Fixtures Marketing v. OPAP*, §40; *Fixtures Marketing v. Veikkaus Oy*, §34; *Fixtures Marketing v. Svenska Spel*, §37.
[31] Annemarie Christiane Beunen, *Protection for Databases. The European Database Directive and Its Effects in the Netherlands, France and the United Kingdom* (Nijmegen: Wolf Legal Publishers, 2007), p. 107.
[32] Beunen, p. 108.
[33] Dutch Supreme Court 22 March 2002, *IERI* 2002, 22 (*NVM v. Telegraaf*), §3.4.1.

the database [is not] of any relevance".[34] This point of view was implicitly confirmed by the CJEU in *British Horseracing Board v. William Hill*, as stated above, finding that the mere fact that the maker of a database is also the creator of its contents does not necessarily preclude him from the sui generis protection; he can still claim that he invested in the collection, verification and presentation of those materials, using resources other than those used to create those materials.[35]

However, borderline cases may exist where the investments in the *creation* of contents coincide with the *collection* thereof, and it is not easy to make a distinction between the costs concerned with the creation of the contents and those related to their collection. It has been proposed that in these circumstances, it would be decisive, in particular, whether the creation of the information is aimed at the creation of a database, including the further expansion and perfection thereof.[36] For example, this would be the case where astronomical measurements and observations are made specifically for the creation of a star catalogue.[37] The creation and collection of data, in such cases, both merge into the primary activity; instead of a spin-off theory, one might speak of a *merge theory* or something similar to that.

Advocate General Stix-Hackl argued that "[a]ll the language versions [of the directive] thus allow of an interpretation according to which […] the protection of the Directive kicks in" if "the creation of data coincides with its collection and screening".[38] It is not clear whether the ECJ would agree with such a point of view. However, the Court has repeatedly emphasised that the Database Directive is intended to "ensure that the person who has taken the initiative and assumed the risk of making a substantial investment in terms of human, technical and/or financial resources in the setting up and operation of a database receives a return on his investment by protecting him against the unauthorised appropriation of the results of that investment".[39] This may ask for a rather broad interpretation, not only of the exclusive rights vested in the maker, but also of the scope of the concept of *investment*. Moreover, it is argued that the fact that the proposal for the directive contained statutory licences where contents could only be retrieved from one source implies that the legislator acknowledged database rights where data were generated by the maker, since it sought to prevent them from monopolising that data.[40]

There are also some counter arguments. First, referring to the last argument above, it is also suggested that the final version of the Database Directive lacks a provision on statutory licences, which thereby suggests that the legislator does *not* acknowledge database rights to be vested in a maker who produces data himself.[41] Second, Recitals 45 and 46 of the Database Directive point out that the sui generis database right "does not in any way constitute an extension of copyright protection to mere facts or data" and that it "should not give rise to the creation of a new right in the

[34] Opinion of Advocate General Stix-Hackl 8 June 2004, C-203/02, §47.
[35] *BHB v. William Hill*, §§36-37.
[36] Spoor, Verkade & Visser (2005), p. 620: „Beslissend lijkt daarom vooral, of het creëren van de informatie gericht is op het tot stand brengen van de databank, inclusief het verder uitbouwen en vervolmaken daarvan."
[37] Spoor, Verkade & Visser (2005), p. 620.
[38] Opinion of Advocate General Stix-Hackl 8 June 2004, C-203/02, §46.
[39] ECJ 19 December 2013, C-202/12 (*Innoweb v. Wegener*), §36; ECJ 9 October 2008, C-304/07 (*Directmedia Publishing*), §33; *Fixtures Marketing v. Veikkaus Oy*, §35; *BHB v. William Hill*, §§32 and 46.
[40] Hugenholtz (2002).
[41] Hugenholtz (2002).

works, data or materials themselves"; conferring a database right on the basis of investments in the creation of data is coming quite close to the creation of such a new right.

The discussion above illustrates the tension that exists between the free flow of information, and the incentive to makers of databases to invest in databases and the monopolisation of information thereby. What if the contents produced are only marketable in the form of a database? This question might become more relevant, for example, in an environment where research has to rely more on private funding. One might think of high-throughput screening techniques to create large databases with information on e.g. biochemical processes. When funded with public money, there is no question of how to recoup the costs of creating such data and their inclusion in a database. However, if such data are produced for commercial purposes, e.g. to let researchers mine the database for interesting insights, marketing them in the form of a database is the only way.

7.2 Exclusive rights

7.2.1 Copyright: reproduction and communication

The European copyright framework has harmonised several exclusive rights to be granted under Member State laws. For works in general, the Copyright Directive harmonised three exclusive rights that are granted to the author: the *reproduction* right, the right of *communication* and *making available to the public*, and the *distribution* right.[42]

7.2.1.1 Reproduction

The reproduction right is formulated broadly as the "exclusive right to authorise or prohibit direct or indirect, temporary or permanent reproduction by any means and in any form, in whole or in part".[43] As discussed above, the CJEU ruled in its Infopaq decision that this right only covers reproductions made of works or parts of works that constitute an author's own intellectual creation.[44] Hence, this originality criterion does not only define the protected subject-matter, but it defines the scope of the reproduction right as well; reproductions of parts of copyright works that do not constitute the author's own intellectual creation do not fall within the scope of the reproduction right.

It is highly likely that many TDM activities fall within the scope of the reproduction right, presuming that the mined contents consist of protected subject-matter; as set out above, this is highly likely in the case of high-level data such as books, newspaper or scientific articles, photos, images, music, or cinematographic works. In the first stage of TDM, identified sources are possibly copied onto the miner's own storage facilities, which would involve the making of reproductions. Subsequently, selection may be made from the collected subject-matter to copy into a new dataset. For the actual analysis that will follow, the executed analysis software required (partial) reproduction to be made in a computer's RAM which may be merely transient and only consist of fragments of works. However, the CJEU has ruled that such fragments are covered by the reproduction right of Article 2 of the Copyright Directive, if they "contain elements which are the expression of the authors' own intellectual creation, and the unit composed of the fragments reproduced simultaneously must be

[42] Articles 2, 3 and 4 of the Copyright Directive.
[43] Article 2 of the Copyright Directive.
[44] CJEU Infopaq, §48.

examined in order to determine whether it contains such elements".[45] If, eventually, TDM results are published, parts of works may be quoted or otherwise included in the publication. In those cases, these reproductions are covered by Article 2 of the Copyright Directive as well, provided that these parts are the author's own intellectual creation.

7.2.1.2 Communication, making available and distribution

According to Article 3(1) of the Copyright Directive, "Member States shall provide authors with the exclusive right to authorise or prohibit any communication to the public of their works, by wire or wireless means, including the making available to the public of their works in such a way that members of the public may access them from a place and at a time individually chosen by them". It follows from the directive's 23rd recital and the CJEU's ruling in *Circul Globus Bucureşti*[46] that this only covers communications to the public where the public is at a place different from where the communication originates. Consequently, exclusive rights to authorise other sorts of communications (to the public), such as public and live performances, remain unharmonised and their existence and interpretation are left at the discretion of the Member States.

As regards the publication of TDM results, this will often take place in the form of an article or report that is either available digitally on the Internet - e.g. through a website where it is either publicly available, or accessible behind a paywall or through a subscription - or on paper. Where such publication includes an original work or part thereof, the making available of that publication will constitute a making available to the public and requires the right holder's authorisation. For the paper version of such a publication, it is the distribution right of the author that will be affected instead. It is harmonised by Article 4(1) of the Copyright Directive and applies, as follows from recital 28, only to physical copies. This exclusive right covers the distribution to the public of an original or copies of a work by sale or otherwise.

7.2.1.3 Exclusive rights in databases

Besides the discussed exclusive rights under general copyright law, the harmonized regime for copyright in databases grants the author of a database a specific set of exclusive rights, although they largely overlap those provided by the Copyright Directive. Article 5 of the Database Directive namely provides the rights to authorise or prohibit:

- the "temporary or permanent reproduction by any means and in any form, in whole or in part" (reproduction right);
- "any form of distribution to the public of the database or of copies thereof" (distribution right);
- "any communication, display or performance to the public" (communication to the public right).

The reproduction right is likely to cover the stages where one or more collections are retrieved in their entirety. Following the Court's reasoning in Infopaq, when parts of databases are reproduced, the reproduction right only cover those parts that are original. Consequently, in later stages where

[45] CJEU 4 October 2011, Joined Cases C-403/08 and C-429/08 (*Premier League*), §159.
[46] CJEU 24 November 2011, C-283/10 (*Circul Globus Bucureşti*).

contents are selected and possibly combined into a new dataset, the original selection and structure might get lost, with the result that "elements on which copyright could have a grasp have become undetectable".[47] This might also apply where databases are not retrieved as a whole and where only certain contents are selected and subsequently retrieved. However, we are not sure whether the Court would interpret reproduction under the Database Directive similarly to its equivalent in the Copyright Directive. Either way, as Triaille et al. have also argued, there is a problem from an evidential point of view:[48] if large amounts of data are collected from many sources, it might be difficult to prove that contents originate from a specific database.

In the typical TDM process, both the distribution and communication to the public rights may only be affected at the stage of publication of any TDM results. This will only be the case where the results contain elements of the selection and arrangement that are considered original, which is, for example, highly unlikely when a small part (of one of the contents) of the databases is quoted or otherwise included in the results.

Where the Copyright Directive lacks the provision for an adaptation right, the Database Directive provides for a right of "translation, adaptation, arrangement and any other alteration" in Article 5(b). It is conceivable that an original database may be adapted in the process of creating the target dataset for the analysis, where, for example, only parts of the contents are selected or where their arrangement has been modified. However, as with the reproduction right, the evidential problem as regards the original source arises.

7.2.2 Database law

According to Article 7(1) of the Database Directive, the maker of a database has the exclusive right to authorise the extraction or re-utilisation of the whole or a substantial - qualitatively or quantitatively assessed - part of the contents of the database. It also includes the repeated and systematic extraction or re-utilisation of non-substantial part, insofar that would "conflict with a normal exploitation of that database or which unreasonably prejudice the legitimate interests of the maker of the database shall not be permitted" (Article 7(5) of the Database Directive). How these exclusive rights are interpreted, is mainly found in the CJEU's decisions.

In *BHB v. William Hill*, the CJEU emphasised Recital 42 of the Database Directive, which states that the objective pursued by the sui generis database right is intended to protect the maker of a database against acts carried out by users that go beyond the legitimate rights and thereby harm the investment. The purpose is to guarantee the maker a return on his investments associated with the creation and maintenance of a database.[49] According to the Court, the purpose of these acts of extraction and re-utilisation is not relevant for the assessment of the scope of the sui generis right.[50] The prohibited acts do not include the consultation of a database, including by third parties who consult a publicly accessible database, whether or not the accessed database concerns the (authorised) re-utilisation by a third party.[51] In the same case, the Court also defined the *lawful user*

[47] Triaille et al. (2014), p.35.
[48] Triaille et al. (2014), p.35.
[49] *BHB v. William Hill*, §46; Recital 48 of the Database Directive.
[50] *BHB v. William Hill*, §§47-48; CJEU 5 March 2009, C-545/07 (*Apis-Hristovich v. Lakorda)*, §46.
[51] *BHB v. William Hill*, §§54-56.

as the user whose access to the contents of a database for consultation purposes follows from a direct or indirect consent of the maker.[52]

7.2.2.1 Right of extraction

The concept of *extraction* refers to "any unauthorised act of appropriation" of a whole database or a part of its contents,[53] meaning that the materials are "placed on a medium other than that of the original database", regardless of whether this is permanent or temporary.[54] If the "physical and technical characteristics" of contents in a protected database are present in another database, this is only evidence of an extraction of contents from the first database, "unless that coincidence can be explained by factors other than a transfer between the two databases concerned".[55]

The extraction right, as well as the re-utilisation right, requires that a substantial part is appropriated. This must be evaluated *quantitatively* and *qualitatively*. Qualitatively refers to "the volume of data extracted from the database and/or re-utilised, and must be assessed in relation to the volume of the contents of the whole [database]".[56] *Qualitatively* refers to the "scale of the investment in the obtaining, verification or presentation of the contents" that are re-utilised, irrespective of whether it forms a substantial part quantitatively.[57] In this regard, *verification* refers to "ensuring the reliability of the information contained in [the] database" and to the monitoring of "the accuracy of the materials collected when the database was created and during its operation".[58] *Obtaining* refers to the searching of independent materials to collect them in the database, and not the investment made in the creation of those materials.[59]

If a database consists of multiple modules, then the substantiality of the extraction from one module needs to be assessed in relation to the whole module; when extractions are made from multiple modules of that database, these have to be assessed in relation to the whole database.[60] Moreover, the *repeated and systematic extractions* of *insubstantial* parts of a database's contents are also prohibited when they "conflict with the normal exploitation of [a] database or […] unreasonably prejudice the legitimate interests of the maker of the database". For example, this is the case when they "have resulted in the reconstruction of a substantial part of those contents".[61]

Retrieving stage

In the course of TDM activities, a lot of extractions may be made when data is retrieved from different sources. Often the whole database is used or retrieved to select data for the further analysis, and if not, there is a high chance that a substantial part is retrieved. There might be cases where only a non-substantial part is extracted from a database, but the retrieval of contents is generally likely to be problematic.

[52] *BHB v. William Hill*, §58.
[53] *BHB v. William Hill*, §67.
[54] *Apis-Hristovich v. Lakorda*, §45.
[55] *Apis-Hristovich v. Lakorda*, §55.
[56] *BHB v. William Hill*, §70.
[57] *BHB v. William Hill*, §71.
[58] *BHB v. William Hill*, §34.
[59] CJEU BHB v. William Hill, §31; CJEU Fixtures Marketing v. OPAP, §40; CJEU Fixtures Marketing v. Veikkaus Oy, §34; CJEU Fixtures Marketing v. Svenska Spel, §37.
[60] CJEU Apis-Hristovich v. Lakorda, §74.
[61] CJEU Directmedia Publishing, §59-60.

Creating target dataset

When creating a target dataset, data is selected from a database to form a new dataset for analysis purposes. This dataset may already have been retrieved in the previous stage, but the miner may also select data from a database from a third-party and subsequently only retrieve the *selected* data. In both cases, the act of selecting and copying contents into a new database will constitute an extraction that requires authorisation when it concerns a substantial part. This is the case even if a copy of the database is obtained with the authorisation of the maker, since the exhaustion only applies to the resale of that *copy*, but not the right to control extraction and re-utilisation.[62]

Mining/analysis

In this stage, the final dataset is being analysed by the mining software. Extractions of the database's contents will be made, since parts of the database are continuously copied in the working memory of the computer. They are temporary, but the extraction right also covers temporary copies.[63] Triaille et al. consider that it is conceivable that no permanent or temporary copies are made at all, for example, when software crawls in the mined data and only "'counts' occurrences or 'registers a link between this data and another [sic] data'".[64] This might indeed be the case. The partial copies that are possibly made into the computer's RAM may be non-substantial in themselves and have not resulted in the reconstruction of a substantial part of the database, since they are only temporarily stored to enable the software to count or make links.

Evaluation or publication

With regard to the evaluation stage, we refer to the discussion under *'Re-utilisation right*'. The same reasoning for the substantiality of the contents that are re-used in the publication of TDM results applies to the extraction rights. The mere incidental inclusion of contents in publications is not very likely to amount to a substantial part of their source database.

7.2.2.2 Right of re-utilisation

Generally, TDM goes beyond the mere consultation of a database. *Re-utilisation* is interpreted as meaning a distribution to the public (of the whole or part) of the contents,[65] and this is fulfilled when the number of persons targeted is indeterminate.[66] Since the contents of a database are only used to be 'read' by mining software, there is no actual disclosure to the researchers themselves, let alone a *public*. It is only in the stage of publication that contents might be disclosed to the public, when they are included in the publication of the results, e.g. quotation or example of data used. In such a case, the only question that remains is whether that would constitute the re-utilisation of a *substantial part*. The latter must be assessed both *quantitatively* and *qualitatively*. Since the first refers to "the volume of data extracted from the database and/or re-utilised, and must be assessed in relation to

[62] Article 7(2)(b) of the Database Directive; *BHB v. William Hill*, §52.
[63] CJEU Apis-Hristovich v. Lakorda, §45.
[64] Triaille et al. (2014), pp.38–39.
[65] CJEU BHB v. William Hill, §67; ECJ 18 October 2012, C-173/11 (*Football Dataco v. Sportradar*), §20; CJEU Innoweb v. Wegener, §37.
[66] CJEU Innoweb v. Wegener, §51.

the volume of the contents of the whole [database]",[67] the incidental inclusion of one or a few records from a database in a publication is not very likely to be substantial from a quantitative perspective.

Qualitatively refers to the "scale of the investment in the obtaining, verification or presentation of the contents" that are re-utilised, irrespective of whether it forms a substantial part quantitatively.[68] In this regard, *verification* refers to "ensuring the reliability of the information contained in [the] database" and to the monitoring of "the accuracy of the materials collected when the database was created and during its operation".[69] *Obtaining* refers to the searching of independent materials to collect them in the database, and not the investment made in the creation of those materials.[70] It is difficult to argue that the incidental inclusion of contents in publications of TDM results are qualitatively substantial parts of the source database. This would mean that a substantial investment is made in that particular data, that does not relate to its creation.

However, it should be noted that in the *Innoweb* case, the CJEU discussed a meta search engine, that enabled the end user to search through multiple databases with one query. This query was 'translated' into the necessary queries for the search engines of the searched databases, providing the user with the search results of multiple database. The Court found that the making available of such a meta search engine constitutes a re-utilisation of substantial parts of the databases concerned, if not their entirety, since it enables the end user to search those entire databases; this finding was not affected by the fact that only parts of those databases were actually consulted and displayed.[71] Although TDM acts themselves may not be affected by the Court's findings, its ruling could be relevant to parties that enable TDM users to search multiple databases for contents with only one query.

7.3 Exceptions

When TDM activities fall within the ambit of an exclusive right under copyright or database law, this does not necessarily mean that they always require authorisation from the author. They may be covered by an exception to the exclusive rights under certain circumstances. The Copyright Directive provides for 20 exceptions of which only one is mandatory for Member States to implement in their national legislations. The Database Directive also provides for some exceptions, both in the context of copyright in databases as well as sui generis rights. In this section, we will discuss the exceptions that are of relevance for TDM. Where the exceptions in both regimes are of a similar nature, they are discussed together. Each section that examines an exception concludes with a discussion of each of the three categories of barriers of our benchmark.

7.3.1 Temporary reproductions

Only the Copyright Directive provides for an exception to the reproduction right for the making of reproductions of transient nature. As mentioned above, it is the only mandatory exception and is

[67] CJEU BHB v. William Hill, §70.
[68] CJEU BHB v. William Hill, §71.
[69] CJEU BHB v. William Hill, §34.
[70] CJEU BHB v. William Hill, §31; CJEU Fixtures Marketing v. OPAP, §40; CJEU Fixtures Marketing v. Veikkaus Oy, §34; CJEU Fixtures Marketing v. Svenska Spel, §37.
[71] CJEU Innoweb v. Wegener, §53.

implemented in all Member States in mostly identical wording.[72] No similar exception is provided for either the copyright or the sui generis regime under the Database Directive, which suggests that it does not apply to copyright in databases;[73] the Norwegian implementation even explicitly states that it does not apply to databases and computer programs.[74] The Commission has proposed the introduction of similar exceptions for copyright and sui generis rights vested in databases.[75]

Article 5(1) of the Copyright Directive provides that temporary reproductions are exempted from the reproduction right of Article 2 when:

- they are transient or incidental;
- they are an integral and essential part of a technological process;
- their sole purpose is to enable
    - a transmission in a network between third parties by an intermediary of a work, or
    - lawful use of a work;
    - they have no independent economic significance.

We shortly discuss the criteria below.

Transient and part of a technological process

Whether this exception is able to cover acts in a TDM process depends on the actual technology used and the way a miner deals with the mined subject-matter. Reproductions made on paper are not regarded transient or incidental,[76] which will not be an issue in the context of TDM that is digital by nature. The making of transient copies is considered an essential part of a technological process when it is carried out *entirely* in the context of implementing a technological process, which excludes reproduction made (partially) outside of that process, and that the technological process "could not function correctly and efficiently without that act".[77] It does not require that no human activity is involved in the process,[78] but reproductions have to be *automatically* deleted,[79] since manual deletion implies a risk that it is stored longer,[80] taking away its transient nature. In that regard, reproductions made in the computer's memory by software in the stage of analysis may be considered to be transient and part of a technological process, since they are automatically removed from the RAM after the analysis and since they are indispensable to the analysis process. In the stage of retrieval of information and creating a target dataset, reproductions

[72] Westkamp (2007), p.12.
[73] See also Beunen (2007), p.27.
[74] Report for Norwegian.
[75] Commission Staff working paper on the review of the EC legal framework in the field of copyright and related rights, SEC(2004) 995, pp.7-8.
[76] CJEU Infopaq, §§68-70.
[77] CJEU 17 January 2012, C-302/10 (*Infopaq II*), §30.
[78] CJEU Infopaq II, §32.
[79] CJEU Infopaq, §65.
[80] CJEU Infopaq, §69.

may be intended to exist only temporarily. However, since they require human intervention to be removed, it is not likely that they are covered by the exception of Article 5(1). Obviously, the same applies for the inclusion of works in TDM publications, which are permanent by nature.

Lawful use

In the context of TDM, a discussion on the purpose of transmission by an intermediary is not relevant. Rather, it is important to consider when there is lawful use of a work, and whether this is the sole purpose of the temporary copy. Recital 33 of the Copyright Directive states that a "use should be considered lawful where it is authorised by the right holder or not restricted by law". In Infopaq II, the Court ruled that the making of an extract of 11 words from a newspaper article was neither restricted by European Union law, nor by Danish legislation. Therefore, it was considered to be lawful use, although it was not authorised by the right holder.[81] There may still be national differences in the interpretation of lawful use: must any use be assessed separately or are all uses by a *lawful user* considered to be lawful when they comply with the other criteria of Article 5(1)?[82] This may bring uncertainty for TDM.

No independent economic significance

The last criterion requires that ephemeral reproductions have no independent economic significance, meaning that "the economic advantage derived from their implementation must not be either distinct or separable from the economic advantage derived from the lawful use of the work concerned and it must not generate an additional economic advantage going beyond that derived from that use of the protected work".[83] It also requires that the works are not modified during the process,[84] which excludes steps in the TDM process where the contents are modified in any way;[85] for example this could be the case where the original subject-matter is enriched with metadata or where unstructured texts are converted into structured data for the purpose of the analysis.

Restrictiveness

The purpose of the exception of Article 5(1) is "to facilitate a technological process and to respond to the needs of modern information technology".[86] This is reflected in Recital 33 of the directive, which states that this exception should particularly apply to the browsing and caching of contents on the Internet. However, the scope of the exception does not seem to provide the flexibility to permit TDM activities for any purpose; it is only the analysis step that is likely to benefit from Article 5(1), which excludes any TDM technology where information must be either retrieved from an external source or where contents must be modified before analysis.

[81] CJEU Infopaq II, §44.
[82] Cf. Westkamp (2007), p.12.
[83] CJEU Infopaq II, §50.
[84] CJEU Infopaq II, §54.
[85] Cf. Triaille et al. (2014), p.48.
[86] European Commission (2004), p.7.

Fragmentation

The literal implementation of Article 5(1) in Member States' laws suggests that no fragmentation would exist regarding this exception. Nevertheless, several criteria may be interpreted in varying ways,[87] in particular the criterion of *lawful use*.

Uncertainty

We identified the concept of *lawful use* as not entirely clear and in need of further explanation. How must this be interpreted in the context of TDM? Is the use to benefit from a copyright exception considered lawful use?[88] Does lawful access to a database with works imply that TDM may be used to extract knowledge, as a user is also allowed to read a book to gain knowledge from reading the text?

7.3.2 Use for teaching and research under copyright law

According to Article 5(3)(a) of the Copyright Directive, Member States are permitted to include an exception exception to both the reproduction right and the communication to the public right for uses of works for for "the sole purpose of illustration for teaching or scientific research". This is only allowed where such uses such uses are justified by the non-commercial purpose to be achieved. Unless impossible, the source and and author must be attributed. Since it is an optional exception, the corresponding exceptions - if any - in - in national copyright laws vary widely both in the scope of beneficiaries and permitted acts. This is well well illustrated by Table 1 - National implementations of the research exception

below, which provides an overview of the Member States' implementations according to the results of the questionnaire that we have conducted.

Country	Rights	Purpose	Object	Attr[89]	Contract derogation	TPM[90]
U = use | R = reproduction | C = communication to the public | M = making available | D = distribution | P = public performance | T = translation | ITS = illustration for teaching and scientific research						
Bulgaria	U	ITS; preparation of analysis, commentary or other scientific research	parts/non-significant number of works in volume	yes, unless[91]	no	no; request
Cyprus	U	ITS	all works	yes, unless	uncertain	
Czech Republic	U	ITS	works	yes, unless	uncertain	

[87] Eechoud et al. (2009), p.111; see also Westkamp (2007), p.12.
[88] Eechoud et al. (2009), p.116.
[89] 'Attr' refers to the presence of an attribution requirement.
[90] The *TPM* column shows whether circumvention of technical protection measures (TPMs) is allowed to benefit from the exception.
[91] 'Unless' refers to the fact that the attribution requirement is not absolute, meaning that, for example, it may not apply where it is impossible or would require disproportionate efforts from the beneficiary of the exception.

Country	Rights	Purpose	Object	Attr[89]	Contract derogation	TPM[90]
U = use | R = reproduction | C = communication to the public | M = making available | D = distribution | P = public performance | T = translation | ITS = illustration for teaching and scientific research						
Estonia	U	ITS; edu. and research institutions	lawfully published works	yes, unless	uncertain	no
France	R/C	illustration in course of teaching and research; public consists primarily of students, teachers and researchers	extracts of works excluded: pedagogical works, sheet music, digital literary works	yes	yes	no; request
Germany	M	research and teaching	all works		uncertain	no; request
Greece	R	teaching	articles lawfully published in newspaper or periodical; short extracts of works, parts of short works or lawfully published works of fine art	yes, unless	uncertain	
Hungary	U	ITS	parts of works	yes, unless	no	no; request
Iceland	U	religious services, classroom instructions or educational broadcasting	parts of works in compilations; individual literary/musical small works; chapters taken from longer works; pictures/drawings	-	no	no; request
Netherlands	R/C	illustration for teaching	parts of works	yes, unless	uncertain	no; request

Country	Rights	Purpose	Object	Attr[89]	Contract derogation	TPM[90]
U = use | R = reproduction | C = communication to the public | M = making available | D = distribution | P = public performance | T = translation | ITS = illustration for teaching and scientific research						
Norway[92]	R	research and teaching; educational/research institutions	works		uncertain	
Poland	R/T	educational institutions, universities and research entities	small works; parts of larger works	yes, unless	uncertain	
Slovakia	R/M/P	ITS	published works		uncertain	yes
Slovenia	R/D/C	issuing textbooks/schoolbooks;	extracts of works; articles in periodic publications; entire visual works		uncertain	no; request
United Kingdom	R	computational analysis; for sole purpose of research	all works	yes, unless	no	

Table 1 - National implementations of the research exception

Restrictiveness

Article 5(3)(a) allows for certain uses *solely* for *scientific research*. It might therefore permit TDM activities when they are carried out in an (academic) research environment, with no other purposes achieved than scientific research. This means that such an exception will lose its relevance for TDM carried out in a context that is not of a 'scientific research nature' or that (partially) serves goals other than scientific research. Also, it excludes any TDM activity that achieves commercial purposes, excluding, for example, any (scientific) research carried out in commercial company to develop products to bring on the market. In the context of TDM, the purpose of "illustration for teaching" seems to lack relevance, since works are used to extract knowledge and patterns and are not themselves used for illustration in teaching activities.

While the scope of this provision is in itself already limited to certain uses, not all Member States have implemented the full scope of this exception and implementations differ largely on the following aspects:

[92] Note that Norway is not a Member State of the EU, but is part of the EEA and has to comply with the rules on copyright law under the *acquis communautaire*.

- *Covered exclusive rights*: while some national copyright laws have a rather neutral wording of the exclusive rights that are covered by the exception, notably by the word "uses", others have restricted the exception to only to apply to, for example, the reproduction right, or the communication or making available to the public right. As discussed in the section on exclusive rights, it is the reproduction right that is predominantly affected by the TDM. Where national implementations of the research exception do not cover this right, it will not be able to cover any TDM activities that involve the making of reproductions.

- *Purpose and beneficiaries of the exception*: the overview in Table 1 also shows that the purpose or beneficiaries of the exception do not always represent those in Article 5(3)(a). Certain countries, such as Greece and The Netherlands, have only covered teaching activities. Even under the umbrella of (scientific) research purposes, the scope in some jurisdictions are narrowed. For example, Estonia appoints certain actors - i.e. educational and research institutions - that may benefit from this exception. Such restrictions could exclude, for example, scientific research carried out in commercial entities, even though non-commercial purposes are achieved. In Slovenia, uses are limited to the specific act of issuing textbooks or school books.

- *Subject-matter covered*: Article 5(3)(a) applies to works in general and this is also the case in some national copyright laws, but the subject-matter covered is more limited in most Member States. For example, France appears to cover only parts of works. The corresponding exceptions in countries such as Bulgaria, Greece, Poland and Slovenia cover both parts of works and small works (often aimed at articles in periodicals), and others include visual works as a whole as well. A requirement often found is that the works used for the given purpose have been lawfully published. For TDM, the research exception might lose relevance where only parts of works may be used, since mining of contents often requires going through the entire works. However, where the notion of 'small works' implies that, for example, journal or newspaper articles are covered, this could cover the situation where TDM is used to merely mine these sorts of subject-matter.

- *Absolute character of the attribution requirement*: in itself, the attribution requirement seems burdensome in TDM where the author(s) of each work in large collections that are mined must be acknowledged, including their source. However, Article 5(3)(a) exempt those cases where this turns out to be impossible. This is similar in most Member States, with some providing that no attribution is required where this would involve disproportionate costs or efforts. There are few states where the attribution requirement appears to have an absolute character, which may discourage researchers from using TDM due to the efforts they would have to make in tracing all authors and sources of the works.

- *Non-commercial*: in all jurisdictions covered in the overview, the exceptions for teaching and research activities only apply where non-commercial purposes are achieved, although slightly different wording may be used for this. TDM carried out for commercial purposes can therefore not benefit from this exception.

- *Others*: other conditions restricting the scope of the exception were also reported. For example, in the Czech Republic, all copyright exceptions are subject to the three-step-test.[93] Consequently, in each case the applicability of the research exception to a certain use has to be assessed, the three-step-test needs to be complied with. Similarly, Slovenia provides for a four-step-test to be applied in individual cases, which is derived from this three-step-test.

The odd man out in the overview of Table 1 is the United Kingdom, where a specific exception for TDM is in force under copyright law. It covers reproductions made to carry out "computational analysis" and is restricted to research purposes. As with all the national exceptions for teaching and research purposes discussed, no commercial purposes may be achieved. Another requirement that is not found in most provisions of our overview above, but which is present in the Database Directive's research exception, is the requirement that only persons having *lawful access* to the works may benefit from this exception. According to the UK *Intellectual Property Office*, this is the case where "researchers have the legal right to access a copyright work to read it"; examples mentioned concern subscriptions to a journal or database and works published under open licences.[94]

Another important factor that may restrict the relevance of a research exception is the extent to which its applicability may be overridden by contract and to what extent the protection of *technological protection measures* (TPMs) covers uses that would normally be permitted under the exception. As regards the contractual overridability, only four countries have reported that derogations by contract are not allowed. Except for France, where contractual derogation is reported to be allowed, it is rather uncertain for the other Member States in our sample; for example, respondents have reported that there is a lack of interpretation by the courts or legislator, as well as a lack of interpretation or consensus in doctrine.

As provided by Article 6(1) of the Copyright Directive, Member States are obliged to provide legal protection against the circumvention of TPMs applied to works and other subject-matter covered under the directive. This 'paracopyright' forms an extra protection layer on top of the copyright in the work,[95] which potentially restricts the user from certain acts that may otherwise be permitted under one of the copyright exceptions. Under Article 6(4), in the absence of any voluntary measures by right holders, Member States must take necessary measures to ensure that users of works may benefit from the exceptions. From our sample, it appears that in most jurisdictions circumvention of TPMs is not allowed when carried out for the purpose of benefiting from the research exception. However, in conformity with the Copyright Directive, they generally provide for a right to request access, whether or not through an assigned committee or body. However, such a process can be burdensome, especially when the TDM user deals with a large amount of works, thereby restricting TDM activities. As a result, right holders may derogate from the research exception by physically or digitally restricting the acts necessary for TDM activities covered by a research exception.

Fragmentation

As shown above, the exceptions vary widely in their scope, which creates a segmented copyright and database law landscape in the EU. This may not be so much of a problem when a miner is strictly

[93] This three-step-test corresponds with Article 5(5) of the Copyright Directive and the three-step-test as found in the major international copyright treaties, such as the Berne Convention and the WIPO Copyright Treaty.
[94] Intellectual Property Office (2014), p.7.
[95] Burk (2003), p.43.

working from one Member State, but difficulties may arise where miners are collaborating on a pan-European level and work from different jurisdictions. In such cases, it may be burdensome for TDM users to assess the lawfulness of their TDM activities when faced with many laws with their own criteria for the research exception to apply, if it exists at all. Thereby, it may also affect pan-European collaborations where miners in 'non-permissive' (meaning probably not permitting TDM) countries may not be in demand as partners. Consequently, this may have a fragmentary effect on the TDM research landscape in Europe as well as the single market for TDM related activities.

Uncertainty

We already observed the fragmentary landscape with regard to the research exception, which in itself also enhances uncertainty as miners may not be aware of copyright law in other countries. However, other elements add to the uncertainty in relation to these exceptions as well, of which some are a shared feature of many exceptions - i.e. the concepts of *lawful use*, *scientific research* and *lawful user* - while other are specific to certain Member States - such as the applicability of a *fair dealing* criterion in the UK and the three-step-test in the Czech Republic.

First, and as Triaille et al. already observed, the phrase "illustration for" that is found in the research exception both in the Copyright and Database Directives, may apply to *teaching* only or to *scientific research* as well.[96] If the latter would be true, they argue, the exception would be of no value for TDM activities.[97] Whether or not the European legislator may have intended the said passage to apply to scientific research as well, it seems that most of the respondents to our questionnaire have not regarded this a problem and that the limited applicability of the exceptions to TDM lies in the other, already identified, limitations of the implemented scopes.

The concept of scientific research is generally not defined in either the copyright and database laws or case law. As Triaille et al. have observed, not all implementations add the term "scientific" before research and when research is to be qualified as "scientific" is also subject to discussion;[98] it could range from a very traditional view where only research within universities is covered, to research by any beneficiary that is regarded to be scientific due to the nature of the research. So if one might conclude that the research exception might be useful for certain TDM activities in a research environment, it is not at all clear what sort of environment or what sort of research is covered.

All implementations of the research exception require the use to be non-commercial. It is very conceivable that academic research within a research university, of which the results are not (commercially) valorised in any way, will be covered by the exception. However, the mere fact that the organisation is not-for-profit or mainly funded by the public does not affect the findings as to its commercial purposes, which is clear from Recital 42 of the Copyright Directive. As Triaille also recognised, it is more likely that no commercial purposes are present to the extent that research entities, for example, mainly carry out fundamental research, in contrast to the case where research is aimed at bringing a new product on the market.[99] However, it becomes problematic where commercial aspects are (remotely) present or where the results of research unintentionally turn out to be commercially viable. It is questionable whether it matters if the initial aim was non-commercial,

[96] Triaille et al. (2014), p.82.
[97] Triaille et al. (2014), p.82.
[98] Triaille et al. (2014), p.60.
[99] Triaille et al. (2014), p.65.

which would render this a rather subjective criterion. It might also bring evidential problems, where the results of research are valorised in a monetary way afterwards: if the onus of proof lies with the TDM researcher, he may have difficulties in providing evidence as to his initial - non-commercial - intentions.

The conditions listed under "Others" in the overview provided in Table 1 also include some that may cause uncertainty. For example, The Netherlands and Greece require that the particular use covered has to comply with "fair practice", which is an open norm and subject to interpretation. However, respondents have not reported any particular issues regarding the interpretation of this criterion.

As already discussed above, uncertainty also exists as regards the extent parties may derogate from the research exception by contract in most jurisdictions.

7.3.3 Use for teaching and research under database law

Article 9(b) of the Database Directive (96/9/EC) also provides Member States with the possibility of implementing a teaching and research exception under the sui generis database regime. Its scope and conditions are almost identical to its copyright equivalent, save for a few exceptions:

- The exception does not cover all exclusive rights, but only the extraction right;
- The attribution requirement is worded in an absolute way, lacking the 'unless impossible' clause in the Copyright Directive;
- Only *lawful users* can benefit from this exception.

Potentially, this exception may cover extractions made in TDM carried out for research purposes and therefore cover the TDM process as a whole, insofar as the re-utilisation of a database's content remains limited to insubstantial parts. In contrast to the research exception in copyright law, it is implemented word-by-word in most Member States. In some countries, the research exception applies to both copyright and database law, thereby being identical in scope for both regimes.[100]

Restrictiveness

For the barriers falling under the category of restrictiveness, we can generally refer to the findings as regards the research exception under copyright law. Nonetheless, as stated above, the research exception under database law is more restrictive as regards:

- The attribution requirement: there seems to be no opportunity to derogate from this requirement when this turns out to be impossible or to involve extreme efforts or costs;
- It does not cover all exclusive rights. Nonetheless, it must be noted that it covers the extraction right, which is most relevant in the context of TDM.

In addition, restrictive factors are found in Czech law, where the three-step-test must be applied in each case, and in Slovenia, where the exception only covers teaching.

[100] This is stated in the reports for France, Iceland and Norway.

While the United Kingdom has implemented a TDM exception in its copyright law, miners have to rely on the research exception with regard to the extraction rights under database law.[101] The exception does implement many of the criteria of Article 9(b) of the Database Directive, but it lacks the requirements for the users to be *lawful*; it rather requires that the database has been made available to the public in any manner. Moreover, it adds that the database right is not "infringed by *fair dealing* [emphasis added]".[102]

Fragmentation

The level of fragmentation as regards the research exception is significantly lower under the sui generis database regime than it is under copyright law. As discussed, there are a few jurisdictions in our sample that differ from other countries to an appreciable extent.

Nonetheless, significant fragmentation exists when comparing the research exception in both regimes, which may raise issues when mined databases both consist of copyrightable works and are protected as such under sui generis database law. Under the Copyright Directive, the exception only permits use for the *sole* purpose of illustration for teaching or scientific research, where its counterpart in the Database Directive lacks the additions of "sole". This implies that scientific research does not have to be the only purpose in order for the exception to cover such uses; if another purpose is served, this might prevent the use from being covered by the exception under the copyright regime. A second difference is found in the exception's attribution requirement. Where the Copyright Directive includes a mitigating factor for the requirement to acknowledge the source and author's name, i.e. exempting the cases where "this turns out to be impossible", the Database Directive lacks such a phrase for its exception under its sui generis regime. As a consequence, the attribution requirement under database law has an absolute character, while its copyright equivalent, both under the Copyright and Database Directives, is less strict.

Uncertainty

In addition to the findings as to the uncertainty with regard to the research exception under copyright law, there is also uncertainty regarding the concept of *lawful user*. It is not only found in the research exception, but in Article 8(1) of the Database Directive as well, which provides that lawful users may extract and re-utilise insubstantial parts of a database. In the context of the latter, Recital 34 of the Database Directive makes clear that when the right holder has made a copy of the database available to the user, that user is regarded a lawful user and may use the database for the purposes and in the way set out in the agreement with the right holder; this also applies where such "use necessitate[s] performance of otherwise restricted acts". Such a description of the lawful user is somewhat problematic in the context of the research exception, especially on the point that the database must be used for the purposes set out in the agreement. If research purposes are excluded for the agreement - or not explicitly addressed, would reproductions for such purposes prevent a user from being lawful?

A similar discussion in doctrine is reported as regards Czech database law. Under Czech copyright law, a 'user' is regarded as an individual that uses a copyright work and thereby is potentially infringing the exclusive rights of the right holder. However, doctrine has rejected the use of such a

[101] See also the report for the United Kingdom.
[102] Section 20 of the UK *Copyright and Rights in Databases Regulations 1997*.

definition as regards the lawful user under database law, since the addition of 'lawful' would create a tautology; a *lawful* user cannot potentially infringe a database right. Furthermore, merely having access to a (copy of a) database is not sufficient according to doctrine under Czech copyright law. The report for The Netherlands shows that the Dutch legislator has provided a broader understanding of the lawful user, including any event where the user possesses a lawfully acquired copy of the database or where the user has access to an online database in accordance with an agreement; the purposes of the agreement do not seem to affect this finding.

As the concept of lawful user is not clear-cut, it is not entirely clear who the beneficiaries of the research exception are under database law. More clarity, preferably on a European level to avoid fragmentation, would clarify in what situations TDM activities would be covered by the exception.

7.3.4 Private copying

The Copyright Directive provides for an exception to the reproduction right regarding the making of copies in the private sphere. Although the Database Directive also provides for a similar exception in Article 9(a), exempting extractions made for private purposes, it only applies to non-electronic databases, thereby lacking any relevance in the context of TDM.

Article 5(2)(b) of the Copyright Directive provides that Member States may provide for an exception for "reproductions on any medium made by a natural person for private use and for ends that are neither directly nor indirectly commercial, on condition that the right holders receive fair compensation". This exception may therefore be useful with regard to TDM activities that involve the making of reproductions.

Restrictiveness

The extent to which this exception is able to cover reproductions made in the course of TDM is particularly limited by the following two factors:

- *Private use*: the requirement of private use limits the scope of the exception to apply in situations where TDM is carried out (entirely) for private use. Insofar as reports have addressed national implementations of the private copying exception, national wording seem to focus on personal use, study or (small scale) research,[103] and sometimes the scope is limited to a few copies.[104] While in The Netherlands every form of collaboration in this regard seems to be excluded, since copies may only be made for the beneficiary's *own* use, the Polish exception is reported to be somewhat broader, covering also copies of works made by a circle of people having personal relationships. However, in general, TDM carried out by groups of people within an institution is not likely to fall under the scope of private copying exceptions.

- *Non-commercial*: reproductions must also be made for non-commercial ends. As with the research exception, it therefore excludes any TDM activity that achieves commercial purposes.

[103] For example, this has been reported for Germany and the Netherlands.
[104] This has been reported for Norway and Poland.

Fragmentation and uncertainty

The scope of the private copying exception already in itself appears to be too restrictive to be of any relevance for users of TDM. It is therefore not relevant to discuss the fragmentation as regards the national implementations thereof, which surely exists,[105] or the uncertainty as to its scope and applicability. The findings as to the contractual overridability and possibilities of lawfully circumventing TPMs for the purpose of benefiting from the exception, as reported by the legal experts, largely apply to the private copying exception as well.

7.3.5 Quotation

A copyright exception for making quotations has been reported for all Member States in our sample. This is not surprising, as Article 10(1) of the Berne Convention, to which all EU Member States are party, provides that "[i]t shall be permissible to make quotations from a work which has already been lawfully made available to the public". The quotation exception is also found in the Copyright Directive, where Article 5(3)(d) provides that

> "*quotations for purposes such as criticism or review, provided that they relate to a work or other subject-matter which has already been lawfully made available to the public, that, unless this turns out to be impossible, the source, including the author's name, is indicated, and that their use is in accordance with fair practice, and to the extent required by the specific purpose*".

In the TDM process, the quotation exception appears particularly relevant where TDM results are published published and extracts of works are occasionally quoted in the publication. Depending on the restrictions of restrictions of the exception and the particular method of TDM, the quotation exception could theoretically theoretically cover reproductions made in the other stages of TDM. Table 2 - National implementations of the quotation exception

provides an overview of the national reports on the quotation exception, which will be used to assess the extent to which the quotation exception may cover TDM.

[105] For example, the remuneration requirement set out in Article 5(2)(c) of the Copyright Directive is applied in varying ways in Member States; see for example World Intellectual Property Organization (2012).

Country	Rights	Purpose/beneficiary	Object	Attr	Permitted proportion
Q = quote or quotation | U = use | R = reproduction | D = distribution | M = making available					
Bulgaria	Q	criticism/review	parts of works	yes, unless	necessary for purpose
Cyprus	Q		passages from published works	yes, unless	justified by purpose
Czech Republic	U	- minor: not specific - major: critique/review/scientific work	- minor: only parts of works in work - major: also small whole works, not exceed extent adequate to given purpose	yes, unless	- minor: justified extent - major: complying with fair practice and required by purpose
Estonia	Q		works, already made available	yes, unless	justified by purpose and idea of work as a whole
France	Q	critical, polemic, educational, scientific, informatory nature	very short excerpts	yes	very short, both in relation to source and work in which incorporated
Germany	R/D/M	referencing	parts of works		justified by purpose
Greece	Q	support case; criticise position author	short extracts of lawfully published works		justified by purpose
Hungary	Q		parts of works	yes	justified by character and purpose of recipient work
Iceland	Q	critical or scientific public discussion or other recognised purpose	any presented work		reasonable length
Italy	Q	criticism and discussion; may not clash with		yes, unless	

Country	Rights	Purpose/beneficiary	Object	Attr	Permitted proportion
Q = quote or quotation | U = use | R = reproduction | D = distribution | M = making available					
		commercial utilisation of work[106]			
Netherlands	Q	in announcement, review, polemic, scientific treatise or comparable purpose	any work lawfully made public	yes, unless	number and size of quoted parts justified by achieved purpose
Norway[107]	Q		parts of issued works; whole work when very short		necessary to achieve desired purpose
Poland		explanation, polemic, critical analysis, and research and teaching; according to rules of genre works	parts of disseminated works		
Slovakia	Q	review and criticism of quoted work	parts of any published work	no?	justified by purpose
Slovenia	Q	illustration, confrontation, reference	parts of works; whole visual works	yes	short quotations in comparison to work as a whole

Table 2 - National implementations of the quotation exception

[106] The report for Italy explains that when the quotation is carried out for teaching or research, such use is not regarded to have a commercial purpose.

[107] Note that Norway is not a Member State of the EU, but is part of the EEA and has to comply with the rules on copyright law under *acquis communautaire*.

Restrictiveness

In conformity with the columns of the table, we discuss the restrictiveness of each factor:

- *Fair practice*: generally, the national exceptions require that quotations are made in accordance with fair practice, which is in conformity with the Copyright Directive. This is in itself not a restricting factor as regards the lawfulness of TDM, but it may depend on the specific case. It is further discussed under *Uncertainty*.

- *Covered acts*: most reported exceptions do not refer to the specific exclusive rights from which quotations are exempted, but rather use a neutral wording like "quotation", "quote" or "use" as acts covered by the exception. According to the German report, only the exception in Germany explicitly refers to the exclusive rights of reproduction, distribution and making available. As such, the issue of the acts covered does not restrict TDM activities in itself.

- *Purpose/beneficiaries*: the purpose of the quotation has the potential to exclude TDM activities when they are not carried out for the purpose described in the exception. The Copyright Directive provides that quotations are made for purposes *such as* criticism and review, therefore illustrating possible purposes but not limiting the possible purposes to those two. Nonetheless, it appears that some countries have chosen a rather limited list of purposes, such as Bulgaria, Germany, Greece, Slovakia, Slovenia or the United Kingdom. Some countries include more open ended concepts in their list of permitted purposes. See for example The Netherlands ("or comparable purpose"), Iceland ("other recognised purpose") and France ("informatory nature"). As a general rule, the exception would only permit TDM activities that relate in some way to criticism and review. Thereby, quotes from works in, for example, research papers would be covered by this exception. However, as regards the reproductions made in earlier stages of TDM, the purposes of criticism and review seem too narrow to sufficiently cover these. With respect to the countries with a broader wording of the exception, it will depend on the actual interpretation in case law - if any.

- *Subject-matter and proportionality*: while reproductions are often made of *whole* works in the TDM process, the quotation exception generally covers only parts of works. Some Member States go beyond short extracts of works. For example, the Czech Republic has a three-stage system for quotations, which goes beyond the mere quoting: it distinguishes between minor quotations, major quotations and those for teaching and scientific purposes. The first seems to conform to the quotation exceptions in most jurisdictions and the latter rather to the research exception in the Copyright Directive; it is therefore already discussed in the section on the teaching and research exception. The second one, 'major' quotations, also covers small *whole* works, in contrast to the exception for 'minor' quotations. In such cases, TDM carried out on, for example, short research papers, might fall within the scope of the exception. Similarly, the Slovenian exception also covers whole visual works, which could be of relevance where photographic works are mined. The size and proportion of the quotation is not fixed, but they generally have to be short. The French exception is even reported to be very short, both in relation to the source and destination of the quote. In such

cases, TDM activities are only covered to the extent that only small parts of works are retrieved and otherwise reproduced in the process.

- *New work*: in addition to the requirements as to the works that are quoted *from*, there can be requirements as to the *destination* of those quotations. For example, the exception in Czech copyright law, for both minor and major quotes, requires these to be part of a new work. This requirement does not particularly restrict quotations made in a publication of TDM research, such as a journal article; it becomes more problematic in cases where reproductions are collected in, for example, a structured database made for the analysis. It would require that such a database is itself a work that is original and involved creative choices regarding the selection and arrangements of its contents. It requires a case-by-case analysis to assess whether this is the case.

- *Attribution*: in conformity with the Copyright Directive, most national quotation exceptions generally require the attribution of source and author, unless this is impossible. As discussed in the context of the research exception, attribution can be extremely burdensome when the user of TDM technology retrieves and mines a large quantity of works; whether the miner will have to attribute the authors and source of all those works depends on the interpretation of *impossible*.

- *Others*: some countries have reported some extra conditions under the quotation exception, that predominantly relate to the moral or personality rights of the author. For example, Iceland requires quotations to be correct and the Dutch exception adds that the author's moral rights are to be respected. This does not seem to be a restrictive factor in the context of TDM. As regards the quotation in TDM publications these may in general be expected to be accurate. Moreover, it is questionable whether reproductions in the preceding stages of the TDM process are able to 'disrespect' the moral rights of the author, when they are merely made for the purpose of analysis itself.

To summarise, while the quotation exception may be very useful to permit quotations in the publication of TDM results, it seems too restrictive for TDM users to rely on such exception; generally, only short works may be used and for a limited set of purposes, where it is possibly required that the destination of those extracts is a work in itself.

Fragmentation

The quotation exceptions as they exist under national copyright laws appear to have many conditions in common, although they are sometimes worded or interpreted somewhat differently and sometimes additional conditions need to be met. Since the exception as provided under the Copyright Directive seems in itself already too restrictive in relation to TDM, we do not regard it as relevant to discuss any barriers that are the result of the fragmentation of national renditions of the exception.

Uncertainty

We observed that quotations, whether they consist of extracts of works or small complete works, are generally required to be short and that their length is not fixed. This may create uncertainties in case where, for example, TDM is carried out on small works and therefore possibly covered by the quotations exception. Will this cover, for example, all scientific papers, or does it make a difference

whether the length is 3,000 rather than 20,000 words? Such uncertainty renders the exception unfit for TDM users to rely on.

Uncertainty also exists where the required purposes for making quotations are not enumerated in an exhaustive way. While TDM may potentially serve such purposes, it will depend on the actual interpretation in case law and may require an assessment of the purposes for each TDM activity. Also in this regard, the exception does not therefore provide a solid basis for the lawfulness of TDM.

Moreover, the issues in the context of contractual derogation and circumvention of TPMs apply here as well. Although Bulgaria, Hungary and Norway have reported that parties may not contractually derogate from the quotation exception, it remains uncertain in most jurisdictions; only France has reported that such derogations are permitted. Generally, TPMs may not be circumvented to benefit from the exception, but some countries[108] have reported that rights exist for beneficiaries to request access or means to do so. In the cases where TDM activities would be covered by the quotation exception, it often remains uncertain whether TDM users may benefit from that where it is contractually restricted and whether they can successfully request the necessary means thereto.

7.3.6 Press exception

The press exception is also among the exceptions that we identified as having the potential to cover TDM acts under certain circumstances, in particular having journalistic TDM in mind. It is found in Article 5(3)(c) of the Copyright Directive, which permits Member States to exempt:

> "*reproduction by the press, communication to the public or making available of published articles on current economic, political or religious topics or of broadcast works or other subject-matter of the same character, in cases where such use is not expressly reserved, and as long as the source, including the author's name, is indicated, or use of works or other subject-matter in connection with the reporting of current events, to the extent justified by the informatory purpose and as long as the source, including the author's name, is indicated, unless this turns out to be impossible*".

Table 3 - National implementations of the press exception

provides an overview of the national implementations, as well as their relevant criteria and conditions, of the press exception in our sample.

[108] As reported for e.g. Bulgaria, Hungary and The Netherlands.

Country	Rights	Purpose	Beneficiary	Object	Topics	Attr	Used proportion
R = reproduction | C = communication | M = making available | U = use | T = translation | D = distribution							
Bulgaria	R/C/M		press	lawfully made available	current economic, political and religion	yes, unless impossible	
Cyprus	R/C/M		press	published articles	current economic, political, religious, broadcast works or other subject-matter of same character	yes	extent justified by informative purpose
Czech Republic	U/T	public security, judicial proceedings, for the purposes of news reporting and other informative purposes		also unpublished works		yes, unless impossible	
France		press review; compare different comments originating from different journalists on same subject/event	despite definition, common view that not necessarily mass media organisations or professional journalists				
Germany	R/C/D	informing		press like media	recent events		

Country	Rights	Purpose	Beneficiary	Object	Topics	Attr	Used proportion
R = reproduction | C = communication | M = making available | U = use | T = translation | D = distribution							
Greece	R/C	informing	mass media	works delivered in public	political speeches, addresses, sermons, legal speeches, or other work of same nature, and summaries or extracts of lectures	yes	extent justified by purpose
Hungary	R/C/M		press	articles, broadcasts	current economic or political topics	yes	
Iceland				newspapers, periodicals, broadcast material	economics, politics, religion	yes	
Italy	R/C	informing		articles	current interest	yes, unless	
Netherlands	U/T		by daily or weekly newspaper, weekly or other periodical, or radio or tv programme or other medium with same function	daily or weekly newspaper, weekly or other periodical radio or tv programme or other medium with same function	news items, miscellaneous items, or articles on current economic, political or religious topics or works of same nature	yes	
Poland		review of publications	journalistic press	short extracts of reports and articles	current events		

Country	Rights	Purpose	Beneficiary	Object	Topics	Attr	Used proportion
R = reproduction | C = communication | M = making available | U = use | T = translation | D = distribution							
Slovakia	R/M/D	informing by press			political speech, public lecture		extent justified by right to inform
Slovenia	R/C/M	informational			media reviews, public speeches, daily news reports	yes	no entire articles

Table 3 - National implementations of the press exception

Restrictiveness

We distinguish between, and discuss, the restrictiveness in the following aspects that correspond to the columns in Table 3:

- *Exclusive rights*: while the harmonising provision in the Copyright Directive refers to the acts of reproduction, communication to the public and making available to the public, this is not necessarily the case in all press exceptions found in national legislation. According to the national reports, most countries do refer to all three acts, or to either at least the acts of reproduction and making available to the public or the acts of reproduction and communication to the public. Those not referring specifically to the acts covered by the author's exclusive rights use neutral wordings such as "use" and in some cases "translations" as well. The acts of reproduction and possible making available in the TDM process are therefore generally not restricted by these factors.

- *Beneficiaries*: most reported exceptions assign the press as the beneficiary of the press exceptions. This potentially includes TDM carried out by users within a press organisation, excluding all non-press users. The Polish exceptions speaks of "journalistic press", which, depending on its actual interpretation, could be more restrictive. The same is true for the Greek exception, which refers to "mass media". The Netherlands appear to have implemented a more open ended concept, adding any "other medium with the same function" to the list of beneficiaries. This is more likely to also cover journalistic and press bodies, or individuals, that fit less in the traditional concept of journalism. TDM carried out by press organisations or individual journalists could potentially fall under the press exception in some Member States, while their position is more doubtful in other jurisdictions.

- *Purpose*: the purposes mentioned in Article 5(3)(c) also mainly relate to journalistic events; this applies for the national implementations as well, although they appear in

different wordings, ranging from broad concepts such as “informational” to more narrow ones such as “informing on recent events”. As soon as TDM is carried out beyond these purposes, the miner cannot rely on this exception as to its lawfulness.

- *Medium*: the type of medium where the used, or mined, contents have appeared is also relevant. The exceptions generally require that they are from specific media, such as newspapers or broadcasts, but also “press like media” as in Germany, which potentially covers a wider range of media. Other countries are even broader, referring to published or lawfully made available works in general.

- *Contents*: the topics covered in the media used are relevant as well. In most national implementations, these generally follow the Copyright Directive: the works used must concern current economic, political and religion related topics or recent events. A few appear to be narrower, such as Slovakia and Slovenia. This may have a restrictive effect as to TDM. For example, if a miner wants to mine a database of newspaper articles, it may not even mine all of its contents. Some articles may not be on topics covered by the press exception. To find out which parts may be mined, and which not, can be a timely and costly effort, possibly preventing a journalist or press body from making use of TDM technology.

- *Proportionality*: some reports show that the national exceptions follow their European counterpart, providing that use under the press exception is permitted to the extent justified by the informatory purpose. This may imply that such use is not allowed where it is not necessary for the informatory purpose. In the context of TDM, this is hard to assess, since, by definition, it is used to explore and find new patterns in existing information sources. Although very theoretically, this may restrict TDM for journalistic uses where, whether or not with hindsight, the TDM seems not to have served the informatory purpose.

- *Attribution*: as with the research exception and the quotation exception, the attribution requirement can have restrictive effects as well. In that regard, we refer to the discussion under the research exception in particular.

- *Contractual derogation*: most implementations provide that the press exception does not apply where right holders of works used have explicitly reserved such use. This may reduce the amount of materials that can be lawfully mined and, when such reservations are not made in a machine-readable way, may raise practical issues regarding sorting out which materials may be used and which may not.

- *Others*: as we already saw with other copyright exceptions, the three-step-test must be applied under Czech copyright law for each individual case to assess whether it is covered under any exception. Such a requirement may restrict the applicability of the press exception to even fewer cases.

To conclude, the press exception has the potential to exempt the journalistic use of TDM in very specific cases. However, the beneficiaries, materials used, as well as the purposes of such use is narrowly defined, making the exception only useful to rely on in a very limited amount of cases where TDM may be used for journalistic purposes.

Fragmentation

Although most Member States have implemented a press exception, their scopes do not seem harmonised on many aspects: they differ, although sometimes only slightly, in the uses covered, beneficiaries and subject-matter.

Uncertainty

Uncertainties that seem to bring most the uncertainty as to the applicability of the press exception to TDM, and that seem to have the most restrictive effect, relate to

- topics that are covered: can journalists mine newspapers in their entirety or only certain sections?
- beneficiaries: can a data journalist, for the purpose of his own weblog, rely on the press exception?
- media: for example, what are "press like media" and do they cover new (digital) modes of publishing news?

7.3.7 Use of insubstantial parts

The scope of the exclusive rights of extraction and re-utilisation under the Database Directive extend only to substantial parts of a database. A contrario, insubstantial parts are not covered by these rights - save for the systematic and repeated extraction or re-utilisation thereof that conflicts with the normal exploitation of the database or unreasonably prejudices the legitimate interests of the database maker. Thereby, the exception of Article 8(1) as regards the extraction and re-utilisation of insubstantial parts of a database might seem somewhat superfluous. Nonetheless, in accordance with the title of Article 8, it is instead formulated as a right of the (lawful) database user. The provision prohibits the maker of a database, that is made available to the public, from "prevent[ing] a lawful user of the database from extracting and/or re-utilizing insubstantial parts of its contents, evaluated qualitatively and/or quantitatively, for any purposes whatsoever". The lawful user therefore has a general right to use an insubstantial part and, according to Article 15 of the Database Directive, any contractual provisions providing to the contrary are null and void.

Restrictiveness

We already discussed the scope of the exclusive rights under database law and the extent to which they restrict TDM. The exception for use of insubstantial parts does not affect those findings. Where databases are protected under the sui generis regime, Articles 8(1) and 15 of the Database Directive make the limited scope of these rights even more absolute, since contractual restrictions extending the prohibited acts to insubstantial parts are not permitted. Nonetheless, restrictiveness can be found in the fact that the provisions of these articles only apply to databases protected under the sui generis regime. This is confirmed by the CJEU in its *Ryanair* decision.[109] While this may seem only logical, the fact that they have the character of protecting the interests of database end users may as well have led to the conclusion that such rights extend to the use of an unprotected database.[110] As a result, the user of a database has stronger access rights to protected databases than unprotected

[109] CJEU 15 January 2015, C-30/14 (*Ryanair*), §45.
[110] Cf. Hugenholtz (2015), 304.

ones, which may restrict the uses of unprotected databases for the purpose of TDM to greater extent.

Despite the prohibitions on contractual derogations to the use of insubstantial parts, TPMs might still be applied to restrict use of a database. In the national reports, we often see that circumvention of TPMs to benefit from the user right is not automatically permitted. As with the other copyright and database law exceptions, users generally have to rely on their right to make a request for means or access.

Fragmentation and uncertainty

The absolute character of Articles 8(1) and 15 of the directive is reflected in that they are implemented in all Member States (of our sample), so as to prohibit contractual deviation. Therefore, there seems to be no fragmentation on the implementation to an appreciable extent. Nonetheless, as identified above, fragmentation exists between the rights of a database user, ensuring he or she can use insubstantial parts of databases protected under database law, but potentially hindering this with respect to unprotected databases. This may also result in uncertainty, since it is not always possible to outwardly assess a database as to whether it might be protected under database (or copyright) law or not. The user can therefore not be sure whether they have the right to (re-)use insubstantial parts and whether this might result in a breach of any contractual provisions prohibiting such use.

Another uncertainty lies in the concept of the lawful user, as already discussed under the teaching and research exception under database law. We therefore refer to Section 7.3.3.

8 DATA PROTECTION LAW

This section identifies the barriers to TDM, according to our benchmark, within the legal framework for data protection in Europe. First, it identifies the barriers that can be found in the general rules of data protection law. This is followed by a section on the 'exception' for *historical, statistical and scientific purposes*, since that concept has the potential to cover TDM under certain conditions and to create some leniency for these activities. We will therefore also identify barriers to this concept separately.

Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data was only very recently adopted (on 27 April 2016), after four years of intense negotiations. Consequently, in view of the short deadline of this deliverable, we can only provide a preliminary evaluation of the provisions contained therein. So far as possible, we will compare the rules previously contained in Directive 95/46/EC, with those of the new instrument. Since the rules in the Regulation only apply from 25 May 2018, we will revisit the current findings in a subsequent version of this deliverable, taking advantage of the legal commentaries that will undoubtedly have been published in the meantime on the topic of the new Regulation. For now, all scholarly commentary referenced inevitably relates to the provisions of Directive 95/46/EC.

8.1 General rules of data protection law

When TDM is carried out on personal data, this is highly likely to involve acts that are relevant under data protection law in the EU Member States. Data protection law is now fully harmonised by the General Data Protection Regulation (2016/679/EC) (GDPR). Two key concepts that trigger the data protection regime are *personal data* and *processing*, as the processing of personal data is subject to the principles and legal requirement provided for in the Regulation.

8.1.1 Principles, obligations and concepts

8.1.1.1 Personal data

It is the identifiable or identified person that is protected by data protection law.[111] Article 4(1) of the Regulation (2016/679/EC) provides that personal data means "any information relating to an identified or identifiable natural person… [where] an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person". It is thus crucial whether a person is identified or identifiable, which might depend on who owns what data. While an internet service provider might be able to identify a specific person from an IP address, the average surfer on the web probably could not.[112] Generally, the equivalent provision of the Directive was broadly interpreted and while certain data may not constitute personal data by

[111] Kuner (2007), p.76.
[112] Cf. Kuner (2007), p.91.

itself, in combination with other data it might enable a party to identify a person or make them "considerably easier" to identify.[113]

Regarding the former provision, article 2(a) of the Directive, the Article 29 Working Party, established to advise on several matters of data protection law, has stated that many cases where data subjects can be identified exist in the online environment,[114] and that "[a]nonymisation is increasingly difficult to achieve with the advance of modern computer technology and the ubiquitous availability of information".[115] Considering the explosive growth in available data, it is conceivable that such cases exist to an even greater extent. In such an environment, it may therefore be safer to assume that data processed in this context constitutes personal data, unless it is evident that it is not.[116] Note that photographs may also constitute personal data.[117]

Under European data protection law, a stricter regime applies to special categories of data, that constitute *sensitive data*; according to Article 9 GDPR, these include personal data revealing:

- racial or ethnic origin
- political opinions
- religious or philosophical beliefs
- trade-union membership, and
- data concerning health or sex life

8.1.1.2 Processing

It is the *processing* of personal data that causes the rights and obligations as regulated by the GDPR to apply. The concept of processing must be interpreted rather broadly, meaning "any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction".[118] Since TDM involves the making of reproductions of subject-matter, as well as the possible publication thereof, it is evident from such a definition that the mining of personal data will constitute processing of personal data. Examples could be the mining of:

- Internet browsing and clicking behaviour
- Customers' buying history
- Patient records
- Research involving human subjects
- Genetic data

[113] Kuner (2007), p.92.
[114] Article 29 Data Protection Working Party (2000), p.21.
[115] Article 29 Data Protection Working Party (2013), p.13.
[116] Cf. Kuner (2007), p.92.
[117] Kuner (2007), p.93.
[118] Article 4(2) of the GDPR.

8.1.1.3 Data controller and data processor

Under data protection law, it is crucial to identify the *data controller* and the *data processor*. This relates to the obligations and liabilities under the Regulation which are primarily aimed at the data controller. While the Regulation provides for some responsibilities for the processor, they mainly follow orders from the controller and are therefore not subject to many obligations.[119] This difference in role is reflected in the definitions of *controller* and *processor*. According to Article 4(7), the controller is "the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data; where the purposes and means of such processing are determined by Union or Member State law, the controller or the specific criteria for its nomination may be provided for by Union or Member State law". Article 4(8) regards the *data processor* to be the "natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller". As follows from the definition of controller ('alone or jointly'), multiple parties might be considered to be the controller of the data, which are then regarded as "co-controllers".[120] When a controller chooses a processor to process data on his behalf, this does not discharge the controller from obligations relating to the security of the data. Article 28(1) GDPR provides that the controller must "use only processors providing sufficient guarantees" relating to the technical and organisational security of the processing of the personal data and he must also ensure that these measures are complied with.

In text and data mining, it will often be the miner who decides which data is processed for what purposes. It is therefore likely that he will be considered the *controller*. It might become more complex where research is carried out within a large consortium of collaborating researchers, who jointly determine what research methods are used and what data is used for what purposes. It depends on the circumstances and on what is mutually arranged among the researchers involved. This also applies to, for example, the situation where a researcher instructs a research assistant to mine a certain dataset or a certain category of data. It might depend on how detailed the instructions were and how much leeway was given to that assistant in assessing whether he is a mere processor or makes decisions in a way that he might be regarded a controller of the data.

8.1.1.4 Applicable law

Among the novelties brought by the Regulation is the territorial scope of the data protection. According to Article 3, the Regulation applies to the processing of personal data in the context of the activities of an establishment of a controller or a processor in the Union, *regardless of whether the processing takes place in the Union or not*. The protection also applies to the processing of personal data of data subjects who are in the Union, by a controller or processor not established in the Union, where the processing activities are related to:

a) the offering of goods or services, irrespective of whether a payment of the data subject is required, to such data subjects in the Union; or

b) the monitoring of their behaviour as far as their behaviour takes place within the Union.

[119] Kuner (2007), pp.69–70.
[120] Kuner (2007), p.70.

This provision was adopted in response to the pre-existing rules which required an examination of the location of the *data controller* and whether the processing of personal data was part of the controller's establishment. The *Google Spain* decision illustrates that this was not always clear-cut. In that decision, the CJEU first regarded the operator (Google) of a search engine to be a data controller.[121] However, the actual controller in this case was Google Inc., which is established in the U.S. and, therefore, the provisions of the Data Protection Directive would not apply in the underlying case. However, the Court found "that processing of personal data is carried out in the context of the activities of an establishment of the controller on the territory of a Member State […], when the operator of a search engine sets up in a Member State a branch or subsidiary which is intended to promote and sell advertising space offered by that engine and which orientates its activity towards the inhabitants of that Member State".[122] The decision showed that the activities of an establishment did not necessarily have to involve the processing of personal data, but that it is sufficient that they are directly or indirectly related to such processing.

8.1.1.5 Principles and obligations

Any processing of personal data must comply with the six main principles provided for by Article 5(1) GDPR, which provides that such data must be:

a) processed lawfully, fairly and in a transparent manner in relation to the data subject ('lawfulness, fairness and transparency')

b) collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes ('purpose limitation');

c) adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed ('data minimisation');

d) accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay ('accuracy');

e) kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed; personal data may be stored for longer periods insofar as the personal data will be processed solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) subject to implementation of the appropriate technical and organisational measures required by this Regulation in order to safeguard the rights and freedoms of the data subject ('storage limitation');

f) processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss,

[121] CJEU 13 May 2014, C-131/12 (*Google Spain*).
[122] CJEU Google Spain, §60.

destruction or damage, using appropriate technical or organisational measures ('integrity and confidentiality').

In addition to these general principles, Articles 13 and 14 GDPR provide that the data controller must inform the data subject of, inter alia, his identity, the purposes of the processing of data, the recipients and categories of personal data, and the existence of the data subject's right of access and right to rectify his data - which is regulated in Article 15.

In the context of TDM, it is important to note that the Article 29 Working Party had emphasised in relation to Article 6(1) of the Directive 95/46/EC that

> "*it has no reason to believe that the EU data protection principles, as they are currently enshrined in Directive 95/46/EC, are no longer valid and appropriate for the development of big data, subject to further improvements to make them more effective in practice. It also needs to be clear that the rules and principles are applicable to all processing operations, starting with collection in order to ensure a high level of data protection*."[123]

This quote illustrates the fundamental nature of the data protection principles, which is not surprising given that the protection of personal data is recognised as a fundamental right under Article 8 of the Charter of Fundamental Rights of the European Union.

8.1.1.6 Legal grounds

Any processing of personal data requires a legal ground. Article 6 provides a limitative list of six grounds that legitimise the processing of personal data:

a) the data subject has given consent to the processing of his or her personal data for one or more specific purposes;
b) processing is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract;
c) processing is necessary for compliance with a legal obligation to which the controller is subject;
d) processing is necessary in order to protect the vital interests of the data subject or of another natural person;
e) processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller;
f) processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data, in particular where the data subject is a child.

Paragraph 4 of Article 6 GDPR further specifies that

> *"Where the processing for a purpose other than that for which the personal data have been collected is not based on the data subject's consent or on a Union or Member State law*

[123] Statement on Statement of the WP29 on the impact of the development of big data on the protection of individuals with regard to the processing of their personal data in the EU, 16 September 2014, p. 2.

which constitutes a necessary and proportionate measure in a democratic society to safeguard the objectives referred to in Article 23(1), the controller shall, in order to ascertain whether processing for another purpose is compatible with the purpose for which the personal data are initially collected, take into account, inter alia:

a) any link between the purposes for which the personal data have been collected and the purposes of the intended further processing;
b) the context in which the personal data have been collected, in particular regarding the relationship between data subjects and the controller;
c) the nature of the personal data, in particular whether special categories of personal data are processed, pursuant to Article 9, or whether personal data related to criminal convictions and offences are processed, pursuant to Article 10;
d) the possible consequences of the intended further processing for data subjects;
e) the existence of appropriate safeguards, which may include encryption or pseudonymisation."

This is slightly different with regard to the processing of sensitive data, which is in principle prohibited by Article 9(1) GDPR. Nonetheless, under certain conditions provided for by Article 9(2) such processing is allowed. First, the explicit consent of the data subject may legitimise the processing of sensitive data, although Member States may provide that such consent does not constitute a legal ground for sensitive data.[124] Second, such processing is permitted where such data is *manifestly made public* by the data subject.[125] Third, other exceptions to this rule relate to the interests of the controller in performing activities under employment law, the protection of the data subject's vital interests and the processing carried out by non-profit entities with a "political, philosophical, religious or trade-union aim". The list of permissible grounds for processing of sensitive data has been augmented in the Regulation to include the following:

a) processing is necessary for reasons of substantial public interest, on the basis of Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject;
b) processing is necessary for the purposes of preventive or occupational medicine, for the assessment of the working capacity of the employee, medical diagnosis, the provision of health or social care or treatment or the management of health or social care systems and services on the basis of Union or Member State law or pursuant to contract with a health professional and subject to the conditions and safeguards referred to in paragraph 3;
c) processing is necessary for reasons of public interest in the area of public health, such as protecting against serious cross-border threats to health or ensuring high standards of quality and safety of health care and of medicinal products or medical devices, on the basis of Union or Member State law which provides for suitable and specific measures to safeguard the rights and freedoms of the data subject, in particular professional secrecy;
d) processing is necessary for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) based on Union or

[124] Article 9(2)(a) of the GDPR.
[125] Article 9(2)(e) of the GDPR.

Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject.

The GDPR also allows Member States to maintain or introduce further conditions, including limitations, with regard to the processing of genetic data, biometric data or data concerning health.

8.1.2 Benchmarking the general regime

Restrictiveness

Data protection law has the potential to restrict TDM activities, meaning that mining of personal data cannot be carried out unreservedly. As appears in part from the aspects of data protection law discussed above, it is restricted predominantly by the following factors:

- principles of processing
- legal grounds
- informatory obligations

As regards the principles of processing, the purpose limitation of Article 5(1)(b) in particular may restrict TDM where the miner has not collected the data from data subjects, but rather through a third party. The purpose limitation provides that personal data may not be processed for purposes incompatible with those for which they were originally collected. As a consequence, the miner may not use the personal data if the purpose for which he or she carries out the TDM activities is not compatible with the purpose for which they were collected. In cases where personal data is retrieved through third parties and the TDM appears compatible with the original purposes of collection, an obligation will still rest on the miner to inform the data subject(s) of, inter alia, his identity and the purposes of the processing. This can be rather burdensome where a large amount of data subjects is concerned.

The legal ground requirements of the Regulation have a restrictive effect on TDM, since the lack of any legal ground renders the TDM activity unlawful. Whereas, under the former directive, a TDM user cannot rely on the legal grounds that relate to their legitimate interests, performance of a contract, a legal obligation, the public interest or the vital interests of the data subject(s), and so must rely on the consent of the data subject, the Regulation has broadened the legal grounds for processing personal data. The same observation can be made with respect to the processing of sensitive data. Although the mining of sensitive data is in principle more restricted than the mining of merely ‘normal’ personal data, the European legislator has introduced new possibilities for exception, mainly in the area of health and medicine.

Even where the processing of personal data in the course of TDM turns out to be lawful, the information obligations of Articles 13 and 14 GDPR can serve as a barrier, in particular where many data subjects need to be informed; the same applies to the access rights of Article 15. As noted above, the number of data subjects is often quite large and may therefore render such obligations and rights impossible or burdensome for the miner to comply with.

The GDPR introduces extensive obligations for data management, provision of information to data subjects and reporting to the “supervisory authority”, which the processor must comply with before

carrying out the processing of personal data. The Regulation also introduces interesting provisions on the development and approval of codes of conduct which may prove useful in the future for TDM activities. More stringent information and reporting obligations may create barriers for a TDM user. Nonetheless, the extent to which these requirements really restrict TDM may depend on several factors.

Fragmentation

Until the recent adoption of the Regulation, although data protection law was highly harmonised in the European Union, it was still subject to national implementation and interpretation of the provisions of the Data Protection Directive. Member States now have two years, until 2018, to make their national system compliant with the new Regulation, before the new instrument formally applies. Until such time, fragmentation can therefore still persist in the national interpretations of several concepts of EU data protection law.

The category of sensitive personal data has at times been subject to national interpretation. For example, this is illustrated by the fact that under Czech data protection law, biometric data - which allows identification of a data subject - has been considered to be sensitive data.[126] Kuner has also identified that not only definitions of sensitive data differed to a great extent in national data protection laws, but the requirements as to their processing as well.[127] As a consequence, personal data might have been regarded as 'normal' data in one jurisdiction, while it was considered sensitive data in another. Moreover, the processing of sensitive data might have been lawful in one Member State, while it was considered unlawful in another due to different requirements. Whether these variations will continue in the intermediate period is unclear.

Second, each Member State has its own national DPA, which creates a fragmentary data protection landscape in the following ways:

- *notification*: different DPAs might apply different (formal) requirements as to the notifications of data processing. This has the potential to restrict use of TDM in an environment where pan-European dataflows and collaboration exist, when DPAs do not streamline their requirements.

- *self-regulation*: it appears from the national reports that national DPAs can play slightly different roles concerning the approval and assistance with regard to the setting up of self-regulation or codes of conduct within certain sectors. Although we did not receive extensive information on this aspect, we saw that DPAs in certain Member States do provide assistance in this context. In Norway, the DPA offers services for the drawing up of codes of conduct. In France and The Netherlands, the respective DPAs also provide the possibility of approval or 'certification' of a code of conduct or other form of self-regulation as being in compliance with the data protection rules. DPAs may apply different requirements to qualify for such approval.

[126] As appears from the report for the Czech Republic.
[127] Kuner (2007), pp.103–106.

Uncertainty

Uncertainty exists in some key concepts of data protection law of which that of personal data is, in particular, crucial in the context of TDM. The definition of personal data appears to be very broad. As noted earlier, it may therefore be safer for a TDM user in many cases to assume that the mined contents are personal data. However, when all obligations and principles of data protection law deter the miner from carrying out mining activities, anonymisation of such data may be helpful. The Article 29 Working Party has recognised that "in practice, there is a very significant grey area, where a data controller releasing the data might believe a dataset is anonymised, but a third party may still be able to identify at least some of the individuals from the data".[128] If the miner cannot be sure that the data concerned is indeed considered to be anonymous - rather than personal - data, such uncertainty may prevent the user from carrying out TDM or other parties, such as investors, may be held back from such activities.

8.2 Concept of historical, statistical and scientific purposes

8.2.1 HSSPs

The barriers that are identified above as regards the general rules of data protection law do not necessarily apply in each specific case. In particular, in the context of TDM carried out for *historical, statistical and scientific* purposes, a 'lighter' regime may apply which we will refer to as the HSSPs. Several principles and obligations of the Regulation do not apply when activities fall within the scope of this concept.

First, the purpose limitation of Article 5(1)(b) considers the processing of personal data for "historical, statistical and scientific purposes" not to be incompatible with the purposes for which they were initially collected. Likewise, where Article 5(1)(e) normally requires that personal data may not be stored longer than is necessary for the purposes for which they were collected and further processed, longer storage is permitted for historical, statistical and scientific use. Also, the information obligations of Article 14 concerning cases where data are not being collected from the data subject do not apply where, "in particular for processing for statistical purposes or for the purposes of historical or scientific research, the provision would involve a disproportionate effort or if recording or disclosure is expressly laid down by law". For all three cases, this special regime is subject to "appropriate safeguards" that Member States are required to lay down.

8.2.2 Benchmark

To benchmark the HSSPs as to the three categories of barriers, the focus is on the concept of the scope and interpretation of the HSSPs in Member States' law and the concept and implementation of appropriate safeguards as developed over the years on the basis of the Data Protection Directive. We raised particular queries on these aspects in the questionnaire sent to the correspondents. The input was not extensive, but was sufficient to demonstrate where the obstacles are as regards TDM.

Restrictiveness

Generally, the national reports show that there is no general interpretation of historical, statistical and scientific purposes; neither is there any general interpretation from the European legislator. We

[128] Article 29 Data Protection Working Party (2013), p.13.

see that national implementations often refer to other legislative acts that deal with very specific sorts of data or purposes. We discuss the concept with respect to its three different aspects:

- *historical*: not much is reported on the interpretation of historical purposes. However, the reports of the Czech Republic and Norway both refer particularly to archiving purposes. Under Czech data protection law, the Act on Archiving and Records Management is relevant in this context. It lists the specific entities that are obliged to store records. They are exempted from the data protection obligations only where these entities process personal data to carry out their legal duty under the act. In the context of TDM, this does not seem to have much relevance, unless the legal duty of such archives is to carry out TDM activities for certain purposes.

- *statistical*: our sample shows that *statistical* commonly refers to state statistics.[129] This is much stricter than the term itself implies, since state statistics is only one of the many purposes for which - and by whom - statistics can be obtained. Note that some reports have not reported any interpretation, some due to the lack thereof; therefore, we cannot state as a general rule that statistical refers to state statistics. We can only state that in certain jurisdictions, the scope of the definition is very narrow.

- *scientific*: the term itself implies that this aspect of the HSSPs has the potential to cover the processing within the course of the TDM process when carried out for (scientific) research purposes. Nonetheless, insofar as national reports have provided any interpretations, this is not necessarily the case. Some broad definitions are provided for the Czech data protection regime, where in scholarly doctrine it is argued that a controller may benefit from the regime for HSSPs where these purposes are recent - in contrast to merely proclaiming scientific purposes. The Icelandic report provides a particular description of what is covered by the concept: research dealing with specific tasks, with a predefined beginning and end, and which answers a particular question or hypothesis and creates new knowledge or increases existing knowledge. In the Netherlands, a governmental decree provides more details on the processing for scientific purposes; while it does not provide a definition thereof, it states that it only covers the processing for "particular research". This implies that there should be a concrete research project for which it is used and that it may not merely be stored in case a relevant research project might pop up.

As regards the concept of "appropriate safeguards", Recital 29 of the Data Protection Regulation states that they "must in particular rule out the use of the data in support of measures or decisions regarding any particular individual". The latter principle is explicitly referred to in the Dutch and Slovakian reports; the Dutch government explicitly mentioned that no use of personal data in support of measures or decisions will exist in the context of scientific research. Other principles reported concern:

- *anonymisation*: this is either reflected in an obligation to separate identifiable data from anonymous data, both legally and by technical means,[130] or in the fact that data need to be

[129] For example, this is reported for Bulgaria, the Czech Republic and France.
[130] This is reported for The Netherlands.

anonymised as soon as possible.[131] The underlying idea seems to be that underlying data may only be kept when it is really necessary for the HSSPs.

- *sole purpose*: according to the Polish report, when data is not used for any purpose other than scientific purposes, such processing is legitimate and therefore meets the safeguards requirement.

- *integrity/security*: in some Member States, appropriate safeguards are connected with the integrity and security or a certain treatment of the processed personal data, where the reports generally refer to the general obligations of the data controller in this context.[132]

- *narrow scope*: in the Czech Republic, the rather narrow interpretation of HSSPs is regarded to be an appropriate safeguard.

- *others*: as reported for Greece, particular safeguards can be imposed by the DPA before it authorises certain processing of personal data.

Generally, the safeguards may impose certain obligations on the TDM user when personal data is to be mined, but these do not appear to be more restrictive than the obligations that already exist under data protection law in general. The only difference is that, in this context, they must be complied with to ensure that the particular processing for HSSPs is carried out in an appropriate manner. While the Regulation more frequently refers to HSSPs, it remains to be seen whether the restrictive interpretation given in some Member States will persist.

Fragmentation

We already noted that no general (European) interpretation of HSSPs exist. Moreover, even within Member States, definitions of the relevant terms - historical, statistical and scientific - are provided in a rather fragmentary way, often referring to specific acts. As a result, the scope of HSSPs differs extensively among Member States. The same applies to the implementation of appropriate safeguards, which may bring different obligations and other safeguards that data controllers, e.g. miners of personal data, have to comply with. This might hinder organisations, (research) institutes, companies or other entities established in multiple Member States from carrying out TDM or cause them to base all their activities in one Member State. The same may apply to pan-European collaboration between users of TDM that would jointly be regarded to be the data controller.

Uncertainty

From the lack of interpretation of HSSPs in many reports, it appears that there exists uncertainty with respect to this concept. This is mostly due to an absence of interpretation by both legislators and courts, although in some cases some doctrine has been developed in scholarly literature. As a result, where TDM concerns the mining of personal data, the concept of HSSPs, and thereby the possible relief from certain obligations and requirements under data protection law, is not sufficiently interpreted to serve as a solid basis for miners to rely on when they potentially fall within its scope.

[131] This is reported for Slovakia.
[132] For example, this is reported for France and Norway.

9 OVERALL FINDINGS FOR LEGAL BARRIERS

This section provides an overview summarising overview all the *legal* barriers that are identified in all European and national copyright law, database law and data protection law regimes. They are classified according to the three categories of barriers in our benchmark. Please note that the results of the previous sections have not taken into account the TDM exception that was recently adopted in France. When we conducted our questionnaire, the TDM exception had only been proposed. For more on the French TDM exception, see Section 10.2

9.1 Restrictiveness

Table 4 - National implementations of the press exception

shows an overview of the barriers identified that fall within the category of restricting TDM, highlighting the aspects of the barriers that appears most problematic.

Regime	Barrier	Consequence
Copyright and database law	Contractual derogation	In most jurisdictions, it is uncertain whether contractual derogations from copyright exceptions are permitted, but in some it is clear that it is actually allowed. This has the potential to restrict TDM even more than the provisions of copyright and database law, as TDM users may be prevented from benefiting from the exceptions that would otherwise cover their activities.
	Circumvention of TPMs[133]	The circumvention of TPMs to benefit from a copyright exception is only reported to be allowed for one Member States in our sample. In all others, it is not permitted. Nonetheless, a right generally exists for beneficiaries of exceptions to request access or means to actually benefit, but this may be burdensome to do for each work that is to be mined. Some reports state that in their respective countries such requests have never been made.
	research exception: - exclusive rights covered - limited beneficiaries - scope of subject-matter - attribution requirement	The limited scope of the research exception in many Member States renders it insufficient to rely on for TDM research. Either the acts themselves are not exempted or the beneficiaries or purposes are very narrowly defined. The non-commercial requirement excludes any TDM research that achieves commercial gains. In some Member States, only parts of works are covered by the exception. The attribution requirement

[133] TPM refers to *technological protection measures*.

Regime	Barrier	Consequence
	- non-commercial - three-step-test	can be burdensome to miners. In certain countries, a three-step-test needs to be applied in every single case, which may restrict the scope even more.
Copyright law	scope of reproduction right	Many acts in TDM involve the making of reproductions and require authorisation from the author of works
	scope of making available right and distribution right	With the publication of TDM results, parts of works might be quoted or otherwise included and will therefore constitute an act of making available (e.g. online) or distribution (e.g. on paper)
	exception for temporary reproductions: - transient copies - non-commercial	Reproductions made in the analysis stage of TDM may be covered, but reproductions made in the other stages are not likely to benefit from this exception. As with many copyright exceptions, the non-commercial requirement excludes TDM activities carried out to gain economic advantage.
	private copying exception: - private use - non-commercial	The scope of the private copying exception is generally restricted to the mere private use. Only in certain cases might it cover TDM where it is carried out in a small and closed environment for research purposes. The non-commercial requirement excludes any use where any direct or indirect economic advantages are gained.
	quotation exception: - purpose - subject-matter - recipient work	The purpose of quotation is often criticism or review, which would cover TDM in very few cases. The subject-matter covered is in all Member States limited to a - sometimes extremely - small proportion, where some allow the quotation of small whole works. TDM of whole works is generally not permitted. Some implementations require the quotes to be part of a new work, which excludes TDM where the mined contents are not part of a new work - which is often the case.
	Press exception: - beneficiaries - purpose - subject-matter	Beneficiaries are limited to the press, with some narrower implementations only referring to either “journalistic press” or “mass media”. The purpose is to inform on recent events, which implies that outdated events are not covered. The subject-matter consists of newspapers or periodicals in terms of media, and must concern current economic, political and religion related

Regime	Barrier	Consequence
		topics in terms of topic; this excludes TDM of other sorts of media (blogs?) and/or covering other topics.
Data protection law	Principles of processing: - purpose limitation - HSSPs[134]	When personal data are mined, virtually anything done with it constitutes processing. This is subject to principles, that may restrict TDM especially where not collected from data subjects themselves and the TDM purposes are not compatible with the purpose for which the data were initially collected. TDM for HSSPs are regarded as compatible; although the scope of HSSPs is generally very limited, it may cover scientific research in some jurisdictions (for others it is less clear).
	Legal grounds: - consent	Consent as a legal ground for processing is difficult where data is obtained from a third party.
	Information obligations	Obligations to inform the data subjects on the processing of their personal data may require significant effort from the TDM user when dealing with lots of data subjects.

Table 4 - National implementations of the press exception

[134] HSSPs refers to *historical, statistical and scientific purposes*.

9.2 Fragmentation

Table 5 - Overview of legal barriers due to fragmentation

shows an overview of the barriers identified that fall within the category of *fragmentation*, thereby restricting TDM. It highlights the most important aspects for each legal regime discussed in the foregoing sections. Note that the findings on the fragmented interpretation and application of the rules data protection may need to be revised as soon as the General Data Protection Regulation will enter into force in April 2018.

Regime	Barrier	Consequence
Copyright law	Implementation of copyright exceptions	Except for the exception for transient copies, all copyright exceptions provided under the Copyright Directive are optional to implement in national copyright laws. A large variation in the implementation of the (teaching and) and research exception exists, which especially serves as a barrier for pan-European collaboration for TDM purposes.
Copyright and database law	Research exception: - EU framework - implementation	In the European framework, fragmentation exists between the research exception under the Copyright Directive and its counterpart under the Database Directive. The latter is stricter on the 'lawful user' and attribution requirement, but more lenient with regard to the 'scientific research' requirement (lacking the addition of "sole"). Thus, while TDM research may be permitted under one regime, the same research may not be permitted under the other. The same situation applies as regards the national implementations, where many Member States have implemented the exception under the Database Directive almost word-for-word, in contrast to the research exception under the Copyright Directive.
Data protection law	Sensitive data	The definition of sensitive data, as well as the requirements for processing, differs in scope in national data protection laws. Therefore, the lawfulness of processing of such data in one Member State does not necessarily imply that it is lawful in another.
	National DPAs - notification - role	Each Member State has a national DPA, which must be notified of certain processing of personal data when the data controller is (also) established in that respective jurisdiction and which may have its own (deviating) requirements in that context. National DPAs also play a different role, if any, in the establishment of industry codes of conduct.

Regime	Barrier	Consequence
	HSSPs	The concept of HSSPs is interpreted in very different ways among national data protection laws, thereby rendering it inappropriate to rely on where the data controller is established in multiple jurisdictions.

Table 5 - Overview of legal barriers due to fragmentation

9.3 Uncertainty

Table 6 - Overview of legal barriers due to uncertainty

shows an overview of the barriers identified that fall under the category of *uncertainty*. They concern the barriers that create uncertainty for TDM users as to the lawfulness of their activities. It highlights the most important aspects for each legal regime discussed in the foregoing sections.

Regime	Barrier	Consequence
Copyright law	Contractual derogation	Generally, there exists great uncertainty in many Member States as to the possibilities to derogate from copyright exceptions by contract. As such contracts are able to go beyond the scope of copyright law, this is important to clarify.
	Exception for temporary reproductions: - lawful use	The concept of lawful use is not entirely clear, which creates uncertainty as to the (TDM) activities that may be covered by the exception.
	Research exception: - scientific research - non-commercial	The definition of scientific research is not clear and might need more clarification. There exists a grey area between non-commercial and commercial purposes, creating uncertainty particularly where research without any commercial aims turn out to be commercially valorised in the end.
	Extra conditions to exceptions	Some Member States are reported to apply extra conditions to all exceptions under copyright law, and sometimes database law, such as a three-step test or similar test, or the payment of any remuneration or compensation. They need to be applied in each individual case, which creates uncertainty as to the extent TDM activities are covered by the exception on which the TDM user relies.
	Quotation: - proportion	A TDM user cannot be sure that the contents to be mined constitute 'quotes', as the exact size may differ for the specific case, depending on applicable law and the size of the source or destination of the quote. The same applies to Member States where quoting of small whole works is permitted.
	Press exception: - beneficiaries - topics	The press exception generally covers traditional press bodies, but it may be uncertain whether new sorts of 'press' or (digital) journalistic activities are covered as well. Moreover, the topics covered by the exception create uncertainty as to whether, for example, a TDM user may mine newspapers in their entirety and of all timeframes or whether this must be restricted to certain

Regime	Barrier	Consequence
		sections of newspaper and to certain sections or topics within the issues.
Database law	Research exception: - lawful user	The concept of lawful user is not clear, creating uncertainty as to the actual beneficiary of the exception. Thus, a TDM is not able to rely on this exception with full - or sufficient - certainty.
Data protection law	Personal data	Although the broadness of the concept of personal data seems to make its scope clear (enough), it creates uncertainty whether and how a TDM user can anonymise his data to not infringe data protection rules.
	HSSPs	The lack of interpretation of HSSPs, either from the European legislator or the national lawmakers and courts, does not provide a solid basis to rely on for TDM users whose activities potentially fall within the ambit of this concept.

Table 6 - Overview of legal barriers due to uncertainty

10 NEW AND UPCOMING TDM EXCEPTIONS IN EUROPE

In addition to existing TDM exceptions in Member States' law, we have noted that legislative proposals are pending, or has been recently adopted, at both national and European level. They aim to reform copyright and database law to include an exception for TDM under those regimes. In this section, we shortly elaborate on these legislative activities.

10.1 European Union

In September 2016, the European Commission issued its proposal to make copyright rules fit with the Digital Single Market – hereafter referred to as the *DSM proposal*.[135] Being part of the measures that must "adapt exceptions and limitations to the digital and cross-border environment", it introduces an exception for TDM. At the time of writing, this exception is being considered and discussed by the European Parliament and by the Council of the European Union. Article 3 of the DSM proposal provides an exception to the reproduction right (under copyright law) and the extraction right (under sui generis database law) for:

- *Reproductions* and *extractions*
- Made by *research organisations*
- In order to *carry out TDM* of works or other subject-matter
- To which they have *lawful access*
- For the purposes of *scientific research*.

As with the current TDM exception in the UK, lawful access and scientific research purposes are a basic requirement in the proposed exception. Where the UK exception only covers the reproduction right under copyright law, the Commission's exception – as the French – also covers the extraction right under the database regime. In stark contrast with the UK and France, the proposed exception puts no restriction on commercial purposes, but instead, restricts the scope of beneficiaries to *research organisations*. This should also cover those organisations where they enter into public-private partnerships (PPPs).

At the time of writing, the legislative process for this TDM exception is at an early stage. Therefore, we do not go into detail about the extent to which this exception would cover TDM and to which extent it would be a barrier to TDM activities not falling under this scope. The final text is likely to change. Nevertheless, we can point out that the mandatory character of this exception – both in the ways that Member States are obliged to implement this exception and that contracts to the contrary are avoid – is an improvement in certainty, in comparison to the current exceptions in the Copyright and Database Directives that are – largely – voluntarily implemented. Further, uncertainty might arise as regards the scientific research requirement, as well as the interpretation of 'research

[135] Proposal for a directive of the European Parliament and of the Council on copyright in the Digital Single Market, COM(2016) 593 final.

organisation': can the Commission ensure that this will indeed cover the majority of research organisations, even when they participate in PPPs?[136]

10.2 France

In our previous Deliverable D3.3, we reported on the so-called *Digital Republic* ("République numérique") bill that was to reform copyright law and the sui generis database regime in France, among other laws, and included a TDM exception. At the time of writing of the updated deliverable D3.3+, the law has passed. As a result, two TDM exceptions are now part of the French Intellectual Property Code: one under the copyright regime and one under the database regime.[137] Both exceptions provide that right holders may not prohibit digital reproductions – of works or databases – to mine text and data *from*, or associated *with*, scientific literature, if the miner has lawful access to the source. Also, the mining must not pursue any commercial purpose, although their respective wordings are slightly different, and must be carried out for the purpose of *public research* (copyright exception) or *in the context of a research* (database right exception). Moreover, both exceptions provide some safeguards and opportunities with regard to conservation and communication of any reproductions made in the TDM process after the TDM activities are carried out.

In comparison to the TDM exception in the UK, its scope is both broader and narrower. The French TDM approach is broader in the sense that it addresses both copyrights and database rights. At the same time, it is also much narrower, as the French exceptions are restricted to a specific subject-matter: text and data from or associated with scientific literature. This does not only increase the *restrictiveness* of the exception. It also increases the *uncertainty* concerning the TDM users who may benefit from the exceptions, since it may be unclear when text and data is actually *associated with* scientific literature, as well as what *scientific literature* exactly comprises. Another narrowing factor in the French TDM exceptions is the addition of the adjective 'digital' to *reproductions*, which means that not *all* reproductions are covered, while creating *uncertainty* on what reproductions are actually covered.

Another important feature of the French TDM exception is the safeguards to ensure reproducibility of research. Under the copyright exception, a decree will set out ways to preserve and communicate TDM research outcomes. In a similar way, under the regime of the database right exception, a decree will assign bodies to take care of the conservation and communication of the reproductions that are the output of TDM research. An advantage of this approach is that (academic) TDM research can be verified and reproduced, while having safeguards preventing that reproductions are used for other purposes than TDM.

10.3 Germany

In January 2017, the Federal Ministry of Justice in Germany issued a draft for a bill to "Align Copyright Law with the Current Demands of the Knowledge Based Society", which also includes an exception

[136] For a short analysis on the possible beneficiaries of the proposed exception, see: Caspers, M., 2016, The Commission's proposed TDM exception: Who's to benefit? (http://www.futuretdm.eu/blog/legal-policies/the-commissions-proposed-tdm-exception-whos-to-benefit/).

[137] Respectively, Articles 122-5 (10) and 342-3 (5) of the French *Code de la propriété intellectuelle*.

for TDM.[138] In order to enable the automatic analysis of large numbers of works for scientific research", Section 60d renders it permissible to:

- Reproduce – including automatically and systematically – source material
  - In order to create, in particular through normalisation, structuring and categorisation, a corpus that can be analysed
- Make the corpus available to the public for a specifically limited circle of persons
  - For their joint scientific research
  - As well as to individual third persons for the purpose of monitoring the quality of scientific research

The exception is restricted to cases where users only pursue non-commercial purposes. Where the mining involves databases protected under copyright or sui generis database rights, the reproductions made will be considered 'normal use' of that database. The German provision also prescribes that once the TDM research is completed, the corpus and reproductions of the sources must be deleted. However, they may be send to libraries, archives, museums or educational establishment for long-term storage. Section 60g provides that contractual provisions contrary to the exception are invalid. According to Section 60h, the exception is subject to the payment of an equitable remunerations, payable through a collecting society; a remuneration right for TDM is currently not found in the TDM exceptions in the UK and France.

In resemblance to the other (proposed) TDM exceptions in EU Member States, the German provision restricts the scope to non-commercial and scientific research purposes. A remarkable difference is that it is narrowly construed to only cover reproductions that are made to create a corpus that can be analysed. It seems that this exception is written with only *text* (and not 'pure' *data*) mining in an academic research context in mind. The verifiability of research – as an important value in science – seems to be supported by the fact that a corpus created in accordance with this exception may be made available to third parties to monitor the quality of the research.

Overall, the exception seems narrow in scope, increasing the *restrictiveness* of the provision. At the same time, it uses terminology that is not common in copyright law, such as *normalisation* and *corpus*, which might add to the uncertainty of the scope of this exception.

10.4 Estonia

In the context of a copyright reform in Estonia, an exception for TDM has been proposed. The Draft Copyright and Related Rights Act of 2014 provides that it is permissible to:[139]

- reproduce and process
- an object of rights
- for "text analysis and data mining"

[138] Text of the draft proposal is available at: https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/RefE_UrhWissG.pdf;jsessionid=717C31AA231CB9A6E1BE65058D11699C.1_cid289?__blob=publicationFile&v=1.
[139] Kelli (2015), p.51.

- on the condition that the author, the name of the work and source are attributed, unless this is impossible, and
- that such mining is not carried out for commercial purposes.

This is similar to the criteria found in the UK and French TDM exceptions. TDM activities were already considered to be covered under the research exception in Estonia and the proposal of this exception serves primarily to provide legal clarity on the matter.[140] In that regard, such an exception seems to have the potential to overcome one of the barriers that we identified with respect to the research exception: legal *uncertainty* in its scope.

11 TDM IN FOREIGN REGULATION

11.1 Introduction

In this section, we discuss how foreign IP legislation deals with TDM; in this context, 'foreign' refers to non-EU and non-EEA legislation. As far as we are aware, with the exception of the Google cases in the United States, there is no specific case law addressing the lawfulness of text and data mining as such. For this reason, the focus of this section is therefore mainly on the extent written laws (may) permit TDM, but where relevant case law is analysed in this context as well. First, we examine the 'fair use' doctrine in the USA and to what extent this doctrine permits TDM activities (for certain purposes). Subsequently, we discuss the situation in China followed by a brief overview of the regimes in two Commonwealth countries, namely Australia and Canada. Finally, we shortly elaborate on the TDM exception in Japanese copyright law.

11.2 „Fair use“: United States

Under US Copyright law, the author is granted – similarly to EU copyright law – the exclusive right to authorise the making of reproductions of his or her works (see §106 of the US Copyright Act). In contrast to the EU *acquis communautaire*, there is no explicit 'making available to the public' right. However, consistent with the “views of Congress, multiple Administrations, appellate courts, and leading academic authorities, the Copyright Office [has concluded] that the exclusive rights of copyright owners set forth under 17 U.S.C. § 106 collectively meet and adequately provide the substance of the making available right”.[141] Like their European equivalents, the exclusive rights are not absolute and several exceptions to those rights exist under US copyright law. By contrast with the Copyright Directive's exhaustive list of permitted exceptions, the US Copyright Act provides for an open-ended 'fair use' exception, which was initially developed in case law and codified in §107 in 1976. It permits the fair use of a copyrighted work for purposes “such as” – hence, not limited to – criticism, comment, news reporting, teaching, scholarship, or research. Four factors are taken into account to assess a particular use under this exception:

1. *the purpose and character of the use: in particular, the commercial nature or nonprofit educational purposes may be relevant in this regard;*

[140] Kelli (2015), p.51.
[141] Pallante (2016).

2. *the nature of the copyrighted work;*
3. *the amount and substantiality of the portion used;*
4. *the effect of the use upon the potential market for or value of the work*

The fact that a work is unpublished shall not itself bar a finding of fair use if such finding is made upon consideration of all the above factors.

In two recent rulings, the Court of Appeals for the Second Circuit has expressly found that text and data mining meets the criteria of fair use and does not, therefore, amount to copyright infringement.[142] In *Authors Guild v. Google*[143] Judge Chin affirmed that Google's use of the copyrighted works in the context of its book scanning and indexing project constitutes "fair use" under copyright law. The court held that Google's digitisation of books is "highly transformative," adds value, serves several important educational purposes, and may enhance the sale of books to the benefit of copyright owners. Again, the fact that Google Books facilitates search, offering an important tool for readers, scholars, researchers, libraries and others to identify and find books, and opens up new fields of research, in particular through text mining, was put forward to demonstrate the transformative character of Google's use of the copyrighted works. On the benefits of the Google Books project for researchers and librarians, Judge Chin wrote:

> *Second, in addition to being an important reference tool, Google Books greatly promotes a type of research referred to as "data mining" or "text mining. Google Books permits humanities scholars to analyze massive amounts of data—the literary record created by a collection of tens of millions of books. Researchers can examine word frequencies, syntactic patterns, and thematic markers to consider how literary style has changed over time.. Using Google Books, for example, researchers can track the frequency of references to the United States as a single entity ("the United States is") versus references to the United States in the plural ("the United States are") and how that usage has changed over time. The ability to determine how often different words or phrases appear in books at different times "can provide insights about fields as diverse as lexicography, the evolution of grammar, collective memory, the adoption of technology, the pursuit of fame, censorship, and historical epidemiology."*[144]

This decision was confirmed by the Court of Appeals for the Second Circuit, in a judgment written by Judge Leval.[145]

11.3 Permitted uses in China

Although (the People's Republic of) China does not provide for a fair use exception as such under its copyright law, it is nevertheless worthy to mention in this context. Similar to other copyright laws, the Chinese Copyright Act grants authors, inter alia, the exclusive rights to authorise and prohibit the

[142] United States Court of Appeals for the Second Circuit, No. 13-4829-cv (*Google Books vs. Authors' Guild*) (16.10.2015); United States Court of Appeals for the Second Circuit, June 10, 2014 (*Authors' Guild of America vs. Hathitrust*), No. 4547-cv., 755 F.3d 87, 91 (2d Cir. 2014).
[143] Authors Guild, Inc. v. Google Inc., 954 F.Supp.2d 282 (N.Y.S.D. 2013).
[144] Authors Guild, Inc. v. Google Inc., 954 F.Supp.2d 282 (N.Y.S.D. 2013), at 288.
[145] Authors Guild, Inc. v. Google, Inc., 804 F.3d 202 (2nd Circuit, 2015).

reproductions and making available of works. As an exception to those rights, use of works are permitted for the user's own "private study, research or self-entertainment", which would, as discussed under similar exceptions under European national copyright laws, not leave much room for TDM. However, according to Geller and Nimmer, the China's Supreme People's Court has opened up the system of exceptions under Chinese copyright law when it issued a policy document in 2011,[146] stating that in circumstances necessary to stimulate technical innovation and commercial development, an act that would neither conflict with the normal use of the work nor unreasonably prejudice the legitimate interest of the author could be deemed 'fair use'. It provided the four factors that we discussed under the US fair use exception.[147] This policy has been confirmed in a 2014 case and, therefore, we consider the Chinese Copyright Act to be more permissive towards TDM than copyright laws in the European Member States.

11.4 „Fair dealing": Commonwealth and Canada

A common feature of copyright laws in the countries of the Commonwealth is that they generally provide for 'fair dealing' exceptions, which are narrower in scope than the concept of 'fair use' under US copyright law. They are tied to certain purposes, while the list of purposes for fair use is open-ended. Where the United Kingdom has a specific exception for TDM, besides several fair dealing exceptions, TDM in other Commonwealth nations must rely on fair dealing provisions. As under the UK Copyright, Designs and Patents Act of 1988, they generally permit the fair dealing for purposes such as criticism, private use and research.[148] They are therefore rather (too) restrictive in scope as regards TDM,[149] which is confirmed by the fact that the UK adopted a specific TDM exception. Canada is an exception to this finding, where recent Supreme Court decisions have favoured in rule of fair dealing for research purposes. In *CCH Canadian Ltd. v Law Society of Upper Canada*, a landmark case[150], the Supreme Court was asked to decide upon the application of the fair dealing defence for purposes of research and private study. The Court ruled that 'these allowable purposes should not be given a restrictive interpretation or this could result in the undue restriction of users' rights'. Therefore, TDM for research purposes may be permitted under Canadian copyright law. For other purposes, the exceptions are probably too restrictive. Bill C-11 of 2012 amended the Canadian Copyright Act, expanding the fair dealing purposes to education, parody and satire as well.[151] However, in our view, such purposes do not seem relevant to cover TDM activities conducted outside the research context.

Contrary to its Canadian counterpart, the Australian fair dealing exception has not received such a broad interpretation from the courts. In 2012, the Australian Law Reform Commission (ARLC) examined the issue of the scope of the fair dealing and other exceptions under the Australian Copyright Act. The ALRC was asked to consider whether exceptions and statutory licences in the *Copyright Act 1968* are adequate and appropriate in the digital environment and whether further exceptions should be recommended. The Final Report was tabled in Parliament on 13 February

[146] Geller & Nimmer (2015), CHI-72.
[147] Song (2011).
[148] Handke et al. (2015), p.14.
[149] Cf. Handke et al. (2015), p.14.
[150] CCH Canadian Ltd. v Law Society of Upper Canada, [2004] 1 SCR 339 [CCH] at para. 54.
[151] http://www.parl.gc.ca/HousePublications/Publication.aspx?DocId=5697419&File=24#1.

2014.[152] The report discussed at length the comparative benefits and drawbacks of introducing a fair use defence or amending the fair dealing defence. The Report contained 30 recommendations for reform, of which the key recommendation was for the introduction of a fair use exception to Australian copyright law. No change has been made in this sense so far to the Australian Copyright Act.

11.5 TDM exception in Japan

As the first country in the world, the Japanese Copyright Act was amended in 2009 to include an exception to permit TDM.[153] As of then, Article 47*septies* of the act provides as follows:

> "*For the purpose of information analysis [...] by using a computer, it shall be permissible to make recording on a memory, or to make adaptation (including a recording of a derivative work created by such adaptation), of a work, to the extent deemed necessary. However, an exception is made of database works which are made for the use by a person who makes an information analysis.*"[154]

At first glance, the exception seems rather broad in scope. First, it lacks a non-commercial requirement, as opposed to the UK TDM exception and the research exception in the Copyright Directive. Second, the information analysis does not necessarily have to be carried for (scientific) research purposes; it is not limited to any purpose for which the analysis is carried out. This fits with the purpose of the Japanese exception to boost the digital economy in Japan,[155] rather than the uptake of TDM for certain purposes such as research. Third, the concept of "information analysis" is interpreted quite broadly, meaning "to extract information, concerned with languages, sounds, images or other elements constituting such information, from many works or other much information, and to make a comparison, a classification or other statistical analysis of such information".[156]

On the other hand, some aspects seem to limit the scope of the Japanese provision. First, the permitted making of a "recording on a memory" suggests that only reproductions made in the computer's RAM are permitted, which would thereby only permit the actual analysis stage in the TDM process and not the preceding stages where information is retrieved and stored. However, this may be rather a matter of translation, since an alternative translation would be "recording on a medium". Second, the phrase "by using a computer" limits the permitted information analysis activities to those carried with a computer; however, we can currently not think of a way to carry out TDM without a computer and, therefore, this does not really seem to limit the scope in practice. On the contrary, the last phrase of the exception does have the potential to limit the range of TDM activities covered, namely it excludes TDM carried out on "database works which are made for the use by a person who makes an information analysis". The meaning of this phrase is quite uncertain,

[152] See Australian Law Reform Commission (2014).
[153] Triaille et al. (2014), pp.10–11.
[154] http://www.cric.or.jp/english/clj/cl2.html.
[155] OECD (2015), p.301.
[156] http://www.cric.or.jp/english/clj/cl2.html.

but our interpretation is that it exempts the mining of works in databases that are created specifically to use for TDM purposes.[157]

Hence, the Japanese TDM exception appears to cover a broad range of TDM exceptions, both commercial and non-commercial, and both for research and other purposes. However, there are some concepts in the exception that are not entirely clear and may cause uncertainty, especially the last phrase of the provision.

11.6 General findings

A quick look at non-EU regulation regarding intellectual property rights shows that there are legal, c.q. copyright, systems that are more permissive towards TDM, although their scope must not be overestimated. The Japanese Copyright Act explicitly permits TDM activities, regardless of its purposes, but its wording leaves some uncertainties. The concept of fair use under US, and possibly Chinese, copyright law does also leave room to allow TDM activities, but is more likely to cover TDM for non-commercial academic research purposes than TDM carried out outside the (academic) research environment for commercial purposes. The fair dealing exception in Commonwealth copyright laws suggests a flexibility similar to the fair use doctrine, but these appear more restrictive and possibly only permit TDM under very specific circumstances. However, under Canadian copyright law, the fair dealing exception for research purposes appears to leave room to permit TDM activities. In general, we can conclude similar to the European framework that much uncertainty exists on the lawfulness of TDM, especially where such activities leave the non-commercial academic environment.

[157] Cf. Triaille et al. (2014), p.11.

12 STAKEHOLDER IP AND PRIVACY POLICIES

12.1 Benchmark for policies

Identifying the barriers to TDM under the current European legal framework only gives a partial view of the situation, as the legal rules can either be strengthened or attenuated through the application of complementary policies. This section focuses on the policies developed and applied by different stakeholders in relation to TDM activities. The question central to this section is: *Do stakeholder policies help to overcome the legal barriers to TDM activities?* For the purpose of answering this question, the section is divided into two sections; the first deals with the way policies address the barriers present in the intellectual property regimes, while the second concerns policies relating to data protection rules.

Perhaps the most clearly favourable policy used to overcome the intellectual property law barriers to TDM is the requirement set forth at different levels that all publicly funded research output, e.g. publications and data, be disclosed under open access conditions. A first subsection explains how open access policies contribute to the lawful conduct of TDM activities. All intellectual property law policies, including those put forward by the European Commission, national funding agencies, institutions and libraries, publishers, and online content providers are thereafter discussed in the light of the following three elements:

- *beneficiaries*: who benefits from the policy? Which users are permitted to carry out the acts permitted?
- *object*: what (protected) subject-matter is covered for the permitted use?
- *use*: what sorts of uses are covered by the policy? Is it unrestricted, is it limited to certain acts or purposes, or is it subject to other conditions?

The policies that deal with data protection rules may, for example, consist of self-regulation, or industry or sector specific codes of conduct. They are evaluated with respect to the extent to which they help overcome the barriers raised by the data protection framework. The focus is on the restrictive and 'uncertain' rules, criteria, exceptions and concepts under data protection law.

12.2 Policies relating to IP rights

The legal barriers in copyright and database law illustrate that they restrict TDM mainly on the aspects of the

- *beneficiaries,*
- *object* (i.e. subject-matter, such as category of works, parts or whole works, etc.), and
- *uses*

they restrict. Our focus will therefore be in particular on these three aspects to evaluate to what extent they encourage or hamper TDM activities.

In general, Open Access (OA) policies and licences have the potential to open up contents for unrestricted access and re-use, thereby also permitting mining and promoting the uptake of TDM by taking away the IP related barriers. Therefore, we will first discuss the meaning of Open Access and

the extent to which it permits TDM, but also the restrictions that may still exist in licences as to the re-use of OA contents. In the following sections on IP related policies we will have a particular focus on the role of Open Access in stakeholder policies.

12.2.1 Open Access policies

In general, Open Access (OA) policies have the potential to open up information. Its general principles are to ensure free accessibility of scientific information, as well as the further distribution and archiving thereof.[158] Three important declarations relating to OA are the *Declaration of the Budapest Open Access Initiative*,[159] the *Bethesda Statement on Open Access Publishing*[160] and the *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities*.[161]

Going 'green' or 'golden'

The main routes to Open Access publishing are the *green road* and the *golden road*. Following the green road, the author of a scientific work makes the publication freely available, either concurring with the publication of that article in a non-OA (peer-reviewed) journal or – more often – after a certain embargo period. This will generally not require any additional costs for the scientific author. The golden road means that publications are published directly by an OA publisher. As a result, there is no embargo period between original publication and the free availability of the author's contribution. However, a fee that is referred to as an *article processing charge* (APC) must often be paid to the publisher, as the free availability of materials shifts the market from the subscribers to the publishing authors. Another option may be available in some cases which is often referred to as hybrid open access. This means that articles within a non-OA journal are made available under an OA-licence against an additional fee payable by the publishing author.

Creative Commons

An often-used licence to make information freely available is the Creative Commons (CC) licence,[162] of which several versions are available - one more permissive than the others. The most permissive licence is a CC-0 licence, by which the author dedicates his work to the public domain; this allows for unrestricted use and re-use of such works. A common feature of all other variants is that works published under a CC licence are free to use, distribute and reproduce, subject to possible restrictions:

- *BY*: stands for the attribution requirement, meaning that the author of the original work must always be acknowledged. This is the only requirement that is present in every CC-licence (except for CC-0).
- *SA*: stands for share alike, meaning that any work built upon the original work must be shared under the same licence.

[158] Guibault (2011), p.139.
[159] http://www.budapestopenaccessinitiative.org/.
[160] http://legacy.earlham.edu/~peters/fos/bethesda.htm.
[161] https://openaccess.mpg.de/Berlin-Declaration.
[162] https://creativecommons.org.

- *ND*: stands for no derivatives, meaning that the original work only may be further distributed without changes. This condition cannot be combined with the SA condition, since the latter is only meaningful as a condition where works are used to build upon.

- *NC*: stands for non-commercial, meaning that the permitted acts may only be carried out for non-commercial purposes.

Call for Open Science

On 5 April 2016, the Netherlands' EU Presidency hosted a conference in Amsterdam titled *Open Science – From Vision to Action*, from which a "living document" resulted, calling for action with regard to Open Science (of which Open Access is an important pillar).[163] This document, the so-called *Amsterdam Call for Action on Open Science*, is based on expert and stakeholder input during the conference and preceding meetings and reports. Among other things, it also addresses the issue of text and data mining. In particular, it urges the European Commissions to put forward it copyright reform proposals in 2016, to promote TDM not only for academic purposes, but "preferably also for societal and commercial purposes". It evidently refers to the TDM exception that was proposed by the Commission in December 2015 and calls upon the EU and national lawmakers to adopt and implement rules that make TDM easier for said purposes. If this exception would be designed accordingly, this would go beyond the non-commercial restriction of the current research exception under the Copyright Directive. In response to this document, the Association of European Research Libraries (LIBER) emphasised that "reform to allow the use of TDM for societal purposes is not 'preferable', it is essential".[164]

Moreover, the Call for Action document calls upon research funders and research organisations to encourage authors to "retain control over their research output", for example, by means of funding conditions, and on publishers to permit TDM by users who have legal access. In addition, LIBER also advocates a more proactive role of libraries in the promotion of new publishing models, "not only providing funds for APCs or as institutional publishers but actively exploring and experimenting with new disruptive publishing models that will be made possible through the opening up of the research lifecycle".[165]

The Hague Declaration

In December 2014, *The Hague Declaration* was drafted by 25 experts – researchers, publishers, lawyers and lecturers – from around the world.[166] The declaration's general aim is to "foster agreement about how to best enable access to facts, data and ideas for knowledge discovery in the Digital Age".[167] At the time of writing, it is signed by 657 individuals and 247 organisations, among which are LIBER Europe,[168] LERU,[169] Open Knowledge,[170] and Creative Commons.[171] With regard to policies, the Hague Declaration advocates that "the right to read includes the right to mine" and that

[163] http://english.eu2016.nl/documents/reports/2016/04/04/amsterdam-call-for-action-on-open-science.
[164] http://libereurope.eu/blog/2016/05/17/liber-response-amsterdam-call-action/.
[165] http://libereurope.eu/blog/2016/05/17/liber-response-amsterdam-call-action/.
[166] http://thehaguedeclaration.com/original-participants/.
[167] http://thehaguedeclaration.com/.
[168] http://libereurope.eu/.
[169] League of European Research Universities: http://www.leru.org/index.php/public/home/.
[170] https://okfn.org/.
[171] See for all signatories: http://thehaguedeclaration.com/all-signatories-organisations/.

policy makers should therefore ensure that content mining is not an infringement of copyright.[172] It recognises that open access (and data) is a "key enabler of content mining", provides that research funders should require that research publication should be made available under CC BY and research data under CC0.

Beneficiaries

OA in general, and the CC licences in particular, do not address a specific category of beneficiaries; it is open for anyone to use material that is published under CC, regardless of the CC licence variant.

Object

The CC licences can be applied to any sort of material, meaning that it is not restricted to, for example, a certain category of works or databases. Nonetheless, Creative Commons recognises that other sorts of open licences are more suitable for software.[173]

Use

The only aspect in our benchmark that may be restricted by CC licences is the permitted use. All licences authorise the use under, *inter alia*, copyright law and sui generis database law, in particular to produce, reproduce and share the original material. They also permit the circumvention of TPMs for the purpose of exercising the licensed rights.[174] We discuss for each condition - i.e. BY, ND, SA and NC - that can be included in a CC licence the possible restrictions to TDM:

- *CC-0*: as CC-0 permits further use and re-use without any restrictions, it fully permits all TDM activities.

- *BY*: we have identified the attribution requirement as a barrier in the context of the exceptions to copyright and database law, especially where this requirement is absolute. However, the CC licence only requires attribution where materials are shared. In the first three stages of TDM where reproductions are made, this requirement is therefore not a problem, since those reproductions are not - necessarily - shared. In cases where they are further shared, i.e. provided to the public,[175] than the miner would have to be aware of his obligation to acknowledge the original author. In the publication stage, where mined materials are included, attribution would be required; we do not regard this to be a barrier to TDM, as this is a general practice as regards quoting and citing in publications.

- *ND*: if the miner is not allowed to make derivatives, this could have consequences particularly for the 'creating a target dataset' stage in the TDM process, where adaptations of mined contents are possibly made. However, the ND requirement only applies to the sharing of adapted materials, which would only be a problem if the miner shared the created target dataset. As long as this happens within a collaborative research group between a very limited number of TDM users, we do not believe that this would constitute sharing, meaning that it is provided to the public.

[172] http://thehaguedeclaration.com/the-hague-declaration-on-knowledge-discovery-in-the-digital-age/.
[173] https://creativecommons.org/software/.
[174] https://creativecommons.org/licenses/by/4.0/legalcode, section 2(a)(4).
[175] https://creativecommons.org/licenses/by/4.0/legalcode, section 1(i).

- *SA*: as we identified in our previous points, adaptations are potentially made in the creation of a target dataset. If this is shared, and not restricted by an ND requirement, then this would require the miner to share the contents of this dataset, and possibly the dataset itself, under the same licence as the original material(s). We do not consider this to be a restriction to TDM, since the mining process is in itself in no way restricted.

- *NC*: the condition that is most restrictive within the context of TDM is the restriction to non-commercial uses. Non-commercial is defined as "not primarily intended for or directed towards commercial advantage or monetary compensation".[176] TDM evidently carried out for commercial purposes is therefore not permitted under CC licences that include the NC requirement, which are CC BY-NC, CC BY-NC-SA and CC BY-NC-ND. Moreover, the definition of *non-commercial* may not provide the required certainty for a TDM user to assess whether his or her activities are regarded as non-commercial or not. The definition only refers to the 'primary intention', which leaves open the issue of commercial gains that, for example, result from a - perhaps unintended - side effect. Hence, it may easily cover academic research that is only published as a journal research paper, but not exploited in any way. On the contrary TDM carried out by, for example, a consultancy company to produce commercial reports may be restricted by licences including the NC requirement.

We can conclude that CC-licences generally permit every act carried out in TDM. The only thing that the TDM user should really be aware of is any NC requirement in the licence; that miner must take a critical look at the intention of the activities.

12.2.1.1 General EU policy as regards TDM

The European Commission is aware of the barriers to TDM activities raised by copyright and has in particular expressed its concerns regarding the scientific community, citing the "lack of a clear EU provision on TDM for scientific research purposes".[177] It has therefore announced that it will consider legislative proposals in 2016 that will include a copyright exception for TDM that permits:

- public interest research organisations
- to carry out text and data mining
- of content they have lawful access to
- for scientific research purposes

It remains to be seen what the actual design of the exception in the proposal will be. At first glance, the current wording seems to allow for a broader range of TDM activities carried out within the academic environment, lacking the non-commercial requirement as provided for by the Copyright Directive.

More generally, the Council of the EU has recently published its *Conclusions on the transition towards an open science system*,[178] where it highlights that "open science entails amongst others

[176] https://creativecommons.org/licenses/by-nc/4.0/legalcode, section 1(i).
[177] http://ec.europa.eu/newsroom/dae/document.cfm?action=display&doc_id=12526, p7.
[178] http://data.consilium.europa.eu/doc/document/ST-8791-2016-INIT/en/pdf, p4.

open access to scientific publications and optimal reuse of research data, citizens science, and research integrity". It aims to promote the use of OA licences, such as Creative Commons, for scientific publications and research data. The Council also refers to the OA provisions in the Horizon 2020 framework and encourages Member States "to work with stakeholders to do the same at the national level on publicly funded research".

12.2.1.2 PSI re-use framework

Mere access to government information is a highly national matter, not harmonised by EU law. The main purpose of such access is to promote transparency of government activities. On the contrary, the *re-use* of government information is regulated and harmonised by the PSI Directive (2003/98/EC), which was revised in 2013.[179] It defines re-use as "[…] the use by persons or legal entities of documents held by public sector bodies, for commercial or non-commercial purposes other than the initial purpose within the public task for which the documents were produced […]". As a general principle, the directive obliges public sector bodies (PSBs) to allow the re-use of PSI to which access is not excluded.[180] The directive sets some obligations regarding the format and charges for the re-use of the documents,[181] and it allows PSBs to impose licence conditions on the re-use if it does not unnecessarily restrict re-use possibilities or competition.[182] The directive also sets out that its obligation "apply only insofar as they are compatible with the provisions of international agreements on the protection of intellectual property rights", and "documents for which third parties hold intellectual property rights" are excluded.[183] Nevertheless, Member States must ensure that documents in which libraries, museums and archives hold intellectual property rights "shall be re-usable for commercial or non-commercial purposes", where the re-use of these documents is allowed.[184]

As a result, the framework does not guarantee the availability of PSI, e.g. for TDM purposes; the re-use obligations only apply to PSI to which access is granted. Nonetheless, it does aim to achieve a wider range of possibilities for the re-use of PSI to "allow European companies to exploit its potential and contribute to economic growth and job creation".[185] Below, we evaluate the potential of the PSI directive to promote or restrict TDM activities, according to our benchmark.

Beneficiaries

The PSI framework does not address any specific beneficiaries; it does not distinguish between type of re-user. On the contrary, it ensures "fair, proportionate and non-discriminatory conditions for the re-use of [PSI]".[186] Therefore, it should not make a difference whether, for example, a TDM user - of any kind - or a journalist seeks to re-use PSI to which access is granted.

[179] The PSI Directive was revised by Directive 2013/37/EU.
[180] Article 3(1) in conjunction with Article 1 of the PSI Directive.
[181] Articles 5 and 6 of the PSI Directive.
[182] Article 8 of the PSI Directive.
[183] Article 1(5) & (2)(b) of the PSI Directive.
[184] Article 3(2) of the PSI Directive.
[185] Recital 5 of the PSI Directive.
[186] Recital 8 of the PSI Directive.

Object

As the framework is limited to the re-use of PSI, the subject-matter that is potentially permissible to mine is obviously limited to such information, i.e. documents held by PSBs. Documents are broadly defined to include "any content whatever its medium (written on paper or stored in electronic form or as a sound, visual or audiovisual recording) and "any part of such content".[187] Generally, the object is restricted by the fact that only documents in electronic form are useful for TDM purposes - unless the miner is willing to put effort into activities such as scanning of paper documents. Practically, the covered subject-matter can therefore be restricted in practice where documents are only available in non-electronic form. There is no obligation to disseminate documents in digital form; Article 4(1) of the PSI Directive only provides that PSBs shall make documents available through electronic means "where possible and appropriate". Although there is an obligation to provide the documents "in any pre-existing format or language, and, where possible and appropriate, in open and machine-readable format together with their metadata", this does not apply where this would require a disproportionate effort - "meaning going beyond a simple operation" - from the PSB.[188]

More specific restrictions as to the objects covered apply:[189]

- the definition of PSBs: documents held by another body are not covered;
- documents the supply of which is an activity falling outside the scope of the public task of the PSB;
- documents in which third parties' intellectual property rights are vested;
- documents where access is restricted due to protection of personal data;
- documents held by public service broadcasters;
- documents held by educational, research and cultural establishments, except for university libraries, libraries, museums and archives.

Use

There are no general restrictions as to the type of re-use. Article 3(1) of the directive explicitly provides that documents shall be re-usable for both commercial and non-commercial purposes. Nonetheless, PSBs may impose conditions applying to the permitted re-use, "where appropriate through a licence".[190] These conditions may not unnecessarily restrict possibilities for re-use or restrict competition, which may bring uncertainties with respect to how these licences may affect TDM in practice. Nonetheless, Member States must ensure that standard licences, whether or not adapted for specific licence applications, "are available in digital format and can be processed electronically".[191] This may be desirable in the context of TDM, as TDM software may automatically detect whether such information may be used for the specific purposes or not.

[187] Article 2(3) of the PSI Directive.
[188] Article 5(1) and (2) of the PSI Directive.
[189] See mainly Article 1 of the PSI Directive.
[190] Article 8(1) of the PSI Directive.
[191] Article 8(2) of the PSI Directive.

12.2.2 OA and TDM in Member States: laws and proposals

12.2.2.1 Germany & The Netherlands: OA provision

Provisions promoting Open Access to scientific publications - following the model of the *green road* - are found both in the copyright laws of Germany and The Netherlands. The German Copyright Act provides that the author of a scientific contribution to a periodical may make the manuscript thereof freely available to the public for non-commercial purposes after an embargo period of 12 months after first publication, when it is created in the context of research that is funded at least 50% by public money; the source of the original publication must be attributed.[192] No derogations to this provision are permitted to the prejudice of the author. It does not *oblige* the author to make the scientific work available; it only allows him so, thereby *indirectly* promoting the availability of OA scientific works that may be freely re-used for TDM.

As of 2015, the Dutch Copyright Act contains a similar provision.[193] The main difference is that the latter does not fix the embargo period, but rather uses the open concept of a "reasonable period", and that it is not required that at least half of the research is funded by public money. What is considered to be a reasonable period depends on the share of funding that comes from public resources and investments made by the publishers. The legislator has commented that under certain circumstances it may be reasonable that the author makes the scientific work freely available immediately after publication by the publisher. Moreover, the Dutch lawmaker also commented that the provision also aims to lead to publishers and authors agreeing on what is a reasonable period in their contractual negotiations.[194] As with its German equivalent, the Dutch provision does only allow, and not oblige, the scientific author to make his or her contribution freely available.

Beneficiaries

In the context of TDM, any category of TDM user that seeks to mine scientific publications may indirectly benefit from the German and Dutch OA provisions. Since they are addressed at the authors of such works, it is only after the specific act of making these works freely available to the public that the general public may benefit from these exceptions; there is no statutory *obligation* to do this. As with OA policies in general, anyone can benefit once the works are made available.

Object

The object of the provisions is restricted to scientific articles. The German provision requires that the scientific work is a contribution in a periodical that appears, at least, bi-annually, while its Dutch equivalent is only applicable to "short scientific works". The Dutch legislator has clarified that the latter particularly refers to scientific articles. Both provisions seem to exclude larger scientific works such as books.

Use

In both Germany and The Netherlands, there is no specific requirement as to the licence used for making the scientific contribution available. The only requirement is that it is made available for free. This leaves room for interpretation and may imply that, while access is provided on a free basis, TDM

[192] https://www.gesetze-im-internet.de/urhg/__38.html.
[193] http://wetten.overheid.nl/BWBR0001886/2015-07-01#HoofdstukIa_Artikel25fa.
[194] https://zoek.officielebekendmakingen.nl/dossier/33308/kst-33308-11?resultIndex=20&sorttype=1&sortorder=4.

acts may be restricted by the author relying on the OA provision. Commercially, the author may not have any interest in doing so, as he or she may only make the work available for non-commercial purposes; therefore, the author may not, for example, require additional fees for the authorisation of mining activities in addition to the free access.

12.2.2.2 France: *République numérique*

The *Digital Republic* ("République numérique") act that was recently adopted in France aims to achieve three main objectives:[195]

- "Wider data and knowledge dissemination"
- "Equal rights for Internet users"
- "Fraternity through an inclusive digital society"

This is to be achieved through several provisions that relate to, inter alia, public sector information (PSI), IP rights and data protection law. First, it seeks to promote a wider dissemination of PSI by extending "the scope of the administrative documents that may already be made available, and which central and local government, and public and private legal entities having a public service mandate, must voluntarily disseminate." Such a provision may enhance the possibilities of mining PSI in the future, as more content is ready available.

Second, the act endorses Open Access to scientific works by providing that the author may make his work available for non-commercial purposes after 12 months from its original publication; this is extended to 24 months for scientific work in the human and social sciences. This is similar to the OA provisions discussed for Germany and The Netherlands, although its Dutch equivalent does not fix the embargo period.

Third, the Digital Public legislation comes with a concept of "data of general interest", providing that public and private entities subsidised by public authorities fall under the scope of the open data policy. Such a provision may promote both access and re-use of data produced by such bodies.

Fourth, the act introduced a TDM exception, as discussed in Section 10.2.

12.2.3 Research funders

Funders of research have the potential to promote the uptake of TDM through the conditions they impose on beneficiaries of research grants. For example, they may oblige the researcher, consortium or other recipients of grants to make their research results – e.g. in the form of research papers – and research data publicly available under an OA licence. This section looks into the policies of research funders and assesses the extent to which they overcome the IP related barriers to TDM.

12.2.3.1 Horizon 2020

Horizon 2020 is the largest research and innovation programme of the EU. It provides a total funding of €80 billion for the public and private sector and runs from 2014 to 2020. Its "goal is to ensure Europe produces world-class science, removes barriers to innovation and makes it easier for the

[195] http://www.republique-numerique.fr/pages/in-english.

public and private sectors to work together in delivering innovation".[196] The programme's first 100 calls have resulted in 3 236 agreements being signed by the end of April 2015, receiving €5.5 billion from the EU.[197] It also promotes Open Access, both to publications and research data arising from projects funded under the programme. According to the Commission, a "[f]uller and wider access to scientific publications and data therefore helps to":[198]

- build on previous research results,
- encourage collaboration & avoid duplication of effort (greater efficiency),
- speed up innovation, and
- involve citizens and society.

It therefore argues that each beneficiary under Horizon 2020 must ensure Open Access to all scientific publications resulting from the funded research.

The rules for participation in the Horizon 2020 Programme are laid down in Regulation 1290/2013/EU (hereafter referred to as H2020 Regulation).[199] Article 43(2) of the Regulation provides that the grant agreements under this programme shall lay down terms and conditions regarding OA publishing and the reimbursement of costs relating to such publishing – which is particular relevant in the context of the golden road that often involves APCs. However, costs incurred for OA publishing upon the "completion of action" may not be dealt with in the H2020 grant agreements. This may be problematic where research results are published after research has been finalised. The Commission has acknowledged this issue in the context of post-grant OA publications in the FP7 programme and has funded a pilot project – OpenAIRE – under which support is offered for OA publication resulting from FP7 projects.[200] The obligation to provide Open Access to scientific publications is laid down in Article 29.2 of the Model Grant Agreement,[201] which obliges the beneficiary to deposit, as soon as possible, a digital version of the publication – the published version or accepted manuscript – in a repository for scientific publications; it has implemented the embargo periods of six months and twelve months from the Regulation.

A novelty in the H2020 framework is the pilot to promote OA to research data arising from H2020 funded projects. According to Article 43(2) of the H2020 Regulation, grant agreements under the programme may include provisions with respect to open access to research results, "in particular in ERC frontier research and FET (Future and Emerging Technologies) research or in other appropriate areas". This pilot is reflected in Article 29.3 of the Model Grant Agreement, for which there two options as to its content: the provision is either "Not applicable" or it may provide that the beneficiary of the grant must deposit its research data, including associated metadata, in a repository which "make[s] it possible for third parties to access, mine, exploit, reproduce and disseminate – free

[196] https://ec.europa.eu/programmes/horizon2020/en/what-horizon-2020.
[197] https://ec.europa.eu/programmes/horizon2020/en/horizon-2020-statistics.
[198] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf.
[199] http://ec.europa.eu/research/participants/data/ref/h2020/legal_basis/rules_participation/h2020-rules-participation_en.pdf.
[200] https://www.openaire.eu/postgrantoapilot.
[201] http://ec.europa.eu/research/participants/data/ref/h2020/mga/gga/h2020-mga-gga-multi_en.pdf.

of charge". If the main action of the funded project would be jeopardised by providing OA to certain parts of the research results, access may be restricted with respect to those parts; the data management plan must provide the reasons for such restrictions.

In general, the Horizon 2020 programme appears to have the potential to remove barriers to TDM with regard to scientific materials, although a distinction must be made between research publications and research data. While the framework ensures OA to scientific publications, whether or not after a certain embargo period, Open Access to the underlying data is not guaranteed (yet). As is recognised by the Commission, the provision of free and open access to research data may involve other issues, such as those related to the protection of personal data.

12.2.3.2 The Netherlands: NWO

The Netherlands Organisation for Scientific Research (NWO) funds "scientific research at public research institutions in the Netherlands, especially universities. [It] focuses on all scientific disciplines and fields of research".[202] The aim of the Dutch State Secretary for Education, Culture and Science is to have all scientific publications, funded by public resources, published under an OA licence by 2024, with an interim target of 60% by 2018. The NWO also holds the view that results – publications as well as other research output – from publicly funded research should be freely accessible worldwide.[203] To serve that purpose, the NWO applies stricter OA rules to its funded research projects as of December 2015, meaning that all publications from NWO funded research must be made freely available immediately upon publication. This can be either through the golden road or the green road, but the NWO promotes the golden road in particular. An incentive fund is set up to compensate researchers – funded by the NWO – for any APCs they must pay for publishing in OA journals. This fund does not support the model of hybrid open access.

The NWO also highlights the importance of Open Access to research data. At the same time, it recognises the challenges of such a policy, e.g. with regard to the confidentiality thereof. It has started a pilot "Data management" project to prepare for future policies on this point.

The policy of the NWO is good news for TDM users who seek to mine scientific publications, but - as with the Horizon 2020 programme - the policy regarding OA to research data is still somewhat in its infancy.

12.2.3.3 United Kingdom: Wellcome Trust

The United Kingdom based Wellcome Trust provides over £700 million in funding per year to support scientific research. It supports changes in copyright legislation to permit use by researchers of TDM technologies on research publications and research data.[204] The Wellcome Trust also requires researchers to provide Open Access to research papers, monographs and book chapters resulting from its funded research.[205] Where a fee must be paid for OA publishing, it requires that the publisher uses a CC BY licence and explicitly refers to the possibility that the materials can be freely re-used for, inter alia, "text- and data-mining purposes". To our knowledge, the Trust has no specific

[202] http://www.nwo.nl/en/about-nwo/what+does+nwo+do.
[203] http://www.nwo.nl/en/policies/open+science.
[204] http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-and-text-mining/index.htm.
[205] http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTD002766.htm.

policy on OA to research data, but requires that researchers make data "available with as few restrictions as possible".[206]

12.2.3.4 United Kingdom: RCUK

The Research Councils UK (RCUK) is a partnership between the United Kingdom's seven Research Councils that funds academic research with an annual budget of £3 billion.[207]

According to the Grant Conditions, a grant beneficiary must publish research results in accordance with the RCUK policy on Open Access. In its policy document on Open Access, the RCUK expects funded researchers to publish peer-reviewed research papers in journals that comply with its Open Access policy.[208] Such journals comply with RCUK's policy if:

- they provide "immediate and unrestricted access" under a CC BY licence on their own websites to the final versions of the published papers and allow those publications to be immediately deposited in other repositories without restriction as to re-use, or
- they consent to the deposit of accepted manuscripts in any repository without restriction to non-commercial re-use after an embargo period, not exceeding six months for STEM publishers and twelve months for publications in the arts, humanities and social sciences.

In this context, the first category of journals would be more appropriate to promote the uptake of TDM, since re-use of materials is not restricted. On the other hand, the latter category allows for publishers to restrict TDM carried out on their materials for commercial purposes.

Although the RCUK holds the view that publicly funded research data should be made openly available with as few restrictions as possible,[209] it does - to our knowledge - not impose specific requirements on funded researchers as regards Open Access to research data.

12.2.3.5 Preliminary findings

Although our sample of research funders, as of now, is rather limited in size, we can already derive some preliminary conclusions. We see that Open Access is expressly endorsed by research funders, but also note that – also in terms of policy – a distinction is made between OA to publications and OA to research data.

OA to publications

The research funders studied have clear OA obligations that ensure that published research papers derived from publicly funded research become freely available to anyone. Generally, this will also allow the re-use of those materials by miners. However, the RCUK deviates somewhat from the other cases, where it also allows research papers to be published under licences that restrict re-use that is commercial; while 'access' can be considered to be open in those cases, it is not necessarily freely re-usable.

[206] http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Guidance-for-researchers/index.htm.
[207] http://www.rcuk.ac.uk/.
[208] http://www.rcuk.ac.uk/documents/documents/TermsandConditionsofResearchCouncilfECGrants-pdf/.
[209] http://www.rcuk.ac.uk/documents/documents/rcukcommonprinciplesondatapolicy-pdf/.

OA to research data

Policies regarding OA to research data are underdeveloped in comparison with OA policies for publications. Although the Horizon 2020 is experimenting with OA obligations as regards such data, other research funders have only expressed principles and viewpoints on this aspect. To our knowledge, no obligations are implemented in concrete policies in this context. Nonetheless, we are aware that OA to research data faces challenges that are not so much present in the case of OA to publications; for example, research data may involve personal data, e.g. in the case of genetic research, interviews or other research with human subjects. It is not surprising that we often see that funders find that research data should be open as much as "possible". Future policy recommendations should take the negative consequences of OA to research into serious consideration.

12.2.4 Institution and library policies

12.2.4.1 LIBER and LERU

The Association of European Research Libraries (LIBER), representing over 400 research libraries in Europe,[210] underlines the importance of Open Access policies and that libraries play a key role in the development of a supporting infrastructure in that context and in advocating Open Science.[211] More specifically, LIBER promotes the uptake of TDM. According to LIBER, "research libraries are at the centre of the data deluge" and that the freedom to carry out TDM on this data "will maximise the return on investment of public money".[212] It points out that the current lack of legal clarity on the lawfulness of TDM impedes the TDM research in Europe; as our study also confirms, LIBER argues that the non-commercial requirement in the research exception is "impossible to clarify". LIBER therefore proposes a mandatory and non-overridable copyright exception to allow the mining ("use computers to analyse") of content that the users has legal access to, regardless of any commercial or non-commercial purposes.

Also LERU, the League of European Research Universities, underscores the legal uncertainty as regards the lawfulness of TDM and proposes that an exception should be made under copyright law for TDM, at least when carried out for research purposes.[213]

As the representing organisations of European universities and libraries are advocating Open Access and legal certainty of TDM (by means of a copyright exception), we are interested in the viewpoint of individual universities and libraries and to *if* and *how* they promote and reflect these principles. We discuss their policies in the following sections.

12.2.4.2 Dutch university libraries

According to the Association of Dutch Universities (VSNU), publicly funded research should be freely available.[214] This can be illustrated by the institutional plan of one of its members, the University of Amsterdam, which states that 'openly sharing information, sources and data contributes to the

[210] http://libereurope.eu/userlist/.
[211] http://libereurope.eu/open-access/; http://libereurope.eu/strategy/strategic-direction-1-enable-open-science/.
[212] http://libereurope.eu/wp-content/uploads/2014/11/Liber-TDM-Factsheet-v2.pdf.
[213] http://www.leru.org/files/publications/AP14_LERU_Roadmap_for_Research_data_final.pdf.
[214] http://vsnu.nl/files/documenten/Factsheets/21_Factsheet_OpenAccess.pdf.

quality of the education and research results' and that the university library makes its data, collection and publications available under OA licence where desired or appropriate.[215]

The association of Dutch libraries and Royal Library (UKB), in collaboration with the VSNU, negotiates with publishers to reach agreements on transition to Open Access. This is part of the 'big deal' negotiations on the universities' access to publisher publications that take place every three to five years.[216] Agreements on OA have been concluded already with Springer, Wiley, SAGE, and there is also an agreement 'in principle' with Elsevier.[217]

While general OA policies as such may promote the mining of scientific publications, it is worth noting that the UKB has also proposed including specific provisions on TDM in the negotiations. The exception as proposed by UKB covers the activities associated with TDM – such as the downloading, extracting, loading and mounting of data on a server – insofar as done for research purposes, with no limitations to, for example, non-commercial research. According to this proposed clause, any output of TDM may be made available to the public, as long as any extracts from the contents are limited, which appears to bear a relation to the quotation exception. A general limitation exists as to the third-party use of TDM output: no third-party "may harvest" any TDM output, regardless of whether any protected elements are present in the output. There exists no obligation on the publisher to guarantee availability of contents for TDM purposes, although the overarching agreement contains warranties as to server up-time etc. The miner is obliged to follow the instructions of the publisher, including technical and security access requirements, according to this clause. The extent to which the clause is adapted or agreed in the negotiations is not clear at the time of writing, since the negotiations where the clause is brought in are currently ongoing.

12.2.4.3 University of Oxford

The University of Oxford endorses Open Access publishing by its researchers.[218] Its Oxford University Press has been publishing OA since 2004, with the majority of the journals currently offering green and golden road OA publishing options to authors.[219] Oxford University also acknowledges that "academics, researchers, staff and students must be free to publish in the form of their choice";[220] it can therefore not force researchers to publish under an OA licence.

Nonetheless, it supports authors to publish OA, for example when required by the funding agency. It set up the Oxford Research Archive (ORA) in 2007, which serves as a single point of public access to digital copies of peer-reviewed articles by Oxford researchers. It actively harvests online repositories,

[215] http://www.uva.nl/binaries/content/assets/uva/nl/over-de-uva/uva-profiel/identiteit-en-missie/instellingsplan-2015-2020-nieuw.pdf?2870292702546.
[216] http://www.vsnu.nl/openaccess.
[217] See: http://www.vsnu.nl/nl_NL/nieuwsbericht/nieuwsbericht/243-wiley-en-nederlandse-universiteiten-sluiten-akkoord-over-open-access-en-wetenschappelijke-tijdschriften.html, http://www.vsnu.nl/nl_NL/nieuwsbericht/nieuwsbericht/218-universiteiten-en-uitgeverij-sage-bereiken-akkoord-open-access.html, http://www.vsnu.nl/nl_NL/nieuwsbericht/nieuwsbericht/241-nederlandse-universiteiten-en-elsevier-bereiken-principeakkoord-over-open-access-en-abonnementen.html.
[218] http://openaccess.ox.ac.uk/wp-uploads/2013/03/Statement-on-Open-Access-at-the-University-of-Oxford-Approved-by-Council-on-11-March-2013.pdf.
[219] https://global.oup.com/academic/open-access/?lang=en&cc=nl.
[220] http://openaccess.ox.ac.uk/wp-uploads/2013/03/Statement-on-Open-Access-at-the-University-of-Oxford-Approved-by-Council-on-11-March-2013.pdf.

indexes and local systems within the university to search for papers produced by Oxford researchers that should be submitted to the ORA.

The University of Oxford has also indicated that it encourages academics to deposit other sorts of research outputs, subject to third-party arrangement or conventions within the discipline.

12.2.4.4 HathiTrust

HathiTrust is a digital repository for digitised contents of university libraries. It is a collaboration between over a hundred universities of which most are US based, but it also includes universities from Canada and Europe. It does not guarantee any unrestricted access to or re-use of the contents of the repository. Users have to assess the applicable licence for each publication themselves.[221] It could be that further authorisation is required for the use of the materials with regard to mining activities, since the applicable licences range from public domain statements to "Available by Permission".[222] HathiTrust applies standardised labels to indicate the sort of licence applicable to each publication. However, it must be noted that copyright assessments are generally made in the light of US copyright;[223] for example, when a work is regarded to have fallen in the public domain due to the lapse of the copyright term, this might therefore be different in other jurisdictions.

Repositories like HathiTrust have the potential to promote TDM, especially in providing access through a comprehensive infrastructure. In the context of overcoming IP related barriers, copyright restrictions may still prohibit the re-use of contents in the repository for TDM purposes; if the right holder has not permitted such use, and it is neither allowed under a copyright (or database law) exception, authorisation must be requested by the miner on a case-by-case basis.

12.2.4.5 Preliminary findings

We see that the sample of research institutions and libraries studied generally highlight the importance of OA, thereby contributing to the promotion of TDM uptake. Not only the aspect of unrestricted 'access' is taken into account, but the unrestricted 're-use' as well, which is important for TDM. The Dutch case of negotiations between the universities and publisher illustrates that TDM has become part of the discussion. As non-OA publishers do not always permit TDM as a general policy, such developments are able to promote TDM, at least, in the academic environment.

12.2.5 Publisher policies

Publisher archives of published journal articles, books, and other contents can be very valuable for TDM purposes, in particular if the miner is able to crawl all contents stored therein and retrieves what is relevant for the particular purpose of the analysis. Thereby, publishers may promote the use of TDM by providing such access and allowing for the re-use of the contents, whether or not subject to certain restrictions. One way to promote such use is to permit such use based on licensing terms. Generally, publishers can allow for TDM re-use of their contents on three levels:

- *general policy*: a publisher can apply general terms and conditions for the use of contents, whether or not restricted to authorised users, in which TDM activities are explicitly allowed

[221] https://www.hathitrust.org/access_use.
[222] https://www.hathitrust.org/access_use.
[223] https://www.hathitrust.org/copyright.

for. When contents are published under an Open Access licence, such activities are permitted, subject to possible restrictions of the specific licence.

- *group licence*: the TDM user may be part of an institution, company or other entity that has concluded an agreement with one or more publishers under which TDM activities are permitted, even though TDM may not be allowed as a general policy by the publisher(s) involved.

- *individual licence*: an individual who seeks to mine publisher databases may also request permission – possibly on a case-by-case basis – to re-use the contents for TDM purposes or may have concluded an individual agreement that permits such activities.

In our view, it is evident that a general policy permitting TDM has the most potential to promote TDM, at least in the context of publisher materials, since it does not require separate negotiations and assessments to conclude agreements allowing such mining and to establish any conditions that apply to the use of contents for such purposes. Moreover, as general policies say most, or much, on publishers' attitudes towards TDM carried out by (authorised) users, we focus in particular on this category.

Table 7: Overview of publisher policies

provides an overview of a sample of publishers and their general policies towards TDM use of their materials. Note that this sample is not exhaustive and consists mainly of scholarly publishers. Also note that some of the publishers that do not permit TDM in their general policies may have open access journals in their portfolio for which TDM activities would generally be permitted. The table only includes OA publications where publishers apply OA licences as a general policy.

Publisher	TDM?	Beneficiaries	Licence/use	Conditions	Other
BioMed Central	yes	all	CC0	wherever possible, cite the source(s) of the data in a derivative work, although this is not a legal requirement	-
eLife	yes	anyone	CC BY	-	-
Elsevier	yes	academic subscribers	text mine subscribed content on ScienceDirect	non-commercial; use of API	other uses or subscribers on case-by-case basis
Emerald Group Publishing	no	-	-	-	-
Hindawi	yes	all	CC BY	-	makes available text

Publisher	TDM?	Beneficiaries	Licence/use	Conditions	Other
					corpus in ZIP file
MDPI	yes	all	CC BY	-	makes available text corpus in ZIP file
Nature Publishing Group	no	-	-	-	-
Open Library of Humanities	yes	all	CC licence of author's choice	not to our knowledge	not to our knowledge
Oxford University Press	no	-	-	-	-
PeerJ	yes	all	CC BY	reserve the right to limit access to users if servers are unable to handle the load	-
Pensoft Publishers	yes	all	CC BY	-	-
PLOS	yes	all	CC BY	-	-
Routledge	no	-	-	-	no general policy to our knowledge
Royal Society	yes	authorised users	use TDM technologies; materials may only be locally stored during lifetime of TDM project	-	-
SAGE	no	-	may not use materials to develop any database or other information, nor for internal use, or create compilations or derivative works of the materials	-	-
Springer	yes	researchers through institution's subscription	text and data mining	non-commercial; quotations in publications of TDM results restricted to	separate permission required for use of images in TDM results

Publisher	TDM?	Beneficiaries	Licence/use	Conditions	Other
				200 characters, 20 words, or 1 complete sentence; cite through DOI link to original content	
Taylor & Francis	no	-	may not use data-mining, robots, or similar data-gathering and extraction tools	-	only applies to free materials; premium contents subject to conditions in subscription agreement through institution
Ubiquity Press	yes	all	CC BY	only search and access through published interfaces	-
Wiley	yes	subscribers and other lawful users	text and data mining	non-commercial scholarly research related to specific projects; quotations in publications of TDM results restricted to 200 characters, 20 words, or 1 complete sentence; only through API	separate permission required for use of images in TDM results
Wolters Kluwer	no	-	-	-	-

Table 7: Overview of publisher policies

Some general findings are apparent from this overview. First, we see that the sample contains a relatively large proportion of OA publishers, which is not necessarily representative of the actual share of publishers in the field that publish under OA. Even though open licences are used, generally CC BY, the use for TDM may be subject to some conditions. For example, we find that some publishers require the user to only make use of APIs or other (graphical) interfaces that are made available to the miner. Thereby, the user may be restricted to the available functionalities in the API as regards the TDM possibilities, but the actual restrictions depend on the available interfaces in the specific case.

Second, as regards non-OA publishers, some permit TDM activities on their materials under certain circumstances, while others do not in their general policy; we are aware that this does not necessarily mean that, for example, in individual cases they are not open to permitting TDM on request. Where mining of publisher materials is allowed, we find that this is generally restricted with respect to the following aspects:

- *beneficiaries*: it is common, and unsurprising, that the beneficiaries are authorised or lawful users. However, some publishers restrict the group of beneficiaries even more to, for example, academic users or researchers through their institution's subscription. As is the case with Elsevier, other sorts of user may request authorisation for TDM, which is assessed on a case-by-case basis.[224] Therefore, such policies generally only permit TDM carried out within the context of academic research.

- *use*: as a main rule in our sample, the terminology used in general policies to allow mining explicitly refers to such activities as "text and data mining", "using TDM technologies" or "text mining", without making a distinction between the several sorts of acts that it may involve. In that regard, the permitted use is formulated in quite a neutral way, covering also future technologies that would fall under the definition of TDM. More restriction is found in the purposes for TDM, and in particular the non-commercial requirement that is found in all policies of the non-OA publishers in our sample that permit TDM. Another restriction that is common in these policies relates to the competitiveness of the TDM user's dataset, which may therefore often not be shared or used outside the specific project for which the materials were mined. In addition, some publishers restrict the proportion of quotations that may be used for the publication of TDM results; the publishers imposing such limitations constrain the size of quotations to either 200 characters, 20 words or one complete sentence, which may be more limited than a quotation exception under copyright law would allow for – depending on the Member State in question.

Third, it is noteworthy that general policies, where they do not permit TDM, often use wording that does not refer to TDM explicitly. Either they narrowly formulate which uses are permitted, such that TDM activities are not covered, or they prohibit acts associated with mining, such as the use of data-gathering or extraction tools, or the use of materials to create any database or compilations (which is often necessary for the analysis). Such terminology mainly affects the initial stage of TDM, where information is being crawled and scraped.

Fourth, we see that often conditions may restrict TDM activities in a more practical way, even in the case of some OA publishers. For example, in some cases materials may only be accessed and retrieved through the application program interface (API) or other (graphical) interface made available by the publisher. Depending on the actual interface, this may restrict the user's mining possibilities. However, few (OA) publishers in the sample make all their materials available in the form of a ZIP file containing the whole text corpus (in either full-text PDF or XML, which is updated on a daily basis.[225]

[224] https://www.elsevier.com/about/company-information/policies/text-and-data-mining.
[225] See for example: http://www.mdpi.com/librarians#corpus.

12.2.6 Online content providers

12.2.6.1 Introduction

This section looks into a few policies applied by larger parties who provide content on the Internet, in whatever form, on which TDM activities can be carried out. It serves to illustrate how policies can, in practice, impede TDM activities, or, on the contrary, promote it despite the presence of legal barriers resulting from intellectual property laws.

12.2.6.2 Twitter

Social networking and microblogging Internet service Twitter enables users to create tweets, i.e. short messages that the user wants to send out. Public tweets are visible for anyone, while protected tweets are only visible to the accepted followers of the user creating the tweet. Although the size of the messages is restricted to 140 characters, they may be protected under copyright law; as we discussed in Section 2, even passages of eleven words may be eligible for copyright protection. Therefore, mining tweets may fall under the scope of the exclusive rights of the right holder. The actual authors of the tweets – being 'works' - are the users who create the tweets.

According to the Twitter's Terms of Service, those users – by posting their tweets – grant Twitter a "worldwide, non-exclusive, royalty-free license (with the right to sublicense) to use, copy, reproduce, process, adapt, modify, publish, transmit, display and distribute such Content in any and all media or distribution methods (now known or later developed)".[226] Assuming that this is a licence with lawful effect, Twitter may sublicense the use, including reproductions and other acts carried out in the course of TDM, to third parties' e.g. miners. This is what Twitter does for third parties who seek to develop services using the Twitter API, through which they have access to the tweets, and user IDs and profiles. The Developer Agreement & Policy provides that users of the API may "[c]opy a reasonable amount of and display the Content".[227] There are restrictions as to the display of the content retrieved through the API, such as its modification for those purposes and the use of Twitter Marks, but no particular restriction is imposed as to the purposes of the use. Nonetheless, there are rate limits that a user of the API c.q. miner should take into account, which restrict the total amount of requests of Twitter contents.

As regards TDM, Twitter's policy seems not to be too restrictive to TDM, although miners will be restricted as to the content that can be received through the API, as well as to the rate limits.

12.2.6.3 Facebook

Online social networking service Facebook offers a Public Feed API, which provides access to user posts, which may be protected by copyright, where the privacy settings are set to 'public'.[228] This could, for example, allow a user of TDM technology to carry out a sentiment analysis on public posts. However, at the time of writing, this API is not publicly accessible; its use is restricted to several media publishers and, according to Facebook's information on the API, other users can currently not apply to use the API.[229] The scraping of Facebook contents, as well as any possible subsequent analysis thereof, is therefore highly restricted. Facebook does offer a "Keyword Insights API", which

[226] https://twitter.com/tos?lang=en.
[227] https://dev.twitter.com/overview/terms/agreement-and-policy.
[228] https://developers.facebook.com/docs/public_feed.
[229] https://developers.facebook.com/docs/public_feed.

provides the results of Facebook's own analysis of trends in the use of terms by its users, but does not allow a TDM user to mine the contents and make an own analysis.[230]

12.2.6.4 Instagram

Instagram is an online social networking service that people can use to make and share photos and videos on mobile devices. Miners may extract valuable insights from these contents, including their metadata, by using TDM technologies. To give access to public contents, Instagram provides for an API which can be used after registration and authorisation of the API user.[231] However, there are restrictions to the use of the API that are relevant in the context of using TDM technologies. In particular, the terms of use prohibit "apply[ing] computer vision technology" on the user contents without prior permission; this restricts the mining to merely other contents accompanying the photographic and video materials such as metadata.[232] Such restrictions seem to be connected with privacy and personal data protection issues, rather than copyright related issues, since use of computer vision technologies may allow one to recognise faces and to identify individuals, as well as possibly other features that relate to such individuals. The terms also restrict the storage of user contents to "the period necessary to provide your app's service"; in the context of TDM, this would probably restrict the storage period to what is necessary to carry out the analysis.

12.2.6.5 Google

Google offers services that may be valuable for third-party users of TDM technology to mine, such as its main search engine, its Google Images, Google Maps, Google Translate, Google News and Google Books. In general, Google prohibits the use of content from its services, and thereby TDM related activities, unless the user, c.q. miner, is authorised by Google or the owner of the respective contents, or when permitted by law.[233] Given the wide variety of services that it offers, additional policies and rules apply to some of them. Below, we will discuss several of those services.

Google News

The Google News service allows users to input search queries to find news items on the Internet. As such, Google is not only a miner itself, by crawling the Internet to find related contents, but also offers a service to users to use the found results. In this, Google is generally not itself be the holder of intellectual property rights in the contents it shows. The respective holders are rather the parties running the news websites from which the news items are crawled. However, further use of Google News contents is highly restricted by Google's terms and conditions, extending only to "display the content of the Service for [the user's] own personal use (i.e. non-commercial use)"; users may not "take the results from the Service and reformat and display them, or use any robot, spider, other device or manual process to monitor or copy any content from the Service".[234] Thereby, none of the acts related to TDM are allowed in the context of mining of contents provided by Google News.

Google Maps, Google Earth and Street View

The services of Google Maps, Google Earth and Street View (hereinafter referred to as Google Maps), offer contents such as street maps, satellite imagery, panoramic views of streets, but also additional

[230] https://developers.facebook.com/docs/graph-api/other-apis.
[231] https://www.instagram.com/developer/authentication/.
[232] https://www.instagram.com/about/legal/terms/api/.
[233] https://www.google.com/policies/terms/.
[234] https://www.google.com/intl/en_us/terms_google_news.html.

services that are based on those contents, such as traffic information, navigation options, and points of interests. Google's terms and conditions for these services apply to "everything you'd find in these products: map and terrain data, imagery, business listings, traffic, reviews and other related information provided by Google, its licensors, and users".[235] Generally, they allow non-commercial uses of the contents in Google Maps. An attribution requirement applies where contents are being displayed by the (re-)user of the contents. The terms also provide for a list of use cases where the use of Google Maps contents is either permitted or not; they allow use in books (limited to 5 000 copies), periodicals, as well as in reports and presentations. It prohibits the use in guidebooks, consumer goods and print advertisements.

Nonetheless, to carry out TDM, it might be necessary to request contents in bulk, which may necessitate the use of an API that is provided for Google Maps. However, the terms applicable to the use of this API prohibit to "pre-fetch, cache, index, or store any Content to be used outside the Service", except for limited amounts that are necessary for improving performance.[236] Moreover, they prohibit the mass downloading or bulk feeding of Google Maps contents. Therefore, the mining of Google Maps contents does not appear to be lawful, unless the miner is explicitly authorised by Google to do so.

Google Translate

Google offers a machine translation engine where users can submit text and that returns a translation of that text in another preferred language. We are not aware of any specific terms that apply to this service, which would mean that Google's general terms apply in this context. Consequently, the contents provided through the Google Translate service may not be re-used for TDM purposes.

However, Google Translate comes with an API that can be used to receive translations of text input. Use of the API is subject to a fee – generally to be paid per query – and comes with an own set of applicable terms and conditions. While those terms do not explicitly prohibit TDM activities or any acts associated with mining, they do prohibit use of the API "to create, train, or improve (directly or indirectly) a substantially similar product or service, including any other machine translation engine".[237] This suggests that TDM would be permitted as long as such activities are not carried out to make other translation services. The miner should be aware of the charges that a large amount of queries can amount to.

Google Books

In the context of the Google Books project,[238] Google has scanned and digitised millions of books. The Google Books service allows users to perform full text searches and to have access to these books. Nonetheless, access can be limited due to copyright restrictions. There are four types of access that Google provides to the books in its digital collection:[239]

- *Full View*: the entire book can be viewed

[235] https://www.google.com/intl/ALL/permissions/geoguidelines.html.
[236] https://developers.google.com/maps/terms.
[237] https://cloud.google.com/translate/v2/terms.
[238] https://books.google.com/intl/en/googlebooks/about/index.html.
[239] http://www.google.nl/googlebooks/library/.

- *Limited View*: a limited number of pages can be viewed
- *Snippet View*: the user can only read a few sentences that surround its search query
- *No Preview*: no access at all to the contents of the book, only information about the book can be viewed

Google does not allow full access to its entire text corpus, which does not allow miners to perform acts that are related to the initial stages of scraping contents and selecting – and transforming – those into a dataset for the purposes of analysis. Nonetheless, Google provides a service to allow users to do some sort of text mining through its "Ngram Viewer".[240] Ngram Viewer enables users to count the occurrences of the terms they put in or a certain sequence of text. That way, most of the TDM activities, and all of the copyright related acts, are carried out by Google's service, and not by the user. The results are presented to the user, who may subsequently interpret them and possibly use them for a publication of some sort.

Thus, Google Books does restrict TDM to its text corpus, but instead provides a TDM service to carry out the mining activities itself. The user is thereby limited to the functionalities of this service.

12.2.7 Findings

In general, online content providers seem rather restrictive to the re-use of contents for mining purposes, although a distinction should be made between the analysis stage in the TDM process, the preceding stages of crawling, scraping and transformation of contents, and the publication of contents. It is often the crawling and scraping that is not allowed or possible, which also renders any subsequent analysis of those contents impossible. Several reasons may underlie these restrictions:

- *Prevention of competitive products*: the content providers seek to prevent that contents are used to rebuild or otherwise create a competing product or service.
- *Maintain the stability, integrity and security of the infrastructure*: use of the infrastructure is limited by its technical capacity. A policy of restricting mass use and extraction of contents may prevent the infrastructure from being over overloaded.
- *Protection of third-party intellectual property rights*: many online services build on or make use of third-party intellectual property rights, or copyrights as the case may be, and may therefore restrict further use of those contents. This is illustrated by the fact that some policies provide that the authorisation of the (third-party) holder of the applicable rights is required for further use.
- *Protection of privacy and personal data*: in some services, personal data may be present in the contents that are provided. For example, images found on Instagram may contain identifiable persons. Instagram's restriction of computer vision technology suggests that it may want to protect the privacy interests of the identifiable persons, e.g. by face recognition. Furthermore, on social media platforms and Google Maps, large amounts of personal data in the form of texts and numbers may be present.

12.3 Data protection policies

[240] https://books.google.com/ngrams.

12.3.1 Introduction

Issues related to data protection law differ fundamentally in their nature from those relating to the identified barriers within intellectual property laws. Whereas the restrictions imposed by copyright and database law generally relate to the economic interests of the right owner, those in data protection law are rather affiliated with the privacy interests of the data subject. Moreover, the economic rights under copyright and database law can be fully transferred or licensed, allowing the transferee or licensee to (sub)license rights to authorise acts, such as those carried out in TDM. For example, if a publisher is transferred the (relevant) exclusive rights by authors, it may permit third parties to carry out the respective acts. On the contrary, the lawfulness of data processing may be much more complex to assess.

First, the data subject's consent is not the only possible legal ground for processing. Second, consent alone does not necessarily render any processing lawful. Third, besides legal grounds for processing, principles and obligations as to the transparency, integrity and safety of personal data always apply. Fourth, TDM often concerns the re-use of data, meaning that it builds on data that is already collected from the data subjects by another party. Consequently, the miner must ascertain that the purposes for which he carries out TDM are compatible with the purposes for which the personal data were collected, unless the respective processing can be considered to be carried out for historical, statistical or scientific purposes.

Hence, where authorisation may be sufficient under copyright law, data protection requires a whole set of rules to be taken into account when processing personal data; this is not different for data processing in TDM. We therefore assess (legal) policies by governments and different stakeholders with respect to how they deal with the following aspects in particular:

- collection of personal data
- legal grounds for processing
- sensitive data
- information obligations with regard to the data subjects
- the data subject's right to object and rectify

The policies that we will address in particular are the codes of conduct and other forms of self-regulation. These policies have the potential to take away uncertainties for stakeholders as regards the lawfulness of their processing activities - one of the largest barriers in data protection law we identified - and they may provide guidance as to dealing with safety and other obligations. As the Article 29 Working Party has emphasised in the context of big data developments, "complying with [the data protection] framework is a key element in creating and keeping the trust which any stakeholder needs in order to develop a stable business model that is based on the processing of such data";[241] in our view, this applies just as well to TDM activities, both in the public and private sector.

[241] Statement on Statement of the WP29 on the impact of the development of big data on the protection of individuals with regard to the processing of their personal data in the EU, 16 September 2014, p. 2.

In this section, we first discuss some findings from the national reports with regard to legal policies applied by governments that provide more (sector) specific rules regarding the processing of personal data. Subsequently, we elaborate on codes of conduct and other forms of self-regulation with regard to the processing of personal data, insofar as they may be relevant in the context of TDM.

12.3.2 Government legal policies

12.3.2.1 Bulgaria

For Bulgaria, it has been reported that a Law on Health permits the provision of personal data to third parties for the purposes of medical statistics or medical scientific research, only after anonymising the data. This raises the same uncertainty that we have identified in Section 2, namely that it may be uncertain for miners whether the data they anonymise is indeed regarded as anonymised within the meaning of data protection law.

12.3.2.2 France

As was reported by our French respondent, the French Data Protection Act contains a specific chapter on the treatment of personal data for the purpose of research, study or evaluations in the domain of health. It requires that automated processing of data is authorised by the Commission Nationale de l'Informatique et des Libertés (CNIL), although any data subject may object to such processing. Where it concerns data of deceased persons, processing is permitted unless the data subject has objected to this in writing during his or her life. As a general rule, persons whose information is collected for the purpose of medical research need to be informed in this regard. When the data is used for historical, statistical and scientific purposes, other than medical research, authorisation is required from CNIL as well, but not from the data subjects.

When TDM is carried out on medical data, which will generally constitute sensitive data, for HSSPs, this specific framework may provide more certainty with respect to the lawfulness thereof, in particular when it is approved by the CNIL, although it may be burdensome for a 'regular miner' to seek authorisation for each mining activity he or she carries out.

12.3.2.3 Iceland

We have learned from the report for Iceland that there is an Act on Scientific Research in the Health Sector, which provides rules on how to deal with health information materials.[242] For example, it provides that:

- health data from each scientific study shall be stored separately in a health database;
- linking together health data on an individual from different studies is prohibited;
- if health information materials are obtained for a specific scientific study, they may be retained no longer than necessary to complete the study, unless participants have granted consent for use in subsequent studies.

Such specific rules may promote TDM, since they may provide certainty and guidance to miners.

[242] https://eng.velferdarraduneyti.is/media/acrobat-enskar_sidur/Health-Sector-Research-Act-No-44-2014.pdf.

12.3.2.4 The Netherlands

As is reported for the Netherlands, the Dutch Civil Code contains a part on contracts for medical treatments. It provides that no permission from the patient is required if, for the purpose of statistical or scientific research in the context of public health, information concerning a patient is provided in the following cases:

- Asking permission is not reasonably possible and the research is carried out in a way that guarantees that the patient's privacy is not harmed;
- Regarding the nature and purpose of the research, the asking of permission cannot be required and the care worker has provided the information in a way that reasonably prevents the identification of individual natural persons.

It is additionally necessary that the research serves a public interest and cannot be carried out without the data concerned, and that a patient has not expressly raised objections to the provision of the information. Any provision of information according to this regulation has to be indicated in the health records. The provision is a lex specialis to the Dutch Data Protection Act, and in particular to the exception to the prohibition of processing of sensitive data.[243]

This specific legal regime for health research may open up TDM possibilities where it normally may be burdensome to obtain consent of the data subjects. However, it appears that for each case the miner must be able to prove that its TDM activities serve a public interest and cannot be carried out without use of the personal data.

12.3.2.5 Norway

The Norwegian Biotechnology Act, as reported by our Norwegian legal expert, provides for a stricter regime than general data protection as regards the use of genetic information. As a main rule, it prohibits use of genetic information outside the context of health service. For any other research on such materials, consent is required from the data subjects.

As a result, this would render re-use of genetic information obtained from a third-party collection rather difficult, if not impossible. Consent for the specific TDM purposes must already have been obtained when the data were collected from the data subjects, or the miner who seeks to mine the third-party collection must go back to every data subject affected by the mining activities; the latter would only be possible when contact information is linked to the genetic information.

12.3.3 Self-regulation

12.3.3.1 The Netherlands: Health Research

The Dutch DPA has approved the Code of Conduct for Health Research,[244] which elaborates not only on the Dutch Data Protection Act, but on the provisions in the Dutch Civil Code - as discussed under 4.3.2. - on contracts for medical treatments as well. It only applies to data to which medical

[243] https://zoek.officielebekendmakingen.nl/kst-25892-3.html.
[244] *Gedragscode Gezondheidsonderzoek*, available at: http://www.federa.org/sites/default/files/bijlagen/coreon/gedragscode_gezondheidsonderzoek.pdf.

confidentiality applies. This code does distinguish between several types of personal data for which different regimes apply:[245]

a) Personal data, for which the concept is identical to Dutch and EU data protection law.

b) Directly identifiable personal data: data that enable the researcher to directly identify the data subject by that data alone or combined with communication data.

c) Indirectly identifiable personal data: data that do not enable the researcher to directly identify the data subject, but nevertheless enable him to do so with the means available to him with no unreasonable amount of time or effort.

d) Communication data: names, initials, gender, date of birth, address, postcode, domicile, phone numbers and similar data necessary for communication and bank account numbers.

e) Coded data: data that does not contain personal data directly identifying the data subject and which is coded, allowing identification of the data subject only by intervention of the provider of the data or a third party holding the decoding key.

The code provides that *anonymous* data may be used for health research and may be stored as long as necessary for that purpose, except where the researcher carries out acts that enable him to identify the data subject. As a primary principle, the code requires that *personal* data may only be processed with the data subject's express and informed consent. There are two exceptions to that rule. The first concerns the use of indirectly identifiable data, whether or not coded, without the data subject's consent. This is permitted when requesting consent cannot be reasonably expected, e.g. because of a disproportional effort, subject to the following conditions:

- The research serves a public (health) interest, which is, in any event, the case where the research is carried out by an institution for scientific research or statistics insofar as it intends to publish the research.
- The research cannot be carried out with the data concerned.
- The data subject has not objected to such use.
- Identification through the data is reasonably prevented in the following way:
    - The researcher seeking to obtain the data concludes a written agreement with the provider thereof, that stipulates that:
        - provider must remove directly identifiable features, and, if applicable, code the data in a way that the researcher is not reasonably able to identify the data subject and that recognition is reasonably prevented.
        - The researcher may not carry out acts that would allow him to identify the data subjects.

[245] Article 1(l)-(p) *Gedragscode Gezondheidsonderzoek*.

- This code of conduct is complied with in its entirety by everyone involved in the research.

The second exception concerns the use of personal data without consent of the data subject. This is only permitted when it cannot be reasonably expected to request consent due to one of the following circumstances:

- Asking permission would cause such a burden to the data subject that permanent harm should be feared.
- The data subject has died, his/her address cannot be traced or the data subject does not respond after, at least, two notices.
- It only concerns the making of a sample and many more data subjects should be asked for permission then necessary for answering the research question, because only a small sample will be involved in the research. This case is subject to the following conditions:
  - The procedure is laid down in research protocol.
  - There is only access with and under the responsibility of the involved care provider.
  - No more data are accessed than necessary for the sample.
  - The researcher signs a confidentiality agreement.
  - After the making of the sample, permission is required for the processing of the data for the research.
  - Asking permission would not be meaningful, because the research is only in a preparatory stage. This is subject to the following conditions:
    - The researcher cannot draft a research protocol without accessing a limited amount of personal data from a limited number of data subjects.
    - These persons cannot reasonably be asked for permission, since the research for which permission is sought cannot be sufficiently determined.
    - There is only access with and under the responsibility of the involved care provider.
    - No more data are accessed than necessary for specifying the research question.
    - The purpose and time of the access of the data are laid down in writing between the researcher and the care provider and the researcher has signed a confidentiality agreement.

Additionally, these cases are all subject to the following conditions:

- The research serves a public (health) interest, which is, in any event, the case where the research is carried out by an institution for scientific research or statistics insofar as it intends to publish the research.

- It appears from the research protocol that the research is sound and meaningful, that it cannot be carried out without the data concerned and it cannot be carried out in a way that interferes less with the data subject's private sphere, e.g. through the use of anonymous data. If any personal data need to be used, use must, as much as possible, be made of indirectly identifiable data and directly identifiable data must only be used where there is no other choice.
- The data subject has not objected to such use.
- Appropriate measures are taken to guarantee the private sphere of the data subject is not disproportionately harmed.

As a further principle, the code of conduct stipulates that there should be a distinction between the communication file and the research file of which the linking takes place by an administrative number which contains no information in itself. If no such separation is made, the file must be destroyed when it is reasonably foreseeable that it will no longer be used for that research; if such separation is made, the communication file must be destroyed as soon it no longer necessary for the purpose of the research to make use of it, after which the researcher is left with a file without directly identifiable data that may be stored as long as it is reasonably foreseeable that it is necessary for the research.

12.4 Overall findings for policy barriers

Although the policies section of this deliverable is of an exploratory nature, we can already draw some general conclusions that are relevant in the context of developing a new policy framework. We made a distinction between policies related to intellectual property laws and policies dealing with data protection rules. They fundamentally differ in that IP mainly protects the economic interests of the right holder, while data protection law is more related to the (informational) privacy of the data subject. For the IP related policies, we assessed how policies dealt with IP and to what extent they permit or restrict TDM. As regards the data protection laws, we looked at legal policies (sector specific rules) dealing with the processing of certain data for certain purposes, as well as self-regulation in certain sectors.

First and generally, we see that stakeholders may apply Open Access (OA) policies to scientific research. Such policies promote TDM as the scope of beneficiaries for such policies is not restricted and all acts within TDM are covered. We see that many stakeholders apply the CC BY licence, which also permits TDM carried out for any purpose. Restrictions only exist with regard to the CC licences, and similar licences, that restrict re-use to non-commercial purposes. A general finding is that a distinction is made between OA to scientific publications and OA to research data. Where OA to scientific publications is already quite common as a policy among publishers, funders and research institutions and libraries, OA to research data is still at a somewhat embryonic stage. While the importance of OA to research data is generally emphasised by stakeholders, it is not implemented as a standard policy. This may be largely due to the fact that the nature of research data is fundamentally different from publications, which raises challenges. For example, research data may consist of personal (and sensitive) data and making them freely available may result in infringing data protection rules.

Second, in the case of publishers, we see that where publishers only publish under OA licences, they commonly use the CC BY licences. In their general policies, certain non-OA publishers permit TDM by subscribed users for non-commercial academic or research purposes, while others do not permit it at all. We acknowledge that, on a case-by-case basis, publishers may be willing to permit TDM, but general policies provide more clarity and make it less burdensome for a miner to mine contents he has lawful or authorised access to.

Third, regarding data protection rules, we can already see that sector specific regulations and policies, as reported by national respondents to the questionnaire we conducted, commonly relate to medical data. In these regulations and policies, we see that the rules regarding the processing, consent required from data subjects, as well as other obligations are more detailed than in the 'normal' data protection laws. While on the one hand detailed rules can be more restrictive to TDM, they can provide guidance - and therefore certainty - to TDM users ensuring that their TDM activities are lawful. In that regard, they can serve as an example for recommendations for policies relating to data protection law.

PART III PRACTICAL BARRIERS

13 INTRODUCTION

Part III of this deliverable is dedicated to the barriers that we have identified within the project, which do not relate to legal regulation or stakeholders' legal policies. This part provides an overview of the more 'practical barriers' that TDM users experience, meaning that societal, technical or economic factors are hindering their activities, rather than regulation, contracts or enforcement of rights. These barriers are mainly identified through our stakeholder engagement activities, such as the Knowledge Cafes, interviews and workshops organised by project partners. The following categorisation of the TDM landscape was used to design and organise these events:[246]

- **Legal and policies**
- **Education and Skills**
- **Economy and Incentives**
- **Technical and Infrastructure**

The first category of barriers is covered by Part II of this deliverable. Each of the following sections of Part III will cover one of the other categories of barrier, starting with an introduction to that category and a description of the nature of these barriers. This is followed by an overview of barriers that we have identified through our tasks in the different work packages within the FutureTDM project.

The benchmark that was introduced and used in Part II to assess the barriers related to regulation and policies will be used to assess the practical barriers as well. This benchmark distinguishes three – non-mutually exclusive – aspects in which factors or actors can impede TDM activities: uncertainty, fragmentation and restrictiveness. See Section 4 for more details on this benchmark. Every section within Part III will conclude with an analysis of all the practical barriers, similar to the one made for the legal barriers, in which they are broken down according to the benchmark that groups barriers into the categories of *uncertainty*, *fragmentation* and *restrictiveness*.

[246] See FutureTDM deliverables D2.2, D2.3 and 4.3.

14 EDUCATION & SKILLS

14.1 Introduction

This section addresses barriers that relate to skills and education among stakeholders. 'Skills and education' refers here not only to educating people in the skills required to become TDM practitioners themselves, but to all aspects of the TDM value chain affected by a lack of knowledge among key stakeholders.

In order to derive value from TDM, there must first be someone with the knowledge to recognise that TDM can be applied to a particular situation. If that person does not have the skills to carry out TDM themselves, they will require access to training, resources, or skilled experts who can assist them. They will require access to knowledge about whether their intended activity is legal. And finally, they will require a combination of data literacy and domain-specific expert knowledge to be able to derive meaningful, actionable insights from the results of TDM.

Furthermore, there are several peripheral aspects of the TDM value chain that require specialised skills and knowledge, separate from the ability to carry out TDM itself.

14.2 Lack of awareness of TDM

The most fundamental barrier to greater uptake of TDM in terms of education is varying awareness and knowledge about TDM as a technology, and the ways in which it can be applied. One learning that emerged from our Knowledge Cafes is that there are potential applications of and benefits to TDM in every subject field and sector of the economy – but these benefits can only be realised if there is awareness across all fields and sectors of how TDM may be applied. In reality, TDM is often still seen as "computer science", a "niche activity", or even a "mystical thing". In order to fully realise its benefits, stakeholders in all areas must be able to recognise situations in which TDM may be applied.

14.2.1 Variation across fields and sectors

In practice, there is significant variation in awareness and knowledge of TDM across different fields. Not all researchers or indeed businesses are aware of the potential benefits of TDM, which limits its implementation. Knowledge and awareness in this context varies significantly by sector, with greater awareness in fields that have traditionally dealt with more quantitative data.

One statistician working with TDM estimated that fewer than one in 20 researchers is carrying out any kind of TDM in their domain; despite the large potential for applications of TDM to statistics, there is low awareness in this domain about TDM and its uses. At the other end of the spectrum, a research content manager noted that TDM methods are routinely used in bioinformatics to gather information.

14.2.2 Limited spread of awareness

To date, the spread of interest in and awareness of TDM has depended in large part on individuals with a personal interest in these new technologies, who have led the uptake of TDM by example. Awareness of the benefits of TDM may spread from these individuals to others within their professional networks in a bottom-up, grassroots fashion, but the reach of such individual networks

is limited. One former social science researcher admitted they had never heard of TDM until they moved to the publishing industry. As a PhD student they had known no researchers or research groups doing TDM, and in their experience the same is true for most researchers.

There have been few top-down initiatives to broaden and increase the awareness of TDM, therefore the fragmentation of TDM awareness across fields and sectors remains. Several stakeholders have expressed a wish for national-level support for dialogue across fields and sectors to address this issue.

Without better awareness of TDM across all fields and sectors, opportunities to benefit from the use of TDM are inevitably missed. Having a high number of people skilled in the application of TDM is not enough to realise the full potential of these technologies. It is also necessary for people in specific fields and sectors to realise when the use of TDM may apply to their field, as only domain experts will have the knowledge to ask meaningful questions that can be solved with TDM technologies.

14.2.3 Need for a coordinated approach

The foundation for realising the full potential of TDM is a society that is data-literate, where people in all fields and sectors of the economy have awareness of how data analysis can benefit them. A theme that emerged particularly at our Knowledge Cafe in Portoroz was that students in all fields should be taught basic literacy in data analysis, in order to at least be aware of the kinds of situations in which technologies like TDM can be applied.

Some universities in Europe are becoming aware of this need and have begun investigating ways to incorporate data literacy into their curricula; Ghent University, for example, has begun treating data literacy as a key skill required by all PhD-level researchers. However, this is the exception rather than the rule. When we contacted European universities looking for examples of strategic integration of TDM- and data-related skills into curricula, most responses we received asked us in turn whether we could share information about progress other institutions had made, in the hope of learning from others' examples. This highlighted a need for better coordination and knowledge sharing around raising TDM awareness.

14.3 Barriers to application of TDM skills

Once a potential application of TDM is identified, people with skills necessary to carry out TDM are required to realise that application. Although off-the-shelf tools are available for applications such as harvesting common kinds of information off the web, existing TDM tools are generally not fit for more specific research purposes. One researcher working with historical documents explained how they had needed to develop entirely new tools, as well as modifying existing tools and methodologies, to meet their TDM needs. Our consultations with other TDM practitioners found that this was not an uncommon experience, and that it is further complicated by a lack of access to information about what tools and methodologies are available for users to work with.

14.3.1 High barriers to entry

The need for at least some programming knowledge to apply many TDM tools creates a high barrier to entry for new users, especially in fields that have not traditionally involved education in or knowledge of data analysis and programming.

At our Knowledge Cafe in Leiden, stakeholders discussed how in some fields – for example astronomy, particle physics, and the pharmaceutical industry – the nature and sheer volume of data that research involves necessitates skills in data analysis. In these cases both university curricula and research skills have by necessity adapted accordingly, and researchers are more likely to have the fundamental skillsets necessary to learn and apply TDM tools and technologies.

In other cases, the need for specialised skills to make use of TDM tools creates a significant barrier to uptake. The veterinary domain was given as an example of a field with high interest in the potential benefits of TDM, but in which domain experts traditionally have little education in programming or related skills required to make use of TDM technologies. One self-described "computer savvy" researcher admitted to still having difficulty working with TDM tools, and generally stakeholders are of the view that TDM tools are too technical for most people to use. This steep learning curve can put people off using TDM entirely.

14.3.2 Rapid progression of technology

The nature of cutting-edge technologies such as TDM is that they progress rapidly, and existing tools and practices are quickly replaced or become obsolete. This presents a challenge to researchers hoping to stay up-to-date with the skills and knowledge necessary for TDM, especially if TDM itself is not their primary research focus, but a tool they can apply to specific problems.

As one researcher in the biomedical field explained, "I can see [this TDM tool] makes things easier, but it's hard to learn, and will soon be obsolete. It's better to wait a year for better, simpler tools." Until tools become significantly easier to use, many others will no doubt continue to feel the same way. In the meantime, education around TDM technologies should therefore focus on principles that can apply to a wide variety of situations, rather than on the use of specific tools which risk becoming obsolete.

14.4 Access to training and support

In situations where stakeholders lack the skills to carry out TDM themselves, they have two options: They can either attempt to learn the necessary skills themselves, or turn to skilled experts to help them. However, there are limitations to each of these solutions. For those who hope to learn how to carry out TDM, it is often unclear whom (if anyone) they can turn to for advice, and expert help can be expensive and difficult to access.

14.4.1 Whom can practitioners turn to?

In multiple Knowledge Cafes, stakeholders reported that it is not clear whom they can look to for help if they do not have the necessary skills to carry out TDM themselves. Research institutions do not generally have explicit support services for TDM, and it is therefore not obvious who in an institution might have an idea of how to find or access training and other resources for TDM.

Some researchers are able to make use of informal personal networks to seek advice, but this varies on an individual basis. The consensus at the Helsinki Knowledge Cafe was that it would be more effective for institutions as a whole to develop skills and resources related to TDM, than for this burden to rest informally on individuals within an institution who have TDM-related skills.

Another option for practitioners is to look at online courses and training resources for self-directed learning, although again many practitioners said that they may not know how to identify appropriate resources.

14.4.2 The role of research libraries

In the absence of other clear options, researchers often turn to their institutional libraries for help with TDM. Research libraries undertake a variety of roles supporting research activities, and are therefore seen by many to be natural candidates for supporting TDM. Again at the Helsinki Knowledge Cafe, stakeholders discussed how libraries might help direct users to information resources, or offer advice to researchers in a secure environment.

At the moment, however, libraries' abilities to fulfil this role vary significantly from institution to institution. Some "forward thinking" institutional libraries have invested significant work in understanding the role of TDM in research, and are well-equipped to direct researchers towards learning resources. However, it was clear from discussions at the Knowledge Cafes in both Helsinki and London that many libraries would have very little idea of where to find such resources, and are unsure what their role should be in supporting TDM. Institutional librarians are not typically taught about TDM or how to support researchers interested in it.

There is a clear need for support for researchers from people knowledgeable in the TDM landscape. Multiple stakeholders suggested this may be a role for "data librarians" to be specifically trained in dealing with data and metadata, among other relevant skills.

14.4.3 Outside of research institutions

Not all of those interested in pursuing TDM belong to a research institution, or have access to a research library. While large industrial organisations may have the resources to invest in dedicated training for the use of TDM, stakeholders such as startups and citizen scientists have much more limited options and face significant barriers.

A Lloyds Bank report on digital behaviour in the UK found that small businesses are increasingly relying on friends, relatives or colleagues (53%) and online searches (42%) for digital advice, while only 3% use formal training courses.[247] Whether due to lack of budget or access to resources, many small businesses choose to rely on low cost or free resources rather than investing in digital skills.

14.4.4 Lack of information about TDM tools

Those with basic programming skills are better placed to make use of TDM technology, however documentation about existing TDM tools can be poor. This makes it difficult to understand which tool(s) might be useful for a specific application – let alone to adapt them for a related purpose. One practitioner in the health sector commented that it is not easy to discover the available tools, or even to find documentation to help explain their outputs.

Stakeholders have expressed frustration at the lack of a list or similar resource describing publicly-available TDM tools for various applications, particularly in languages other than English. Lack of such

[247] Lloyds Bank UK Business Digital Index 2016, October 2016 (retrieved 26 March 2017)

information limits TDM practitioners' ability to discover and assess which tools might be appropriate for a given data set and task.

Simpler, more user-friendly TDM tools would help to reduce this skills-based barrier to uptake of TDM – but only to an extent. The specific needs of specific use cases will almost inevitably require some tweaking of existing tools, and clear, readily accessible documentation will be required to address this issue. The issue poor documentation is discussed in further detail in Section 16.4.

14.4.5 TDM skills in the education system

Outside of the field of data science itself, specific education in TDM skills is rare even in fields that traditionally deal with large amounts of data. As described in Section 14.2, awareness and adoption of TDM in research tends to be driven by individuals with a particular interest in the technology, who go on to influence their immediate professional networks. Stakeholders reported that a significant barrier to TDM was the need to educate *all* researchers in the potential uses of TDM tools and technology – one content provider commented that at the moment, "If researchers want to learn to do TDM, they're on their own."

Some university initiatives and courses specifically teaching TDM technologies are beginning to emerge, for example at the School of Computer Science at the University of Manchester, home of the UK's National Centre for Text Mining (NaCTeM). However, while these offer specialised education in TDM as a subject in its own right, they are less useful for addressing the needs of stakeholders whose primary focus is not TDM, but for whom TDM may be a useful tool to answer questions or solve problems – for example researchers and startups who wish to employ TDM techniques.

This can be a difficult problem to address as this kind of TDM use does not fall within traditional academic field boundaries, but rather is a skill that can apply to all fields. Curriculum changes of this breadth would require collaboration and strategic goals across multiple university departments.

Stakeholders also report a skills gap between industry needs and academic training, when it comes to TDM skills. At the Knowledge Cafe in Potoroz, attendees noted that on the one hand the skills that industrial and commercial stakeholders require with respect to TDM aren't necessarily being taught at universities, and on the other, industrial and commercial stakeholders are unsure which universities to approach, or indeed which departments within a university to approach, to find researchers with skills that can help address their use cases.

One stakeholder who works with large datasets of open, user-generated data particularly emphasised the need for graduates to understand concepts rather than specific tools. In their experience, skilled graduates tend to have focussed more on web and application development rather than understanding algorithms. There is a need for greater clarity from industry about career opportunities and needs in data analytics, and for universities to respond to those needs with courses better tailored to meet them.

14.4.6 Data literacy in the education system

As discussed in Section 14.1, supporting TDM skills and education requires more than just education in skills to actually carry out TDM. One TDM practitioner who works with geospatial data lamented the fact that people with undergraduate degrees in geography rarely have what he would consider fundamental knowledge about things such as common data file types. This lack of data literacy makes the learning curve to carry out TDM that much steeper.

Several EU countries have already begun taking steps towards introducing data literacy into early education, for example including programming courses at a primary school level,[248] which TDM stakeholders believe will help support a new generation of TDM practitioners. Attendees at our Knowledge Cafe in Berlin reported that in Germany there is already a bridge between the economy and information technology, in terms of adjustments in curricula at German universities – however this connection is lacking in other fields.

Some believe that there is a need to take this even further, and introduce compulsory programming education in all advanced research programmes, to give all researchers the skillsets to apply and use TDM tools. Consensus at the Potoroz Knowledge Cafe was that at the very least, information literacy projects need to include data analysis. In the USA, there are several funding bodies investing in TDM infrastructure and education, including the National Endowment for the Humanities (NEH), the National Institutes of Health (NIH), and the Andrew W. Mellon Foundation.

14.4.7 Access to TDM experts

Rather than developing skills to carry out TDM activities themselves, another option for stakeholders is to consult experts in TDM to help address a specific application. Many stakeholders have expressed a need for access to experts in the theory, practice, and sustainability of TDM. One TDM practitioner involved in analysing scientific literature predicted that in the future researchers will be able to hire experts for TDM analysis as easily as they can hire statistical experts today. At the moment though, access to experts in TDM can be limited by a number of factors.

The first is cost. TDM and data analysis are highly sought-after skills in the commercial world, and many academics and startups cannot afford commercial rates to consult TDM experts. TDM experts can make more money in industry than academia, and many therefore choose the former over the latter.

The second is discoverability. As in many other aspects of the TDM landscape, there is little in the way of a coordinated effort or resource to help those interested in TDM connect with experts who may be able to address their needs. Again, there is a need for stakeholders to be able to discover and consult experts across fields and sectors.

The third is the skills gap between TDM experts and domain experts, which will be discussed in more detail in Section 14.6.1.

At the Helsinki Knowledge Cafe, some suggested that research libraries might become TDM practitioners themselves, and act as expert TDM partners in research projects involving the use of TDM, however, as yet no libraries have attempted to actively fill this role.

14.5 Legal knowledge about TDM

The legal concerns and barriers with respect to TDM are discussed in detail in Part I of this report. As can be seen from the length and detail of Part I, understanding the legal issues involved in carrying out TDM is not a simple matter. Stakeholders at the Leiden Knowledge Cafe gave the example of

[248] Computer programming and coding in schools - an emerging trend, March 2014 (retrieved 26 March 2017)

holders of archived materials, who have engaged full-time expert legal advice to understand the legality of access to and reuse of their materials.

This option would not be available to many stakeholders – such as in startups and academia – due to lack of resources. Even when resources are available, attendees at the Berlin Knowledge Cafe said it is sometimes difficult to find legal experts with IT knowledge and awareness who can advise on TDM activities.

14.5.1 Lack of awareness of the law

One TDM practitioner from the not-for-profit sector reported that in their experience there is very little understanding of intellectual property, licensing laws, or the legal realities of TDM. Researchers may not be aware that what seems logical or sensible to them may not necessarily be legal.

The recent introduction of a UK exception to copyright for TDM activities is thought to have helped raise awareness around this issue among stakeholders interested in TDM in the UK. However, there is still a lack of knowledge about how data protection laws apply to TDM – many assume that information available publicly online is free to use however they like, when this is not always the case.

14.5.2 Chilling effect

Where people lack knowledge or education about what is or is not legal, this lack of legal clarity has a chilling effect on the uptake of TDM. Universities and publicly-funded bodies are particularly risk-averse when it comes to the law, opting to err on the side of caution rather than risk legal battles with content owners where the legal status of their research is uncertain.

One researcher admitted they simply go ahead with TDM they are unsure about, since "...a lawyer's confirmation would take forever, and they'd probably just say no." However not all practitioners can afford to take this kind of legal risk. In general stakeholders reported that many researchers don't pursue TDM (or limit their work to open data) due to legal issues; for academics in particular, uncertainty about whether they would be able to use or publish their results is a significant barrier.

14.5.3 Limited access to legal knowledge

Where stakeholders do not have the resources to solicit expert legal advice, many again rely on informal personal networks for information; our case study of Mediately is one example of a startup relying on the legal expertise of family and friends. Those working within academic institutions may also turn to their institutional libraries for support.

The limitations to each of these approaches have already been discussed: personal networks may not include people with the necessary knowledge, and libraries' training and ability to answer legal questions related to TDM vary from institution to institution. Libraries are again seen as well-placed to support researchers, and some libraries have invested significant effort in understanding whether publishers' licences allow TDM activities; one social science researcher reported that their librarian is "well aware of the licensing landscape" around TDM. Others, however, have few resources to devote to understanding relevant issues and supporting researchers in this context.

Where there is a lack of legal resources available, researchers may end up spending large amounts of time trying to understand and negotiate licences, rather than actually carrying out research. This burden can be a major barrier for smaller project groups. As one professor in the biology domain put

it, "I can't give a PhD student this project, because they'll waste a year working on copyright." Another former academic researcher commented on how much more efficient their work became once they moved to a company with dedicated legal experts to handle the legal aspects of TDM.

14.6 Drawing meaningful, actionable insights from TDM

The ultimate benefits of TDM stem from providing insights that can drive meaningful business decisions or research. However, gaps in knowledge and awareness can limit stakeholders' ability to make full use of the results of TDM activities, and their ability to turn information into strategy. In the words of one startup that offers TDM services to other businesses, "The real barrier is that customers don't know what they want - it's like trying to sell cool bicycles when there are no roads."

14.6.1 Skills gap with domain experts

As mentioned in 14.2, only domain experts will have the knowledge to ask meaningful questions in their subject areas that can be solved with TDM technologies – and similarly, domain experts will be best placed to connect the results of TDM to actionable insights in their domains.

For this to be possible, however, domain experts and technical experts must be able to understand each other's needs and challenges.

One researcher focussing on geographic data suggested that creating overlap in both directions would help bridge this gap: for example, by giving GIS (geographic information systems) students training in data and databases, and computer science students training in mapping and coordinates. A provider of TDM and data analytics tools and services agreed that there is a need for data scientists with skills beyond the technical, for example understanding of policy objectives, or degrees in arts or humanities.

As discussed previously, this would require researchers in all fields to have a fundamental understanding of data literacy, and the principles and tools that underpin TDM analysis. Again, this is more common in some fields than others; 'data-savvy' domain experts are much rarer in less traditionally data-intensive fields such as humanities, where many researchers "live in analogue worlds".

In the strategic communication industry, for example, nearly half of stakeholders surveyed by the European Communication Monitor 2016 reported "lack of analytical skills to make sense of big data" as a major challenge, making this the most commonly cited challenge to working with big data.[249] Even in this domain where many define themselves as "information experts", there is a clear need for better data analytical skills.

Where there is a gap between the skills of TDM practitioners, and the knowledge of domain experts, there is the potential for drawing inaccurate conclusions based on misunderstandings of TDM processes, analyses and limitations.

[249] European Communication Monitor 2016 Results, July 2016 (retrieved 26 March 2017)

14.6.2 Decision-making in industry

Similarly, for industries to make full use of the potential of TDM technologies, strategic decision-makers must understand the potential benefits of TDM for their business well enough to be motivated to integrate TDM into their processes and culture.

Stakeholders at our Berlin Knowledge Cafe believed that more 'data people' are reaching positions of influence within commercial sectors, but that it will still take significant effort to effect cultural change in larger companies, and upskill employees to understand the benefits of data-driven decision making and integrating data analysis into operational activities. The Open Data Institute highlights some of the key barriers to implementing data-based decision-making and strategy:

> *First, culture change is messy and uncertain. 70% of all culture change programmes fail. Second, senior data strategists have too narrow a focus. 77% were focussed on data governance alone as their first priority. Third, data officers tend to come from backgrounds without organisational change experience. 64% of CDOs came either from IT or data.*[250]

At the moment, although 90% of ICT decision-makers believe that big data is relevant to their business, only 39% are currently implementing big data strategies.[251] Barriers to adoption cited include legacy infrastructure, data governance, and lack of skills; a majority of decision-makers admitted they do not have the skills to implement big data (66%) or to capitalise on the data they collect (69%).

Although improved profitability is the biggest perceived benefit of big data, the return on investment for implementing TDM-based strategies is not always clear. A stakeholder involved in artificial intelligence and natural language processing noted that without clear benefits, larger companies are less likely to disrupt existing data value chains, as this can be an expensive undertaking.

14.6.3 Turning research into results

One final but crucial limitation to deriving economic benefits from TDM is the barrier between academic research and commercial benefits. Stakeholders report that while TDM researchers are constantly developing cutting edge technology and algorithms, they lack the skills to convert them into products that could generate revenue and benefit others. One data analytics startup suggested a need for teaching entrepreneurship to researchers in areas such as social sciences and life sciences to help address this barrier.

This barrier is further complicated by the legal uncertainty highlighted in Section 7, with respect to when TDM research activities may benefit from certain copyright and database right exceptions. Where these exceptions are restricted to non-commercial research, it is unclear to what extent, if any, the results of research carried out under these exceptions may later be commercialised or adapted for commercial purposes. As a result, the current legal framework limits the interface between academia and commercial products and services, making it more difficult to convert TDM research into economic benefits.

---

[250] The cultural chasm: why your strategy will fail without a data-literate organisation, October 2016 (retrieved 26 March 2017)

[251] Trends vs Technologies 2016: A research report from Capita in partnership with Cisco, August 2016 (retrieved 26 March 2017)

14.7 Skills and knowledge in other parts of the TDM value chain

There are many ways in which skills and knowledge in peripheral aspects of the TDM value chain can have an effect on the uptake of TDM.

14.7.1 Infrastructure

One clear finding from two of our Knowledge Cafes (in Leiden and Portoroz) was that good infrastructure for TDM relies on having people with the skills and understanding to build and maintain that infrastructure. As discussed in Section 16, there are significant technical and infrastructure barriers to overcome to better support TDM activities, and addressing these requires specialised knowledge and skills.

It is important to recognise that developing infrastructure requires a different set of skills to TDM research and development itself. As one Knowledge Cafe attendee put it, "You don't get publications out of setting up infrastructure." Building data infrastructure is a separate career path to TDM research and development, and supporting data infrastructure requires its own set of funding mechanisms as discussed in Section 15.2.

14.7.2 Data management

Likewise, managing data in ways that support TDM requires knowledge and understanding of the issues described in Section 14.5. Stakeholders from multiple sectors reported a need to train researchers and other content owners in managing data and information. Although many research funders now mandate data management plans for storing and sharing data, most researchers who generate research data are not educated in best practices for storing and sharing data, and rules on storing and sharing data vary from institution to institution. This leads to fragmentation of data infrastructure and creates barriers to access to data - especially bulk access for TDM.

Even outside of academia, in a survey of 461 data management professionals the most cited barrier to successful management of big data within organisations was inadequate staffing or skills, cited by 40% of respondents.[252]

14.7.3 Data sharing

Stakeholders report that in some cases, there is a need for better education on the actual processes, risks and benefits of sharing data for TDM. At the moment, people sharing data and other kinds of digital content tend to think primarily of individual use, rather than of TDM. This in part contributes to poor data management with respect to TDM, as discussed above and in Section 14.7.2.

In academia, one publisher noted that older researchers in particular are conservative about allowing others to access or re-use their data, with one survey finding that almost half of researchers would not want others to use their work in TDM[253]. However a research funder suspected this conservatism may be based on misconceptions about the potential harm of releasing work under

[252] TDWI Best Practices Report: Managing Big Data, October 2013 (retrieved 26 March 2017)
[253] Taylor & Francis Open Access Survey (PDF), June 2014 (retrieved 26 March 2017)

licences that support TDM. As they explained, "TDM is not a new use of material; it's just faster than going through a million printouts with a highlighter."

There is therefore scope for better education around data sharing, both to promote better data management practices, and to encourage content owners to release more data under licences that support TDM activities.

14.7.4 Sharing of skills and best practices

As mentioned in several sections of this report, gaps and fragmentation in skills and knowledge are an issue in many areas of the TDM value chain. Coordinated approaches to support the sharing of learnings and best practices across fields, economic sectors, institutions, and even countries would help address many of these barriers, and increase the uptake and benefits of TDM.

Among other things stakeholders have suggested supporting national centres of expertise or knowledge sharing, national or international support networks, and events such as hackathons and practitioners' workshops, to help broaden the TDM community.

14.8 Benchmark

Barrier	Uncertainty	Fragmentation	Restrictiveness
Awareness of TDM and its potential	Lack of knowledge and awareness about what TDM is, and what it can do	Significant variation in awareness among sectors, with more awareness in areas that are traditionally data-driven	Lack of coordinated support and initiatives across sectors limits the spread of awareness
Lack of skills in programming and data analysis	Uncertainty about where information and knowledge about data analysis tools can be found	Lack of skills to adapt tools and processes for different purposes means tools remain isolated in one field or application, rather than finding broader use Skills are more likely to be taught in more 'technical' fields,	Complexity of tools means lack of relevant skills constitutes a high barrier to entry for new users, if tools are too technical for them to understand
Lack of training and support	Unclear whom to turn to for advice when needing expert advice or support Unclear where to find appropriate online courses or other support for self-directed learning	Access to training and support relies largely on personal networks, leading to individual pockets of support but no broader knowledge-sharing	Cost of access to training or expert support can be prohibitively expensive for some users
Lack of knowledge and support from (research) libraries	Libraries are often uncertain or unaware of how to direct researchers to relevant learning resources, and may have little understanding of	Depends entirely on the institution and internally-motivated initiatives to upskill librarians to support TDM	Libraries are often a researcher's first port of call for support; without knowledge of how to support TDM, researchers cannot get the support

	the role of TDM in research		they need from libraries
Lack of knowledge and information about TDM tools	Unclear where to find support or documentation for TDM tools	No centralised or curated resource for locating TDM tools; they are published and stored in many different places	Limited knowledge and information about TDM tools makes them difficult to use or adapt, which has a chilling effect on their use
Lack of TDM skills and data literacy in education	Industry uncertain which academic departments or researchers to turn to for TDM support Lack of clarity about TDM career opportunities	Gap between skills needed for industry and other applications of TDM, and what is taught in education systems	Lack of appropriate skills among graduates limits how much TDM can actually be carried out in practice
Lack of knowledge about legal aspects of TDM	Uncertainty on what is permitted under law has a chilling effect on the use of TDM	Access to legal knowledge varies depending on personal networks, institutional support, etc.	Lack of legal understanding and awareness prevents some people from carrying out TDM Lack of resources to pay for expert legal advice can limit practitioners' ability to carry out TDM
Lack of awareness and knowledge among strategic decision-makers	Uncertainty about benefits and return on investment means decision-makers are less likely to risk disrupting existing processes	Skills gaps between data scientists and decision-makers inhibits translation of data analysis into insight and action	Lack of understanding of the value of data, and resultant lack of cultures where data drives decision-making, limits integration of data analysis into operational activities
Lack of ability to turn research into results	Uncertainty about whether results of research can be (legally) commercialised	Gap between cutting-edge TDM technologies developed within universities and their application in industry	Lack of skills to convert research into products limits economic benefits of TDM research
Lack of knowledge about risks of data sharing	Uncertainty or ignorance of the legal and practical risks of sharing data with others limits willingness to share data		Risk-aversion in data sharing results in less data available for TDM activities

Table 8 - Overview of barriers in skills and education

15 FUNDING & INDUSTRY

15.1 Introduction

This section covers the barriers that relate to the costs and benefits of TDM activities, as well as barriers that are particular for TDM use in the commercial field, covering businesses and industries.

15.2 (Research) funding & investments

15.2.1 High costs…

Carrying out TDM demands a fairly high amount of (monetary) resources. Different stages of the TDM process require considerable investments in both obtaining sources and technology to enable mining.[254] First, data and contents as sources to be mined must be acquired. If they are freely available on the web and delivered by one provider, then there is not much time and costs involved. However, if information must be acquired from any different sources, costs are associated with the finding of information, the negotiation processes to get access and permission for use in TDM, as well as with the actual fee for obtaining that access and permission.

When those sources have been obtained, they require considerable storage capacity, as well as computing power, to manage and analyse. This technical infrastructure also comes at a price. Furthermore, TDM software needs to be executed on these computer systems, enabling the researcher to actually analyse the content (s)he has collected. Some tools are available for free, e.g. under free and/or open source licenses, while others are commercial and require investments by the TDM user; or they do not exist at all, and TDM users must spend a lot of time on developing such tools on their own.

For private companies, it is sometimes mainly a matter of discovering and understanding their own information. Many companies do not know what data they have and even have less comprehension of how to build an infrastructure that holds all the data and allows for its understanding. Integrating their own data with external sources is also very much challenging and requires investment.

15.2.2 Hence the need for research funding

From the above, it becomes obvious that, within a research project, budget needs to be reserved to carry out TDM. Just as other types of research that require, for example, expensive laboratories furnished with measurement equipment need to reserve budget for that in their grant applications, so must research using TDM as a method to create new knowledge. It is therefore not surprising that researchers have signalled that they need to spend parts of funding on infrastructure and acquiring data.[255] However, they are often denied funding, because their proposed research was not believed to be academic by research funders or it was not focused on new applications. Such TDM research is outside of the scope of funders' research agendas, which do not seem to recognise the strong

[254] FutureTDM Deliverable D2.3, p. 17.
[255] FutureTDM Deliverable D4.3, p. 18-19.

potential and benefits of the use of TDM methods for research.[256] Therefore, researchers have difficulties in finding funding to use TDM or develop TDM technologies.

15.2.3 ...and private funding

Evidently, TDM is not only a costly undertaking for (academic) researchers, but for companies and other organisations as well. If such companies are to use TDM to, for example, enhance and optimise their commercial or internal activities and processes, or support their decision taking, they need to make investments in data sources, infrastructure and software tools as well. In contrast to academic research, where academic potential does not automatically translate into funding of TDM, this is more likely in a business context: if there is money to gain (or spare), companies may be more willing to invest. Non-academic TDM practitioners have expressed their willingness to pay for quality data.[257]

On a worldwide scale, investments in TDM by private companies have been estimated to have amounted to $6.2 billion in total in 2016, with a total expenditure of $23.8 billion on the global Big Data Market.[258] Large investments are found mainly in data storage, and much less in analytics of data, illustrating "how big data and analytics are still very much in their infancy".[259] Because large investments are required, it is often the large companies who have the proper resources to carry out TDM activities.[260]

In cases where there is a lack of investment, several causes explain why this is the case. In the course of FutureTDM's stakeholder engagement and economic research activities, we have identified at least two major causes:

1. Difficulties in determining the value of TDM in a business context
2. Mismatch between expectations of TDM results and the actual benefits of TDM

Determining the value in a business context

As discussed above, TDM activities may require considerable investments in data sources and technology. In a business context, therefore, companies want to know or predict what value they can extract from TDM activities, before they seriously consider such investments. Currently, many businesses are grappling with the issue of not being able to appropriately determine the return on investment,[261] which makes investing in TDM resources a risky undertaking. In this context, a comparison has been made with marketing, where the actual advantages of investments therein cannot be measured beforehand; at best, only rough estimates can be made.[262]

In many cases, when decisions are finally taken to start a TDM pilot, these projects are mainly driven from the data itself and not connected with any specific business case or business model. The real potential of TDM for that particular business thereby remains unrevealed. This leads to disappointed

[256] FutureTDM Deliverable D2.3, p. 17-18.
[257] FutureTDM Deliverable D2.3, p. 18.
[258] FutureTDM Deliverable D5.2, p. 59.
[259] FutureTDM Deliverable D5.2, p. 36.
[260] FutureTDM Deliverable D5.2, p. 31.
[261] FutureTDM Deliverable D5.2, p. 31-32.
[262] FutureTDM Deliverable D5.2, p. 54.

decision-makers and the cessation of the TDM (pilot) project, without having ever progressed beyond the experimental stage and before any maturity level of TDM can be reached.[263]

Expectations vs. reality

The previous section showed that investments in TDM pilot projects do not always result in satisfying results for management in companies and other organisations. The fact that many TDM project never progress beyond the experimental stage, illustrates how expectations do not always meet reality. Decision makers within businesses may expect too much, or rather too *soon*, from the output of the TDM activities.[264] While the real benefits for the company will manifest in the long term, TDM activities are reported to be cut off in earlier stages based on short term expectations.

In a business context, it seems that awareness of TDM opportunities exists, but also that there is a lack of skills and knowledge among those who are making the decisions on investments in and use of TDM technology within the company, as discussed in Section 14.6. There is not always *full* awareness of the *actual* benefits of TDM for their business and the potential of data as a business asset. Also, there is a general feeling in the commercial world, that one of the main obstacles to translating data into value is actually lack of mining and analytical skills, and lack of data savvy culture in the organisations. The latter concerns barriers relating to skills and education, which are discussed in much more detail in Section 14.

15.3 Data silos

15.3.1 Internal data silos

In many cases, companies do not acquire data sources for TDM, but rather work with data they generated themselves, e.g. customer data or data on business processes. These data are not always shared within the organisation and remain within the division they are created;[265] and sometimes these data are not even reused within that division when data analysis activities are subcontracted.[266] In those cases, companies are not able to unleash the full potential of their own generated data. After all, the potential of TDM lies particularly in the volume and inclusiveness of the data, which cannot be achieved when data is not shared amongst departments. These so-called 'data silos'[267] are therefore barriers to TDM, not only for businesses, but for every organisation where data is generated, or collected, and kept only within one division. This phenomenon prevents data from being translated into actionable intelligence and companies from reaching economies of scale, as they receive only little return on investments.[268]

15.3.2 Organisational silos

Data siloing is not only a problem that manifests itself within companies; even when data is shared across the company - rather than being kept in one department - often, maximum value is still not extracted from this data.[269] But the use of data across borders and sectors of the economy may do

[263] FutureTDM Deliverable D5.2, p. 40.
[264] FutureTDM Deliverable D5.2, p. 53.
[265] FutureTDM Deliverable D5.2, p. 79, 83.
[266] SWD(2017)2 final, p. 9.
[267] FutureTDM Deliverable D5.2, p. 43, 79.
[268] FutureTDM Deliverable D5.2, p. 79.
[269] SWD(2017)2 final, p. 8.

so. For example, farming sensor data can be shared with software developers to create applications for harvesting optimisation,[270] or public transport scheduling information can be shared with route planning services to optimise traveling. Unfortunately, such opportunities are not available when companies and other organisations do not have access to such data. This limits, or even holds back, (commercial) opportunities to enable reuse of such data in downstream markets,[271] and prevents individual organisations and society from benefitting from the full potential value in data that would be unleashed by mining that data.[272]

15.3.3 Data localisation

Another problem, which results mostly from legal barriers, derives from legal uncertainty when sharing data across European borders. As identified by the European Commission, there is a perception among companies that localising services in one country is safer. This is not necessarily true; in this context, the lack of transparency of applicable rules is pointed out as a root cause for those concerns. When companies believe that they are less likely to breach any rules when they localise their data in one EU member state, data cannot flow across borders to benefit companies throughout Europe. At the same time, the organisation holding the data is not able to commercialise the data and extract maximised value out of it. In turn, this results in limited access to data and prevents potential TDM activities from being carried out.

15.3.4 Personal data

Please note that, where personal data is involved, sharing of data amongst divisions within an organisation, or sharing across companies, should be considered very carefully, in order to not breach any data protection or privacy laws or run counter to any ethical principle. Section 8 of this report provides a detailed overview of issues relating to data protection law.

[270] SWD(2017)2 final, p. 8.
[271] SWD(2017)2 final, p. 9.
[272] SWD(2017)2 final, p. 7.

15.4 Benchmark findings

Our Barrier Benchmark enables us to categorise the barriers discussed in this section. Below, Table 9 summarises the barriers and shows the nature of the barrier according to the benchmark.

Barrier	Uncertainty	Fragmentation	Restrictiveness
Lack of research funding	The lack of research funding is, at least partly, the result of lack of awareness among research funders on the potential value of TDM in academic research		Restricts resources that are necessary to carry out TDM in an academic environment
Lack of investments	Decision makers do not fully grasp the potential of TDM and do not have the patience to await positive longer-term results	TDM requires substantial investments, putting larger companies in a better position than SMEs and startups	Without proper investments, organisations are not able to carry out TDM at all
Data siloing		As data is 'siloed' within organisations, or not shared among organisations, there is a fragmented data landscape, preventing maximisation of value extraction from data	Within organisations, divisions do not have access to data held by other divisions, limiting TDM opportunities
			Companies and other organisations do not have opportunities to acquire access to data held by other companies, limiting TDM opportunities
Data localisation	Uncertainty about legal rules (covered in Part II on legal barriers)	Companies constrained by cross-border information law, creating a fragmented market in the EU	No access to data from companies in European member states held by a company in another member state

Table 9 – Overview of barriers in funding and industry

16 TECHNICAL & INFRASTRUCTURE

16.1 Introduction

The terms *Text Mining* and *Data Mining* have been independently used in previous years before being compiled into one single term. *Data Analysis* has been suggested as a wider term encompassing *Text Mining*, *Data Mining* and *Text and Data Mining*[273]. The nature of data being processed has determined the terminology used: *Text Mining* is the analysis of textual data, as well as all other forms of data converted to text (e.g. audio transcripts), while *Data Mining* started from mining databases and evolved to encompass mining all forms through which information can be transmitted: sounds, videos, images, graphs, numbers, chemical compounds, likes, clicks, etc. *Text and Data Mining* encompasses unstructured data (e.g. textual data) and structured data (e.g. numerical data). Text mining (usually) involves natural language processing techniques with the ultimate goal to turn text into structured data for further analysis.

As regards the technical and infrastructural aspects, the landscape is fragmented, which leads to barriers that are heavily intertwined and therefore, extremely difficult to be viewed strictly independently. The barriers concern

- data, content and language/knowledge resources (dictionaries, thesauri, ontologies, or other types of structured language data)
- tools and services for their processing,
- documentation of data, tools and services, and
- technical facilities and infrastructures for their deposition, storage, processing, maintenance and sharing.

There is an obvious proliferation of data globally; the same holds for TDM tools and services, which evolve in parallel with the growing data, but also for documentation instruments/facilities (standards, metadata models, licensing schemes) and for depositing, sharing and distribution facilities (repositories, aggregators, protocols and computing infrastructures).

Despite the rapid growth, the TDM field does not seem to have reached a maturity level of convergence; it is characterised by a fragmentation of sources which is explicable, as it is due to the differences attested in various domains as regards policies, procedures and technical competencies related to collection, processing and sharing of datasets. The datasets and the relevant tools are variably deposited and organised in a high number of repositories and aggregators, employing varying techniques, methods and implementation strategies. This clearly manifests the need for interoperability; additionally, TDM is also delimited by legal regulations and, in many cases, by ethical restrictions.

Discoverability, but also accessing and processing of data is closely linked to the issue of identification, which is further hampered by the emergence of dynamic data (e.g. twitter streams

[273] 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.

change every second, but so do meteorological data as well). The issue of persistent identification of datasets and tools/services is still a research issue, despite some recently emerging good practices, like handles and DOIs.

As regards efficient storage, although storage capacity has tremendously increased, the issue of persistent storage of data has not been successfully tackled. Closely connected to identification and discovery of data and tools is the extensive use of appropriate metadata; curation and maintenance of data and metadata is of utmost importance. Interoperability of metadata schemas is also an ongoing concern. Evidently, the demand for efficient dealing with the increasing amount of data has put the spotlight on a series of technical issues, which are related to all stages of the resources (data and tools/services) life-cycle, i.e.

- the collection, documentation, storage, curation and maintenance (on the providers' side),
- licensing, sharing or distribution (on the providers' or the distributors' side), and finally
- the actual use and re-use, either of data or of tools/services (on the end-users' side).

Furthermore, the case of multilingualism in Europe entails the importance of language resources and tools for all EU languages and renders the development and maintenance of related services' infrastructures indispensable.

The following sections discuss the technical and infrastructural barriers affecting all TDM entities mentioned above and depicted in Figure 4.

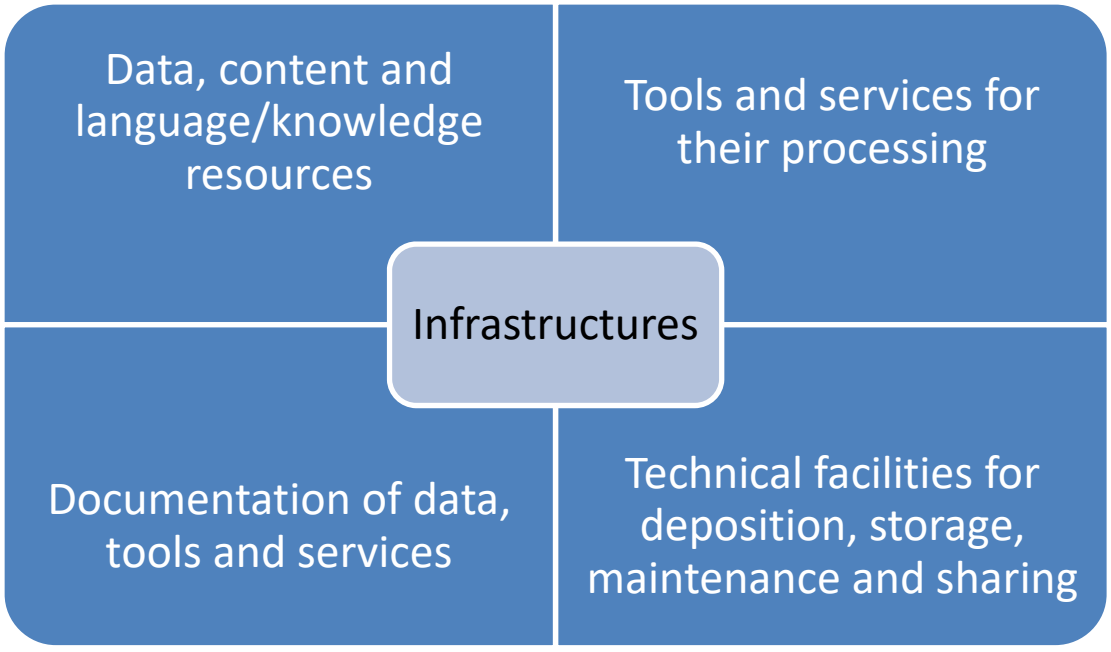

Figure 4 – Landscape of infrastructures, data, resources and technologies

16.2 Availability & Discoverability

Availability is a barrier that varies across domains, types of resources and tools and, of course, languages. Certain domains, scientific fields or languages dearly suffer from lack of digital resources and/or tools, while others have abundance thereof. A more detailed discussion on the issue of digital readiness is provided in Section 0 below.

Availability of data, however, does not entail easily discoverable data. Discoverability of data refers to the easiness with which the prospective user identifies the desired/appropriate dataset, knowledge resource or tool/service. Obviously, discoverability is a requirement not only for TDM data, tools and services, but also for data storing facilities: how does the user know where to look? How do they know how to search in order to discover the needed material or tool? How easy is it to navigate through possibly relevant data sources, registries, repositories and infrastructures in order to discover the desired one?

Discoverability is the first barrier that hinders the uptake of TDM practices in the sense that, if not catered for, it prohibits the procedure at the very initial step, namely the identification of data and relevant tools.

16.3 Accessibility

Technical

Accessibility of existing data and TDM tools has two aspects: the technical and the legal aspect. The technical aspect of the accessibility barrier refers to the degree of ease with which the user uses, accesses or downloads the desired dataset, tool or service, once identified. Barriers hindering ease of access generally relate to the user experience, i.e.

- lack of user-friendliness in the design of the TDM tool/service and the user interface,
- lack of simple procedures and clear guidelines as regards modes of use, and
- lack of appropriate documentation (as also discussed in Section 14.4 as well as further down in 16.4)

Furthermore, accessibility is burdened by user-frustrating obstacles often posed by the TDM providers, such as restrictions on the acceptable number of processes, limits on the size of data downloaded etc.

Accessibility has greatly improved since the development of facilities that collect, document, store and distribute data and related tools. These facilities constitute a valuable resource in the domain of TDM, publishing, archiving, discovery and long-term maintenance and curation of huge amounts of digital data and processing tools and services; they provide the infrastructure for describing and documenting, storing, preserving, and making this information publicly available in an open, user-friendly and trusted way.

Legal

The second aspect of accessibility, the legal one, is viewed from the perspective of legal clarity. For an in-depth analysis of the legal and policy barriers, we further refer to Part II of this report, but we highlight some barriers (again) here in the context of accessibility. It is not always clear whether specific licensing arrangements are required or whether the use of data is covered by copyright

limitations and exceptions or fair dealing provisions. These are not harmonised either at the international or at the European level, causing further uncertainty between data and tools users. The use of software under collective authorship, for example, is controlled by different laws according to the authors' nationality. In addition, accessibility is hampered in the case of commercial products which are offered either under restrictive licenses or at high cost.

The legal aspect of the accessibility barrier is also tightly connected to Open Access. The importance of mineable data and TDM tools is widely recognised in the EU, as attested by the Open Access policies such as the "PSI Directive" on the reuse of Public Sector Information,[274] as well as its updates in 2013.[275] The emphasis is on data deposition, aggregation and sharing as well as the prioritisation of services over such data by all stakeholders along the value chain. Accessibility, sharing, reuse and generation of new insights out of big volumes of data are, however, hampered by a rather vague and incomprehensive legal framework.[276] This Directive has given a tremendous impetus to the trend for open data, appropriately organised and stored and available for re-use. Open data portals are maintained by national governments[277] as well as by the EU's Open Data Portal[278]. In the EU Open Data portal, which is the single point of access to data produced by the institutions and other bodies of the European Union, the data contained is free to use, reuse, link and redistribute for commercial or non-commercial purposes.

Even in the case of access to public open data, however, legal barriers do exist. The degree of openness and the rights of use differ; users might be allowed to simply download a dataset, or access the data through an interface or even need to sign an (open) license agreement.

16.4 Documentation & Metadata Curation

Documentation includes all the necessary material providing information on the correct implementation, integration, and usage of both technologies and datasets. Problems affecting the quality of documentation are versioning and fragmentation (i.e., different versions being available in various locations). The absence of documentation or its non-systematic updating significantly hamper discoverability and accessibility, and finally, the actual use of the dataset or the tool.

The adherence to widely acknowledged standards, best practices and metadata models for the description of datasets and tools/services guarantees their accessibility, their use, re-use and re-purposing, but also the interoperability between data, between tools/services and, most crucially, between data and tools/services.

Metadata descriptions constitute the means by which data and tools producers describe their resources and users identify the resources they seek (at the level of human users). Of utmost importance is, however, the deployment of metadata descriptions by computational infrastructures hosting TDM data and tools, in order to calculate the technical compatibility between them (or the lack thereof). Crucial features for efficient metadata models are

[274] http://www.epsiplatform.eu/content/eu-psi-directive-200398ec.
[275] https://ec.europa.eu/digital-single-market/en/european-legislation-reuse-public-sector-information.
[276] See our thorough analysis of legal barriers in Part II of this report.
[277] For a list of government portals, see https://open-data.europa.eu/en/about.
[278] https://open-data.europa.eu/en/data.

- expressiveness (covering any type of resource)
- extensibility (modularity of the schema allowing for future extensions)
- semantic clarity
- flexibility (allowing exhaustive but also minimal descriptions)
- interoperability (guaranteed through mappings to widely used schemas).

16.5 Interoperability

Interoperability, which can often be a very demanding requirement, especially in the interdisciplinary framework, is the ability of multiple systems to exchange information, integrate components and use mutually intelligible data formats. It is clear that interoperability affects the whole design of a system, especially as regards the mutual integration of resources and tools between systems in view of a broader infrastructure ecosystem; the issue of interoperability is interrelated with the existence and the adoption of common standards and best practices.[279]

Registries, repositories and platforms are built on the basis of sharing resources coming from various sources and depend for this purpose on interoperability. The discovery of data, tools and services is facilitated by the metadata for their documentation maintained by *registries*; registries usually do not host the data, tools/services themselves. The hosting of actual datasets, tools and services is the task of *repositories*, while the interaction between tools/services and data resources is the task of *platforms*, which enable running of web services either on data included in the platform or uploaded by the user; see FutureTDM Deliverable D4.1, Section 3.2.2 on 'Tool hosting facilities'.[280] *Aggregators* harvest existing repositories, providing unique points of access to a great variety of repositories and their data. *Research infrastructures* provide facilities, resources and related services used by the scientific community to conduct top-level research in their respective fields, ranging from social sciences to astronomy, genomics to nanotechnologies[281]. The RIs consist of repositories and aggregators, coupled with user services for the facilitation of data identification, access and processing; crucial are the RIs' standardisation efforts concerning the adoption of common practices across scientific domains. The importance of RIs lies not only in the expertise of human resource or the technology and tools developed and used but, most notably, in the volume of data collected from experiments, measurements and observations. Research infrastructures are a treasury of data; the notion of data has evolved from referring to a means for doing research to an independent entity on its own right.

The absence of interoperability is due to (but also leads to) architectural mismatches, incompatibility, and data heterogeneity (variety of data formats). Interoperability depends on the existence and the adoption of common standards and best practices; therefore, consistency in the use of standards (preferably open) and standardisation at all levels (data sets, processing tools, metadata and also architectures of infrastructures) is highly recommended (see D4.1, Section 3.1.1 - Overview of data sources).

[279] Monachini et al. (2011).
[280] FutureTDM Deliverable D4.1.
[281] http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=what.

16.6 Language

Language is an idiosyncratic feature of textual content, the processing of which entails the necessity of available tools and supporting language data for the language in which the original content is rendered. As such, language barriers also infiltrate the TDM process in two different ways, demanding:

- the availability of language resources (datasets and tools/technologies) in a certain language or the different manifestations of language(s), i.e. sublanguages of specific domains, text types and language registers (*linguistic coverage*) and
- the need for specific language datasets in order to train or improve tools and technologies (*tool performance*)

Text mining necessitates the existence of appropriate tools and resources for each natural language. Even language-independent tools may depend on language data for their training; some of them make use of lexical/knowledge resources such as lexica, thesauri, ontologies, gazetteers etc.; others are trained on texts of a specific domain, which renders them inadequate for processing texts of a different domain. In this sense, TDM tools face all problems posed by language processing.

A large number of tools are open-source; these are, naturally, popular in the research domain. On the other hand, there are also proprietary tools, technologies or frameworks for text and data mining. Both open-source and proprietary make use of (reference) language datasets such as thesauri, terminological and lexical resources, glossaries, ontologies and language corpora, for the tasks of (for example) data indexing and classification, named entity recognition and annotation.

Digital readiness of languages

Text mining requires large amounts of written or spoken data, which, in the case of languages with relatively few speakers, it is difficult to acquire. Besides the size of the language speaking community, additional factors hampering digital readiness are the complexity of the respective language and the existence of active research in the public or private domain concerning the specific language. The notion of **digital readiness** of a language defines the degree to which a natural language is supported as regards language technology; that is, what are the available natural language processing tools, for which technologies and applications, and how many language resources exist to support technology development for this language? It is evident that European languages are far from being at the same level of 'digital readiness'; this came as the outcome of a study conducted in the framework of the META-NET project,[282] which is documented in the META-NET White Paper Series.[283]

16.7 Quality

The performance of TDM technologies is highly affected by data input. If the data are not of good quality, the tools output/performance is negatively affected. Problems arise not only from raw data, but also from data which have been inconsistently annotated or data with erroneous metadata descriptions which makes it impossible for the tools to handle them or even accept them as input.

[282] www.meta-net.eu.
[283] http://www.meta-net.eu/whitepapers/overview.

The quality issue constitutes a barrier that should be taken into account at the very early stages of the TDM process. Datasets need to be cleaned before being used. Appropriate platforms should be developed for annotating/correcting datasets. Guidelines for annotations should be strict and clear and implemented accordingly. Metadata of data and tools should be publicly shared, open to re-use and correction in order to satisfactorily depict their content and use. The principles of consistency and completeness in these processes are critical for the construction of quality resources.

Quality also relates to the use of extensive documentation and versioning of tools and services, which guarantee ease of re-use.

However, the concept of *quality* can be parameterised in relation to the notion of *purpose*: *a dataset or a tool/service is good for what? For which purpose*? What is of good quality for a specific task might not be acceptable for a different task. This dictates the measurement of quality on the basis of concrete use requirements.

16.8 Sustainability

Storing datasets, tools and services and offering them to the user community demands sustainability; this, in turn, demands long-term planning, dedication and specific mandates to the organisations undertaking the task.

The technical barriers concerning sustainability have to do with (a) the persistence of the data and tools stored and (b) the robustness of the facilities used for their storage. Factors that render the data and tools non-sustainable are:

- absence of documentation and proper metadata,
- lack of mechanisms for their long-term preservation and discoverability (such as, for example, persistent identifiers, updating and versioning)
- lack of curation and maintenance.

On the other hand, barriers for the sustainability of the facilities offering TDM datasets and tools/services relate to the rather heavy investment needed for the safeguarding of the facilities, the stored material and the users; the facilities need to invest in

- implementing explicit authentication and authorization mechanisms,
- organizing regular self-assessment cycles,
- having concrete data management plans in place, which take into account legal, privacy and ethic constraints, and last, but not least,
- planning the economic sustainability of the infrastructure.

16.9 Digital unawareness of users

A TDM user can range from a developer, i.e., a TDM programmer or a TDM-unaware person with no related skills. TDM users may belong to scientific or commercial domains which deal with high

volumes of data (raw, analysed, textual or in other media, etc.) which demand the use of TDM; see FutureTDM Deliverable D4.1, Section 3.2 on 'TDM Technology in Europe'.[284]

Some users are still unaware of the existing TDM technologies, the possibilities they offer and how to utilise them, or sometimes they find it difficult to use the unfriendly or complicated technologies/ workflows/ infrastructures.

Consequently, the combination of unskilled users with non-user-friendly procedures or with infrastructures with inadequate support for the non-expert user poses another discouraging obstacle to the uptake of TDM.

More on lack of skills and knowledge can be found in Section 14.

[284] FutureTDM Deliverable D4.1.

16.10 Overall findings

Below, Table 10 summarises the technical and infrastructure barriers that are identified throughout this section.

Barrier	Uncertainty	Fragmentation	Restrictiveness
Availability and discoverability		Some sectors or domains have abundance, while other have scarcity in data, tools or resources.	With no data, tools or resources available or discoverable, access – and thus TDM – is not possible.
Accessibility	Lack of simple procedures and clear guidelines as regards modes of use. Lack of proper documentation of TDM software. Lack of user friendliness in design of TDM software/service user interfaces.		User access limited by e.g. limited amount of downloads or downloaded data package sizes.
Quality of documentation		Different versions of documents are available at different locations.	Absence of or outdated documentation restricts use of dataset or TDM tool.
Poor curation of metadata		Results in lack of interoperability between different datasets, different tools, as well as between data and tools.	Lack of adherence to standards, best practices and metadata models result in less accessibility of data and tools for TDM.
Lack of interoperability		Architectural mismatches, incompatibility, and data heterogeneity result in lack of interoperability.	
Lack of availability of language resources		Tools trained on specific domains not adequate for different domains. Less digital readiness for smaller languages.	Limited or no training of language tools.

Poor quality of (meta)data		Dataset or tool may be of good quality for one purpose, but not necessarily for another purpose.	TDM tools may not handle or accept data. May result in lack of sustainability, because accessibility to data in the future is worsened.
Lack of or poor data management plans			Lack of sustainability of data and tools.
Inadequate or no technical or economic sustainability of infrastructure			No availability of data and tools in the long term

Table 10 – Overview of barriers in technologies and infrastructures

PART IV CONCLUSION

17 SUMMARISING AND CONCLUDING REMARKS

In this report, we identified a broad range of factors that hinder the uptake of TDM in Europe. The legal and policy barriers are identified largely through legal (desk) research and research in existing stakeholders' policies, as well as a comprehensive questionnaire conducted among legal experts in sixteen Member States. Extensive stakeholder engagement activities, such as organising workshops and stakeholder interviews, as well literature research and economic research, carried out throughout the several work packages in FutureTDM has resulted in an extensive overview of the 'practical' barriers that were identified in the course of those activities.

Legal and practical barriers

Except for the stakeholder policies, all fields of barriers are classified in three categories of barriers:

- Restrictiveness,
- Fragmentation, and
- Uncertainty

In Table 11 below, we provide a simplified and generalised overview of all sorts of barriers that have been identified within each category.

Domain	Restrictiveness	Fragmentation	Uncertainty
Legal	TDM activities under scope of exclusive rights in copyright and sui generis database regimes, due to broadly defined rights and narrowly construed exceptions.	Copyright and database right exceptions vary substantially in scope and interpretation among Member States.	Unclear rules and concepts in data protection law. Scope of exceptions uncertain under copyright and database regimes.
Education & Skills	There is a lack of TDM skills among employees and hiring external expertise can be expensive. No cultural change within organisations to upskill employees, preventing data analysis from being integrated into operational activities.	Gaps between academia and industry, as well as between sectors and domains, due to different skills, sector specific skills, or application of tools and data from one field into another.	Lack of clarity on legal rules and (legal) risks of sharing data with others. Unclear whom to turn to for expertise or advice on TDM.
Funding & Industry	Lack of funding and investments in TDM, while investments in TDM tools, data and infrastructure are necessary.	Data 'siloing' and data localisation prevent extraction of potential value from data – both internally or by third parties	Lack of awareness of – or patience for – potential of data among those investing in or funding TDM.
Technical & Infrastructure	Limited accessibility of data, tools and resources, due to poor	Data, tools and resources may be language specific and not usable for other	Unclear procedures, guidelines and documentation on TDM

	documentation, poor quality of materials or lack of language support.	languages, often with little support for smaller languages. Lack of standardisation and interoperability between datasets and tools.	tools and datasets. TDM tools lack user-friendliness in their design, making them less accessible for users.

Table 11 – Overall overview of barriers

Policy barriers

We assessed the policy barriers in a different way. Because we studied how the stakeholder policies deal with legal rights and obligations, and therefore have the potential to overcome (some of the) legal barriers, we rather evaluated them on the extent to which they allow or promote TDM. The policies were assessed on the following factors:

i. *Who*, i.e. which (sort of) TDM users, benefits (potentially) benefits from this policy?
ii. *What* sort of subject-matter (that is to be mined) is covered by this policy?
iii. Which sorts of TDM *uses* does this policy cover, and is it restricted to certain purposes or is it subject to certain conditions?

As a more general finding, we found that Open Access (OA) policies – at least under a CC-BY licence, which we found to be very common – generally:

i. Benefit every user,
ii. Can cover any sort of subject-matter, and
iii. Permit all sorts of uses and acts, carried out for any purpose with no restrictions.

Thereby, such policies allow TDM to be carried out on the subject matter licensed. In the context of scientific publishing, we noticed that OA publishing is gaining a foothold, enhancing TDM possibilities. However, not all publishers have an OA policy, so the 'TDM friendliness' of publisher policies was assessed on the individual licenses applied per publisher (in our sample). While there are publishers who permit TDM in their general terms and conditions, their conditions and beneficiaries vary; together with publishers who do not allow TDM activities on their publications, this results in a fragmented landscape of TDM licenses. This becomes problematic when a TDM user seeks to combine many different sources and needs to research many policies and, if necessary, negotiate individual TDM licenses.

In the context of data protection, the majority of sector specific regulations and policies relate to medical data. These regulations and policies are often a detailing of the general rules and principles of data protection law. This level of detail can cause them to be more restrictive to TDM, but they can also provide more guidance - and therefore certainty - to TDM users ensuring that their TDM activities are lawful.

18 NEXT STEP: A NEW POLICY FRAMEWORK

So far, this report has only identified and explained the barriers to TDM. The next step is to overcome these barriers to promote the uptake of TDM in Europe, and unleash the full potential of TDM for

research in Europe and the European economy. Work package 5 deals with actually elaborating on the ways and means to achieve this uptake. In that context, the next step will be to design a new policy framework, consisting of high-level principles and recommendations that stakeholders should take into account, if they want to contribute to the uptake of TDM in Europe.

LIST OF REFERENCES

Literature and reports

Article 29 Data Protection Working Party, 2013. *Opinion 06/2013 on open data and public sector information ('PSI') reuse, 5 June 2013*, Available at: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp207_en.pdf.

Article 29 Data Protection Working Party, 2000. *Working Document. Privacy on the Internet. An integrated EU Approach to On-line Data Protection, 21 November 2000*, Available at: http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2000/wp37en.pdf.

Australian Law Reform Commission, 2014. *Copyright and the Digital Economy (ALC Report 122)*, Available at: http://www.alrc.gov.au/publications/copyright-report-122.

Beunen, A.C., 2007. *Protection for databases. The European Database Directive and its effects in the Netherlands, France and the United Kingdom*, Nijmegen: Wolf Legal Publishers.

Burk, D.L., 2003. Anti-circumvention misuse. *IEEE Technology and Society Magazine*, 22, pp.40–47.

Eechoud, M. van, 2012. Along the Road to Uniformity – Diverse Readings of the Court of Justice Judgments on Copyright Work. *jipitec*, 3(1), pp.60–80.

Eechoud, M. van et al., 2009. *Harmonizing European copyright law: the challenges of better lawmaking* P. B. Hugenholtz, ed., Alphen aan den Rijn: Kluwer Law International.

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

European Commission, 2004. Commission Staff working paper on the review of the EC legal framework in the field of copyright and related rights, 19 July 2004. Available at: http://ec.europa.eu/internal_market/copyright/docs/review/sec-2004-995_en.pdf.

Geller, P.E. & Nimmer, M.B., 2015. *International copyright law and practice*, New York: Matthew Bender.

Guibault, L., 2011. Owning the right to open up access to scientific publications. In L. Guibault & C. Angelopoulos, eds. *Open content licensing*. Amsterdam: Amsterdam University Press, pp. 137–168.

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.

Hugenholtz, P.B., 2015. Annotatie bij Hof van Justitie van de EU 15 januari 2015 (Ryanair Ltd / PR Aviation BV) en Hoge Raad 17 januari 2014 (Ryanair Ltd / PR Aviation BV). *Nederlandse Jurisprudentie*, 2015(34/35).

Hugenholtz, P.B., 2002. De spin-off theorie uitgesponnen. *AMI*, 2002(5), pp.161–166.

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

Kelli, A., 2015. The conceptual bases for codifying Estonia's IP law and the main legislative changes: From the comparative approach to embedding drafted law into the socio-economic context. *International Comparative Jurisprudence*, 1(1), pp.44–54.

Kuner, C., 2007. *European Data Protection Law*, Oxford: Oxford University Press.

Monachini, M., Quochi, V., Calzolari, N., Budin, G., Caselli, T., Choukri, K., Francopoulo, G., Hinrichs, E., Krauwer, S., Lemnitzer, L., Mariani, J., Odijk, J., Piperidis, S., Przepiorkowski, A., Romary, L., Schmidt, H., Uszkoreit, H., and Wittenburg, P., 2011. *The standards' landscape towards an interoperability framework*, Available at: http://www.flarenet.eu/sites/default/files/FLaReNet_Standards_Landscape.pdf.

OECD, 2015. *Data-Driven Innovation: Big Data for Growth and Well-Being*, Available at: http://www.oecd-ilibrary.org/content/book/9789264229358-en.

Pallante, M., 2016. *The making available right in the united states: A report from the Register of Copyrights*, Available at: http://www.copyright.gov/docs/making_available/making-available-right.pdf.

Rosati, E., 2013. *Originality in EU Copyright: Full Harmonisation Through Case Law*, Edward Elgar Publishing.

Song, S.H., 2011. Reevaluating fair use in china - A comparative copyright analysis of chinese fair use legislation, the U.S. fair use doctrine, and the European fair dealing model. *IDEA: The Intellectual Property Law Review*, 51(3).

Spoor, J.H., Verkade, D.W.F. and Visser, D.J.G., 2005. *Auteursrecht. Auteursrecht, naburige rechten en databankenrecht*, Deventer: Kluwer.

Triaille, J.-P., D'Argenteuil, J. de M. & Francquen, A. de, 2014. *Study on the legal framework of text and data mining (TDM)*, Available at: http://ec.europa.eu/internal_market/copyright/docs/studies/1403_study2_en.pdf.

Westkamp, G., 2007. *The Implementation of Directive 2001/29/EC in the Member States, Part II*, Available at: http://bit.ly/1LZzoui.

World Intellectual Property Organization, 2012. *International Survey on Private Copying: Law & Practice 2015*, Available at: http://www.wipo.int/export/sites/www/freepublications/en/copyright/1037/wipo_pub_1037.pdf.

FutureTDM Deliverables

All our public deliverables are available at our Knowledge Libarary on the FutureTDM platform: http://www.futuretdm.eu/knowledge-library/.

Frew, H., 2015. *Stakeholder involvement roadmap and engagement strategy*, FutureTDM Deliverable D2.2.

Boom, F. van den, 2016. *Report on stakeholder mobilisation and perceptions*, FutureTDM Deliverable D2.3.

Eskevich, M. and Bosch, A. van den, 2016. *Research report on TDM landscape in Europe*, FutureTDM Deliverable D3.1.

Caspers, M. and Guibault, L., 2016. *Baseline report of policies and barriers of TDM in Europe*, FutureTDM Deliverable D3.3.

Piperidis, S., Pouli, K., Gavriilidou, M., Galanis, D. and Bakagianni, J., 2016. *European landscape of TDM applications report*, FutureTDM Deliverable D4.1.

Boom, F. van den, 2016. *Compendium of best practices and methodologies*, FutureTDM Deliverable D4.3.

Caspers, M. and Guibault, L., 2016. *FutureTDM policy framework*, FutureTDM Deliverable D5.1.