M4.3 - Specification of semantic artefact description

Alejandra Gonzalez-Beltran, Antony Wilson, Biswanath Dutta, Daniel Garijo, Clement Jonquet, Yann Le Franc, María Poveda-Villalón

FAIR-IMPACT
Expanding FAIR solutions across EOSC

Project Title	Expanding FAIR solutions across EOSC
Project Acronym	FAIR-IMPACT
Grant Agreement No.	101057344
Start Date of Project	2022-06-01
Duration of Project	36 months
Project Website	https://fair-impact.eu/

1 M4.3 - Specification of semantic artefact description

Work Package	**WP 4, Metadata and Ontologies**
Lead Author (Org)	**Alejandra Gonzalez-Beltran[1] (UKRI-STFC), Antony Wilson[2] (UKRI-STFC)**
Contributing Author(s) (Org)	**Biswanath Dutta[3] (ISI), Daniel Garijo[4] (UPM), Clement Jonquet[5] (INRAE), Yann Le Franc[6] (eSDF), María Poveda-Villalón[7] (UPM)**
Due Date	**2024-02-29**
Date	**2024-02-29**
Version	**V1.0**
DOI	**10.5281/zenodo.10725304**

Dissemination Level

X	PU: Public
	PP: Restricted to other programme participants (including the Commission)
	RE: Restricted to a group specified by the consortium (including the Commission)
	CO: Confidential, only for members of the consortium (including the Commission)

[1] https://orcid.org/0000-0003-3499-8262
[2] https://orcid.org/0000-0001-7336-4823
[3] https://orcid.org/0000-0003-3059-8202
[4] http://orcid.org/0000-0003-0454-7145
[5] https://orcid.org/0000-0002-2404-1582
[6] http://orcid.org/0000-0003-4631-418X
[7] https://orcid.org/0000-0003-3587-0367

Funded by the European Union

Funded by
the European Union

2 Versioning and contribution history

Version	Date	Author	Notes
0.0	2022-2023	Participants in task T4.2.3	Contributions to work on T4.2.3 that are summarised in this milestone, as documented in meeting notes and GitHub repositories
0.1	Jan-Feb 2024	Alejandra Gonzalez-Beltran, Antony Wilson	Document outline and first draft of sections
0.3	Feb-2024	Biswanath Dutta, Daniel Garijo, Clement Jonquet, Yan Le Franc, María Poveda-Villazón	Contributions to the document and review of all the content
1.0	Feb 2024	Alejandra Gonzalez-Beltran, Antony Wilson	Final review addressing comments/feedback

Disclaimer

FAIR-IMPACT has received funding from the European Commission's Horizon Europe funding programme for research and innovation programme under the Grant Agreement no. 101057344. The content of this document does not represent the opinion of the European Commission, and the European Commission is not responsible for any use that might be made of such content.

3 Table of Contents

List of Diagrams

Funded by the European Union

Terminology

Terminology/Acronym	Description
API	Application Programming Interface
DCAT	Data Catalogue Vocabulary
EOSC	European Open Science Cloud
FAIR	Findable Accessible Interoperable and Reusable
RDA	Research Data Alliance
RDF	Resource Description Framework
LOT	Linked Open Terms
MOD	Metadata for Ontology Description and Publication
OWL	Web Ontology Language
SA	Semantic Artefact
SKOS	Simple Knowledge Organization System
SSSOM	Simple Standard for Sharing Ontology Mappings
TTL	Terse RDF Triple Language
VSSIG	Vocabulary and Semantic Services Interest Group
WP	Work Package

Funded by the European Union

4 Introduction

Semantic artefacts (SA) are key for the description of data and for making data FAIR (findable, accessible, interoperable and reusable) [1]. SA is a broader term to include ontologies, terminologies, taxonomies, thesauri, vocabularies, metadata schemas and semantic standards. Describing SAs is fundamental to make them FAIR themselves. The Metadata for Ontology Description and Publication (MOD[8]) was developed to provide the vocabulary required to describe ontologies, and Semantic Artefacts in general. The Data Catalogue Vocabulary (DCAT[9]) [2] was designed to describe datasets and resources that can be catalogued.

This milestone presents a DCAT-based standard for description of Semantic Artefacts and their catalogues, building on the MOD vocabulary as well as recommendations and outcomes from the FAIRsFAIR project[10] and the Research Data Alliance Vocabulary and Semantic Services Interest Group (RDA VSSIG[11]). This milestone also makes a distinction between the MOD specification and a series of mappings with other vocabularies, presented in a machine-actionable way. Last but not least, we describe a methodology for MOD profiles and how to formalise them in a machine-actionable and composable way.

The next step (deliverable D4.3) related to this milestone will be to specify a common Application Programming Interface (API) for interoperability of SA catalogues in the European Open Science Cloud (EOSC) ecosystem and beyond, building from MOD descriptions of SAs. The API for SA-catalogues will enable interoperability and unified access to their content, enabling seamless querying and use by stakeholders independent of domain. The API will be adopted by FAIR-IMPACT T4.2's use case SA-catalogues and it will be publicly available for other catalogues to deploy; via the API, other registries could consume content from multiple SA-catalogues. The implementation of this API will be the topic of an upcoming FAIR-IMPACT Open Call.

5 Description of the Milestone

This milestone presents a first version of a specification for Semantic Artefacts descriptions, relying on a new version of the Metadata for Ontology Description and Publication Ontology (MOD) (version 3.0).

The milestone emphasises on the importance of Semantic Artefacts descriptions in the context of producing data that follows the FAIR guiding principles, and presents an historical perspective of the evolution of the MOD vocabulary, which has been originally developed [3] in a convergence approach considering three aspects: by exploring existing vocabularies that

[8] MOD: https://github.com/FAIR-IMPACT/MOD
[9] DCAT version 3: https://www.w3.org/TR/vocab-dcat-3/
[10] FAIRsFAIR project: https://fairsfair.eu/
[11] RDA VSSIG: https://www.rd-alliance.org/groups/vocabulary-services-interest-group.html

Funded by the European Union

are useful to describe SAs, by analysing existing SAs descriptions and by studying the description properties offered by SA-catalogues

Here we present the latest design decisions that lead to MOD version 3.0, which has also been complemented with a set of mappings to other vocabularies and representation of MOD profiles.

There are other tasks and their milestones/deliverables within the FAIR-IMPACT project that are related to this work, as follows:

- M4.2 Processes and tools to engineer FAIR Semantic Artefacts [3];
- M5.3 Semantic Artefact assessment methodology [5]
- Upcoming D4.5 Guidelines and methodology to create, document and share mappings and crosswalks
- Upcoming D4.3 Specification of shared metadata description of Semantic Artefacts and their catalogues, including common reference API

5.1 Role of the milestone

This milestone has been driven by three objectives:

i) specify a new version of the MOD vocabulary as an extension of the DCAT vocabulary by addressing outstanding design issues,

ii) a clear identification between MOD elements and mappings to other vocabularies, by representing the mappings using the SSSOM standard,

iii) present a machine-actionable representation of an MOD profile.

5.1.1 Means of verification

The required means of verification for this milestone is the publication of this report together with the artefacts presented in the related GitHub repositories:

i) MOD repository: https://github.com/FAIR-IMPACT/MOD

ii) MOD mappings: https://github.com/FAIR-IMPACT/MOD-mappings

iii) An MOD profile: https://github.com/FAIR-IMPACT/MOD-FAIRsFAIR-profile

Funded by the European Union

6 FAIR Semantic Artefacts

This milestone relies on the definition of Semantic Artefact historically proposed by the H2020 FAIRsFAIR project[12], and endorsed by the EOSC Interoperability Framework [6], as follows: "*Semantic Artefact is defined as a machine-actionable and readable formalisation of a conceptualisation enabling sharing and reuse by humans and machines. These artefacts may have a broad range of formalisation, from loose set of terms, taxonomies, thesauri to higher-order logics, and include the concepts/terms/classes constituting these. Moreover, Semantic Artefacts are serialised using a variety of digital representation formats, e.g., RDF Turtle, OWL- RDF, XML, JSON-LD*" [7].

The description of SAs is an important element required to make a Semantic Artefact FAIR, following the findability F2 principle "Data are described with rich metadata" and reusability R1 principle "(Meta)data are richly described with a plurality of accurate and relevant attributes".

The FAIRsFAIR project work on FAIR Semantic Artefacts [7], [8], [9] focused on providing practical recommendations on how to make SAs FAIR, and these best practices and recommendations were discussed extensively with the community. To implement one of the recommendations (P-Rec 3 "A common minimum metadata schema must be used to describe semantic artefacts and their content"), FAIRsFAIR also developed and presented a minimal schema for describing Semantic Artefacts [6][10]. In addition, the project presented a service architecture for harmonising SA repositories metadata with this minimum metadata schema and published the resulting harmonised metadata in a FAIR Data Point. This service architecture, called FAIRCat, has been developed from the Federated FAIR Data Space technology, developed in EOSC Pillar [11]. This work has been done in collaboration with a dedicated discussion group with the RDA Vocabulary and Semantic Service Interest Group. During a dedicated final workshop, the model based on DCAT and MOD2.0 was presented to the various communities and was evaluated according to a simple use-case i.e. searching and retrieving ontologies. Based on this use-case, the 70+ participants from 17 communities identified the metadata elements of the model which should be mandatory, recommended and optional. The mandatory set of fields representing the minimum metadata model. Based on voting results, we created two machine actionable descriptions of the minimum metadata schema, an OWL version and a SHACL version. These machine actionable versions were used to aggregate 5 different SA repositories with the FAIRCat platform and to publish the harmonised metadata into a FAIR Data Point[13]. In this milestone, we present the FAIRsFAIR work[14] as a profile for the MOD specification (see Section 7.3).

Other initiatives also worked on recommendations on making SAs FAIR [12], [13], [14], [15] and tools [16], [17] to verify their FAIRness.

[12] https://fairsfair.eu/
[13] https://fdp-semdcat-app.vps.esciencedatafactory.com/
[14] https://github.com/FAIRsFAIR/SemanticDCAT-AP/

Funded by the European Union

Within the FAIR-IMPACT project, we are working on different aspects of Semantic Artefacts building on previous work:

- This work (M4.3) focuses on a specification for describing Semantic Artefacts, a presentation of a couple of profiles of such specification, and mappings between the specification and other relevant vocabularies
- M4.2 [4] proposes a FAIR-by-design methodology for developing vocabularies and ontologies (formalised in the RDF/RDFS and OWL representation languages) that can be extended to address other types of SAs.
- M5.3 [5] presents a generic methodology for assessing the FAIRness of Semantic Artefacts that groups categories of tests consistently; this is achieved by building on the Linked Open Terms (LOT) [18] methodology, which splits the assessment into smaller parts, considering code, content, metadata and a new module on FAIR assessment.

7 The Metadata for Ontology Description and Publication (MOD) Ontology

7.1 MOD Previous Work

In this section, we present a historical view on how MOD was developed. The initial version (MOD 1.0[15]) was presented at the Dublin Core conference in 2015 [19]. It introduced the Metadata for Ontology Description and publication (MOD) analysing the need for a controlled vocabulary for cataloguing ontologies, its design principles and methodology. The article reported the survey results of ontology users, which consisted of ontology researchers and practitioners with diverse educational backgrounds, their search behaviour, search criteria and parameters. It also reported on the metadata vocabulary usage by the thirteen ontology libraries (now called Semantic Artefact catalogues in EOSC), such as BioPortal[16], EBI OLS[17], AberOWL,[18] LOV[19], etc. for ontology description.

MOD 1.2 [3] revised the previous version by considering 23 existing metadata vocabularies with relevant properties for describing Semantic Artefacts, as well as studied metadata usage analytics within ontologies and ontology repositories from three different perspectives: analysis of existing metadata vocabularies to describe ontologies; analysis of present use of metadata vocabularies for ontology descriptions (more in section 8.3); and analysis of metadata representation within ontology repositories. This new version of MOD (1.2) proposed 88 properties intended as a vocabulary for the annotation and description of semantic artefacts to be used by ontology engineers and ontology libraries. The article

[15] https://www.isibang.ac.in/ns/mod/1.0/
[16] https://bioportal.bioontology.org/
[17] www.ebi.ac.uk/ols
[18] http://aber-owl.net
[19] https://lov.linkeddata.es

Funded by the European Union

documented the selection criteria for inclusion of terms from other existing vocabularies to MOD.

At the time, discussions with the RDA VSSIG started to consider a community-driven standardisation effort, collaborative design and adoption of MOD future versions. In addition, AgroPortal [20] implemented MOD 1.2 and identified other properties not already available in that version of MOD [21]. MOD 1.4[20] was produced by incorporating the new properties identified by the AgroPortal work bringing MOD's list of description properties to 127.

In 2020, as part of the RDA VSSIG and FAIRsFAIR work, a proposition for MOD 2.0 was made that considered relating MOD classes to the DCAT vocabulary for description of data catalogues and catalogues of resources, from DCAT version 2.0. This work was presented and discussed in multiple RDA and FAIRsFAIR meetings[21] in 2020 and 2021 then later described in [10]. Initial versions of MOD defined the *mod:Ontology* class that was later replaced by *mod:SemanticArtefact*. In MOD 2.0, the properties from the other vocabularies not maintained any more by their original creators (*i.e.* OMV, DOOR and VOAF) were redefined within the MOD namespace, which since MOD 2.0 relies on the W3ID platform for persistent identifiers (https://w3id.org/mod).

7.2 MOD Version 3.0

In this section, we describe how the new version of MOD (version 3.0)[22] was designed and implemented, by formalising MOD relationship with the DCAT vocabulary for resource catalogues and clearly distinguishing between the elements that are part of the MOD vocabulary definition and additional terms from other ontologies that were mapped to MOD, in order to facilitate:

i) the specification of a new version of the MOD vocabulary as an extension of the DCAT vocabulary by addressing outstanding design issues,

ii) a clear identification between MOD elements and mappings to other vocabularies, by representing the mappings using the SSSOM standard,

iii) present machine-actionable representations of MOD 3.0 profiles.

7.2.1 Design Decisions

The WP4 members involved in this task met on multiple occasions to reach consensus on the approach to be taken around multiple design issues, which were documented in the open GitHub repository using the Discussions feature[23]. This section describes each of the design issues and the decisions taken.

[20] MOD versions: https://github.com/FAIR-IMPACT/MOD/tree/main/versions
[21] https://fr.slideshare.net/jonquet/presentation-fairsfair-workshop-june-2021
[22] https://github.com/FAIR-IMPACT/MOD/
[23] MOD Design Discussions: https://github.com/FAIR-IMPACT/MOD/discussions

Funded by the European Union

7.2.1.1 What class of DCAT does mod:SemanticArtefact specialise?

The discussion on this designed issue was captured online[24] and determines what is the DCAT class that *mod:SemanticArtefact* should specialise. Two DCAT classes were identified as possible parent classes of the *mod:SemanticArtefact*, the *dcat:Resource* and *dcat:Dataset*. A *dcat:Resource* is "something that can be catalogued" and a *dcat:Dataset* is a "collection of data".

We had multiple discussions on this design issue (FAIR-IMPACT project meetings related to T4.2.3), the discussion above, as well as discussion in the DCAT issue w3c/dxwg#1576[25].

While some team members, who considered Semantic Artefact as a collection of concepts and relations, were happy with deriving *mod:SemanticArtefact* from *dcat:Dataset*, others preferred to distinguish between a dataset and a Semantic Artefact, so preferred to derive from *dcat:Resource* that is DCAT's extension point.

However, for Semantic Artefacts, we still want to take advantage of the dichotomy between abstract entity and distribution provided by the DCAT modelling, so the suggestion was to still use the *dcat:distribution* property for Semantic Artefacts. In effect, it was decided to explicitly specialise *dcat:Resource*, but using *dcat:distribution* implies that *dcat:SemanticArtefact* can be inferred as a *dcat:Dataset* in the current interpretation of DCAT. This seemed to be a good compromise with those who disagree to make the subclass explicit, while we wait for a resolution on the discussion with the DCAT group.

Thus, we represented:

```
mod:SemanticArtefact
        a rdfs:Class ;
        a owl:Class ;
        rdfs:subClassOf dcat:Resource ;
```

and still use the distribution predicate to refer to multiple representations of the Semantic Artefact. This representation is available in MOD 3.0.

7.2.1.2 How to capture the notion of distribution?

The discussion on this designed issue was captured online[26]. Three different proposals were made.

Distribution - Proposition 1

- *mod:SemanticArtefact* specialises *dcat:Resource* only
- *mod:SemanticArtefactDistribution* is defined independently of *dcat:Distribution*

[24] Design issue #1: https://github.com/FAIR-IMPACT/MOD/discussions/34
[25] https://github.com/w3c/dxwg/issues/1576
[26] Design issue #2: https://github.com/FAIR-IMPACT/MOD/discussions/35

Funded by the European Union

- If one connects other objects (e.g., *schema:Dataset*, *void:Dataset*) to *dcat:Dataset* then a *mod:SemanticArtefact* can not benefit from their properties.

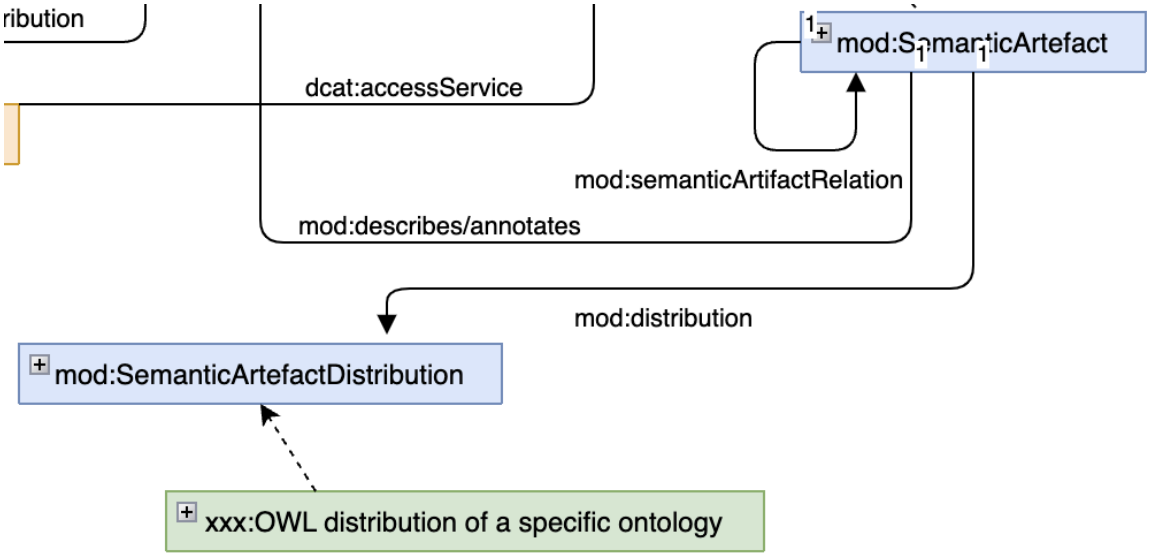

Diagram 1. Distribution - Proposition 1

Distribution - Proposition 2

- SA becomes a subclass of dcat:Dataset and thus SADistribution becomes a subclass of *dcat:Distribution*
- The idea of keeping the distribution is still there but is a SA a dataset ?
- If we connect *schema:Dataset* to *dcat:Dataset* then a SA can still benefit from the properties
- We probably do not need the *mod:distribution* property anymore

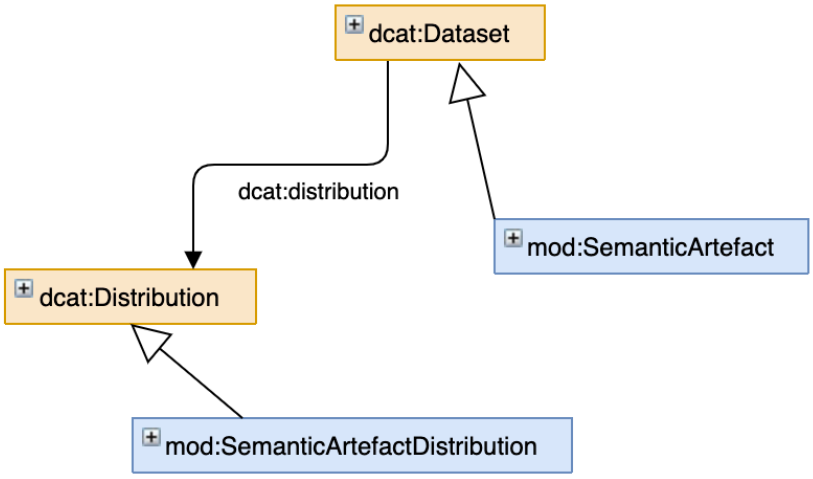

Diagram 2. Distribution - Proposition 2

Funded by the European Union

For this proposition, the properties for *mod:SemanticArtefactDistribution* need to be considered (see Diagram 3 listing those properties).

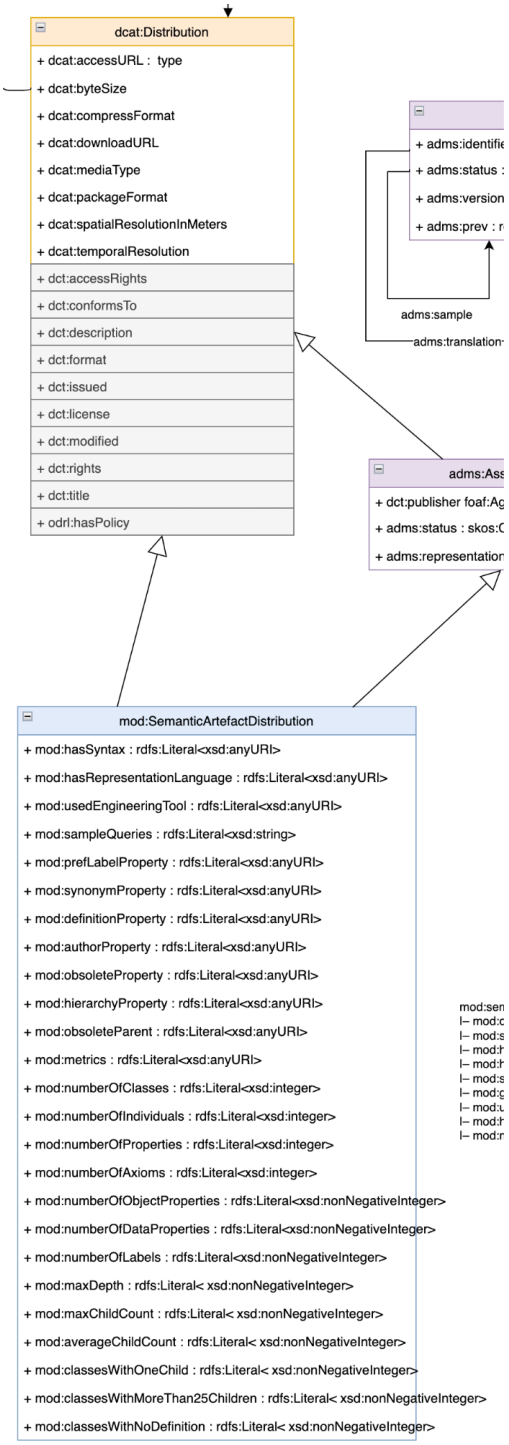

Diagram 3. Distribution - List of Properties

Distribution - Proposition 3

- Get rid of SemanticArtefactDistribution completely, but then it does not make sense to follow DCAT anymore.

The discussion around this design issue is related to the previous design issue, where *mod:SemanticArtefact* is now represented as a subclass of *dcat:Resource*, and not *dcat:Dataset* explicitly, even though the use of the property *dcat:distribution* is still encouraged (and with the current DCAT representation, this implies that *mod:SemanticArtefact* will be inferred as a *dcat:Dataset*).

The group agreed that it is important for MOD to follow the approach of DCAT and distinguish between SA and SADistribution, thus *mod:SemanticArtefactDistribution* will extend *dcat:Distribution*.

The properties listed for *mod:SemanticArtefactDistribution* were analysed to make sure they make sense at the distribution level. The property *dcat:distribution* will be used, and there is no need for a *mod:distribution* property. Thus, Proposition 2 was accepted.

7.2.1.3 What other classes (not in DCAT) can mod:SemanticArtefact specialise?

The discussion on this designed issue was captured online[27]

Semantic Artefact Specialisation - Proposition 1

- SA specialises all the pre-existing semantic objects (e.g., *owl:Ontology*) as it was done in MOD1.4 with the object *mod:Ontology*. Then, all the properties of all the objects are available.

[27] Design issue #3: https://github.com/FAIR-IMPACT/MOD/discussions/36

Funded by the European Union

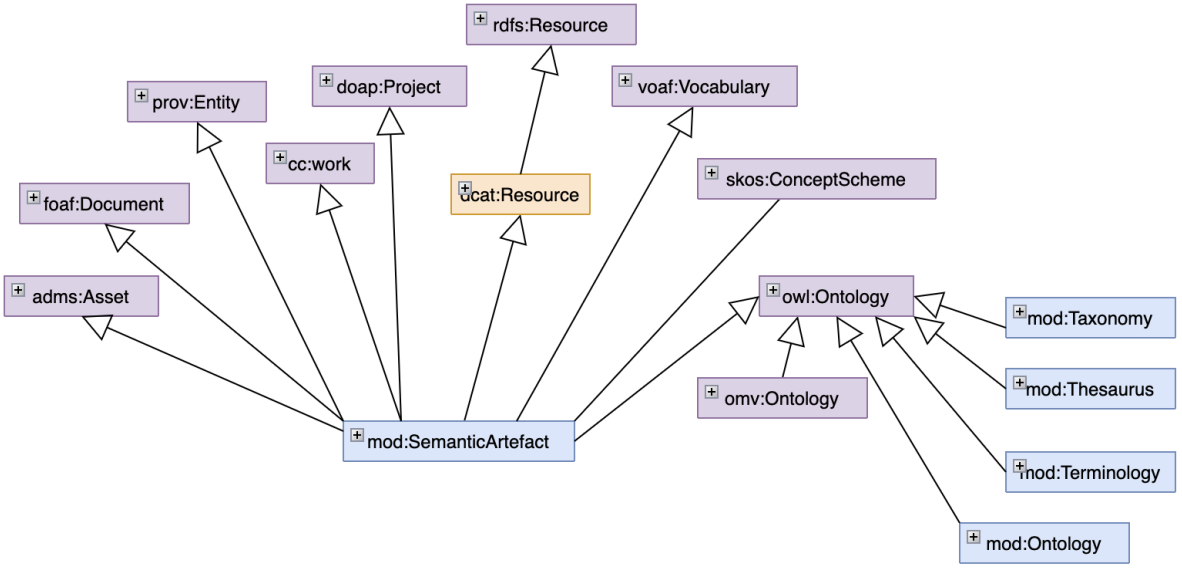

Diagram 4. Semantic Artefact Specialisation - Proposition 1

Semantic Artefact Specialisation - Proposition 2

- SA specialises even more pre-existing semantic objects but not the pre-existing ones that were representing a type of SA (*owl:Ontology*, *skos:ConceptScheme*, *voaf:Vocabulary*)

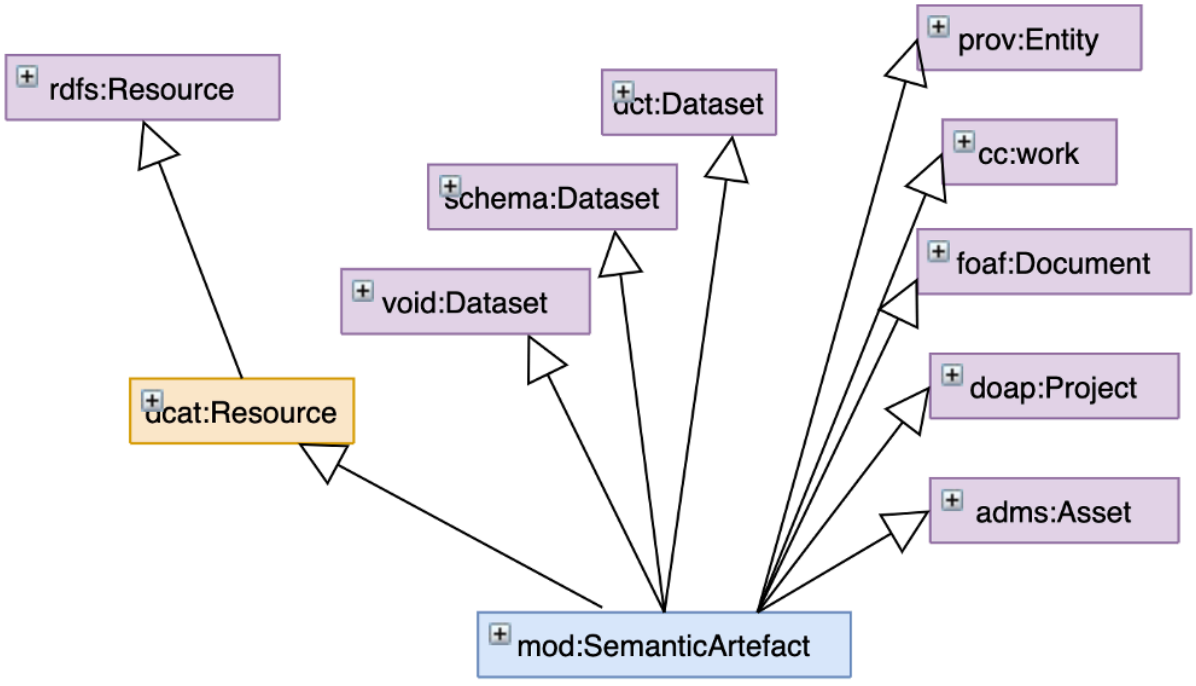

Diagram 5. Semantic Artefact Specialisation - Proposition 2

Funded by the European Union

Semantic Artefact Specialisation - Proposition 2-bis

- The same as proposition 2 but SA supersedes the pre-existing ones.

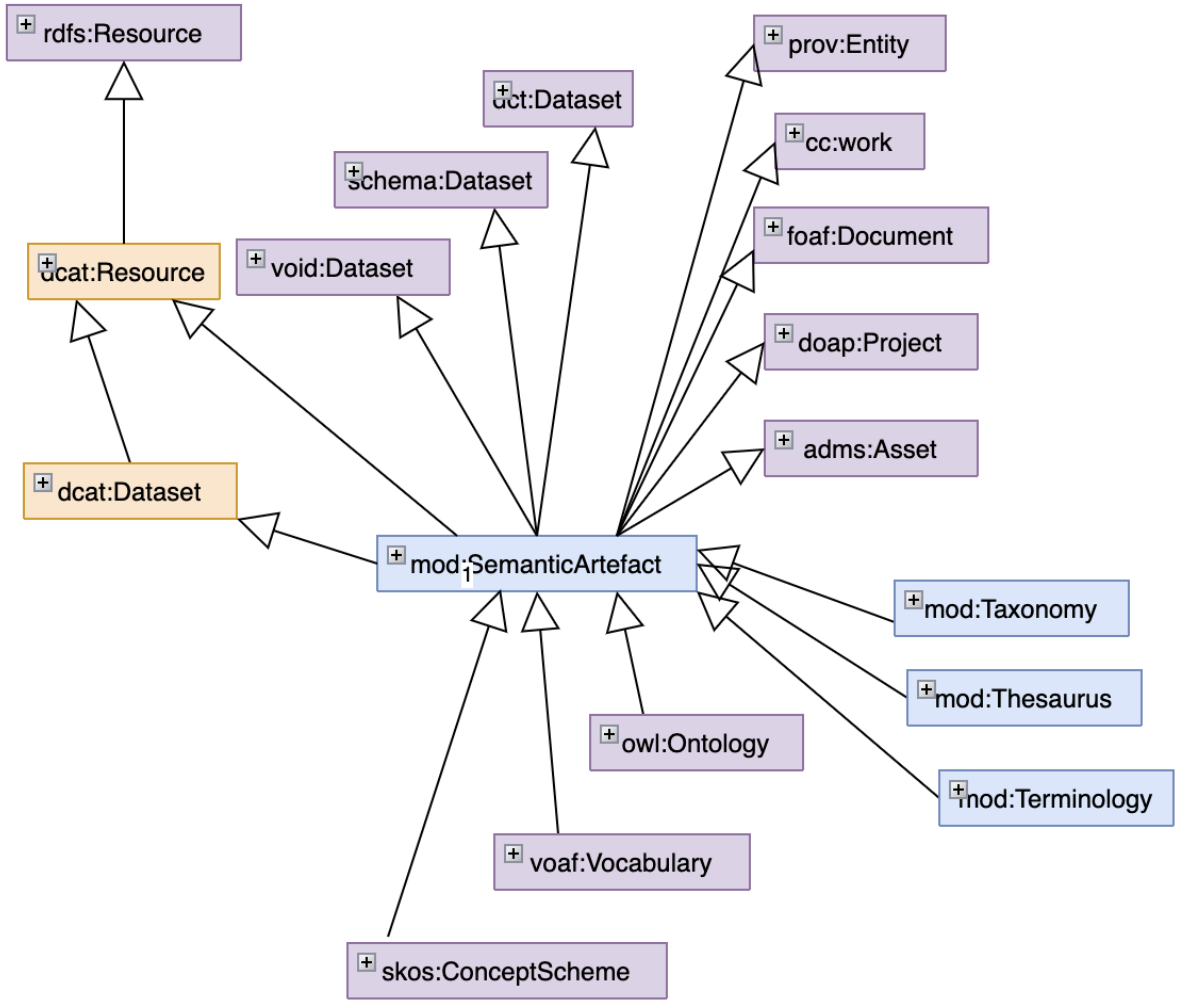

Diagram 6. Semantic Artefact Specialisation - Proposition 2-bis

The conclusion around this design issue considered:

- *mod:SemanticArtefact* must not specialise *owl:Ontology* or *rdfs:Resource*
- Any other specialisation, when considering other vocabularies, should be discussed as part of an MOD profile and/or MOD mappings with other vocabularies, rather than being part of MOD definitions.

7.2.2 Implementation

The new MOD version 3.0 was developed and hosted in the GitHub repository: https://github.com/FAIR-IMPACT/MOD

Funded by the European Union

The pre-release for MOD version 3.0 is available at: https://github.com/FAIR-IMPACT/MOD/releases/tag/v3.0

It is available using the turtle format as default syntax, incorporating all the conclusions from the design issues discussion.

7.3 MOD Profiles

We adopted the definition of profiles by the Profiles Vocabulary [22], which refers to "A [data/application] specification that constrains, extends, combines, or provides guidance or explanation about the usage of other [data/application] specifications".

Here we consider MOD as a specification, “A basis for comparison; a reference point against which other things can be evaluated.” [22], and we have defined a profile to demonstrate how to use the Open Geospatial Consortium (OGC) building blocks approach[28,29].

The OGC building blocks approach has its foundation on the reusability of a set of requirements, so that they can be applicable to other contexts. The basic principles are[30]: i) a building block a standard or a part of a standard, ii) one or more building blocks can be integrated into new applications, iii) each building block represents a testable interface component.

To demonstrate this approach with profiles, we produced a FAIRsFAIR MOD profile, which is defined by considering the results in the vote organised within a FAIRsFAIR workshop for the endorsement level mandatory-recommended-optional for each property to describe a semantic artefact, as reported in [10]. The article refers to Consensus as:

consensus = (0.333 + percentage of votes for the winning option (mandatory, recommended or optional) − sum of percentages of the non-winning options) / 1.333

To create the FAIRsFAIR MOD profile[31], we considered as required those properties that have consensus > 50%, for the mod:SemanticArtefact class, as no properties had consensus >50% for mod:SemanticArtefactDistribution.

For future work, we will continue to explore this approach, and look to implement other potential MOD profiles[32], and explore ways to combine them. We will also explore the connection with previous work, such as the FAIRsFAIR SemanticDCAT-AP[33] work.

[28] https://blocks.ogc.org/
[29] https://github.com/opengeospatial/bblock-template
[30] https://blocks.ogc.org/
[31] https://github.com/FAIR-IMPACT/MOD-FAIRsFAIR-profile
[32] https://github.com/FAIR-IMPACT/MOD/discussions/56
[33] https://github.com/FAIRsFAIR/SemanticDCAT-AP/

Funded by the European Union

8 MOD Mappings

8.1 Introduction

Work has been undertaken to produce machine-readable mappings which will enable users to map ontology metadata properties from existing metadata vocabularies and standards to MOD (see more details in Section 7), formalising previous mappings that were part of previous versions of MOD and doing an additional analysis.

As part of previous work, a TTL description of MOD included more than 200 mappings from MOD properties to other properties. Note that in most of the cases, the property suggested by MOD is not declared within the MOD namespace, so the mappings are representing MOD point of view on which properties could be used to encode an information. See Section 8.2 for a description of how these mappings were formalised.

For example, the *mod:status*, declared in the MOD namespace is proposed by MOD to encode the "The status of the current version of the ontology (alpha, beta, production, retired)." We identified 3 equivalent properties: *adms:status*, *idot:status* and *omv:status* respectively from the ADMS vocabulary, IDOT vocabulary and the historical OMV ontology. We thus have to represent 3 mappings each of them with a source property within the MOD namespace.
On the other hand, MOD suggests to use the *dcterms:creator* property to encode the "An entity primarily responsible for making the resource." and does not declare any new property in the MOD namespace. We identified 7 equivalent properties: *doap:maintainer, foaf:maker, omv:hasCreator, pav:authoredBy, pav:createdBy, prov:wasAttributedTo and schema:author* from multiple various metadata vocabularies. We thus have to represent 7 mappings but none of them are with a source property in the MOD namespace. In other words, MOD proposes these mappings that we believe are relevant in the context of semantic artefact descriptions; but might not be relevant in any other context.

As part of M5.3, Daniel Garijo carried out an audit of ontology metadata [5]. Subsequently, as part of this work, over 600 properties were mapped to MOD [23]. See Section 8.3 for more details on the analysis of ontology metadata and mappings.

Mappings from these two sources have been combined to produce a single set of mappings, which in turn has been converted into a machine-readable format, see Section 8.4. These mappings were to MOD 2.0 terms, as MOD 3.0 was still under development. As part of the ongoing development, the mappings will be updated to point to MOD 3.0 terms.

8.2 Existing MOD Mappings

Previously, as part of MOD version 1.4 then 2.0, a turtle (*TTL)* file was created that included 200+ mappings.[34] In fact, counting the suggested 127 properties, and all the mappings MOD2 included 346 properties that a system could use to manage ontology metadata

[34] https://github.com/FAIR-IMPACT/MOD/blob/main/mod-v2.0_profile.ttl

Funded by the European Union

descriptions. This enabled, for instance, AgroPortal to recognize 346 properties in semantic artefact source files and map them to the metadata model based on the 127 MOD1.4 properties [21].

The mappings existing in the file (MOD ttl[35]) have been analysed and combined with the ontology metadata audit (see next section) and represented using a machine-readable representation of mappings (see Section 8.4).

8.3 Ontology Metadata Audit

As part of the methodology for FAIR semantic artefact assessment [5], a metadata landscape analysis was performed over 1961 unique OWL and RDFS ontologies and 587 SKOS vocabularies in order to determine common metadata practices across different semantic artefacts. This analysis extends the work done by [21] with 805 ontologies taken from different sources (AgroPortal, BioPortal, MMI ORR and miscellaneous online servers).

The analysis retrieved semantic artefacts from popular community catalogues and repositories: Linked Open Vocabularies (1495 ontology and vocabulary versions were found),[36] Archivo (1750),[37] BioPortal (976),[38] EcoPortal (23),[39] IndustryPortal (45),[40] and MedPortal (54).[41] Additional ontologies were found, and subsequently downloaded, by searching in w3id.org (687)[42] and OnToology (160).[43] Since many semantic artefacts were present in more than one registry (e.g., Archivo incorporates many LOV ontologies), the final set consisted of 2784 *files* (with their corresponding metadata). After removing artefacts with duplicate identifiers, the final number of semantic artefacts was reduced to ***1961 OWL ontologies and 587 SKOS vocabularies*** [5].

Next, we performed a metadata audit on the downloaded resources. We counted the number of occurrences of each metadata property per semantic artefact. For example, if an ontology defines three authors, the “author” property would only be counted once, in order to avoid over-representation. The results of this step resulted in a list of 622 unique properties, including properties with typos (e.g., misspelt) or referencing non-resolving vocabularies.

- All properties were manually aligned to MOD 2.0 by three different human annotators, linking all properties (when possible) to the existing MOD property (which itself included mappings to other properties, as explained in previous Section) [23]. Disagreements in the annotation process were resolved by iterating discussion until an agreement was reached. This human-readable mapping is available online:

[35] https://raw.githubusercontent.com/FAIR-IMPACT/MOD/main/mod-v2.0_profile.ttl
[36] https://lov.linkeddata.es/dataset/lov/
[37] https://archivo.dbpedia.org
[38] https://bioportal.bioontology.org/
[39] https://ecoportal.lifewatch.eu/
[40] http://industryportal.enit.fr/
[41] https://medportal.bmicc.cn/
[42] https://github.com/perma-id/w3id.org
[43] https://ontoology.linkeddata.es/

Funded by the European Union

https://github.com/FAIR-IMPACT/MOD-mappings and was used to generate the SSSOM mappings described in Section 8.4.

8.4 Production of MOD mappings with SSSOM

The SSSOM[44] representation language and format has been recommended by FAIR-IMPACT T4.4 to represent the mappings.

"The Simple Standard for Sharing Ontology Mappings (SSSOM) [24] is an initiative to provide a minimal and standard set of elements for the dissemination of mappings between ontology terms, to ensure a reliable interpretation of generated mappings and to enable sharing and data integration between people and applications"[45].

The main items of the data model are subject_id, subject_label, predicate_id, object_id, object_label and mapping_justification. In addition there are provenance fields for the individual mappings, author_id, reviewer_id and mapping_provider.

SSSOM has very few required fields and many optional fields. Only a limited number of the optional fields were used when producing the MOD mapping. One of the outputs of T4.4 will be a list of recommended fields that should be included in SSSOM mappings.

8.4.1 Mappings From the MOD ttl

The mappings from the MOD2.0 *ttl* file[46] were extracted and put into a spreadsheet. Because aligning multiple metadata vocabularies where properties are explicitly defined differently (rdf:property, owl:AnnotationProperty, owl:DataProperty or owl:ObjectProperty) with mixed domains and ranges, MOD authors decided to choose a semantically loose relationship to encode those "mappings". Indeed, the initial relationship used was "dcterms:relation". We have now decided to require a more specific relationship. Assigning new relationships was a manual process. The properties, on either side of the relationship, were resolved and any descriptions of the propertiesterms read, before assigning a new relationship. The relationships used were taken from Simple Knowledge Organization System (SKOS):

- skos:relatedMatch
- skos:closeMatch
- skos:exactMatch
- skos:narrowMatch
- skos:broadMatch

There were 207 mappings extracted from the MOD2.0 *ttl* file. The relationships can be broken down as follows:

47 skos:relatedMatch
60 skos:closeMatch

[44] SSSOM: https://mapping-commons.github.io/sssom/home/
[45] SSSOM specification: https://mapping-commons.github.io/sssom/spec/
[46] https://github.com/FAIR-IMPACT/MOD/blob/main/mod-v2.0_profile.ttl

Funded by the European Union

89 skos:exactMatch
11 skos:narrowMatch

8.4.2 Mappings for the Ontology Audit

The output from the ontology audit included mappings of the properties identified in the source files to MOD 2.0 properties, although the relationship was not defined for the mappings. In this phase 255 mappings from the ontology metadata audit were added to the spreadsheet produced from the MOD *ttl* mappings (Section 8.4.1). From this 37 duplicates were removed. Where there were conflicts the mappings from the MOD 2 ttl were used. The relationships between the remaining properties were then assigned to one of the SKOS relationships listed above.

It should be noted that many of the metadata properties taken from the ontology source files had errors including URI's that did not resolve, properties not present in the ontologies or vocabularies. For a number of these it was easy to spot errors in their URI's, for example including "#" when it should not be there or using a property that is not even in the vocabulary, e.g., in Dublin Core we found "author", rather than "creator". For this phase we mapped properties where it was easy to determine what the property should have been. This still left over 200 terms that have not been mapped. By including erroneous properties in the mappings, we provide a means to use these erroneous properties.

We are showing 2 examples of incorrect URIs in use in ontology source files below. The first one has an incorrect namespace (including "#"), and the second one uses "author" instead of "creator":

- http://purl.org/dc/terms/#license
- http://purl.org/dc/terms/author

8.4.3 Producing a SSSOM Mapping

The SSSOM cookie cutter[47] was used to set up a GitHub repository , MOD-mappings[48], to host the mappings. This repository includes tools to generate a SSSOM file from a tsv file and validate the generated SSSOM file.

The workflow is as follows:

- Production of the mappings is done in a spreadsheet, we have used a Google Sheet.
- The data is exported as a *tsv* file, *mod_mappings.tsv*, which is put in the *sources* directory in the *MOD-mappings* GitHub project. Also in this directory is the *YAML* file *mod_mappings_metadata.yml*, which contains the curie map.
- A new mappings file, *mappings/mod_mappings.sssom.tsv*, is generated using "sh odk.sh make mappings"
- The new mappings file is then validated by running "sh odk.sh make test"

[47] https://github.com/mapping-commons/mapping-commons-cookiecutter
[48] https://github.com/FAIR-IMPACT/MOD-mappings

Funded by the European Union

Included in the repository are Git Hooks to run the validation when a new mapping file is committed.

The outcome of this process is a validated SSSOM file that can be used by other applications.

8.4.4. Future Mapping Work

The main product of this milestone is a collection of mappings to MOD 2.0 terms. With the release of MOD 3.0 these mappings will be updated to use MOD 3.0 terms. In addition, we need to follow the outcomes of T4.4 in terms of the metadata that should be included with the SSSOM mapping.

9 Conclusions and next steps

This milestone is about a specification for describing semantic artefacts based on the MOD vocabulary.

The main outcomes from this milestone are:

i) specify a new version of the MOD vocabulary as an extension of the DCAT vocabulary by addressing outstanding design issues,

ii) a clear identification between MOD elements and mappings to other vocabularies, by representing the mappings using the SSSOM standard,

iii) present machine-actionable representations of MOD profiles.

For the next steps, we will continue to refine the MOD specification, and we will produce an Application Programming Interface for semantic artefact catalogues. This will be described in a deliverable extending this milestone.

Funded by the European Union

10 References

[1] M. D. Wilkinson *et al.*, 'The FAIR Guiding Principles for scientific data management and stewardship', *Sci. Data*, vol. 3, no. 1, Art. no. 1, Mar. 2016, doi: 10.1038/sdata.2016.18.

[2] R. Albertoni, D. Browning, S. Cox, A. N. Gonzalez-Beltran, A. Perego, and P. Winstanley, 'The W3C Data Catalog Vocabulary, Version 2: Rationale, Design Principles, and Uptake', *Data Intell.*, pp. 1–37, Dec. 2023, doi: 10.1162/dint_a_00241.

[3] B. Dutta, A. Toulet, V. Emonet, and C. Jonquet, 'New Generation Metadata Vocabulary for Ontology Description and Publication', in *Metadata and Semantic Research*, vol. 755, E. Garoufallou, S. Virkus, R. Siatri, and D. Koutsomiha, Eds., Cham: Springer International Publishing, 2017, pp. 173–185. doi: 10.1007/978-3-319-70863-8_17.

[4] M. Poveda-Villalon *et al.*, 'M4.2 - Processes & tools to engineer FAIR semantic artefacts', Dec. 2023, doi: 10.5281/ZENODO.10551054.

[5] D. Garijo, M. Poveda-Villalón, P. Flohr, A. Gonzalez-Beltran, Y. le Franc, and M. Verburg, 'M5.3 Semantic artefact assessment methodology', Aug. 2023, doi: 10.5281/ZENODO.8305173.

[6] European Commission. Directorate General for Research and Innovation. and EOSC Executive Board, *EOSC interoperability framework: report from the EOSC Executive Board Working Groups FAIR and Architecture.* LU: Publications Office, 2021. Accessed: Feb. 29, 2024. [Online]. Available: https://data.europa.eu/doi/10.2777/620649

[7] Y. L. Franc, L. Bonino, H. Koivula, J. P.-V. Essen, and R. Pergl, 'D2.8 FAIR Semantics Recommendations Third Iteration', Feb. 2022, doi: 10.5281/ZENODO.6675295.

[8] W. Hugo, Y. Le Franc, G. Coen, J. Parland-von Essen, and L. Bonino, 'D2.5 FAIR Semantics Recommendations Second Iteration', Dec. 2020, doi: 10.5281/ZENODO.5362010.

[9] Y. Le Franc, J. Parland-von Essen, L. Bonino, H. Lehväslaiho, G. Coen, and C. Staiger, 'D2.2 FAIR Semantics: First recommendations', Mar. 2020, doi: 10.5281/ZENODO.5361930.

[10] C. Jonquet, B. Dutta, L. O. B. da S. Santos, R. Pergl, and Y. Le Franc, 'Common Minimum Metadata for FAIR Semantic Artefacts', in *2nd Workshop on Ontologies for FAIR and FAIR Ontologies (Onto4FAIR 2023)*, Sherbrooke, Canada: C. Trojahn and L. O. Bonino da Silva Santos and G. Guizzardi and C. Jonquet, Jul. 2023. Accessed: Jan. 24, 2024. [Online]. Available: https://hal.science/hal-04106533

[11] L. Candela, L. Frosini, F. Mangiacrapa, O. Rouchon, Y. Le Franc, and B. Toulemonde, 'EOSC-Pillar D5.6 FAIR Research Data Management Tool Set Update', Sep. 2021, doi: 10.5281/ZENODO.5720411.

[12] D. Garijo and M. Poveda-Villalón, 'Best Practices for Implementing FAIR Vocabularies and Ontologies on the Web', in *Studies on the Semantic Web*, G. Cota, M. Daquino, and G. L. Pozzato, Eds., IOS Press, 2020. doi: 10.3233/SSW200034.

[13] M. Poveda-Villalón, P. Espinoza-Arias, D. Garijo, and O. Corcho, 'Coming to Terms with FAIR Ontologies', in *Knowledge Engineering and Knowledge Management*, vol. 12387, C. M. Keet and M. Dumontier, Eds., in Lecture Notes in Computer Science, vol. 12387. , Cham: Springer International Publishing, 2020, pp. 255–270. doi: 10.1007/978-3-030-61244-3_18.

[14] S. J. D. Cox, A. N. Gonzalez-Beltran, B. Magagna, and M.-C. Marinescu, 'Ten simple rules for making a vocabulary FAIR', *PLOS Comput. Biol.*, vol. 17, no. 6, p. e1009041, Jun. 2021, doi: 10.1371/journal.pcbi.1009041.

Funded by the European Union

[15] F. Xu, N. Juty, C. Goble, S. Jupp, H. Parkinson, and M. Courtot, 'Features of a FAIR vocabulary', *J. Biomed. Semant.*, vol. 14, no. 1, p. 6, Jun. 2023, doi: 10.1186/s13326-023-00286-8.

[16] E. Amdouni, S. Bouazzouni, and C. Jonquet, 'O'FAIRe makes you an offer: metadata-based automatic FAIRness assessment for ontologies and semantic resources', *Int. J. Metadata Semant. Ontol.*, vol. 16, no. 1, pp. 16–46, 2022, doi: 10.1504/IJMSO.2022.131133.

[17] D. Garijo, O. Corcho, and M. Poveda-Villal{\'o}n, 'FOOPS!: An Ontology Pitfall Scanner for the FAIR Principles', presented at the International Semantic Web Conference (ISWC) 2021: Posters, Demos, and Industry Tracks, in {CEUR} Workshop Proceedings, vol. 2980. CEUR-WS.org, 2021. [Online]. Available: http://ceur-ws.org/Vol-2980/paper321.pdf

[18] M. Poveda-Villalón, A. Fernández-Izquierdo, M. Fernández-López, and R. García-Castro, 'LOT: An industrial oriented ontology engineering framework', *Eng. Appl. Artif. Intell.*, vol. 111, p. 104755, May 2022, doi: 10.1016/j.engappai.2022.104755.

[19] B. Dutta, D. Nandini, and G. K. Shahi, 'MOD: Metadata for Ontology Description and Publication', *Int. Conf. Dublin Core Metadata Appl.*, pp. 1–9, Sep. 2015.

[20] C. Jonquet *et al.*, 'AgroPortal: A vocabulary and ontology repository for agronomy', *Comput. Electron. Agric.*, vol. 144, pp. 126–143, Jan. 2018, doi: 10.1016/j.compag.2017.10.012.

[21] C. Jonquet, A. Toulet, B. Dutta, and V. Emonet, 'Harnessing the Power of Unified Metadata in an Ontology Repository: The Case of AgroPortal', *J. Data Semant.*, vol. 7, no. 4, pp. 191–221, Dec. 2018, doi: 10.1007/s13740-018-0091-5.

[22] 'The Profiles Vocabulary'. [Online]. Available: https://www.w3.org/TR/dx-prof/

[23] D. Garijo, A. Beltran, and A. Wilson, 'Ontology metadata categorization'. Zenodo, Aug. 29, 2023. doi: 10.5281/ZENODO.8299390.

[24] N. Matentzoglu *et al.*, 'A Simple Standard for Sharing Ontological Mappings (SSSOM)', *Database*, vol. 2022, p. baac035, May 2022, doi: 10.1093/database/baac035.

Funded by the European Union