HEALTHYCLOUD
Health Research & Innovation Cloud

D3.2 Guidelines to standardise metadata templates and assessment of FAIRness maturity levels Version 1.2

Document Information

Contract Number	965345
Project Website	http://www.healthycloud.eu/
Contractual Deadline	M23, January 2022
Dissemination Level	PU - Public
Nature	R - Report
Author(s)	Celia Alvarez-Romero (SAS) Alicia Martínez-García (SAS)
Contributor(s)	Pascal Derycke (Sciensano) Carlos Luis Parra Calderón (SAS) Irene Kesisoglou (Sciensano) Shona Cosgrove (Sciensano)
Reviewer(s)	Laura Portell (BSC) Hanna Tolonen (THL)
Keywords	Health data management, health data infrastructure, health data hub, FAIR principles, FAIRness assessment, metadata.

Notice: *The HealthyCloud project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement Nº965345*

©2021 HealthyCloud Consortium Partners. All rights reserved.

Change Log

Version	Author	Date	Description of Change
v0.1	Celia Alvarez-Romero (SAS)	26.09.2022	Initial Table of Content and first draft.
v0.2	Celia Alvarez-Romero (SAS)	30.09.2022	Methods and Results sections.
v0.3	Celia Alvarez-Romero (SAS) Alicia Martínez-García (SAS)	05.10.2022	Methods and Results sections.
v0.4	Celia Alvarez-Romero (SAS) Alicia Martínez-García (SAS) Pascal Derycke (Sciensano) Irene Kesisoglou (Sciensano) Shona Cosgrove (Sciensano)	16.11.2022	Methods and Results sections.
v0.5	Celia Alvarez-Romero (SAS) Pascal Derycke (Sciensano) Irene Kesisoglou (Sciensano) Shona Cosgrove (Sciensano)	15.12.2022	First content for review.
v0.6	Laura Portell (BSC) Hanna Tolonen (THL)	15.12.2022 - 10.01.2023	Reviews.
v1.0	Celia Alvarez-Romero (SAS) Pascal Derycke (Sciensano) Irene Kesisoglou (Sciensano) Shona Cosgrove (Sciensano)	10.01.2023	Covering reviewers' comments. Second draft sent to HealthyCloud Coordinators.
v1.1	HealthyCloud Coordinators	10.01.2023 - 24.01.2023	Reviews.
v1.2	Celia Alvarez-Romero (SAS)	25.01.2023	Final version submitted.

Table of contents

Executive Summary

The main aim of **HealthyCloud WP3** is to carry out a health data landscape analysis, aiming to capture the European health data collections available for research purposes, evaluate their FAIRness level and determine the feasibility to perform individual level data linkages. Within this work, Task 3.3 focuses on capturing insights from FAIRness maturity evaluation to define metadata templates as functions of the health-related data framework.

Therefore, the result of **deliverable D3.2** is a set of **guidelines on best practices for FAIR-health metadata templates**, comprising the use of interoperable vocabularies and metadata schemes for health-related data, carrying out a comprehensive study of current metadata standards and quality to make data FAIR in the health field.

As conclusions, recommendations for FAIRness assessment are gathered, including tools to carry out an automatic assessment of datasets against the FAIR principles. Besides, we can conclude the use of a descriptive metadata record template compatible with the Data Catalog Vocabulary specifying an Application Profile (DCAT-AP) standard.

The DCAT-AP standard is presented since it allows users to describe their datasets in a consistent and standardised way, making it easier for others to discover and reuse the data. Likewise, DCAT-AP is a flexible and extensible vocabulary, allowing users to define their own terms and properties in addition to the standard terms and properties defined in the vocabulary.

1. Introduction

Improving the **Findability, Accessibility, Interoperability, and Reusability** (FAIR principles) [1] of health data stored in data infrastructures for secondary use and meeting the appropriate requirements is crucial to facilitate and improve research and innovation in the field of health. A recent example would be the COVID-19 crisis, where international cooperation achieved rapid sequencing and epidemiological studies that resulted in the fastest body of published work in history for any disease [2, 3].

Digitalised health data was a prominent but insufficient part of the response to the COVID-19 crisis. The visible rise of telemedicine was a great success story as well, but addressing the pandemic required much more support and a revamping of health data infrastructures. The inadequacies of health data infrastructures, cultural barriers, out-of-date policies and misaligned incentives must be addressed. A key lesson learnt from the COVID-19 pandemic was that accessible, interoperable, and readily analysable health data, and the development of the appropriate architectural framework to support data sharing were essential to support the response to COVID-19 crisis[4].

In this sense, the deliverable **D3.2 'Guidelines to standardise metadata templates and assessment of FAIRness maturity levels'** covers a set of recommendations on the best practices for FAIR-health metadata templates. Besides, D3.2 gathers the main results of HealthyCloud Task 3.3 'Reference guidelines to standardize metadata templates for health-related data for uplifting and creating FAIR data catalogues' where researchers focus on capturing insights from the FAIRness maturity evaluation of data collections carried out in Task 3.2 'Landscape analysis of FAIRness levels of health-related data using a catalogue matrix' to define metadata templates as functions of the health-related data framework (including data quality).

On the one hand, in order to address the D3.2 purpose, the **milestone MS3.3** 'Guidelines: standardised guidelines for FAIRness maturity levels completed' was reached at month M15 (May 2022), providing a set of recommendations that can be applied to assess the FAIRness maturity levels in health data infrastructures, taking into account the specific requirements for health data due to its sensitive nature. That is, the milestone MS3.3 has been the basis for deliverable D3.2.

Finally, previous initiatives on **metadata templates** have been explored to be included as references of the D3.2 guidelines.

2. Methods

The final aim of the related Task 3.3 is to capture insights from FAIRness maturity evaluation to define metadata templates as functions of the health-related data framework. Part of those conclusions are collected in the deliverable D3.2, aiming to provide a set of clear guidelines to standardise metadata templates and the assessment of FAIRness maturity levels.

In this section, references and previous works are described in order to gather the recommendations that are part of the D3.2 guidelines.

2.1. Guidelines to assess the FAIRness maturity levels

Building on the initial work carried out on MS3.3 'Guidelines: standardised guidelines for FAIRness maturity levels completed', we further develop here those aspects to provide a set of clear guidelines to apply the FAIR principles in health data infrastructures at European level.

Firstly, to reach the milestone MS3.3, meetings with experts from other initiatives related to FAIR methods models were organised in WP3 to learn more from them. Concretely, researchers from the Research Data Alliance Working Group (RDA WG) on FAIR Data Maturity Model [5], GO FAIR [6], IMI2 FAIRplus [7], H2020 FAIR4Health [8], Population Health Information Research Infrastructure (PHIRI) Joint Action [9] and the European Joint Programme on Rare Diseases (EJP RD) [10], among others, presented the methods and FAIR metrics applied by them.

Secondly, deliverable D3.2 is based on the project completed for the FAIRplus Fellowship Program [11] where one of the researchers of the SAS team has participated together with another researcher from the BSC team. Specifically, the project included a set of clear guidelines to apply the FAIR principles in health data infrastructures at a European level. This work was presented as part of the MS3.3 report.

An analysis of the available literature, tools already developed and other processes and workflows used in previous initiatives was carried out in order to review the landscape and conclude the recommendations to be applied in the guidelines of the deliverable D3.2.

2.2. FAIR-health metadata templates

A FAIR metadata catalogue allows datasets to be more discoverable via a user-friendly portal or programmatic access mechanisms, e.g. REST API. Concretely, HealthyCloud researchers work in the specifications that the future FAIR data portal should have including the definition of meta-catalogues for the aggregation of FAIRified data collections (deliverable D6.2 'Specifications for the FAIR data portal').

Metadata is a number of descriptors providing the right context to understand the associated data so that it can be discovered, accessed and reused. Metadata can be categorised by administrative metadata (What technical information is linked to the dataset?), descriptive metadata (How to discover, identify and access the dataset?) and structural metadata (How is the dataset organised?).

A metadata standard is a specific guide that supports the harmonised creation of metadata records for datasets. Metadata elements are grouped into sets designed for a specific purpose and given a standard name and definition. Rules on what content must be included, what syntax must be used, or a controlled vocabulary can also be included in a metadata standard. Authors of 'How to FAIR' [12] provide more information on metadata standards and ontologies [13]. Briefly, authors show that FAIR is a set of principles that makes your research efficient, transparent and sustainable, aiming to make data more FAIR to improve research data management and safeguard research data for the future. Besides, to aid other research supporters and research data managers, the authors provide lists of available standards, extensions, tools, and use cases, since using a metadata standard and/or an ontology commonly used is the best practice.

Among the considered projects and initiatives, the ones below are working on different metadata aspects addressing the same objectives as deliverable D3.2. They were analysed and the main references are included below.

- RDA WG on Research Metadata Schemas [14], whose purposes are: "(i) to identify and bridge gaps in existing schemas commonly used for research data, by bringing together communities who are working with such vocabularies to document research data and related resources; and (ii) to provide guidelines for those communities whose needs are not addressed by existing metadata schema such as schema.org, and provide guidelines on proposing extensions."
- RDA Metadata Standards Directory Working Group [15] that "is supported by individuals and organisations involved in the development, implementation, and use of metadata for scientific data. The overriding goal is to develop a collaborative, open directory of metadata standards applicable to scientific data that can help address infrastructure challenges".
- CEDAR Metadata Center [16] that “is making data submission smarter and faster, so that scientific researchers and analysts can create and use better metadata. Through better interfaces, terminology, metadata practices, and analytics, CEDAR improves metadata from provider to end user".
- The HealthData@EU pilot project [17] will extend the descriptive metadata standard for health data based on the Data Catalog Vocabulary, specifying an Application Profile (DCAT-AP). The pilot also aims to build a European metadata-based catalogue to facilitate the discovery of health-related data for secondary use. Concretely, DCAT-AP is a W3C metadata recommendation for publishing data on the Web. DCAT-AP is defined in Resource Description Format (RDF) and reuses the Dublin Core Metadata standard. DCAT-AP is designed to facilitate interoperability between data catalogues published on the Web. By using DCAT-AP to describe datasets in data catalogues, publishers increase discoverability and enable applications to easily consume metadata from multiple catalogues. It further enables decentralised publishing of catalogues and facilitates federated dataset search across sites [18].

3. Results

This section describes the main outcomes after carrying out the methods presented in the previous section.

3.1. Guidelines to assess the FAIRness maturity levels

Below, D3.2 guidelines include recommendations to address the Findability, Accessibility, Interoperability, and Reusability of data, as well as to assess their FAIRness maturity level.

Findability

Findable means that the data can be discovered by both humans and machines, for instance by exposing meaningful machine-actionable metadata and keywords to search engines and research data catalogues. The data are referenced with unique and persistent identifiers, e.g. DOIs or handles, and the metadata include the identifier of the data they describe. Therefore, it is essential that the metadata is well described, including a persistent unique identifier (PID), and the identifier of the data they describe. Also, it is strongly recommended that every health data infrastructures have a metadata catalogue that helps researchers find data. In this sense, metadata standards ensure the correct and proper use and interpretation of data by other researchers, increasing the discoverability of data collections from different sources and federated data searches. For that, a descriptive metadata standard that can be used is DCAT-AP [19, 20], which is designed to facilitate interoperability between data catalogues published.

Likewise, data and metadata should be accessible, registered or indexed in an open registry, such as the ELIXIR core data resources [21] or any other catalogue of repositories suitable for the health research field.

Accessibility

The strong recommendation here is related to data access procedures or protocols defined and publicly accessible. Health data infrastructures have to define the mechanisms and formal procedures to data access and transfer through a secure processing environment. That is, the physical or virtual environment and organisational means to provide the opportunity to re-use data in a manner that allows for the operator of the secure processing environment to determine and supervise all data processing actions, including to display, storage, download, export of the data and calculation of derivative data through computational algorithm [22].

For that, data access procedures or protocols should include: communications protocols, clear data access policies and legal basis, describing the conditions to get access to the data and data sharing agreements. Besides, for the data access, the data access application is reviewed and assessed by a Data Access Committee (DAC) previously defined in the health data infrastructure.

Interoperability

Both secondary use of data and data sharing lead to the need for data to be interoperable to enable data storage, analysis and processing. In this sense, it is worth mentioning the recommendation of using internationally recognised standards, ontologies and vocabularies in the health field to facilitate the health data sharing and usage. For that, the data and metadata representation and meaning have to be formal, clear and unambiguous, and the vocabularies used have to follow FAIR principles. For instance, the Ontology Lookup Service (OLS) [23] is a search engine that facilitates the finding of domain ontologies that best describe specific data.

On the other hand, the Fast Healthcare Interoperability Resources (FHIR) is a next generation standards framework created by HL7, even being mainly an standard for primary use, is also used for secondary use [24]. In addition, HL7 FHIR includes some interoperability with the Observational Medical Outcomes Partnership

(OMOP) Common Data Model (CDM) aims to transform data contained within health databases into a common format and to use common terminologies [25]. OMOP CDM is an open community-driven data model designed to standardise the structure and content of observational data and to enable efficient analyses that can produce reliable evidence. A central component of the OMOP CDM is the Observational Health Data Sciences and Informatics (OHDSI) standardised vocabularies. The OHDSI vocabularies allow organisation and standardisation of medical terms from different data sources to be used across the various clinical domains of the OMOP common data model.

In conclusion, data and metadata representation and standardised vocabularies using a Common Data Model (CDM) ensures meaningful, comparable and reproducible research results. Concretely, standardised metadata is the first step towards interoperability of data because it represents what kind of data exists, how different variables are described and what kind of code systems are used. Then the next step is to standardise the actual data.

Reusability

Data sharing and reuse facilitates the scientific discovery and optimizes the research purposes. Defining a clear and accessible data sharing, reuse and usage licences in a regulatory framework is essential for this objective. For instance, the Creative Commons licences are helpful for that [26, 27].

Therefore, to increase data reuse, it is fundamental to publish metadata in a way that increases data usage, such as the inclusion of the results on publicly accessible and findable repositories (e.g., Zenodo).

Assessment of FAIRness maturity levels

There are multiple resources to assess FAIRness maturity levels, starting with the **HealthyCloud FAIRness assessment tool** [28] developed during Task 3.2. The results from the evaluation using this tool were included in deliverable D3.1 [29]. This FAIRness assessment tool [28] was developed by adapting the already existing ARDC FAIRness evaluation tool [30]. It was then applied using information from the health data collections that answered the WP3-WP4 survey (included as Annex 1 of deliverable D3.1). This tool presents the results of the FAIRness assessment as a percentage score for each letter of FAIR separately, as well as the total value, based on the answers related to the principles underpinning Findable, Accessible, Interoperable and Reusable.

In addition, another set of tools and recommendations to assess the FAIRness maturity levels in health data infrastructures have been analysed and presented below.

On the one hand, there are **tools to carry out an automatic assessment** of datasets against the FAIR principles:

- FAIREvaluator [31] which provides a registry and online execution functions to evaluate resources FAIRness against Collections of Maturity Indicator Tests: Maturity Indicator Tests, Community-defined Collections of Maturity Indicator Tests, and Quantitative FAIRness evaluations of a Resource based on these Collections: https://fairsharing.github.io/FAIR-Evaluator-FrontEnd/#!/
- FAIRshake [32] that provides a catalogue of community-contributed ways to characterise FAIRness. Using FAIRshake (https://fairshake.cloud/), a variety of biomedical digital resources can be manually and automatically evaluated for their level of FAIRness.
- FAIRassist tool [33] (under development: https://fairassist.org/#!/) to help users understand how to achieve a state of "FAIRness", and how this can be measured and improved. The focus is on manual questionnaires, checklists and automated tests that help users understand how to achieve a state of "FAIRness", and how this can be measured and improved.
- Australian Research Data Commons (ARDC)' FAIR data self-assessment tool [30] to perform an online assessment of how FAIR a research dataset is and get practical tips on how to enhance its FAIRness, through answering questions related to the FAIR principles: https://ardc.edu.au/resource/fair-data-self-assessment-tool/

On the other hand, the level of data FAIRness can be measured by **metrics and indicators** related to the conformance of data objects to the FAIR principles. Below essential references are included.

- Indicators related to each specific aspect of FAIR proposed by the Research Data Alliance (RDA) Working Group (WG) on FAIR data maturity model. This RDA WG divides the FAIRness into 6 levels, from 0 (not FAIR at all) to 5 (completely FAIR). It uses arbitrary scoring to measure conformance to the FAIR principles [34].
- Indicators selected and provided by the FAIRplus project [35], aligned to the set of RDA data maturity indicators, generated by community agreement [36, 37].
- Guidance on how to implement responsible and FAIR approaches to research assessment [38] provided by the Wellcome foundation [39]. The central components of the guidance draw on Declaration on Research Assessment (DORA)'s core principles to be explicit about the criteria used to evaluate research productivity and to recognize the value of all relevant research outputs (for example publications, datasets and software), as well as other

types of contributions, such as training early-career researchers and influencing policy and practice.

3.2. Guidelines for FAIR-health metadata templates

As mentioned above, before researchers request access to data, they should be able to know what data is available. So, data infrastructures should provide researchers with what content is available. This is where published and accessible metadata plays an important role in facilitating this purpose. For instance, some relevant descriptive metadata in this regard could be: title, acronym, URL, Digital Object Identifier (DOI), references, medical domain, description, funding, etc.

This section presents a descriptive metadata record template compatible with the DCAT-AP standard since we can state that it is the metadata standard of choice for several projects, including the HealthData@EU [17].

The DCAT-AP is a specification for metadata records based on the Data Catalogue Vocabulary developed by the W3C Government Linked Data Working Group. It provides to data and metadata portals in Europe a semantic interoperability framework on the basis of reuse of established controlled vocabularies (e.g. EuroVoc) and the mappings to existing metadata vocabularies (e.g. StatDCAT, GeoDCAT,...). DCAT-AP allows users to describe their datasets in a consistent and standardised way, making it easier for humans and machines to discover and reuse their data. DCAT-AP is one standard for Linked Data and the Semantic Web providing a universal vocabulary for descriptive metadata of datasets and data services.

The DCAT-AP "dataset" class contains 2 mandatory properties (i.e.: Title and Description) and has 7 recommended and 26 optional properties with various cardinalities. For instance, the property "Has version" from 1 to *n* refers to a related Dataset that is a version, edition, or adaptation of the described Dataset. If the DCAT-AP classes "Catalog" and "Dataset" assure findability and discoverability of data, the recommended DCAT-AP class "Distribution" is the gateway to the description of the dataset at variable level.

DCAT-AP offers the possibility to enrich datasets with linked metadata. Thanks to the DCAT-AP RDF vocabulary, i.e. linked metadata, the content is made machine readable. Table 1 presents the properties of the class dataset, of the version 2.1.0 for the DCAT-AP for data portals in Europe (https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe/releases).

Table 1 shows some examples of information per property are provided, as well as a turtle file (Annex 1 of this document) as it is one of the possible serialisations for exchanging metadata records between data portals (i.e.: harvesting metadata).

For each cell: property name, usage note, [URI, range, cardinality], example.

Title This property contains a name given to the Dataset. This property can be repeated for parallel language versions of the name. [dct:title rdfs:Literal 1..n] **Ex: Edinburgh Ovarian Cancer Database**
Description This property contains a free-text account of the Dataset. This property can be repeated for parallel language versions of the description. [dct:description rdfs:Literal 1..n] **Ex: Established in 2019, the Ovarian Cancer Database builds on 40 years of data collection across the region of the South East Scotland Cancer Network. The database holds data on diagnosis, treatment and outcomes of patients undergoing care within the region.**
Contact point This property contains contact information that can be used for sending comments about the Dataset. [dcat:contactPoint vcard:Kind 0..n] **Ex: https://www.ed.ac.uk/information-services/about/contacting-is**
Dataset distribution This property links the Dataset to an available Distribution. [dcat:distribution dcat:Distribution 0..n]
Keyword/ tag This property contains a keyword or tag describing the Dataset. [dcat:keyword rdfs:Literal 0..n] **Ex: OVARY, OVARIAN, FALLOPIAN TUBE, PERITONEUM, PERITONEAL, CANCER, SARCOMA, CARCINOSARCOMA, ADENOCARCINOMA, SEROUS, ENDOMETRIOID, CLEAR CELL**

Publisher

This property refers to an entity (organisation) responsible for making the Dataset available.

[dct:publisher foaf:Agent 0..1]

Ex: https://dataloch.org/

Spatial/ geographical coverage

This property refers to a geographic region that is covered by the Dataset.

[dct:spatial dct:Location 0..n]

Ex: http://publications.europa.eu/resource/authority/country/GBR

Temporal coverage

This property refers to a temporal period that the Dataset covers.

[dct:temporal dct:PeriodOfTime 0..n]

Theme/ category

This property refers to a category of the Dataset. A Dataset may be associated with multiple themes.

[dcat:theme, subproperty of dct:subject skos:Concept 0..n]

Ex: http://purl.obolibrary.org/obo/MONDO_0008170

Access rights

This property refers to information that indicates whether the Dataset is open data, has access restrictions or is not public.

[dct:accessRights dct:RightsStatement 0..1]

Ex: https://www.wiki.ed.ac.uk/display/CAN/Governance

Creator

This property refers to the entity responsible for producing the dataset.

[dct:creator foaf:Agent 0..n]

Ex: NHS Lothian and the University of Edinburgh

Conforms to

This property refers to an implementing rule or other specification.

[dct:conformsTo dct:Standard 0..n]

Documentation

This property refers to a page or document about this Dataset.

[foaf:page foaf:Document 0..n]

Frequency

This property refers to the frequency at which the Dataset is updated.

[dct:accrualPeriodicity foaf:Document 0..1]

Ex: http://publications.europa.eu/resource/authority/frequency/UPDATE_CONT

Has version

This property refers to a related Dataset that is a version, edition, or adaptation of the described Dataset.

[dct:hasVersion dcat:Dataset 0..n]

Identifier

This property contains the main identifier for the Dataset, e.g. the URI or other unique identifier in the context of the Catalogue.

[dct:identifier rdfs:Literal 0..n]

Is referenced by

This property is about a related resource, such as a publication, that references, cites, or otherwise points to the dataset.

[dct:isReferencedBy rdfs:Resource 0..n]

Is version of

This property refers to a related Dataset of which the described Dataset is a version, edition, or adaptation.

[dct:isVersionOf dcat:Dataset 0..n]

Landing page

This property refers to a web page that provides access to the Dataset, its Distributions and/or additional information. It is intended to point to a landing page at the original data provider, not to a page on a site of a third party, such as an aggregator.

[dcat:landingPage foaf:Document 0..n]

Ex: https://www.wiki.ed.ac.uk/display/CAN/Edinburgh+Cancer+Informatics+Wiki

Language

This property refers to a language of the Dataset. This property can be repeated if there are multiple languages in the Dataset.

[dct:language dct:LinguisticSystem 0..n]

Ex: https://publications.europa.eu/resource/authority/language/ENG

Other identifier

This property refers to a secondary identifier of the Dataset, such as MAST/ADS , DataCite , DOI , EZID or W3ID .

[adms:identifier adms:Identifier 0..n]

Provenance

This property contains a statement about the lineage of a Dataset.

[dct:provenance dct:ProvenanceStatement 0..n]

Ex: CLINIC , PRIMARY CARE , OUTPATIENTS , IN-PATIENTS , SERVICES , PHARMACY

Qualified attribution

This property refers to a link to an Agent having some form of responsibility for the resource.

[prov:qualifiedAttribution prov:Attribution 0..n]

Qualified relation

This property provides a link to a description of a relationship with another resource.

[dcat:qualifiedRelation dcat:Relationship 0..n]

Related resource

This property refers to a related resource.

[dct:relation rdfs:Resource 0..n]

Release date

This property contains the date of formal issuance (e.g., publication) of the Dataset.

[dct:issued rdfs:Literal typed as xsd:date, xsd:dateTime, xsd:gYear or xsd:gYearMonth 0..1]

Ex: 23/11/2022

Sample

This property refers to a sample distribution of the dataset.

[adms:sample dcat:Distribution 0..n]

Source

This property refers to a related Dataset from which the described Dataset is derived.

[dct:source dcat:Dataset 0..n]

Spatial resolution

This property refers to the minimum spatial separation resolvable in a dataset, measured in metres.

[dcat:spatialResolutionInMeters rdfs:Literal typed as xsd:decimal 0..1]

Temporal resolution

This property refers to the minimum time period resolvable in the dataset.

[dcat:temporalResolution rdfs:Literal typed as xsd:duration 0..1]

Type

This property refers to the type of the Dataset. A recommended controlled vocabulary data-type is foreseen.

[dct:type skos:Concept 0..n]

Ex: http://edamontology.org/format_3752

Update/ modification date This property contains the most recent date on which the Dataset was changed or modified. [dct:modified rdfs:Literal typed as xsd:date, xsd:dateTime, xsd:gYear or xsd:gYearMonth 0..1] **Ex: 23/11/2022**
Version This property contains a version number or other version designation of the Dataset. [owl:versionInfo rdfs:Literal 0..1]
Version notes This property contains a description of the differences between this version and a previous version of the Dataset. This property can be repeated for parallel language versions of the version notes. [adms:versionNotes rdfs:Literal 0..n]
Was generated by This property refers to an activity that generated, or provides the business context for, the creation of the dataset. [prov:wasGeneratedBy prov:Activity 0..n]

Table 1: Examples of information per property

In the context of HealthyCloud, as summary of the main results of D3.2, the following Table 2 gathers a set of recommendations as part of the guidelines to standardise metadata templates and assessment of FAIRness maturity levels.

Recommendations to standardise metadata templates
Use a descriptive metadata record template compatible with the DCAT-AP standard.
Recommendations for the assessment of FAIRness maturity levels
Metadata catalogue where metadata is well described, including a PID, and the identifier of the data they describe.
Data and metadata should be accessible, registered or indexed in an open registry.

Data access procedures or protocols defined and publicly accessible.
Data access procedures should include: communications protocols, clear data access policies and legal basis, describing the conditions to get access to the data and data sharing agreements.
Data and metadata representation and meaning have to be formal, clear and unambiguous, using internationally recognised standards, ontologies and vocabularies (e.g. using CDM).
Define a clear and accessible data sharing, reuse and usage licences in a regulatory framework (e.g. applying Creative Commons licences).
Inclusion of the results on publicly accessible and findable repositories.
There are multiple resources to assess FAIRness maturity levels, starting with the HealthyCloud FAIRness assessment tool and other tools, metrics and indicators related to the conformance of data objects to the FAIR principles (Section 3.1 > Assessment of FAIRness maturity levels of this document).

Table 2: Summary of guidelines to standardise metadata templates and assessment of FAIRness maturity levels

4. Conclusions

The deliverable **D3.2 'Guidelines to standardise metadata templates and assessment of FAIRness maturity levels'** covers a study on recommendations for metadata templates standardisation and FAIRness maturity levels assessment.

Firstly, Section 3.1 gathers the most relevant recommendations proposed for addressing the **findability, accessibility, interoperability and reusability** of health data collections of HealthyCloud. Besides, Section 3.1 drafts a set of tools and metrics for the **FAIRness assessment** of data collections of HealthyCloud, including the main references and indicators used in the research field. All these tools are valid for the data collections in HealthyCloud. Further information and how they work can be found in the references and links.

Secondly, Section 3.2 presents a **descriptive metadata record template** compatible with the **DCAT-AP** standard since we can state that it is the metadata standard of choice for several projects, including the HealthData@EU [17]. DCAT-AP offers the possibility to enrich datasets with linked metadata. Concretely, Table 1 presents the properties of the class dataset, of the version 2.1.0 for the DCAT-AP for data portals in Europe, and shows some examples of information per property are provided. For each cell: property name, usage note, [URI, range, cardinality], example.

As main conclusions, Table 2 shows a summary of the results of D3.2 as guidelines to standardise metadata templates and assessment of FAIRness maturity levels.

Finally, in order to complete the Task 3.3, deliverable D3.3 'Landscape analysis using a health related-data catalogue matrix' will include a comparison of the FAIRness evaluation tools mentioned in this document. Deliverable D3.3 will be submitted at month M26 (April 2023).

5. References

[1] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

[2] Kinsella, C. M., Santos, P. D., Postigo-Hidalgo, I., Folgueiras-González, A., Passchier, T. C., Szillat, K. P., ... & Martí-Carreras, J. (2020). Preparedness needs research: How fundamental science and international collaboration accelerated the response to COVID-19. PLoS Pathogens, 16(10), e1008902. https://doi.org/10.1371/journal.ppat.1008902

[3] Besançon, L., Peiffer-Smadja, N., Segalas, C., Jiang, H., Masuzzo, P., Smout, C., ... & Leyrat, C. (2021). Open science saves lives: lessons from the COVID-19 pandemic. BMC Medical Research Methodology, 21(1), 1-18. https://doi.org/10.1186/s12874-021-01304-y

[4] Lee, P., Abernethy, A., Shaywitz, D., Gundlapalli, A. V., Weinstein, J., Doraiswamy, P. M., ... & Madhavan, S. (2022). Digital Health COVID-19 Impact Assessment: Lessons Learned and Compelling Needs. NAM Perspectives.

[5] RDA WG on FAIR Data Maturity Model: https://www.rd-alliance.org/groups/fair-data-maturity-model-wg#:~:text=The%20RDA%20FAIR%20Data%20Maturity%20Model%20Working%20Group%20develops%20as,maturity%20level%20of%20a%20dataset.

[6] GO FAIR Initiative: https://www.go-fair.org/

[7] FAIRplus project website: https://fairplus-project.eu/

[8] FAIR4Health project website: https://fair4health.eu/

[9] PHIRI project: https://www.phiri.eu/

[10] European Joint Programme on Rare Diseases (EJP RD): https://www.ejprarediseases.org/

[11] FAIRplus Fellowship Program: https://fairplus-project.eu/get-involved/fellowship

[12] D.B. Deutz, M.C.H. Buss, J. S. Hansen, K. K. Hansen, K.G. Kjelmann, A.V. Larsen, E. Vlachos, K.F. Holmstrand (2020). How to FAIR: a Danish website to guide researchers on making research data more FAIR. https://doi.org/10.5281/zenodo.3712065

[13] More on metadata standards and ontologies: https://howtofair.dk/links-additional-reading/#more-on-metadata-standards-and-ontologies-

[14] RDA WG on Research Metadata Schemas: https://www.rd-alliance.org/groups/research-metadata-schemas-wg

[15] RDA Metadata Standards Directory Working Group: https://www.rd-alliance.org/groups/metadata-standards-directory-working-group.html

[16] CEDAR Metadata Center: https://metadatacenter.org/

[17] HealthData@EU pilot project website: https://www.ehds2pilot.eu/

[18] Data Catalog Vocabulary (DCAT) - Version 2. W3C Recommendation 04 February 2020: https://www.w3.org/TR/vocab-dcat/

[19] Data Catalog Vocabulary (DCAT). Version 2: https://www.w3.org/TR/vocab-dcat-2/#introduction

[20] Data Catalog Vocabulary (DCAT). Version 3: https://www.w3.org/TR/vocab-dcat-3/

[21] ELIXIR Core Data Resources: https://elixir-europe.org/platforms/data/core-data-resources

[22] HealthyCloud Glossary of commonly used terms in the field of health data research - developed by the EU project HealthyCloud: https://zenodo.org/record/6787119#.Y8feBRfMKUk

[23] EMBL-EBI Ontology Lookup Service: https://www.ebi.ac.uk/ols/index

[24] HL7 FHIR: https://www.hl7.org/fhir/license.html

[25] OMOP Common Data Model: https://www.ohdsi.org/data-standardization/the-common-data-model/

[26] Creative Commons licenses: https://creativecommons.org/licenses/

[27] License RDF: https://wiki.creativecommons.org/wiki/License_RDF

[28] HealthyCloud FAIRness assessment tool. https://doi.org/10.5281/zenodo.7038397

[29] HealthyCloud D4.1 – Recommendations for integration in HealthyCloud, including an analysis of data hub patterns of governance. https://healthycloud.eu/wp-content/uploads/2022/11/D4.1.pdf

[30] ARDC FAIR Data Self Assessment Tool: https://ardc.edu.au/resources/aboutdata/fair-data/fair-self-assessment-tool/

[31] FAIR Maturity Evaluator: https://fairsharing.github.io/FAIR-Evaluator-FrontEnd/#!/

[32] FAIRshake tool: https://faircookbook.elixir-europe.org/content/recipes/assessing-fairness/fair-assessment-fairshake.html

[33] FAIRassist tool: https://fairassist.org/#!/

[34] FAIR Data Maturity Model. Specification and Guidelines. https://doi.org/10.15497/RDA00050

[35] FAIRplus Project website: https://fairplus-project.eu/

[36] FAIRplus: D3.2 IMI FAIR Metrics Publication. https://zenodo.org/record/4428633#.YnJlnnZBzIV

[37] FAIRplus FAIR indicators. https://fairplus.github.io/fairification-results/2020-10-11-FAIRplus-indicators-v0.1/

[38] Guidance for research organisations on how to implement responsible and fair approaches for research assessment: https://wellcome.org/grant-funding/guidance/open-access-guidance/research-organisations-how-implement-responsible-and-fair-approaches-research

[39] Wellcome foundation website: https://wellcome.org/

Annex 1

Turtle file of the DCAT-AP example provided in Table 1.

```
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix ldp: <http://www.w3.org/ns/ldp#> .

<http://URL.eu/dataset/f415a71e-02cc-4aad-9d40-05af6be80fea> a
   dcat:Resource, dcat:Dataset;
  dcterms:accessRights <https://www.wiki.ed.ac.uk/display/CAN/Governance>;
  dcterms:language
<https://publications.europa.eu/resource/authority/language/ENG>;
  dcterms:license <http://rdflicense.appspot.com/rdflicense/cc-by-nc-nd3.0>;
  <https://w3id.org/fdp/fdp-o#metadataIdentifier>
<http://URL.eu/dataset/f415a71e-02cc-4aad-9d40-05af6be80fea#identifier>;
  <https://w3id.org/fdp/fdp-o#metadataIssued>
"2022-12-13T20:59:58.709Z"^^xsd:dateTime;
  <https://w3id.org/fdp/fdp-o#metadataModified>
"2022-12-13T20:59:59.451Z"^^xsd:dateTime;
```

```
  dcterms:isPartOf
<http://URL.eu/catalog/03de9bcb-a074-4966-923c-cc0e9c1baa5f>;
  <http://semanticscience.org/resource/SIO_000628>
<http://URL.eu/dataset/f415a71e-02cc-4aad-9d40-05af6be80fea/metrics/445c0a7
0d1e214e545b261559e2842f4>,

<http://URL.eu/dataset/f415a71e-02cc-4aad-9d40-05af6be80fea/metrics/5d27e8
54a9e78eb3f663331cd47cdc13>;
  dcterms:spatial
<http://publications.europa.eu/resource/authority/country/GBR>;
  dcat:contactPoint
<https://www.ed.ac.uk/information-services/about/contacting-is>;
  dcat:keyword " ADENOCARCINOMA"@en, " CANCER"@en, "
CARCINOSARCOMA"@en, " CLEAR CELL"@en,
    " ENDOMETRIOID"@en, " FALLOPIAN TUBE"@en, " OVARIAN"@en, "
PERITONEAL"@en, " PERITONEUM"@en,
    " SARCOMA"@en, " SEROUS"@en, "OVARY"@en;
  dcat:landingPage
<https://www.wiki.ed.ac.uk/display/CAN/Edinburgh+Cancer+Informatics+Wiki>;
  dcat:theme <http://purl.obolibrary.org/obo/MONDO_0008170>;
  dcat:distribution
<http://URL.eu/distribution/29501e79-c6f1-42f2-b599-1822723a66db>;
  <http://www.w3.org/2000/01/rdf-schema#label> "Edinburgh Ovarian Cancer
Database";
  dcterms:description "Established in 2019, the Ovarian Cancer Database builds on
40 years of data collection across the region of the South East Scotland Cancer
Network. The database holds data on diagnosis, treatment and outcomes of
patients undergoing care within the region."@en;
  dcterms:hasVersion "1.0";
  dcterms:issued "2022-11-23T00:00:00.000Z"^^xsd:dateTime;
  dcterms:modified "2022-11-23T00:00:00.000Z"^^xsd:dateTime;
  dcterms:publisher [ a foaf:Agent;
      foaf:name "NHS Lothian and the University of Edinburgh"
    ];
  dcterms:title "Edinburgh Ovarian Cancer Database"@en;
  dcterms:conformsTo
<http://URL.eu/profile/2f08228e-1789-40f8-84cd-28e3288c3604> .

<http://URL.eu/dataset/f415a71e-02cc-4aad-9d40-05af6be80fea/metrics/445c0a7
0d1e214e545b261559e2842f4>
  <http://semanticscience.org/resource/SIO_000332>
<https://www.ietf.org/rfc/rfc3986.txt>;
  <http://semanticscience.org/resource/SIO_000628>
<https://www.ietf.org/rfc/rfc3986.txt> .
```

```
<http://URL.eu/dataset/f415a71e-02cc-4aad-9d40-05af6be80fea/metrics/5d27e8
54a9e78eb3f663331cd47cdc13>
  <http://semanticscience.org/resource/SIO_000332>
<https://www.wikidata.org/wiki/Q8777>;
  <http://semanticscience.org/resource/SIO_000628>
<https://www.wikidata.org/wiki/Q8777> .

<http://URL.eu/dataset/f415a71e-02cc-4aad-9d40-05af6be80fea#identifier>
  a <http://purl.org/spar/datacite/Identifier>;
  dcterms:identifier
"http://URL.eu/dataset/f415a71e-02cc-4aad-9d40-05af6be80fea" .

<http://URL.eu/distribution/> a ldp:DirectContainer;
  dcterms:title "Distributions";
  ldp:membershipResource
<http://URL.eu/dataset/f415a71e-02cc-4aad-9d40-05af6be80fea>;
  ldp:hasMemberRelation dcat:distribution;
  ldp:contains
<http://URL.eu/distribution/29501e79-c6f1-42f2-b599-1822723a66db> .

<http://URL.eu/profile/2f08228e-1789-40f8-84cd-28e3288c3604>
<http://www.w3.org/2000/01/rdf-schema#label>
    "Dataset Profile" .
```