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 1](#_Toc1063732952)

[2.- Introduction 1](#_Toc219351545)

[3.- Datasets 2](#_Toc917695749)

[4.- Architecture 2](#_Toc1789177546)

[4.1.- Processing pipeline 3](#_Toc1516624893)

[4.2.- Web frontend 4](#_Toc930121338)

[5.- Conclusion 5](#_Toc1982477233)

[Annex A: Support4LHS FAIR data questionnaire 6](#_Toc1789501579)

[Annex B: FAIR principles 6](#_Toc2026691064)

# 1.- Executive summary

This deliverable is the Data Management Plan (DMP) for the Support4LHS project. Its aim is to collect the descriptions of the datasets of the project and to provide an overview of the technical decisions necessary to publish the datasets following FAIR principles (Findable, Accessible, Interoperable, Reusable). This document is alive and it will be updated through the project life span.

# 2.- Introduction

The Objective 3.1 (*Design and implementation of a FAIR Data Management Plan*) of the project Support4LHS comprises the publication of 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 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, 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, 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.

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

# 3.- Datasets

In order to obtain a preliminary list of possible datasets, an informal enquiry was made, obtaining the results depicted in Table 1.

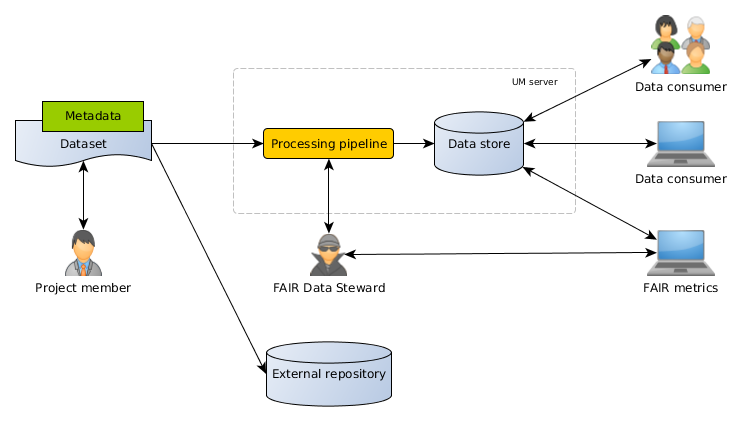
|  |  |
| --- | --- |
| **Objective** | **Datasets** |
| O1 Knowledge-driven methods to support process mining in healthcare |  |
| O1.1 Knowledge and standards-based transformation methods for process mining data ingestion in healthcare | No datasets produced |
| O1.2 Extraction of knowledge graphs 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 clincal 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 objetive.*

The details of the prospective datasets to be FAIRified were obtained through the form described in Annex A, and the answers, comprising the actual catalog of datasets, is described in Annex B.

# 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 and also illustrates 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.*

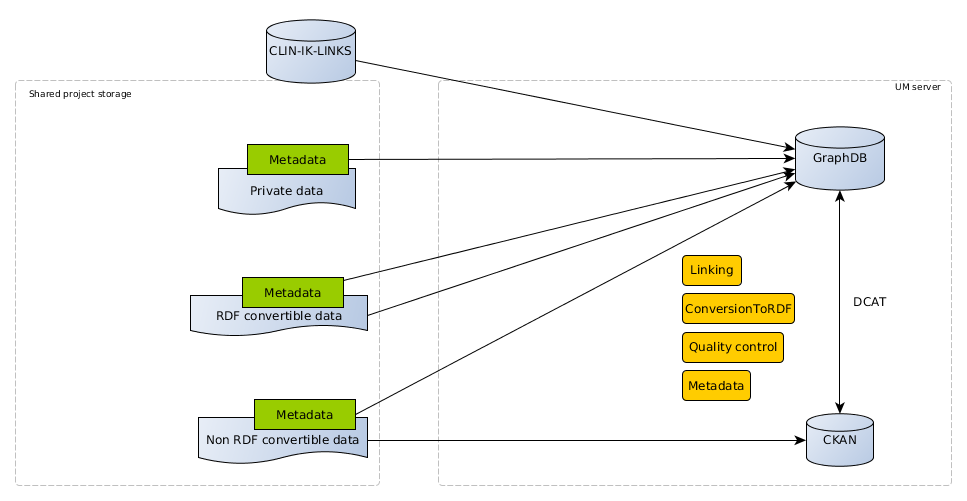
The process is divided into two main steps:

**Processing pipeline:** it acquires the data from the project members, it processes it, and it stores 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]](#footnote-9730). 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 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 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 (meta)data. The storage is implemented by GraphDB[[2]](#footnote-17770) for RDF based data and CKAN[[3]](#footnote-10086) for file-based data.



*Figure 2: processing pipeline. The data to be published is stored in a shared resource (e.g. FTP server) 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 kicks-off 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:

**CLIN-IK-LINKS datasets:** the CLIN-IK-LINKS platform (Maldonado et al., 2020) will store Clinical Process Models (CPMs) produced in the project as Knowledge Graphs (KGs). Since the platform will offer REST APIs to access those elements, a connector will be developed to insert the KGs 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[[4]](#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 not feasible 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[[5]](#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[[6]](#footnote-20674), VoID[[7]](#footnote-16129), PROV[[8]](#footnote-21939), and Creative Commons[[9]](#footnote-12352) vocabularies, apart from any other vocabularies already present in the datasets, and it will follow the FAIR Data Point metadata specification[[10]](#footnote-2507) as a guide (Jacobsen et al., 2020). 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[[11]](#footnote-1606)) or Java (Using RDF4J[[12]](#footnote-31141)).

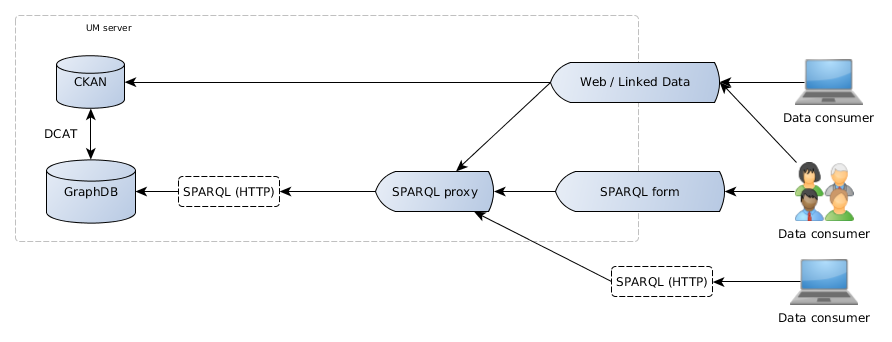
**Quality control:** SHACL[[13]](#footnote-18578) will be used to ensure the quality of the produced RDF, especially in the case of Metadata. Quality control for data will be limited due to reduced resources.

**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[[14]](#footnote-16402) 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 ensure the implementation of the principle F1 (*(Meta)data are assigned a globally unique and persistent identifier*) using W3ID identifiers[[15]](#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 (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[[16]](#footnote-10982).

**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[[17]](#footnote-1368) or AtomGraph Processor[[18]](#footnote-11583).

Apart from the FAIR publication frontend described in Figure 3, a static web will be served 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[[19]](#footnote-7057).

# 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 data according to FAIR principles, collecting any changes to the procedures and decisions described herein.

References

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>

# Annex A: 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?**

# Annex B: Support4LHS FAIR data questionnaire results

**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**

**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?**

**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?**

**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?**

# Annex C: FAIR principles

The GO FAIR initiative[[20]](#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. A server purchase is detailed in the Grant Agreement. [↑](#footnote-ref-9730)
2. https://graphdb.ontotext.com/ [↑](#footnote-ref-17770)
3. https://ckan.org/ [↑](#footnote-ref-10086)
4. https://www.w3.org/TR/rdf11-concepts/ [↑](#footnote-ref-32314)
5. https://github.com/common-workflow-language/cwltool [↑](#footnote-ref-1839)
6. https://www.w3.org/TR/vocab-dcat-2/ [↑](#footnote-ref-20674)
7. https://www.w3.org/TR/void/ [↑](#footnote-ref-16129)
8. https://www.w3.org/TR/prov-o/ [↑](#footnote-ref-21939)
9. https://creativecommons.org/ns [↑](#footnote-ref-12352)
10. https://specs.fairdatapoint.org/ [↑](#footnote-ref-2507)
11. https://rdflib.dev/ [↑](#footnote-ref-1606)
12. https://rdf4j.org/ [↑](#footnote-ref-31141)
13. https://www.w3.org/TR/shacl/ [↑](#footnote-ref-18578)
14. http://silkframework.org/ [↑](#footnote-ref-16402)
15. https://w3id.org/ [↑](#footnote-ref-28839)
16. https://triply.cc/docs/yasgui [↑](#footnote-ref-10982)
17. https://github.com/zazuko/trifid [↑](#footnote-ref-1368)
18. https://github.com/AtomGraph/Processor [↑](#footnote-ref-11583)
19. https://github.com/mikel-egana-aranguren/SUPPORT4LHS-FAIR-data [↑](#footnote-ref-7057)
20. https://www.go-fair.org/fair-principles/ [↑](#footnote-ref-6621)