SUPPORT4LHS

Process Mining and Knowledge Representation technologies to Support the Learning Health System

Deliverable 3.1 - Specification of the Data Management Plan (DMP)

Project deliverable

Authors: Mikel Egaña Aranguren (UPV/EHU), ???,

[1.- Executive summary 1](#_Toc1531280118)

[2.- Introduction 1](#_Toc228176571)

[3.- Datasets 2](#_Toc856064865)

[4.- Architecture 2](#_Toc837912187)

[Processing pipeline 2](#_Toc823652587)

[Web frontend 4](#_Toc415038207)

[5.- Conclusions 5](#_Toc329874453)

[6.- References 5](#_Toc54360834)

[Annex A: Support4LHS FAIR data questionnaire 5](#_Toc2067874064)

[Annex B: FAIR principles 5](#_Toc927834810)

# 1.- Executive summary

# 2.- Introduction

The objective 3.1 ("Design and implementation of a FAIR Data Management Plan") of the project Support4LHS aims at publishing the datasets of the project following the FAIR principles (Wilkinson et al., 2016). In order to accomplish this, as specified in the grant agreement, two main tasks need to be fulfilled:

**(1) Design and creation of a Data Management Plan (DMP).** The DMP is presented in this deliverable (*D 3.1 - Specification of the Data Management Plan (DMP))*. This DMP is conceived as means to support the whole life cycle of the project data that will be collected, processed or generated. Therefore, the document is alive, and it will be updated through the life span of the project, accommodating changes that will be presented in the *Deliverable 3.2 - Final report on the Data Management Plan and degree of accomplishment of FAIR principles*.

**(2) Implementation and deployment of the DMP.** The ideas reflected in the DMP will be implemented during the project, and the evaluation of the results of this implementation presented in the *Deliverable 3.2 - Final report on the Data Management Plan and degree of accomplishment of FAIR principles*. However, an architectural overview is provided in this DMP.

The remainder of the document is organised as follows:

**Section 3 ("Datasets")** provides a catalog of the datasets that are expected to be produced in the project, with information collected through the form described in Annex A.

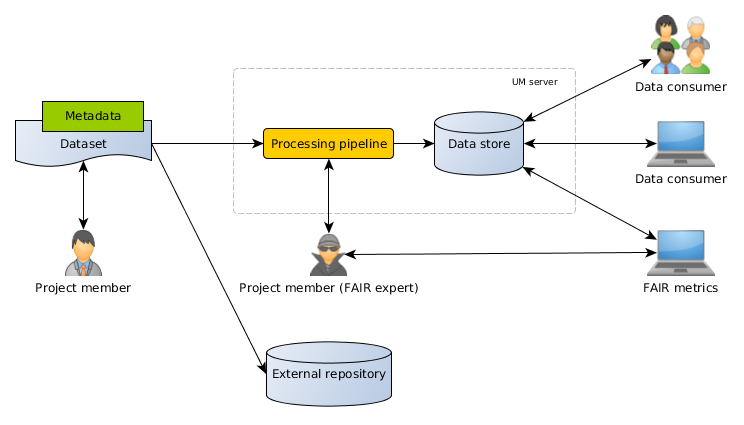
**Section 4 ("Architecture")** describes the overall technical setting designed to capture and publish the data, including details on the resources allocated for the processes.

**Section 5 ("Conclusions")** wraps the document with final considerations for the future development of the DMP and its implementation.

# 3.- Datasets

# 4.- Architecture

The basic architecture of the "FAIRification" framework that will be implemented, explained in this section, also illustrates how the most salient points of the Objective 3.1 of the grant agreement will be realized. The main task of the FAIRification framework is to process the data produced in the project and publish it according to FAIR principles, as illustrated in Figure 1 (More details are provided in the following figures).



*Figure 1: basic architecture of the FAIRification framework.*

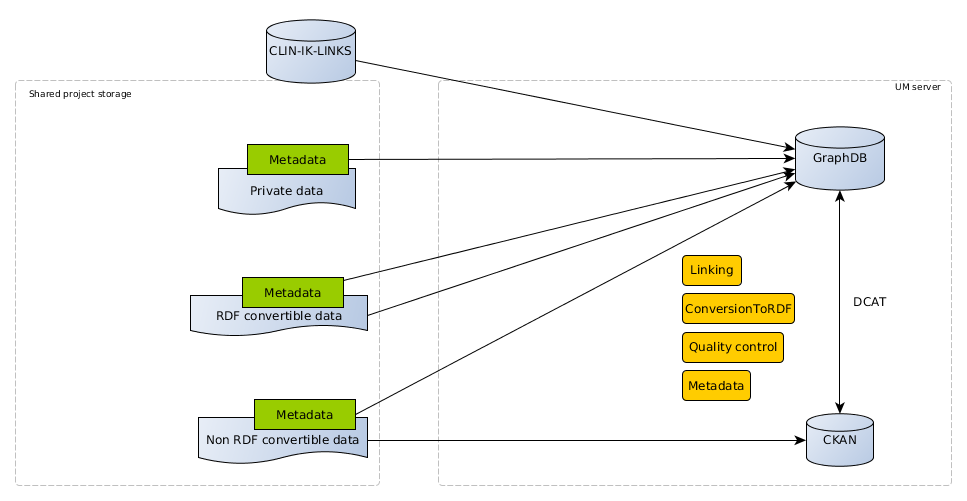
The process is divided into two main steps:

**Processing pipeline:** it acquires the data from the project members, it processes it, and it stores it in the data store. The project members of Objective 3.1 oversee the processing pipeline. The pipeline is deployed in a UM (Universidad de Murcia) server. More details are provided bellow.

**Web frontend:** The data stored by the processing pipeline is published according to FAIR principles in a web frontend at the UM server. Such publication is targeted at external clients, both humans (Other scientists) and, more importantly, computational agents. The FAIR metrics framework used to evaluate the "FAIR level" achieved also consumes data from this frontend, and it is used iteratively by the project members to adjust the FAIRification process (Wilkinson et al., 2019). More details are provided bellow.

## Processing pipeline

The processing pipeline is described in Figure 2. The pipeline comprises the processing proper (Linking, quality control, etc.) and the storage of data and metadata. Such storage is implemented by GraphDB[[1]](#footnote-17770) for RDF based data and CKAN[[2]](#footnote-10086) for file-based data.



*Figure 2: processing pipeline.*

The pipeline kicks-off when a project member provides a new dataset, from the ones described in section 3, in the shared project storage. There are four types of datasets with regards to their treatment by the processing pipeline:

**CLIN-IK-LINKS datasets:** the CLIN-IK-LINKS platform will store Clinical Process Models (CPMs) and Knowledge Graphs (KGs) produced in the project. Since the platform will offer REST APIs to access those elements, a connector will be developed to insert those elements directly into the GraphDB Triple Store.

**Private data:** clinical data tends to be protected by strict legislation. In this case, since the data cannot be published, some metadata will be collected in RDF and stored in GraphDB, specially, but not only, referring to possible access methods (e.g., contact information for the person responsible for data access in a hospital). Storing the Metadata in GraphDB, regardless of the data storage, also ensures the application of principle A2 ("Metadata are accessible, even when the data are no longer available").

**Public, RDF convertible data:** public data that can be fully published and it is already available in RDF[[3]](#footnote-32314) (Resource Description Framework) or it is feasible to convert to RDF. In this case both data and metadata will be stored in GraphDB.

**Public, non RDF convertible data:** public data that can be fully published but it is not available in RDF or it is very difficult to convert to RDF. In this case the metadata will be stored in GraphDB, with pointers to a CKAN server, in which the data, in its original form, will be stored. The CKAN DCAT extension will be used to synchronise GraphDB and CKAN at the metadata level, an ensure that the FAIR principle F3 ("Metadata clearly and explicitly include the identifier of the data they describe") is implemented.

The processing will be implemented with tailored programs, using cwltool[[4]](#footnote-1839) (Common Workflow Language tool) as a framework for combined execution in workflows and provenance. The processing pipeline comprises the following specific processes:

**Metadata.** This process ensures that the published metadata will have a minimum quality, either by transforming the existing metadata or adding new metadata items to implement principle F2 ("Data are described with rich metadata") and R1 ("(Meta)data are richly described with a plurality of accurate and relevant attributes "). This Metadata baseline will imply the use of the DCAT[[5]](#footnote-20674), VoID[[6]](#footnote-16129), PROV[[7]](#footnote-21939), and Creative Commons[[8]](#footnote-12352) vocabularies, apart from any other vocabularies already present in the datasets. The use of this vocabularies will also entail the application of principle I2. ("(Meta)data use vocabularies that follow FAIR principles").

**Conversion to RDF:** in order to apply principle I1. ("(Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation") metadata and data (To the extent possible) will be converted to RDF. The conversion to RDF will be performed with tailored programs, written either in Python (Using RDFLib[[9]](#footnote-1606)) or Java (Using RDF4J[[10]](#footnote-31141)).

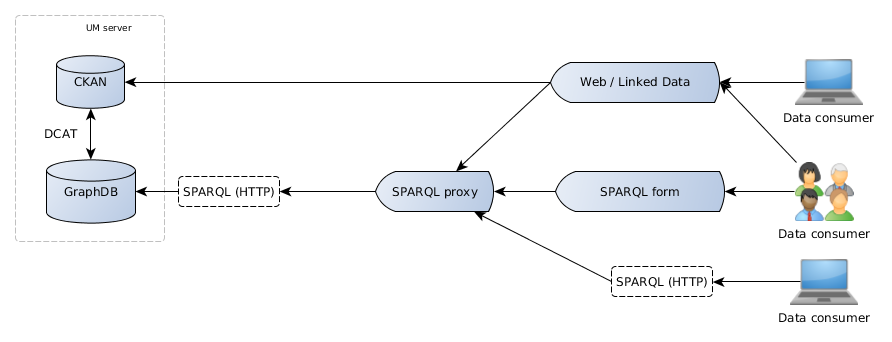
**Quality control:** SHACL[[11]](#footnote-18578) will be used to ensure the quality of the produced RDF, especially in the case of Metadata.

**Linking:** in order to apply principle I3 (" (Meta)data include qualified references to other (meta)data") RDF links will be added to metadata, and to a lesser extent also to data, through the SILK[[12]](#footnote-16402) platform or manually.

## Web frontend

The publication frontend offers the possibility of consuming the stored data through different interfaces, suitable for different clients. All the channels of communication are based on the HTTP protocol, applying principle A1. ("(Meta)data are retrievable by their identifier using a standardised communications protocol").

The publication frontend ensures the implementation of the principle F1. ("(Meta)data are assigned a globally unique and persistent identifier") using W3ID identifiers[[13]](#footnote-28839). In order to apply principle F4. ("(Meta)data are registered or indexed in a searchable resource") the provided content will be annotated with JSON-LD scripts that follow the bio-schema and schema vocabularies, in order to be crawled in a structured way by the most common search engines.



*Figure 3: publication frontend.*

YASQUE, Zazuko, ...

Apart from the FAIR publication frontend described in Figure 3, a static web will be offered with the following content: A Web page describing the access to the data for humans but specially machines. This Web page will also include the DMP and any other data-related information: for example, pointers to any external repositories used for depositing project data (e.g. Zenodo, Dataverse, etc.). Benchmarks based on FAIR metrics.

A GitHub project has been set up for the development of the FAIR publication framework[[14]](#footnote-7057).

# 5.- Conclusions

6.- References

References

Wilkinson, M. D., Dumontier, M., Sansone, S., Bonino da Silva Santos, Luiz Olavo, Prieto, M., Batista, D., . . . Schultes, E. (2019). Evaluating FAIR maturity through a scalable, automated, community-governed framework. *Scientific Data, 6*(1), 1-12. doi:10.1038/s41597-019-0184-5

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., . . . Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship : Comment. *Scientific Data, 3*, 1-9. Retrieved from <https://www.narcis.nl/publication/RecordID/oai:library.wur.nl:wurpubs%2F501704>

# Annex A: Support4LHS FAIR data questionnaire

# Annex B: FAIR principles

The GO FAIR initiative[[15]](#footnote-6621) offers a suitable explanation of the FAIR principles, as follows:

**Findable**

F1. (Meta)data are assigned a globally unique and persistent identifier

F2. Data are described with rich metadata (defined by R1 below)

F3. Metadata clearly and explicitly include the identifier of the data they describe

F4. (Meta)data are registered or indexed in a searchable resource

**Accessible**

A1. (Meta)data are retrievable by their identifier using a standardised communications protocol

A1.1 The protocol is open, free, and universally implementable

A1.2 The protocol allows for an authentication and authorisation procedure, where necessary

A2. Metadata are accessible, even when the data are no longer available

**Interoperable**

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (Meta)data use vocabularies that follow FAIR principles

I3. (Meta)data include qualified references to other (meta)data

**Reusable**

R1. (Meta)data are richly described with a plurality of accurate and relevant attributes

R1.1. (Meta)data are released with a clear and accessible data usage license

R1.2. (Meta)data are associated with detailed provenance

R1.3. (Meta)data meet domain-relevant community standards

1. https://graphdb.ontotext.com/ [↑](#footnote-ref-17770)
2. https://ckan.org/ [↑](#footnote-ref-10086)
3. https://www.w3.org/TR/rdf11-concepts/ [↑](#footnote-ref-32314)
4. https://github.com/common-workflow-language/cwltool [↑](#footnote-ref-1839)
5. https://www.w3.org/TR/vocab-dcat-2/ [↑](#footnote-ref-20674)
6. https://www.w3.org/TR/void/ [↑](#footnote-ref-16129)
7. https://www.w3.org/TR/prov-o/ [↑](#footnote-ref-21939)
8. https://creativecommons.org/ns [↑](#footnote-ref-12352)
9. https://rdflib.dev/ [↑](#footnote-ref-1606)
10. https://rdf4j.org/ [↑](#footnote-ref-31141)
11. https://www.w3.org/TR/shacl/ [↑](#footnote-ref-18578)
12. http://silkframework.org/ [↑](#footnote-ref-16402)
13. https://w3id.org/ [↑](#footnote-ref-28839)
14. https://github.com/mikel-egana-aranguren/SUPPORT4LHS-FAIR-data [↑](#footnote-ref-7057)
15. https://www.go-fair.org/fair-principles/ [↑](#footnote-ref-6621)