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 3](#_Toc132391246)

[2.- Introduction 5](#_Toc2132926119)

[3.- Datasets 5](#_Toc921869820)

[4.- Architecture and processes 5](#_Toc1043703698)

[5.- Conclusion 5](#_Toc103628451)

[6.- References 5](#_Toc1440617340)

[Annex A: Support4LHS FAIR data questionnaire 6](#_Toc353318005)

# 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 principles1. 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: *Deliverable 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 implementation and deployment of the DMP will be carried out during the project, and the evaluation of its results presented in the *Deliverable 3.2 - Final report on the Data Management Plan and degree of accomplishment of FAIR principles*. However, an architectural overview of the processes designed to capture and publish the datasets is provided in this DMP.

The remainder of the document is organised as follows:

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

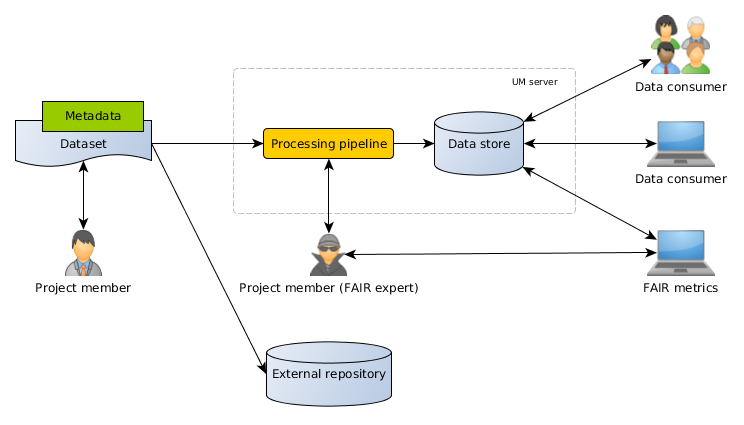
**Section 4 ("Architecture")** describes the overall technical setting designed to capture and publish the data, including details on the resources allocated for the processes.

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

# 3.- Datasets

# 4.- Architecture

The basic architecture of the "FAIRification" framework that will be implemented also illustrates how the most salient points of the grant agreement will be realized (w.r.t objective 3.1). 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 following figures).



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

The process is divided into two main steps:

1) 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.

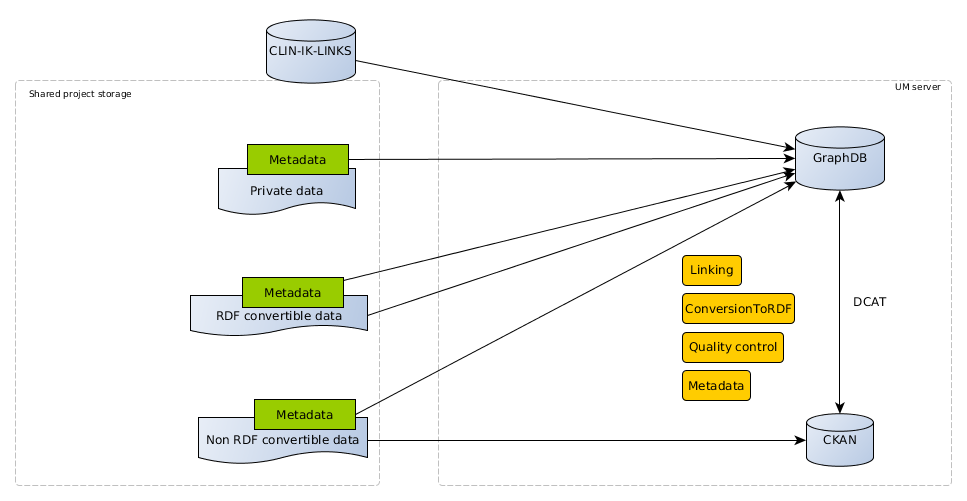
2) The data stored in the data store by the processing pipeline is published according to FAIR principles in a web frontend at the UM server. Such publication is targeted at external clients, both humans (Other scientists) and, more importantly, computational agents. The FAIR metrics framework used to evaluate the "FAIR level" also consumes data from this frontend, and it is used by the project members to adjust the FAIRification process.

The processing pipeline is described in Figure 2. The pipeline comprises the processing module (Linking, quality control, etc.) and the storage of data and metadata. Such storage is implemented by:

GraphDB: a Triple Store for storing RDF based data.

CKAN: a

The pipeline kicks-off when a project member provides a new dataset, from the ones described in section 3, in the shared project storage.



*Figure 2: processing pipeline.*

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).

**Public, RDF convertible data:** public data that can be fully published and it is already available in RDF 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 pipeline comprises the following specific processes:

Metadata. This process ensures that:

a) The necessary metadata, in DCAT (RDF) form, is added

VoID, PROV

b) The metadata already present conforms with DCAT (Via conversion, if necessary)

Quality control: SHACL will be used to ensure

Metadata addition

Metadata processing

Conversion to RDF

The publication frontend ...

*Figure 3: publication frontend.*

Grant agreement ...

# 5.- Conclusions

ok

6.- References   
1. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR guiding principles for scientific data management and stewardship : Comment. *Scientific data*. 2016;3:1-9.

# Annex A: Support4LHS FAIR data questionnaire

# Annex B: FAIR principles