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), EDUARDO ILLUECA (UMU)

1.- Executive summary

This deliverable is the Data Management Plan (DMP) for the Support4LHS project. Its aim is two-fold:

1. To collect the provisional descriptions of the datasets that will be produced during the project in a catalogue.
2. To provide a technical overview of the publication of such datasets following FAIR principles (Findable, Accessible, Interoperable, Reusable).

This document will be updated throughout the project life span, since the datasets themselves and the publishing process will change considerably.

2.- Introduction

The Objective 3.1 (*Design and implementation of a FAIR Data Management Plan*) of the project Support4LHS comprises the publication of project datasets following the FAIR principles (Wilkinson et al., 2016). The datasets are produced and/or collected by other project members as part of the development of their respective objectives.

The project members of the Objective 3.1 are regarded as **Data Stewards** and will take responsibility for the FAIR publication process, overseeing the whole process (Jacobsen et al., 2020). In order to accomplish it, 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 a means to support the whole life cycle of the project data. Therefore, the document is alive, and it will evolve 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, in Section 4.

The remainder of the document is organised as follows:

Section 3 ("Datasets") provides a catalogue of the datasets that are expected to be produced in the project.

Section 4 ("Architecture") describes the overall technical setting designed to capture and publish the data.

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

3.- Datasets

In order to achieve an adequate level of FAIRification basic information about the datasets that will be published must be captured in a catalogue. The details of the prospective datasets to be FAIRified were obtained through the form described in Section 3.2 , and the answers, comprising the actual catalogue of datasets, is described in Section 3.3. A preliminary list is described in Section 3.1.

3.1.- Preliminary list of datasets

In order to obtain a preliminary list of possible datasets, an informal enquiry was directed to the project members, obtaining the results depicted in Table 1. This includes the names of the datasets not to be published.

Objective	**Datasets**
O1 Knowledge-driven methods to support process mining in healthcare	
O1.1 Knowledge and standards-based transformation methods for process mining data ingestion in healthcare	No datasets produced
O1.2 Extraction of knowledge graphs from clinical process models	Clinical Process Mining Graphs
O1.3 Interoperability framework based on knowledge graphs for clinical process models	OWL ontologies
O1.4 Knowledge Graph methods for analysing clinical process models	Metrics, alignments
O1.5 Exploitation of knowledge models for the explainability of process mining models	Alignments
O2 Data-driven methods to support the discovery of healthcare processes	
O2.1 New process mining methods and metaphors for the discovery of clinical processes from EHR data	Metaphors (Format To be Defined)
O2.2 Integration framework for process mining models and clinical guideline models	Not applicable, Clinical Guideline Models
O2.3 Interactive process mining methods based on integrated process mining models and clinical guideline models	Clinical process circuits, Process Mining Models
O2.4 Analysis of compliance based on integrated process mining models and clinical guidelines	No datasets produced
O3 Data management and application to use cases	
O3.1 Design and implementation of a FAIR DMP	No datasets produced
O3.2 Interactive process mining for the discovery of clinical processes in selected scenarios	Clinical process circuits, Process Mining Models
O3.3 Development of knowledge models for selected scenarios	Clinical Process Mining Graphs, Clinical Guideline Models, alignments, annotations
O3.4 Exploitation of process mining models and knowledge models in selected scenarios	No datasets produced

Table 1: preliminary list of datasets for each objective.

3.2.- Support4LHS FAIR data questionnaire

The aim of this questionnaire is to create a catalogue with the datasets that will be produced as part of the Support4LHS project, in order to assess the necessary adjustments to the FAIR data publication framework developed as part of Objective 3.1 (Design and implementation of a FAIR

Data Management Plan). The datasets will be published, to the extent possible, respecting the FAIR principles (Findable, Accesible, Interoperable, Reusable). The results of this questionnaire will be included in the Deliverable 3.1 (Specification of the Data Management Plan (DMP)) as catalogue of datasets.

It is assumed that the leader of each objective is aware of the datasets that the objective will produce, so this form is aimed at objective leaders. Even though it is impossible for an Objective leader to foresee all the datasets, the more details are provided herein, the less adjustments necessary for the FAIR data publication framework.

If your objective will produce more than one dataset, fill the form for each dataset (But please try to group the data in as few datasets as possible).

Any questions can be directed to mikel.egana@ehu.eus.

Dataset ownership

Questions about who owns the data

1.- What objective of the project are you the leader of?

2.- Name, institution, and email of the person responsible (Objective leader)

Dataset description

Basic information about the dataset

3.- Name of the dataset

4.- Origin of the dataset

5.- Dataset purpose: what is the aim of this dataset? What is it going to be used for?

6.- Dataset format: what is the expected format of the dataset? (File formats, relational databases, etc)

7.- Data "shape": What are the data types of the dataset? (e.g., data types of the columns on a spreadsheet)

8.- Data volume: what is the expected size of the dataset?

9.- Updates: what is the expected update frequency for the dataset?

Dataset interoperability

Information about the current interoperability level of the dataset

10.- Is the dataset already published following FAIR principles? Where?

11.- Does the dataset reuse other datasets?

12.- Does the dataset include rich metadata? In which form?

13.- Does the dataset follow naming conventions or any other community standards of its domain?

14.- Does the dataset include vocabularies, ontologies, or any other standards at metadata or data level? Which ones?

15.- Does the dataset include new ontologies developed specifically for it? Where are they published?

16.- Is the dataset properly versioned?

17.- Does the dataset include an explicit and machine-readable license?

18.- Are there any special methods or software needed to access the data? Which ones?

Dataset publication

Information about the specifics of publishing the dataset following FAIR principles

19.- Is there an embargo period in which the dataset, or its metadata, cannot be published? (For example, if you prefer to wait for manuscript acceptance in a journal)

20.- Do you agree to publish the dataset and/or its metadata following FAIR principles? Specify any limitations to publish the datasets and associated metadata, detailing any alternative access methods. For example, in a clinical dataset it might be impossible to publish the data to respect patients' privacy, but the method to access it is still valuable (For example the email of the person in charge of the committee to access the data might be provided as basic metadata)

21.- Is the dataset protected by any regulation or special legislation related to personal data? Which one?

22.- Is the dataset going to be published in a specialised public repository (e.g., like UniProt - https://www.uniprot.org/-)? Which one? When is the expected publication date?

23.- I commit to post the dataset (or its metadata, in case of private data) and its updates to the shared resource that will be set up to that end (FTP server, OneDrive shared folder, etc.). (Name and email)

24.- Has any anonymisation been applied to the dataset to preserve patient privacy?

3.3.- Support4LHS FAIR data catalogue (questionnaire results)

O1.2 Extraction of knowledge graphs from clinical process models

1.- What objective of the project are you the leader of?

O1.2 Extraction of knowledge graphs from clinical process models.

2.- Name, institution, and email of the person responsible (Objective leader)

Jose Antonio Miñarro Giménez, Universidad de Murcia, jose.minyarro@um.es.

3.- Name of the dataset

ClinicalPMGraph.

4.- Origin of the dataset

The dataset is generated from a Clinical Process Model that is extracted from clinical notes.

5.- Dataset purpose: what is the aim of this dataset? What is it going to be used for?

The aim of this dataset is the semantic formalization of procedural stages of patient. This dataset can be used to compare the actual clinical process with clinical guidelines.

6.- Dataset format: what is the expected format of the dataset? (File formats, relational databases, etc)

We will provide the model in a OWL file. The complete dataset will be provided in a graph database.

7.- Data "shape": What are the data types of the dataset? (e.g., data types of the columns on a spreadsheet)

The most of the data types will be sort strings but some numerical values will also be included.

8.- Data volume: what is the expected size of the dataset?

The data volume is limited to 10-20 process nodes as they represent aggregated data.

9.- Updates: what is the expected update frequency for the dataset?

Each time a new or updated of a clinical process model is provided.

10.- Is the dataset already published following FAIR principles? Where?

Not yet.

11.- Does the dataset reuse other datasets?

It will use clinical process model generated in O2.1.

12.- Does the dataset include rich metadata? In which form?

It will include mappings to medical ontologies to represent details of PROCEDURES, SUBSTANCES, ...

13.- Does the dataset follow naming conventions or any other community standards of its domain?

Not yet decided.

14.- Does the dataset include vocabularies, ontologies, or any other standards at metadata or data level? Which ones?

Yes, SNOMED CT.

15.- Does the dataset include new ontologies developed specifically for it? Where are they published?

Clinical Process Model Ontology is going to be defined. It has not been published yet.

16.- Is the dataset properly versioned?

To be decided.

17.- Does the dataset include an explicit and machine-readable license?

To be decided.

18.- Are there any special methods or software needed to access the data? Which ones?

It will need to have a compatible graph database.

19.- Is there an embargo period in which the dataset, or its metadata, cannot be published? (For example, if you prefer to wait for manuscript acceptance in a journal)

It will need to have a compatible graph database.

20.- Do you agree to publish the dataset and/or its metadata following FAIR principles? Specify any limitations to publish the datasets and associated metadata, detailing any alternative access methods. For example, in a clinical dataset it might be impossible to publish the data to respect patients' privacy, but the method to access it is still valuable (For example the email of the person in charge of the committee to access the data might be provided as basic metadata)

To be decided.

21.- Is the dataset protected by any regulation or special legislation related to personal data? Which one?

To be decided.

22.- Is the dataset going to be published in a specialised public repository (e.g., like UniProt - https://www.uniprot.org/-)? Which one? When is the expected publication date?

To be decided.

23.- I commit to post the dataset (or its metadata, in case of private data) and its updates to the shared resource that will be set up to that end (FTP server, OneDrive shared folder, etc.). (Name and email)

Jose Antonio Miñarro Giménez, jose.minyarro@um.es.

24.- Has any anonymisation been applied to the dataset to preserve patient privacy?

I believe in O2.1 some anonymisation and aggregation is applied to preserve patient privacy.

O1.3 Interoperability framework based on knowledge graphs for clinical process models

1.- What objective of the project are you the leader of?

O1.3 Interoperability framework based on knowledge graphs for clinical process models

2.- Name, institution, and email of the person responsible (Objective leader)

jfernand@um.es.

3.- Name of the dataset

OWL ontologies.

4.- Origin of the dataset

This dataset will be developed by the consortium and may reuse content from existing ontologies.

5.- Dataset purpose: what is the aim of this dataset? What is it going to be used for?

The purpose is to represent the common semantic framework for the project, which consists on Clinical Process Patterns which will serve to standardize the specification of particular content related to clinical process models.

6.- Dataset format: what is the expected format of the dataset? (File formats, relational databases, etc)

OWL.

7.- Data "shape": What are the data types of the dataset? (e.g., data types of the columns on a spreadsheet)

There is no initial limitation on the data types if can be expressed in OWL and are useful for clinical process models.

8.- Data volume: what is the expected size of the dataset?

Around 10 files.

9.- Updates: what is the expected update frequency for the dataset?

The frequency will be variable during the project.

10.- Is the dataset already published following FAIR principles? Where?

No.

11.- Does the dataset reuse other datasets?

Probably.

12.- Does the dataset include rich metadata? In which form?

OWL metadata.

13.- Does the dataset follow naming conventions or any other community standards of its domain?

Yes.

14.- Does the dataset include vocabularies, ontologies, or any other standards at metadata or data level? Which ones?

Yes, but all of them cannot be specified at this time.

15.- Does the dataset include new ontologies developed specifically for it? Where are they published?

They are ontologies themselves, that will be published using the project tools.

16.- Is the dataset properly versioned?

Not yet versioned.

17.- Does the dataset include an explicit and machine-readable license?

Not yet.

18.- Are there any special methods or software needed to access the data? Which ones?

Standard methods for OWL content.

19.- Is there an embargo period in which the dataset, or its metadata, cannot be published? (For example, if you prefer to wait for manuscript acceptance in a journal)

No.

20.- Do you agree to publish the dataset and/or its metadata following FAIR principles? Specify any limitations to publish the datasets and associated metadata, detailing any alternative access methods. For example, in a clinical dataset it might be impossible to publish the data to respect patients' privacy, but the method to access it is still valuable (For example the email of the person in charge of the committee to access the data might be provided as basic metadata)

Yes.

21.- Is the dataset protected by any regulation or special legislation related to personal data? Which one?

No.

22.- Is the dataset going to be published in a specialised public repository (e.g., like UniProt - https://www.uniprot.org/-)? Which one? When is the expected publication date?

Not clear if there is an appropriate resource for this kind of content.

23.- I commit to post the dataset (or its metadata, in case of private data) and its updates to the shared resource that will be set up to that end (FTP server, OneDrive shared folder, etc.). (Name and email)

Yes.

24.- Has any anonymisation been applied to the dataset to preserve patient privacy?

Not needed.

O2.2 Integration framework for process mining models and clinical guideline models

1.- What objective of the project are you the leader of?

O2.2 Integration framework for process mining models and clinical guideline models

2.- Name, institution, and email of the person responsible (Objective leader)

Universitat Jaume I, marcos@uji.es

3.- Name of the dataset

Clinical Guideline Models

4.- Origin of the dataset

manual modelling

5.- Dataset purpose: what is the aim of this dataset? What is it going to be used for?

validation framework

6.- Dataset format: what is the expected format of the dataset? (File formats, relational databases, etc)

PROforma, TP-VML, other file formats

7.- Data "shape": What are the data types of the dataset? (e.g., data types of the columns on a spreadsheet)

n/a

8.- Data volume: what is the expected size of the dataset?

low

9.- Updates: what is the expected update frequency for the dataset?

medium

10.- Is the dataset already published following FAIR principles? Where?

no

11.- Does the dataset reuse other datasets?

no

12.- Does the dataset include rich metadata? In which form?

if needed, a medical terminology/ontology code

13.- Does the dataset follow naming conventions or any other community standards of its domain?

yes, own conventions

14.- Does the dataset include vocabularies, ontologies, or any other standards at metadata or data level? Which ones?

no

15.- Does the dataset include new ontologies developed specifically for it? Where are they published?

no

16.- Is the dataset properly versioned?

yes

17.- Does the dataset include an explicit and machine-readable license?

no, only as comments

18.- Are there any special methods or software needed to access the data? Which ones?

yes, PROforma, TP-VML, other tools

19.- Is there an embargo period in which the dataset, or its metadata, cannot be published? (For example, if you prefer to wait for manuscript acceptance in a journal)

yes, 5 years

20.- Do you agree to publish the dataset and/or its metadata following FAIR principles? Specify any limitations to publish the datasets and associated metadata, detailing any alternative access methods. For example, in a clinical dataset it might be impossible to publish the data to respect patients' privacy, but the method to access it is still valuable (For example the email of the person in charge of the committee to access the data might be provided as basic metadata)

yes

21.- Is the dataset protected by any regulation or special legislation related to personal data? Which one?

n/a

22.- Is the dataset going to be published in a specialised public repository (e.g., like UniProt - https://www.uniprot.org/-)? Which one? When is the expected publication date?

no

23.- I commit to post the dataset (or its metadata, in case of private data) and its updates to the shared resource that will be set up to that end (FTP server, OneDrive shared folder, etc.). (Name and email)

yes

24.- Has any anonymisation been applied to the dataset to preserve patient privacy?

n/a

O1.4 Knowledge graph methods for analysing clinical process models

1.- What objective of the project are you the leader of?

O1.4 Knowledge graph methods for analysing clinical process models

2.- Name, institution, and email of the person responsible (Objective leader)

Manuel Quesada Martínez, Universidad de Murcia

3.- Name of the dataset

CPM-KG metrics

4.- Origin of the dataset

Developed in the project

5.- Dataset purpose: what is the aim of this dataset? What is it going to be used for?

This dataset will include the quality metrics associated with knowledge graphs, which will help to characterise the knowledge graphs

6.- Dataset format: what is the expected format of the dataset? (File formats, relational databases, etc)

RDF

7.- Data "shape": What are the data types of the dataset? (e.g., data types of the columns on a spreadsheet)

Strings, URI, floats

8.- Data volume: what is the expected size of the dataset?

At least 20 values per knowledge graph. It will depend on the number of metrics and graphs

9.- Updates: what is the expected update frequency for the dataset?

Every 6 months

10.- Is the dataset already published following FAIR principles? Where?

No

11.- Does the dataset reuse other datasets?

No

12.- Does the dataset include rich metadata? In which form?

Metadata for describing the metrics will have to be developed, such as an ontology for describing the metrics

13.- Does the dataset follow naming conventions or any other community standards of its domain?

Not aware of any community convention or standard in this area

14.- Does the dataset include vocabularies, ontologies, or any other standards at metadata or data level? Which ones?

Not aware of any community convention or standard in this area

15.- Does the dataset include new ontologies developed specifically for it? Where are they published?

If so, they will be published in the project repository

16.- Is the dataset properly versioned?

The dataset does not exist yet

17.- Does the dataset include an explicit and machine-readable license?

The dataset does not exist yet

18.- Are there any special methods or software needed to access the data? Which ones?

The dataset does not exist yet, but special methods are not expected to be needed

19.- Is there an embargo period in which the dataset, or its metadata, cannot be published? (For example, if you prefer to wait for manuscript acceptance in a journal)

No

20.- Do you agree to publish the dataset and/or its metadata following FAIR principles? Specify any limitations to publish the datasets and associated metadata, detailing any alternative access methods. For example, in a clinical dataset it might be impossible to publish the data to respect patients' privacy, but the method to access it is still valuable (For example the email of the person in charge of the committee to access the data might be provided as basic metadata)

Yes

21.- Is the dataset protected by any regulation or special legislation related to personal data? Which one?

No

22.- Is the dataset going to be published in a specialised public repository (e.g., like UniProt - https://www.uniprot.org/-)? Which one? When is the expected publication date?

No

23.- I commit to post the dataset (or its metadata, in case of private data) and its updates to the shared resource that will be set up to that end (FTP server, OneDrive shared folder, etc.). (Name and email)

Yes

24.- Has any anonymisation been applied to the dataset to preserve patient privacy?

Not applicable

O1.4 Knowledge graph methods for analysing clinical process models (II)

1.- What objective of the project are you the leader of?

O1.4 Knowledge graph methods for analysing clinical process models

2.- Name, institution, and email of the person responsible (Objective leader)

Manuel Quesada Martínez (Universidad Miguel Hernández) mquesada@umh.es

3.- Name of the dataset

CPM-KG Alignments

4.- Origin of the dataset

Developed by the project

5.- Dataset purpose: what is the aim of this dataset? What is it going to be used for?

Representing links between the content of different knowledge graphs

6.- Dataset format: what is the expected format of the dataset? (File formats, relational databases, etc)

RDF

7.- Data "shape": What are the data types of the dataset? (e.g., data types of the columns on a spreadsheet)

URI

8.- Data volume: what is the expected size of the dataset?

Hundreds of mappings

9.- Updates: what is the expected update frequency for the dataset?

Every 6 months

10.- Is the dataset already published following FAIR principles? Where?

Not yet

11.- Does the dataset reuse other datasets?

The dataset of knowledge graphs of the project

12.- Does the dataset include rich metadata? In which form?

It will, in RDF

13.- Does the dataset follow naming conventions or any other community standards of its domain?

The use of standards mapping representation approaches is expected

14.- Does the dataset include vocabularies, ontologies, or any other standards at metadata or data level? Which ones?

To be determined

15.- Does the dataset include new ontologies developed specifically for it? Where are they published?

Not expected

16.- Is the dataset properly versioned?

Versioning policy to be determined

17.- Does the dataset include an explicit and machine-readable license?

License to be determined

18.- Are there any special methods or software needed to access the data? Which ones?

Not expected

19.- Is there an embargo period in which the dataset, or its metadata, cannot be published? (For example, if you prefer to wait for manuscript acceptance in a journal)

No

20.- Do you agree to publish the dataset and/or its metadata following FAIR principles? Specify any limitations to publish the datasets and associated metadata, detailing any alternative access methods. For example, in a clinical dataset it might be impossible to publish the data to respect patients' privacy, but the method to access it is still valuable (For example the email of the person in charge of the committee to access the data might be provided as basic metadata)

Yes

21.- Is the dataset protected by any regulation or special legislation related to personal data? Which one?

No

22.- Is the dataset going to be published in a specialised public repository (e.g., like UniProt - https://www.uniprot.org/-)? Which one? When is the expected publication date?

No

23.- I commit to post the dataset (or its metadata, in case of private data) and its updates to the shared resource that will be set up to that end (FTP server, OneDrive shared folder, etc.). (Name and email)

Yes

24.- Has any anonymisation been applied to the dataset to preserve patient privacy?

Not applicable

O2.1 New process mining methods and metaphors for the discovery of clinical processes from EHR data

1.- What objective of the project are you the leader of?

O2.1 New process mining methods and metaphors for the discovery of clinical processes from EHR data

2.- Name, institution, and email of the person responsible (Objective leader)

Carlos Fernandez Llatas, UPV, cfllatas@itaca.upv.es

3.- Name of the dataset

Methaphors Set

4.- Origin of the dataset

Manual Modeling

5.- Dataset purpose: what is the aim of this dataset? What is it going to be used for?

Graphical view of semantic nodes

6.- Dataset format: what is the expected format of the dataset? (File formats, relational databases, etc)

Vectorial Images

7.- Data "shape": What are the data types of the dataset? (e.g., data types of the columns on a spreadsheet)

Images

8.- Data volume: what is the expected size of the dataset?

low

9.- Updates: what is the expected update frequency for the dataset?

low

10.- Is the dataset already published following FAIR principles? Where?

no

11.- Does the dataset reuse other datasets?

no

12.- Does the dataset include rich metadata? In which form?

no

13.- Does the dataset follow naming conventions or any other community standards of its domain?

own conventions

14.- Does the dataset include vocabularies, ontologies, or any other standards at metadata or data level? Which ones?

no

15.- Does the dataset include new ontologies developed specifically for it? Where are they published?

no

16.- Is the dataset properly versioned?

yes

17.- Does the dataset include an explicit and machine-readable license?

no

18.- Are there any special methods or software needed to access the data? Which ones?

PMApp

19.- Is there an embargo period in which the dataset, or its metadata, cannot be published? (For example, if you prefer to wait for manuscript acceptance in a journal)

yes, until published

20.- Do you agree to publish the dataset and/or its metadata following FAIR principles? Specify any limitations to publish the datasets and associated metadata, detailing any alternative access methods. For example, in a clinical dataset it might be impossible to publish the data to respect patients' privacy, but the method to access it is still valuable (For example the email of the person in charge of the committee to access the data might be provided as basic metadata)

yes

21.- Is the dataset protected by any regulation or special legislation related to personal data? Which one?

no

22.- Is the dataset going to be published in a specialised public repository (e.g., like UniProt - https://www.uniprot.org/-)? Which one? When is the expected publication date?

no

23.- I commit to post the dataset (or its metadata, in case of private data) and its updates to the shared resource that will be set up to that end (FTP server, OneDrive shared folder, etc.). (Name and email)

yes

24.- Has any anonymisation been applied to the dataset to preserve patient privacy?

N/A

O2.3 Interactive process mining methods based on integrated process mining models and clinical guideline models

1.- What objective of the project are you the leader of?

O2.3 Interactive process mining methods based on integrated process mining models and clinical guideline models

2.- Name, institution, and email of the person responsible (Objective leader)

Carlos Fernandez-llatas, UPV, cfllatas@itaca.upv.es

3.- Name of the dataset

Clinical Process Circuits

4.- Origin of the dataset

Manual Modeling

5.- Dataset purpose: what is the aim of this dataset? What is it going to be used for?

Formalization of declarative methods for filtering

6.- Dataset format: what is the expected format of the dataset? (File formats, relational databases, etc)

Own language

7.- Data "shape": What are the data types of the dataset? (e.g., data types of the columns on a spreadsheet)

Compiler

8.- Data volume: what is the expected size of the dataset?

low

9.- Updates: what is the expected update frequency for the dataset?

low

10.- Is the dataset already published following FAIR principles? Where?

no

11.- Does the dataset reuse other datasets?

no

12.- Does the dataset include rich metadata? In which form?

no

13.- Does the dataset follow naming conventions or any other community standards of its domain?

Own Standards

14.- Does the dataset include vocabularies, ontologies, or any other standards at metadata or data level? Which ones?

no

15.- Does the dataset include new ontologies developed specifically for it? Where are they published?

no

16.- Is the dataset properly versioned?

yes

17.- Does the dataset include an explicit and machine-readable license?

yes, is a compliler machine readable

18.- Are there any special methods or software needed to access the data? Which ones?

is deployed trough dll

19.- Is there an embargo period in which the dataset, or its metadata, cannot be published? (For example, if you prefer to wait for manuscript acceptance in a journal)

Until publication

20.- Do you agree to publish the dataset and/or its metadata following FAIR principles? Specify any limitations to publish the datasets and associated metadata, detailing any alternative access methods. For example, in a clinical dataset it might be impossible to publish the data to respect patients' privacy, but the method to access it is still valuable (For example the email of the person in charge of the committee to access the data might be provided as basic metadata)

yes

21.- Is the dataset protected by any regulation or special legislation related to personal data? Which one?

no

22.- Is the dataset going to be published in a specialised public repository (e.g., like UniProt - https://www.uniprot.org/-)? Which one? When is the expected publication date?

no

23.- I commit to post the dataset (or its metadata, in case of private data) and its updates to the shared resource that will be set up to that end (FTP server, OneDrive shared folder, etc.). (Name and email)

n/a

24.- Has any anonymisation been applied to the dataset to preserve patient privacy?

n/a

O2.3 Interactive process mining methods based on integrated process mining models and clinical guideline models (II)

1.- What objective of the project are you the leader of?

O2.3 Interactive process mining methods based on integrated process mining models and clinical guideline models

2.- Name, institution, and email of the person responsible (Objective leader)

Carlos Fernandez llatas, UPV, cfllatas@itaca.upv.es

3.- Name of the dataset

Timed Proces Automaton Model

4.- Origin of the dataset

Format of the Process Discovery records created by PMApp

5.- Dataset purpose: what is the aim of this dataset? What is it going to be used for?

Define the format of the data

6.- Dataset format: what is the expected format of the dataset? (File formats, relational databases, etc)

Class

7.- Data "shape": What are the data types of the dataset? (e.g., data types of the columns on a spreadsheet)

n/a

8.- Data volume: what is the expected size of the dataset?

low

9.- Updates: what is the expected update frequency for the dataset?

low

10.- Is the dataset already published following FAIR principles? Where?

Formalims published https://ieeexplore.ieee.org/document/5961241 Code not published

11.- Does the dataset reuse other datasets?

No

12.- Does the dataset include rich metadata? In which form?

Yes, URIs and values

13.- Does the dataset follow naming conventions or any other community standards of its domain?

Own conventions

14.- Does the dataset include vocabularies, ontologies, or any other standards at metadata or data level? Which ones?

If Needed URIs to other standards

15.- Does the dataset include new ontologies developed specifically for it? Where are they published?

not decided yet

16.- Is the dataset properly versioned?

yes

17.- Does the dataset include an explicit and machine-readable license?

License to be determined

18.- Are there any special methods or software needed to access the data? Which ones?

deployed as DLL

19.- Is there an embargo period in which the dataset, or its metadata, cannot be published? (For example, if you prefer to wait for manuscript acceptance in a journal)

Already published

20.- Do you agree to publish the dataset and/or its metadata following FAIR principles? Specify any limitations to publish the datasets and associated metadata, detailing any alternative access methods. For example, in a clinical dataset it might be impossible to publish the data to respect patients' privacy, but the method to access it is still valuable (For example the email of the person in charge of the committee to access the data might be provided as basic metadata)

No

21.- Is the dataset protected by any regulation or special legislation related to personal data? Which one?

No

22.- Is the dataset going to be published in a specialised public repository (e.g., like UniProt - https://www.uniprot.org/-)? Which one? When is the expected publication date?

No

23.- I commit to post the dataset (or its metadata, in case of private data) and its updates to the shared resource that will be set up to that end (FTP server, OneDrive shared folder, etc.). (Name and email)

N/A

24.- Has any anonymisation been applied to the dataset to preserve patient privacy?

N/A

O3.2 Interactive process mining for the discovery of clinical processes in selected scenarios

1.- What objective of the project are you the leader of?

O3.2 Interactive process mining for the discovery of clinical processes in selected scenarios

2.- Name, institution, and email of the person responsible (Objective leader)

Vicente Traver, UPV, vtraver@itaca.upv.es

3.- Name of the dataset

Clinical Process Circuits of Pilots IPIs

4.- Origin of the dataset

Definition of professionals

5.- Dataset purpose: what is the aim of this dataset? What is it going to be used for?

Definition of Clinical Process Circuits that are defines for each problem

6.- Dataset format: what is the expected format of the dataset? (File formats, relational databases, etc)

Plain text

7.- Data "shape": What are the data types of the dataset? (e.g., data types of the columns on a spreadsheet)

Declarative Sentences

8.- Data volume: what is the expected size of the dataset?

Low

9.- Updates: what is the expected update frequency for the dataset?

Medium

10.- Is the dataset already published following FAIR principles? Where?

No

11.- Does the dataset reuse other datasets?

no

12.- Does the dataset include rich metadata? In which form?

no

13.- Does the dataset follow naming conventions or any other community standards of its domain?

own standars

14.- Does the dataset include vocabularies, ontologies, or any other standards at metadata or data level? Which ones?

No

15.- Does the dataset include new ontologies developed specifically for it? Where are they published?

no

16.- Is the dataset properly versioned?

yes

17.- Does the dataset include an explicit and machine-readable license?

no

18.- Are there any special methods or software needed to access the data? Which ones?

Compiler created on O2.3

19.- Is there an embargo period in which the dataset, or its metadata, cannot be published? (For example, if you prefer to wait for manuscript acceptance in a journal)

yes, until published

20.- Do you agree to publish the dataset and/or its metadata following FAIR principles? Specify any limitations to publish the datasets and associated metadata, detailing any alternative access methods. For example, in a clinical dataset it might be impossible to publish the data to respect patients' privacy, but the method to access it is still valuable (For example the email of the person in charge of the committee to access the data might be provided as basic metadata)

yes

21.- Is the dataset protected by any regulation or special legislation related to personal data? Which one?

No

22.- Is the dataset going to be published in a specialised public repository (e.g., like UniProt - https://www.uniprot.org/-)? Which one? When is the expected publication date?

No

23.- I commit to post the dataset (or its metadata, in case of private data) and its updates to the shared resource that will be set up to that end (FTP server, OneDrive shared folder, etc.). (Name and email)

24.- Has any anonymisation been applied to the dataset to preserve patient privacy?

No

O3.2 Interactive process mining for the discovery of clinical processes in selected scenarios (II)

1.- What objective of the project are you the leader of?

O3.2 Interactive process mining for the discovery of clinical processes in selected scenarios

2.- Name, institution, and email of the person responsible (Objective leader)

Vicente Traver, UPV, vtraver@itaca.upv.es

3.- Name of the dataset

Process Mining Models of pilots

4.- Origin of the dataset

Result of process Discover over the data

5.- Dataset purpose: what is the aim of this dataset? What is it going to be used for?

Show the data to the expert

6.- Dataset format: what is the expected format of the dataset? (File formats, relational databases, etc)

JSON

7.- Data "shape": What are the data types of the dataset? (e.g., data types of the columns on a spreadsheet)

JSON

8.- Data volume: what is the expected size of the dataset?

Big

9.- Updates: what is the expected update frequency for the dataset?

Depend on the problem, even daily

10.- Is the dataset already published following FAIR principles? Where?

No

11.- Does the dataset reuse other datasets?

No

12.- Does the dataset include rich metadata? In which form?

Yes, URIs and values

13.- Does the dataset follow naming conventions or any other community standards of its domain?

not decided

14.- Does the dataset include vocabularies, ontologies, or any other standards at metadata or data level? Which ones?

not decided

15.- Does the dataset include new ontologies developed specifically for it? Where are they published?

not decided

16.- Is the dataset properly versioned?

yes

17.- Does the dataset include an explicit and machine-readable license?

no

18.- Are there any special methods or software needed to access the data? Which ones?

PMApp

19.- Is there an embargo period in which the dataset, or its metadata, cannot be published? (For example, if you prefer to wait for manuscript acceptance in a journal)

Yes it cant be published has sensible data

20.- Do you agree to publish the dataset and/or its metadata following FAIR principles? Specify any limitations to publish the datasets and associated metadata, detailing any alternative access methods. For example, in a clinical dataset it might be impossible to publish the data to respect patients' privacy, but the method to access it is still valuable (For example the email of the person in charge of the committee to access the data might be provided as basic metadata)

No, it can't be published

21.- Is the dataset protected by any regulation or special legislation related to personal data? Which one?

Yes LOPD

22.- Is the dataset going to be published in a specialised public repository (e.g., like UniProt - https://www.uniprot.org/-)? Which one? When is the expected publication date?

No

23.- I commit to post the dataset (or its metadata, in case of private data) and its updates to the shared resource that will be set up to that end (FTP server, OneDrive shared folder, etc.). (Name and email)

24.- Has any anonymisation been applied to the dataset to preserve patient privacy?

no, it will be not published

O1.5 Exploitation of knowledge models for the explainability of process mining models

1.- What objective of the project are you the leader of?

O1.5 Exploitation of knowledge models for the explainability of process mining models

2.- Name, institution, and email of the person responsible (Objective leader)

Begoña Martínez Salvador

3.- Name of the dataset

Clinical Process Knowledge Graph Annotations

4.- Origin of the dataset

Semi-automatic process developed in the project

5.- Dataset purpose: what is the aim of this dataset? What is it going to be used for?

Explainability of the clinical process models

6.- Dataset format: what is the expected format of the dataset? (File formats, relational databases, etc)

RDF

7.- Data "shape": What are the data types of the dataset? (e.g., data types of the columns on a spreadsheet)

URI, String

8.- Data volume: what is the expected size of the dataset?

Medium

9.- Updates: what is the expected update frequency for the dataset?

Every 6 months

10.- Is the dataset already published following FAIR principles? Where?

Yes, expected

11.- Does the dataset reuse other datasets?

All datasets related to Clinical Process Knowledge Graphs

12.- Does the dataset include rich metadata? In which form?

Yes

13.- Does the dataset follow naming conventions or any other community standards of its domain?

Explainable AI conventions, if they exist

14.- Does the dataset include vocabularies, ontologies, or any other standards at metadata or data level? Which ones?

Biomedical ontologies and terminologies

15.- Does the dataset include new ontologies developed specifically for it? Where are they published?

To be determined

16.- Is the dataset properly versioned?

Yes

17.- Does the dataset include an explicit and machine-readable license?

No

18.- Are there any special methods or software needed to access the data? Which ones?

Specific editors for RDF data

19.- Is there an embargo period in which the dataset, or its metadata, cannot be published? (For example, if you prefer to wait for manuscript acceptance in a journal)

No

20.- Do you agree to publish the dataset and/or its metadata following FAIR principles? Specify any limitations to publish the datasets and associated metadata, detailing any alternative access methods. For example, in a clinical dataset it might be impossible to publish the data to respect patients' privacy, but the method to access it is still valuable (For example the email of the person in charge of the committee to access the data might be provided as basic metadata)

Yes

21.- Is the dataset protected by any regulation or special legislation related to personal data? Which one?

No

22.- Is the dataset going to be published in a specialised public repository (e.g., like UniProt - https://www.uniprot.org/-)? Which one? When is the expected publication date?

Yes

23.- I commit to post the dataset (or its metadata, in case of private data) and its updates to the shared resource that will be set up to that end (FTP server, OneDrive shared folder, etc.). (Name and email)

Yes

24.- Has any anonymisation been applied to the dataset to preserve patient privacy?

N/A

O1.5 Exploitation of knowledge models for the explainability of process mining models (II)

1.- What objective of the project are you the leader of?

O1.5 Exploitation of knowledge models for the explainability of process mining models

2.- Name, institution, and email of the person responsible (Objective leader)

Begoña Martínez-Salvador

3.- Name of the dataset

Clinical Guideline Models - Process Mining Models alignments

4.- Origin of the dataset

Manual modelling

5.- Dataset purpose: what is the aim of this dataset? What is it going to be used for?

Explainability of clinical process mining models

6.- Dataset format: what is the expected format of the dataset? (File formats, relational databases, etc)

To be determined

7.- Data "shape": What are the data types of the dataset? (e.g., data types of the columns on a spreadsheet)

To be determined

8.- Data volume: what is the expected size of the dataset?

Medium

9.- Updates: what is the expected update frequency for the dataset?

Medium

10.- Is the dataset already published following FAIR principles? Where?

No

11.- Does the dataset reuse other datasets?

Yes. Clinical guideline models, process mining models

12.- Does the dataset include rich metadata? In which form?

Yes

13.- Does the dataset follow naming conventions or any other community standards of its domain?

Probably, own conventions

14.- Does the dataset include vocabularies, ontologies, or any other standards at metadata or data level? Which ones?

Probably, clinical terminologies

15.- Does the dataset include new ontologies developed specifically for it? Where are they published?

No

16.- Is the dataset properly versioned?

Informally

17.- Does the dataset include an explicit and machine-readable license?

No

18.- Are there any special methods or software needed to access the data? Which ones?

yes, own visualization tool

19.- Is there an embargo period in which the dataset, or its metadata, cannot be published? (For example, if you prefer to wait for manuscript acceptance in a journal)

yes, 5 years

20.- Do you agree to publish the dataset and/or its metadata following FAIR principles? Specify any limitations to publish the datasets and associated metadata, detailing any alternative access

methods. For example, in a clinical dataset it might be impossible to publish the data to respect patients' privacy, but the method to access it is still valuable (For example the email of the person in charge of the committee to access the data might be provided as basic metadata)

Yes

21.- Is the dataset protected by any regulation or special legislation related to personal data? Which one?

N/A

22.- Is the dataset going to be published in a specialised public repository (e.g., like UniProt -https://www.uniprot.org/-)? Which one? When is the expected publication date?

No

23.- I commit to post the dataset (or its metadata, in case of private data) and its updates to the shared resource that will be set up to that end (FTP server, OneDrive shared folder, etc.). (Name and email)

Yes

24.- Has any anonymisation been applied to the dataset to preserve patient privacy?

N/A

O3.3 Development of knowledge models for selected scenarios

1.- What objective of the project are you the leader of?

O3.3 Development of knowledge models for selected scenarios

2.- Name, institution, and email of the person responsible (Objective leader)

Jesualdo Tomás Fernández Breis, Universidad de Murcia, jfernand@um.es

3.- Name of the dataset

Use case scenario

4.- Origin of the dataset

Developed in the project

5.- Dataset purpose: what is the aim of this dataset? What is it going to be used for?

Validation and demonstration of the project

6.- Dataset format: what is the expected format of the dataset? (File formats, relational databases, etc)

Formats in which Clinical Process Models Knowledge Graphs (including alignments and annotations), Process Mining Models and Clinical Guideline Models will be represented

7.- Data "shape": What are the data types of the dataset? (e.g., data types of the columns on a spreadsheet)

Datatypes associated with datasets:Clinical Process Models Knowledge Graphs (including alignments and annotations), Process Mining Models and Clinical Guideline Models

8.- Data volume: what is the expected size of the dataset?

Medium

9.- Updates: what is the expected update frequency for the dataset?

Every year

10.- Is the dataset already published following FAIR principles? Where?

This will depend on each subdataset

11.- Does the dataset reuse other datasets?

This will depend on each subdataset

12.- Does the dataset include rich metadata? In which form?

This will depend on each subdataset

13.- Does the dataset follow naming conventions or any other community standards of its domain?

This will depend on each subdataset

14.- Does the dataset include vocabularies, ontologies, or any other standards at metadata or data level? Which ones?

This will depend on each subdataset

15.- Does the dataset include new ontologies developed specifically for it? Where are they published?

This will depend on each subdataset

16.- Is the dataset properly versioned?

The versioning policy will be determined

17.- Does the dataset include an explicit and machine-readable license?

The license will be determined

18.- Are there any special methods or software needed to access the data? Which ones?

This will depend on each subdataset

19.- Is there an embargo period in which the dataset, or its metadata, cannot be published? (For example, if you prefer to wait for manuscript acceptance in a journal)

This will depend on each subdataset

20.- Do you agree to publish the dataset and/or its metadata following FAIR principles? Specify any limitations to publish the datasets and associated metadata, detailing any alternative access methods. For example, in a clinical dataset it might be impossible to publish the data to respect patients' privacy, but the method to access it is still valuable (For example the email of the person in charge of the committee to access the data might be provided as basic metadata)

Yes

21.- Is the dataset protected by any regulation or special legislation related to personal data? Which one?

This will depend on each subdataset

22.- Is the dataset going to be published in a specialised public repository (e.g., like UniProt - https://www.uniprot.org/-)? Which one? When is the expected publication date?

No

23.- I commit to post the dataset (or its metadata, in case of private data) and its updates to the shared resource that will be set up to that end (FTP server, OneDrive shared folder, etc.). (Name and email)

Yes

24.- Has any anonymisation been applied to the dataset to preserve patient privacy?

This will depend on each subdataset

4.- Architecture

The datasets of the project will go through a "FAIRification" process in order to be published following FAIR principles. Such process will be inspired by the generic workflow described in (Jacobsen et al., 2020) and implemented in a FAIRification framework. The basic architecture of the FAIRification framework is explained in this section illustrating how the most salient points of the Objective 3.1 of the grant agreement will be realized. The basic structure of the FAIRification framework is illustrated in Figure 1 (More details are provided in the following figures).

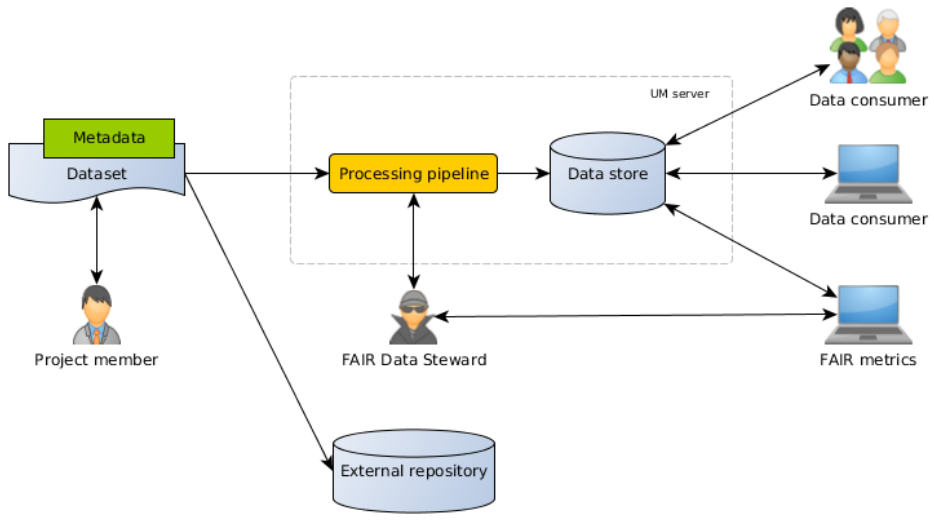

Figure 1: basic architecture of the FAIRification framework.

The process is divided into two main steps:

Processing pipeline: the processing pipeline acquires the data from the project members, processing it, and storing it in the data store. The FAIR Data Stewards oversee the processing pipeline. The

pipeline is deployed in a UM (Universidad de Murcia) server[1]. More details are provided in Section 4.1.

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 Maturity Indicators" achieved consumes data also 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 in Section 4.2.

4.1.- Processing pipeline

The processing pipeline is described in Figure 2. The pipeline comprises the processing of (Linking, quality control, etc.) and the storage of (meta)data. The storage is implemented by GraphDB[2] for RDF based data and CKAN[3] for file-based data.

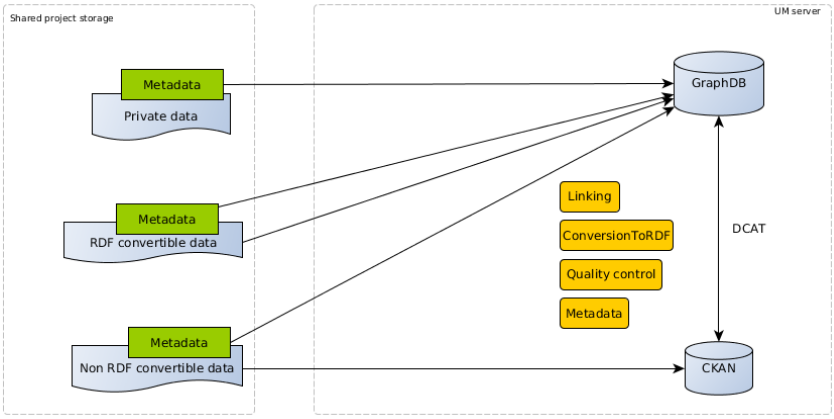

Figure 2: processing pipeline. The data to be published is stored in a shared project storage for FAIR Data Stewards to process (Manually and automatically). After processing the data, it is stored in GraphDB and/or CKAN, depending on its nature.

The pipeline is launched when a project member provides a new dataset, from the ones described in section 3, in the shared project storage (FTP server, Cloud Drive, etc.). There are four types of datasets with regards to their treatment by the processing pipeline:

Private data: clinical data tends to be protected by strict legislation. In this case, since the data cannot be published, minimum 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[4] (Resource Description Framework) or it is feasible to convert to RDF. In this case both data and metadata will be stored in GraphDB.

[1] A server purchase is detailed in the Grant Agreement.
[2] https://graphdb.ontotext.com/
[3] https://ckan.org/
[4] https://www.w3.org/TR/rdf11-concepts/

Public, non RDF convertible data: public data that can be fully published but it is not available in RDF or it is not feasible to convert to RDF. In this case the metadata will be stored in GraphDB, with pointers to a CKAN[5] server, in which the data, in its original form, will be stored. The CKAN DCAT extension[6] 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[7] (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 entail the use of the DCAT[8], VoID[9], PROV[10], and Creative Commons[11] vocabularies, apart from any other vocabularies already present in the datasets, and it will follow the FAIR Data Point metadata specification[12] as a guide (Jacobsen et al., 2020). By using this vocabularies principle I2 will be applied. (*(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[13]) or Java (Using RDF4J[14]).

Quality control: SHACL[15] will be used to ensure the quality of the produced RDF, especially in the case of Metadata. Quality control for data will be limited due to reduced resources.

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[16] platform or manually.

4.2.- Web frontend

The publication frontend will offer the possibility of consuming the stored data through different interfaces, suitable for different clients, as shown in Figure 3. 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 will also ensure the implementation of the principle F1 (*(Meta)data are assigned a globally unique and persistent*

[5] https://ckan.org/
[6] https://extensions.ckan.org/extension/dcat/
[7] https://github.com/common-workflow-language/cwltool
[8] https://www.w3.org/TR/vocab-dcat-2/
[9] https://www.w3.org/TR/void/
[10] https://www.w3.org/TR/prov-o/
[11] https://creativecommons.org/ns
[12] https://specs.fairdatapoint.org/
[13] https://rdflib.dev/
[14] https://rdf4j.org/
[15] https://www.w3.org/TR/shacl/
[16] http://silkframework.org/

identifier) using W3ID identifiers[17]. 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[18] scripts that follow the bio-schema[19] and schema[20] vocabularies, in order to be crawled in a structured way by the most common search engines (Jacobsen et al. 2020).

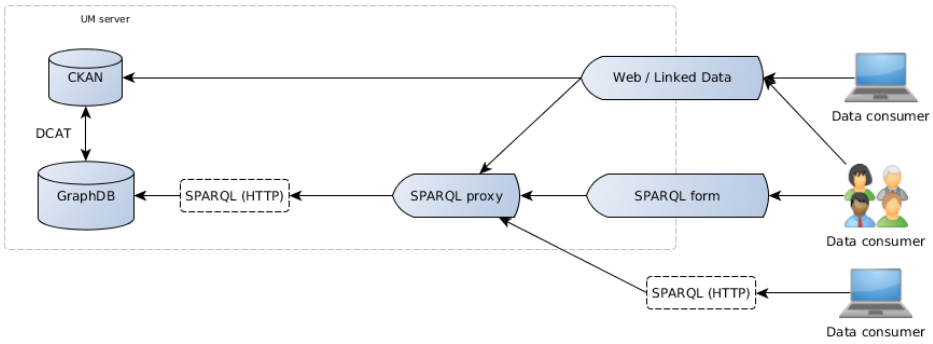

Figure 3: publication frontend. The publication frontend offers the stored data through different interfaces for humans but specially for machines.

The frontend comprises the following elements:

SPARQL proxy: it redirects any SPARQL queries to GraphDB with the appropriate security settings. It will be created for the project.

SPARQL form: it offers a human friendly interface to pose SPARQL queries. It will be based on YASGUI[21].

Web / Linked Data server: it will process direct web calls producing a redirection to CKAN or a Linked Data item request (SPARQL DESCRIBE query), as appropriate. It will be created for the project based on existing tools like Trifid[22] or AtomGraph Processor[23].

Apart from the FAIR publication frontend described in Figure 3, a static web page will be published 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 and benchmarks based on FAIR metrics.

A GitHub project has been set up for the development of the FAIR publication framework[24].

5.- Conclusion

This document constitutes the Data Management Plan for the Support4LHS project. During the three years of the project this DMP will guide the publication of the project datasets according to FAIR principles, collecting any changes to the procedures and decisions. The result of the application of FAIR principles will be automatically and rigorously evaluated.

References

[17] https://w3id.org/
[18] https://www.w3.org/TR/json-ld11/
[19] https://bioschemas.org/
[20] https://schema.org/
[21] https://triply.cc/docs/yasgui
[22] https://github.com/zazuko/trifid
[23] https://github.com/AtomGraph/Processor
[24] https://github.com/mikel-egana-aranguren/SUPPORT4LHS-FAIR-data

Jacobsen, A., Kaliyaperumal, R., da Silva Santos, Luiz Olavo Bonino, Mons, B., Schultes, E., Roos, M., & Thompson, M. (2020). A generic workflow for the data FAIRification process. *Data Intelligence, 2*(1-2), 56-65. doi:10.1162/dint_a_00028

Maldonado, J. A., Marcos, M., Fernández-Breis, J. T., Giménez-Solano, V. M., Legaz-García, M. d. C., & Martínez-Salvador, B. (2020). CLIN-IK-LINKS: A platform for the design and execution of clinical data transformation and reasoning workflows. *Computer Methods and Programs in Biomedicine, 197*, 105616. doi:10.1016/j.cmpb.2020.105616

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