Deliverable 5.1 FutureTDM policy framework

Caspers, M.; Guibault, L.

Publication date
2016
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 5.1

FutureTDM policy framework

Project

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

Consortium:

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

Stichting LIBER, The Netherlands, (LIBER)

Open Knowledge, UK, (OK/CM)

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

The British Library, UK, (BL)

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

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

Ubiquity Press Limited, UK, (UP)

Fundacja Projekt: Polska, Poland, (FPP)

Deliverable

Number:	**D5.1**
Title:	**FutureTDM policy framework**
Lead beneficiary:	**UVA**
Work package:	WP5: ELABORATE: Legal framework, policy priorities, roadmaps and practitioner guidelines
Dissemination level:	Public (PU)
Nature:	Report (RE)
Due date:	30.08.2016
Submission date:	07.12.2016 (Due date is extended as defined in the first Review Meeting)
Authors:	**Marco Caspers**, UVA **Lucie Guibault**, UVA
Contributors:	**Kiera McNeice**, BL **Freyja van den Boom**, OK **Helen Frew**, LIBER **Susan Reilly**, LIBER **Kanella Poulli**, ARC **Jan Strycharz**, FPP **Olga Jurkowska**, FPP **Burcu Akinci**, SYNYO
Review:	**Olga Jurkowska**, FPP **Jan Strycharz**, FPP
Thanks to:	Rossitza Atanassova, Marloes Colen, Rory McNeice

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

Table of Contents

Table of Figures

Table of Tables

1. AIMS AND STRUCTURE

The exponential growth of data in the digital age has led to the development of powerful techniques for effectively harnessing digital information and discovering new knowledge. In this context, text and data mining (TDM) enables businesses, governments, journalists, researchers etc. to analyse and extract insights and knowledge, and exploit diverse and complex datasets from various digital media. By *text and data mining*, we mean any form of text and data analysis, including their preparatory activities such as retrieving contents and data and transforming them into workable datasets.[1] Unfortunately, the present use of TDM in Europe is significantly lower than in the US and Asia, in part perhaps due to limitations imposed by the European legal framework. In this context, the FutureTDM project identifies and reduces the barriers that inhibit the uptake of TDM within Europe.

To this end, we have studied the legal and policy barriers to TDM, mapped the TDM landscape in terms of available tools and applications, engaged with stakeholders on many formal and informal occasions, and carried out extensive economic research on the market and (potential) economic impact of TDM.[2] From these activities, we have identified what is holding back the uptake of TDM activities in Europe in several fields of society, such as academia, industry and government. With these key barriers identified, we can elaborate on the basic principles for further actions to overcome these barriers.

In this report, we introduce a policy framework that can be used for future TDM practices in Europe. It consists of a hierarchical model of principles and recommendations (Figure 1), where the recommendations are concerned with regulatory and policy, as well as recommendations for 'interventions' or actions by influential stakeholders.

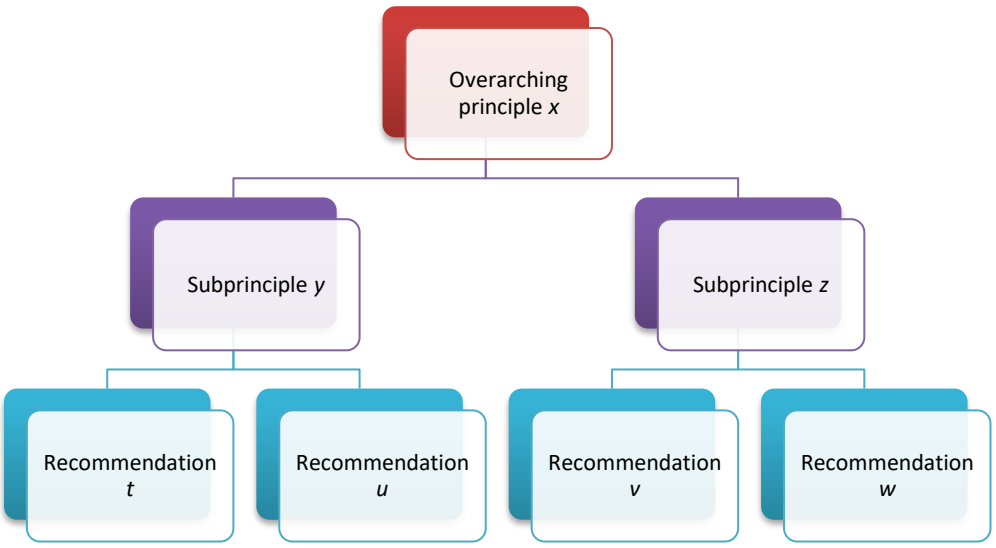

Figure 1: Hierarchy of policy framework

This report is divided into three parts. In Part I, we first outline and explain the TDM value chain, from the creation of content that can be mined to the creation of value through the results of TDM activities.

[1] For more on our definition and conception of TDM, see FutureTDM deliverable D4.1 *European Landscape of TDM Applications Report*, section 2.2.
[2] All public reports of the FutureTDM project can be found at: http://www.futuretdm.eu/knowledge-library/.

We then identify the stakeholders that have the potential to influence the activities taking place in each link of the chain. These are the stakeholders that are appropriate to address in our recommendations further in the deliverable.

Part II is the core of this report and introduces the FutureTDM policy framework. It is subdivided into different sections and starts with identifying the overarching, cross-domain principles that underlie the whole policy framework and its domain-specific principles and recommendations. These principles go beyond sector-specific barriers and must be considered in policy formation in any section or field of industry or research.

This is followed by sets of *subprinciples* and recommendations tailored to specific barriers, where a distinction is made between barriers in the following fields:

- Legal rules and policies
- Education, skills and awareness
- Economy and incentives
- Technical and infrastructure

In contrast to the overall principles, these subprinciples must be considered when seeking to overcome certain barriers in specific sections or domains of industry and science. For each subprinciple, we make recommendations addressed at specific stakeholders who have the potential to contribute to the uptake of TDM in Europe. These recommendations are not meant to be exhaustive or mutually exclusive. Other recommendations could be made as well, and other stakeholders could also help in overcoming barriers to TDM in the EU. However, with the principles and subprinciples in mind, any stakeholder seeking to overcome TDM barriers should be able to find guidance in undertaking action.

Part III presents a short summary and overall findings regarding the principles and recommendations, as well as further FutureTDM activities that will build on this policy framework.

PART I: TDM VALUE CHAIN & INFLUENTIAL STAKEHOLDERS

If we are to increase text and data mining in the EU, we must look beyond the action of mining itself: our findings have shown that barriers to TDM also exist in activities upstream and downstream of the text and data mining itself. The TDM value chain includes all these activities, thereby providing a series of distinct policy opportunities that could have a positive impact on TDM uptake. In this Part, we first outline the TDM value chain, identifying all steps in the process that relate to the *input* for TDM, TDM itself, and the TDM *output*. We then identify relevant stakeholders, so that more targeted policy recommendations can be tailored to these groups in upcoming deliverables and tasks in the project.

2. THE TDM VALUE CHAIN

In general, we can identify four steps in the TDM process as illustrated below. Each step is explained in more detail in sections 2.1 to 2.4.

Figure 2: TDM value chain

2.1 Content creation

The existence of content is a *must* for TDM: without content, there is nothing to be mined. Therefore, the TDM value chain starts with the creation of content. In this regard, content refers to any sort of information, work or data that may be the subject of mining activities. We can generally make a distinction between two categories of content:[3]

- ***Low-level* data**: this refers to what is understood as *data* in a narrow sense of the word, referring mostly to 'raw' data and facts, such measurement points, GPS coordinates, names, addresses, health information, genetic data, etc.
- ***High-level* data**: this refers to contents that are rather 'created' by human beings by some means of intellectual labour, and represent more than just a fact, characteristic, attribution or other single piece of information. This will generally include written works, videos, music or other sounds, images, diagrams, websites etc.

In general, high-level data attracts copyright, where low-level data does not. Nevertheless, both collections of low-level and/or high-level data are protectable subject matter in copyright and sui

[3] Cf. FutureTDM deliverable D3.3 *Baseline Report of Policies and Barriers of TDM in Europe* (available at http://www.futuretdm.eu/knowledge-library/), section 2.1.1.1.

generis database law. When referring to data or content, we will refer to both low-level and high-level data in the context of this report.

Although the content creation phase is not a part of the TDM process itself, some principles may affect certain elements and stakeholders in this phase, since a solution to overcome a TDM barrier may require some action from stakeholders involved in content creation. The two main categories of stakeholders associated with content creation are the actual *creators* and those who enable creation as *funders* of those creators.

2.2 Content dissemination

Content dissemination entails the activities of parties that serve as the link between the content that is created in the foregoing phase and the use of that content for TDM. Dissemination of content for TDM will predominantly take place in an online environment, where data is transported over the Internet in several ways – such as institutional repositories, publisher databases, email, P2P, cloud storage, etc. However, it is not completely ruled out that content delivery may still take place on physical carriers.

Stakeholders in the content dissemination step are essential in delivering content to TDM users, preferably in a useful format and within a proper time frame. Stakeholders may be able to exercise influence on the format in which content is delivered, as well as the accessibility, discoverability and searchability of content. This step enables TDM users to carry out the first steps of the TDM process, where content is being aggregated from different sources.

2.3 Text and data mining (TDM)

The third step is the actual phase where content is being analysed. In some cases this follows transformation and alteration of contents and datasets into useable formats; in this context, a useable format is one that allows TDM software to analyse the content according to the algorithm that the miner chooses or designs. The group of TDM users may comprise several sorts of individuals or organisations, such as academics, journalists, companies, citizens, governments, etc.

The influential stakeholders in this stage comprise of service providers that offer the tools for TDM users to carry out TDM, particularly TDM software developers providing TDM as a software tool or cloud service. These do not necessarily have to be commercial entities, since large volunteering communities may work on open source TDM tools as well.

2.4 Value creation

In this context, we refer to “value” in the broadest sense of the word, not just in terms of monetary value. Three ways in which value can be created through TDM activities are:

- Dissemination of new knowledge
- Use of insights and knowledge to improve or guide the strategy and policy of the business, organisation or government
- Marketing products and services that are based on TDM activity and its outcomes

In several sectors, such as journalism or academia, knowledge dissemination is generally the main goal of TDM; knowledge is retrieved through the use of TDM technologies, and needs to be disseminated to an audience through, for example, journal or newspaper articles, books, or the research data itself. Stakeholders involved in knowledge dissemination, as well as the activities therein, have a large overlap with those involved in the earlier steps of content creation and content dissemination – in this case they may create and disseminate new content that results from TDM activities. This means that certain stakeholders may be involved in several parts of the value chain; for example, an academic may be both a content creator and a TDM practitioner who consumes created content. In this way, they act as a very influential stakeholder to enable (the ease of) TDM for the next 'TDM cycle'.

Especially in commercial industries, the ultimate goal of TDM is often to derive insights that can be used to improve or guide business strategy. Examples include process optimisation, drug discovery, personalised advertising, sentiment analysis, and supporting high-level decision making by providing insights into the performance of the company. Stakeholders in this category can be influential by dealing with the expectations of TDM outcomes, and ensuring TDM output meets demands.

TDM activities themselves can form the basis of new products and services offered by commercial organisations ranging from start-ups to technology giants. These technologies offer a new kind of value to all sectors of the economy; in many cases, companies looking to carry out TDM to gain business insights for internal purposes make use of TDM products and services provided by third parties. These new products and services may arise from commercial research and development, or from TDM research carried out in an academic environment via technology transfer processes, both of which have the potential for significant innovation and societal impact. Examples of such products and services are search engines,[4] machine translation,[5] or 'digital lawyers'.[6]

3. INFLUENTIAL STAKEHOLDERS

Early in the project, we identified stakeholders that are involved at some point in the TDM value chain for the purposes of stakeholder engagement. We provided an overview of these stakeholders in deliverable D2.2.[7] Based on this overview of stakeholders and additional findings in the project, we have identified several categories of stakeholders that have the potential to influence activities in one or more links in the TDM value chain. Table 1 presents an overview of these stakeholders and the respective activities where they may promote TDM, including examples of each category. This overview merely serves to illustrate to whom recommendations in the following sections can be made. Given the high-level nature of this policy framework, we do not address all (listed) stakeholders in our recommendations. Nevertheless, any stakeholder seeking to positively influence the advances of TDM uptake in Europe should be able to find guidance in the principles and recommendations formulated in this document.

[4] E.g. Google (https://www.google.nl), DuckDuckGo (https://duckduckgo.com), Yahoo (https://www.yahoo.com) or Bing (https://www.bing.com).
[5] E.g. Google Translate (https://translate.google.com).
[6] E.g. ROSS (http://www.rossintelligence.com).
[7] FutureTDM deliverable D2.2 *Stakeholder Involvement Roadmap and Engagement Strategy* (http://www.futuretdm.eu/knowledge-library/), section 3.5.

Stakeholder categories	Examples	1. Content creation	2. Content dissemination	3. TDM	4. Value creation
Research community	Researchers and their associated organisations Research councils Research institutes Professional associations Universities Scientific organisations Data scientists Natural language processing researchers Semantic web scientists Information retrieval community Research libraries	(Enable the) production of research data and publications	-	Use TDM in research Fund research using TDM Provide information and advice to researchers on TDM and licensing	Publish about TDM findings Enable (OA) publication of TDM research publications and data
TDM content providers	Publishers (e.g. STM, press, OA, digital) National and university libraries (and their representative organisations) School and public libraries Research infrastructures (e.g. repositories/scholarly communication networks)	-	Disseminate content to (end-)users, retailers, etc.	Provide APIs or TDM services to TDM users to enable mining of their databases and contents	
General businesses	Energy Financial Health care IT and software, Telecommunication Retail Manufacturing	Produce data, e.g. on customers, manufacturing processes, growing processes, contents in form or reports, etc.	Share data with companies, users and other organisations Collect and access (paid) data	Carry out TDM	Use TDM internally for process optimisation, management decisions, targeted advertising, etc. Share and publish results or insights

Stakeholder categories	Examples	1. Content creation	2. Content dissemination	3. TDM	4. Value creation
	Consultancy firms				
Internet businesses	Weblogs Social media platforms Microblogging platforms Forums Websites Access & hosting providers Cloud storage and data centres Search engines Information aggregators Any website	Create digital content	Make content available and accessible on platforms, in databases, repositories, etc. Provide storage for content	Use TDM to enhance or to be able to provide certain services Use TDM for monitoring unlawful or undesirable content Use TDM for other reasons	Provide services or products based on TDM technologies, such as search engines, content/news aggregation, and recommendations
Funders	Public (research) funders Private funders of TDM initiatives or companies using TDM	-	-	Fund (scientific) research using TDM Determine conditions for funding	Provide funding for publication of TDM research and research data Determine conditions for funding
Policy shapers	EU institutions National governments Intergovernmental organisation European data protection board (EDPB) National data protection authorities (DPAs) Advocacy groups	-	Design rules and conditions for making available (certain) content	Design rules for using content to carry out TDM	Design rules for using and publishing findings and data from TDM activities and research
Skills providers	Universities and colleges (Research) libraries High schools Lower schools	-	-	Provide training in, or information or advice on TDM principles, tools, licenses and technologies	Provide training in interpreting TDM results, as well as the impact and limitations of TDM

Stakeholder categories	Examples	1. Content creation	2. Content dissemination	3. TDM	4. Value creation
	Private training companies				
TDM software providers	Large software companies Open source community SMEs and start-ups developing TDM software	-	-	Provide TDM tools/software Draft documentation	-
TDM service providers	Big data analytics providers Data services	-	-	Provide TDM and analytics services	-
Citizens	Citizen scientists Citizen journalists Citizens in general	Create digital content Source of (personal) data	Publish digital content, e.g. in the form of blog post, social media, etc.	Use TDM to discover/confirm trends, patterns and knowledge	Publish about TDM findings
Public-private connectors	University technology transfer offices (TTOs) Public-private partnerships (PPPS) Contracted research Career centres	-	Share information and knowledge between universities and private sector	Share TDM skills between universities and private sector Mobility of TDM skilled employees between university and private sector	Share TDM outcomes between universities and private sector
Public administration	Public sector bodies Tax administration Meteorological institutes Land registries Other public services	Create datasets that are the results or an externality of the main activities or services Creating datasets and digital versions of (case) laws, regulation, etc.	Making datasets available Making (case) law, regulation, policy documents, etc. available	-	-

Table 1: Influential stakeholders

PART II: FUTURETDM POLICY FRAMEWORK

This Part sets out the policy framework that will underlie all upcoming FutureTDM activities and deliverables and guide relevant stakeholders in their endeavours to advance uptake of TDM in Europe. It is the result of our extensive research in policies and regulations, and our intensive discussions and interviews with stakeholders.

We first set out the *overall principles*, which are the core of the whole policy framework and resonated in all conversations with stakeholders and policy makers. From the overall principles, we subsequently derive *subprinciples* for each category of barriers addressed. In the context of these subprinciples, we provide recommendations for the influential stakeholders that are in the best position to deal with the specific barrier and corresponding principle that should be adhered to. All sections are structured in the same way. Each section begins with a recap of the barriers that are addressed and explained in more detail in D3.3. Please note that these barriers are not necessarily exhaustive and may not be experienced by all stakeholders. All identified barriers are the result of desk research - e.g. legal and economic - and stakeholder engagement through knowledge cafés, roundtables, workshops and interviews set up and organised by FutureTDM. As a result, while the highlighted barriers reflect the most important issues in practice, they are not exhaustive and barriers may not be experienced by all relevant stakeholders. The recap of barriers is only meant to illustrate the important and relevant barriers in a certain category; for a more comprehensive overview of barriers, we refer to D3.3 Baseline report of policies and barriers of TDM in Europe.

The recap is then followed by a subsection corresponding to each overall principle, in which recommendations are made for specific stakeholders to implement the respective subprinciple. We have identified some relevant (categories of) stakeholders in different links of the TDM value chain in section 3. We have codified the different links by using colours that match the column colours in **Error! Reference source not found.** of section 3. When stakeholders are addressed in the following sections, the recommendations are provided with text in the corresponding colours to indicate in which link of the value chain the respective stakeholders can exert their influence.

Figure 3: Legend for colour coding the stakeholders

4. PRINCIPLES

4.1 Overall principles

In the first year of our project, we have investigated barriers to the uptake of TDM, both from a theoretical perspective, as well as through stakeholder engagement. These barriers have also been summarised in D3.3, where they have been categorised into three categories of barriers based on how they obstruct TDM uptake:[8]

- *Uncertainty*: this category includes uncertainties as to how, why and if TDM can be carried out, as well as the lack of awareness of different aspects of TDM.
- *Fragmentation*: this refers to the fragmentation in the TDM landscape, which prevents TDM from being carried out across e.g. national borders, scientific domains, companies or fields of expertise.
- *Restrictiveness*: the last category refers to direct limitations to possibilities to carry out TDM, either by restrictive laws, lack of expertise, limited (financial) resources, etc.

To determine the overall principles of a new policy framework, we have sought to respond to these three types of barriers. Through our extensive listening exercises, analysis of the research landscape and drawing from the expertise of the consortium and the project's expert advisory board (EAB), there emerged three clear overarching principles that will help to guide our recommendations for increasing TDM uptake. These will continue to be validated in our further outreach to stakeholders in the FutureTDM project. All of this has led to the following three overall principles:

- *AWARENESS AND CLARITY*: aims to solve the barriers related to uncertainty, but also the lack of awareness.
- *TDM WITHOUT BOUNDARIES*: aims to solve the issues of a fragmented TDM landscape.
- *EQUITABLE ACCESS*: aims to deal with restrictions – either legal, practical, economic or technical – to TDM.

Consequently, we have a set of overarching principles that are very high-level; they will need to be distilled into more concrete principles that stakeholders can work with. At the same time, they do serve as an 'answer' to all three types of barriers, making sure each type is covered and addressed in the policy framework set out in this report. Figure 2 illustrates the three overall barriers and the corresponding principles to overcome them. Below, we provide a short explanation for each principle.

[8] See note 3.

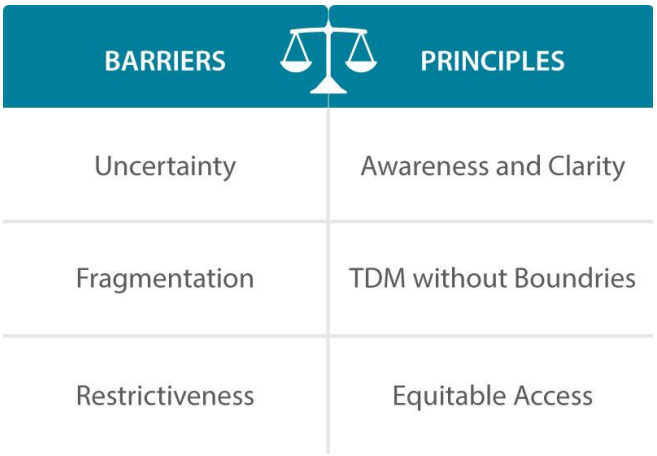

Figure 4: Types of barriers and corresponding overall principles

Principle 1 – AWARENESS AND CLARITY

Information on TDM is crucial for a flourishing TDM environment in Europe. Key factors are awareness and clarity. Awareness comprises all aspects of TDM, ranging from the legal aspects to the economic and scientific potential of TDM, and relates to all stakeholders involved, ranging from TDM users to investors and policy makers. It refers to awareness of both the (potential) advantages of TDM, as well as the risks and obstacles that currently exist in the context of TDM. For example, among the advantages are the boost to innovation and knowledge production and the economic potential in better decision-making and process optimisation. Among the risks are the uncertainty as to the lawfulness of TDM under intellectual property laws and data protection regimes, as well as the lack of standardisation, interoperability and availability of data. Before stakeholders can take advantage of TDM and address the risks thereof, there need to be awareness of these. Therefore, we include *awareness* as the first principle.

Awareness by itself is not sufficient for (potential) TDM users to become well-acquainted with the ins and outs of TDM, when information – such as legal rules, documentation for TDM tools, or specifications of an API – lacks clarity. Therefore, *clarity* constitutes a key element in the informational principle.

Principle 2 – TDM WITHOUT BOUNDARIES

This principle must be interpreted in the broadest possible sense. *TDM without boundaries* refers to all aspects that (potentially) create a fragmented TDM landscape. Some important factors identified in the FutureTDM project causing a fragmented TDM landscape include:

- *Differences in national regulations, such as copyright, database and data protection laws*: TDM might be lawful in one country, but unlawful in another. Thereby, national regulations create a fragmented TDM landscape in a geographical sense.
- *Different requirements in different fields of industry, research, or applications or purposes for the use of TDM*: many purposes, fields and applications of TDM result in the existence of many different TDM software tools, algorithms, and data formats, which all require their own expertise or instructions. This creates a fragmented TDM landscape in the sense of its application.

- *Different standards*: differences in standards – and lack of standardisation – is partly the result of the previous factor. Nonetheless, according to our findings in the project, it also relates to the lack of incentives to standardise.
- *Language differences*: being a feature peculiar to Europe, the existence of many different languages creates a fragmented landscape for TDM in particular where it concerns the field of language processing.

Principle 3 – EQUITABLE ACCESS

Access to data, text or any sort of work or information is a prerequisite for TDM to take place. Therefore, access to such content[9] is a very important and evident principle. Each case where no access to mineable content can be achieved is, in principle, undesirable. However, we must also take into account other interests at stake, especially those parties (spending resources in) creating and disseminating content. So, where free and unrestricted access would seem the main principle from a TDM user perspective, it is not for other stakeholders. It must be taken into account that an incentive to *create* and *disseminate* content must be in place to guarantee the creation and availability of content in the long term. That is where the aspect of *equity* comes into play. It means that we strive for the most unrestrictive access possible, while maintaining a fair balance with other stakeholders who may have a legitimate interest in restricting access to a certain extent. For example, there are conflicting interests in addressing the legal barriers, but also in standardisation issues – which may be the result of different requirements from TDM users. The principle of equitable access thus puts emphasis on a balanced approach to removing and overcoming a restrictive TDM environment.

4.2 Purpose of overall principles

The overall principles cover all principles discussed in this report. They cover all principles that must be regarded when solutions are considered to overcome barriers to TDM. Therefore, the three formulated principles in section 4.1 will recur in every section of this deliverable dedicated to a particular barrier. For each (category of) barrier(s), we seek how to derive barrier-specific principles from the overarching principles. This contributes to a consistent policy framework.

[9] In this report, *content* must be interpreted in the broadest possible sense, referring to any sort of data, text, image, video, sound, or piece of information that can be mined.

5. LEGAL RULES AND POLICIES

5.1 Recap of barriers

Copyright & database law	• Uncertain: scope of exceptions often unclear • Fragmented: different national scope of exceptions • Restrictive: broad reproduction right and narrow exceptions
Data protection law	• Uncertain: how to comply • Fragmented: different national interpretations, bodies and policies • Restrictive: TDM of personal data potentially unlawful
Documentation	• Uncertain: not always clear whether TDM permitted • Fragmented: TDM practitioners have to deal with many different licence conditions • Restrictive: TDM often not permitted under non-OA licences

5.2 AWARENESS AND CLARITY

Subprinciple 1	Make clear rules on copyright, database and data protection law
● @lawmaker:	Minimise borderline cases where it is unclear whether copyright or related laws apply, by clearly defining terms and concepts (e.g. *research organisation*, *text and data mining* or *scientific research purposes*) in law-making[10]
● @policy makers:	Provide explanatory documentation to accompany relevant laws, that TDM practitioners can refer to for guidance[11]
Subprinciple 2	Create guidelines on the law: what is permitted and what not?
● @FutureTDM:	Provide practitioner guidelines on how the law works, what the rules are and how to best comply[12]

[10] Although borderline cases will always exist, the legislator should make appropriate decisions to minimise the amount of such cases.

[11] Sometimes, the law needs to work with open norms or vague concepts to allow some flexibility in the course of time and with the advance of technology. In those cases, policy makers can provide for more guidance and certainty on what is lawful, and what is not, through explanatory documentation.

[12] For example, upcoming FutureTDM deliverable D5.3 *FutureTDM practitioner guidelines* would be a means to achieve this.

@European Data Protection Board & national data protection authorities:	• Provide general guidelines on TDM to help practitioners comply with data protection law and other laws relating to information privacy and confidentiality • Offer certification of data research and/or self-regulation and/or codes of conduct concerning TDM research or activities dealing with personal data, to provide certainty for practitioners • Create guidelines on what is to be understood as *archiving purposes* and *historical, statistical and scientific purposes* to help practitioners understand when these apply[13]
@ professional associations representing 'personal-data intensive' companies and research institutes:	Draft self-regulation or codes of conduct ensuring compliance, in particular with information privacy and data protection rules, for companies and research institutes to follow[14]
@ research institutes & (start-up) companies	Implement ethical and legal review practices for large research projects dealing with big data[15]
@research libraries & librarians	Create a physical or virtual centre or platform for knowledge on TDM licences and rules for researchers interested in TDM to consult
@ research funders	When paying for researchers to publish their results open access, require that grant receivers use well-known OA licenses, attached to the relevant publication or dataset in human- and machine-readable metadata, to help harmonise licences in OA content. Where necessary provide guidelines to help researchers do this.
@ all content creators and all content providers/ disseminators	Use clear licence statements in the metadata and in the text, preferably well-known ones and in machine-readable form, to improve accessibility and re-usability of content by TDM processes[16]

[13] When TDM is carried on personal data for such purposes, it is subject to a more lenient regime under General Data Protection Regulation (and under the Data Protection Directive 95/46/EC). See also FutureTDM deliverable D3.3, section 2.2.2.

[14] Inspiration can be drawn from comprehensive self-regulation in the field of health research. Cf. FutureTDM deliverable D3.3, section 3.2.3.

[15] Similarly, the Royal Netherlands Academy of Arts and Sciences recommends, for example, institutes or departments that are involved in informatics research should install an "Ethical Review Board for Informatics (ERBI)". See its report on *Ethical and legal aspects of informatics research*, p.9. Available at https://www.knaw.nl/shared/resources/actueel/publicaties/pdf/20160919-eng-advisory-ethische-en-juridische-aspecten-van-informaticaonderzoek-web.

[16] Cf. RDA-CODATA Legal Interoperability Interest Group 2016, p. 5, where it recommends "Principle Four: State the rights transparently and clearly".

5.3 TDM WITHOUT BOUNDARIES

Subprinciple 1	Create harmonised and mandatory rules on copyright, database and data protection law
@European lawmaker:	• Introduce a harmonised, mandatory exception to copyright for TDM activities, and minimise the leeway for deviating interpretations by national legislators and courts to ensure harmonisation in practice as well as in principle • Ensure this TDM exception allows for the same exempted activities under both copyright and database law • Ensure this TDM exception does not discriminate between commercial and non-commercial organisations or activities, as this distinction may not always be possible in practice • Ensure this TDM exception applies only in cases where practitioners have lawful access to content, to respect the rights of content owners[17] • Ensure that TDM results can be disseminated while protecting the interests of other stakeholders • Create a fully harmonised data protection regime and minimise deviating interpretations of the General Data Protection Regulation to ensure harmonisation in practice as well as in principle
Subprinciple 2	Harmonise licences for TDM[18]
@content creators and providers:	As much as possible, minimise variation among TDM licences, and minimise the number of different TDM licences used to reduce the complexity of re-using content for TDM (with deviation only for reasons of privacy, data protection, etc.)
@research institutions:	Where practical, avoid signing licenses that prohibit or restrict TDM, to avoid unnecessary restrictions on content
Subprinciple 3	Harmonise interpretations of key principles in data protection law
@ EDPB:	Achieve consensus on key concepts such as *personal data*, *historical, statistical and scientific purposes* and *archiving purposes* to reduce uncertainty for TDM practitioners
Subprinciple 4	Harmonise rights in research data[19]
@ research funders:	Align research data sharing and research publishing conditions required of research grant recipients to ensure that all public research in Europe is available under the same – TDM friendly - conditions

[17] In any event, there is *lawful access* when it is permitted by the law or by the rightsholder.
[18] Cf. RDA-CODATA Legal Interoperability Interest Group 2016, p. 5, where it recommends "Principle Five: Promote the harmonization of rights in research data".
[19] We have adopted this principle from RDA-CODATA Legal Interoperability Interest Group 2016, p. 6.

@ European lawmaker:	Discuss with Member States the benefits of mandatory rules that ensure that a researcher may publish publicly-funded research publications in open (university) repositories or scholarly communications networks for non-commercial purposes, with consideration for reasonable embargo periods[20]

5.4 EQUITABLE ACCESS

Subprinciple 1	Ensure legal rules reflect a fair balance between the interests of TDM practitioners and rightsholders (of copyright and database rights), and reaffirm that ideas and facts as such are not protected
@EU lawmaker:	Introduce a mandatory copyright exception for TDM activities that: • permits uses that do not trade on the underlying creative and artistic expression of the content analysed • requires *lawful access*[21] by the user • is not overridable by contract[22] • only permits technical measures that are necessary, reasonable and proportionate to guarantee the security and integrity of content providers' infrastructure[23] Make sure that circumvention of technical protection measures (TPMs) and digital rights management (DRM) is permitted,[24] without harming the said security and integrity
@ EU and national governments:	In relation to the previous recommendation, provide guidelines and best practices on what is regarded to be reasonable and proportional in the context of (technical) measures to guarantee the security and integrity of the system[25]

[20] Germany and the Netherlands are an example of how this can be done through copyright legislation. See for more details: FutureTDM deliverable D3.3. Also, France has a similar provision that was recently introduced Article 30 the *Digital Republic* act (https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000033202746&dateTexte=&categorieLien=id) and Italy as well (http://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:legge:2013;112, Article 4).

[21] "Lawful access" should be interpreted as access that is permitted by the author or by the law.

[22] In this context, inspiration can be drawn from Article 15 in conjunction with Articles 6(1) and 8 of the Database Directive (96/9/EC), which provides for 'rights' of lawful users of a database that are not overridable by contract.

[23] This is currently proposed as well in Article 3(3) of the Commission's *Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market*, COM(2016) 593 final.

[24] As appeared from FutureTDM deliverable D3.3, section 2.1., national rules on TPMs currently do not guarantee that individuals who should benefit from copyright and database right exceptions are prevented from that.

[25] FutureTDM deliverable D3.3, section 2.1, showed that there is already a lot of uncertainty on the circumvention of TPMs to benefit from exceptions, and new rules concerning technical measures might create an extra layer of uncertainty.

Subprinciple 2	Ensure rules and policies on data protection reflect a fair balance between privacy interests of data subjects and value creation from TDM
@ lawmakers:	Make sure that TDM research is allowed when individuals' privacy interests are not severely affected – and make clear when that is the case – to avoid unnecessary restrictions on TDM activity
@European Data Protection Board & national data protection authorities:	Provide for procedures that promote access to data for TDM research, while safeguarding the privacy interests of data subjects
Subprinciple 3	Encourage researchers to freely share scientific publications and research outcomes, especially when they are publicly funded, so they can be mined without restrictions[26]
@ research funders:	• Encourage publications from publicly-funded research to be published under an Open Access licence, such as CC-BY 4.0 (and no previous versions, since they do not include *sui generis* database rights), to increase access to content for TDM activities[27] • Make sure that researchers have resources to publish in OA journals when (reasonable) article processing fees apply[28] • Where practical, mandate timely deposit of a peer-reviewed version of publications in an open repository, to provide access to research results for TDM activities
@ research institutions:	• Develop OA policies, that take into account impact of research and the benefits of open access to research for TDM[29] • Provide researchers with sufficient information on OA licences, OA journals and repositories to make informed decisions about OA publishing
@researchers:	Where practical, make research outputs available in a publicly available repository under standard OA licences to improve accessibility for TDM

[26] This should enhance the *accessibility* and *re-usability* of scientific works, which are two key factors in the FAIR data principles and promote the mining thereof. Cf. Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, and others, 'The FAIR Guiding Principles for Scientific Data Management and Stewardship', *Scientific Data*, 3 (2016), 160018 https://doi.org/10.1038/sdata.2016.18.

[27] Some research funders already impose similar requirement. See FutureTDM deliverable D3.3, section 3.1.3., for more on this.

[28] For example, see FutureTDM deliverable D3.3, section 3.1.3.2.: The Netherlands Organisation for Scientific Research has set up an incentive fund to compensate researchers for APCs.

[29] See FutureTDM deliverable D3.3, section 3.1.1., on the permissibility of TDM under OA licenses.

Subprinciple 4	Encourage open access to underlying research data, especially in publicly-funded research[30]
@research funders:	Require research projects to include a comprehensive data management plan, with emphasis on the value of open data approaches, that considers and implements safeguards that address conflicting interests related to privacy, data protection and confidentiality
Subprinciple 5	Guarantee reproducibility of TDM research
@lawmakers:	Ensure that a TDM exception allows the retention of copies for reproducibility and verifiability of TDM research[31]
@publishers and other content providers:	Allow TDM practitioners to retain long-term copies of content used for TDM for the purpose of reproducibility and verifiability of the TDM research[32]

[30] This should enhance the *accessibility* and *re-usability* of research data, which are two key factos in the FAIR data principles. Cf. Wilkinson, Mark D., Michel Dumontier, Ijsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, and others, 'The FAIR Guiding Principles for Scientific Data Management and Stewardship', *Scientific Data*, 3 (2016), 160018 https://doi.org/10.1038/sdata.2016.18.

[31] This may be accompanied with - formal or technical - measures to guarantee that these reproductions are not 'abused', e.g. for commercial purposes. The recently introduced TDM exception in France provides for similar safeguards, see *LOI n° 2016-1321 du 7 octobre 2016 pour une République numérique*, Articles L. 122-5 and L. 342-3.

[32] See note 22.

6. EDUCATION, SKILLS AND AWARENESS

6.1 Recap of barriers

Education	• Uncertainty: poor awareness of potential benefits • Restrictive: lack of education in data management and legal/licensing issues
Skills	• Fragmented: poor understanding of domain-specific needs; poor understanding in academia of knowledge transfer requirements; skills gap on both sides between domain experts and TDM experts • Restrictive: lack of access to skilled data analysts; high barrier to entry for use of complex tools; lack of skilled in-house experts
Knowledge access	• Uncertainty: lack of knowledge around whom to consult with queries • Fragmented: best practices/techniques not shared across domains/sectors • Restrictive: access to skilled practitioners prohibitively expensive for academics

6.2 AWARENESS AND CLARITY

Subprinciple 1	Ensure all citizens in society have awareness and basic understanding of the potential impact and value of (big) data
@ educational institutions and governments:	Establish (and mandate) training and courses in data literacy, "computational thinking" and fundamental ideas about data science early in the educational system, as well as in lifelong learning, to foster societies that are better able to recognise the potential value of data and data analysis
Subprinciple 2	Ensure all researchers and industries – including those in traditionally less data-driven fields – have awareness and basic understanding of the potential uses and benefits of TDM (OR: "Sell the concept better")
@FutureTDM:	Provide evidence of the positive results of TDM, including the economic growth potential, to demonstrate the concrete value of TDM
@FutureTDM:	Provide examples, best practices and use cases to sell the 'TDM concept', particularly in underrepresented fields, to ensure all sectors are able to benefit from the use of TDM
@ universities, research organisations, library associations, medical community, businesses, and members of the content mining community	Advocate the benefits of content mining as trusted voices in their respective communities

@ universities and higher education institutions:	Integrate education about data management and analysis across all subjects, not just traditionally data-driven fields, to ensure all sectors can benefit from the use of TDM
Subprinciple 3	Ensure everyone working with data has access to education, information and experts to consult about TDM, e.g. as regards legal issues, tools, algorithms, ethics, etc.
@ libraries:	Serve as information hubs and provide training for researchers and citizen scientists in data, tools and TDM services, as well as advice on licences and legal concerns around TDM
@university senior management:	Consider implementing services beyond what libraries can provide, for example dedicated centres supporting TDM, to further support researchers
@governments:	Establish central, comprehensive, accessible information resources for people interested in TDM, managed by an institution that can connect expertise on a national and international level, to make it simple for TDM practitioners to access comprehensive guidance and support
Subprinciple 4	Raise awareness of the benefits that sharing research data can bring to researchers and to others
@research libraries:	Help researchers to develop skills in research data management (RDM) by: • Explaining the benefits of sharing research data • Sharing best practices around making research data discoverable and re-usable • Building open science partnerships to demonstrate their value
Subprinciple 5	Ensure users have realistic expectations of technical capabilities and limitations of TDM tools/services
@developers/ providers of TDM tools and services:	Inform clients about what they can expect – and *not* expect – as output from TDM activities, so that clients can make informed decisions about how to employ the use of TDM. For example, if appropriate, this could be achieved by pursuing openness of algorithms used, to reveal the choices made in the analysis which influence the outcomes of TDM

6.3 TDM WITHOUT BOUNDARIES

Subprinciple 1	Make it easy for existing and potential TDM practitioners to discuss and share skills and best practices across disciplines and sectors
@universities, and research libraries:	Set up internal communication channels and platforms where TDM researchers and practitioners from different disciplines can discuss and share their skills and best practices, to foster cross-disciplinary knowledge transfer
@professional associations:	Establish communication channels among TDM practitioners in different companies, disciplines and sectors of the economy, to foster cross-disciplinary knowledge transfer
@universities:	Promote public-private partnerships (PPPs) in TDM research to bridge gaps in skills and needs between industry and academia and increase the impact of TDM research

● @TDM researchers and users in academia and industry:	Participate in knowledge and skills exchanges between companies and/or researchers through conferences and seminars devoted to data mining, to facilitate sharing of best practices
Subprinciple 2	Ensure education in TDM prepares practitioners for the diversity of TDM tools and applications
● @educational institutions:	Provide courses that prepare practitioners for a diverse TDM landscape by focussing on core principles and skills that can be applied to a variety of different tools in practice
● @ educational institutions, industry and researchers:	Enter into dialogue so that courses in TDM prepare students for a diverse labour market

6.4 EQUITABLE ACCESS

Subprinciple 1	Promote exchanges of skills and knowledge between companies and universities
● @universities:	Enhance knowledge transfer activities regarding text and data mining technologies, theories and other knowledge, e.g. through networking with TDM companies, carrying out contract research, participating in PPPs to develop technologies and knowledge, and improving the mobility of employees between universities and industries[33]
● @companies and universities:	Initiate, support and collaborate on private learning initiatives that work at the interface of academia and industry[34]
Subprinciple 2	Ensure that minority groups are not disproportionately disadvantaged or discouraged from TDM activities
● @ universities and industry:	Promote and raise visibility of underrepresented minorities in data analytics – e.g. as role models – who carry out interesting TDM projects
● @ FutureTDM:	Shared best practices of TDM should also promote the work of TDM practitioners from underrepresented minorities in the field

[33] Cf. the different ways of knowledge transfer as identified by Finne, Håkon, Adrian Day, Andrea Piccaluga, André Spithoven, Patricia Walter, and Dorien Wellen, *A Composite Indicator for Knowledge Transfer: Report from the European Commission's Expert Group on Knowledge Transfer Indicators*, 2011 https://ec.europa.eu/research/innovation-union/pdf/kti-report-final.pdf, section 2.1.

[34] An example of such an initiative is S2DS http://www.s2ds.org.

7. ECONOMY AND INCENTIVES

7.1 Recap of barriers

Investment issues	• Uncertainty: lack of awareness of TDM's gains for companies | troubles calculating ROI in TDM projects • Fragmented: investments not sufficient in all parts of the value chain
Limited use of potential	• Uncertain: lack of understanding TDM • Fragmented: organisational 'data silos' across different sectors and businesses
Supply & demand	• Uncertainty: companies are hesitant to invest largely in TDM activities that many not produce expected results
Decision making	• Uncertain: how to relate TDM to data-based management | how TDM brings organisational value

7.2 AWARENESS AND CLARITY

Subprinciple 1	Promote TDM and the value it can bring to businesses and organisations
@ FutureTDM and government agencies:	• Disseminate knowledge on TDM use cases and their effects and utilisations in a business context • Disseminate success stories to show how TDM can benefit companies
@ research funders:	• Promote research that showcases and studies the financial benefits of TDM for SMEs • Promote research that develops state-of-the-art ways of calculating return on investments in TDM for companies
@ governments, professional organisation, advocacy groups, businesses and universities:	Organise hackathons and similar events to promote TDM and create awareness of its opportunities[35]
Subprinciple 2	Promote a data-savvy culture so that all stakeholders are aware of the potential benefits of TDM
@ businesses:	• Make analytics-based decision-making part of an organisational culture • Broaden analytical capabilities portfolio to cover all areas from descriptive, diagnostic, predictive and prescriptive analysis, to human input, decision and, finally, action

[35] An example of such an event is the yearly *EU Hackathon*, see e.g. the 2016 version http://2016.euhackathon.eu.

@ FutureTDM, academia and consulting firms:	• Show specific ways of calculating the positive effects of introducing TDM in organisations • Develop and promote a better understanding of the business opportunities enabled by TDM • Identify, validate and promote sustainable business models based on TDM in different sectors

7.3 TDM WITHOUT BOUNDARIES

Subprinciple 1	Promote sharing among data 'silos' of different businesses and sectors
@ industry:	Promote and disseminate knowledge on TDM usage in business, and how it can be enhanced using common standards - for example in APIs, structuring of datasets, metadata or shared ontologies
Subprinciple 2	Promote data savviness across departments
@ industry:	Invest in talent acquisition and staff training, making staff more data-savvy across all departments of the organisation, to ensure everyone understands the value of incorporating data use into business activities

7.4 EQUITABLE ACCESS

Subprinciple 1	Promote sharing among data 'silos' of different businesses and sectors
@ research funders:	Provide more funding for research that uses TDM, taking into consideration the necessary budget for all stages of the TDM value chain (including infrastructure, storage and scaling)
Subprinciple 2	Dedicate more funding to companies creating value from TDM
@ governments:	Set up platforms where TDM companies, academic researchers and capital investors can meet and discuss the advantages of TDM products and services, and articulate their potential value to funders
Subprinciple 3	Provide more recognition and acknowledgement for TDM uptake
@ universities, research organisations, research funders and businesses:	Introduce incentives to reward those who are using TDM, e.g. by noting and commending TDM in evaluation of processes and proposals

8. TECHNICAL AND INFRASTRUCTURE

8.1 Recap of barriers

Data(sets)	• Fragmented: data heterogeneity • Restrictive: poor quality of data, annotations and metadata
Tools & Infrastructure	• Uncertain: user unfriendly interfaces • Fragmented: architectural mismatches | incompatibility
Languages	• Fragmented: lack of availability of language resources • Restrictive: availability of language resources
Documentation	• Uncertain: vagueness • Fragmented: mismatch between documentation and tool versions • Restrictive: absence of documentation

8.2 AWARENESS AND CLARITY

Subprinciple 1	Provide clear documentation and user manuals for TDM tools, technologies and datasets, for other (interoperability-seeking) developers
@developers:	• Write clear and well-written specifications for TDM tools, to help others use them • Keep documentation up-to-date and accessible from a single, easy-to-find access point
@creators of datasets and metadata curators:	Maintain clear and up-to-date specifications and guidance for the use of annotations and other metadata schema, to help content owners use consistent metadata
@FutureTDM:	Share best practices around documentation of TDM tools and methods
Subprinciple 2	Data(sets) should be consistent and complete
@data(base) producers:	As much as possible, provide for 'clean' datasets – that is, datasets that minimise the amount of processing and normalisation necessary for TDM activities
@developers, research institutions, libraries and their representing organisations:	• Develop appropriate platforms for annotating, amending and normalising datasets, to help create interoperable and re-usable data • Draft guidelines for annotations that are strict and clear, and implemented accordingly
@developers and researchers	Publicly share metadata of datasets, open to re-use and correction, to make it easier to accurately identify and understand the contents and context of datasets

Subprinciple 3	Minimise barriers to entry for the use of TDM by lay-users or those with limited computational skills
@developers:	Create user-friendly TDM tools, workflows and infrastructure, e.g. through user-friendly interfaces, for the benefit of users with limited computational skills
@research infrastructures:	Adopt standards for interoperability to link datasets from various sources, and provide access to these via open, user-friendly APIs

8.3 TDM WITHOUT BOUNDARIES

Subprinciple 1	Encourage consistency in the use of standards[36]
@ registries, repositories and industry:	Standardise data formats, communications protocols and middleware used by different components of a system to improve interoperability and make it easier to connect and use data from a variety of sources
Subprinciple 2	Use *open* standards
@ content creators and providers:	Provide datasets in open standards, instead of proprietary standards, to ensure everyone wishing to use those datasets has access to the relevant standards
@research funders:	Require publicly-funded research to use open standards for tools and dataset formatting, to ensure data is as accessible as possible
Subprinciple 3	Make TDM stronger among all languages
@ developers:	Adjust and tune existing tools to support more European languages, to ensure as wide a range of users as possible has access to these tools
@governments:	Promote or incentivise support for more European languages with funding, contests or other instruments that award these efforts made by developers
Subprinciple 4	Ensure standards reflect the large variety of TDM tools and applications
@ developers:	When working together on developing standards, take into account the variety of their users and applications, to ensure that standards are actually applicable to the breadth of the TDM landscape
Subprinciple 5	Create a common infrastructure for all sciences
@ EU government:	Fund, enable, promote and/or initiate a common infrastructure where researchers can share, store, and access research outputs and data *inter alia* for TDM purposes. Make sure that current initiatives, such as OpenMinTeD[37] and the European Open Science Cloud,[38] succeed

[36] Standard setting in the context of (*big) data technologies* is also among the European Commission's priorities. See European Commission, 'Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions "ICT Standardisation Priorities for the Digital Single Market" COM(2016) 176 Final', 2016 http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=15265, section 3.1.
[37] http://openminted.eu.
[38] http://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud.

8.4 EQUITABLE ACCESS

Subprinciple 1	Increase access to and enhancement of TDM tools by making them available under open source licences
● @ research funders:	Require that TDM tools and technologies developed through publicly-funded research are made available under an open source licence, to maximise the value they offer to society

PART III: CONCLUSION/SUMMARY

This FutureTDM policy framework highlights the principles that should, in our view, be taken into account when stakeholders develop new policies concerning the use and promotion of TDM. This framework aims to contribute to the uptake of TDM, in particular within the EU. By formulating the principles on a high level, any stakeholder should be able to find guidance in the overall principles. Since these principles are crystallised into more detailed subprinciples to apply to specific domains or specific types of barriers, stakeholders should find even more guidance in that. In addition, we have provided more concrete recommendations that are meant to set priorities as to where the focus should be for action. At the same time, these recommendations can serve as an example for other stakeholders to show how certain principles can be implemented into actions.

This policy framework also forms the basis for upcoming deliverables in Work Package 5. First, the principles and recommendations put forward in this report will inform the practitioner guidelines we will develop in deliverable D5.3. In the current report, we have included ourselves – FutureTDM – as an influential stakeholder in certain recommendations to develop practitioner guidelines that may be useful in certain fields to promote uptake of TDM. Second, we will prepare fundamental policy priorities, and develop future research agendas and roadmaps for increasing uptake of TDM in deliverable D5.4. The principles of this policy framework form the normative framework for D5.4, while the recommendations already indicate the direction of policy priorities, research agendas and roadmaps. Needless to say, FutureTDM continues to engage with stakeholders to evaluate our principles and recommendations to make sure that our goal of promoting TDM uptake in Europe can be achieved.