Data Management Guidelines for Researchers

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!

2. Background

2.1 What is data management?

According to DAMA, The Global Data Management Community,[1] "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."[2]

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!

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.

[1] https://www.dama.org/

[2] 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.)*

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".

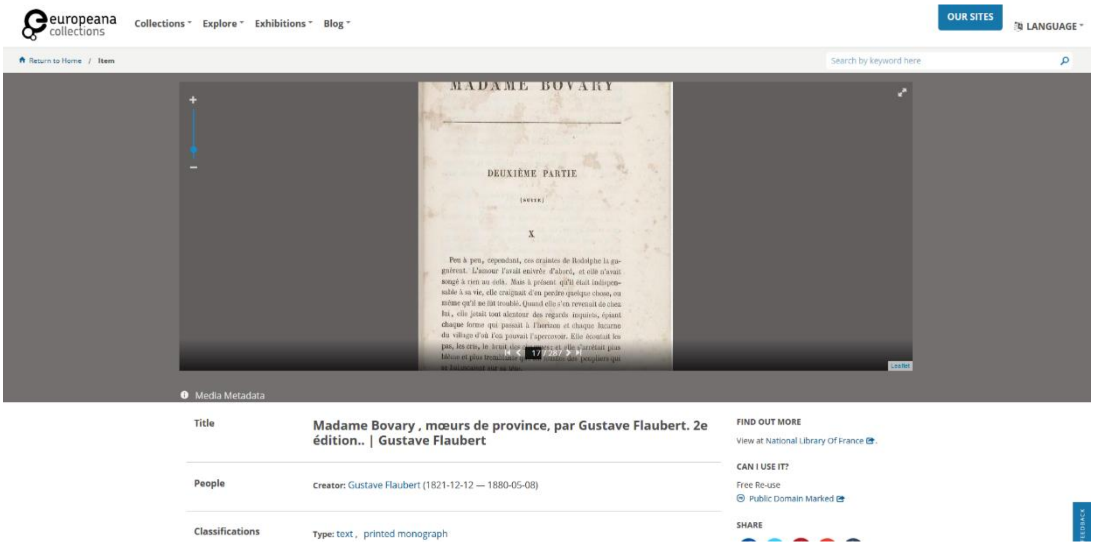
Human vs. Machine access and use of data

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[3] 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:[4]



MADAME BOVARY

DEUXIÈME PARTIE

(SUITE)

X

Peu à peu, cependant, ces craintes de Rodolphe la gagnèrent. L'amour l'avait enivrée d'abord, et elle n'avait songé à rien au delà. Mais à présent qu'il était indispensable à sa vie, elle craignait d'en perdre quelque chose, ou même qu'il ne fût troublé. Quand elle s'en revenait de chez lui, elle jetait tout alentour des regards inquiets, épiant chaque forme qui passait à l'horizon et chaque lucarne du village d'où l'on pouvait l'apercevoir. Elle écoutait les pas, les cris, le bruit des charrues; et elle s'arrêtait plus blême et plus tremblante que les feuilles des peupliers qui se balançaient sur sa tête.

Title: Madame Bovary , mœurs de province, par Gustave Flaubert. 2e édition.. | Gustave Flaubert

People: Creator: Gustave Flaubert (1821-12-12 — 1880-05-08)

Classifications: Type: text , printed monograph

FIND OUT MORE
View at National Library Of France.

CAN I USE IT?
Free Re-use
Public Domain Marked

SHARE

Figure 1: *Madame Bovary* as provided by the Europeana website

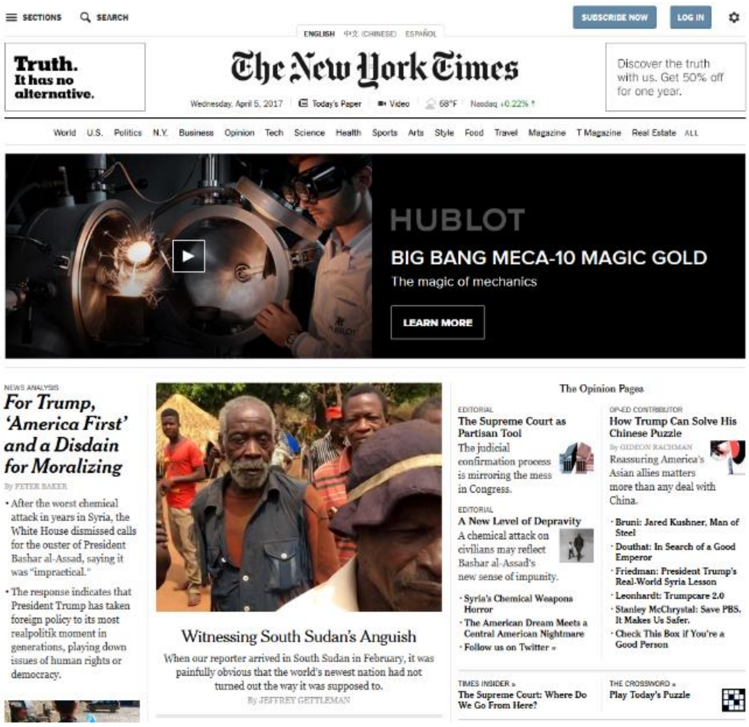
The human user has no problem reading this version, which is free for re-use, but it is completely inappropriate for TDM processes. This has to do with its format, which the computer does not recognise as text (which can be analysed by TDM processes), but a jpg image file which is unprocessable by most 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

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**,[5] 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.

[3] http://www.europeana.eu/portal/en
[4] Madame Bovary (accessed 17 April 2017)
[5] https://www.nytimes.com

This project has received funding from the European Union's Horizon 2020 (H2020) Research and Innovation Programme.



Truth.
It has no
alternative.

Discover the truth with us. Get 50% off for one year.

NEWS ANALYSIS

For Trump, 'America First' and a Disdain for Moralizing

By PETER BAKER

- After the worst chemical attack in years in Syria, the White House dismissed calls for the ouster of President Bashar al-Assad, saying it was "impractical."
- The response indicates that President Trump has taken foreign policy to its most realpolitik moment in generations, playing down issues of human rights or democracy.

Witnessing South Sudan's Anguish

When our reporter arrived in South Sudan in February, it was painfully obvious that the world's newest nation had not turned out the way it was supposed to.

By JEFFREY GETTLEMAN

The Opinion Pages

EDITORIAL

The Supreme Court as Partisan Tool

The judicial confirmation process is mirroring the mess in Congress.

EDITORIAL

A New Level of Depravity

A chemical attack on civilians may reflect Bashar al-Assad's new sense of impunity.

- **Syria's Chemical Weapons Horror**
- **The American Dream Meets a Central American Nightmare**
- **Follow us on Twitter »**

OP-ED CONTRIBUTOR

How Trump Can Solve His Chinese Puzzle

By GIDEON RACHMAN

Reassuring America's Asian allies matters more than any deal with China.

- **Bruni: Jared Kushner, Man of Steel**
- **Douthat: In Search of a Good Emperor**
- **Friedman: President Trump's Real-World Syria Lesson**
- **Leonhardt: Trumpcare 2.0**
- **Stanley McChrystal: Save PBS. It Makes Us Safer.**
- **Check This Box if You're a Good Person**

TIMES INSIDER »

The Supreme Court: Where Do We Go From Here?

THE CROSSWORD »

Play Today's Puzzle

Figure 2: 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 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;

This project has received funding from the European Union's Horizon 2020 (H2020) Research and Innovation Programme.

- 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.

2.3 What is a Data Management Plan?

As the EU *Guidelines on FAIR Data Management in Horizon 2020 (Version 3.0)*[6] 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;
- 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

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

- 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.

3. What are the benefits of data management?

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.

This project has received funding from the European Union's Horizon 2020 (H2020) Research and Innovation Programme.

- **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.

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. 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.

4.1 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*[7], 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.

[7] 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)

This project has received funding from the European Union's Horizon 2020 (H2020) Research and Innovation Programme.

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,[8] 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[9]).

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.

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

[8] https://europepmc.org/

[9] https://zenodo.org/

This project has received funding from the European Union's Horizon 2020 (H2020) Research and Innovation Programme.

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

This project has received funding from the European Union's Horizon 2020 (H2020) Research and Innovation Programme.

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.[10] 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[11], 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[12] (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),[13] AGPL (GNU Affero General Public

[10] Insofar as it is not possible to re-identify the data subjects from the metadata; see the FutureTDM Legal Guidelines for TDM Practitioners for guidelines on protecting personal data.

[11] For a list of such tools see the Australian National Data Service (ANDS) Guide to Sharing and Publishing Sensitive Data.

[12] https://creativecommons.org/

[13] http://opensource.org/licenses/gpl-license.php

This project has received funding from the European Union's Horizon 2020 (H2020) Research and Innovation Programme.

License),[14] Apache Licence 2.0,[15] BSD,[16] GFDL (GNU Free Documentation License),[17] or LGPL (GNU General Public License)[18] 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-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."[19] 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.[20] 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),[21] the Handle System,[22] Persistent Uniform Resource Locators (PURLs),[23] Uniform Resource Names (URNs),[24] and Extensible Resource Identifiers (XRIs).[25] DataCite[26] 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.

4.3 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.

[14] http://www.gnu.org/licenses/agpl-3.0.html
[15] http://www.apache.org/licenses/LICENSE-2.0.html
[16] https://opensource.org/licenses/BSD-2-Clause
[17] http://www.gnu.org/copyleft/fdl.html
[18] http://www.gnu.org/licenses/lgpl.html
[19] https://en.wikipedia.org/wiki/Persistent_identifier
[20] A 2015 assessment of 180,000 web links cited in research articles found that 24.5% of them were unavailable.
[21] https://en.wikipedia.org/wiki/Digital_Object_Identifier
[22] https://en.wikipedia.org/wiki/Handle_System
[23] https://en.wikipedia.org/wiki/Persistent_uniform_resource_locator
[24] https://en.wikipedia.org/wiki/Uniform_Resource_Name
[25] https://en.wikipedia.org/wiki/Extensible_Resource_Identifier
[26] https://www.datacite.org/

This project has received funding from the European Union's Horizon 2020 (H2020) Research and Innovation Programme.

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.

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 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.

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	• Adopt versioning model	• Comply with guidelines

This project has received funding from the European Union's Horizon 2020 (H2020) Research and Innovation Programme.

	method/model		
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

6. Conclusions

Good data management is a prerequisite to share research data in an effective way. As discussed in section 3, 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,[27] 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.

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

7. Further materials

Data management standards

- 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
- 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

This project has received funding from the European Union's Horizon 2020 (H2020) Research and Innovation Programme.