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 6.4

FutureTDM policy frameworks, roadmaps, and practitioner guidelines (Additional Report)

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:	**D6.4**
Title:	**FutureTDM policy frameworks, roadmaps, and practitioner guidelines**
Lead beneficiary:	**ARC**
Work package: awareness raising	WP6: Empower: Collaborative knowledge base, open information hub,
Dissemination level:	Public (PU)
Nature:	Other (O) Additional Report
Due date:	31.07.2017
Submission date:	31.07.2017

Authors: **Kanella Pouli,** ARC
Burcu Akinci, SYNYO

Contributors: **Kiera McNeice,** BL

Review: **Bernhard Jäger,** SYNYO

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 Contents

1 INTRODUCTION

This report presents the outcomes of the FutureTDM project as concerns the **elaborated frameworks**, **policy priorities**, **roadmaps** and **practitioner guidelines** as they have been published on the FutureTDM platform via informative and insightful infographics and visualisations to appeal to a dynamic range of users and stakeholders.

The **FutureTDM platform** consists of two parts: the main dissemination platform namely **Open Information Hub** and the **Collaborative Knowledge Base** where the collection of knowledge and information resources on Text and Data Mining is displayed in a user-friendly structure with coherent and up-to-date view on the TDM landscape. The platform[1] was launched on April 1st, 2017 and has the following structure:

1. **Home page**: it serves as an entry point, provides information on the project, and has the latest updates on events, newsletters, blog posts and videos (see Figure 1).

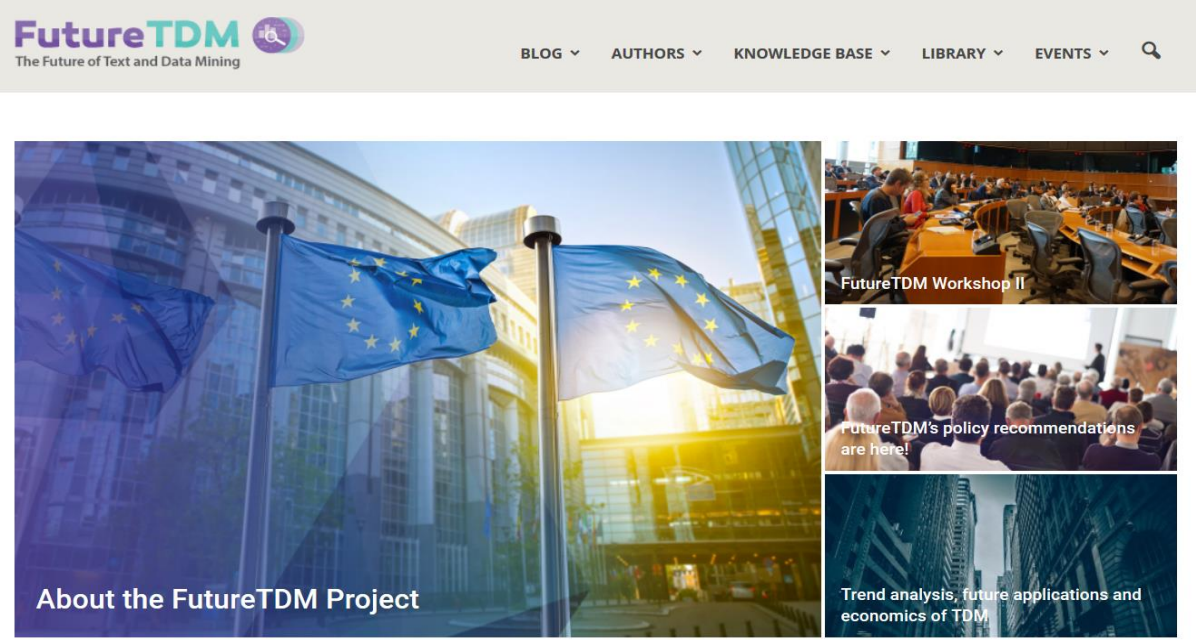


Figure 1: The FutureTDM home page

2. **Blog**: it contains personal opinions of the authors on important TDM issues as well as some of the project outcomes under the categories listed below:
   - Community & Events
   - Education & Skills
   - Legal & Policies
   - Research & Insights

[1] The platform was created using standard open access software and designed with a flexible and scalable architecture to guarantee a long-term sustainable usage. It was launched eight months earlier than initially planned. It is described in detail in deliverables 6.1 Web-based FutureTDM Collaborative Knowledge Hub, and 6.3 Research projects directory and best practice library (Progress Report).

- Stories & Cases
- Technologies & Tools

3. **Authors**: this section contains lists of the organizations, the projects and the people (internal and external experts) who are involved in a variety of ways with the FutureTDM project.
4. **Knowledge Base**: the content of the Knowledge Base has been derived from various project deliverables[2]; it is organized in the following sections and hopefully serves as a reference point for researchers, practitioners and any stakeholder interested in the field.
    - **TDM Projects** is a directory of the most important TDM related projects (see Figure 2)

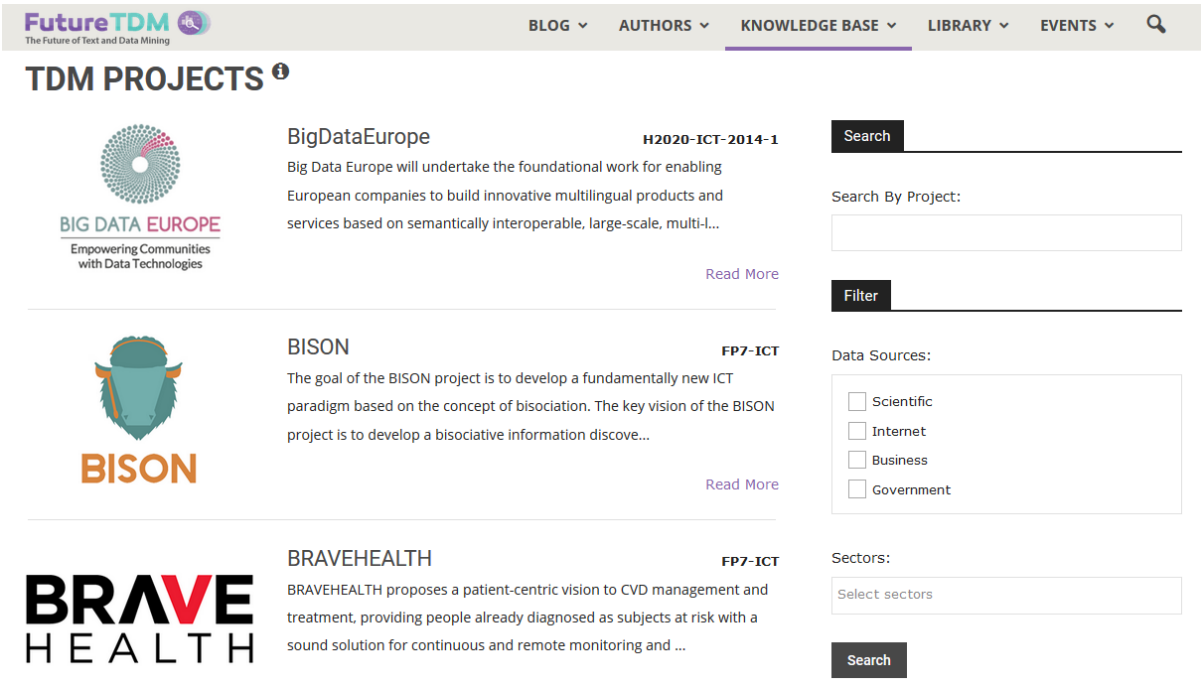


Figure 2: The TDM Projects page

- **Expert Navigator** is a directory of the organisations working with TDM
- **TDM Methods** is a directory of the commonly used methods for TDM (see Figure 3)
- **TDM Tools** is a directory of tools used for TDM,
- the **Policy Framework** section is dedicated to recommendations for TDM practices (for a detailed description see section Policy Framework),
- the **Practitioner Guidelines** section contains all the sets of guidelines to help address specific challenges to greater use of TDM technologies (for a detailed description see section Practitioner Guidelines), and
- the **Roadmap** section outlines a plan to increase the uptake of TDM technologies in Europe (for a detailed description see section Roadmap).

[2] The Knowledge Base presents collections derived from WP4, specifically D4.2 Collections of relevant projects, (not publicly available), which encompasses individual persons or groups of people focusing on TDM (i.e. experts, organizations and project collections) as well as technologies and resources useful for current and future TDM practitioners (i.e. methods and tools lists).

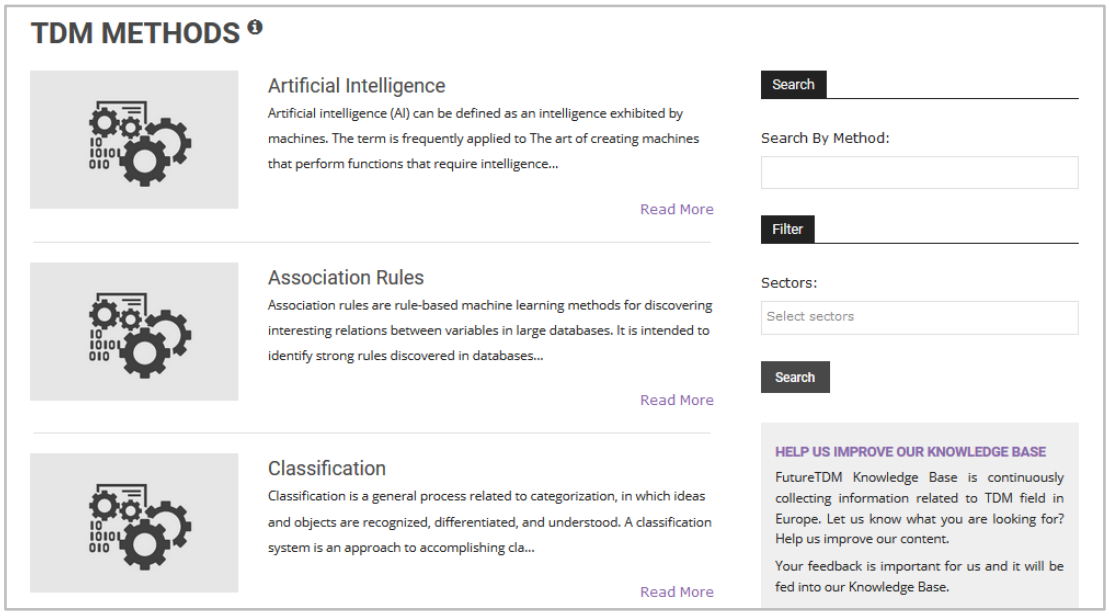


Figure 3: The TDM Methods page

In addition functionalities have been created to facilitate search in the directories which are dedicated to projects, experts, methods and tools (see Figure 2: The TDM Projects page Figure 2 and Figure 3): the user can search either by using a name or by filtering results according to data sources of usage/specialization, or even specific sectors and/or application fields of interest.

Below (See Figure 4) detail page of one of the TDM Tools, ContentMine, can be seen.

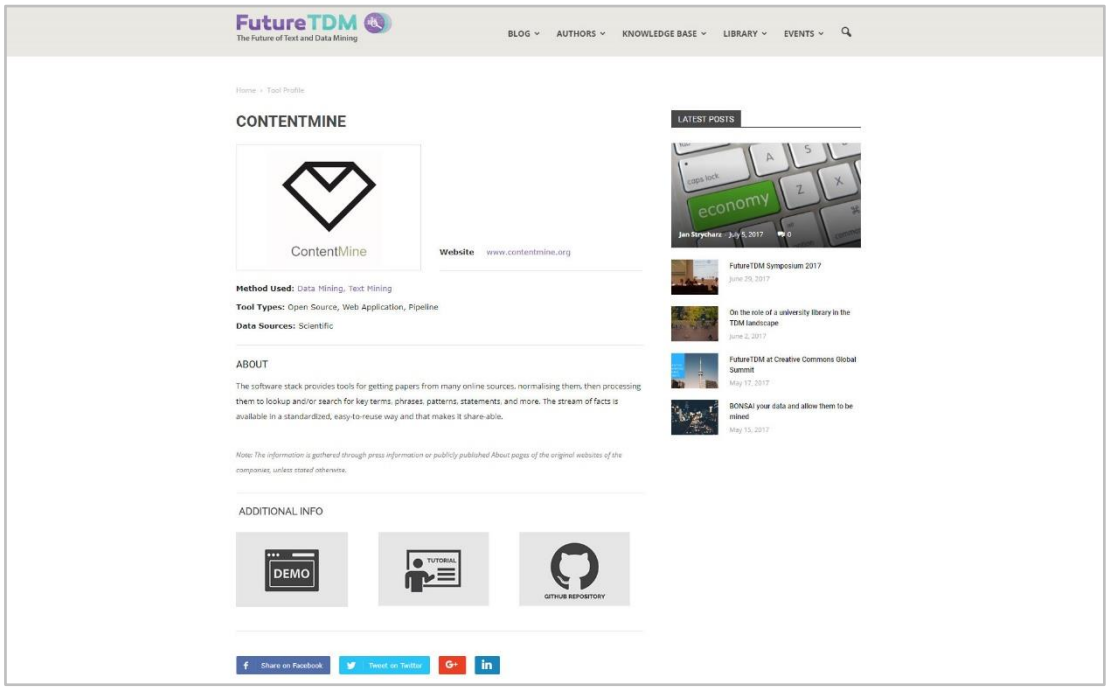


Figure 4: TDM Tools Detail Page

5. The **Library** is the section dedicated to all the project produced material: the publicly available deliverables which can be found under the category "expert reports", the flyers (available at "Dissemination outputs") as well as all the Awareness Sheets.
6. **Events**: this section is subdivided into the Knowledge Cafés[3], Community Events and Event photos, providing information on the various events organized by FTDM or to which FTDM participated.

Besides the strictly informative nature of the platform on issues related to TDM, special care has been given to create divisions providing insights and solutions to TDM practitioners and stakeholders. These are the sections dedicated to the **policy framework**, the **guidelines** and the **roadmap** which are described in detail in the following chapters. As dictated by the Description of Action, special design and architecture has been used to make the visualization of the project results as explicit and comprehensive as possible. Different colours have been chosen for each of the aforementioned sections, each one being associated with the FTDM logo: **purple** for the policy framework, **light blue** for the guidelines and **light green** for the roadmap.

[3] Throughout the first half of 2016, FutureTDM had been running a series of Knowledge Cafés across Europe, the results of which are presented in **D2.3 Report on stakeholder mobilisation and perceptions**, and summarized in this section of the platform.

2 POLICY FRAMEWORK

The ongoing dialogue with the various TDM groups throughout the project[4], as well as the close collaboration with the advisory board of the consortium, ensured various stakeholder perspectives and highlighted the problems encountered by the TDM community. The barriers encountered were grouped in four broad categories:

a) Legal Rules & Policies
b) Education, Skills & Education
c) Economy & Incentives
d) Technical & Infrastructure

To overcome these barriers a hierarchical model of principles and recommendations has been created and presented in the **Policy Framework** section of the Knowledge Base on the platform.

A guide on how to read the recommendations is provided by the drop-down heading "HOW TO READ" (see Figure 5). The overall principles underlying the barriers in all categories are (as explained in D5.1 FutureTDM Policy Framework):

- **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, and
- **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.

The recommendations created by FTDM experts are based on the principles of:

- **Awareness & Clarity**, which aims to solve the barriers related to uncertainty, but also the lack of awareness,
- **TDM Without Boundaries**, which aims to solve the issues of a fragmented TDM landscape, and
- **Equitable Access**, which aims to deal with restrictions -either legal, practical, economic or technical- to TDM.

[4] The outcomes of consultations with stakeholders are presented in deliverables **D2.3 Report on stakeholder mobilisation and perceptions**, **D2.4 Workshop summary report 1**, **D2.5 Workshop Summary report 2**, **D4.3 Compendium of Best Practices and-Methodologies**, **D4.5 Compendium of Best Practices and Methodologies** (the deliverables which are publicly available can be downloaded from the **Open Knowledge Library**). They are licensed under a Creative Commons Attribution 4.0 International License (CC-BY) and appropriate attribution to the FutureTDM project is mandatory when sharing and adopting the provided material.

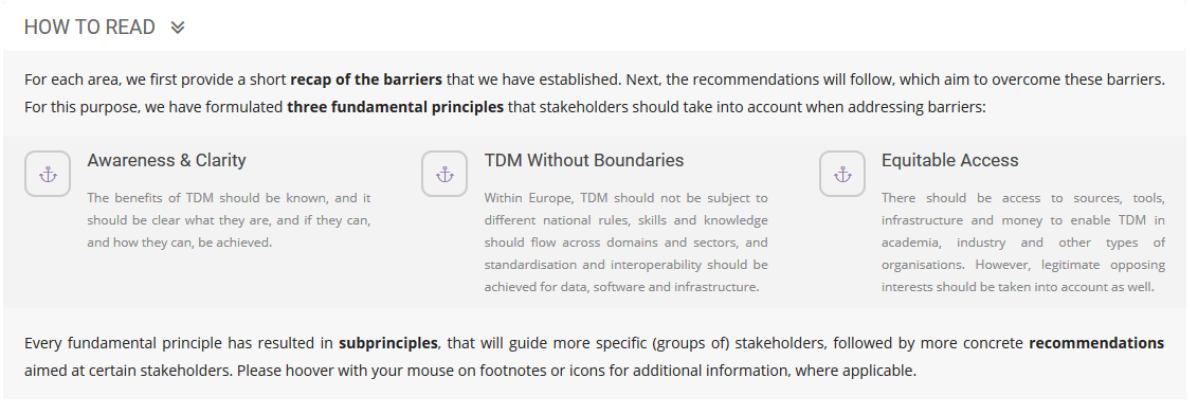


Figure 5: Instructions on how to read the policy framework

The platform offers the possibility to navigate through the four categories, with a recap of barriers for each one preceding recommendations (see Figure 6).

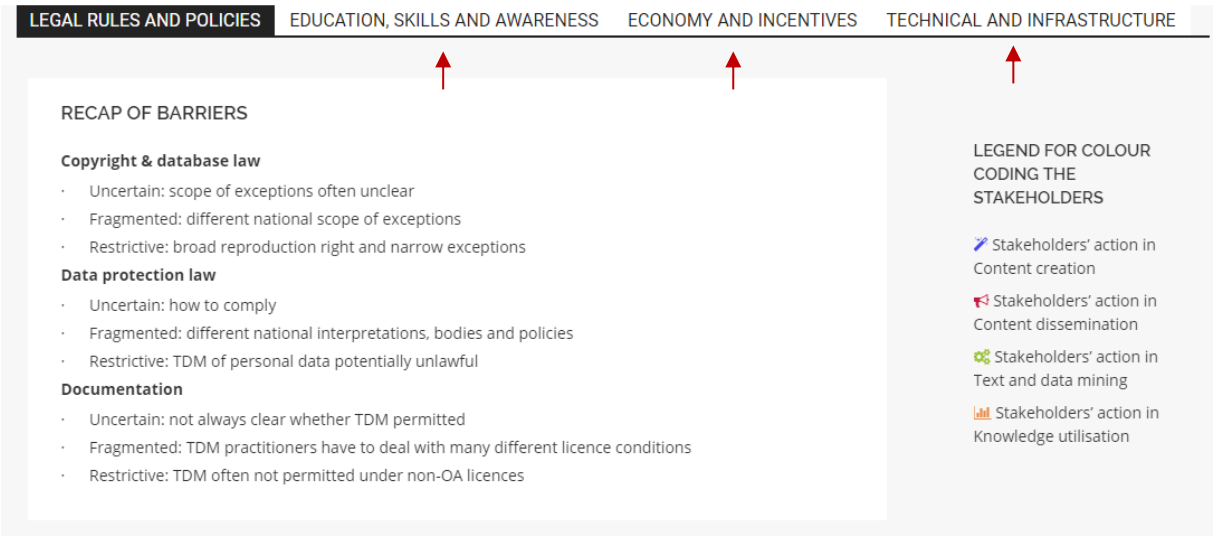


Figure 6: Barriers and the legend for color coding the stakeholders

All recommendations for actions are accompanied by symbols which are explained by a legend appearing on the top right of each category section (see). This legend makes reference to the steps of the TDM process, known as the TDM value chain[5] (see Figure 7) where different groups of TDM stakeholders are involved.

[5] For more information please see D5.1 FutureTDM Policy Framework.

Figure 7: The TDM value chain

Therefore, in any given category (Figure 9 showcases an example from **Technical and Infrastructure**) first the principle under consideration is introduced (here it is **AWARENESS AND CLARITY**) and then the existing subprinciples are highlighted, each one in a purple box: Subprinciple 1 (Clear documentation) and Subprinciple 2 (Consistency and Completeness of datasets). After each subprinciple is briefly explained, the appropriate categories of stakeholders are mentioned with a symbol (see Figure 7) representing the area where the action should be made: content creation, content dissemination, text and data mining or knowledge utilization (this area is equivalent to the value creation step of the TDM value chain).

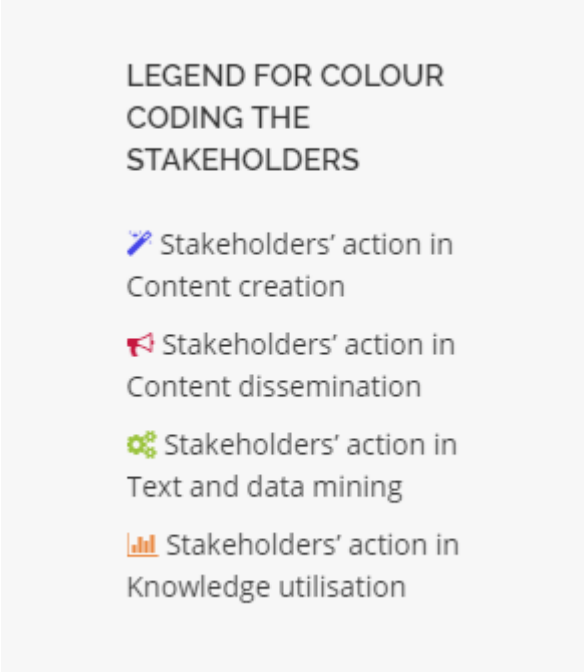


Figure 8: Legend for coding the stakeholders

In the case of subprinciple 1, it is developers who are responsible for writing clear documentation (and updating it regularly) and this affects the actual Text and Data Mining phase; on the other hand creators of datasets and metadata curators, should take action in both the stages of actual TDM and Knowledge Utilization.

The use of symbols and colours, as well as the common idea on which the information in all categories is organized and presented, facilitates the understanding of complex issues and problems which are

intertwined and which involve the active participation of multiple groups of stakeholders to overcome them.

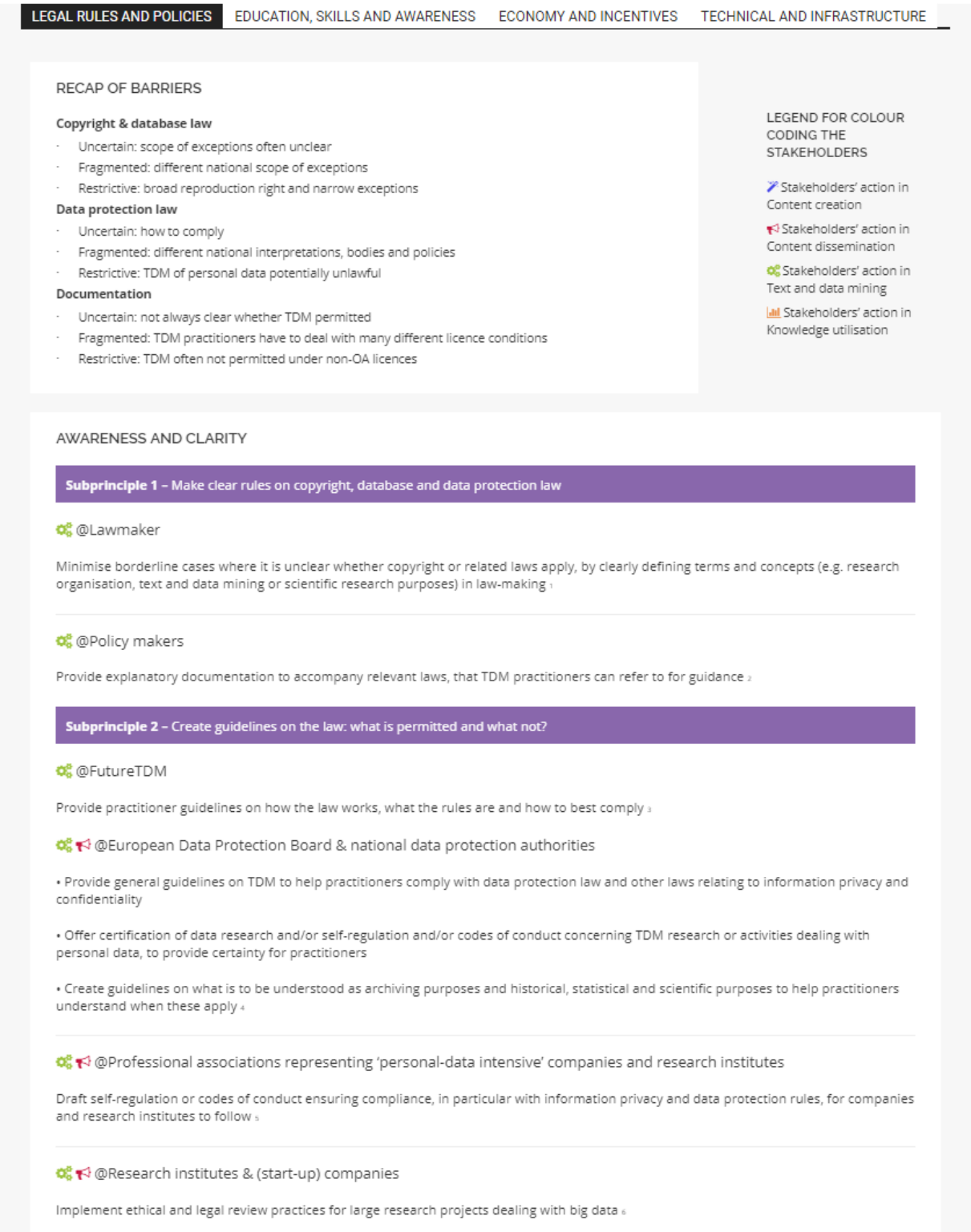


Figure 9: Recommendations concerning awareness and clarity in Technical and Infrastructure

3 PRACTITIONER GUIDELINES

The Knowledge Base section devoted to the **practitioner guidelines** has been created to help address specific challenges to greater use of text and data mining (TDM) technologies. The FutureTDM project has produced several sets of guidelines[6] which are grouped in four categories (see Figure 10) addressing mainly, but not exclusively, data providers and users as anyone else, from academia to industry to the general public, who is interested in TDM. In specific one can find guidelines for:

- TDM practitioners on legal issues
- TDM practitioners on issues related to licences and licencing
- Researchers on Data Management
- Universities on curricula and education issues

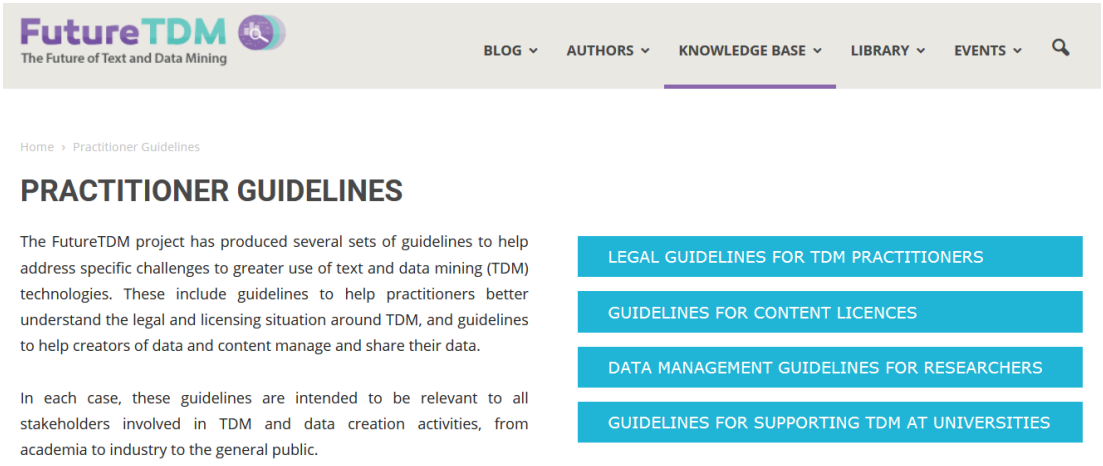


Figure 10: Practitioner Guidelines

The first set of guidelines is entitled **Legal Guidelines for TDM Practitioners** and is based on the realization that the legal landscape around TDM is more complicated than TDM practitioners may believe. For this reason, it provides an overview of the legal landscape around TDM (e.g. what is public domain information, what is personal data etc.) so that legal risk is minimized and practitioners are protected. These guidelines are not intended to be comprehensive legal advice though; they, rather, aim to give TDM practitioners a foundational overview of relevant legal considerations and to help understand when it might be necessary to seek expert legal advice. In order to make it more comprehensible, information is presented either as a step to step plan, as for example when approaching content for mining, or broken down to questions and answers (see Figure 11).

[6] For a detailed description of the guidelines, please see D5.3 FutureTDM practitioner guidelines, available at http://www.futuretdm.eu/wp-content/uploads/FutureTDM_D5.3-FutureTDM-practitioner-guidelines.pdf

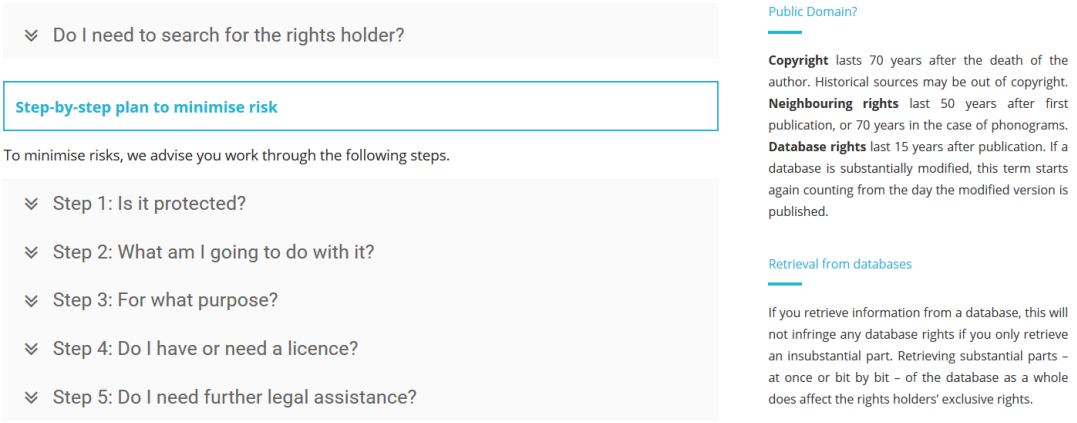


Figure 11: Legal Guidelines for TDM Practitioners

There is also an easily memorizable distinction of DO's and DON'Ts with lists of the permitted vs non-accepted TDM policies and practices (see Figure 12) as well as a link at the bottom of the page to the relevant deliverable for those who wish to delve into more detail.

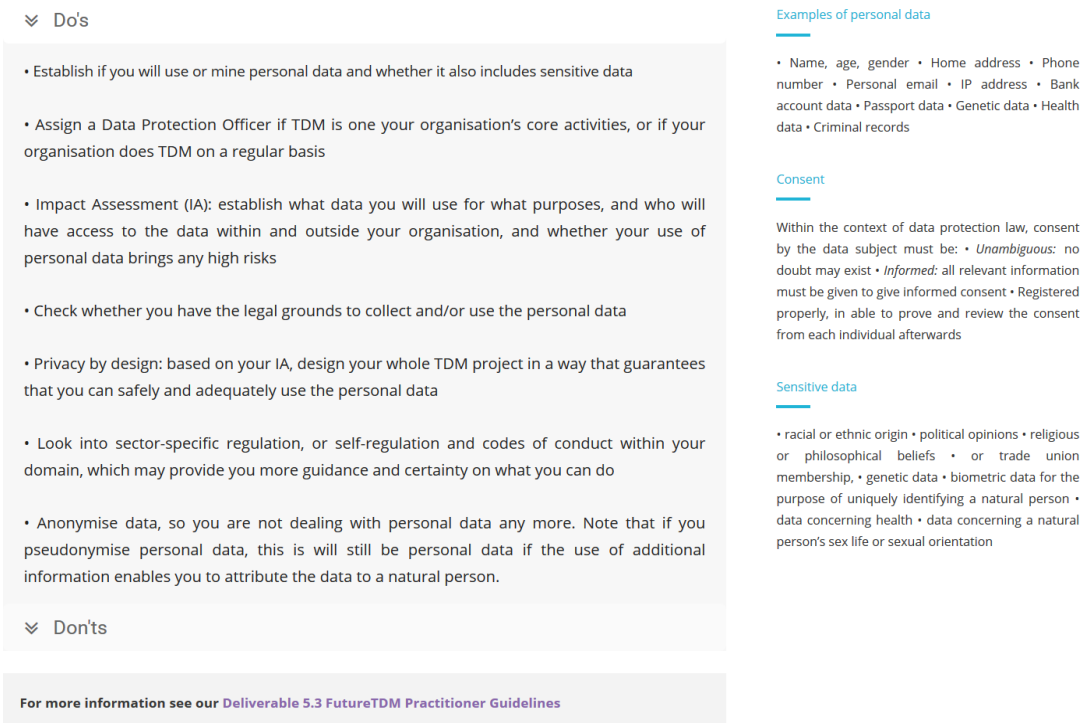
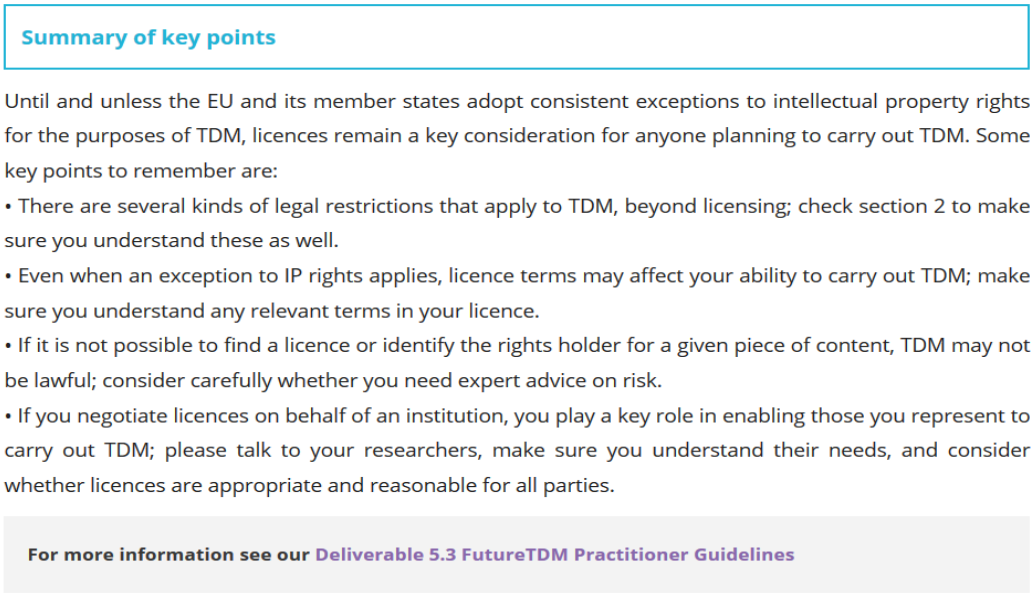


Figure 12: Do's and Don'ts in the Legal Guidelines section

The following set of guidelines is entitled **Guidelines for Content Licences** and is dedicated to various intellectual property (IP) laws that can be relevant to the use of content for TDM activities. It follows the same structure with questions and answers described earlier and in addition it offers a summary of the most important key points to be remembered (see Figure 13).



Summary of key points

Until and unless the EU and its member states adopt consistent exceptions to intellectual property rights for the purposes of TDM, licences remain a key consideration for anyone planning to carry out TDM. Some key points to remember are:

• There are several kinds of legal restrictions that apply to TDM, beyond licensing; check section 2 to make sure you understand these as well.

• Even when an exception to IP rights applies, licence terms may affect your ability to carry out TDM; make sure you understand any relevant terms in your licence.

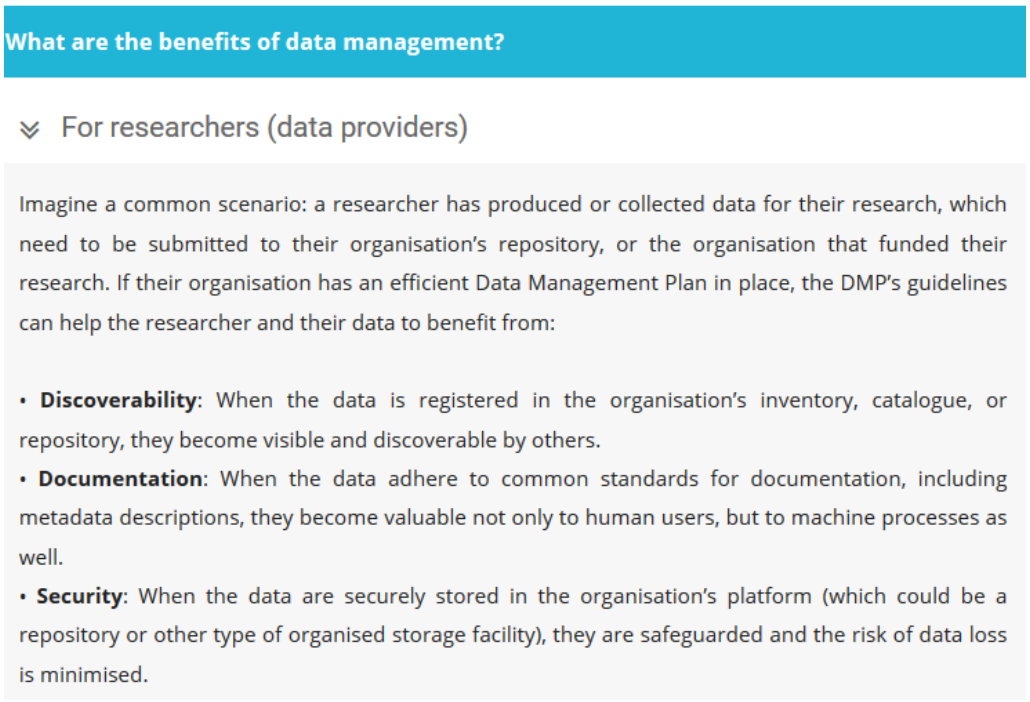
• If it is not possible to find a licence or identify the rights holder for a given piece of content, TDM may not be lawful; consider carefully whether you need expert advice on risk.

• If you negotiate licences on behalf of an institution, you play a key role in enabling those you represent to carry out TDM; please talk to your researchers, make sure you understand their needs, and consider whether licences are appropriate and reasonable for all parties.

For more information see our Deliverable 5.3 FutureTDM Practitioner Guidelines

Figure 13: Summary of key points in the Guidelines for Content Licences section

Next, there is a set of guidelines aiming primarily at academic researchers who collect, create, store and share data (Data Management Guidelines for Researchers) in order to help them make sure their data is reusable. This subsection clarifies what data management is and reveals all the benefits deriving from it.



What are the benefits of data management?

For researchers (data providers)

Imagine a common scenario: a researcher has produced or collected data for their research, which need to be submitted to their organisation's repository, or the organisation that funded their research. If their organisation has an efficient Data Management Plan in place, the DMP's guidelines can help the researcher and their data to benefit from:

• **Discoverability**: When the data is registered in the organisation's inventory, catalogue, or repository, they become visible and discoverable by others.

• **Documentation**: When the data adhere to common standards for documentation, including metadata descriptions, they become valuable not only to human users, but to machine processes as well.

• **Security**: When the data are securely stored in the organisation's platform (which could be a repository or other type of organised storage facility), they are safeguarded and the risk of data loss is minimised.

Figure 14: Data management benefits

This subsection addresses specific questions, such as how to identify a repository in an appropriate domain for data storage, or guidelines on how to secure validation and quality of data and metadata

(see Figure 15). All the guidelines have been created in order to make data by researchers as valuable for reuse as possible.

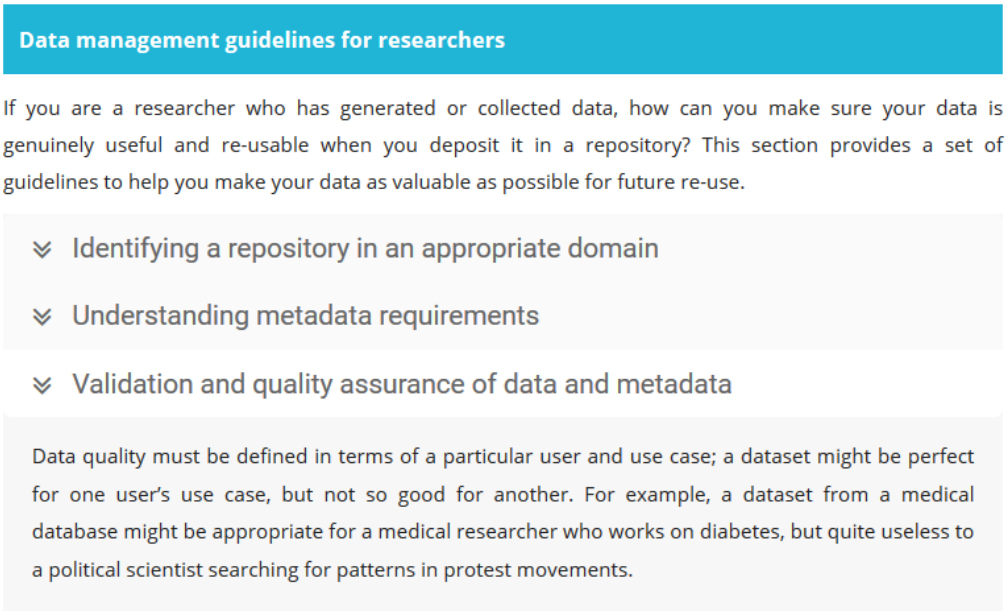


Figure 15: Data management guidelines for researchers

Last, there is subsection devoted to Universities (**Guidelines for Supporting TDM at Universities**) as their role has been identified of crucial importance for the TDM uptake. Here the focus is on the necessary skills and education programmes to provide future users with them (see Figure 16).

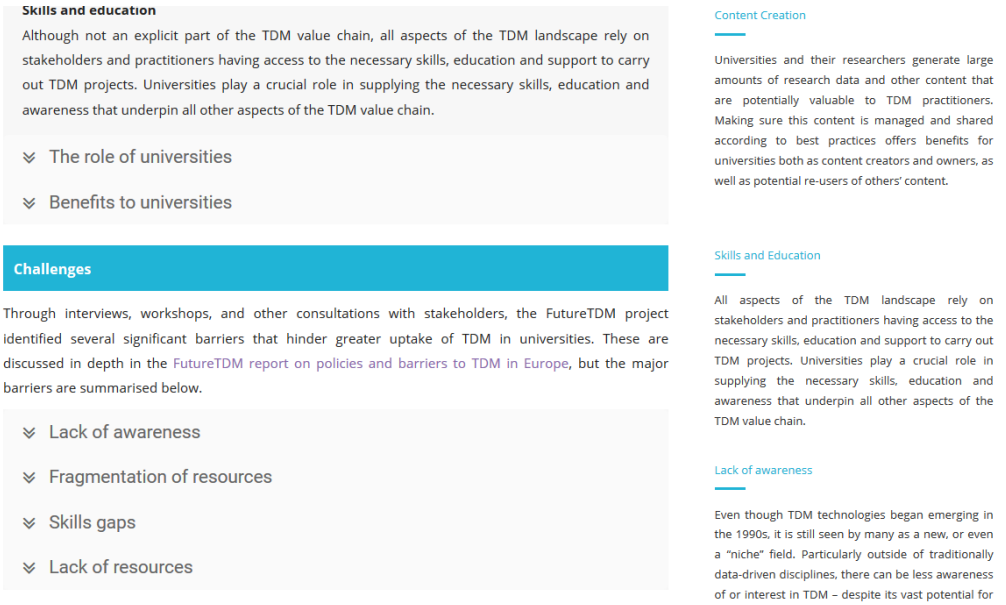


Figure 16: Guidelines for supporting TDM at Universities

Again, the possible benefits from university engagement with TDM is highlighted (see Figure 17). As explicitly stated "Data is the new IT" and Universities should be prepared to provide researchers with the necessary tools to make the best use of it.

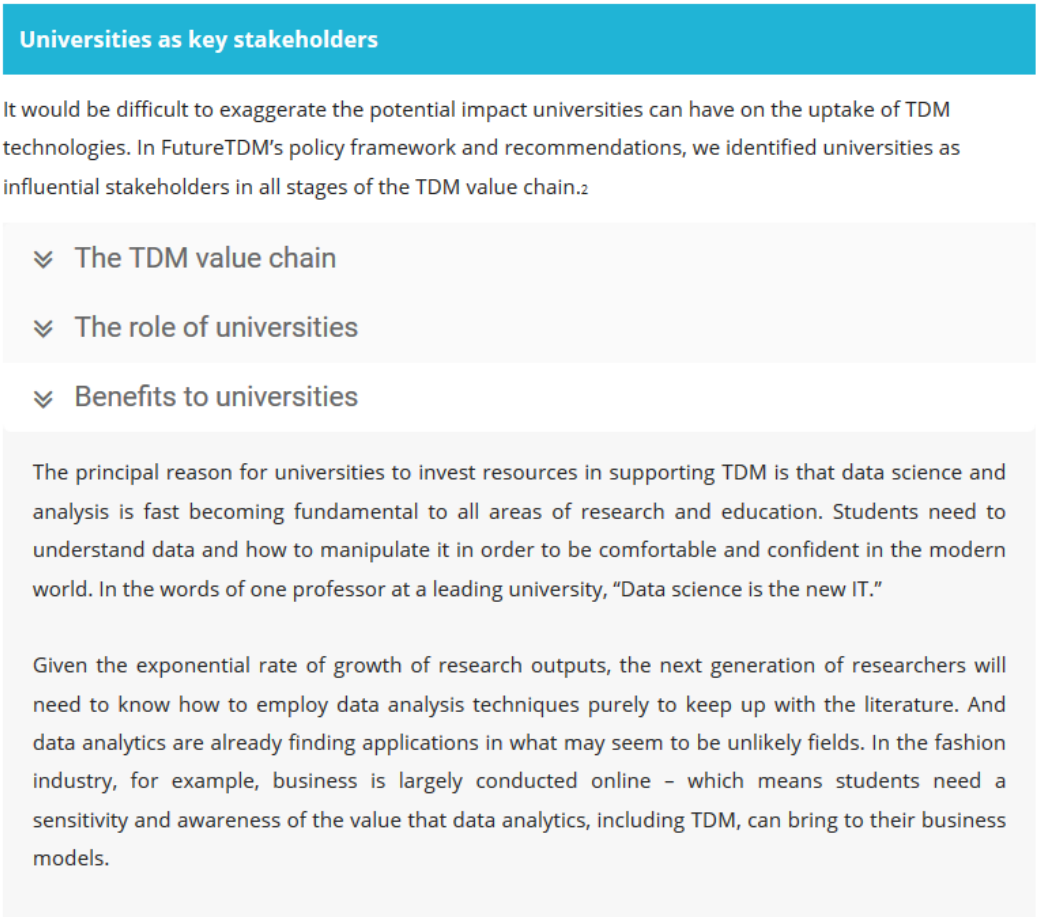


Figure 17: Benefits to universities

4 ROADMAP

As mentioned in the previous chapters, over the past 22 months, the project has engaged with stakeholders from all across Europe[7] in order to develop a comprehensive understanding of the current TDM landscape, the barriers that exist, and a policy framework through which to address those barriers[8]. FutureTDM has created a roadmap that brings together the outcomes of all the research and mobilization activities of the FutureTDM project to present an EU-level path to stimulating TDM opportunities in Europe. This **roadmap** has its own dedicated section in the platform (see Figure 18): a plan is outlined to increase the uptake of TDM technologies in Europe, focusing on three conceptual phases:

- **Phase I: Content Availability**
- **Phase II: Support Early Adopters**
- **Phase III: The Next Generation**

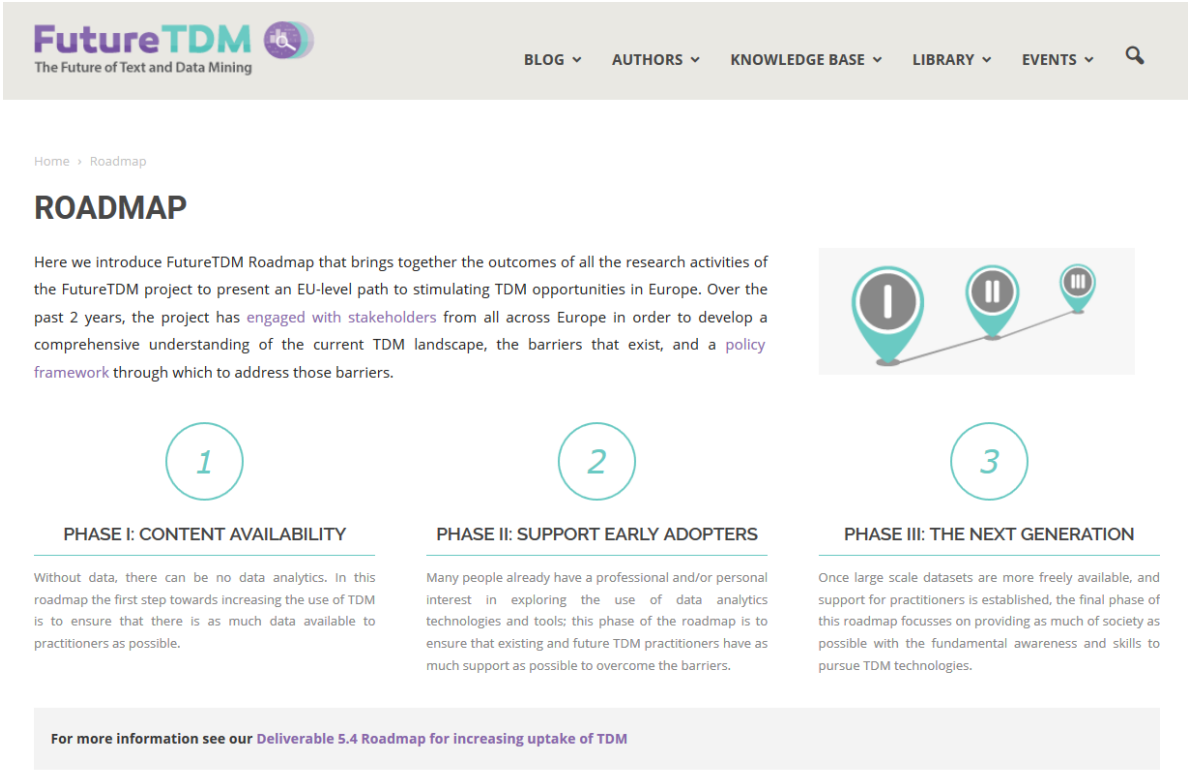


Figure 18: Roadmap

Each phase of the roadmap is divided into five sections (see Figure 19):

- A single statement of the overall **Objective** of this phase;
- An outline of the current **Situation** regarding TDM, in the context of this phase;
- A breakdown of the main **Challenges** to achieving this phase's objective;

[7] At FutureTDM Knowledge Cafes and other community events
[8] D5.1 FutureTDM Policy Framework

- Key **Principles** to address each challenge highlighted;
- **Activities** that the Commission and other parties should engage in to implement those principles and meet the phase's objective.

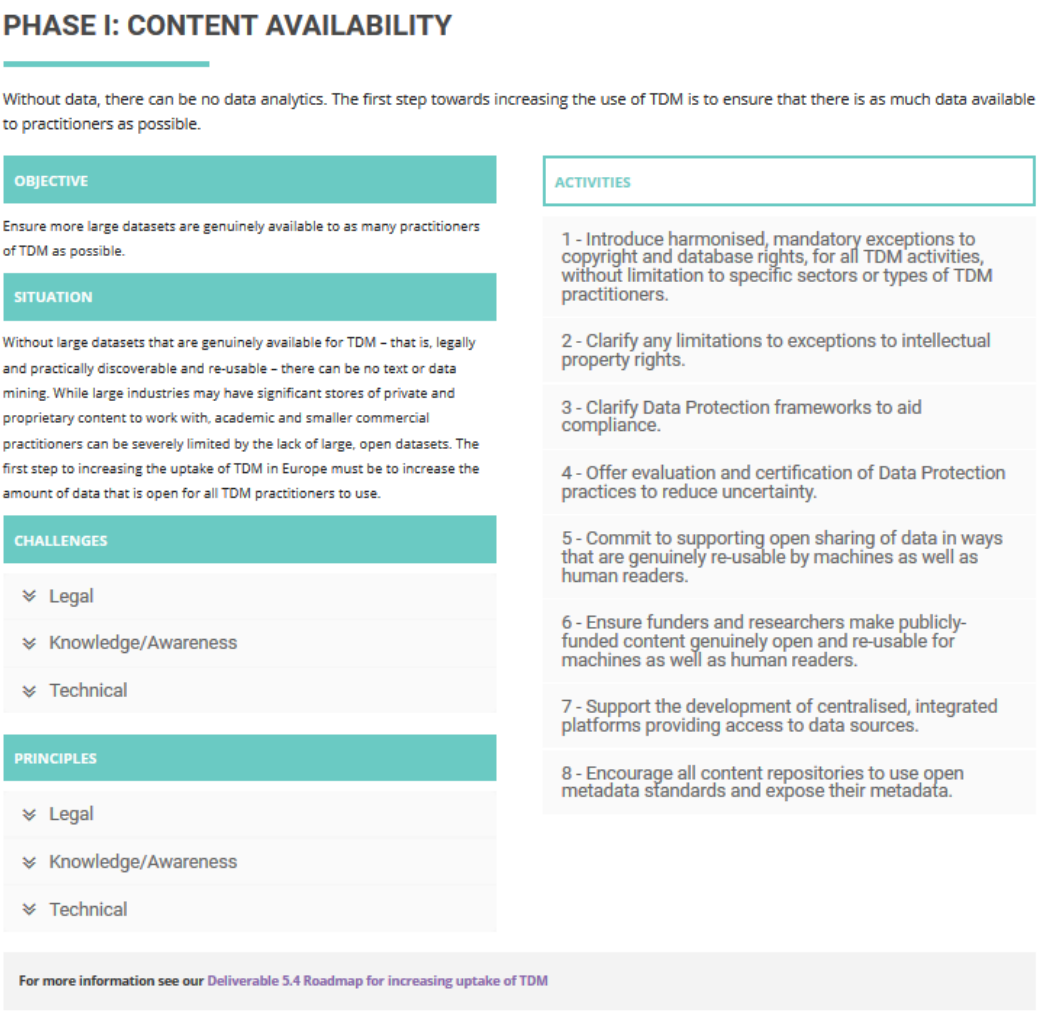


Figure 19: Roadmap Phase I: Content Availability

The challenges, principles and suggested activities outlined are all based on the extensive research and stakeholder engagement carried out over the course of the FutureTDM project – particularly the FutureTDM Policy Framework, which gives a more comprehensive overview of the ways in which all stakeholders in the TDM landscape can help to support the uptake of TDM technologies in Europe.

The Roadmap is available on the FutureTDM platform to increase the visibility and accessibility of these guidelines. The respective deliverable is also available online at the bottom of each roadmap webpage.

5 CONCLUSION

This stable and flexible FutureTDM platform wishes to empower practitioners with the resources, guidelines and solutions to promote the uptake of TDM in Europe. The platform was presented latest in the FutureTDM Symposium which took place within the iDSC Conference in Salzburg Austria on June 12th - 13th 2017 and received very positive comments. As indicated in almost all pages of the platform, feedback from stakeholders and any suggestions for improvement are welcome.

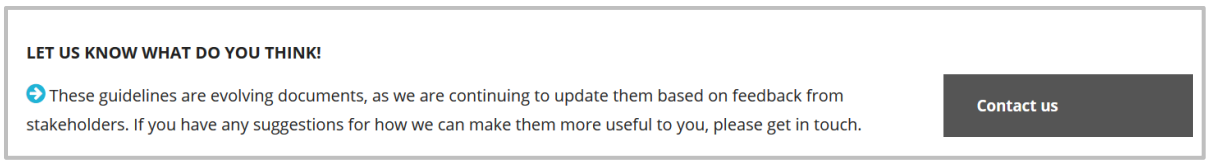


Figure 20: Feedback requested on the FTDM platform

The platform will continue to be revised and updated, as it has throughout the project, in order to maintain a coherent and up-to-date view on the European TDM landscape.