METIS

Research and Innovation Action (RIA)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 945121

Start date : 2020-09-01 Duration : 48 Months

Data Management Plan

Authors : Mr. Gilles QUENEHERVE (LGI)

METIS - Contract Number: 945121

Project officer: Katerina PTACKOVA

Document title	Data Management Plan
Author(s)	Mr. Gilles QUENEHERVE
Number of pages	17
Document type	Deliverable
Work Package	WP1
Document number	D1.4
Issued by	LGI
Date of completion	2021-03-24 13:56:08
Dissemination level	Public

Summary

This is deliverable ?D1.4 Data Management Plan? of the METIS project. This document is structured in three distinct parts. The first part provides some basic contextualisation of a data management plan in H2020 projects and reminds some key definitions necessary to understand the deliverable correctly. The second part aims at explaining the scope of METIS Data Management Plan, in other words, the purpose of data collection in the frame of METIS, the type, format and the origin of the data to be collected and the use and re-use of the data. The third part details the various actions METIS will implement to make its data findable, accessible, interoperable and available for re-use. According to the EU?s guidelines regarding the DMP (European Commission, 2016), the document may be updated - if appropriate - during the project lifetime (in the form of deliverables). DMPs should, therefore, have a clear version number and include a timetable for updates.

Approval

Date	By
2021-03-24 15:41:54	Dr. Irmela ZENTNER (EDF)
2021-03-24 16:09:47	Dr. Irmela ZENTNER (EDF)

METIS

Seismic Risk Assessment
for Nuclear Safety

Research & Innovation Action

NFRP-2019-2020

Data Management Plan

Deliverable D1.4

Version N°1

Author: Gilles Quénéhervé (LGI)

This project has received funding from the Horizon 2020 programme under grant agreement n°945121. The content of this presentation reflects only the author's view. The European Commission is not responsible for any use that may be made of the information it contains.

Disclaimer

The content of this deliverable reflects only the author's view. The European Commission is not responsible for any use that may be made of the information it contains.

Document Information

Grant agreement	945121
Project title	Methods And Tools Innovations For Seismic Risk Assessment
Project acronym	METIS
Project coordinator	Dr. Irmela Zentner, EDF
Project duration	1st September 2020 – 31st August 2024 (48 months)
Related work package	WP 1 Management
Related task(s)	Task 1.6 – Data Management
Lead organisation	LGI
Contributing partner(s)	EDF
Dissemination level	Public

Abbreviations and Acronyms

Acronym	Description
WP	Work Package

Table of content

List of figures

List of tables

Summary

This is deliverable "D1.4 Data Management Plan" of the METIS project.

This document is structured in three distinct parts. The first part provides some basic contextualisation of a data management plan in H2020 projects and reminds some key definitions necessary to understand the deliverable correctly. The second part aims at explaining the scope of METIS Data Management Plan, in other words, the purpose of data collection in the frame of METIS, the type, format and the origin of the data to be collected and the use and re-use of the data. The third part details the various actions METIS will implement to make its data findable, accessible, interoperable and available for re-use.

According to the EU's guidelines regarding the DMP (European Commission, 2016), the document may be updated - if appropriate - during the project lifetime (in the form of deliverables). DMPs should, therefore, have a clear version number and include a timetable for updates.

Keywords

Data, Knowledge, Information, Data Management

1. INTRODUCTION

1.1. Background of the Data Management Plan in H2020

This document is the METIS Data Management Plan (DMP), a deliverable required by the European Commission for every project participating in the Open Research Data (ORD) Pilot. Since 2017, the European Commission has extended this pilot project to all thematic areas of H2020, making METIS part of this pilot. According to the European Commission[1], Open access (OA) refers to the practice of providing online access to scientific information that is free of charge to the end-user and reusable. 'Scientific' refers to all academic disciplines. In the context of research and innovation, 'scientific information' can mean:

- peer-reviewed scientific research articles (published in scholarly journals);
- research data (data underlying publications, curated data and/or raw data).

The rationale behind the Pilot project is that data management is not a goal in itself but a key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse. It follows from this that the ORD pilot aims to improve and maximise access to and re-use of research data generated by H2020 projects while balancing openness and protection of scientific information, commercialisation and Intellectual Property Rights, privacy concerns, etc[2].

Data Management Plans (DMPs) are a key element of good data management. As part of making research data findable, accessible, interoperable and re-usable (FAIR), a DMP should include information on the data life cycle:

- the handling of research data during and after the project.
- what data will be collected, processed or generated.
- what methodology and standards will be applied.
- whether data will be shared/made open and how.
- how data will be curated and preserved.

While open access to research data has become applicable by default in H2020, the Commission acknowledges that there could be good reasons to keep the research data confidential. Therefore, the Commission also provides opt-out possibilities.

The METIS DMP was written in reference to the task 1.4 of the Grant Agreement on data management activities.

1.2. Definition

Before explaining the consortium strategy in terms of Data Management, several terms must be defined:

- Data: Data refers to unstructured facts and figures, which are not organised in any way and which provide no further information regarding patterns, context, etc. For instance, data on production, demand, results from technical tests and so on, is unstructured data.
- Information: For data to become information, it must be contextualized, categorized, calculated and condensed. Information thus paints a bigger picture; it is data with relevance and purpose. It may convey a trend in the environment, or perhaps indicate a pattern of sales for a given period of time.
- Knowledge: Knowledge is closely linked to doing and implies know-how and understanding. The knowledge possessed by every individual is a product of his/her experience and encompasses the norms by which s/he evaluates new inputs from his/her surroundings. For instance,

knowledge is related to the know-how acquired in R&D projects, commercial activities or the expertise that is inherent to each partner[3].

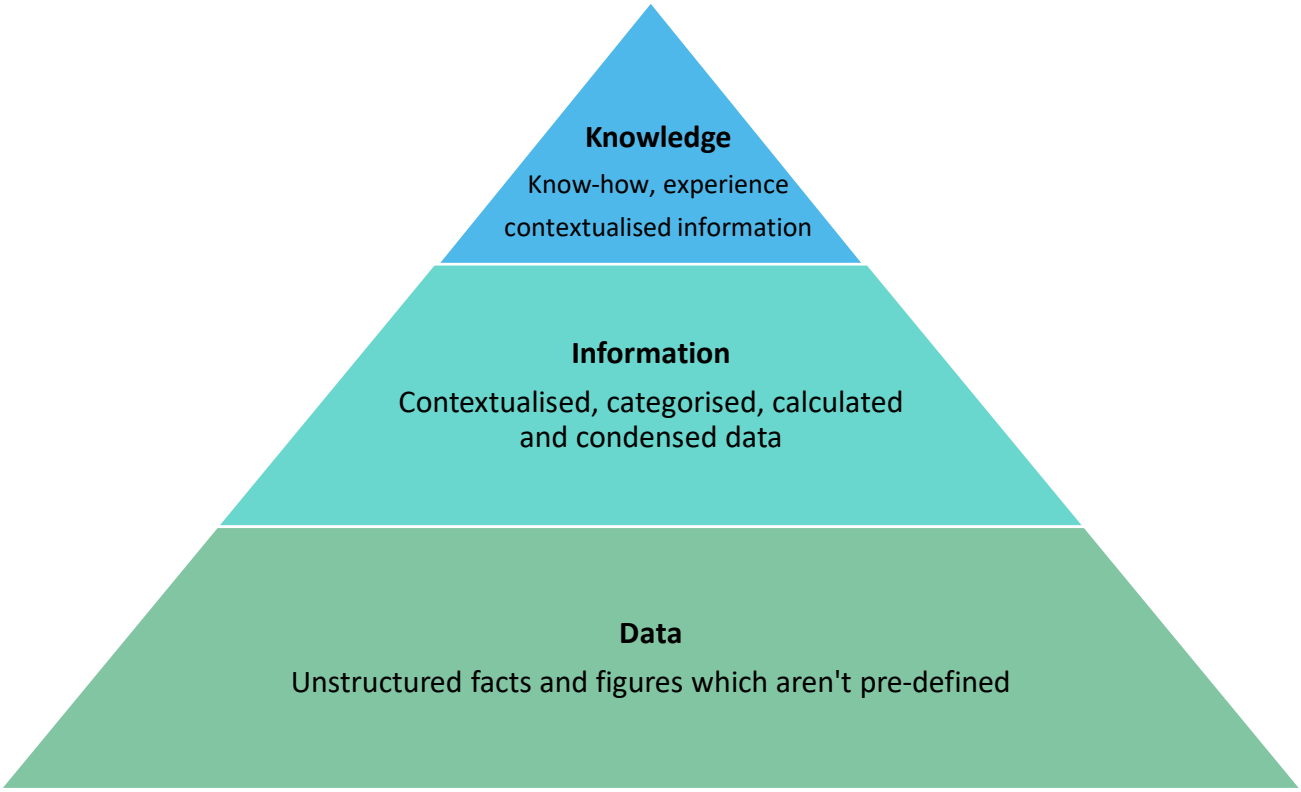

Figure 1: Knowledge Management – Definitions and hierarchy

This present DMP will mainly deal with how the data will be managed and will mention superficially the links with knowledge.

- Data codebook: A codebook is an essential document that informs the data user about the study, data file(s), variables, categories, etc., that make up a complete dataset. The codebook may include a dataset's record layout, list of variable names and labels, concepts, categories, cases, missing value codes, frequency counts, notes, universe statements, and so on[4].
- Data set: a data set is a collection of data. Most commonly a data set corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable. The data set lists values for each of the variables, such as height and weight of an object, for each member of the data set. The data set may comprise data for one or more members, corresponding to the number of rows[5].

1.3. Versions

According to the EU's guidelines regarding the DMP (European Commission, 2016), the document may be updated - if appropriate - during the project lifetime (in the form of deliverables). The DMP should be updated as a minimum in time with the periodic evaluation of the project. If there are none, such an update needs to be made in time for the final review at the latest.

The DMP is intended to be a living document in which information can be made available on a finer level of granularity through updates as the implementation of the project progresses.

DMPs should, therefore, have a clear version number and include a timetable for updates.

2. Scope and lifecycle of the METIS data

2.1. General framework for data collection

In this section, the data to be collected in each WP of the METIS project will be presented and described in order to define the purpose of the collection as well as to previously define the type, format and origin.

2.1.1. WP1 – Project Management

No specific research data is planned to be collected from WP1, as it is devoted to management. Personal data from project partners, advisory boards, will be collected. The Data Management Plan (DMP) will be developed and updated when relevant.

2.1.2. WP2 – Dissemination, exploitation & training

WP2 aims at maximising the project's impact within the nuclear community, in particular by disseminating the outcomes of the project across the nuclear community, with particular focus on industry, safety. A survey for the End-user group will be created to make sure that the project meets expectations and facilitate effective involvement of end-users and regulators throughout the project programme. Information collected will therefore be mainly dedicated to stakeholder involvement.

2.1.3. WP3 – Case study for implementation and application of METIS results

WP3 aims is to provide a coherent framework for integration of all WPs final results, where the final inputs and outputs of WPs are logically linked. WP3 guarantees that all METIS developments, scientific and engineering results are tested and verified in a real case study. WP3 will assures: the information transfer between WPs to facilitate the implementation in a case study, all data set related with the case study necessary for the implementation. A platform for storing and transferring data will be performed. Stored data concerns mainly data provided by private companies: technical reports, site maps, basins geometries, soils, structures and equipment data, models (PSHA, PSA…), seismic records, etc.

2.1.4. WP4 – Seismic Hazard Analysis

The objective of WP4 is the construction of methods and open-source tools for performing Seismic Hazard Analyses supporting modern Seismic Probabilistic Safety Assessments for critical facilities. In this WP, data will mainly be used for modelling the seismic hazard using tools such as the OpenQuake engine. Data produced as part of WP4 will be used in WP5.

2.1.5. WP5 – Ground motion selection for engineering analyses including site response

The goal of WP5 is to develop and apply a methodology for selecting sets of site hazard consistent ground motions appropriate for SSCs response and fragility analyses. The sets of ground motion time histories are selected from databases that represent a wide range of intensity levels. Both recorded ground motions (e.g., PEER NGA EAST and WEST2, and RESORCE) and synthetic ground motions from WP4 will be considered.

2.1.6. WP6 – Beyond Design and Fragility Analysis

WP6 will develop, verify, validate reliable mechanical models and failure criteria for detailed/specific fragility SSCs. WP6 will use different techniques (e.g. quantify aleatory and epistemic uncertainties, develop multi-dimensional (vector-based) fragility evaluation methods, implement Bayesian updating of

models and fragilities, etc.) to perform a holistic performance assessment. Based on the developed models and methodologies, fragility curves will be computed.

2.1.7. WP7 – PSA Tools and Methodology

WP7 aims to develop an open-source PSA tool and to address technical issues of seismic PSA by improving this code. This includes the following objectives: Integration of high (seismic driven) probability basic events in PSA model and results, creation of a single user-friendly database the mathematical expression of each component's seismic failure probability. This includes the assessment of shared parameters:

- Consideration in PSA computations an epistemic uncertainty associated with uncertainties of hazard and fragility curves;
- Validation and testing models against data;
- Implementation and testing new modelling options such as the consideration of vector valued hazard and fragility and aftershocks.

Type of data associated with WP7 includes: input data related to the case study from WP3 and WP6, research and output data. Research data include open-source PSA tool(s) and probabilistic models, Output data contain results of calculation.

2.2. Use and Re-Use of the data

The data collected and generated by the consortium will be useful to the development of further activities related to other WPs within the project, to specific end-users and to some EU regulators.

As part of WP2, data will be re-usable only within the consortium, and for METIS related activities, such as newsletter dissemination, posters presentation, invitation to events or specific news on the website.

WP3 handles the exchange of data related to common METIS case study.

As part of WP6, the sharing within the consortium of the model and fragility data produced by TUK is essential, as the analysis of the fragilities are needed for the achievement of the project objectives. Since the created data describes in detail the considered site, data sharing outside the project will be at the discretion of each NPP operator.

Under WP7, input data related to the case study is restricted only for the METIS project use. Research data will be processed and analysed using different languages (e.g. Python, C ++). Output data will be generated by probabilistic modelling computer software (e.g. SAPHIRE, Risk Spectrum, SCRAM, Andromeda, etc.). Formats used for the output datasets will be: ASCII, Microsoft excel files, proprietary code dependent formats. Raw data will be stored and managed by each project partner who produced them. Most relevant research and/or output data will be reported in the project deliverables. The large part of the WP7 deliverables have a public dissemination level and will be accessible on the project website what will favour the re-use of the data.

3. FAIR Data Management in METIS

In compliance with the ORD pilot, every H2020 project is required to draft a DMP in order to make the data findable (1), accessible (2), Interoperable (3) and available for re-use (4).

3.1. Making data findable

This section will provide insight into how METIS intends to make it easier to find data collected or produced by the consortium. The way to proceed in order to achieve this goal is to describe properly the content of the data sets.

3.1.1. Storing the data with datasets

- **Name and Reference code of datasets**

In order to imbue the names of datasets with easily identifiable meaning that conveys important information, the following naming convention shall apply:

CountryCode.DataOwner.Openness.Title

CountryCode: this string identifies the country to which the data pertains/where the data was collected using the ISO 3166 Alpha-2 coding system.

DataOwner: this string identifies the project partner in METIS that is associated with the dataset (data collector/custodian) using the official abbreviated partner names.

Openness: this string determines whether a given dataset is intended to be shared with the public as Open Data. It may take the following values:

1. Open: can be accessed, used and shared by anyone without limitations, accessible on the internet in a machine-readable format, free of restrictions on use in its licensing)
2. Shared: available to use, but not under an open data license. Restrictions on its use or reproduction may apply (limited to a given group of people or organisations, may not be reproduced without authorisation, etc.)
3. Closed: can only be accessed by its subject, owner or holder

Title: a short and descriptive string to identify the contents of the data

Using these strings, the name of a dataset would look like this:

FR.LGI.Open.CommuteHouseholdSurvey

A dataset with this name would describe a household survey on commuting preferences conducted in France and curated by LGI.

- **Description of the data**

Each data set that will be collected, processed or generated within the project will be accompanied by a brief table description. The following detailed information sheet will be produced for every dataset to be produced/collected/curated in the project.

Name of the data set[6]	Complete title of the data set

Description	• A brief, easy to understand description of what the dataset contains and what it will be used for in the project • A list of institutions to whom the data set could be useful outside the project • Whether the dataset has been/will be used for a scientific publication (if yes, brief details about the content and journal) If the dataset is collected, a brief description of its origin and how it was collected will be provided
Media Type	The physical medium of the content representation, e.g., video, image, text, numerical data, n-grams, etc.
Language(s)	The language(s) of the resource content
Use & re-use	Foreseen use of the resource for which it has been produced
Size	Size of the resource with regard to a specific size unit measurement in the form of a number
Format/license	The format in which the data will be available (e.g. .xls, .csv, .txt) will be provided. The license to be used will also be provided.
Version Number	Specify the version number of the document

Table 1: Table specifying the content of a dataset

If a dataset is directly collected, the origin of the data set will also be provided.

3.2. Making data openly accessible

Open accessibility of the data is the second key aspect for making data FAIR. This section will describe the type of data to be made available, its location and the procedure to obtain it.

Several degrees of accessibility are identified below¸ including both open access and restricted access.

3.2.1. Data licensing

Data licensing standards are used to layout the openness of data sets in concrete terms. There are many types of licenses to choose from, and this document will not cover them in depth. The table below provides a summary of common data licenses that will be considered for use in the project (based on definitions from opendefinition.org):

Name	Domain	Attribution	Share-alike*	Notes
Creative Commons CCZero (CC0)	Content, data	N	N	All rights (including those of attribution) waived
Open Data Commons Public Domain Dedication and Licence (PDDL)	Data	N	N	All rights (including those of attribution) waived

Creative Commons Attribution 4.0 (CC-BY-4.0)	Content, data	Y	N	Credit must be given, a link to the license must be provided, changes made must be indicated. If these terms are not followed, license may be revoked
Open Data Commons Open Database License (ODbL)	Data	Y	Y	Credit must be given, share-alike must be assured, data may be redistributed using DRM as long as a DRM-free version is also released

Table 2: Example of licenses

*Share-alike is the requirement that any materials created using the given dataset must be redistributed under the same license

3.2.2. Datasets that could be made openly accessible in METIS

Data producer	Brief description of dataset	Foreseen use & re-use	Possibility to share the data beyond the consortium
TUK	**Mechanical building models** For selected SSCs mechanical models will be created. The building data will come from the available information for the site.	**Foreseen Use:** Models will be used for fragility calculation, selection of failure relevant IMs and ground motions, uncertainty quantification and further research within the project. Therefore, data will be part of the case study.	**Re-use**: data sharing will be at the discretion of each NPP operator for data on their sites
TUK	**Fragility curves** For the generated model's fragility curves will be computed.	**Foreseen Use:** Fragility curves will be used for the PSA within WP 7 and case study within WP 3	**Re-use**: data sharing will be at the discretion of each NPP operator for data on their sites

Table 3: Datasets that could be made openly accessible in METIS

3.2.3. Datasets to remain confidential

Most of the data produced by METIS will be public. They are two main exceptions:

The products of WP5, namely the datasets of selected ground motions, cannot be made public at least not in their entirety. The reason is that some of the ground motions cannot be distributed freely by third parties like the project partners but only by the organizations that own them.

Second, as part of WP6, if it turns out NPP operators do not want data on their sites to be disclosed, this data will remain confidential, deposited on the METIS Electronic Collaborative Content Platform (ECCP), proposed by LGI, which enables the collaborative storage and the exchange of documents among the consortium. The identity of the person accessing the data will be ascertained during the registration process on the ECCP.

3.2.4. Data storage in METIS

After collection, data will be generally organised in Excel files and Word documents.

METIS will use Zenodo to systematically publish open data, open access presentations and public deliverables in order to maximise re-use and promote the project results. Prior to any upload, open publications on Zenodo will have to be approved by the Steering committee of METIS. If requested, LGI may provide its support and advice to the partners prior to the publication on Zenodo.

On the other hand, restricted data will be available on the FLEXX for at least five years after its collection.

3.3. Making data interoperable & increasing re-use

3.3.1. Making data interoperable

As described on 3.1, Standard vocabulary may be used on a case-by-case basis to make the data interoperable between researchers, institutions, organisations, countries, etc.

- A list of acronyms and/or abbreviations will be provided at the beginning of every report
- For surveys, standard definitions for entities such as trips, trip chains etc. will be adopted.

3.3.2. Restrictions for re-use

To be compliant with the new General Data Protection Regulation, data generated through interviews and surveys (e.g. as part of WP2) will not be re-used directly due to privacy concerns. To allow re-use and avoid loss of research data, two different techniques could be used to disseminate its data while abiding by regulations on privacy.

1) *Anonymization of data*[7]

"Anonymization" of data means processing it with the aim of irreversibly preventing the identification of the individual to whom it relates. Data can be considered anonymised when it does not allow identification of the individuals it is related to, and no individuals can be identified from the data by any further processing of that data or by processing it together with other information which is available or likely to be available.

There are different anonymization techniques. Here are the two most relevant:

- Generalisation: generalising data means removing its specificity. For example, in the case of a table containing household income levels, with 4 figures mentioned: $164,000, $58,543, $90,893, and $232,234. One way of generalising this numbers would be to write that the values are "more than $150,000, less than $60,000, between $90,000 and $100,000, and more than $225,000" respectively. Essentially it means taking exact figures, establishing a baseline category, and then obfuscating the data by assigning it to one of the categories in order to remove any sense of specificity from it.
- K-anonymity; A release of data is said to have the *k*-anonymity property if the information for each person contained in the release cannot be distinguished from the other individuals whose information also appear in the release. For instance, in a table composed of six attributes (Name, Age, Gender, State of Domicile, Religion and Disease), removing the name and the religion column while generalising the age is a way to effectively k-anonymise the data.

Other techniques, such as "masking" or "pseudonymisation", which are aimed solely at removing certain identifiers, may also play a role in reducing the risk of identification. In many cases, these techniques work best when used together.

2) Pseudonymisation of data

"Pseudonymisation" of data means replacing any identifying characteristics of data with a pseudonym, or, in other words, a value which does not allow the data subject to be directly identified.

Although pseudonymisation has many uses, it should be distinguished from anonymization, as it only provides limited protection for the identity of data subjects in many cases as it still allows identification using indirect means. Where a pseudonym is used, it is possible to identify the data subject by analysing the underlying or related data.

PNO may use this technique if it decides to disseminate its databases for stakeholder analysis.

3.3.3. Archiving and preservation

It is of the utmost importance for METIS' consortium to keep the data available for the partners after the end of the project.

As already mentioned in 3.2, to ensure medium-term preservation of the datasets, the consortium agrees to keep raw data on the ECCP for a minimum duration of 5 years. Anonymised data to be publicly shared by the consortium will be stored on Zenodo, which is a multi-functional open platform recognised by OpenAIRE and the European Commission.

These two types of storage have the advantage to be free of charge. It should, however, be noted that files exceeding 30 MB cannot be stored on the ECCP.

4. Human resources

Every Work Package Leader in METIS will be responsible for the data management within its own Work-Package.

5. Ethical aspects

5.1. GDPR

This Data Management Plan (DMP) was drafted and updated taking into account the General Data Protection Rules (GDPR) for the collection, storage and re-use of the data, in line with the following general principles.

Personal data shall be:

1. processed lawfully, fairly and in a transparent manner in relation to the data subject ('lawfulness, fairness and transparency');
2. collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes ('purpose limitation');
3. adequate, relevant and limited to what is necessary for relation to the purposes for which they are processed ('data minimisation');
4. accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay ('accuracy');
5. kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed; personal data may be stored for longer periods insofar as the personal data will be processed solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) subject to implementation of the appropriate technical and organisational measures required by this Regulation in order to safeguard the rights and freedoms of the data subject ('storage limitation');
6. processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures ('integrity and confidentiality')[8].

6. Bibliography

[1] European Commission. Directorate-General for Research & Innovation (March 2017). H2020 Programme Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020. Retrieved from: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

[2] European Commission. Directorate-General for Research & Innovation. (July 2016). H2020 Programme – Guidelines on FAIR Data Management in Horizon 2020. Retrieved from: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[3] Alan Frost. Defining Knowledge, Information, Data. (2017). Retrieved from: http://www.knowledge-management-tools.net/knowledge-information-data.html

[4] DDI Alliance. Create a codebook. Retrieved from: http://www.ddialliance.org/training/getting-started-new-content/create-a-codebook

[5] Rajit Dasgupta. 19 Free Public Data Sets for your first data science project. Springboard.com. (October 2015). Retrieved from : https://www.springboard.com/blog/free-public-data-sets-data-science-project/

[6] Maria Koutsombogera & Stelios Piperidis (Athena RC). D3.1 Data Management Plan. Cracker, Cracking the language barrier. (June 2015). Retrieved from: https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5a0203500&appId=PPGMS

[7] Data Protection Commissioner. Anonymization and Pseudonymization. Retrieved from: https://www.dataprotection.ie/docs/Anonymisation-and-pseudonymisation/1594.htm

[8] Intersoft Consulting. GDPR. Article 5. (2018) Retrieved from : https://gdpr-info.eu/art-5-gdpr/