CYBECO

Supporting Cyberinsurance from a Behavioural Choice Perspective

D2.2: Data Management Plan

Due date: M3

Abstract: D2.2 establishes a Data Management Plan (DMP) for the CYBECO project compliant with the Horizon 2020 FAIR Data Management Principles. These principles state that, in general terms, the research data should be findable, accessible, interoperable and reusable. As there will be many datasets and data exchanges during the project, it is key that this data is treated correctly.

This document is based on the Guidelines on FAIR Data Management in Horizon 2020 and is intended to be a living document that will be updated with further detail as the project progresses.

The current document is the first version (i.e., DMP v.1), which is also deliverable D2.2 that shall be submitted to the European Commission in the third month of the project (M3).

Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Document Status

Document Title	Data Management Plan
Version	1.0
Work Package	2
Deliverable #	2.2
Prepared by	CSIC
Contributors	A. Couce Vieira, D. Ríos Insua, N. Vasileiadis, J. Vila
Checked by	
Approved by	
Date	Jul 31, 2017
Confidentiality	PU

Document Change Log

Each change or set of changes made to this document will result in an increment to the version number of the document. This change log records the process and identifies for each version number of the document the modification(s) which caused the version number to be incremented.

Change Log	Version	Date
First draft	0.1	Jul 7, 2017
Final draft	0.9	Jul 27, 2017
Final version	1.0	Jul 31, 2017

Table of Contents

1 Introduction

1.1 Objective and Scope

The objective of this deliverable is to establish a Data Management Plan (DMP) for the CYBECO project compliant with the *Horizon 2020 FAIR Data Management Principles*. These principles establish that, in general terms, the project's research data should be findable, accessible, interoperable and reusable. The partners will bring specific datasets towards the project, and there will be many more data exchanges during the actual project. It is key that this data is treated correctly, to prevent the leaking of IP or commercially sensitive information from one of the partners by another partner.

This plan is based on the *Guidelines on FAIR Data Management in Horizon 2020* [1] and its annex the *Horizon 2020 FAIR DMP template* [1]. It further details the data management contents of the *CYBECO Proposal* [2]. The DMP is intended to be a living document, and it will be updated with further detail as the project progresses and when significant changes occur. Therefore, it will have several versions and includes an update procedure.

The current report is the first version (i.e., DMP v.1), which is also deliverable D2.2 that shall be submitted to the European Commission in the third month of the project (M3).

1.2 Document Structure

The document is structured as follows:

- Sect. 2 describes the procedure for implementing and updating the DMP.
- Sect. 3 presents the general aspects of the DMP, covering how the project as a whole would manage repositories, open access, data security and data from research with human participants.
- Sect. 4 synthesises the adherence of CYBECO datasets to the FAIR principles on making the research data findable, accessible, interoperable and reusable. It also discusses the adherence to other supporting principles, namely, resource allocation, data security and ethical aspects. The specific information for the different datasets is provided in their Dataset Record, in the annex.
- Sect. 5 provides a list with the different datasets of the project. Additionally, the data management information for each dataset is detailed in its Dataset Record (provided as annexes of this document). The current version of the DMP, the initial plan, only details one dataset: the internal repository that will centralise the project documents and most of its datasets.

2 Implementation and update procedure

The DMP plan is implemented by evaluating the different datasets created by the project regarding the FAIR principles, as well as an evaluation of the overall data management within the project.

Specifically, the implementation of the DMP will consist of the following steps:

1. Creation of a ***Data Repository*** in a private part of the CYBECO website. Unless there would be a possible conflict with confidentiality, security or commercial sensitivity, all data needed to validate the results as presented in any of the publications will be made available through an ***open research data repository*** in the CYBECO website as soon as possible. The URL for the website is www.cybeco.eu, whereas the URL to access the private part is www.cybeco.eu/private-area/repository.

2. Each partner should fill a ***Dataset Record*** for each of the datasets they create. We define a dataset as any collection of research data that is particular or special from the data management perspective. This means that data about different topics might be grouped in a dataset if no particular aspect makes its management different (e.g., confidentiality, security, intellectual property). Annex 1 provides a template of the Dataset Record.

3. Once the Dataset Record is filled, each partner should store them in the Dataset Repository alongside the actual dataset.

4. Some datasets might have specific data management policies or procedures (e.g., the experiments). If possible, each partner should upload those policies and procedures too.

5. On a regular basis, CSIC, with the support of TREK, will review these records to update the DMP accordingly and ask for additional feedback to the partners. As a minimum, the DMP should be updated in the context of the periodic evaluation/assessment of the project. If there are no other periodic reviews envisaged within the grant agreement, an update needs to be made in time for the final review at the latest.

Additionally, the consortium will agree and specify, in the next project meeting (second half of 2017), the following data management aspects:

- How many years the data will be preserved after the end of the project and how the data will be curated for that long-term preservation.

- Identification of the relevant datasets that the project will generate. With special emphasis on protection (e.g., privacy and intellectual property) and public sharing (for both scientific and general usage).

3 Summary of the Data Management Plan

The DMP details the following aspects: datasets, standards and metadata, data sharing, identification of repositories, long-term preservation and associated costs. The project datasets will be kept in a repository in a private part of the CYBECO website (hereafter the CYBECO Repository), which will allow for all data to be identifiable, but also include information about whether this data is commercially sensitive and so on to allow for proper sharing of the information amongst the partners and the public at large. This central repository will ensure long-term preservation of the data, and will also be secured using relevant methods for access protection and backup.

The main reference for the DMP are the *Guidelines on FAIR Data Management in Horizon 2020* [1] and its annex the *Horizon 2020 FAIR DMP template* [1]. These provide a sufficient high-level procedure for data management.

However, the approach of the plan is bottom-up: each dataset will be evaluated and prepared based on its security, privacy, technical and dissemination needs. Those datasets without critical aspects will follow the H2020's FAIR guidelines. However, some datasets would have additional specific data management procedures. Two of these specificities are privacy and ethical aspects. Datasets generated from research with human participants will follow stringent procedures as specified in Sect. 3.2. Another factor that makes more convenient an individual data management procedure is that each domain has different publication or metadata standards or procedures. Thus, for maximising the access and utility of our public datasets, it is important to follow these domain-specific procedures.

3.1 Open access to research data

Unless there would be a possible conflict with confidentiality, security or commercial sensitivity, all data needed to validate the results as presented in any of the publications will be made available through an open research data repository in the CYBECO website as soon as possible. Likewise, other elements, such as software tools or equipment, will also be provided in the same repository.

Any sensitive data may be masked and made anonymous to protect the sensitivity of data, while still allowing this to be used by other projects in the future. Such sensitive data could also fall under an embargo period, the length of which will be determined by the potential commercial development based on this data, such as e.g. IP protection.

Some data may not be made available at all to the public due to its commercially sensitive or security nature, to ensure that the project delivers long-term profitable development for the commercial partners. The same applies to IP brought into and developed during the

project. If during the project certain IP would restrict the intended availability of some of the outputs, then a sample code approach will be used to overcome this problem. Such sample code, as e.g. also used in the standardisation of the MPEG format, allows for a functional model to be presented, while the freely available code would not contain all possible optimisations. Hence, commercially and security sensitive information can be retained and secured accordingly while the open source tools would still be functional.

Following the dissemination plan of CYBECO (D8.1), datasets associated with scientific publications are especially relevant. Peer-reviewed scientific publication must be made openly available and free of charge, online, for any user. Therefore, datasets and tools needed for make the paper reproducible will be provided.

3.2 Data management of research with human participants

As declared in the ethical self-assessment, we shall perform research with human participants, specifically with volunteers for social and human sciences research and that personal data collection/processing will be involved. These activities will be performed by DEVSTAT and UNN, which have experience in performing this type of research with the highest ethical standards. Their research protocols will fully comply with the principles of the *Declaration of Helsinki (1989)*, the *Universal Declaration of Human Rights (UNESCO, 1948)* and the *Agreement for the Human Rights Protection in Biology and Biomedicine (Oviedo, 1997)*, and the CYBECO charter of ethics for experiments [2].

Behavioural economic studies

Several considerations will be made to minimise confidentiality issues with the participants in the behavioural economics studies. First, the amount of personal information required will be limited to the absolute minimum. Second, personal information will be collected without unique identifiers attached to the data, or known to the researcher. Although consent forms will include the participant name, these personal identifiers will not be linked to the recorded data. Third, each participant data will be associated with an alphanumeric code to remain anonymous in all stages of the research protocol. The identifying list will be stored in a safe and separate area from the study data.

Security measures for storage and handling of subject data will be carefully considered: experimental data will be originally recorded in a computer without Internet access and with restricted access to researchers involved in the project. Access passwords are necessary to log in the experimental computerised setup. Participant recordings will be removed from the computer when the experiment has finished with each participant to increment the security of this data set.

Psychology-led studies

Similar considerations will be in place for the psychology-led studies, which adheres, in addition, to the ethical code of practice as laid down by the *British Psychological Society*. Although the stored interview data may hold information unique to particular participants and so strict consent protocols will be devised around the storage and use of this data and the right to withdraw. Participants will be free to withdraw from the experiment at any time. It will be made clear that participation is voluntary, and that deciding to quit from the study is affecting only the amount of payment, but a minimum wage is guaranteed at the end of the first session.

3.3 Data management security

Information will be handled as confidentially as possible in accordance to applicable national regulations on data protection, to the *Directive 95/46/EC* of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data (OJ 23 November 1995, No L. 281 pp. 0031-0050) and to the *Directive 2001/20/EC* of the European Parliament and of the Council of 4 April 2001 on the approximation of the laws.

Because of its very own nature, security is a key issue within CYBECO. As stated, the Executive Board of the Project will act as well as Security Scrutiny Committee, identifying issues that should remain at confidential level. In particular, some details in connection with the experiments and the product will remain at such level for security reasons, as stated in the work plan. Besides, the CSIC team includes a specialist in data protection, J.A. Rubio, who will take care of data protection issues.

4 Adherence to the FAIR principles

This section synthesizes the adherence of the CYBECO datasets to the FAIR principles on making the research data findable, accessible, interoperable and reusable. The specific information for the different datasets is provided in their Dataset Record, in the annexes.

This section will evolve as the CYBECO project grows.

4.1 Making data findable

- CYBECO will create a repository for the project partners and an open research data repository for the public.
- Datasets and documents will contain version numbers, metadata and keywords for identification. Internally, following the structure of the project organisation. The public available datasets will, additionally, use identifiers such as DOI and metadata that facilitates a clear identification and citation by external users.

4.2 Making data openly accessible

- All data needed to validate the results of CYBECO will be made openly available unless there would be a possible conflict with confidentiality, security or commercial or intellectual property aspects.
- Sensitive datasets may be masked, made anonymous, or presented as a sample to protect the sensitivity of data, while still allowing this to be used by the public.

4.3 Making data interoperable

- The public available datasets will follow standardized formats that facilitate their interoperability and reusability. First, by using highly interoperable formats such as .sql, .csv, or .xml. Second, by producing “tidy data” [3] so that the datasets are easy to edit and visualize.

4.4 Making data reusable

- Sensitive data may fall under an embargo period for determining whether and how this data will be made public.

4.5 Additional aspects of data management

Resource allocation

- Long-term preservation of the CYBECO website and, thus, the repositories.
- Preservation of back-ups of the datasets.

Data security

- Datasets in the CYBECO repository will include information about whether the data is sensitive and the type of sensitive information (e.g., personal data, intellectual property, commercial).
- Website hosted in European servers.
- Use of secure methods for access and backups of the CYBECO repositories.
- The CSIC team includes a specialist in data protection, J.A. Rubio, who will take care of data protection issues.

Ethical aspects

- Following the ethical self-assessment, we have declared that we shall perform research with human participants and, thus, personal data collection and processing.
- The data management of research with human participants will be performed by DEVSTAT and UNN, which have experience in performing this type of research with the highest ethical standards. Further details of the data management of this research is provided in Sect. 3.2.

5 Datasets

The initial DMP identifies one special dataset: the internal repository. During the life of the project this list will grow to include other datasets. Each of the dataset is further detailed in its correspondent annex.

Dataset	Description	Annex
Internal repository	Central repository of all the datasets that can be shared between the project consortium (CO), restricted to include additional third parties (PP, RE) and to the public (PU).	2

6 Conclusion

This document presented the deliverable D2.2 with the initial version Data Management Plan (DMP v.1), to be submitted to the European Commission in the third month of the project. It presented the general data management plan that the CYBECO Consortium established for the project and the updating procedure so that the DMP is updated throughout the project.

Further versions of the DMP will involve the creation of Dataset Registers for each of the datasets created by the project and the refinement and updating of this document in accordance with the progress of the project.

7 References

[1] European Commission, *Guidelines on FAIR Data Management in Horizon 2020, Version 3.0,* 2016.

[2] CYBECO Consortium, *Proposal 740920 'CYBECO' for the Horizon 2020 Program,* 2016.

[3] H. Wickham, “Tidy Data,” *Journal of Statistical Software*, vol. 59, no. 10, pp. 1 -- 23, 2014.

8 Acronyms and Abbreviations

CYBECO	Acronym for the H2020 Project 'Supporting Cyberinsurance from a Behavioural Choice Perspective'.
DMP	Data Management Plan
H2020	Horizon 2020 Research and Innovation Program
UNN	University of Northumbria at Newcastle

Annex 1: Dataset record template

Data Summary	
Name of dataset	**A short and descriptive name for the dataset**
Work Package(s)	Work packages for which the dataset is created
Task(s)	Tasks for which the dataset is created (if can be detailed)
Nature of the data	Data collection / Data generation
Purpose of the data	What is the purpose of the dataset?
Relation to the project	What is the relation of the dataset to the objectives of the project?
Type of data	What type of data contains the dataset?
Data format	What is the format of the dataset?
Reuse in project	Will be the dataset reused in other parts of the project?
Reuse outside the project	Will be the dataset reused outside the project?
Origin of the data	What is the origin of the data?
Expected size of dataset	What is the expected size of the data in observations/instances? And in storage size (Mb, few Gb, dozens of GB, …)?
Data utility	To whom might it be useful?
Making data findable	
Discoverability	Are the data discoverable with metadata?
Identifiability	Are the data identifiable and locatable using a standard identification mechanism (e.g. persistent and unique identifiers such as DOI)?
Naming	Does the data follow naming conventions? What conventions do you follow?
Keywords	Will search keywords be provided that optimise possibilities for re-use?
Versions	Do you provide clear version numbers?
Metadata	What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.
Making data accessible	
Availability	Will the dataset be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why clearly separating legal and contractual reasons from voluntary restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out.
Accessibility	How will the dataset be made accessible (e.g. by deposition in a repository)?

Repository	Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible. Have you explored appropriate arrangements with the identified repository?
Access support	What methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?
Access policy	How will access be provided in case there are any restrictions? If there are restrictions on use, how will access be provided? Is there a need for a data access committee? Are there well-described conditions for access (i.e. a machine-readable license)? How will the identity of the person accessing the data be ascertained?
Making data interoperable	
Interoperability	Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)? What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?
Standardisation	Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability? In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?
Making data reusable	
License	How will the data be licensed to permit the widest re-use possible?
Timing	When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.
Third-parties	Are the data produced and used in the project usable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.
Length of time	How long is it intended that the data remains re-usable?
Quality	Are data quality assurance processes described?
Allocation of resources	
Costs	What are the costs for making data FAIR in your project? How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).
Responsible(s)	Who will be responsible for data management in your project?
Long-term preservation	Are the resources for long-term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?
Data security	
Provisions	What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

Certified repositories	Is the data safely stored in certified repositories for long-term preservation and curation?
Ethical aspects	
Ethical and legal issues	Are there any ethical or legal issues that can have an impact on data sharing? These can be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action.
Informed consent	Is informed consent for data sharing and long-term preservation included in questionnaires dealing with personal data?
Other procedures for data management	
Other data management procedures	Do you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones? Provide that procedure if possible.

Annex 2: Internal Repository

Data Summary	
Name of dataset	**Internal repository** (www.cybeco.eu/private-area/repository)
Work Package(s)	All (managed by WP1)
Task(s)	N/A
Nature of the data	Data collection
Purpose of the data	Central repository for all the datasets that can be shared between the project consortium (CO), with additional third parties (PP, RE) and with the public (PU).
Relation to the project	Facilitate the generation and sharing of documents and datasets.
Type of data	Files
Data format	Folder-based
Reuse in project	It encompasses all the project
Reuse outside the project	As a repository, it is only dedicated to the CYBECO project.
Origin of the data	Project partners
Expected size of dataset	Dozens of GB
Data utility	The repository is necessary to the project partners.
Making data findable	
Discoverability	Yes: titles, descriptions, changelog, authors and meta-tags.
Identifiability	Yes: as above.
Naming	Yes: Project tasks, deliverables and versions of documents.
Keywords	Yes
Versions	Twice: Version number in the deliverables, and the different versions that the repository saves when a user updates a file.
Metadata	Project deliverable, version of document, keywords, etc.
Making data accessible	
Availability	As internal repository, just to the project partners. We will make particular documents available to the public through other means (e.g., website).
Accessibility	Deposition in a repository for internal partners.
Repository	The repository has been designed, and is managed, by TREK. It is based on well-established technologies (Joomla and jDownloads).
Access support	The access is through the web. The repository platform allows users to create a synchronised local folder.

Access policy	The access requires login. The repository platform allows the implementation of additional access measures (e.g., captcha, access rights, additional passwords).
Making data interoperable	
Interoperability	As internal reposotory, interoperability is based on the usage of common file types (e.g., pdf, doc, xml, sql, xls, tex).
Standardisation	There are no explicit vocabulary or ontology schemes.
Making data reusable	
License	The license for deliverables will be the specified in the project proposal. For other documents and datasets, the license established by their creators.
Timing	As internal repository it will last the life of the project plus the years agreed by the consortium for its persistence.
Third-parties	As internal repository, no. We will make available particular datasets and documents through other means (e.g., website).
Length of time	To be agreed by the project consortium.
Quality	As repository, the quality policy involves keeping the repository clean from irrelevant or duplicated documents, maintaining a clear and consistent identification and naming of files and ensuring a fair use of the space.
Allocation of resources	
Costs	Persistence around 150-250 EUR per year.
Responsible(s)	CSIC is responsible for the data management plan. TREK responsible for the repository.
Long-term preservation	The cost will depend on the length of the preservation, which has to be agreed by the project consortium.
Data security	
Provisions	Repository has backups and security measures (e.g., user groups, captchas, changelogs, logins).
Certified repositories	To be discused by the consortium based on the agreed length of preservation.
Ethical aspects	
Ethical and legal issues	Yes. As central repository, the considerations specified in Sect. 3 of the DMP for the project might apply in some aspects to contents stored in the repository.
Informed consent	To be specified in the datasets from research with human participants.
Other procedures for data management	
Other data management procedures	As repository, no.