Deliverable D2.2 – Report on DOME 4.0 FAIR compliance

Responsible Partner:	UKRI	07-11-2024
Contributor(s):	UKRI	07-11-2024
Reviewer(s):	INTRA (Kostas Sipsas), UCL (Adham Hashibon)	29-11-2024
Coordinator:	CMCL	29-11-2024
Dissemination Level:	Public	
Due Date:	M48 (November, 2024)	
Submission Date:	29.11.2024	

Project Profile

Programme	Horizon 2020
Call	H2020-NMBP-TO-IND-2020-twostage
Topic	DT-NMBP-40-2020 Creating an open marketplace for industrial data (RIA)
Project number	953163
Acronym	DOME 4.0
Title	Digital Open Marketplace Ecosystem 4.0
Start Date	December 1st, 2020
Duration	48 months

This document is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 953163. It is the property of the DOME 4.0 consortium and do not necessarily reflect the views of the European Commission.

Document History

Version	Date	Author	Remarks
V0.1	2024-11-05	Noel Vizcaino Silvia Chiacchiera Ilian Todorov Vasily Bunakov Martin Uhrin (acknowledgement of previous content)	Version for reviewer partners based on previous internal reports
V0.2	2024-11-25	Kostas Sipsas	Review
V0.3	2024-11-29	Adham Hashibon	Review
V0.4	2024-11-29	Willem van Dorp	Formatting and finalization

Executive Summary

This document summarizes advice on data/asset FAIRness and its evaluation. Some pre-existing tools are analyzed in detail and advice is given to data providers and connector developers to increase the FAIRness of their data.

Table of Contents

List of Figures

List of Tables

1. Introduction

The DOME 4.0 Task 2.2" FAIR Data Monitoring and Auditing Service" commenced as planned in June 2021, by looking into relevant projects and community effort such as in RDA for FAIR assessment. This line of work was further pursued and culminated in November 2021 by compiling an internal report with a list of relevant projects and initiatives. This document builds on that report and following versions. Accordingly, we start by listing initiatives and their (data/asset) FAIRness evaluation tools. Then in section 2 we give the FAIRness codes, a set of aspects typically used to actually measure FAIRness. Beside their general definition, we provide for each of them concrete examples of the affected metadata. In Sec. 3 we discuss best practices for metadata and in Sec. 4 links to other DOME tasks. Metadata can be provided in different ways, we focus on Sec. 5 on the JSON-LD format and provide advice on its generation, with emphasis on relevant scenarios (e.g. tabular scientific data). In Sec. 6 we discuss in detail the main FAIRness evaluation tools. Finally, we draw our conclusions and summarize the lessons learnt.

1.1 Relevant initiatives and tools

1.1.1 Preferred initiative

The most mature software tool, originating in the aforementioned projects, is F-UJI [5] (developed by FAIRsFAIR) which implements most of the metrics developed by the RDA group. We started looking into F-UJI code, which proved revealing as RDA-developed metrics leave much room for their interpretation, and only by looking into the code could one really make sense of how a particular FAIR metric was actually interpreted.

Another tool called FAIR Evaluator [6] was developed by EOSC-Synergy along the same lines based on the RDA metrics, but it looks less mature than F-UJI and deemed not worth reusing. However, EOSC-Synergy tried to move one step further than FAIRsFAIR did and investigated developing another set of FAIR metrics beyond the assessment of a dataset but devoted to that of an entire data repository. This can be a promising line of work conceptually, and, to some extent, of practical interest as we may want to evaluate certain repositories where DOME is fetching data from, then just "trust" them, i.e. automatically assign "default" FAIR metrics to datasets originating from a certain "FAIR" repositories.

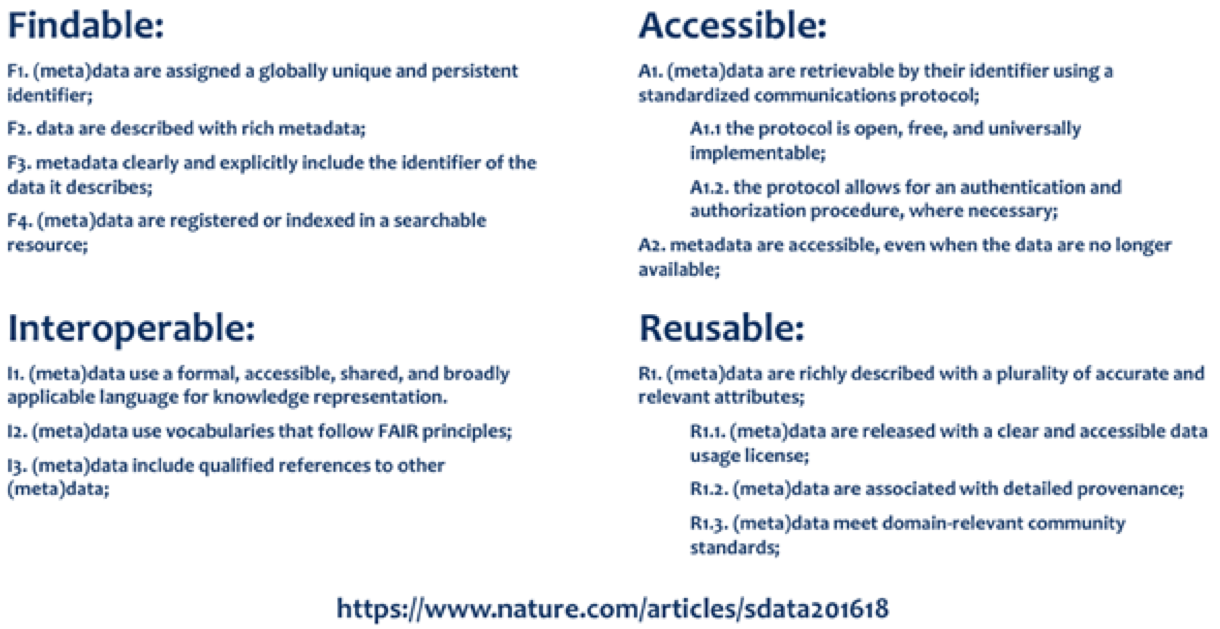
1.1.2 Other initiatives

Of a particular importance are the outcomes of the RDA FAIR Data Maturity Model Working Group [2] that produced a set of FAIRness indicators using a regular RDA process for community engagement. This set of indicators underpins the actual software tools developed by FAIRsFAIR [3] and EOSC-Synergy [4] projects. It is worth noting that both projects ended in 2022, so in a way the DOME T2.2 has taken over from them, regarding investigating viable FAIR measurement approaches.

We further investigated a recent work carried out by FAIRsharing.org [7] who developed FAIRsharing FAIR Evaluation Services [8] that encompass various resources and guidelines to assess the FAIRness of digital assets. One of the most promising resources developed by FAIRsharing.org is their List of Maturity Indicators [9]. It contains some popular indicators and allows to register your own FAIR maturity indicator then refer to it using a standard FAIRsharing citation mechanism, including a DOI assigned to an indicator. This bears a good potential for defining FAIR metrics that reflect a specific notion of FAIRness well-fit with DOME design and purposes, and further the usage of these metrics in FAIR assessment tools and semantic assets such as ontologies.

2. FAIRness codes

There is a list of FAIRness codes produced and accepted by many research groups. This takes many shapes with different nuances. The codes classify FAIRness by narrowing it down into a specific subtopic, while remaining quite abstract. No concrete implementations are addressed so the ideas can be easily ported to other vastly different solutions.



Findable:

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

F2. data are described with rich metadata;

F3. metadata clearly and explicitly include the identifier of the data it describes;

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

Accessible:

A1. (meta)data are retrievable by their identifier using a standardized communications protocol;

A1.1 the protocol is open, free, and universally implementable;

A1.2. the protocol allows for an authentication and authorization 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;

https://www.nature.com/articles/sdata201618

Figure 1: FAIR principles from GO-FAIR. Source: https://www.go-fair.org/fair-principles/

For further details and reference about these practices see:

- 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
- FAIRMetrics. Detailed description of the metrics with fine-grained codes here: https://github.com/FAIRMetrics/Metrics/blob/master/MaturityIndicators/Gen1/ALL.pdf

We can map FAIR codes and their FAIR descriptions to match the affected elements in the example metadata examples produced. Note that FAIR codes are broad and abstract; they also are implementation independent.

The FAIR codes below are used in e.g., the actual implementation source code of both **F-UJI (FAIRsFAIR)** and **EOSC-Synergy**.

Table 1: FAIR CODES to metadata. Example mapping (Note: @id aliased to id and @type to type)

FAIR CODE	Brief explanation	Metadata Scope	Example metadata affected (DCAT2/JSON-LD)
FM1-F1A	GUID used as part of dataset IRI	Dataset ID	Top object “@id” IRI (e.g. contains UUID4 as a global identifier)
FM1-F1B	The IRI must be permanent (Persistent Identifier)	Fixed scheme	Top object “id” IRI.
FM-F2	Rich metadata	yes	All
FM-F3	Data includes identifiers	yes	All object "@id" "identifier” (list)
FM-F4	Data indexed in a searchable resource.	Vertical slice. A service provides access.	All, but all ‘@id’ essential.
FMA1	Retrievable by ID using standard protocol	Yes (HTTP) RDF model (JSON-LD).	All object "@id" but likely by dataset.
FM-A1.1	Open, free etc protocol	Yes (HTTP) RDF model (JSON-LD).	All object "@id" but likely by dataset.
FM-A1.2	Protocol allows for authentication and authorisation	Yes (HTTP) Perhaps combining with IDs and JWT. (RBAC)	N/A (Handled by another service) *Metadata may or may not be involved.*
FM-A2	Metadata is accessible even when data are no longer available	Metadata and data separated and independent but connected.	Metadata and actual data files separated.

			Metadata **still accessible** when the data is not. By design. *'accessURL' 'downloadURL'* fail, not available or blank scenarios.
FM-I1	It uses language for knowledge representation	Ontologies: DCAT2, EMMO (and associated vocabs) *EuroSciVoc* or others.	All, by design.
FM-I2	Vocabularies used are themselves FAIR	Yes	All, by design.
FM-I4	It uses qualified references to other metadata	ror.org, ORCiD, OpenID, SPDX (licences) URLs, etc	All '@id' Mostly user provided.
FM-R1	Richly described with accurate and relevant attributes	Yes	All, by design.
FM-R1.1	Open licences URL: SPDX, Creative Commons Other: URL must be provided	Yes	From the 'license' object: the '@id'. User provided.
FM-R1.2	Detailed provenance must be possible	Using IDs. DCAT.	*Metadata may or may not be involved.* All '@id' tracking. User provided. DCAT provisions for provenance and lineage. Source user provided. e.g. 'source' (not in example)
FM-R1.3	Meets community standards	W3C standards based.	All

The DOME 4.0 connectors (cf. DOME Deliverable 3.4) must meet these requirements when producing FAIR metadata for DOME 4.0. We recall that heir data and model catalogue use the DOME 4.0 Ecosystem Ontology to classify exemplars.

3. State-of-the-art metadata

3.1 5stardata

Sir Tim Berners-Lee's endorsed classification of data on the web to encourage **users** to provide data in files in specific formats so they are open, and thus, facilitating interoperability:

- There is a 5-level assessment: https://5stardata.info/en/
- To retain the control of the quality of our own metadata.
- With lesser barrier of entry for newer users.
- To avoid needless data conversions (when possible).
- To enrich the data incrementally and progressively.
- The users can enrich the metadata online, in a safe way.
- To avoid needless obsolescence.
- Semantic linked data is a top-level solution.

The general implication of the above is that users will have to provide, eventually (online), all free-text entries, any categorisation, provenance, etc. The inputs can be rationalised: *e.g. a drop down is safer than a textbox.*

3.2 Google web metadata

Google supports both *schema.org* and W3C DCAT serialised as JSON-LD. Google and other search engines are promoting these standards for scientific datasets, among others uses. This approach is quite W3C RDF centric.

- It is also adding W3C CSVW support: https://developers.google.com/search/docs/advanced/structured-data/dataset#approach
- A W3C standards ecosystem is preferred to be consistent but is not required.
- Multilingual support, by design, makes any DOME platform intelligence multilingual on inception. (JSON-LD has multilingual support built-in)

Further information: https://developers.google.com/search/docs/appearance/structured-data/dataset

3.3 Permanent Identifier (PID) strategy recommendations

These FAIR practices are based on national Permanent Identifiers (PID) strategies. F-UJI will look that the IRIs are unique global IDs and, if they are de-referenceable URLs, fetch and assess the data payload.

3.3.1 For dataset metadata resources

- ✓ <base> could be e.g. http://dome4.com/ or http://example.com/ (or any other).
- ✓ The @id of the dataset is dataset IRI = <base>dataset/<**uuid**>
- ✓ The @id of a file (distribution) is distribution IRI = <base><distribution>/<**filehash**>

3.3.2 For data files resources

- ✓ **accessURL: M**eant for API access = <base>data/<**filehash**>.
- ✓ ***checksum :***is From the SPDX vocabulary (which is also for open licences). Used by e.g. **DCAT-AP**. A A minimal requirement would be to use the MD5 hash digest (filehash) to *ID the file* in a distributed setting (for anti-tampering the SHA-512 algorithm should instead be used).
- ✓ **downloadURL** for web clients of any kind = <base>download/<**filehash**>

The “dataset” string literal in the remote resource addresses is not an arbitrary name as it is based on the concept of DCAT dataset. Similarly other entities of note may be identified. For scientific entities a SKOS base controlled vocabulary may be needed. The DOME 4.0 Ecosystem Ontology should be used as a reference. The objective is to create a flexible yet clear resource address space. They will also play a role as IDs in the Knowledge Graph as store in any triplestore. The database will reconcile the IDs on ingestion, making the graph paths deeper.

4. Links with other tasks in DOME 4.0

A good amount of work related to T2.2 has been done in the context of T3.2 Ecosystem Information Model and T4.1 Metadata, Data Acquisition, Curation, and Communication.

The idea in the context of T3.2 is to make DOME data FAIR “by design” leveraging DCAT2 (and associated ontologies) serialised as JSON-LD 1.1. Conversions to other RDF serialisations can be offered to clients by the RDF database (e.g., via a web server). In this way, FAIR principles could be “embedded” in the DOME information model and in its implementation.

There is a natural connection of T2.2 to T4.1, too, as data samples collected can be used as an input to FAIR assessment tool. Implementation of metadata based on raw data provided by industrial partners.

There is an example of a command line tool created in 2021 (updated 2022) to generate JSON-LD and DCAT2 metadata from local raw data (provided by industrial partners). The rationale was that the partners do not need to provide any metadata that could be automatically generated. Also, the safety that they would not need to provide the data to be hosted was a concern. Links to their data store in their remote servers may be added later.

An experimental command line tool was developed to demonstrate how to automate the generation of DCAT2 serialised as JSON-LD. The results have been demonstrated in various tests by ingesting the JSON-LD serialisation in a triplestore (RDF database).

The command line tool is in the Ontology-matters branch in the project GitHub repository with the name metadata-generation.

DOME 4.0 requires the connector developers responsible for producing complying metadata based on their own sources. They should also add provenance information as recommended by W3C DCAT2 using W3C PROV-O.

5. Metadata structure: JSON-LD illustration

JSON-LD is an RDF and a JSON document, making both data views viable for processing. It constitutes one of the possible valid RDF serialisations, others exist (e.g., TTL notation).

JSON(-LD) metadata example(s) automatically generated from the s*ample datasets from industrial partners (*Task 4.1) have been shared.

Comments:

- Aliases have British spelling but not the types.
- JSON-LD does some automatic interpretations implicitly: e.g., all plain strings are **`xsd:string`.**

Table 2: Dataset Metadata (top level object)

Property	Type	Cardinality	Content Policy
@context	Object or IRI to the semantic context with our DCAT2 profile.	0..1	Currently **nested** in the metadata but it can be a URLpointing to the JSON-LD semantic context file **(dome-ld-core.jsonld)**. It can be supplied in various ways.

id (@id)	URI	1	To self-reference the metadata itself. It should be permanent
type (@type) Not to be confused with *dct:type*	*dcat:Dataset*	1	DCAT, Fixed
identifier	Ordered list of IDs (URIs preferred)	1..*	Good practice to include dataset id as first element. It includes copies in other repositories if available.
landingPage	URI	1	The URI for the dataset web page if any.
language	string or better URI	1	Related to contents. Two letter country code or three letter. (ISO 639-1, ISO 639-2)
title	string	1	Free text
description	string	1	Free text
keyword *(no typo!)*	Unordered list of strings	0..*	Free text
theme	Unordered list of skos:Concept belonging to a formal vocabulary. E.g. EuroSciVoc IDs (or any other).	1..*	Compact URIs, any vocab
emmo (It could also be in *distribution*) If this *alias* is inconvenient, we could use *emmoMetadata*	EMMO metadata graph (ontology)	1..*	Own @context. Nested (preferred) This could be large and is completely independent from DCAT2. Serialised as JSON-LD.
conformsTo	dct:Standard	0..*	Collection of Objects with URI to standard.
publisher	foaf:Agent	1..*	Collection of Objects with URI to Agent. Includes foaf:Person and foaf: Organization
creator	foaf:Agent	1..*	Collection of Objects with URI to Agent.
qualifiedAttribution	prov:Attribution	0..*	Collection of objects (Attributions, W3C PROV-O provenance) Roles should be an official vocab e.g. ISO 19115 roles. It should be noted that Agents can be Software Services, Organizations or Individuals.
contactPoint	foaf:Individual	1	Official contact for the **owner** of the rights of dataset.

licence	dct:LicenseDocument	1	It must contain permanent URI to licence. Public licences preferred. It could be aliased to British spelling like in the pan-european DCAT-AP
rights	string	0..1	Further rights statements.
accessRights	Text but may be part of a **vocabulary**	1	• Public • Private • Embargoed • etc
created, modified, issued	All timestamps with timezone	1	ISO 8601 timestamps
temporal	dct:PeriodOfTime	1	DCAT, Fixed Applicability range (time) ISO 8601 timestamps
accrualPeriodicity	URI	0..1	Update frequency (DCT vocabulary). https://www.dublincore.org/specifications/dublin-core/collection-description/frequency/
distribution	dcat:Distribution	1..*	DCAT, fixed. Collection of file metadata.

Table 3: Distribution metadata (data files metadata). Note: all free text (including labels) could ideally be multilingual.

Property	Type	Cardinality	Content Policy
type	dcat:Distribution	1	DCAT, Fixed
id	URI	1	ID of the data file.
title	text	0..1	Title of the datafile
fileName	string	1	Inc. dataset relative path, recommended.
accessURL	URI	1	e.g. For API use
downloadURL	URI	1	Browser friendly URI
mediaType	text	1	IANA media types. In lieu of format (precedence)
byteSize	Coerced to xsd:Decimal (From xsd:string)	1	We will use just bytes (file data size)
checksum	MD5 hash	1	Not part of plain DCAT2.

			Helps with versioning, provenance, integrity and security. And it is universal.
conformsTo	dct: conformsTo	1	Collection of Objects with URI to standard related to this data file.
tableSchema	csvw:Schema	0..1	Not part of plain DCAT2. However, CSV is one of the lowest semi-structured open file formats we could expect. **Only if** mediaType is 'text/csv' and is **RFC 4180 standard compliant**.

The metadata resulting from the above schema template generated by automated means (from provided data), when possible. The UUID(s) as filename per dataset and any hash digest as file data ID to be stored persistently. A JSON-LD processor can reconstruct the full IRI using the proper @context section.

It should be noted that that file **format** is omitted since **mediaType** takes precedence if IANA media type identifiers are available, which is a common occurrence for most generic files.

5.1 Parameters, variables, and multilingual considerations

5.1.1 CSV format metadata (W3C CSVW)

To cover for basic metadata about a csv file:

The column names as: `name, xsd:string` (canonical, processable)

```
"tableSchema": {
            "@type": " Schema",
            "columns": [
              {
                "@type": " Column",
                "name": "id",
              },
              {
                "@type": " Column",
                "name": "kinetic_energy",
              }]
}
```

(Note: the command line tool will only generate CSW for folders not for zip files)

5.1.2 Dealing with parameters and S.I. units

Useful user-contributed, added-value *recommended* incremental additions:

- Description (multilingual, free text)
- Titles of columns (multilingual, free text)
- **Data types** (double, string, etc) : https://www.w3.org/TR/tabular-metadata/#fig-datatypes
- Valid ranges, min, max, etc.
- **Units of measure**. No mandatory standard. We could use SDMX, W3C Data Quality Vocabulary (DQV), etc.

```
{

 "@type": " Column",
 "name": "kinetic_energy",
 "titles": { "en": "Kinetic Energy"}
          "es": "Energía Cinética"},
          "no": "Kinetisk Energi"},
          "de": "Kinetische Energie"},
        },
 "datatype": "xsd:double"

}
```

(We can add column multilingual dct:description).

Developers may extend the metadata and polish it as we discover and add new features to DOME 4.0 at later stages.

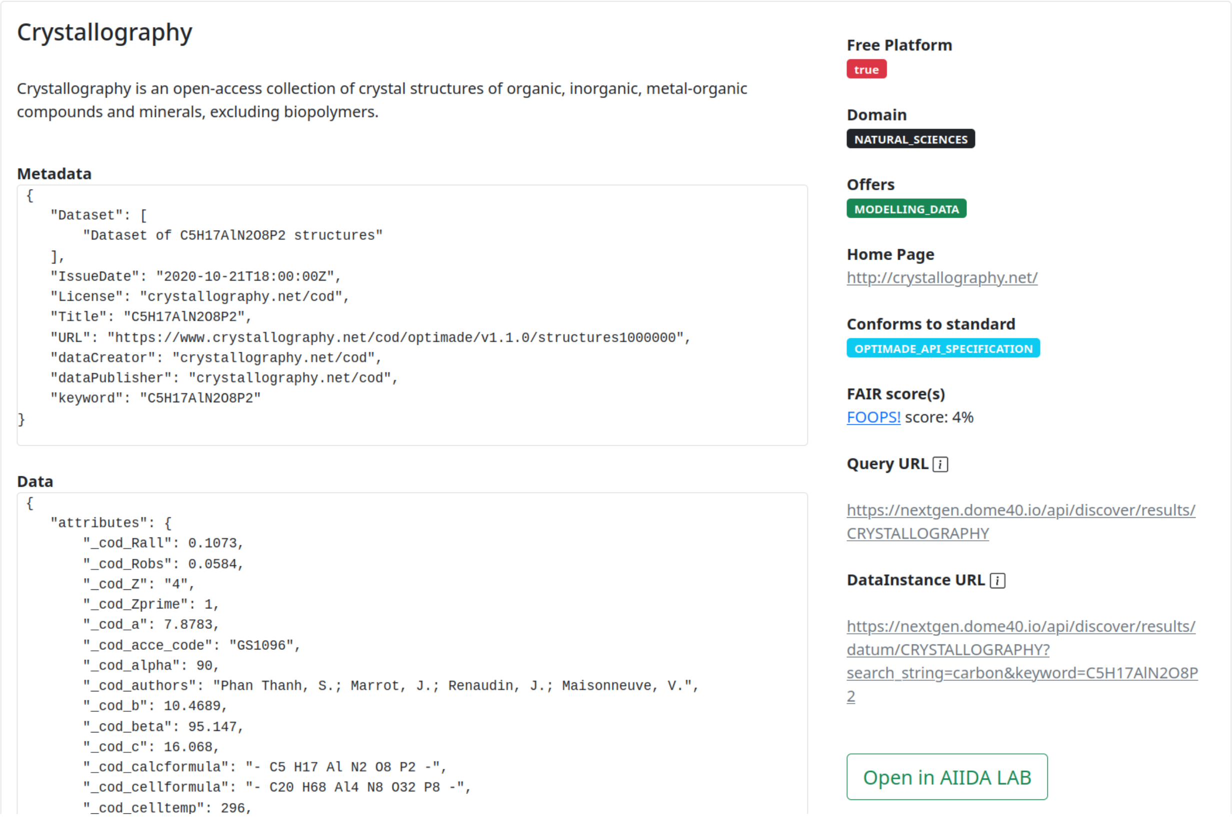
Enriching with further scientific metadata

Beyond any advice so far, the Dome 4.0 Ecosystem Ontology (cf. DOME 4.0 Deliverable 3.2) should be used at any opportunity to harmonise metadata across systems. The connector developers should take advantage of this.

6. FAIR assessment tools

When performing a search on the DOME 4.0 platform and selecting one of the results, DOME provides a FAIR assessment results list for it, see Fig. 2. On the right we find an information section. The “FAIR score(s)” list all the tools and scores we found so far for this record. The default tool currently used is

FOOPS![1]. FU-UJI was recommended to be deployed. Also, other tools may feature eventually. However, we should consider that the DOME 4.0 *connectors* have some standards compliance requirements.



Crystallography

Crystallography is an open-access collection of crystal structures of organic, inorganic, metal-organic compounds and minerals, excluding biopolymers.

Metadata

```
{
   "Dataset": [
       "Dataset of C5H17AlN2O8P2 structures"
   ],
   "IssueDate": "2020-10-21T18:00:00Z",
   "License": "crystallography.net/cod",
   "Title": "C5H17AlN2O8P2",
   "URL": "https://www.crystallography.net/cod/optimade/v1.1.0/structures1000000",
   "dataCreator": "crystallography.net/cod",
   "dataPublisher": "crystallography.net/cod",
   "keyword": "C5H17AlN2O8P2"
}
```

Data

```
{
   "attributes": {
       "_cod_Rall": 0.1073,
       "_cod_Robs": 0.0584,
       "_cod_Z": "4",
       "_cod_Zprime": 1,
       "_cod_a": 7.8783,
       "_cod_acce_code": "GS1096",
       "_cod_alpha": 90,
       "_cod_authors": "Phan Thanh, S.; Marrot, J.; Renaudin, J.; Maisonneuve, V.",
       "_cod_b": 10.4689,
       "_cod_beta": 95.147,
       "_cod_c": 16.068,
       "_cod_calcformula": "- C5 H17 Al N2 O8 P2 -",
       "_cod_cellformula": "- C20 H68 Al4 N8 O32 P8 -",
       "_cod_celltemp": 296,
```

Free Platform
true

Domain
NATURAL_SCIENCES

Offers
MODELLING_DATA

Home Page
http://crystallography.net/

Conforms to standard
OPTIMADE_API_SPECIFICATION

FAIR score(s)
FOOPS! score: 4%

Query URL ⓘ

https://nextgen.dome40.io/api/discover/results/CRYSTALLOGRAPHY

DataInstance URL ⓘ

https://nextgen.dome40.io/api/discover/results/datum/CRYSTALLOGRAPHY?search_string=carbon&keyword=C5H17AlN2O8P2

Figure 2: Screen Capture from stored records.

The FAIRsFAIR project, F-UJI tool, is comprehensive and supports several serialisations It is aimed at datasets. There are other tools, like FOOPS!, which target ontologies. Each tool has a different interpretation of FAIR and a different intended use. A low FAIRness score could be caused by the data indeed not being FAIR or by the inability of the tool to process the data properly, e.g., by using unexpected ontologies. In contrast to RDF and XML, JSON serialisations are also problematic because they are a significant downgrade and require more discipline to create a meaningful semantic view to be serialised (with no support from the standard).

Connector developers are advised to meet the standards discussed so far, prominently:

- Valid IRIs (RFC 3987) where expected and as a minimum.

[1] Tool by the Ontology Engineering Group - Universidad Politécnica de Madrid

- W3C RDF serialisations (including the recommended W3C JSON-LD)
- W3C DCAT2 for datasets.
- W3C PROV-O for provenance.

It is fair to say that the records provided by the connectors are *transient data* part of a dataset. If we consider that these datasets are mere snapshots of the data after a query to external services. However, if we think about anything under a unique global ID or a PID identifying a dataset, then it belongs to an entity in the Knowledge Graph. E.g., anything under a CAS number [2] is indeed a dataset (in the DCAT sense) including all the potential information associated with it. We may also interpret it as some other entity without pointing to any data files. Both views may be part of the Knowlege Graph.

The choice of the tools to make a FAIR assessment is also a concern. If we are not using dataset then F-UJI will not make much sense. An optimal strategy will involve connector developers to have an explicit agreement on the target data records (enforced somehow via standards, but true compliance may be an issue).

6.1 EOSC-Synergy tool

The EOSC-Synergy project has implemented the FAIRness codes as python functions.

6.1.1 External communication

A. The task was presented in KeXS (Knowledge Exchange Space) workshop organized by OntoCommons project in July [10]. The following areas of collaboration with OntoCommons have been considered: first, Dome 4.0 could serve as a platform for industrial data exchange with good attention to interoperability; second, it could be a TLO (Top Level Ontology) Use Case; third, it could support FAIRness assessment that is also generally in line with OntoCommons aspirations.

   In connection with FAIRness assessment and FAIRification, the idea expressed by the workshop participants was to fully align, or at least connect FAIRness and data valorisation agendas: FAIRification can act as a "quality certification" adding value to data deployed for commercial purposes. It increases trust and confidence in data quality and guarantees usability, even though it does not guarantee data quality. In other words, or rather in industry's commercial terms, FAIRification holds marketing potential, which in return supports making FAIR a standard.

B. Application was made to join the EOSC Task Force on FAIR Metrics and Data Quality [11]. The group is heavily oversubscribed and we were not granted membership but they are going to set up a mailing list to keep us informed.

[2] https://dome40.eu/semantics/dome4.0_core#CAS_NUMBER under the DOME 4.0 Ecosystem Ontology

6.2 FAIRsFAIR: F-UJI (recommended)

There is excessive or unjustified use of "*regexing*" and string manipulation, which should be a last resort. This does not really pose a problem from a user perspective, if it is reliable.

If any URL identifying a dataset cannot be accessed over the internet it will treat the dataset as "not FAIR". There also is a public repository requirement. Public access is not necessary and eventual access when needed suffices. IRIs inside the returned document need to be valid and, in addition, on point.

Any problem parsing the metadata will create gaps by skipping the next logical stages of the assessment (key data cannot be reached), leading to extremely low scores.

Anything that is not what the parser will expect will get low scores. Also, parsing local deployments is not as straightforward as it should be.

6.2.1 F-UJI evaluation

PROs :

- The tool is very comprehensive. (Note: just this justifies the use)
- The modularisation of some aspects.
- Deals with the networking (Generated from OpenAPI specification).
- Deals with parsing of (unknown) metadata.
- Supports several serialisations.

CONS:

- Regexing and string manipulation. There are better ways.
- In the case of list of standards, just any absence should not make it less FAIR. E.g., using websockets instead of http. (note: just an example, as it is supported)
- External online tools dependency.
- The inevitable amount of "boilerplate code".

By design it is not intended for near realtime/high load situations. We welcome the great amount of "boilerplate code" to deal with networking/parsing. However, the FAIR core and the actual assessment, is what matters.

Any data coming from the triple store should have a high chance of getting a relatively high score. To get there the integration of metadata should happen primarily in the triplestore.

6.2.2 Summary of F-UJI FAIR tool assessment.

A URL to the dataset is (HTTP) POSTed to the <site>/evaluate where the tool is hosted, and a JSON returns with the assessment of all codes. Under “score_percent” percent key the “FAIR” entry gives the total score as a percentage.

A deployable version is publicly available here:

https://github.com/pangaea-data-publisher/fuji

It is possible for the connector developers to assess their metadata, in development phase, by using the UI powered version hosted here:

https://www.f-uji.net/

6.3 FOOPS!

FOOPS! assessment is available by default, or it will be complementary, depending on the data assessed. FOOPS! is geared to ontologies and vocabularies (rather than e.g. datasets), and its main purpose it to identify ontology pitfalls. Connectors developers need to make sure they are standards compliant (because their solutions are fully custom). The scores are orientative and not absolute. This is typical of FAIR metrics but in this case, the nature of the data assessed may not be the be a good match for this tool.

It is feasible to assess ontologies stored in DOME 4.0 by users and linked them to the tool for assessment. However, this is out of scope.

Automatically deciding which data is assigned to which assessment tool is also a challenging problem. Particularly if key standards are not followed. Therefore, the user needs to interpret the scores sensibly.

7. Conclusions / Next steps

In this document we have provided an overview of initiatives and tools to evaluate data/asset FAIRness. We have then detailed the inner workings of some tools, explaining how the abstract FAIRness codes are in practice evaluated. The learnings from this task have overtime been communicated to the platform core developers, affecting multiple aspects of the platform design. The current version of the DOME 4.0 platform contains an assessment of datasets returned by external data sources, via the DOME 4.0 connectors. We should encourage the connector developers to adopt the required standards to increase the interoperability within DOME 4.0. Standards are fundamental for any large software development and a level of compliance may be required. This will increase the usefulness and thus the value of DOME 4.0.

7.1 Metadata enrichment case: e.g. PSDS ChASe

We can link this IDs (IRIs) with the raw data provided by The Physical Sciences Data-science (PSDS) Chemical Availability Search (ChASe). We were able to create a serialisation of the database in JSON-LD to be ingested into a triplestore in DOME 4.0. Therefore, all accepted IDs will grow the knowledge graph and link with extra information to any existing asset. The converted data has been shared with partners at an earlier stage.

It is important to use a common ontology for known unique IDs as e.g. the DOME 4.0 Ecosystem ontology. This is a goal for the entire system. Commercial information will be reached by adding information linked to such IDs.

7.2 Character Encoding issues

A warning going further. Character encoding issues may surface when fetching text data from third party sources. People use different computer systems or operating systems which may use different encoding. UTF-8 is the standard (and lingua franca in this context) for the internet.

They contribute data online:

- via UI (possibly with bytes of information, invisible to the user)
- Using tools with the wrong encoding setting or when exporting text.
- via APIs

Eventually this contaminated information gets stored in the database of a large repository. This may cause all sort of system wide issues.

This is a concern for any project sourcing external text data.

7.3 Other considerations

The task T2.2 investigated samples of data collected by T4.1. However, the metadata that can be feed to to tools such as F-UJI [5] was absent.

We need to reconsider the suitability of metrics that underpin F-UJI (those inherited from RDA works) and on the need to develop our own metrics (registration in FAIRsharing [9] is possible).

Liaison of T2.2 with T3.2 continued, with more clear separation of concerns and separation of work. One innovative area could be exploring an opportunity for developing specific semantic assets – "FAIRness vocabulary" or "FAIRness ontology" – and their integration in the information model developed by T3.2.

It will be conceptually important, also valuable for DOME positioning among other platforms, to consider T2.2 from the data valorisation point of view, which was feedback on our presentation in the KeXS workshop [10]. In practical terms, this could mean measuring FAIRness of data to be ingested in DOME as "entry control", then measure data FAIRness when the data is ingested and potentially enriched with quality metadata, links and annotations. If we can get evidence this way that DOME raises FAIRness of data ingested, this could be one way to prove the DOME platform value; this could be a part of the DOME "value proposition".

An interesting topic to pursue would be measuring FAIRness not "intrinsically" by applying tools to data assets, but "extrinsically" which in turn could have two major flavours:

- assessment of repositories and "trusting" them by assignment of "default" FAIR metrics to data originating from them,
- looking into data search / access requests and data downloads statistics as evidence of data reuse that gives an indication that data is FAIR (in the eye of a data consumer, as she found a way to Find, Access and Reuse data – perhaps Interoperate with it, too, if we can capture traces of certain data integrations with other data).

8. Lessons learnt

8.1 Measuring FAIRness

FAIRness can be measured. However, it is hard to find absolute measurements and things remain relative (e.g., to a given framework/community), particularly when we consider semantics. In practice, syntactics is unavoidable in technical solutions. The provided scores and percentages should be taken with a pinch of salt.

8.1.1 Issues with IRIs

The metadata must contain valid IRIs (or eventually resolve to a valid ones). Otherwise, tools will ignore that part of the data. If the IRI is a URL and more data is expected but the fetching fails, there will be another information gap.

If the IRI is valid but not what is semantically expected, the situation may be worse depending on how the tools process the IRI.

If the IRIs do not use global unique identifiers (inc. hash digest) or PIDs they will eventually break the graph creating gaps which could be critical.

An example of absolute in the sense that if X over Y IRIs are valid, we have and absolute measure (inc. percentages) only in terms of valid IRIs. If 5/10 of all IRIs in a document are valid that is more informative that stating 50%, so we know how many failed.

IF X over Y IRIs are part of a valid PID scheme then we can conclude and absolute measure of "belonging to that set". Not all PID schemes are formal schemes. DOI IRIs are an example of a formal PID scheme.

These issues will prevent tools from reaching further stages while the rest of the serialisation may be correct.

8.1.2 Accepted ontologies

If the metadata uses ontologies that are not in the list supported by the tool the scores will be low and it may not be representative.

Also, more "supported ontologies" (let us say number of namespaces, cardinality) does not correlate with more FAIRness in any way. However, it measures something different without necessarily putting a label (semantic intensity?).

The quality of the ontologies or their importance is subjective. Also, ontologies from a scientific domain may be preferred over other candidates. It may not be completely obvious which ones should feature.

Another issue is the level of compliance with a standard. Here percentages are also not ideal. Moreover, the difficulty implementing the standards may require a progressive approach. This involves cherry-picking (think e.g. properties) as much as with namespaces.

8.1.3 Automation

Automating the assessment of many records is challenging but desirable. It is also quite important to automatically fix most things if possible (non-trivial). This means the assessment itself needs to be processable to a large degree (non-trivial). We must reduce human intervention to the strictly necessary.

8.2 Beyond connectors: A universal plug-in system

Based on the DOME 4.0 showcase 7 experience, diverse sources can be integrated seamlessly into a coherent JSON-LD output for each record. The sources could be via DOME 4.0 or external sources, balancing between them as desired.

One of the main advantages of DOME 4.0 is that it abstracts away different API interfaces via its public API. Different connector handles their connection to external services on. This aggregate to a library of access within DOME 4.0. However, this can be further improved by *decoupling* the connectors from DOME 4.0.

Connector developers are currently responsible for providing provenance data and standards compliance, but we may streamline these aspects into a plugin system. There may be standardised:

- **Inputs**: URL centric. (as IDs). JSON-LD payload in lieu of any API SDL. We are documenting the queries to third party APIs.
- **JSON Pre-processin**g done in *JSONPath*. (For sites with JSON responses)
- **Outputs**: (Permanent ID associated with JSON-LD payload)
- The plugin should be of a **drop-in** nature. It should include metadata to automatically report capabilities. We may also use the supported introspection capabilites of the language used, if applicable.
- No prior Knowledge of internal DOME 4.0 code should be needed.

This may be a lot of work but could be worth it. Furthermore, making it a formal standard from a known body may be considered.

9. Deviations from Annex 1

There are no deviations from Annex 1.

10. References

[1] FAIR metrics and measurement tools

[2] [2] RDA FAIR Data Maturity Model Working Group

[3] [3] FAIRsFAIR project. https://www.fairsfair.eu/

[4] [4] EOSC-Synergy project. https://www.eosc-synergy.eu

[5] [5] F-UJI Automated FAIR Data Assessment tool. https://www.fairsfair.eu/f-uji-automated-fair-data-assessment-tool

[6] [6] FAIR Evaluator https://github.com/EOSC-synergy/FAIR_eva

[7] [7] FAIRsharing service. www.fairsharing.org

[8] [8] FAIRsharing FAIR Evaluation Services. https://fairsharing.github.io/FAIR-Evaluator-FrontEnd/

[9] [9] FAIRsharing list of maturity indicators. https://fairsharing.org/standards/?q=&selected_facets=type_exact:metric

[10][10] KeXS workshop. https://ontocommons.eu/news-events/events/creating-knowledge-exchange-space-data-management-and-documentation-kexs-0

[11][11] EOSC Task Force on FAIR Metrics and Data Quality charter. https://www.eosc.eu/sites/default/files/tfcharters/eosca_tffairmetricsanddataquality_draftcharter_20210614.pdf

11. Acknowledgement

The author(s) would like to thank the partners in the project for their valuable comments on previous drafts and for performing the review.

Project partners:

#	Type	Partner	Partner full name
1	SME	CMCL	Computational Modelling Cambridge Limited
2	Research	FHG	Fraunhofer Gesellschaft zur Förderung der Angewandten Forschung E.V.
3	Research	INTRA	Intrasoft International SA
4	University	UNIBO	Alma Mater Studiorum – Universita di Bologna
5	University	EPFL	Ecole Polytechnique Federale de Lausanne
6	Research	UKRI	United Kingdom Research and Innovation
7	Large Industry	SISW	Siemens Industry Software NV
8	Large Industry	BOSCH	Robert Bosch GmbH
9	SME	UNR	Uniresearch B.V.
10	Research	SINTEF	SINTEF AS
11	SME	CNT	Cambridge Nanomaterials Technology LTD
12	University	UCL	University College London

This document is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 953163. It is the property of the DOME 4.0 consortium and do not necessarily reflect the views of the European Commission.

12. Table of Abbreviations

Abbreviation	Explanation
API	Application Programming Interface.
COAR	Confederation of Open Access Repositories.
CSVW	W3C CSVW Namespace Vocabulary for tabular data on the web.
DCAT	W3C Data Catalog Vocabulary.
DCAT-AP	Pan-European governmental DCAT profile for public institutions.
DCAT-US	US government DCAT profile for public institutions.
DOI	Digital Object Identifier from CrossRef.
DQV	W3C Data Quality Vocabulary.
FAIR	Findable, Accessible, Interoperable, Reusable.
FOSS	Free and Open-Source Software.
I/O	Input/Output.
InChI	International Chemical Identifier.
InChIKey	Hashed version of the InChI
IRI	Internationalised URI. RFC 3987.
ISO 17369:2013	
ISO 639-1	Two characters language ID.
ISO 639-2	Three characters language ID.
ISO 8601	Datetime standard (RFC 3339).
ISO/IEC 5962:2021	
IUPAC	International Union of Pure and Applied Chemistry
JSON	JavaScript Object Notation.
JSON-LD	W3C JSON for Linked Data.
JSONPath	A JSON query standard. RFC 9535.
KG	Knowledge Graph.
MD5	Hash algorithm. RFC 1321.
OKF	Open Knowledge Foundation.
ORCID	Open Researcher and Contributor ID
PID	Permanent Identifier.
PROV-O	W3C Provenance Vocabulary.
RBAC	Role Based Access Control.
RDF	Resource Description Framework.
Schema.org	Semantic linked data schemas.
SDMX	Statistical Data and Metadata Exchange
SHA	Family of hash algorithms. RFC 4634.
SMILES	Simplified Molecular-Input Line-Entry System.
SPDX	Software Package Data Exchange.
STFC	Science and Technology Facilities Council.
Unicode	Universal Coded Character Set.
URI	Uniform Resource Identifier.
URL	Uniform Resource Locator.
UTF-8	Superset of ASCII.
W3C	Word Wide Web Consortium.

Annex 1