Intelligent Assistants for Flexibility Management (Grant Agreement No 957670)

D1.1 Initial data management plan

Date: 2021-05-27

Version 1.0

Published by the iFLEX Consortium

Dissemination Level: PU - Public

Co-funded by the European Union's Horizon 2020 Framework Programme for Research and Innovation under Grant Agreement No 957670

Document control page

Document file: D1.1 Initial data management plan
Document version: 1.0
Document owner: **JSI**

Work package: WP1 Project coordination
Deliverable type: ORDP

Document status: ☒ Approved by the document owner for internal review
☒ Approved for submission to the EC

Document history:

Version	Author(s)	Date	Summary of changes made
0.1	Dušan Gabrijelčič (JSI)	2021-03-24	Initial ToC
0.2	Dušan Gabrijelčič (JSI)	2021-05-14	Version ready for review
1.0	Dušan Gabrijelčič (JSI)	2021-05-27	Final version submitted to the European Commission

Internal review history:

Reviewed by	Date	Summary of comments
Trine F. Sørensen (IN-JET)	2021-05-25	Minor comments and corrections.
Roman Tomažič (ZPS)	2021-05-26	Review, minor comments

Index:

1 Executive summary

The deliverable presents the iFLEX project initial data management plan. The plan is based on Horizon 2020 Manual (EC, 2020) which lays out basic guiding principles and introduces essential findable, accessible, interoperable and re-usable (FAIR) data attributes and some extended views related to data management, like security and privacy. Mentioned concepts are briefly introduced in the deliverable and interwoven with the project proposed management plan.

A core element of the data management plan is the data management life-cycle. The deliverable proposes the project data life-cycle and describes its essential steps of data identification, collection, processing, storage, usage, sharing, archiving and destroying. Brief project data summary is provided based on current work on iFLEX Assistant architecture reported in D2.3 (iFLEX D2.3, 2021).

The deliverable advances the project data life-cycle with a proposal of data management process. The process has four phases: initialization, assessment, management and reporting. It is iterative and will produce two revisions of the data management plan besides this deliverable. The process is augmented with a simple data management template which will be used to document data sets addressed by the project. The template extends the current data summary as is reported in the deliverable D2.3 (iFLEX D2.3, 2021).

The initialisation phase of the data management plan has been carried out by defining the data management methodology which is based on a combination of data life-cycle and data management process. All three other phases have started through the work proposed in the deliverable. In the next period the summary of the project data will be further broadened through assessing the information on project data sets. The management phase will continue with stimulation, guidance and control of the process of data sets assessment and sharing. The information on the data management process and life-cycle will be passed to the project partners through presentations and workshops which will stimulate their participation in the process. The results will be reported in a new version of this deliverable at M18.

2 Introduction

This deliverable presents the iFLEX project data management plan. The document has a goal of initiating the work on data management and setting the general guiding principles for the work. The document should help the project, project partners and practitioners in the pilots to systematically document the data collected through the project work and utilize it ethically and with FAIR principles (Findable, Accessible, Interoperable and Re-usable) in mind during and after the project lifespan.

2.1 Context and scope

The deliverable is the first of three deliverables that will be provided during the project at M6, M18 and M36 of the project. The context of the first version of the deliverable are early stages of the development in the project. Recently a deliverable D2.1 (iFLEX D2.1, 2021) on use cases and requirements has provided a general scope of the project work. A deliverable D2.3 (iFLEX D2.3, 2021), published in April, provides an initial common architecture of the iFLEX framework. Initial pilot specification has been provided in the deliverable D7.1 (iFLEX D7.1, 2021), which gave first insight into the data that will be collected and processed in the pilots. Deliverables D10.1 (iFLEX D10.1, 2020) and D10.2 (iFLEX D10.2, 2020) deal with personal data management and provide as well procedures and mechanism for privacy regulation compliant data management and use. All deliverables are closely related to the data management. They define general procedures on how and who will use the data and which data will be processed and collected in the project. All stated sources will be used to scope the data management plan work.

2.2 Content and structure

The deliverable will first introduce general guiding principles in Section 3, define the project data file-cycle in Section 4 and initiate a brief data summary in Section 5. THE FAIR data management principles will be introduced in Section 6. Extended views on data management, covering ethical aspects, data security and other issues will be presented in Section 7. Initial data management plan will be provided in Section 8 where a simple data management template will be presented and allocation of resources in the project to data management explained. Finally, conclusions and future work will be given in Section 9.

3 Guiding principles

According to H2020 manual definition (EC, 2020) a Data Management Plan (DMP) is a key element of good data management. A DMP describes the data management life cycle for the data to be collected, processed and/or generated by the project. As part of making research data findable, accessible, interoperable and re-usable (FAIR), a DMP should include information on:

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

To be in-line with the guiding principles the deliverable will provide current view on the project data life-cycle in Section 4 and data summary in Section 5. Based on this information it will be possible to answer the questions set above and provide the data management plan in Section 8.

4 Data life-cycle

Figure 1 shows both the data life-cycle in yellow and the data management plan in blue. In this section only the data life-cycle will be explained, while the data management plan part will be addressed in Section 8.

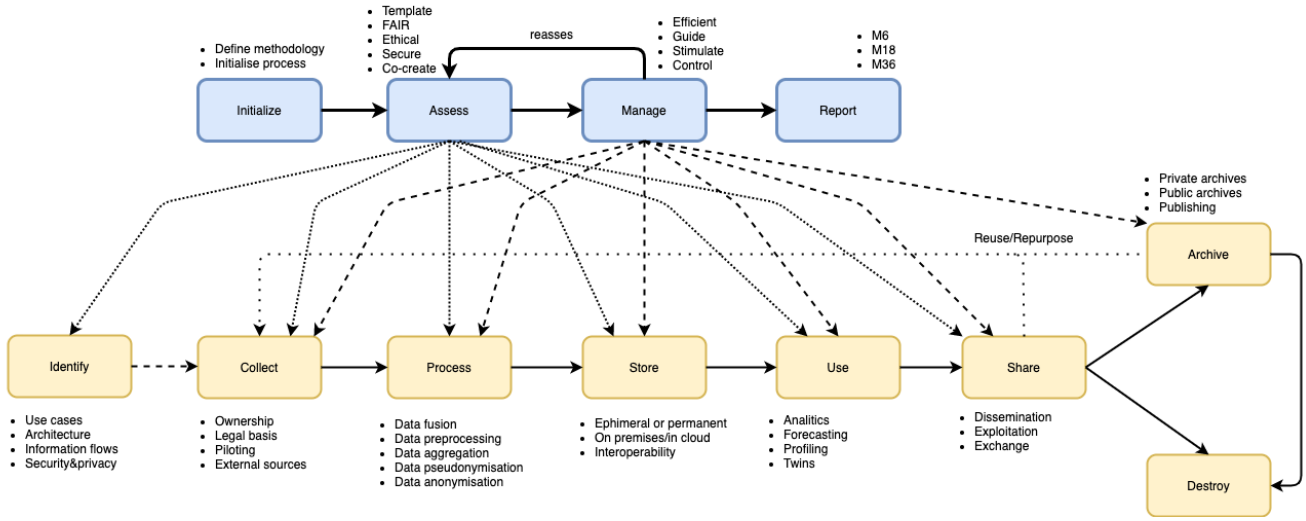

Figure 1: iFLEX project Data life-cycle and management process

The data life-cycle starts with an initial step of data identification. Important parts of this step have been or are currently in a process of initial implementation:

- *Use cases*: the project use cases have been defined in the D2.1 (iFLEX D2.1, 2021). The use cases provide essential information between whom, when and why information/data in the project will be exchanged among stakeholders,
- *Architecture*: the architecture defined in the deliverable D2.3 provides initial information on which elements of the iFLEX platform will process the data and how,
- *Security & privacy*: already in the WP10 Ethics and privacy deliverables D10.1 (iFLEX D10.1, 2020) and D10.2 (iFLEX D10.2, 2020) have provided foundations on ethics and privacy requirements and their fulfilment. The forthcoming security and privacy deliverable D4.7 (iFLEX D4.7, 2021) will extend the WP10 deliverables view and provide initial information on security and privacy mechanisms needed to fulfil security and privacy requirements as set in D2.3 (iFLEX D2.3, 2021),
- *Information flows*: the deliverables mentioned in the initial phase provide enough information to depict central information flows of the project identifying origins of the information and the ways and points of its processing and transformation throughout the project platform.

The data life-cycle beyond identify phase is defined as common life-cycle of mature process of data collection, processing and sharing. The phases, aligned with the project internal steps and plans are as follows:

- *Collect*: during the collection phase the data is being collected in the pilots. Initial pilot specification has been provided in (iFLEX D7.1, 2021) which had already hinted where, when and which data will be colleted. For this phase is of utmost importance to understand an ownership of the data and legal basis allowing to collect the data. Proper informed consent procedure is a must as is explained in D10.1 (iFLEX D10.1, 2020). The life-cycle can cover other important data from external sources which is not privacy sensitive, such as weather or CO2 emission data,
- *Process*: the processing phase involves basic processing and preparation of the data for future use in the system and beyond. The processing can involve pseudonymization or anonymization, and general data pre-processing, fusion and aggregation. Results of the processing can be already used in the system without any further changes to the data,
- *Store*: after the processing the data is stored and ready for the later use. The storage can be for a long term or just ephemeral in case of streaming and relaying the data among system elements. The system can use a private storage when the storage is under control of the data owner or the data controller or

a public storage in a cloud. In all cases the storage needs to guarantee target security requirements as set in D2.3 (iFLEX D2.3, 2021),

- *Use*: the core step where the data is transformed into a product via analytics, forecasting, profiling, twin emulation, etc. The data usage allows providing anticipated benefits for end users and other stakeholders of the system,
- *Share*: the products and the data are shared in this phase with intended audiences. Dissemination, exploitation and exchange of data and services with other projects are planned in the project. Procedures needs to be defined to share the data with external entities in a secure and privacy aware manner. In general, only anonymized data should be shared with the external entities. In such case the data export can be reused or repurposed in another data life-cycle as is indicated in Figure 1 with a dashed feedback loop. The sharing of the data will be tracked to keep an archive of data set, anonymization technique, including with a version and the time of the export,
- *Archive*: one of the possibilities to share the data for long term is to archive it in a private or public archive. There it is accessible for internal or wider audiences. Typical examples are publishing archives of scientific publications, accompanied with research data. The archives can in a similar manner as in Share phase case become a source of data for another data life-cycle,
- *Destroy*: if the data is not shared anymore and there is no purpose to be archived it can get destroyed. Personal data will get destroyed after the period specified in the informed consent used to set legal basis to collect the data from the end users. The project or project members can decide that some data or publication archived is not suitable any more for publication. In this case the data and related archive can be destroyed at well.

5 Data Summary

Initial information on data collected and processed in the project data life-cycle is given in the deliverable D2.3 (iFLEX D2.3, 2021), see a summary table 4 of information view for details. The table provides information summary on data items identified in a contextual and functional view, their description and data type. It doesn't provide only the information on data being collected but the data being generated as well. Some of the data has been already collected and being in process of collection by partners in the project, like Advanced Metering Infrastructure (AMI) smart metering data. Some data is from outside sources, like weather data and CO2 emissions data. The size of the data is currently not defined yet. The utility of the data will be further determined through the D2.1 use cases studies and realisation in technical work packages and pilots.

We will argue at this point that this table is a starting point of data management plan and process as is laid out in Section 8. What the data items still need are properties definition related to FAIR data management and Extended view as are discussed in Sections 6 and 7. The summary view will be replaced in further iterations with data collection information provided in this document appendix, based on data management template as is defined in Section 8.1.

6 FAIR data management

The term FAIR data management refers to following data management attributes: findable, accessible, interoperable and reusable data. Below basic introduction for FAIR data attributes based on Horizon 2020 Data Management Manual (EC, 2020) will be provided.

Findable attribute refers to possibility to make data discoverable via metadata or standard identification mechanisms. The metadata should support not only the project needs but as well potential re-use. Whenever possible already standardised metadata and data format should be selected and extended, if needed. If there are estimated significant benefits from simplifying the conventions they could be considered as well. A possibility to have multiple revisions of metadata and related data should result in a clear versioning system proposed in the data management plan.

Accessible attribute refers to a selection process of data that will be made openly available as default. If the data cannot be provided in open manner an explanation should be provided for such restriction. For open data it should be described how it will be published; open, standard and certified repositories are preferred to non-certified ones. If there is a specialized software needed to access and read the data such software should be preferably provided under open-source license. When the data is openly published it should be provided under well-defined open license.

Interoperability attribute requires open and standardised data and metadata formats compliant as much as possible with available open software applications. The attribute should guarantee open exchange of the data with other researchers, organisations and institutions.

Reusable attribute measures the re-usability of the data. The attribute can be affected with a selection of data license, time-frame and actual actuality of the data release, quality, usability of the data, etc. It is needless to say that the other three attributes can have considerable effect on the achieved level of re-usability.

7 Extended views

There are a number of extended views that could affect the data management. *Data security* is important both for the data collected, stored, processed and used in the project as well for the data shared outside the project. Security of the in-project use will be covered by the forthcoming deliverable D4.7 on Secure consumer data management module (iFLEX D4.7, June 2021). Security aspects of the sharing of the data will be covered by the data management plan and revisions of this deliverable.

Ethical aspects cover legal and ethical issues that can have impact on data sharing. For the iFLEX project, the data privacy is of particular importance. According to the data management plan no personal data will be shared outside the project. All the shared data will be either not private or anonymized. The data management will include control over the sharing of data and approval will be needed to share the data. The approving process will be supported by Ethical Advisory Board (EAB) as is defined in the deliverable D10.2 (iFLEX D10.2, 2020).

There could be *other views* that should be considered as well, like national, sectorial, company, institutional or departmental procedures or policies related to data management. If there are any such views they should be considered and reported together with the data.

8 Data management process

The data management plan should consider all information provided in previous Sections. A general process of the data management is presented in Figure 1, in the top line denoted with blue boxes. The process consists of the following phases:

- *Initialize*: the phase when the data management plan methodology is defined and process of data management initialised. This deliverable boosts the initialisation phase and defines necessary concepts and process to start data management,
- *Assess*: the assess phase carefully monitors the phases of the data life-cycle as is presented in Figure 1. During the assessment a data management template is used to record information about the data passed through the life-cycle. The template is defined in Section 8.1. Together with the template the FAIR principles and Ethical and Security views are evaluated as well. The assessment is done by the project partners with closest understanding of the data under consideration. At each data management process cycle care will be taken to involve end users into a co-creation of the data management process. The co-creation involvment will be included in each cycle report,
- *Manage*: the management phase manages the data management process as well influences the data life-cycle stages. The process should be managed in a way to be efficient and lean. It should guide the data life-cycle stages and stimulate proper assessment of the data. Sharing of the data should be encouraged through internal assessment, selection and recommendation of potential repositories for publishing the data. The guiding is done through project general and technical meetings as well dedicated data workshops. The management will take care of control both of quality and quantity of the shared data as well in coordination with Ethics Advisory Board about security and privacy aspects of the shared data. The management phase triggers assessment phase before every reporting phase to reassess the data templates and to add potentially new data to the collection of the templates. Data Protection Impact Assessments will also be carried out. The management phase will be active constantly during the project lifetime,
- *Report*; the reporting phase is done at scheduled intervals. After this deliverable the next realises of the Data Management Plan are expected at M18 and M36 of the project.

8.1 Data management template

The data management template is a simple table extending the data summary initial table as introduced in the data summary Section 5. The summary table is extended with a data identifier, FAIR, Extended views, owner and status rows as is shown in Table 1. The column(s) of the table present one or more data sets considered in the project. The table could be split in multiple tables.

Table 1: The data management template

Field	Data set
Column 1 text	Column 2 text
Identifier	Unique identifier of the data set, consisting of combination of abbreviation and version number: ABBR-x.y
Name	The name of the dataset
Change log	Change log of the data set, includes the date of change and changes description
Description	The description of the dataset, similar to one provided in the deliverable D2.3 (iFLEX D2.3, 2021)
Type	The type of the dataset, similar to one provided in the deliverable D2.3 (iFLEX D2.3, 2021)

FAIR	Conformance to FAIR principles as discussed in Section 6, denoted with letters F, A, I, R, if the attribute is achieved.
Extended	Denotes a status of the Extended views, denoted with letters: • S: security of the exported data set provided • P: data privacy noted • A: the data set is anonymized • O: other issues exist
Owner	The owner or caretaker of the dataset, one of the partners in the project
Status	Status if the dataset has been shared or not, together with the repository

The design of the table is deliberately simple to ease tracking the data sets assessed in the project. Not all data sets will reach the level to be shared. For the shared ones the template information will be extended to match target repository requirements. Extended dataset information together with information on repository will be provided in an appendix of one of revisions of this deliverable.

8.2 Allocation of resources

Current allocation of resources is limited to the manpower withheld within WP1 for the purpose of data management. Three man-months are allocated to JSI, authoring institution of this deliverable. The resources will be used for preparing the deliverables and for performing the data management tasks as have been laid out in the data management plan in Section 8. Potential costs of open access publishing have been accounted for in some of the partner's budget. The resources for long term preservation haven't been discussed yet.

9 Conclusions and future work

The deliverable presented basic guidelines for data management and introduced FAIR data management principles. Based on current and planned work in the project a data life-cycle was proposed and corresponding data management process defined. To facilitate data assessment a data management template has been introduced, matching current data summary and extended with essential attributes.

The data management process has been initialized and assessment and management phases of the process started. In the next period we expect current data summary to be extended to proposed template and in number of data sets and details provided.

10 List of figures and tables

10.1 Figures

10.2 Tables

11 References

(iFLEX D2.1, 2021) Use cases and requirements. The iFLEX project deliverable D2.1, January 2021.

(iFLEX D2.3, 2021) Common architecture of iFLEX Framework. The iFLEX project deliverable D2.3, April 2021.

(iFLEX D7.1, 2021) Initial pilot specification. The iFLEX project deliverable D7.1, February 2021.

(EC, 2020) Horizon 2020 Manual. [Online, accessed 26. 3. 2020] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

(iFLEX D10.1, 2020) D10.1 H - Requirement No. 1. The iFLEX project deliverable 10.1, December 2020.

(iFLEX D10.2, 2020) D10.2 POPD - Requirement No. 2. The iFLEX project deliverable 10.2, December 2020.

(iFLEX D4.7, 2021) Secure consumer data management module. The IFLEX project deliverable D4.7, in preparation, target release June 2021.