

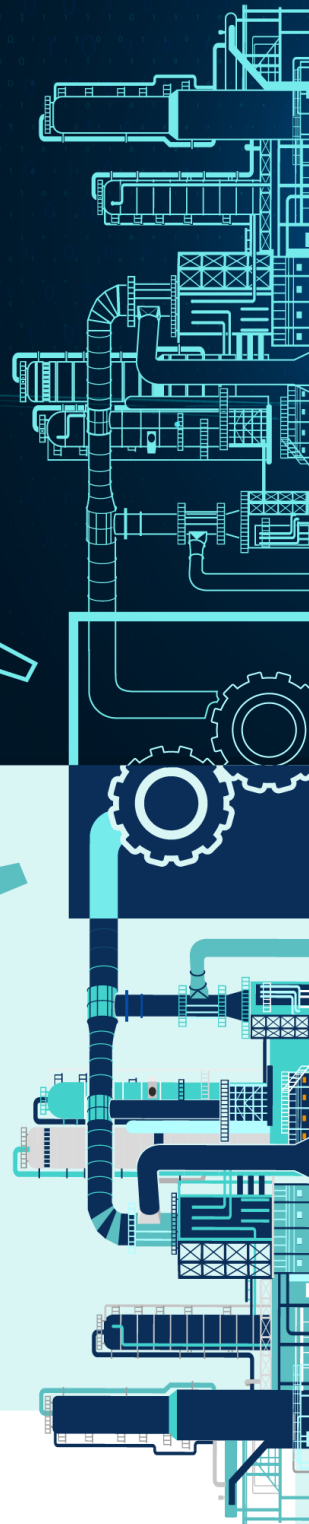


FACTLOG
www.factlog.eu

Ref. Ares(2020)2303147-30/04/2020



ENERGY-AWARE FACTORY ANALYTICS PROCESS FOR INDUSTRY



Deliverable D9.3

Initial Data Management Plan

Version
1.0

Lead Partner
MAG

Date
30/04/2020

Project Name
FACTLOG – Energy-aware Factory Analytics for Process Industries



FACTLOG has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 869951

Call Identifier H2020-NMBP-SPIRE-2019	Topic DT-SPIRE-06-2019 - Digital technologies for improved performance in cognitive production plants
Project Reference 869951	Start date November 1 st , 2019
Type of Action IA – Innovation Action	Duration 42 Months

Dissemination Level

X	PU	Public
	CO	Confidential, restricted under conditions set out in the Grant Agreement
	CI	Classified, information as referred in the Commission Decision 2001/844/EC

Disclaimer

This document reflects the opinion of the authors only.

While the information contained herein is believed to be accurate, neither the FACTLOG consortium as a whole, nor any of its members, their officers, employees or agents make no warranty that this material is capable of use, or that use of the information is free from risk and accept no liability for loss or damage suffered by any person in respect of any inaccuracy or omission.

This document contains information, which is the copyright of FACTLOG consortium, and may not be copied, reproduced, stored in a retrieval system or transmitted, in any form or by any means, in whole or in part, without written permission. The commercial use of any information contained in this document may require a license from the proprietor of that information. The document must be referenced if used in a publication.

Executive Summary

This deliverable provides the first version of the *Data Management Plan (DMP)* of the FACTLOG project. It describes what kind of data will be generated or collected during the project implementation and how these data are then managed and published.

Such information could be the scientific publications issued, white papers published, Open Source code generated, mock-up datasets used for supporting the development process etc. The list of research data expected during the project consists of open source software components, original research data and anonymous statistics. These datasets are expected to be collected during the validation and evaluation phase and are therefore subject to change, considering also the definition of the FACTLOG business models and sustainability plans.

The publishing platforms used are the project website, OwnCloud platform, Zenodo for long-term archiving (as suggested by the EC), and GitLab for open-source code. All these platforms can be accessed openly.

Revision History

Revision	Date	Description	Organisation
0.1	02/03/2020	ToC and first version	MAG
0.2	03/04/2020	Second version with some additions	MAG
0.3	27/04/2020	Contributions from partners	ALL
1.0	30/04/2020	Final version	MAG

Contributors

Organisation	Author	E-Mail
MAG	Armend Duzha	armend.duzha@maggioli.it
MAG	Kostas Kalaboukas	kostas.kalaboukas@maggioli.it
AUEB	Stavros Lounis	slounis@aueb.gr
UNIPi	Pavlos Eirinakis	pavlose@unipi.gr
EPFL	Jinzhi Lu	jinzhi.lu@epfl.ch
Qlector	Jože Rožanec	jose.rozanec@qlector.com
Qlector	Klemen Kenda	klemen.kenda@qlector.com
JSI	Mitja Jermol	mitja.jermol@ijs.si
JSI	Aljaž Košmerlj	aljaz.kosmerlj@ijs.si
JEMS	Johannes Kmrc	johannes@jems.eco
TUPRAS	Melike Onat	melike.onat@tupras.com.tr
BRC	Alexander Adams	alexander.adams@brc.ltd.uk
PIA	Alessandro Canepa	alessandro.canepa@piacenza1733.it
CONT	Alin Popa	alin.3.popa@continental-corporation.com
UNP	Tiago Teixeira	tiago.teixeira@unparallel.pt

Table of Contents

1	Introduction	7
1.1	Purpose and Scope	7
1.2	Structure of the Document.....	7
2	Data Management Strategy	8
2.1	Data Management Process	8
2.2	Roles and Responsibilities	9
3	Data Archiving and Preserving Infrastructure	10
3.1	Project Website	10
3.2	OwnCloud.....	10
3.3	Zenodo	11
3.4	GitLab.....	11
4	Data Summary	12
4.1	Public deliverables.....	12
4.2	Research datasets.....	13
4.3	Open source software components	14
4.4	Scientific publications	15
5	FAIR Data	16
5.1	Making Data Findable, including provisions for metadata	16
5.2	Making Data Openly Accessible	16
5.3	Making Data Interoperable	17
5.4	Increase Data Re-Use (Though Clarifying Licenses).....	18
6	Allocation of Resources	19
7	Data Security	20
8	Ethical Aspects	21
	References	22
	Appendix I – FAIR Data Questionnaire	23

List of Tables

Table 1 – List of FACTLOG public deliverables	12
Table 2 – Datasets to be generated during FACTLOG lifetime.....	14
Table 3 – List of FACTLOG Open Source software components	14
Table 4 – Dataset Metadata.....	17

1 Introduction

1.1 Purpose and Scope

The purpose of this document is to describe how data will be handled during the project duration. For all data generated or collected during the FACTLOG project a description will be provided including the source of the data, the standards and metadata used for data preservation and maintenance, as well as the process of how this data will be exploited and/or shared/made accessible for verification and re-use, in accordance with the EC Guidelines on FAIR Data Management in Horizon 2020 [1].

This deliverable is 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 and when significant changes occur. This is the 1st version of the document which is going to be updated on a regular basis and the final version will be submitted in month M42.

1.2 Structure of the Document

The rest of the document is structured as follows:

- Section 2 defines the data management process that will be applied to all data collected or generated during the project, as well as the tools to be used in order to ensure that the data will be exploited and/or shared/made accessible for verification and re-use along with the data preservation and maintenance processes.
- Section 3 provides a summary of the datasets that will be generated during the project, including information such as types and formats, expected size of the datasets and data utility.
- Section 4 outlines how the research data will become FAIR.
- Section 5 describes the allocation of resources required to make data FAIR.
- In Section 6, the provisions regarding the curation and preservation of the data during and after the end of the project are provided.
- Section 7 presents the ethical principles that the partners should comply.

2 Data Management Strategy

2.1 Data Management Process

The FACTLOG's Data Management Process is defined as a step-wise approach for each result generated or collected during the project runtime. The following questions must be answered to classify the different datasets:

1. Does a result provide significant value to others or is it necessary to understand a scientific conclusion?

If this question is answered with yes, then the result is classified as public (granted for open access). If this question is answered with no, the result is classified as non-public. Such a result could be code that is very specific to FACTLOG platform (e.g. a database initialization) which is usually of no scientific interest to anyone, nor does it add any significant contribution.

2. Does a result include personal information that is not the author's name?

If this question is answered with yes, the result is classified as non-public. Personal information beyond the name must be removed if it should be published according to the ethical principles of the project.

3. Does a result allow the identification of individuals even without the name?

If this question is answered with yes, the result is classified as non-public. As such, in order to make it publishable, the included information must be reduced to a level where single individuals cannot be identified. This can be performed by using established anonymization techniques to conceal a single user's identity, e.g. abstraction, dummy users, or non-intersecting features.

4. Can a result be abused for a purpose that is undesired by society in general or contradict with societal norms and FACTLOG's ethics?

If this question is answered with yes, the result is classified as non-public.

5. Does a result include business or trade secrets of one or more partners of FACTLOG?

If this question is answered with yes, the result is classified as non-public, except if the opposite is explicitly stated by the involved partners. Business or trade secrets need to be removed in accordance to all partners' requirements before it can be published.

6. Does a result name technology that are part of an ongoing project-related patent application?

If this question is answered with yes, then the result is classified as non-public. Of course, results can be published after patent has been filed.

7. Does a result break national security interests for any project partner?

If this question is answered with yes, the result is classified as non-public.

2.2 Roles and Responsibilities

The Data Management Plan aims to identify the project outputs to be disseminated as well as to decide on the way and means of their Open Access (if applicable). To ensure it, a dedicated time slot will be reserved at each of the project plenary meetings. The EC will be informed about related work done and publications provided in the management reports.

Individual responsibilities on data management in the consortium are:

- **Data Management Plan Leader** (MAG) – to prepare and lead related discussions at the relevant project meetings and to maintain the channels for dissemination of project outcomes.
- **Scientific Manager** (AUEB) – to identify data collected by the project and technical project outcomes eventually suitable for publication
- **Dissemination Manager** (UNP) – to identify publications suitable for publication in the considered repositories and maintain FACTLOG inputs for the Open Access
- **All partners** – to identify own project results suitable for publication.

Moreover, each partner has to respect the policies set out in this DMP. Datasets have to be created, managed and stored appropriately and in line with applicable legislation. Validation and registration of datasets and metadata is the responsibility of the partner that generates the data in the WP. Metadata constitutes an underlying definition or description of the datasets and facilitate finding and working with particular instances of data.

Backing up data for sharing through Open Access repositories is the responsibility of the partner possessing the data.

Quality control of these data is the responsibility of the relevant WP leader, supported by the Data Manager.

If datasets are updated, the partner that possesses the data has the responsibility to manage the different versions and to make sure that the latest version is available in the case of publicly available data.

Last but not least, all partners must consult the concerned partner(s) before publishing data in the open domain that can be associated to an exploitable result.

By taking into account the methodological framework, we proceed with the early definition of data sets to be considered as part of the DMP in the following section.

3 Data Archiving and Preserving Infrastructure

The Data Archiving and Preserving infrastructure consists of web-based platforms that provide long-term open access to all generated or collected results of the project. The following list presents the platforms to be used during the project and describes their concepts for publishing, storage, and backup.

3.1 Project Website

FACTLOG has designed and setup a project webpage. It describes the objectives and the general approach of FACTLOG and will update target audience on its development status. A dedicated section for downloads is used to publish reports and white papers as well as scientific publications (in pre-camera ready form, or through links to the publisher's websites in case these are not open access). All documents are published using the portable document format (PDF). All downloads are enriched by using simple metadata information, such as the title and the type of the document. All information on the project website can be accessed without creating an account. The webpage is backed up once per month.

The FACTLOG website will be available during the project runtime, and will still be available for at least two year after the official project end.

URL: <http://www.factlog.eu/>

3.2 OwnCloud

OwnCloud is a suite of client-server software for creating file hosting services and using them. OwnCloud is functionally very similar to the widely-used Dropbox, with the primary functional difference being that OwnCloud is free and open-source, and thereby allowing anyone to install and operate it without charge on a private server, with no limits on storage space (except for disk capacity or account quota) or the number of connected clients.

In order for desktop machines to synchronize files with their OwnCloud server, desktop clients are available for PCs running Windows, OS X, FreeBSD or Linux. Mobile clients exist for iOS and Android devices. Files and other data (such as calendars, contacts or bookmarks) can also be accessed using a web browser without any additional software. Any updates to files are pushed between all computers or mobile devices connected to a user's account.

The OwnCloud platform for FACTLOG is hosted by MAG and runs on a server at the company premises in Italy, therefore keeping all data on an own managed server. The platform is securely backed in the MAG system infrastructure and holds all project-related data, including deliverables and publications. The FACTLOG OwnCloud platform will not duplicate any project-related information to external servers, such as issues, requirements, product code, or deployment information. The OwnCloud platform will be available during the whole project duration and for at least one year after the project end.

URL: <https://docbox.maggioli.it/>

3.3 Zenodo

Zenodo is an open repository which helps researchers to share research results in a wide variety of formats for all scientific disciplines. It was developed by the OpenAIRE+ project and is maintained by CERN using one of Europe's most reliably hardware infrastructures.

Zenodo not only supports the publication of scientific papers or white papers, but also the publication of any structured research data (e.g. using XML). Zenodo provides a connector to GitHub that supports open collaboration for source code and versioning for all kinds of data. All uploaded results are structured by using metadata, like for example the contributors' names, keywords, date, location, kind of document, license, and others. Considering the language of textual metadata items, English is preferred. All metadata is licensed under CC license (Creative Commons 'No Rights Reserved'). The property rights or ownership of a result does not change by uploading it to Zenodo.

All public results generated or collected during the FACTLOG project will be uploaded to Zenodo for long-term storage and open access.

URL: <https://zenodo.org/communities/factlog/>

3.4 GitLab

GitLab is a well-established online repository which supports distributed source code development, management, and revision control. It is primarily used for source code data. It enables world-wide collaboration between developers and provides also some facilities to work on documentation and to track issues. GitLab provides paid and free service plans. Free service plans can have any number of publics, open-access repositories with unlimited collaborators. Private, non-public repositories require a paid service plan. Many open-source projects use GitLab to share their results for free. The platform uses metadata like contributors' nicknames, keywords, time, and data file types to structure the projects and their results. The service is hosted by MAG in Italy.

All source-code components that are implemented during the project and decided to be public will be uploaded to an open access GitLab repository.

URL: <http://git.maggioli.it/>

4 Data Summary

In this section, a list of all existing or foreseeable results for dissemination is presented, separated into public deliverables, publications and open research data. For each result and in accordance to the FAIR data management guideline [2] we provide a description, name the standards used for storage and metadata (to make data findable and interoperable), and define which open access platform is chosen.

4.1 Public deliverables

The following table presents the list of public deliverables of the AEGIS project.

Table 1 – List of FACTLOG public deliverables

ID	Title	Due date
D1.1	Reference Scenarios, KPIs and Datasets	31 July 2020
D1.2	Cognitive Factory Framework	31 October 2020
D2.1	Analytics System Requirements and Design Specification	31 August 2020
D2.2	Analytical Platform for Process Industry	31 October 2021
D2.4	Anomaly detection System	30 June 2021
D3.3	Factory Knowledge for Cognition	30 June 2020
D4.1	Process Modelling Methodology	31 May 2021
D4.2	Knowledge Graph Modelling	28 February 2022
D4.3	Systemic Cognitive Models	30 June 2022
D4.4	KG-based Analytics for Process Optimisation	30 June 2022
D5.1	Real-Time Re-Optimisation Algorithms	30 June 2021
D5.2	Robust and Energy-Aware Planning and Scheduling	30 November 2021
D5.3	FACTLOG Optimization Toolkit and Service	31 August 2022
D6.1	Data Collection Framework (Interim Version)	31 January 2021
D6.2	Data Collection Framework (Final Version)	30 April 2022

ID	Title	Due date
D6.3	Integration Guidelines (Interim Version)	31 Mar 2021
D6.4	Integration Guidelines (Final Version)	31 Aug 2022
D6.5	Integrated Package and Platform (Interim Version)	31 Mar 2021
D6.6	Integrated Package and Platform (Final Version)	31 Aug 2022
D7.3	Cognitive Transformation Report (Interim Version)	30 Sep 2021
D7.4	Cognitive Transformation Report (Final Version)	28 Feb 2023
D7.5	Validation and Impact Assessment Report (Interim Version)	31 Oct 2021
D7.6	Validation and Impact Assessment Report (Final Version)	30 Apr 2023
D8.1	Project Website	31 January 2020
D8.3	Dissemination and Communication Activities Report V1	31 Oct 2020
D8.4	Standardization Activities Report V1	31 Oct 2020
D8.6	Dissemination and Communication Activities Report V2	31 Oct 2021
D8.7	Standardization Activities Report V2	31 Oct 2021
D8.9	FACTLOG Training and Educational Material V1	31 Oct 2021
D8.10	Dissemination and Communication Activities Report V3	30 Apr 2023
D8.11	Standardization Activities Report V3	30 Apr 2023
D8.13	FACTLOG Training and Educational Material V2	30 Apr 2023

4.2 Research datasets

During the project lifetime several datasets from various consortium members, representing different domains, will be produced. The datasets to be generated are listed in Table 2.

As the project evolves, this table will be updated with additions or modifications of datasets. The final table will be presented in the final version of the DMP.

Table 2 – Datasets to be generated during FACTLOG lifetime

Name of the Dataset	Responsible Partner	Accessibility
SynDi_Input	JEMS	Confidential
SynDi_Operation	JEMS	Confidential
SynDi_Output	JEMS	Confidential
TUPRAS_data	TUPRAS	Confidential
BRC_data	BRC	Confidential
PIACENZA_data	PIA	Confidential
CONTINENTAL_data	CONT	Confidential

As we can see in Table 2, the aforementioned datasets are divided into two categories regarding their accessibility, namely Confidential and Open Access. ‘Open Access’ datasets will be included in the Open Research Data Pilot (ORDP) and will be available for public sharing. On the other hand, ‘Confidential’ datasets will be only shared or become accessible after a proper agreement is signed.

4.3 Open source software components

The analysis is performed by taking into account the list of exploitable components, defined in the DOA as open source components. A summary is presented in the following table:

Table 3 – List of FACTLOG Open Source software components

ID	Title	Responsible partner(s)
S01	Analytics library for process industries	JSI
S02	Relational learning library for KG-based analytics	JSI
S03	Optimization-as-a-Service Framework	MAG, AUEB, UNIPI
S04	Messaging and Service Bus	MAG

While FACTLOG project is committed to submit most of the software components as Open Source, the final decision will be made along with the definition of the IPR management schema in WP8.

4.4 Scientific publications

In this section, we are going to include all scientific publications produced during the project duration. There are already some publications submitted by the consortium partners in scientific journals and conferences:

- Jinzhi Lu, Xiaochen Zheng, Ali Gharaei, Kostas Kalaboukas, Dimitris Kiritsis (2020). “Cognitive twins for supporting decision-makings of internet of things systems”. Proceeding of 5th International Conference on the Industry 4.0 Model for Advanced Manufacturing.
- Jože M. Rožanec, Lu Jinzhi, Aljaž Košmerlj, Klemen Kenda, Kiritsis Dimitris, Viktor Jovanoski, Jan Rupnik, Mario Karlovcec, and Blaz Fortuna (2020). “Towards Actionable Cognitive Digital Twins for Manufacturing”, Proceeding of ESWC 2020 conference.
- Eirinakis, Pavlos, Kalaboukas, Kostas, Lounis, Stavros, Mourtos Ioannis, Rožanec Jože M., Stojanovic Nenad, Zois Georgios (2020). “Enhancing Cognition for Digital Twins”. Proceeding of IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)
- Calefato C., Vallini M., Fasana S. (2020). “The cognitive retrofit revolution inside industry: how to modernize factories by Cognitive Digital Twins”. XII Congresso Nazionale SIE2020 (XII National Congress SIE 2020 – Italian Association of Ergonomics). Waiting for notification.

5 FAIR Data

FAIR data principles [2] apply to available for public use datasets. The rest of the datasets are considered as confidential due to internal regulations and/or legal reasons that data providers ought to comply with. Confidential datasets will be either shared within the consortium or become accessible in-house after a proper agreement is signed. The data management policy that will be followed for all the project datasets is described in the following sections.

5.1 Making Data Findable, including provisions for metadata

The standardized naming convention for all the project datasets will include the following information:

1. A unique chronological number of the datasets in the project.
2. The name of the dataset.
3. The acronym of the project.
4. The number of the related work package and deliverable/task.
5. A version number for each new version of the dataset that will be incremental at each revision.

Dataset's internal reference number:

01_NameOfTheDataset_FACTLOG_WPX_DX.X_v1.0.xls

DOIs can be used for each dataset that is uploaded to an online data repository to achieve effective and persistent citation. Each dataset's DOI should be used to all the related publications so that readers will be able to link them with the underlying datasets.

All the project datasets must be described with metadata. Metadata is a set of data that provides context or additional information about other data. Metadata gives the ability to other researches to find data in an online repository which increases the reusability of the dataset. By providing detailed and rich metadata, researchers can define easier if the corresponding dataset is relevant to their research. An additional use of these metadata except from data ingestion and data reusability is the provision of a concrete view of the data being used in the project, independently from the accessibility restrictions.

Moreover, search keywords describing the dataset or content of the data will be provided when a dataset is uploaded to a repository aiming at optimising possibilities of re-use.

5.2 Making Data Openly Accessible

FACTLOG consortium will ensure open access to all peer-reviewed scientific publications relating to its results. The decision on whether to publish through open access will have to account for the potential necessity for protection of sensitive data. The consortium will follow 'green' open access in the publications.

The repositories that will be used for the project datasets have been decided by the consortium. Zenodo [3], an online repository, will be used for the generated open research data during the lifetime of the project. Zenodo can easily share datasets in various sizes and formats, provide flexible licensing, and access and re-use of research data.

Confidential datasets, on the other hand, cannot become public due to privacy restrictions that the partners ought to comply with. These datasets will be uploaded in the project SVN repository in order to be available for the consortium.

GitLab [4] is a version-control online repository which supports distributed source code development and management. FACTLOG's produced datasets could be built on the GitLab platform to raise awareness of the project, increase its impact and ensure its long-term sustainability. GitLab will also host parts of some of the FACTLOG open-source code components that will be implemented during the project.

Research data needed for validation of results presented in scientific publications will be uploaded to Zenodo as soon as possible. In case an embargo period should be applied before the publication of the results, data will be deposited in the project SVN repository. Nevertheless, information about the restricted data along with the related metadata will be published in Zenodo at the same time with the publication.

As mentioned before, some real datasets cannot be shared either with the consortium or publicly due to internal regulation constraints. Moreover, when possible, anonymization, aggregation, minimization or sampling techniques may be applied to real data during the project's implementation to guarantee the preservation of user's personal, sensitive data.

5.3 Making Data Interoperable

To increase the interoperability of the provided data, commonly used vocabularies for the metadata within the datasets will be used. Interoperability will be ensured by using the same standards for data and metadata capture/creation to all datasets.

Table 4 presents the general overview, the content and the technical description and the access of the data.

Table 4 – Dataset Metadata

Name	Name of the dataset
Creator / Responsible Partner	Name of the partner responsible for the data created
Data Identifier	Dataset's internal reference number
DOI	(if applicable)
Description	A brief description of the dataset
Work Package/Deliverable	Associated work package and deliverable
Source	How the data have been generated
Processing	How the data have been processed
Repository	The repository where the data will be uploaded

Language	All languages used in the dataset
Code List	Explanation of codes or abbreviations used
Type	Types of the data
Format	Formats of the data
Expected size	An approximation of the size of the dataset
Keywords	Keywords describing the content of the data
Version	Unique identifier for each version of the dataset
Date of Repository Submission	Release date (preferred format dd-mm-yyyy)
Necessary Software	Necessary software needed to create, view or analyse data
Access Information	Any rights information on the use of the data

5.4 Increase Data Re-Use (Though Clarifying Licenses)

Uploading the datasets to the Zenodo repository will help towards making data reusable. Datasets will be made available to third parties after their generation and will remain public for a considerable time period after the completion of the project. However, possible restrictions such as embargo periods or restrictions from editors of scientific journals and organizers of conferences will be examined on a case-by-case basis. The owner of the data, the beneficiary that generates it, is responsible for maintaining the data after the end of the project.

A Data Quality Assurance Process (DQAP) will be followed in order to ensure quality of the data that are generated during the lifetime of the project. The dimensions to be measured are: i) Validity: Data should be sufficiently accurate/valid for the intended use; ii) Reliability: Data collection methods must remain stable over time; iii) Precision: Data detail is sufficient to represent the phenomenon of interest; iv) Integrity: Accuracy and consistency of data must be maintained and assured; v) Timeliness: Data should be regularly collected, up-to-date and available when needed and vi) Completeness: All required data elements, records, and values must be known. Reproduction of research results is possible.

6 Allocation of Resources

Costs for making data FAIR are mainly related to personnel costs. Each of the project partners will need to prepare the data for publication, update and maintain data, perform data hosting and backup, data sharing, and security, etc. The costs for these actions will be covered by the project funds.

The costs related to long-term preservation of data after the end of the project are difficult to be estimated in this version of the Data Management Plan. The aim of the consortium is to preserve the data for a sufficient time period. We should mention that preserving datasets on the Zenodo repository where a single dataset file does not exceed 50 GB, is free of charge. Moreover, internal datasets of the project will be stored and preserved in cloud-based collaboration platform OwnCloud hosted by MAG.

7 Data Security

Datasets produced by the consortium members will be stored at the responsible partners' local private repositories, which are secure and non-accessible to the public. The appropriate procedures for recovery, secure storage and transfers are followed by the project partners and all of them are GDPR compliant.

Shared datasets among the consortium members will be stored in the project OwnCloud. OwnCloud is the most widespread open source software for cloud-based collaboration platforms that offers real data security and privacy. All datasets are stored in one central location – protected from unauthorized access. Clients and apps can be used for accessing data from anywhere.

The necessary data anonymity will be ensured. Proper pseudonymisation/anonymization techniques will be applied to guarantee the preservation of anonymity of the user personal, sensitive data. All personal data obtained within the project will be transmitted to partners within the consortium only after anonymization or pseudonymization techniques will be applied.

8 Ethical Aspects

One of the main ethical concerns is privacy and personal data protection of individuals regarding both research activities and potential implementation of research results. The FACTLOG consortium is fully aware of the risks for the individual's privacy and shall conduct an in-depth risk assessment placing data protection by design and by default inherent in the set-up of the project

FACTLOG is intended to ensure respect for the ethical principles and fundamental rights embedded in the regulatory framework of the European Union, including the Charter of Fundamental Rights of the European Union as well as the European Convention on Human Rights. The data processing activities within our project will be carried out in accordance with the updated privacy rules as specified in Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data and repealing Directive 95/46/EC (GDPR) and the ePrivacy Directive 2002/58/EC.

Moreover, the corresponding national data protection legislation should be taken under consideration, and all legal documents and certifications required for compliance with such legislation will be obtained. The Parties who provide or make available to any other Party shared information containing Personal Data must have: (i) the authority and/or the authorisation to disclose the aforementioned information; (ii) obtained appropriate informed consents from all the data subjects involved, or from any applicable institution and (iii) a confirmation that there is no restriction that would prevent any other Party from using the shared information. Data protection by design and by default will be at the core of the research and development work as well as the project outputs.

The personal data will be anonymized whenever possible and homomorphic encryption will be used for pseudo-anonymization. Identifiable data will be dissociated from the rest of the collected data in a separate database during data collection. Personal data will be processed under FACTLOG using a neutral code with the aim to render data non-attributable to any natural person.

Due to the way the research will be carried out, consortium partners making decisions about the collection and processing of personal data, in order to achieve the various goals and objectives of the project, will be deemed joint data controllers. As required by the GDPR, an agreement outlining the allocation of the obligations and responsibilities of the joint data controllers within FACTLOG, will be created. A key aspect of this arrangement is regular and effective communication between data controllers to ensure a consistent and effective approach.

FACTLOG implementation does not anticipate or plan any transfer of personal data to third parties. Relevant legal foundations, appropriate safeguards and compliance measures will be identified and implemented if data sharing is deemed necessary at a later stage. These safeguards will be documented in the final version of the Data Management Plan.

References

[1] Guidelines on Data Management in H2020, Version 3.0, 26 July 2016

http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[2] FAIR Data Principles (FORCE11):

<https://www.force11.org/group/fairgroup/fairprinciples>

[3] Zenodo: <https://zenodo.org>

[4] GitLab: <https://about.gitlab.com/>

Appendix I – FAIR Data Questionnaire

Partner Name	
Contact person (e-mail)	
Name of the dataset	
A – Data Summary	
What is the purpose of the data collection/generation and its relation to the objectives of the project? ⁱ	<i>Please Specify</i>
What type of data you will generate/collect? ⁱⁱ	<i>Please Specify</i>
What formats of data you will generate/collect? ⁱⁱⁱ	<i>Please Specify</i>
Will you re-use any existing data and how? ^{iv}	<input type="checkbox"/> No <input type="checkbox"/> Yes (please specify):
What is the origin of the data? ^v	<i>Please Specify</i>
What is the expected size of the data? ^{vi}	<i>Please Specify</i>
To whom might the data be useful? ^{vii}	<i>Please Specify</i>
Will you apply data aggregation, minimization and anonymization methods to the data? ^{viii}	<input type="checkbox"/> No <input type="checkbox"/> Yes (please specify):
B – FAIR Data Findable, Accessible, Interoperable, Reusable	
B1 - Making data Findable	
Are the data produced and/or used in the project discoverable with metadata?	<input type="checkbox"/> Yes <input type="checkbox"/> No
What metadata will be created?	<i>Please Specify</i>
Do metadata standards exist in your discipline? ^{ix}	<input type="checkbox"/> Yes (please specify): <input type="checkbox"/> No (please specify):
Will search keywords be provided that optimize possibilities for re-use? ^x	<input type="checkbox"/> Yes <input type="checkbox"/> No
B2 - Making data openly Accessible	
Accessibility ^{xi}	<input type="checkbox"/> Public data <input type="checkbox"/> Confidential data
If certain datasets cannot be shared (or need to be shared under restrictions), explain why ^{xii}	<i>Please Specify</i>
How will the data be made accessible ^{xiii}	<i>Please Specify</i>
Which repository will be used? ^{xiv}	<i>Please Specify</i>
What methods or software tools are needed to access the data?	<i>Please Specify</i>
Is documentation about the software needed to access the data included?	<input type="checkbox"/> Yes <input type="checkbox"/> No
Is it possible to include the relevant software (e.g. in open source code)?	<input type="checkbox"/> Yes <input type="checkbox"/> No (please specify):
Where will the data and associated metadata, documentation and code be deposited? ^{xv}	<i>Please Specify</i>
If there are restrictions on use, how	<i>Please Specify</i>

will access be provided?			
B3 - Making data Interoperable			
File format			
Spreadsheet: <input type="checkbox"/> ODS <input type="checkbox"/> XLS <input type="checkbox"/> CSV Documentation: <input type="checkbox"/> DOC <input type="checkbox"/> PDF <input type="checkbox"/> TXT <input type="checkbox"/> HTML	Structured data: <input type="checkbox"/> XML <input type="checkbox"/> JSON Geographical data: <input type="checkbox"/> DXF <input type="checkbox"/> SHP <input type="checkbox"/> GEOJSON	Image: <input type="checkbox"/> JPG <input type="checkbox"/> TIFF <input type="checkbox"/> PNG Video: <input type="checkbox"/> WEBM <input type="checkbox"/> MP4 <input type="checkbox"/> MKZ	Other (please specify):
What data and metadata vocabularies, standards or methodologies will you follow?		<i>Please Specify</i>	
B4 - Increase data Re-useable			
How will the data be licensed to permit the widest re-use possible?^{xvi}		<i>Please Specify</i>	
Data owner^{xvii}		<i>Please Specify</i>	
When will the data be made available for re-use?^{xviii}		<i>Please Specify</i>	
Are the data produced and/or used in the project useable by third parties, in particular after the end of the project?^{xix}		<input type="checkbox"/> Yes <input type="checkbox"/> No (please specify):	
How long is it intended that the data remains re-usable?^{xx}		<i>Please Specify</i>	
C – Allocation of Resources			
What are the costs for making data FAIR in the project?		<i>Please Specify</i>	
How will these be covered?^{xxi}		<i>Please Specify</i>	
What are the resources for long term preservation?^{xxii}		<i>Please Specify</i>	
D – Data Security			
Where will the datasets be stored or are already stored?		<i>Please Specify</i>	
What provisions are in place for data security (including data recovery, secure storage and transfer of sensitive data^{xxiii}		<i>Please Specify</i>	
Is the data safely stored in certified repositories for long term preservation and curation?		<i>Please Specify</i>	
Are you GDPR Compliant?		<i>Please Specify</i>	
E – Other Issues			
Do you make use of other national/funder/sectorial procedures for data management?^{xxiv}		<i>Please Specify</i>	

- ⁱ State the purpose of the data collection/generation.
- ⁱⁱ Define the nature of the data, e.g. quantitative, qualitative, numeric, text, audio, structured, unstructured, etc.
- ⁱⁱⁱ Define the format of the data, e.g. txt, xls, doc, pdf, etc.
- ^{iv} Specify if existing data is being re-used (if any).
- ^v Specify the origin of the data.
- ^{vi} State the expected size of the data (if known).
- ^{vii} Outline the data utility, to whom will it be useful, e.g. research groups, private sector, citizens, etc.
- ^{viii} Define if privacy preservation techniques are required before making your data open.
- ^{ix} Define the metadata standards you use. If there are no standards in your discipline describe what type of metadata will be created and how.
- ^x Outline the approach towards search keyword (e.g. search keywords will be provided when the dataset will be uploaded to the repository).
- ^{xi} Will data produced and/or used in the project be made openly available?
- ^{xii} If some data is kept closed provide rationale for doing so, clearly separating legal and contractual reasons from voluntary restrictions.
- ^{xiii} Specify how the data will be made available, e.g. by deposition in a repository.
- ^{xiv} Specify which repository will you use, e.g. Zenodo, open-access journal, project website, etc.
- ^{xv} E.g. Zenodo, open-access journal, project website, etc.
- ^{xvi} License conditions: Copyright, Creative Commons, Open License, etc. A list of licenses can be found here <https://opendefinition.org/licenses/>
- ^{xvii} List the data owner, the copyright owner, the intellectual property owner.
- ^{xviii} E.g. Immediately, after the end of the project (specify the exact time), along with the publication of main results, etc. 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.
- ^{xix} If the re-use of some data is restricted, explain why.
- ^{xx} Specify the length of time for which the data will remain re-usable, e.g. 5 years after the conclusion of the project.
- ^{xxi} 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).
- ^{xxii} Resources: Costs and potential value, who decides and how what data will be kept and for how long.
- ^{xxiii} Address data recovery as well as secure storage and transfer of sensitive data.
- ^{xxiv} Indicate if you must adhere to other policies and procedures for data management.