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 D5.3

FutureTDM Practitioner Guidelines

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:	D5.3
Title:	**FutureTDM practitioner guidelines**
Lead beneficiary:	**The British Library**
Work package:	WP5: Elaborate: Legal Framework, policy priorities, roadmaps and practitioner guidelines
Dissemination level:	Public (PU)
Nature:	Report (RE)
Due date:	30.04.2017
Submission date:	01.05.2017
Authors:	**Kiera McNeice**, BL **Marco Caspers**, UVA **Maria Gavriilidou**, ARC
Review:	**Marco Caspers**, UVA

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. INTRODUCTION

This document represents the first iteration of a series of practitioner guidelines for the increased usage and implementation of text and data mining (TDM) in various sectors. They include:

- In section 2, guidelines to help TDM practitioners understand the legal landscape around TDM and reduce legal uncertainty.
- In section 3, guidelines for those who wish or need to license others' content for TDM, detailing considerations that may be relevant to content licences.
- In section 4, guidelines outlining best practices for data management, to help researchers manage and share their content in ways that make it genuinely valuable for re-use for TDM.
- In section 5, guidelines outlining how universities can take practical steps towards supporting TDM in a coordinated, centralised way.

These guidelines will be further disseminated as separate documents via the FutureTDM website,[1] blog, and dedicated awareness sheets that will be shared with relevant stakeholders. They will also be supplemented by case studies covering illustrative examples, again distributed via the FutureTDM website and awareness sheets.

The guidelines presented in this first iteration will be updated throughout the remainder of the project, based on feedback and further engagement with stakeholders, to ensure that they are as relevant and useful as possible.

[1] http://www.futuretdm.eu/knowledge-library/

2. LEGAL GUIDELINES FOR TDM PRACTITIONERS

2.1 Introduction

The legal landscape around TDM is more complicated than TDM practitioners may realise. In fact, there are many instances in which TDM is potentially unlawful. These guidelines are intended to give practitioners an overview of the legal landscape around TDM, so that they can be aware of potential legal issues and minimise legal risk.

In the absence of clear legal exceptions, intellectual property rights, including copyright, neighbouring rights, and *sui generis* database rights, will almost certainly be relevant when working with content created by others; these are discussed in section 2.3. TDM practitioners should also be careful to respect personal data and the privacy of any data subjects; data protection laws and best practices are discussed in section 2.4.

These guidelines are not intended to be comprehensive legal advice. Rather, they aim to give TDM practitioners a foundational overview of relevant legal considerations, to help understand when it might be necessary to seek expert legal advice.

2.2 Relevant legal considerations

To identify the legal risks of TDM, we first need to understand the activities involved in the TDM process. The outline in Figure 1 shows four general phases in the TDM process, with examples of acts that may be carried out in each phase. (Not everyone carrying out TDM will necessarily need to do all of these things, depending on the type of TDM and the purpose for which it is carried out.)

These phases will be referred to throughout these guidelines.

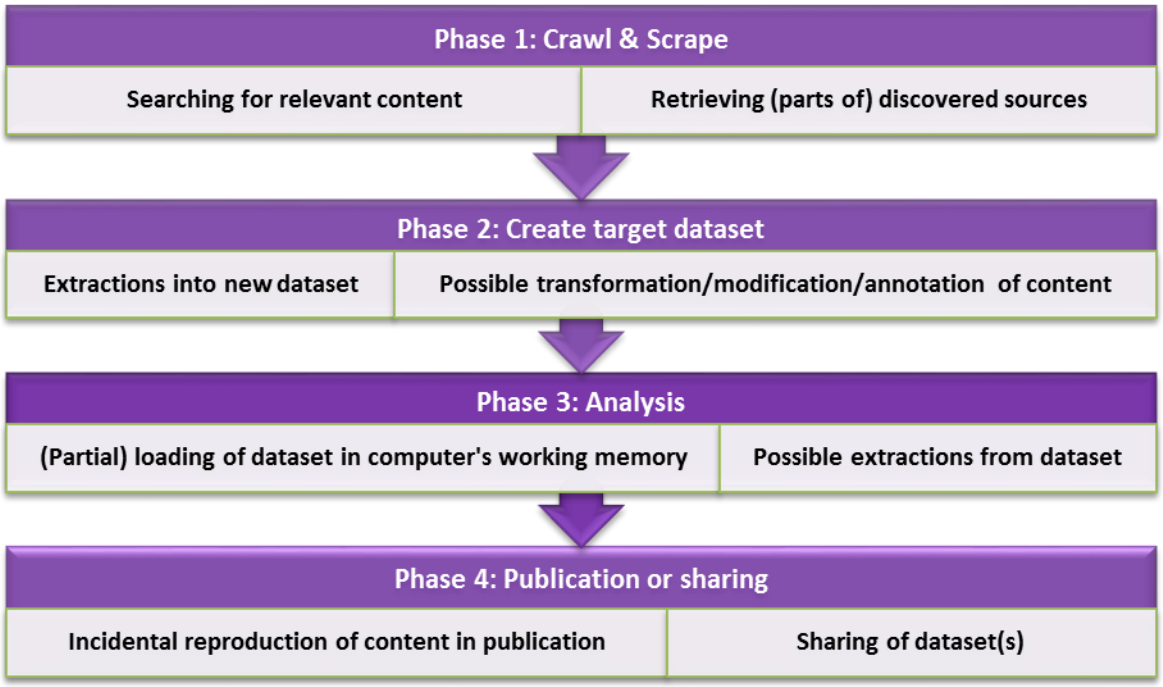

Figure 1: Generalised overview of TDM processes

Before starting any TDM project, it is important to assess the potential legal issues of your project, and plan to avoid or minimise legal risks in your project design. For this purpose, some key questions you should ask about any TDM undertaking are:

- What sort of content am I going to use, and is it protected by or subject to any regulation?
- What sort of acts will I be carrying out on the content, and are these acts subject to specific rules under the relevant regulation?
- How should I deal with any applicable regulation to prevent or minimise the risk of my TDM project being rendered unlawful?
- In which cases should I turn to professional legal advice?

These guidelines should help you to carry out a rough evaluation of the legal risks of your TDM project, and to assess whether you should seek further legal advice.

When it comes to protected content, the two most common legal regimes that miners – at least in Europe – will face in practice are:

1. **Intellectual property rights**, more specifically copyrights, neighbouring rights and database rights
2. **Data protection rules**

We will look at these regimes separately to help answer the questions posed above.

2.3 Mining others' intellectual property

Many TDM activities are carried out using content that is the intellectual property of other people, and subject to intellectual property (IP) rights.

2.3.1 When are IP rights relevant?

When mining content, there are three kinds of protection you need to consider: *copyrights, neighbouring rights* and *database rights*. These are the intellectual property rights that may be attached to the content you are intending to mine. It is important to establish whether any of these rights exist in the content you will be mining, because if they do, you might need permission from the right holders involved.

***Note that facts and data are not creative expressions, and do not attract copyright. Pure 'data mining' is therefore less likely to infringe copyright, except for the copyrights possibly existing in the collection of those data. Conversely, 'text mining' – including mining of other rich contents, such as images, films and music – is highly likely to be affected by copyright or neighbouring rights. In both text and data mining, you should always be aware of database rights in the collections of data, text or other contents.*

If you are dealing with any content similar to the examples in Figure 1, you should be aware that your TDM project could potentially infringe IP rights if you do not have permission from the rights holder. The following sections will help you evaluate whether you need to undertake further action.

2.3.2 Do I need to search for the rights holder?

If you have determined that you are dealing with protected content, you should establish what you are going to do with that content, and verify whether this is something that needs the consent of the IP rights holders. This is generally the case when you *copy*, either permanently or temporarily, or *publish* those contents in whole or in part.

Copying content: In phases 1 to 3 of Figure 1, TDM activities usually involve making copies of content or (parts of) databases, ranging from retrieving copies from one or more sources, to transforming the contents into a (formalised) dataset that will be loaded into the computer's working memory when performing TDM analysis.

Publishing content: If you are planning to share or disseminate any of your TDM results, or the underlying data or content sources, this is likely to be considered "publishing" - and will need permission in most instances as it relates to the exclusive rights of the rights holder to control the communication or redistribution of their content.

These acts need to be authorised by rights holders, unless special exceptions apply. Despite the existence of common European rules on copyrights and database rights, the applicability of and scope of these exceptions vary significantly across national borders. This means that if you work in multiple countries or collaborate with foreign colleagues even within the EU, you will need to assess any relevant exceptions for each country you are operating in.[2]

As of April 2017, only the UK and France[3] have introduced exceptions in their laws that specifically allow you to use content for TDM without permission from rights holders. The UK exception only applies to copyright law, although a general non-commercial research exception exists for database rights in the UK. The French TDM exception applies to both copyright and database rights. In both countries, due to restrictions within the European Copyright Directive, these exceptions are limited to TDM for non-commercial and scientific research purposes where users have lawful access to content – for example because they have subscriptions to journals, or because they are freely available websites on the internet. These exceptions may benefit for example university researchers, whose research is for non-commercial scientific purposes. However, it is not entirely clear-cut when these *non-commercial* and *scientific research* conditions are met. For example, researchers involved in consortia with industry partners cannot be sure that they can benefit from such an exception.

In many European countries, other exceptions may also exist if you use content for:

- *Private and non-commercial purposes*: This may allow you to do text mining for your own private use.
- *Non-commercial research or teaching purposes in general*: some EU member states have an exception for certain acts carried out for research, some for teaching, and some for both. The scope of these is often very narrow and therefore unlikely to cover a full TDM process, if at all.
- *Temporary copies necessary to enable lawful use of a work*: This exception exists in all EU countries and may in many cases permit the part of the TDM process where the contents are temporarily loaded into the computer's working memory, although uncertainty exists regarding the extent to which this exception allows this.

These exceptions generally only permit TDM under either very specific circumstances, or one or a few phases in the TDM process.[4]

2.3.3 Step-by-step plan to minimise risk

To minimise risks, we advise you work through the following steps.

Public domain?

Copyright lasts 70 years after the death of the author. Historical sources may be out of copyright.

Neighbouring rights last 50 years after first publication, or 70 years in the case of phonograms.

Database rights last 15 years after publication. If a database is substantially modified, this term starts again counting from the day the modified version is

[2] You can find an overview of implementations of exceptions at http://copyrightexceptions.eu.

[3] At the time of writing, a decree that would further detail the application of the French TDM exception was rejected by the Conseil D'Etat. Therefore the exception, despite being in the Intellectual Property Code, is not in force yet.

[4] For a detailed review of laws and policies affecting TDM, see FutureTDM Deliverable D3.3 (PDF)

Step 1: Is it protected?

Establish whether the content to be mined is potentially protected by any copyrights, neighbouring rights or database rights. If yes, establish whether the corpus or whole body of contents is in the public domain, because all rights have lapsed.

Step 2: What am I going to do with it?

If you carry out any of the first three steps of the TDM process (Figure 1), you are likely to make copies that are subject to any right holders' approval. Such approval is also necessary when you publish or share TDM results, when these results contain original or modified versions of the contents you mined.

Retrieval from databases

If you retrieve information from a database, this will not infringe any database rights if you only retrieve an insubstantial part. Retrieving substantial parts – at once or bit by bit – of the database as a whole **does** affect the rights holders' exclusive rights.

Step 3: For what purpose?

Approval is not necessary when your activities are subject to an exception. For example, in some European countries, this may be the case when you make reproductions (such as those in steps 1 to 3) for non-commercial private or research purposes. No exception will apply if you share the full set of contents that you mined, but quoting from works in, for example, a research paper might be permitted in many European (and other) countries. Further, the sharing of facts and aggregated data (such as statistical representations), and new knowledge (such as newly created semantic annotations for TDM) always remains free if no original content is being shared.

Step 4: Do I have or need a licence?

In most cases, especially outside of a non-commercial private or research context, you cannot rely on exceptions to IP rights within Europe. Therefore, you should check whether you have an appropriate licence to mine the content. Even when you might benefit from an exception, that exception might be overridable by the terms of your contract. Therefore: always check your licences!

For our licensing guidelines, please see section 3.

Step 5: Do I need further legal assistance?

It is always better to be safe than sorry. If you have any doubt whether you:

- deal with protected sources,
- can rely on any exception, or
- should have a licence,

please consult a legal expert within your organisation or seek advice from an external lawyer. If you are a TDM user who belongs to an academic or other institution, please check with the right person within your organisation about your licences to use content. It might be even safer to take this step before going through the other steps!

2.3.4 Further materials

European harmonising directives

- Copyright & Neighbouring Rights: Copyright Directive 2001/29/EC | http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32001L0029
- Database Rights: Database Directive 96/9/EC | http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A31996L0009
    - And its national transitions: http://eur-lex.europa.eu/legal-content/EN/NIM/?uri=CELEX:31996L0009

Information on national copyright laws

- National copyright exceptions: http://copyrightexceptions.eu
- Sources for national IP legislation: http://www.wipo.int/wipolex/en/
- UK IPO on copyright and research: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/375954/Research.pdf
- Tool for calculating if works are out of copyright: http://outofcopyright.eu

Other

- FutureTDM deliverable elaborating on legal barriers: http://www.futuretdm.eu/knowledge-library/?b5-file=2374&b5-folder=2227

2.4 Mining personal data

Important!

These guidelines alone are not sufficient to tell you how you should work with personal data, as this must be assessed carefully on a case-by-case basis. The guidelines are rather meant as an introduction to the principles and duties of data protection law. We always recommend you integrate data protection principles in the whole design of your TDM project, and always consult a data protection expert within or from outside of your organisation before you commence any TDM project involving personal data.

2.4.1 When is data protection relevant?

In Europe, you have to comply with specific regulations when you are dealing with (or 'processing') personal data. This means that when you mine any data relating to individuals, you should be aware of the rights and duties that come with it. Personal data is any data that relates to an identified or identifiable living[5] person, and can cover any sort of data as long as it enables you to directly or indirectly identify an individual. It also includes opinions about living individuals. Virtually *anything* you do with personal data is bound by European data

Examples of personal data

- Name, age, gender
- Home address
- Phone number
- Personal email
- IP address
- Bank account data
- Passport data
- Genetic data
- Health data
- Criminal records

[5] Be aware that specific (self-)regulation or codes in some countries c sed persons.

protection rules, ranging from collecting and storing data to modifying or removing them. Therefore, if you deal with personal data in any of the phases in the TDM process (see Figure 1**Error! Reference source not found.**), you will need to comply with data protection law.

2.4.2 Principles and duties of European data protection law

Collection and further use

For the purpose of these guidelines, we distinguish three types of data use in the context of mining:

1. *Collection of personal data*: Retrieving any personal data directly from individuals or other sources (re-use of data).
2. *Use of personal data*: mining by you, someone else within your organisation, or on your behalf, of the retrieved data.
3. *Transfer of data*: transferring data to other parties.

Data minimisation vs. maximisation

There is a peculiar contradiction between the *data maximisation* (collecting and using as much data as possible) goal that makes big data and TDM so valuable, and the *data minimisation* principle of data protection regulation. The data minimisation principle entails the following:

- Personal data should only be collected for specified, explicit and legitimate purposes.
- Further use of data should be carried out in a manner compatible with the purposes for which they were collected (purpose limitation).
- Any use must be adequate, relevant and limited to what is necessary for those purposes.

Consent

Within the context of data protection law, consent by the data subject must be:

- Unambiguous: no doubt may exist
- Informed: all relevant information must be given to give informed consent
- Registered properly, in able to prove and review the consent from each individual afterwards

Rights and duties

Personal data may only be processed on the basis of one of the following legal grounds:

- *Consent*: the person (data subject) to which the data relates has given their consent for the specified purposes.
- *Contract*: the use of the data is necessary to comply with a contract to which the data subject is party.
- *Legitimate interest*: you have a legitimate interest in using the data, which overrides the fundamental rights and interests of the data subjects, although public sector bodies may not rely on this anymore from May 2018.
- *Compliance with legal obligations*, *protection of the vital interests of the data subject*, or *performance of public interest task by official authority*: these grounds will generally not be relevant in the context of TDM.

Other duties:

- Notify the relevant data protection authority that you process personal data. This *general* obligation will be abolished as of May 2018, and be replaced by procedures and mechanisms that rather focus on types of data use involving high risks. For example, notifications will have to be made in case of data breaches.

- Inform data subjects of your activities, if they are not already informed.

Rights of the data subject:

- Right to be informed
- Right to access their data
- Right to object to use of their data

Mining sensitive data

European data protection law has a stricter regime for dealing with *sensitive* data. This is generally prohibited, unless you have legal grounds.

Sensitive data

- racial or ethnic origin
- political opinions
- religious or philosophical beliefs
- or trade union membership,
- genetic data
- biometric data for the purpose of uniquely identifying a natural person
- data concerning health
- data concerning a natural person's sex life or sexual orientation

2.4.3 Special provisions for research

On several aspects, the data protection framework provides for a lighter regime when personal data is used for *scientific or historical research* purposes. We give a few examples.

Purpose limitation and storage

Data must be processed for no other purposes than those for which the individual has given their consent to. With scientific research, however, it is often not possible to fully identify the purposes for which personal data are collected. Here the data protection framework has some leeway for scientific research: Further processing of collected data for scientific or historical research purposes will be considered to be compatible with the initial purposes for which the data is collected. Further, where data may normally be stored no longer than necessary for these initial purposes, longer storage is permitted when solely for scientific or historical research.

What is "research"?

It is not entirely clear what exactly constitutes "scientific" or "historical" research. However, when you carry out academic research adhering to academic standards, it will be more likely that this is regarded to be scientific research, than research carried out in an industrial context.

Be aware of special rules

When you use personal data in your scientific research, be sure to check whether in your research domain, or for the type of data you use, particular (domain-specific) regulation, self-regulation or codes of conduct apply. Such special rules commonly exist for the use of medical data or patient records.

Data not collected from individuals

When data is re-used from other sources, the TDM researcher has not collected the data from the individuals themselves. Normally in such cases, they will have to inform all involved data subjects on the use of the data. However, in the particular case of using data for the purpose of scientific or historical research, a TDM researcher will not have to do this if it would be impossible or require disproportionate effort.

Right to be "forgotten"

Data subjects have a right to obtain the erasure of their personal data – to be "forgotten" – for example when the storing of their data is no longer necessary

for the purposes for which it was collected or otherwise processed. This does not apply where the use of personal data is necessary for scientific or historical research purposes.

Safeguards

When personal data is being used for scientific or historical research, this research must be bound by appropriate safeguards, to respect the rights and freedoms of data subjects. If identification of data subjects is no longer necessary to fulfil the research purposes, this data shall be used in a manner that does not enable identification (through *pseudonymisation* or *anonymisation* of the data).

2.4.4 Do's and don'ts

We cannot provide general guidelines on how each TDM project should deal with personal data, since this largely depends on the scale, nature and purpose of the TDM activities, as well as on the nature and source of the personal data. Dealing with data protection law and ethics is very complex and we therefore strongly recommend you always consult an expert in this area when designing your TDM project. This section provides lists of *do's* and *don'ts* to give you some guidance as to the most important aspects of dealing with personal data in your TDM project.

Do's

- Establish if you will use or mine personal data and whether it also includes sensitive data
- Assign a Data Protection Officer if TDM is one your organisation's core activities, or if your organisation does TDM on a regular basis
- *Impact Assessment* (IA): establish what data you will use for what purposes, and who will have access to the data within and outside your organisation, and whether your use of personal data brings any high risks
- Check whether you have the legal grounds to collect and/or use the personal data
- *Privacy by design*: based on your IA, design your whole TDM project in a way that guarantees that you can safely and adequately use the personal data
- Look into sector-specific regulation, or self-regulation and codes of conduct within your domain, which may provide you more guidance and certainty on what you can do
- Anonymise data, so you are not dealing with personal data any more. Note that if you pseudonymise personal data, this is will still be personal data if the use of additional information enables you to attribute the data to a natural person.

Don'ts

- Only think of data protection issues when you actually start to mine
- Collect data and just assume that it does not concern any personal data
- Store and retain all data just because it may be useful in the future
- Randomly transfer or provide access to any data to third parties
- Re-use data from one project in another one, without making sure this is compatible with data protection rules, *even though* you had made sure that the use in the first project was compatible
- Share any personal data with the public, without proper consultation
- Make decisions affecting the data subject based solely on automated processing of their personal data – this is prohibited
- Ignore data subjects' requests to access, rectify or erase data
- Transfer data outside the EU

2.4.5 Further materials

European legislation

- Data Protection Directive 95/46/EC: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:31995L0046
  - This directive does not apply directly to European citizens and organisations, but is implemented in the national laws of its member states
- General Data Protection Regulation: https://publications.europa.eu/en/publication-detail/-/publication/3e485e15-11bd-11e6-ba9a-01aa75ed71a1/language-en
  - Will apply from May 25, 2018
  - From that date, will repeal the above-mentioned Data Protection Directive

Institutions

- List of national Data Protection Authorities: http://ec.europa.eu/justice/data-protection/article-29/structure/data-protection-authorities/index_en.htm
- Glossary of data protection terms: https://edps.europa.eu/data-protection/data-protection/glossary_en
- UK ICO *Guide to data protection*: https://ico.org.uk/for-organisations/guide-to-data-protection/

Other

- FutureTDM deliverable elaborating on legal barriers: http://www.futuretdm.eu/knowledge-library/?b5-file=2374&b5-folder=2227

3. GUIDELINES FOR CONTENT LICENSEES

3.1 Introduction

In section 2 of this report, we discuss the various intellectual property (IP) laws that can be relevant to the use of content for TDM activities. In some cases, exceptions to these laws apply, making it possible to re-use others' content for TDM without needing specific permission from rights holders.

In other cases, however, the law does not allow TDM without permission from rights holders. And even in cases where legal exceptions allow for TDM without explicit permission from rights holders, there may be scope for some aspects of a licence or contract to limit how you may use content for TDM.

In these guidelines, we will discuss what you should be aware of when licensing content for use in TDM. If you plan to carry out a TDM project – or if you are involved with helping others access content for TDM, for example via an institutional library – these guidelines aim to help you understand what kinds of licences can best support TDM activities, and what you may want to consider if you need to negotiate licence agreements with rights holders for TDM.

3.2 Do I need a licence?

This is a question you should consider in the planning stages of any TDM project. Section 2.3.3 sets out a step-by-step plan to minimise legal risk in a TDM project, including deciding whether or not you need a licence for your planned activities.

To recap briefly, there are two main ways in which TDM activities may come into conflict with IP rights:[6]

- Copying content: TDM activities usually involve making copies of content, databases, or parts thereof to transform and use for TDM analysis.
- Publishing content: When publishing the results of your TDM activities, you may want to reproduce parts of the contents used in your analysis.

If you or the researchers you support plan to carry out a TDM project, you should consider whether the project will involve copying and/or publishing parts of others' content. If so, it may be the case that a legal exception allows you to carry out these activities without permission from the rights holder (see section 2.3.2). Otherwise you will need permission from the rights holder to carry out TDM on their content, and should check whether this is allowed by any existing licence you may have. If not, you will need to negotiate an appropriate licence.

3.2.1 Limitations to exceptions

Even in cases where legal exceptions to IP rights apply, they may not actually allow all the TDM activities you plan to carry out. Exceptions can be limited if they are overridable by contract, or if they have caveats that leave scope for the rights holder to define other limitations. In these cases,

[6] Other laws may affect what you can do with other people's content, including Data Protection laws around privacy – see Section 2 for details of other legal considerations.

the terms of any licence or agreement you have with the rights holder might still limit what kinds of TDM are permissible.

Copying content

In the UK, a specific exception from copyright applies if you are copying content for the purpose of non-commercial TDM research. This exception also includes a clause that ensures it cannot be overridden by any contract.[7] Without such a clause, any future exceptions adopted by the EU or its member states could be overridden by the terms of licences agreed with rights holders.

The UK exception also includes the caveats that the exception only applies in cases where TDM practitioners have "lawful access" to content, and that practitioners must not circumvent "technical protection measures" (TPMs) in order to make copies of content. In the case of content protected by IP rights, lawful access is of course defined by the terms of your agreement with the rights holder. For example, in the case of the internet this means … Rights holders also have scope to decide what technical protection measures are appropriate to protect the security and stability of their content and services. Therefore even if you are carrying out TDM under a legal exception, you should be aware of:

- The terms under which the rights holder has granted you lawful access to their content; these may include limitations on who can be considered an authorised user, how many authorised users may lawfully access their content, and from where they may access it.
- The details of any technical protection measures that apply to their content; these may include user authentication processes to verify lawful access, or limitations on the amount of content you may access in a given timeframe. (However note that these protection measures must not "unreasonably" prevent you from benefiting from the exception.)

These terms may restrict your ability to carry out TDM, even under a copyright exception.

Publishing content

Various legal exceptions may allow you to reproduce and publish parts of the original contents used in your TDM analysis, alongside the results of that analysis. You can find a detailed discussion of these exceptions in the FutureTDM report on policies and barriers of TDM in Europe.[8]

One example is a legal exception that allows you to reproduce small pieces of content for the purposes of quotation. However if you plan to rely on such an exception to publish quotes or excerpts along with your TDM results, you should check that the quotation exception that applies to you is not overridable by

Limitations to quotation

"Publications or analyses resulting from TDM of subscribed content may include quotations from the original text of up to 200 characters, or 20 words, or 1 complete sentence."

– Springer's TDM Policy

[7] *"To the extent that a term of a contract purports to prevent or restrict the making of a copy which, by virtue of this section, would not infringe copyright, that term is unenforceable."* Copyright, Designs and Patents Act 1988, section 29A

[8] FutureTDM Deliverable D3.3 (PDF)

contract. If the exception can be overridden by contract, you will need to check the terms of your licence with the rights holder – they may for example only allow quotes of a limited number of words or characters, as in publisher Springer's text- and data-mining policy.[9]

3.3 What's in a licence?

You may already have a licence or agreement with the rights holder whose content you would like to use for TDM. If so, it is important to understand what this licence permits. While these guidelines cannot cover every possible licence clause you may encounter, the below examples should give an idea of the sorts of restrictions or freedoms a licence may specify. If you are unsure whether your licence allows your intended TDM activities, we recommend contacting the rights holder directly or getting expert legal advice.

3.3.1 The licence imposes explicit restrictions on TDM

In some cases, rights holders will explicitly address TDM in their licences, and the limitations under which TDM is permissible. Unless a legal exception applies and cannot be overridden by contract, you must abide by the rights holder's conditions when carrying out TDM on their content.

As discussed above, even if an exception applies and cannot be overridden by contract, the rights holder may still have some limited ability to define the ways in which you may use their content for TDM. This may include applying reasonable technical protection measures to protect their content and system, or limiting the types or number of authorised users who have access to the content. TDM often requires the use of large corpora of content from multiple sources, and some rights holders may impose restrictions on how much content you can access within a given timeframe to avoid overloading their servers.

Licence restrictions on TDM

"The user may not ... use any robots, spiders or other automated downloading programs, algorithms or devices to search, screen-scrape, extract, or index any Elsevier web site or web application"

– Elsevier subscription agreement

Example: Elsevier

Elsevier's Text and Data Mining Service Agreement permits licensees to "continuously and automatically extract semantic entities from full-text articles retrieved through the TDM service",[10] referring to an API service[11] Elsevier provides for the purpose of TDM. This API is provided via a separate technical infrastructure in order to cater for the potentially higher volume needs of TDM activities, as well as to ensure Elsevier can identify and prevent nefarious or unlawful access to their content.

The provision of an API for TDM activities may be seen as a

[9] Springer's text- and data-mining policy (retrieved 30 August 2017)
[10] Elsevier Text and Data Mining Service Agreement (retrieved 30 August 2017)
[11] http://www.developers.elsevier.com

reasonable technical protection measure to protect the stability of Elsevier's other servers,[12] in which case you would need to comply with the terms of Elsevier's licence agreement, and use their API whenever you access their content for TDM.

Elsevier's licence agreement also imposes limits on how you may reproduce excerpts from content you have used for TDM, which you may want to quote or publish alongside the results of your analysis. These are limited to "a maximum length of 200 characters surrounding and including the text entity matched", and must be accompanied by a DOI link back to the full text. Any reproduced excerpts of Elsevier's content must also be assigned a CC-BY-NC licence.

Elsevier's licence agreement also imposes limits on how you may reproduce excerpts from content you have used for TDM, which you may want to quote or publish alongside the results of your analysis. These are limited to "a maximum length of 200 characters surrounding and including the text entity matched", and must be accompanied by a DOI link back to the full text. Any reproduced excerpts of Elsevier's content must also be assigned a CC-BY-NC licence.

Unless you can make use of a legal exception that cannot be overridden by contract, you must also comply with these restrictions when using Elsevier's content for TDM.

3.3.2 The licence explicitly supports TDM

In some cases rights holders take an explicitly permissive stance towards TDM, supporting TDM activities in principle as well as in practice.

Example: The Royal Society

The Royal Society's policy on data mining states explicitly, "We support the stance that the right to read is the right to mine."[13] This means that anyone who has lawful access to read The Royal Society's content may also carry out TDM activities on this content.

Pro-TDM policies

"Members of subscribing institutions have our permission to mine journal content for either commercial or non-commercial purposes."

– The Royal Society open data policy

The other terms of their policy also demonstrate a permissive stance towards TDM practices:

- The policy asks TDM practitioners to cite papers used in analyses "where possible", which implicitly recognises that for TDM activities involving many thousands of papers, it may not be practical to cite every paper included in a corpus for analysis.
- Although The Royal Society has an automatic lock-out for downloads beyond a certain limit, they explicitly offer to help users wishing to carry out TDM by working together to manage their system load.

[12] Interpretation and enforcement of laws around technical protection measures varies among EU member states. In the case of the UK copyright exception for TDM, any API limitations should be assessed to understand if they unreasonably prevent researchers from making use of the exception – for example by limiting the API to 5000 results per query.

[13] The Royal Society Open data policy (retrieved 20 April 2017)

3.3.3 The licence is ambiguous or unclear

In some cases, the wording of a licence may make it difficult to determine what kinds of TDM activities are permitted. In Elsevier's Frequently Asked Questions page about text and data mining,[14] for example, one response states, "You are free to commercialize your own findings," while another states, "we provide TDM access for non-commercial purposes." If you are unsure about the specific details of your licence, you should contact the licensor to clarify what they do and do not permit.

3.3.4 The licence does not address TDM, or there is no licence available

Although data mining began to emerge as a technology in the late 1980s,[15] many rights holders have been slow to adapt to this technology and do not yet explicitly address TDM in their licences or other policies. There are also some kinds of content for which licences are not routinely provided at all, for example the content of web pages. In cases where no legal exceptions apply, this makes it difficult to assess whether your intended use of content for TDM is permitted.

If you are unsure what licensing terms apply to a given set of content, either because the licence is unclear or because you cannot find one, the safest option is always to contact the rights holder to ask permission to access and use their content. Section 3.4 below discusses some considerations you may want to take into account to negotiate reasonable and proportionate licences with rights holders.

In many cases, it may not be practical or possible to contact or even identify the rights holder. Consider a TDM project that involves mining content from hundreds of thousands of public web pages: it would not be practical to try to find licence or contact details from the rights holders of every single website accessed.

In these cases, the safest option is to assume that you are not permitted to use their content for TDM, unless you can make use of legal exceptions.

In practice, however, the risks associated with interfering with intellectual property rights depend on whether the rights holder objects to your use of their content and takes action against you. With this in mind, some people choose to go ahead and risk using content for TDM without explicit permission from the rights holder, under the assumption that the rights holder would not object to the use of their content.

In the case of web mining, for example, some consider a website's Robots Exclusion Protocol[16] to be an informal declaration of what content the website owner permits automated processes to access. A Robots Exclusion Protocol is not a formal licence, but many web crawlers (including those operated by Google) rely on the Robots Exclusion Protocol to determine which content they access and process.

---

[14] https://www.elsevier.com/about/our-business/policies/text-and-data-mining/text-and-data-mining-faq

[15] *"In the late 80s Data Mining term began to be known and used within the research community by statisticians, data analysts, and the management information systems (MIS) communities."* The History of Data Mining (accessed 20 April 2017)

[16] The Robots Exclusion Protocol, also known as a *robots.txt file*, is a file hosted on a web domain that lists which sections of that domain may not be accessed by "robots" or automated web-crawling processes.

You should be aware that although this may be a low-risk practice, it is not a zero-risk practice, and relevant regulations vary among different EU member states. We urge you to seek expert advice before considering carrying out TDM activities without a legal exception or clear permission from the rights holder.

3.4 Negotiating licences

For an individual researcher or small business, negotiating an appropriate licence with each and every rights holder whose content you may wish to use for TDM can be a prohibitive drain on resources. Unfortunately given the lack of consistent, unambiguous legal exceptions across EU member states, you may find this is the only way to ensure your TDM activities are legal.

Those who negotiate licences on behalf of institutions – for example libraries, universities, or consortia thereof – can have a much greater impact on reducing barriers to TDM, by considering licences in terms of access and permissions for TDM as well as for individual use.

If you represent a library, university, or other institution involved in negotiating content licences on behalf of others, we strongly urge you to read and consider whether licences your institutions agree to are a *reasonable* and *appropriate* balance of the needs of the researchers you represent, and the needs of the rights holders, in the context of TDM.

Below we discuss some of the aspects of TDM that rights holders may want to address in licences, and the impact these may have on licensees' TDM activities. If in doubt – talk to your researchers! They will be able to help you understand their needs regarding access to and use of content for their TDM projects.

3.4.1 Purpose of TDM activities

Rights holders may want to apply different permissions to their content depending on whether it is used for "commercial" or "non-commercial" purposes. They may wish to restrict activities that involve accessing/copying their content for analysis, reproducing excerpts of their content after analysis, or both. You should consider whether this distinction is reasonable or practical.

The activities of for-profit industry and businesses will of course generally be considered commercial. Conversely for some academic researchers, it may be clear that their use of TDM is purely non-commercial. But in cases where researchers are partly funded by industry, or collaborate with commercial partners, or are developing new technologies that may become the foundation of spin-out companies, it is less clear where the distinction between commercial and non-commercial research lies. Especially in research, where the potential applications of new knowledge or technologies may not be known at the beginning of a TDM project, restricting TDM to "non-commercial" purposes may have unforeseen impacts.

3.4.2 Usage and activity monitoring

Rights holders may wish to monitor, to some degree, bulk access to their content for TDM purposes. Reasons for this may include ensuring that they can identify fraudulent or malicious access to content, as well as understanding the needs and behaviour of licensees to develop and provide new products and services.

You should consider whether the nature and extent of monitoring of access to content is reasonable and appropriate. Particularly in a research context, overly detailed monitoring of researchers' behaviour may raise ethical questions about academic freedoms.

3.4.3 Reproducing content

Rights holders have a need to protect their intellectual property from redistribution that would impact the value of the original works, and may therefore restrict how much of their content may be re-published following TDM analysis, in the form of quotations or other excerpts. They may also require TDM practitioners to attribute appropriate credit to the rights holder, in cases where excerpts of original content are reproduced.

You should consider whether restrictions on reproducing excerpts of original content are reasonable and appropriate. It may be useful to consider whether the intended reproductions are likely to impact the value of the original works.

You should also consider whether requirements to attribute credit are reasonable and practical, particularly given that TDM projects may involve corpora of hundreds of thousands of documents.

3.4.4 Technical limitations

Rights holders may wish to apply technical protection measures to protect their content. These may be to ensure that only authorised users access their content, or to prevent systems and servers from being overloaded by large-scale access to their content.

You should consider the impact of technical protection measures on TDM users, particularly in the context of the large-scale access to content that TDM typically requires.

3.5 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 further information and guidelines, check the FutureTDM website at http://www.futuretdm.eu/.

3.6 Appendix: Analysis of Selected TDM Licences

To supplement the guidelines above, we have considered several TDM licences and how their terms might affect your ability to carry out TDM activities. These licences were identified through the

CrossRef Metadata API as the most frequently cited TDM licences, after excluding Open Access licences.

As discussed above, if you cannot rely on a legal exception for your TDM activity, or if the legal exception can be overridden by contracts, you will need to make sure you comply with the terms of any relevant licences. Please note however that this overview is not intended to be legal advice. Interpreting licences can be a complex process, and you should consult an expert if you are unsure about the lawfulness of your planned TDM activity.

This analysis has been carried out in good faith and represents our honest interpretation of the licences considered. If you believe any licence terms have been misunderstood or misrepresented please let us know.[17]

3.7 Considerations for TDM Licences

In analysing licences for TDM, we considered the following questions which may be important for TDM practitioners:

- **Where is the licence?** (Can it be easily accessed and read online?)
- **Do I need to check other licences or documents?** (Does this licence/agreement supersede any others, or do I also need to check the terms of any other agreements my institution may have made with this publisher?)
- **Does this licence affect my use of OA content?** (Are there terms in this licence that apply even to TDM carried out on open access content from this publisher?)
- **Is TDM permitted?** (Is any TDM permitted by existing licences/agreements, or do I need a separate agreement?)
- **Can I carry out TDM for any purpose?** (Am I limited for example to using TDM only for non-commercial purposes?)
- **Do I need to tell anyone what I am doing with TDM?** (Do I need to tell the publisher about the kinds of TDM activities I intend to carry out?)
- **Are my TDM activities monitored?** (Will the publisher be able to see which content I access or use?)
- **Can I access any content I like?** (Am I able to use all content I have lawful access to for TDM?)
- **Can I access content any way I like?** (Am I limited to accessing content in a particular way specified by the publisher?)
- **Are there limits on how much content I can access, or how quickly?**
- **Do I need to ask or inform anyone before carrying out TDM?**
- **Are there limitations on the types of TDM analysis I can perform?** (Does the publisher allow all kinds of computational analysis, or just specific activities?)

[17] office@**futuretdm**.eu

- **Are there restrictions on how I can store and share datasets I'm using for TDM?** (Does the publisher restrict how I can share content used for TDM with, for example, colleagues within my institution?)
- **Are there restrictions on how I can share new knowledge I generate as a result of TDM?** (That is, are there any restrictions on what I can do with novel insights or information I generate as a result of my TDM analysis?)
- **Am I required to share the outputs of my TDM research?** (Does the publisher require me to show them the results of my TDM analysis?)
- **Can I support my results with excerpts from the content I have mined?** (Does the publisher restrict how much content I can quote or reproduce alongside the results of my analysis?)
- **Can I retain datasets for verifiability and reproducibility of my results?**
- **Do I have any other responsibilities or obligations?**

The answers to these questions are laid out in the table below, based on our interpretation of the relevant licences, agreements, or other guidance provided by each publisher. The table is colour-coded as follows:

Ideal for TDM activities	Close to ideal for TDM activities	Some negative implications for TDM activities	Very restrictive for TDM activities

	Elsevier	Wiley	Springer	APS	Emerald	IOP	IUCr
Where is the licence?	Online	Online	Online	Online	Online, with FAQs	Online	Online
Do I need to check other licences or documents?	**NO** The TDM Agreement supersedes any and all prior and contemporaneous agreements	**UNCLEAR** The click-through TDM agreements says that it supersedes all other prior and contemporaneous agreements, but also that it is superseded by any separate TDM agreement	**YES** The TDM clause may not have been included in existing SpringerLink subscription agreements, but can be added by existing subscribers	**YES** Unless explicitly licensed under a different licence, TDM requires a separate written bilateral agreement	**NO** Not mentioned in policy	**YES** If your institution has a separate licence, the terms of that licence prevail in case of any conflict with the TDM policy	**NO** Not mentioned in policy; users without subscriptions to IUCr journals can request TDM access
Does this licence affect my use of OA content?	**NO** individual OA licences supersede anything the contrary in the TDM Agreement	**NO** If more permissive licences apply, you may use content in accordance with article-level permissions	**NO** TDM of OA Springer content is usually allowed without restrictions	**NO** Does not apply to content with CC-BY licences	**NO** Not mentioned in policy	**YES** You must ask IOP to remove technical protection measures, even of OA content, if "large amounts of data" are to be extracted	**NO** OA content may be mined without restriction provided proper attribution is given
Is TDM permitted?	**YES**	**YES**	**YES**	**NO** Not without a specific licence agreement	**YES**	**YES**	**YES**
Can I carry out TDM for any purpose?	**NO** You may not extract, develop or use the dataset for any direct or indirect commercial activity	**NO** You may only text and data mine Wiley content for non-commercial scholarly research related to specific projects; direct or indirect commercial purposes require prior written consent from Wiley	**NO** You may only access content for TDM for the purpose of non-commercial research	**?** Depends on agreement	**NO** TDM rights are granted purely for internal non-commercial research purposes	**NO** You may only access content for the purpose of non-commercial TDM; specific terms must be negotiated with IOP for TDM for commercial purposes	**NO** Content may only be mined for non-commercial purposes
Do I need to tell anyone what I am doing with TDM?	**MAYBE** You must provide TDM output and any related content to Elsevier on request	**NO** Not mentioned in policy	**NO** Not mentioned in policy	**?** Depends on agreement	**NO** Not mentioned in policy	**YES** You are required to tell IOP the purpose for which you want to carry out TDM	**NO** Not mentioned in policy

Are my TDM activities monitored?	**YES** You are required to use an API key; Elsevier maintains information about you which may be used in aggregate, and may be used to promote Elsevier offers to you	**YES** You are required to use an API key	**NO** No authentication is required when retrieving SpringerLink content for TDM	**?** Depends on agreement	**NO** No authentication is required for CrossRef's TDM API	**PARTLY** You must provide IOP with your name, institution, and the titles, years and issues of journals you wish to mine	**NO** No authentication or monitoring mentioned in policy
Can I access any content I like?	**NO** You are licensed solely to access content made available via the CrossRef API	**NO** You may only access content made available via APIs	**YES** You may use all subscribed content	**NO** Full-text PDFs will only be made available for TDM one year after the date of publication	**YES** Emerald suggests using CrossRef's TDM service to identify and access content	**YES** No restrictions mentioned in policy	**YES** No restrictions mentioned in policy
Can I access content any way I like?	**NO** You are licensed to use a set of proprietary APIs to access data; you may not use any automated programs to search or scrape any Elsevier web site or application	**NO** You must access content using a Wiley-approved API, and may not bypass the API; you may not use any automated programs to search or scrape Wiley content	**YES** You are encouraged but not required to download content directly from the SpringerLink platform; friendly DOI-based URLs are provided, tools and methods are suggested, and no API key is required	**?** Depends on agreement	**YES** You are encouraged to use CrossRef's TDM services, but not forbidden from accessing content in other ways	**YES** No access methods specified in policy	**YES** You are encouraged to access content from Crystallography Journals Online, but not forbidden from accessing content in other ways
Are there limits on how much content I can access, or how quickly?	**UNCLEAR** No details are provided about rate limiting through Elsevier's API	**SOMETIMES** You must abide by any rate-limiting which may be conveyed from time to time	**VOLUNTARY** You are asked to be considerate and limit your download speed to a reasonable rate	**?** Depends on agreement	**SOME** There are no hard limits on the number of items that may be downloaded, but you may be blocked if your downloading constitutes unfair usage	**YES** The IOPscience platform blocks systematic downloading of content unless you ask for these technical limits to be removed	**VOLUNTARY** You are asked to limit your downloading speed to a reasonable rate
Do I need to ask or inform anyone before carrying out TDM?	**NO** Not mentioned in policy	**NO** Not mentioned in policy	**NO** Not mentioned in policy	**?** Depends on agreement	**ADVISED** You are advised to inform Emerald you wish to mine their site to avoid being blocked due to unfair usage	**YES** You are required to contact IOP to arrange for technical limits to be removed temporarily and allow server loads to be managed, if "large amounts of data" are to be extracted	**NO** Not mentioned in policy

Are there limitations on the types of TDM analysis I can perform?	**YES** You are licensed to extract semantic entities for the purpose of recognition and classification of relations and classifications between them	**NO** You are licensed to carry out computational analysis including but not limited to identification of entities, structures and relationships	**NO** None mentioned in policy	**?** Depends on agreement	**SOME** Licence includes specific definitions of TDM activities and outputs; these are broad but you must not perform systematic or substantive extracting of content	**NO** None mentioned in policy	**NO** None mentioned in policy
Are there restrictions on how I can store and share datasets I'm using for TDM?	**NO** None mentioned in policy	**YES** You may load and technically format content on your servers for use for specific TDM projects; you may not otherwise create any form of central repository for Wiley content, or any product or service that could potentially substitute any existing Wiley services	**NO** None mentioned in policy	**?** Depends on agreement	**SOME** You may load and technically format XML content on your server, PC or laptop to enable access and use of content for allowed TDM purposes; you may not make results of TDM outputs available on any externally facing server or website	**NO** None mentioned in policy	**NO** None mentioned in policy
Are there restrictions on how I can share new knowledge I generate as a result of TDM?	**YES** Results may be used by you and your company or institution, but may not be used in a way that would compete with existing Elsevier products; a specific proprietary notice must be used when sharing results externally	**YES** You may communicate TDM outputs as part of original non-commercial research, including in articles about that research	**NO** None mentioned in policy	**?** Depends on agreement	**YES** There are no restrictions on where and how you can publish your research results, but you may not make results of TDM outputs available on any externally facing server or website	**YES** Anything generated directly by TDM must be licensed under either a CC-BY or CC-BY-NC-ND licence	**NO** None mentioned in policy
Am I required to share the outputs of my TDM research?	**YES** You must provide TDM output and any related content to Elsevier on request to ensure compliance with their agreement	**NO** Not mentioned in policy	**NO** Not mentioned in policy	**?** Depends on agreement	**NO** Not mentioned in policy	**NO** Not mentioned in policy	**NO** Not mentioned in policy

Can I support my results with excerpts from the content I have mined?	**YES** Limited to query-dependent text of a maximum length of 200 characters surrounding the semantic entity matched, or bibliographic metadata; must include a DOI link to the original material	**YES** Limited to brief quotations as permitted under national copyright laws; must include a DOI link to the original material	**YES** Limited to quotations of up to 200 characters, 20 words, or one complete sentence; must include a DOI link to the original material	**?** Depends on agreement	**YES** You can use snippets up to a maximum of 200 characters, provided these are referenced as you would reference a copyright work; you must contact Emerald if larger extracts are exceptionally required	**YES** Outputs can include snippets of up to 200 characters; mined text or data should include a DOI link to the original material wherever reasonably practical	**YES** Brief extracts from articles may be included without revision or modification in publications, with a full bibliographic reference to the original source
Can I retain datasets for verifiability and reproducibility of my results?	**NO** You may not substantially retain the dataset; all Elsevier content stored for TDM must be permanently deleted on termination of the agreement	**NO** You must delete all Wiley content downloaded for TDM on completion of any specific TDM project, or on termination of the agreement with Wiley	**UNCLEAR** Not mentioned in policy	**?** Depends on agreement	**NO** You may not substantially retain content; all copies of Emerald content that have been locally loaded for TDM must be destroyed on termination or expiry of this licence	**UNCLEAR** Not mentioned in policy	**UNCLEAR** Not mentioned in policy
Do I have any other responsibilities or obligations?	**YES** You are responsible for complying with data protection and relevant privacy laws when using or processing personal data	**YES** You must implement and maintain data security measures to protect Wiley content in line with international industry standards; you are responsible for complying with data protection and relevant privacy laws when using or processing personal data	**NO** None mentioned in policy	**?** Depends on agreement	**NO** None mentioned in policy	**NO** None mentioned in policy	**NO** None mentioned in policy

4. DATA MANAGEMENT GUIDELINES FOR RESEARCHERS

4.1 Introduction: Why care about data management for TDM?

These guidelines are intended to give an introduction to the principles of data management. They are aimed primarily at academic researchers who collect, create, store and share data, to give you an idea of how you can make sure your data is genuinely reusable, particularly for text and data mining (TDM) projects. However, the general principles of best practices in data management apply to all cases of storing and sharing content.

Accessing and using content for TDM often involves quite different processes to those used by an individual reader or researcher. New TDM technologies are being developed every day, and managing your data with TDM in mind means you will be better able to use these technologies to discover new knowledge from your data in the future.

Don't let your valuable data lie underused in poorly accessible formats – start thinking about and planning data management for TDM!

4.2 Background

4.2.1 What is data management?

According to DAMA, The Global Data Management Community,[18] "Data management is the development, execution and supervision of plans, policies, programs and practices that control, protect, deliver and enhance the value of data and information assets."[19]

In the specific case of research data, good data management fundamentally aims to make sure that research data are managed according to legal, statutory, ethical and funding body requirements. This means that good data management is relevant to all stages of the data lifecycle, from the procedures of data planning (specifying the types of data to be used) through:

- data generation, collection and organisation;
- documentation and metadata usage;
- curation, maintenance and preservation; and
- ultimately, policies for publishing, sharing and providing access to data.

Data management, particularly the curation and preservation of data, is valuable for two key reasons. Firstly, it allows third parties to validate experimental methods and results. And secondly, it allows for the re-use and re-purposing of data in other contexts, including other disciplines, with different research goals. Many would say that the data on which a research project has been based are as important as the scientific results themselves!

[18] https://www.dama.org/
[19] DAMA-DMBOK Guide (Data Management Body of Knowledge) Introduction & Project Status *(Note: PDF no longer available online at https://www.dama.org; definition taken from Wikipedia.)*

As just one example, better sharing of medical data, such as clinical trial data, could boost scientific progress by exposing these data to secondary analyses. Additional findings could well lead to new knowledge and improvement in public health outcomes, which would not be possible if data were not shared in re-usable formats.

4.2.2 What data management means for TDM

The importance of data for TDM

TDM cannot happen without access to large amounts of data. This data could be of any type – from scientific data, to data related to aspects of everyday life, in domains from meteorological, to biological, to economic and geographical data. With this in mind, it is more than obvious that data management is of crucial importance for TDM.

Although huge amounts of data are produced globally on a daily basis, only a small part of that data is widely known, let alone published and accessible in realistic and practical terms. Data creation is a time-consuming and expensive process, involving not just simple data collection, but additional steps of data curation, metadata addition and annotation, maintenance and preservation, and – last but not least – legal clearance of data.

In many scientific fields, we have already seen that data and related services create added value when they are opened and shared for secondary purposes, from fundamental research to the development of innovative technologies and applications.

Therefore, there is a need for appropriate tools and mechanisms (scientific, technical, legal, organisational – and even social) which will allow efficient access to, sharing of, re-use and re-purposing of data. This all starts with an appropriate Data Management Plan, which we will discuss in the following sections.

Before we proceed, it is crucial to make an important distinction between access to and re-use of data in the context of TDM, as opposed to access and re-use by human users, since accessing content for TDM purposes is a very different process to accessing it for individual reading.

It is perhaps trivial to note that in the second case, the user of the content is human, while in the former, the "user" is a computer tool, service or application that performs a specific task of processing on structured or unstructured (textual) data. But the needs of humans in one case, and computers or algorithms in the other, can be very different in terms of data management. This distinction is not always fully appreciated.

Take the case of text mining, which involves transforming text into structured data that can be used for further analysis, usually with the help of:

- natural language processing (NLP) such as part-of-speech tagging, syntactic parsing, semantic analysis, named entity recognition, automatic summarization, machine translation, etc.;
- statistical processing of data (language or numerical), for example to identify tendencies and trends;
- advanced pattern recognition algorithms which sift through large amounts of data to assist in discovering previously unknown information about, e.g. customer behaviour;
- data clustering techniques, to find similarities between objects in the data;

- machine learning algorithms for knowledge discovery;
- and more.

For this kind of text mining to be possible, data (in this case *text*) needs to be provided in the right formats and with the right metadata for machine processes to "understand".

Human vs. Machine access and use of data

It is clear from the above that access to and use of content and data in the framework of TDM requires an entirely different approach to data, in terms of the tools used for accessing and processing data, but also in terms of data management, which needs to be reflected in Data Management Plans.

To further clarify the distinctions between the two processes (human and TDM access to data), let us consider some examples:

The issue of file format

The Europeana collections[20] make available thousands of books (among other cultural items). These are provided digitally, documented with metadata, and provide information on rights of re-use. For example, Flaubert's *Madame Bovary* is available in French via the Europeana website:[21]

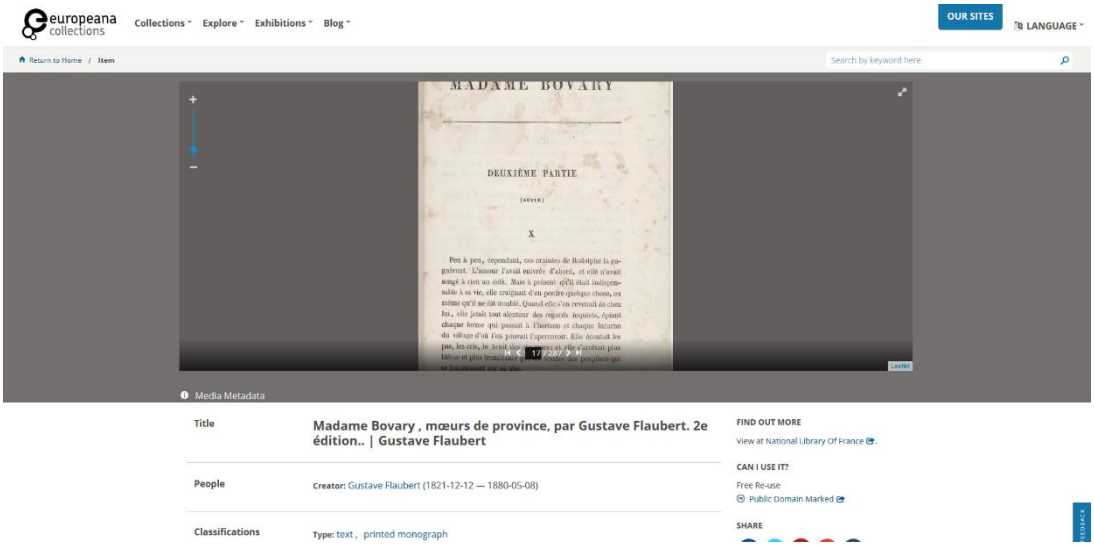

Figure 2: *Madame Bovary* as provided by the Europeana website

The human user has no problem reading this version, which is free for re-use, but it is completely inappropriate for TDM processes. This has to do with its format, which the computer does not recognise as text (which can be analysed by TDM processes), but a jpg image file which is unprocessable by most TDM tools. Tools do exist for converting images to text, such as OCR, but these are imperfect and add an extra source of complexity and higher risk of errors to any TDM analysis.

Usability of web pages

[20] http://www.europeana.eu/portal/en
[21] Madame Bovary (accessed 17 April 2017)

The online version of a newspaper poses no problems to human users, and, since it is open digital text, one might expect it to be a perfect source for TDM. Indeed, newspapers are a valuable source for TDM, but they are far from perfect. The front page of **The New York Times**,[22] for example, needs extensive cleansing before it is usable for TDM purposes: header, footer, banners, login buttons and similar items would have to be removed before passing the pages on to TDM processes. This can certainly be done and dedicated software that removes the so-called boilerplate material is extensively used. However, the original version is not directly useful to TDM processes, and this again requires extra time and work on the part of the TDM practitioner.

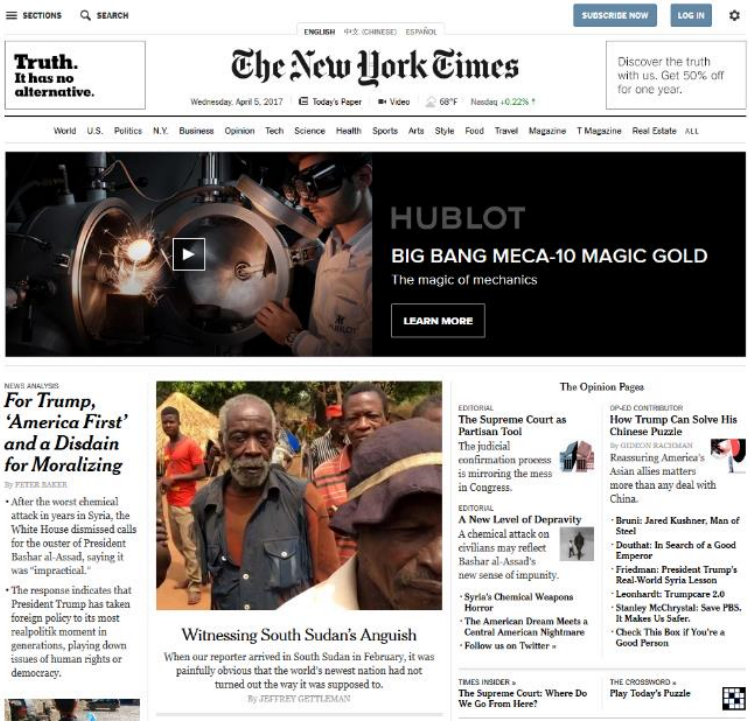

Figure 3: The *New York Times* website

In both these cases it is clear that **"*digital" does not mean "TDM ready".***

The importance of metadata

Most data sharing and distribution platforms think primarily about the needs of the human user, not TDM. For human users, the metadata schemas adopted for data description can afford to be minimal, as the human user is cognitively capable of filling in missing information or deducing it from the available metadata and data. This is not the case, however, with TDM: poor metadata reduce the visibility of the data, and inadequate metadata descriptions render the data difficult to identify, index, classify, retrieve and process.

For example, if the metadata description of a medical dataset does not explicitly say that it belongs to the medical domain, a human user will generally be able to identify that it is medical data. In the case of TDM, however, if the data are not classified as medical, even if their provider deposits them in a medical thematic repository, tools at various levels will fail to recognise them. The missing

[22] https://www.nytimes.com

metadata will mean that harvesting tools looking for medical texts will not recognise them, and tools specifically developed for the processing of medical texts will also fail.

If the description of a dataset does not include licence information in its metadata, the situation is worse: neither humans nor TDM processes will be able to work out what rights they have to reuse the dataset. The human user could contact the owner (if available) and find out, but a TDM tool or service which looks for publicly available data, again, would fail. Absence of metadata about the author of a publication would result in the publication not being linked to the author, and therefore not discoverable to people or tools looking for publications by this author.

Machine readable metadata are used for:

- data and services description, curation, and updating;
- data identification and retrieval;
- browsing catalogues and inventories (as search filters);
- uploading and downloading of data;
- interoperability checking;
- efficient licensing schemas;
- persistent identification;
- persistent connection of data with their creator and other bibliographical information; and
- harvesting by aggregators and other infrastructures.

To sum up, the tools and mechanisms mentioned above (which allow efficient access to, sharing, re-use and re-purposing of data), and more importantly data management in general, need to take into consideration the differing needs of different users and uses of data. This will be discussed in more detail below.

4.2.3 What is a Data Management Plan?

As the EU *Guidelines on FAIR Data Management in Horizon 2020 (Version 3.0)*[23] tell us, a Data Management Plan (DMP) "…describes the data management life cycle for any data to be collected, processed and/or generated by a Horizon 2020 project. As part of making research data Findable, Accessible, Interoperable and Re-usable (FAIR), a DMP should include information on:

- the handling of research data during and after the end of the project;
- what data will be collected, processed and/or generated;
- which methodology and standards will be applied;
- whether data will be shared/made open access; and
- how data will be curated and preserved (including after the end of the project)."

The EU Guidelines particularly stress the importance of:

- open access (while respecting existing copyright restrictions);
- data discoverability, through metadata and persistent identifiers;
- interoperability allowing data exchange and re-use, by adherence to common standards and best practices for data description;

[23] Guidelines on FAIR Data Management in Horizon 2020 (accessed 17 April 2017)

- usage of standard vocabularies; and
- use of certified deposition mechanisms and infrastructural facilities that cater for data curation, maintenance, security, storage and long term preservation, as well as for user management (authentication and authorisation).

Data Management Plans (DMPs) are produced by organisations, projects and companies dealing with data of any type, as well as by their funders. Researchers need to follow DMPs when preparing their research data, both to organise their data and to deposit their data to a repository or infrastructure. DMPs define the purpose of data collection and generation, the types and formats of the data to be collected, their size and the target users, the mode of distribution (if planned), and the preservation model adopted.

The features of the DMP become requirements for data to be used or included in a project: any data to be included have to follow the plan's specific recommendations. This means that these recommendations also act as guidelines for the prospective data providers as regards their data.

The data lifecycle broadly includes the stages of:

- data creation (data generation or collection, cleaning, rendering to the appropriate format and documentation through metadata); and
- data storage and maintenance, curation, preservation and sharing.

The first stage may be the responsibility of many different data providers who offer their data for deposition, while the second stage is catered for by data hosting repositories and infrastructures. For a data lifecycle to function successfully, data providers and data hosts must work together to manage data.

The keys to successful collaboration between data providers and hosting repositories and infrastructures are:

- clear specifications for data preparation (collection, cleaning), and data types needed;
- clear metadata for data descriptions, which allow TDM tools to interpret the interoperability and processability of the data (i.e. whether a dataset can be processed by a specific tool), and which data may be harvested – these metadata essentially constitute the "credentials" for the data to enter the TDM world;
- clear specifications for permitted reuse of the data, and well-defined licensing schemas; and
- clear-cut and bilaterally acknowledged responsibilities for both parties.

If all stakeholders keep these goals in mind, this will help to ensure that any collected or created data are (re-)usable by repositories, infrastructure, and TDM processes.

4.3 What are the benefits of data management?

4.3.1 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.
- **Maintenance and preservation**: When the data are maintained and preserved by the procedures put in place by the storage facility, individual researchers are relieved of this burden.
- **Deployment of powerful computational facilities**: The computational facilities of the organisation, in terms of storage capacity and processing power, greatly exceed those of any individual researcher.
- **Processability and interoperability**: By adhering to standards and by using metadata descriptions, the data become interoperable with TDM tools and technologies, and processable for further investigations.
- **Lawful sharing**: By adhering to the repository's deposition guidelines, the researcher (in collaboration with the repository) ensure that access, sharing and distribution of the data are respecting all relevant legislation and legal procedures.
- **Recognition**: The data and the provider are permanently connected through the repository; in other words, the researcher's ownership of the data is manifest and unquestionable.
- **Citation and publicity**: The data and their provider appear in the organisation's catalogues. This brings them publicity, and the common practice of harvesting among infrastructures significantly increases this publicity.
- **Added value**: Re-use and pre-purposing of the data adds value to it, through the discovery of new modes of use and research perspectives.
- **New collaborations**: By sharing their data, the researcher increases their chances of discovering new collaborations, possibly even across disciplines, which can lead to new discoveries, shed light on different aspects of the original data, and produce new research results or technological applications.

4.3.2 For data users

Almost anyone can be a user of data, from researchers, to private companies, to the general public and citizen scientists. When data are stored in accordance with a good Data Management Plan, all potential users can benefit from:

- **Access to large amounts of data, tools and technologies**: Sharing data provides users with access to much more data than they could ever create or collect on their own.
- **Ease of identification and access**: Data stored in official catalogues (rather than personal computers) and accessible through a simple user interface are easier to find. When they are accompanied by metadata descriptions and relevant documentation, users can easily identify and assess how appropriate the data is for their needs.

- **Persistence**: When data are permanently stored by a repository committed to their maintenance and preservation, the is less risk of users finding and identifying a dataset which later disappears.
- **Licences or explicit terms of use**: When data come with a licence or with terms of use, explicitly defining the actions a user can legally perform with the data, users face less uncertainty about whether they have the right for personal use only, re-distribution of the data, production of derivative datasets, etc.
- **New collaborations**: The opportunities for creating new collaborations is bi-lateral; users may identify interesting datasets and/or tools and technologies which could lead to new collaborations with the data owner.

4.4 Data management guidelines for researchers

If you are a researcher who has generated or collected data, how can you make sure your data is genuinely useful and re-usable when you deposit it in a repository? This section provides a set of guidelines to help you make your data as valuable as possible for future re-use.

4.4.1 Identifying a repository in an appropriate domain

In *The Guidelines on the Implementation of Open Access to Scientific Publications and Research Data in Projects supported by the European Research Council under Horizon 2020*[24], the European Research Council (ERC) strongly encourages ERC-funded researchers to use discipline-specific subject repositories for their publications, and provides a list of recommended repositories.

Subject repositories (also called thematic or disciplinary repositories) host depositions of publications and/or research data in a specific domain, regardless of the author's institutional affiliation. A well-known example is Europe PubMed Central,[25] a repository of content from the life sciences domain.

You should try to identify the most appropriate subject repository in your domain, where you can deposit your data. Subject repositories provide requirements for what kinds of data they accept, which will be reflected in the repository's metadata schema; you can use these as guidelines for submitting your data.

If there is no appropriate discipline-specific repository, you can also make your data available in an institutional repository or in domain-independent centralised repositories such as Zenodo[26]).

4.4.2 Understanding metadata requirements

It is important to use the right metadata elements for your dataset. Some metadata elements are common to all data types; these are usually administrative elements, and give information on phases of the resource's life cycle (e.g. creation, validation, usage, distribution and licensing). Other metadata elements are only relevant to specific types of data, such as captures for audio, video and image resources, linguistic annotation for textual corpora, etc.

[24] Guidelines on the Implementation of Open Access to Scientific Publications and Research Data in projects supported by the European Research Council under Horizon 2020 (accessed 17 April 2017)
[25] https://europepmc.org/
[26] https://zenodo.org/

Some elements can therefore be inappropriate for the description of certain datasets – for example *minutes* is an appropriate unit when referring to the size of an audio dataset, but not when referring to the size of a textual dataset.

The guidelines below list the most important requirements for the creation and management of metadata for all types of data.

Specify the types of data that will be created

Research data can include statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, video recordings and images. Textual data can be data in their 'own right', or they can be discussions or annotations related to other types of data, such as descriptions of images or videos, film scripts, or transcripts of audio data.

When planning how to deposit your data, you will need to specify the type and medium of the data, for example: text, audio, video, image, or numerical. This will make sure the any potential (re-)users, either human or machine, will have the means to judge whether the data are appropriate for their needs.

Specify the provenance of data

Data offered for deposition should specify their origin and the how they were created: whether they were generated by the researcher, collected by the researcher but created by other(s), or whether they constitute results of processing on other primary data.

Specify the size of the data

The size of the data is crucial for understanding how representative TDM computations and results might be, as well as for training of TDM tools. Size is also important for storage purposes and to check whether a given TDM tool can work with the data – some TDM processes have limits to the size of data they can deal with. Size can be measured in any unit appropriate for the specific data type: words, tags, minutes, video frames, pages, etc.

Specify the data format

People who work with TDM often talk about "machine-readable" data. This means data that is in a format which can be used and understood by a computer. For this to be possible, the format of the data must follow accepted standards or broadly used best practices. Scanned versions of printed material, for example, do not fulfil this requirement, especially when no use is made of OCR technologies.

There are many benefits to standardised data formats; they help guarantee usability, re-usability, long-term storage, curation and preservation of the data. Because formats that render the data machine-readable and processable are easily interpretable by software tools and services, they can be smoothly mapped to other formats if needed or migrated to new formats when these are developed.

Non-proprietary software and formats based on open standards, using standard character encodings, are highly recommended; examples are:

- Text: plain text (txt), ASCII, HTML, XML,
- Character encoding: Unicode UTF-8

- Audio: aiff, wav
- Containers: tar, gzip, zip
- Databases or excel files: XML or CSV are preferable to native binary formats
- Open Document Format (ODF)

If your data does not adhere to common standards and best practices, this diminishes the possibility of processing it using TDM techniques, or any other type of software developed by other interested parties (researchers or industry).

Specify any specific tools needed to access the data

If the data have a specific access tool, user interface, or environment without which the data are inaccessible, this needs to be specified. Users need to know that if they download the data without the respective tool, the data will be non-usable. On the other hand, if users need to download a whole environment in order to have access to the data, they should be informed in advance so that they can calculate the storage capacity and computational power needed.

We should also point out that in the case of data accessible through APIs, access should be enabled in ways that allow bulk downloads. Restricting access to, for example, X papers per Y seconds may be acceptable for human users, but given the scale of some TDM projects, this may mean it would take a long time to download the necessary data – even years!

Keep track of changes by versioning

Some data and tools can be static, meaning that they do not evolve after their creation. This is true for digital versions of literary texts, for example, which once created as digital objects remain unchanged during their lifetime. However, there are cases where data and tools evolve over time. These changes need to be reflected in a proper versioning schema, which relates the newer version to the older.

Example cases include:

- Language data, medical data, weather data, social data – any type of data which are constantly produced, and where each new dataset supersedes or complements the previous one. Whether the older version becomes obsolete as it is replaced by the new one, or whether the new version includes the older version, should be reflected in the versioning schema.
- Datasets which get evaluated, corrected, or enriched. In the case of language data, for example, a dataset containing newspaper issues could be enriched with new material; another dataset might be corrected for spelling errors or syntactically annotated. All these changes in the initial version should be specified in the metadata, so that the users know what they have to deal with in each case.
- Similarly, in the case of software tools, versioning is indispensable in order to give the users adequate information about their functionalities.

Versioning will improve research efficiency and protect the authenticity of the data, especially in the case of shared data and tools. In particular, versioning is important for reproducibility and verifiability purposes.

Respect sensitive data

Sensitive data are data through which, for example, an individual, a process, a location can be identified, and where such identification may be unwanted (creating an ethical issue) or illegal (creating a legal issue). According to the law and research ethics, sensitive data cannot be shared on an *"as is"* basis. However, it is not illegal to publish the metadata alone, including a description of the data.[27] This aids discoverability of the dataset without risking disclosing sensitive or personal data. Necessary steps for the protection of sensitive data are:

- acquiring unambiguous consent about data sharing;
- applying an appropriate licence, with restrictions on access if necessary; and
- protecting people's identities by anonymising data where necessary.

Many tools and techniques exist for data anonymisation and de-identification[28], both open source or proprietary. However, data providers should maintain responsibility for anonymisation, given that the tools may not provide fully anonymised results. Note that pseudonymisation, in which individuals may be identified (again) with additional information by the TDM user, is not sufficient.

In cases of collaboration with industry, researchers may also generate commercial or confidential data which should also be treated as sensitive, and the sharing of which may be restricted by the terms of the collaboration.

Specify a licence

Data should be accompanied by information about the licence under which they are published, stating explicitly the terms of use permitted by the rights holder. Ideally, research data should permit the widest reuse possible, including derivative works (new datasets or tools based on the original). This might not be legally possible in all cases, due to existing restrictions on the data, but whatever the conditions of use are, they should be explicitly stated in the licence text.

Best practice on licensing is to use broadly standardised licensing models, such as Creative Commons (CC) licenses[29] (with the motto "When we share, everyone wins") for data, and FOSS licences (Free and Open Source Software), such as GPL (GNU General Public License versions),[30] AGPL (GNU Affero General Public License),[31] Apache Licence 2.0,[32] BSD,[33] GFDL (GNU Free Documentation License),[34] or LGPL (GNU General Public License)[35] for software.

Legal information should be an integral part of metadata. Using clear licence statements in the metadata, preferably well-known ones in machine-readable form, improves accessibility and re-

[27] Insofar as it is not possible to re-identify the data subjects from the metadata; see section 2.4 for guidelines on protecting personal data.
[28] For a list of such tools see the Australian National Data Service (ANDS) Guide to Sharing and Publishing Sensitive Data.
[29] https://creativecommons.org/
[30] http://opensource.org/licenses/gpl-license.php
[31] http://www.gnu.org/licenses/agpl-3.0.html
[32] http://www.apache.org/licenses/LICENSE-2.0.html
[33] https://opensource.org/licenses/BSD-2-Clause
[34] http://www.gnu.org/copyleft/fdl.html
[35] http://www.gnu.org/licenses/lgpl.html

usability of content by TDM processes by enabling tools to directly detect whether they may process the data.

Use a permanent identification mechanism

"A persistent identifier (PI or PID) is a long-lasting reference to a document, file, web page, or other object."[36] As its name clearly states, the PID serves the purpose of long term, unique identification and citation of digital objects on the Internet, so that users can unambiguously locate them. PIDs were introduced as an answer to the problem of broken URL links.[37] Broken links exist on the Internet for a variety of reasons:

- the data are no longer online for various reasons (to make space for more recent data, no-one maintains it, etc.);
- the data has changed location, so the old URL does not work;
- the URL domain itself has become inaccessible.

This means that URLs (network addresses pointing to resources) cannot guarantee persistent access to those resources; persistent identifiers aim to address this problem.

Best practice regarding the use of PIDs is to assign PIDs both to data and metadata records. These PIDs should be suitable for both human and machine interpretation.

Dedicated institutions exist to issue persistent identifiers. The most common are Digital Object Identifiers (DOIs),[38] the Handle System,[39] Persistent Uniform Resource Locators (PURLs),[40] Uniform Resource Names (URNs),[41] and Extensible Resource Identifiers (XRIs).[42] DataCite[43] is a non-profit organisation that provides persistent identifiers (DOIs) for research data, with the goal of helping the research community locate, identify, and cite research data.

The use of PIDs has obvious benefits for worldwide identification, use, and citation of datasets and tools.

4.4.3 Validation and quality assurance of data and metadata

Data quality must be defined in terms of a particular user and use case; a dataset might be perfect for one user's use case, but not so good for another. For example, a dataset from a medical database might be appropriate for a medical researcher who works on diabetes, but quite useless to a political scientist searching for patterns in protest movements.

Content-wise, data quality needs to be defined as 'operational usability'. Data quality metrics are therefore domain-specific, based on data type, research domain and intended use.

[36] https://en.wikipedia.org/wiki/Persistent_identifier
[37] A 2015 assessment of 180,000 web links cited in research articles found that 24.5% of them were unavailable.
[38] https://en.wikipedia.org/wiki/Digital_Object_Identifier
[39] https://en.wikipedia.org/wiki/Handle_System
[40] https://en.wikipedia.org/wiki/Persistent_uniform_resource_locator
[41] https://en.wikipedia.org/wiki/Uniform_Resource_Name
[42] https://en.wikipedia.org/wiki/Extensible_Resource_Identifier
[43] https://www.datacite.org/

Data quality extends to and is affected by metadata quality: the data should bear valid metadata, as detailed as possible, including production date, ownership and contact information.

Metadata should also be accompanied by a licence (preferably an open licence, such as CC-BY) to maximise their usability for TDM, and should also be harvestable, in order for them to be included in the inventories of other infrastructures, aiding data visibility and publicity.

4.4.4 Data security and sustainability

The data provider or creator's responsibility is in the preparation of datasets with the extensive documentation described above, and accurate and up-to-date metadata. Data security, curation, maintenance and sustainability are the responsibility of the hosting infrastructure or repository.

4.5 Summary of data management guidelines

The following table lists the data management guidelines described above, and the possible roles in each case for three stakeholders involved with the data lifecycle: funders, repositories and infrastructures, and researchers who deposit data.

	Funders	Repositories	Researchers
Thematic repository	Request deposition, suggest repository	• Provide concrete guidelines	• Identify relevant repository
Metadata model	Recommend if needed	• Provide metadata model with concrete guidelines for depositors • Export metadata and use standard protocols for metadata harvesting	• Adopt model • Comply with guidelines • Convert/map existing metadata to suggested model
Data types	Recommend if needed	• Recommend • Specify acceptable types	• Comply with guidelines • Convert existing data types to requested standards
Data size	Recommend if needed	• Recommend • Specify acceptable size • Specify size units	• Comply with guidelines
Data provenance	Request data history	• Recommend • Specify mode of provenance tracking	• Comply with guidelines
Data format	Request use of standards	• Define standards adopted	• Comply with guidelines • Convert existing data formats to requested standards
Access tools	Request deposition if needed	• Demand deposition of access tools together with data	• Specify if data need specific access tools • Deposit access tools to repository
Versioning	Request versioning method/model	• Adopt versioning model	• Comply with guidelines
Sensitive data	• Request relevant policy • Impose adherence to relevant legislation	• Define relevant policy • Provide tools for data de-identification/anonymisation	• Comply with guidelines • Implement policy for dealing with sensitive data • Ensure data is anonymised
Licensing	Request/promote Open Access	• Define licensing schema • Provide licensing tools, licence-selecting wizards	• Comply with guidelines • Provide legal information about data

Persistent identification of data (PID)	Request use of PID	• Select PID provider • Adopt PID schema • Issue PIDs to depositors' data	• Comply with guidelines • Demand PIDs from repository
Quality assurance of data and metadata	Request relevant policy	• Define policy • Implement quality assurance methods and validation / evaluation tools • Check data and metadata quality	• Comply with guidelines • Implement data quality assurance prior to deposition or accept validation by repository
Data security, maintenance and sustainability	Request relevant policy and implemented procedure	• Provide policy, methods and tools for data security, maintenance and sustainability	• Check repository's policy prior to depositing

Table 1: Summary of data management guidelines

4.6 Conclusions

Good data management is a prerequisite to share research data in an effective way. As discussed in section **Error! Reference source not found.**, sound data management procedures result in:

- increase of data quality
- increase of research efficiency
- exposure of research data and results through sharing and dissemination
- facilitation of reproducibility of experimental procedures
- facilitation of validation and verification of results
- increase of interoperability between data and between data and tools
- improvement of repositories' and infrastructures' operation

All of these help to create scientific and economic value. Particularly given the tremendous potential of TDM technologies to create value,[44] it is important to design and follow a good Data Management Plan when starting any research project, to ensure the data you create and collect will be as valuable as possible.

4.7 Further materials

- RDA Metadata Standards Directory | http://rd-alliance.github.io/metadata-directory/
- European Commission H2020 Manual on Open Access and Data Management | http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/open-access_en.htm

Guidelines for data management plans

- University of Twente Guidelines Data Management Plan | https://www.utwente.nl/igs/datalab/datamanagement/guidelinesdmp/
- ICPSR Guidelines for Effective Data Management Plans | https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/dmp/
- University of Oregon Libraries Research Data Management Best Practices | https://library.uoregon.edu/datamanagement/repositories.html

[44] FutureTDM Deliverable D5.2: Trend analysis, future applications and economics of TDM (PDF)

- University of the Witwatersrand, Johannesburg Digitisation, Preservation, Curation and Data Management: Research Data Policies and Best Practices/Guidelines | http://libguides.wits.ac.za/c.php?g=145348&p=953464
- DMPTool Data Management General Guidance | https://dmptool.org/dm_guidance
- Data management – Wikipedia | https://en.wikipedia.org/wiki/Data_management
- University of Queensland Library Research data management: Get started | http://guides.library.uq.edu.au/research-data-management/get-started

5. GUIDELINES FOR SUPPORTING TDM AT UNIVERSITIES

5.1 Introduction

Text and data mining can refer to a broad range of different activities, but fundamentally it involves using computer algorithms to analyse content and generate new knowledge. Computers are able to process data on a much larger scale than any human reader, and assess more variables across more datasets. The potential benefits to researchers are vast – but uptake and use of TDM technology in Europe is thus far lagging behind other areas of the world.

The FutureTDM project has produced several sets of guidelines to help address specific challenges to greater use of text and data mining (TDM) technologies. These include guidelines to help practitioners better understand the legal and licensing situation around TDM, and guidelines to help creators of data and content manage and share their data. In each case these guidelines are intended to be relevant to all stakeholders involved in TDM and data creation activities, from academia to industry to the general public.

One finding that emerged from the FutureTDM project, however, was that universities in particular have a key role to play as stakeholders in the TDM landscape. These guidelines aim to highlight just how and why universities play such a crucial role, and provide practical advice on what universities can do strategically to support TDM within their own institutions and across the wider TDM landscape.

As we will explain, supporting TDM and related skills[45] has the potential to bring a wide range of benefits to universities and their students and researchers.

5.2 Universities as key stakeholders

It would be difficult to exaggerate the potential impact universities can have on the uptake of TDM technologies. In FutureTDM's policy framework and recommendations, we identified universities as influential stakeholders in all stages of the TDM value chain.[46]

5.2.1 The TDM value chain

Content creation

Universities and their researchers generate large amounts of research data and other content that are potentially valuable to TDM practitioners. As discussed in our Data Management Guidelines, making sure this content is managed and shared according to best practices offers benefits for universities both as content creators and owners, as well as potential re-users of others' content.

[45] Foundational skills that are relevant to TDM include general data literacy, best practices in data management, and data science and analytical skills. These are collectively described as data- or TDM-related skills and practices throughout these guidelines.

[46] See the FutureTDM Policy Framework as outlined in Deliverable D5.1 (PDF) for a detailed discussion of the TDM value chain, and an overview of the roles various stakeholders can play in supporting the uptake of TDM.

Content dissemination

Universities are increasingly playing a role in the dissemination of content created by their staff and researchers. Again, by following best practices for data management and sharing, universities can maximise the visibility and value of their shared content.

Text and data mining

The use and development of text and data mining technologies is a rapidly growing field in universities around Europe. Whether in dedicated departments or research groups, or as part of broader programmes of data science and related technologies, many applications of TDM technology are happening within and around universities.

Value creation

Similarly, universities play a key role in the translation of TDM research into new knowledge, insights and technologies. Whether this be through disseminating the results of TDM research projects, or turning newly-developed TDM tools and algorithms into spin-out companies, universities play a key role in generating scientific and economic value through the use of TDM.

Skills and education

Although not an explicit part of the TDM value chain, all aspects of the TDM landscape rely on stakeholders and practitioners having access to the necessary skills, education and support to carry out TDM projects. Universities play a crucial role in supplying the necessary skills, education and awareness that underpin all other aspects of the TDM value chain.

5.2.2 The role of universities

As well as their direct involvement in the above aspects of the TDM landscape, the structure of universities, and their positioning in wider social and economic landscapes, make them uniquely placed to support the uptake of TDM technologies:

- As cross-disciplinary bodies, universities can facilitate the sharing of knowledge and skills across multiple domains, helping to bridge the gaps between areas that are typically more advanced in data-related skills, and those that are less likely to work with large-scale digital data.[47]
- As institutions, universities negotiate licences for access to content on behalf of their researchers, and can use their bargaining position to ensure those licences allow TDM.[48]
- Universities are trusted voices in the community, and have large networks through which they can communicate and demonstrate the value of TDM technologies and their applications.
- Through public-private partnerships with industry, universities can ensure that students and researchers learn data-related skills that are relevant and valuable to industry and the economy.

[47] See section 14 of FutureTDM Deliverable D3.3+ for a detailed report on the skills gaps across different disciplines in TDM education and skills.
[48] For guidelines on how appropriate licences can help support TDM, see our Guidelines for Licensees.

Improving awareness and support for TDM and related skills and technologies in universities will therefore bring further benefits beyond just the aspects of the TDM value chain in which universities are directly involved.

5.2.3 Benefits to universities

The principal reason for universities to invest resources in supporting TDM is that data science and analysis is fast becoming fundamental to all areas of research and education. Students need to understand data and how to manipulate it in order to be comfortable and confident in the modern world. In the words of one professor at a leading university, "Data science is the new IT."

Education in data analytics is fast becoming essential, even in what may seem to be unlikely fields. In the fashion industry, for example, business is largely conducted online – which means students need a sensitivity and awareness of the value that data analytics, including TDM, can bring to their business models.

"Data science is the new IT."

More broadly though, universities benefit from the aggregated impact of their researchers,[49] and TDM has the potential to increase the progress of research exponentially. From helping researchers to sort through the ever-growing volume of academic literature, to developing entirely new research and analysis techniques based on new algorithms, TDM can on the one hand save time and money, and on the other become a foundation for all kinds of innovations and research discoveries. Supporting and encouraging training in TDM and related skills will make university graduates better prepared to operate in a big data world. TDM technologies will help to make research more effective and efficient, uncovering new insights and saving time and money through automated processes. We should not be content to leave the teaching of TDM and related skills to just a few specialised universities, when the potential value to all research fields and sectors of the economy is so significant.

5.3 Challenges

Through interviews, workshops, and other consultations with stakeholders, the FutureTDM project identified several significant barriers that hinder greater uptake of TDM in universities. These are discussed in depth in the FutureTDM report on policies and barriers to TDM in Europe,[50] but the major barriers are summarised below.

5.3.1 Lack of awareness

Even though TDM technologies began emerging in the 1990s, TDM is still seen by many as a new, or even a "niche" field. Particularly outside of traditionally data-driven disciplines, there can be less awareness of or interest in TDM – despite its vast potential for applications in all areas. However, awareness does vary by institution, and in some cases digital humanities programmes are already developing and implementing applications for TDM.

[49] The LERU Roadmap towards Open Access, June 2011 (pdf)
[50] FutureTDM Deliverable D3.3+

5.3.2 Fragmentation of resources

In the context of supporting TDM, the sheer scope and breadth of activities that universities typically engage with are potentially one of their greatest strengths, allowing them to bridge gaps between different fields and roles. But actually, overcoming those gaps can equally be one of the greatest challenges to implementing new policies to support TDM.

Depending on a given university's structure, there may be multiple groups interested in, or even already pursuing, TDM-related initiatives and ideas. This can lead to confusion and duplicated effort – but most significantly, without a central or coordinated approach to supporting TDM, progress and learnings are not shared among different parts of the university.

This fragmentation is a significant barrier to the uptake of TDM, and makes it difficult for anyone interested in TDM to know whom to turn to for support or advice – whether they be a researcher looking to develop TDM applications themselves, a researcher looking for others to help apply TDM solutions, or even a librarian looking to better understand TDM technology. Time and again, the feedback from stakeholders consulted during the FutureTDM project was that the absence of coordinated support from their universities forces them to rely on ad hoc personal networks for information and support regarding TDM.

5.3.3 Skills gaps

A further challenge is bridging the gap between students or researchers in a given domain, and experts in data-related skills. Just as researchers may not have the expertise to understand TDM tools and applications, experts in TDM may not have the domain-specific knowledge to understand how those tools can be applied in a given discipline. As one university librarian commented, "You almost need two specialities," to be able to understand how TDM can be utilised to solve problems within a given discipline.

5.3.4 Lack of resources

The lack of awareness, and fragmentation of existing interest in TDM, has obvious repercussions in a broader lack of resources supporting TDM and related skills. Without funding for dedicated roles to support TDM, universities are left to rely on staff with the ability and motivation to support TDM in their own time, on top of other responsibilities.

This is of course difficult for existing staff to manage on top of their existing workloads, as supporting TDM not only requires learning about these new and emerging technologies, but also coordinating among different stakeholders who may need to be consulted for a TDM project, and advocating for further uptake of and support for TDM. There is a crucial need for roles that have the time and backing to properly support and coordinate TDM activities across universities, as well as staff with a range of capabilities who can bridge the gaps between domain experts and TDM experts.

5.4 Paths forward

As of April 2017, few universities have taken concrete steps towards designing or implementing strategic policies to support TDM. However, a large number of the stakeholders we talked to expressed a desire to do more to support TDM within their institutions, as well as an interest in learning from any progress other universities have made.

In this section, we suggest some steps you or your institution might take to work towards developing and implementing policies and strategies to support data science and analytics. We heard from many stakeholders that there is still considerable foundational work to be done around awareness and implementation of data literacy, data management, and data science before addressing support for TDM specifically; the example cases below therefore focus largely on supporting these foundations of TDM. However, the general principles in each step should apply to supporting TDM and related skills at any stage.

More detailed case studies will be available on the FutureTDM website.[51]

5.4.1 Demonstrate need

In 2015, Ghent University carried out a survey and series of interviews across all research and education faculty to understand what skills the university library, and the institution as a whole, should focus on investing in and supporting. Among other things, the results of this project highlighted a need for better skills in data management and data science.

The university library used the results of this project as evidence to investigate how education in data management skills could be introduced into core university curricula. Teaching of some of these skills has now been implemented in doctoral schools, and ultimately the library's project aims to make these part of the curricula of every Master's degree by 2018 (with the depth and detail covered varying by field).

Although not explicitly educating students about TDM itself, laying the foundations for good data management practices is a key first step towards supporting greater use of data analytics. More generally, consulting with the university community and gathering evidence to support a need for better investment in data-related skills is a strategy that can be used to drive policies of investment and support in this area.

5.4.2 Involve all stakeholders

As discussed above, within a university there are likely to be many people and communities who stand to gain from better investment in data-related skills, or who would be impacted by policies in this area. Without buy-in from all these stakeholders, it is difficult to gather support for new strategies and policies around data science and analytics. These stakeholders may include researchers in all fields, from sciences to humanities; librarians and library staff; heads of faculty in education, as well as research; students; IT departments; and other specialised support groups for activities such as software development and e-research.

Several universities have described planning or holding workshops to bring together stakeholders from libraries, IT departments, and research and education faculties to ensure that all parties' needs and concerns are understood. In some cases, these have been supplemented by ongoing working groups, to continue evaluating and refining plans for policies around data. These sorts of initiatives to "get everyone on the same page" are key to developing strategies to support data science and analytics.

[51] http://www.futuretdm.eu/awareness-sheets/

5.4.3 Understand relevant university processes

Each university has its own organisational structure, and its own internal hierarchies; it would not be possible for these guidelines to suggest a general strategy for introducing data-related skills to education curricula or specific research departments. It is therefore a crucial first step to consult and communicate with education and research faculties to understand how their processes work, in order to devise a centralised, coordinated plan to integrate better support for TDM and data science into those areas.

Looking again at Ghent University, the university library initially had little insight into how the education arm of the university was structured or organised. Through consultations with the directors of research and education of every faculty of the university, they developed an understanding of the "learning lines" followed through degree structures, and the basic competencies these cover. While some competencies were discipline-specific, many were general skills taught across all degrees. By identifying places where skills around data literacy and management might fit into existing competencies, the library was able to build a case to suggest places to incorporate these skills into existing education streams – and also to highlight gaps in existing "learning lines" and competencies.

In general terms, understanding how education policy and practice is structured at your university will help you to understand and express how education in data-related skills might be implemented, in a way that education policy-makers understand and recognise.

5.4.4 Consolidate information about resources

As discussed above, support for different aspects of TDM, where it exists, is often fragmented across multiple departments of a university. Many people are unsure of whom they can turn to for help within their own institutions, and what resources might be available to them. Bringing together information about and access to these resources in a centralised, coordinated place makes it much easier for anyone interested in TDM to understand whom they can turn to for support.

This applies to more than just the technical aspects of carrying out TDM itself; researchers may need advice on how to manage, store and share their data, or expert legal advice to understand whether their intended use of TDM is lawful. Researchers should be able to quickly and easily discover how their libraries, IT departments, legal teams, and any other relevant departments can help support TDM.

5.4.5 Identify early adopters and champions

As with any emerging technology, there are likely to be people within your university community who are already interested in TDM and related skills.

The University of Cambridge's "Data Champions"[52] are a group of volunteers who are given support and freedom to host workshops, and create and disseminate materials around data management best practices. The Data Champions programme began with a call for volunteers from within the university's research community to help support data management. These self-identified Data Champions bring with them the domain-specific knowledge and expertise necessary to understand

[52] http://www.data.cam.ac.uk/intro-data-champions

what data management means in practical terms within their specific disciplines, and as researchers themselves, they also have a trusted voice among their peers and colleagues.

The Data Champions programme has thus far seen significant progress in encouraging awareness and best practices around data management. Conversely, the same university's attempts to train librarians in domain-specific skills to better understand the data management needs of specific disciplines were less successful.

Rather than focussing exclusively on teaching domain-specific expertise to librarians, or IT or other support staff, identifying and working with researchers who are keen to understand and promote the importance of data-related skills appears to be a much more effective step towards bridging the gap between researchers and data experts.

5.4.6 Introduce appropriate incentives

Where funding is available, this can be used to hire experts for dedicated roles to support TDM and related skills. For example, the Delft University of Technology, with support and funding from senior management, has recently begun a hiring process for "Data Stewards" to help support best practices in data management at the university. If you are considering a programme of support for TDM, it is worth considering funding sources both internal and external to the university.

However, other incentives can also be used to encourage people within the university community to support TDM and related practices. At the University of Cambridge, volunteer Data Champions are given public recognition via the university website,[53] as well as internal recognition in the form of reference letters to their heads of department. At University College London, PhD students in computer science have been working to create data science education materials aimed at a pre-university level, and have been happy to volunteer their time towards this project as a public good.

The best way to understand what incentives would encourage stakeholders within the university community to help support TDM and related practices is, of course, to consult with them.

5.4.7 Share your results and progress

Thus far, few universities have taken concrete steps towards policies supporting TDM. Many more are interested in supporting TDM in principle, but uncertain how to go about developing and implementing policy in this area. If you or your institution has had success in developing strategies around supporting TDM and data-related skills, we encourage you to share and promote your success stories with others.[54]

5.5 Summary of key points

Universities are uniquely placed to support the uptake of TDM, but to fully capitalise on this potential ultimately requires a central, coordinated approach to bring together information, resources, and people in this field. Some ways that you can gather and encourage support and awareness around TDM at your university include:

[53] http://www.data.cam.ac.uk/datachampions
[54] We welcome everyone to share their experiences with TDM on the FutureTDM blog.

- carrying out surveys or interviews to demonstrate a need for TDM support;
- involving and engaging with all stakeholders across the university;
- understanding the organisation of education and research faculties, and how TDM and related skills might fit into existing processes;
- bringing together information about and access to resources, to make these easily discoverable by people interested in TDM;
- finding and identifying early adopters who are personally interested in promoting or pursuing TDM;
- introducing incentives for people to engage with TDM-related initiatives; and
- sharing experiences and success stories so that others can learn from them.

Case studies of how particular universities have begun to address support for TDM will be shared via the FutureTDM blog[55] and awareness sheets.[56]

[55] http://www.futuretdm.eu/category/blog/
[56] http://www.futuretdm.eu/awareness-sheets/

6. CONCLUSIONS

As discussed in the introduction, this document represents a first iteration of practical guidelines for stakeholders in the TDM value chain. They will be updated throughout the remainder of the project, and supplemented by case studies disseminated via the FutureTDM website.[57]

[57] http://www.futuretdm.eu/knowledge-library/