Joint Proceedings of the Onto4FAIR 2023 Workshops

Cassia Trojahn, Luiz Olavo Bonino da Silva Santos, Giancarlo Guizzardi, Clement Jonquet, Megan Katsumi, Emilio Sanfilippo, Jennifer D'souza, Anisa Rula

Joint Proceedings of the Onto4FAIR 2023 Workshops

Collocated with 13th International Conference on Formal Ontology in Information Systems (FOIS 2023) and 19th International Conference on Semantic Systems (SEMANTICS 2023)

Proceedings

Edited by
Cassia Trojahn, Luiz Olavo Bonino da Silva Santos, Giancarlo Guizzardi, Clement Jonquet, Megan Katsumi, Emilio Sanfilippo, Jennifer D'Souza, Anisa Rula

Preface

Making the resources produced by researchers fully reusable and understood requires specific efforts. The Findable, Accessible, Interoperable and Reusable (FAIR) principles were elaborated to address these issues, describing a set of requirements for resource reusability and interoperability. These principles have been gaining increasing attention in a range of different areas and applications. One the one hand, a key aspect is the ability of properly and semantically describing resources, in particular with the help of ontologies. On the other hand, ontologies themselves have to be compliant with the FAIR principles.

This series of workshops has the following goals: (a) to bring together leaders from academia, industry and user institutions to discuss the adoption of FAIR principles in real-world requirements. (b) to serve to inform industry and user representatives about existing research efforts that may meet their requirements. (c) to investigate how the FAIR principles are supported by the use of schemes, vocabulaires, and ontologies that ideally are themselves FAIR.

In 2023, we have a twin workshop at both FOIS and SEMANTICS conferences. The primary aim is to bring the gap between the scientific and the practitioner/industry sides, respectively, where we would take the greatest and latest advances in the state of the art to industry and bring back the practitioners' needs and challenges to the scientific community of figure out a solution.

This proceedings brings together works presented at both editions: 3rd edition at SEMANTICS and 2nd edition at FOIS. Both editions received 3 submissions where 2 of them have been accepted. For Onto4FAIR@FOIS, one of the papers has been accepted only for presentation following the reviewers recommendation and has not been included in the proceedings.

Each submission were reviewed by three collaborators from our program committee. Our sincere thanks to the members of the Program Committee and all the authors who submitted their work and participated in the event.

Both workshops also counted with an invited talk entitled A FAIR Catalog of Ontology-Driven Conceptual Models by Tiago Sales at FOIS and Claudenir M. Fonseca at SEMANTICS, both researchers at Semantics, Cybersecurity and Services (SCS), University of Twente, Enschede, The Netherlands, involved in the FAIR catalog projet. We warmly thank the speakers for their presentations.

Cassia Trojahn, Luiz Bonino, Giancarlo Guizzardi, Clement Jonquet.
Onto4FAIR workshop organisers

Organization

Workshop chairs

Megan Katsumi	University of Toronto, Canada
Emilio Sanfilippo	ISTC-CNR, Trento, Italy
Jennifer D'Souza	Technische Informationsbibliothek (TIB) Hannover, Germany
Anisa Rula	University of Brescia, Brescia, Italy

Workshop organisers

Cassia Trojahn	Institut de Recherche en Informatique de Toulouse, France
Luiz Olavo Bonino da Silva Santos	University of Twente, Leiden University Medical Centre, the Netherlands
Giancarlo Guizzardi	University of Twente, the Netherlands
Clement Jonquet	French National Research Institute for Agriculture, Food and Environment, Mathematics, Informatics and Statistics for Environment and Agronomy research unit, Montpellier, France

Program Committee

Alejandra Gonzalez-Beltran	University of Oxford, United kingdom
Emna Amdouni	Technological Research Institute SystemX
María Poveda-Villalón	Universidad Politécnica de Madrid, Spain
Tiago Prince Sales	University of Twente, The Netherlands
Nathalie Aussenac-Gilles	IRIT CNRS, France
Sophie Aubin	National Research Institute for Agriculture, Food and Environment (INRAE), France

Table of Contents

Papers Onto4FAIR@FOIS

Papers Onto4FAIR@SEMANTICS

Common Minimum Metadata for FAIR Semantic Artefacts

Clement JONQUET, [a,b,1] https://orcid.org/0000-0002-2404-1582
Biswanath DUTTA, [c] https://orcid.org/0000-0003-3059-8202
Luiz O. BONINO DA SILVA SANTOS, [d,e] https://orcid.org/0000-0002-1164-1351
Robert PERGL, [f,g] https://orcid.org/0000-0003-2980-4400
Yann LE FRANC [f,1] https://orcid.org/0000-0003-4631-418X
[a] *LIRMM, University of Montpellier & CNRS, France*
[b] *MISTEA, University of Montpellier, INRAE & Institut Agro, France*
[c] *DRTC, Indian Statistical Institute, Bangalore, India*
[d] *Semantics, Cybersecurity & Services, University of Twente, The Netherlands*
[e] *Biosemantics group, Leiden University Medical Center, The Netherlands*
[f] *e-Science Data Factory, France*
[g] *Faculty of Information Technology, CTU in Prague, Czech Republic*

Abstract.
Semantic interoperability is crucial for the FAIR Principles and strongly relies on Semantic Artefacts that also need to be FAIR. To achieve this, semantic artefacts require rich, structured, and interoperable metadata. The challenge lies in determining the threshold for "rich metadata" and agreeing on a common minimum set. The H2020 FAIRsFAIR project and the RDA *Vocabulary Semantic Services Interest Group* addressed this question by developing a "minimal metadata model" for semantic artefacts. In this paper, we present background information, methodology, discussions and workshops which contribute to the establishment of the *FAIRsFAIR minimum metadata profile for semantic artefacts.* We present an extension of the Metadata for Ontology Description and Publication Ontology (MOD2.0) incorporating this profile as well as its implementation (SemanticDCAT-AP) and its use to build FAIRcat, a prototype of a FAIR Data Point harvesting the content of multiple semantic artefact catalogues.

Keywords. Minimal metadata model, Semantic artefacts, Ontologies, Vocabularies, Metadata, FAIR, FAIRness assessment.

1. Introduction

Semantic interoperability is at the very core of the FAIR Data Principles [1] and as in any interoperability effort, it requires agreement on how the resources or artefacts supporting "semantics" are described. In all domains, many vocabularies, terminologies, ontologies or more largely semantic artefacts[2] are produced to represent and annotate data to make them more interoperable. Semantic artefacts have even become a master element to achieve FAIRness and have been discussed as digital objects that themselves need to be FAIR [2–4].

[1] Corresponding authors: C. Jonquet (jonquet@lirmm.fr), Y. Le Franc (ylefranc@esciencefactory.com).

[2] Semantic artefact is a broader term, originally proposed in [2], and more and more used to include ontologies, terminologies, taxonomies, thesauri, vocabularies, metadata schemas and standards.

However, in order to properly follow the FAIR principles, semantic artefacts need rich, structured and interoperable metadata, which is also a necessary condition for machine-actionability [1]. One of the main challenges for implementing the FAIR principles, whether for semantic artefacts or for any kind of data, is to determine the threshold for "rich metadata" mentioned in principle F2. It requires communities to agree on a common and minimal metadata schema that could be used as a threshold for FAIR. Reaching an agreement on such a common metadata schema and representation improve systems' interoperability by allowing the development of client applications that would need to read/parse only one representation format. On top of this, by agreeing on a minimum set of metadata for semantic artefacts, the same systems would guarantee every time an application encounters the metadata of a semantic artefact, it would know that minimally certain information would be available.

The task 2.2 of the H2020 FAIRsFAIR project was dedicated to establish prerequisites for better semantic interoperability by developing recommendations and facilitating uptake of good practices to make semantic artefacts compliant with the FAIR principles [2–4]. Among the recommendations produced by the project, *P-Rec 3* required the creation of "*A common minimum metadata schema* to *be used to describe semantic artefacts and their content."* Such a minimum set of metadata (also called a "minimal metadata model") for semantic artefact was missing. To fill in this gap, the task established, in collaboration with a wide range of communities, a first version of this model presented here. This collaboration was supported by the RDA Vocabulary Services Interest Group's (VSSIG) Ontology Metadata task group which was discussing an extension of the pre-proposed *Metadata for Ontology Description and Publication Ontology* (MOD) model [5]. This extension would both: (i) offer a review of all metadata properties available for semantic artefacts (similar to a "maximal metadata model") and (ii) revise the MOD model as an extension of DCAT2.

This joint effort had the overall goal to enable the implementation of FAIRness assessment methods that would be capable of establishing some kind of base threshold for a semantic artefact to be FAIR, as for example the grid proposed in [6]. In doing this, one challenge was to identify both: (i) generic metadata properties for digital objects that would apply to semantic artefacts (e.g., creator, identifier, or license); and (ii) metadata properties that would be specific to semantic artefacts (e.g., representation language). Furthermore, a shared understanding has to be reached on the value, necessity and feasibility of the key metadata properties. Indeed, MOD v1.4, released in 2018, contained 128 properties to describe semantic artefacts. Those were taken from 15 "crosswalked" metadata vocabularies[3] and would come as this, without prioritization, so that a developer would face a concrete problem of identifying which are the key/most important of these properties. Of course, nothing would restrict the usage of more metadata properties for richer description.

Therefore, the subject of minimal metadata model for semantic artefact was discussed openly and publicly in multiple workshops and meetings organized by the FAIRsFAIR task 2.2 and RDA VSSIG task groups. Finally, a workshop organized June 4th 2021 –in which more than 30 participants from around 20 different communities contributed (out of 76 attendees)– helped us to come to the *FAIRsFAIR minimum metadata profile for semantic artefact* presented in this article.

[3] By "crosswalked", we mean that we have identified 346 metadata properties in those metadata vocabularies that, once mapped (crosswalks identified), bring us to 128.

In the following, we present background information on the subject of FAIR semantic artefacts (Section 2), then we introduce our working methodology (Section 3). We briefly present the MOD2.0 proposed model for semantic artefact and their catalogues done by extending DCAT2 (Section 4).[4] Then, based on the two new main objects introduced by this model, `mod:SemanticArtefact` and `mod:SemanticArtefactDistribution`, we present the metadata properties that were gathered to describe them and eventually voted to decide the level of requirement: Mandatory or Recommended or Optional (Section 5). Then, we explain how we have integrated the requirements in MOD2 and also developed two machine-actionable representations of SemanticDCAT-AP, an experimental application profile used by FAIRcat, a tool to aggregate and align repository metadata content to DCAT for publication in a FAIR Data Point (Section 6). Finally, we conclude and present some perspectives (Section 7).

2. Background

Before the FAIR Principles, a recommendation for publishing RDF vocabularies was produced in 2008 by the W3C Semantic Web Deployment Working Group.[5] Then in 2014, the 5-stars LOD principles of Berners-Lee [7] were specialized for linked data vocabularies [8] as five rules to follow for creating and publishing "good" vocabularies. The degree to which the FAIR principles align and extend the 5-star open data principles was also later in studied [9, 10] and [6] presented after. In 2017, the *Minimum Information for Reporting an Ontology* initiative published the MIRO guidelines for ontology developers when reporting an ontology in scientific reports [11]. These guidelines refer to 34 information items (such as "ontology name," "ontology license," "ontology URL") and specify the level of importance (must, should, optional) for each individual information item. This work was significant but was never put in perspective with the FAIR principles. In MOD, where the authors reviewed which properties of MOD v1.4 could "help" addressing which MIRO guidelines (cf. example in section 6.1).

In 2017, Dutta et al. [5] reviewed and harmonized existing metadata vocabularies and proposed a unified *Metadata for Ontology Description and Publication Ontology* (MOD) model to facilitate manual and automatic ontology descriptions, identification, and selection. MOD is not another standard nor another metadata vocabulary, but more an aggregated set of identified properties one can use to describe a semantic resource.[6] MOD 1.4 was used in AgroPortal to implement a richer, unified metadata model [12].

Then, since 2020, we have seen four parallel initiatives that investigated the question of FAIR semantic artefacts:

- In March 2020, the FAIRsFAIR H2020 project delivered the first version of a list of 17 recommendations and 10 best practices recommendations for making semantic artefacts FAIR [2]. For each recommendation, the authors provided a detailed description associated with a list of related supporting technologies or technical solutions proposed by different communities.

[4] This model is still being consolidated (now in the context of the Horizon Europe FAIR-IMPACT project: https://github.com/FAIR-IMPACT/MOD), but the minimal model can be presented here independently.

[5] https://www.w3.org/TR/swbp-vocab-pub

[6] For instance, MOD does not require the use of a specific authorship property but rather encodes that `dc:creator`; `schema:author`, `foaf:maker`, or `pav:createdBy` can be used to say so.

- Later in 2020, Garijo et al. [13] produced "guidelines and best practices for creating accessible, understandable and reusable ontologies on the Web." In another position paper, Poveda et al. [14] completed their work with a qualitative analysis of how four ontology publication initiatives cover the foundational FAIR principles. They proposed some recommendations on making ontologies FAIR and listed some open issues that might be addressed by the semantic Web community in the future. In October 2021, Garijo et al. proposed FOOPS! a Web service for assessing an ontology regarding the FAIR principles [15].
- Late 2020, Cox et al. proposed guidelines ("10 simple rules") for making a vocabulary FAIR (https://fairvocabularies.github.io/makeVocabularyFAIR) and transform vocabularies that are not available following Web standards [16]. However, the authors do not explain how the proposed rules are aligned to each individual FAIR principle.
- A list of functional metrics and recommendations for *Linked Open Data Knowledge Organization Systems* (LOD KOS) was proposed in 2020 [17].
- In the end of 2020, DBPedia Archivo [18], an ontology archive, was released to help developers and consumers to implement FAIR ontologies. The prototype automatically discovers, downloads, archives, and rates new ontologies (https://archivo.dbpedia.org). Unfortunately, this work had not been inspired by existing research methodologies/tools.
- In 2021, Amdouni et al., introduced an "integrated quantitative FAIRness assessment grid for semantic resources [6]. This work was nourished and aligned with relevant state-of-the-art initiatives for FAIRness assessment: the RDA FAIR Data Maturity Model, the RDA Sharing Rewards and Credit evaluation table, the 5-stars for vocabulary as well as FAIRsFAIR and Poveda et al. recommendations cited above. The grid dispatches different credits to each FAIR principle, depending on its importance –according to pre-existing initiatives– when assessing FAIRness.
- Early 2022, the same authors proposed a metadata-based automatic FAIRness assessment methodology for ontologies and semantic resources called *Ontology FAIRness Evaluator* (O'FAIRe), based on the grid described previously [19]. The methodology projects the 15 foundational FAIR principles for ontologies, and proposes 61 questions, among which 80% are based on the resource metadata descriptions. The methodology has been (partially) implemented in AgroPortal [20] and is currently being transferred to other OntoPortal-based ontology repositories.

In conclusion, each of these initiatives reviewed somehow –more or less directly– some metadata properties associated to multiple criteria required to produce a so-called "FAIR semantic artefact"; however, none of these approaches explicitly list a minimal set of metadata properties to be considered FAIR and took the responsibility to qualify these properties as mandatory-recommended-optional. We believed: (i) a consensual approach, based on informal feedback and voting was actually a good way to converge and (ii) our current research projects and working groups were offering a relevant context for discussion such a consensus.

3. Methodology

In 2020-2021, in parallel with the RDA VSSIG Ontology Metadata Task Group which was working on defining MOD2 presented in Section 4, the H2020 FAIRsFAIR project organized three public workshops to eventually produce the minimal model presented in

Section 5. These workshops involved ontologists, knowledge engineers and semantic artefact catalogue providers:

- On April 29th, 2020 (~30 participants): the objective was to present and discuss the first set of 17 "general recommendations" and 10 "best practices recommendations" for FAIR semantic artefacts [2] and in particular about P-Rec3 on metadata for FAIR Semantic artefacts. The recommendations were also made publicly available for comments on GitHub.[7] The outcomes of this workshop as well as the discussion on GitHub and in subsequent RDA task groups meetings contributed to revise the recommendations and produce a second version [3]
- On October 15th, 2020 (~30 participants): the objective of this second workshop was to collect feedback on the first version of the recommendations and to establish the alignment of the recommendations with the RFC 2119 (MUST, SHOULD, SHALL). The outcomes of this workshop contributed to the second version of the recommendations [3].
- On June 4th, 2021 (~80 participants): the objective was then to determine a set of key metadata properties to build a minimum metadata profile for semantic artefacts, setting up a threshold on FAIRness. In this workshop, the participants voted to decide if each property should be optional, recommended or mandatory. During the votes the participants focused on the meaning of the properties i.e., the information they encode, but not necessarily on the metadata vocabularies providing a formal property to encode this information. At the beginning of the workshop, all attendees were made aware of the idea of data modeling and a good familiarity with DCAT was suggested as these were needed to actually contribute during the voting session. They were then presented with a simple use-case to support the voting: *what would be the necessary fields for retrieving semantic artefacts?* Both DCAT and the MOD2 proposition were thoroughly presented, then workshop participants were asked only to share responses for which they considered themselves to have sufficient expertise and awareness to make an informed contribution. They were guided through this process by the organizers (authors).

4. A proposed model for semantic artefact and their catalogues (MOD2)

MOD2.0 was proposed in 2020 as a new version of the *Metadata for Ontology Description and Publication Ontology,* structured as an extension of DCAT.[8] The Data Catalog Vocabulary (DCAT) is an RDF vocabulary designed to facilitate interoperability between data catalogues published on the Web. By using DCAT to describe datasets in catalogues, publishers increase discoverability and enable applications to consume metadata from multiple catalogues. The key idea in extending DCAT was to view semantic artefacts as "datasets of knowledge entities" that can be available in multiple "distributions" and can be "cataloged" in repositories such as BioPortal [21] or the Ontology Lookup Service [22].[9]

In designing MOD2, several design issues –not necessarily discussed in this paper– were raised including: (i) how (and is it necessary) to "extend" the notion of distribution?;

[7] https://github.com/FAIRsFAIR-Project/FAIRSemantics/issues

[8] MOD2 proposition was released in 2020 on GitHub and presented in multiple workshops and talks. However, the new model made of the 5 new classes presented here has never been published in a scientific communication yet.

[9] We now use the expression "Semantic Artefact Catalogue".

(ii) which classes inside DCAT and outside, the `mod:SemanticArtefact` class should explicitly extend or supersede?; (iii) which metadata properties from other vocabularies are available to describe semantic artefacts?; (iv) are there properties from outdated and not maintained metadata vocabulary that MOD could adopt?. The five key-classes from DCAT were finally specialized by creating new classes in the MOD namespace, as illustrated in Figure 2:

- `mod:SemanticArtefact`: A collection of knowledge entities (classes, properties, concepts, terms, mappings), produced and curated by a single or multiple agents, and available for access or download in one or more representations. This is typically the class of any knowledge organization systems or resources such as ontologies, vocabularies, concepts schemes, thesauri, terminologies, etc. For example, the AGROVOC thesaurus (http://aims.fao.org/aos/agrovoc) or the CODO ontology (https://w3id.org/codo).
- `mod:SemanticArtefactDistribution`: A specific representation of a semantic artefact. Typically, the class of any possible distributions or issuances of the semantic artefacts. It could be used to distinguish either multiple versions of a semantic artefact or different format/representation available. For example, "the version 1.3 of CODO in OWL"; or the "AGROVOC Core distribution in SKOS with TTL syntax".
- `mod:SemanticArtefactCatalog`: A curated collection of metadata about semantic artefacts. Typically, the class of repositories, libraries or services hosting and maybe also serving various semantic artefacts. For example, the NCBO BioPortal repository or the AgroPortal vocabulary and ontology repository or the NERC Vocabulary Server.
- `mod:SemanticArtefactCatalogRecord`: A record in a catalog, describing the registration of a single semantic artefact. Typically, the class of the entries for semantic artefacts inside catalogues i.e., when a catalogue hosts a semantic artefact, it is often concretely materialized by a record describing the artefact following the catalogue metadata model. For example, the record for CODO in BioPortal (https://bioportal.bioontology.org/ontologies/CODO) or the record for AGROVOC in AgroPortal (http://agroportal.lirmm.fr/ontologies/AGROVOC).
- `mod:SemanticArtefactService`: A collection of operations providing access to one or more semantic artefacts or SemanticArtefact-based processing functions/services. Typically, the class of the services offered for semantic artefacts. For example, the REST API of BioPortal, the SPARQL endpoint of AgroPortal, or the browsing user interface of a SKOSMOS based service, a FAIR data point.

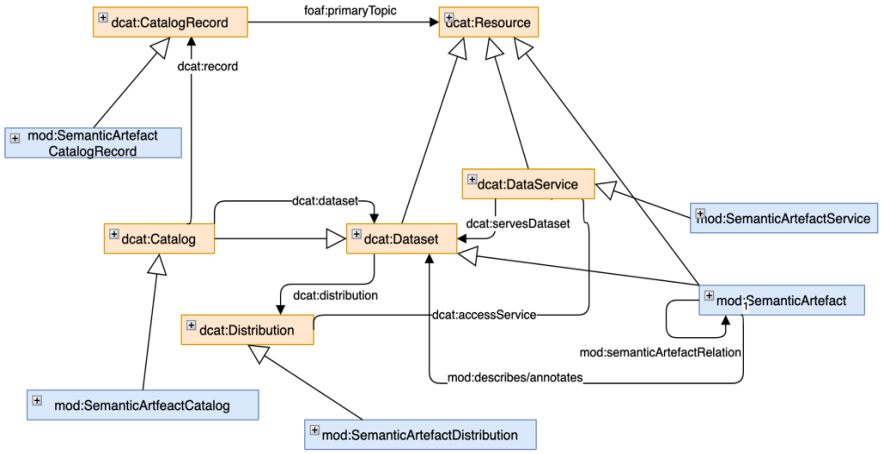


Figure 1. MOD2 proposed model for semantic artefact and their catalogues.

By inheriting from `dcat:Resource` (itself an `rdfs:Resource`) and `dcat:Dataset`, the `mod:SemanticArtefact` and the mod:`SemanticArtefactDistribution` classes could be described by the properties compatible with and suggested by DCAT. Plus, we decided to adopt and generalize in the MOD namespace the properties from OMV [23], DOOR [24] and VOAF (http://purl.org/vocommons/voaf) as those vocabularies were very specific to ontologies/vocabularies but are not maintained anymore.[10] Finally, we also had to re-incorporate the metadata properties from the previous versions of MOD (v1.4) to the relevant class (SemanticArtefact or Distribution), as MOD v1.4 did not distinguish the two. These three steps gave us a set of 92 properties (for `mod:SemanticArtefact`) and 46 properties (for `mod:SemanticArtefactDistribution`) from which we selected respectively a subset of 41 (Table 1) and 24 (Table 2) properties for the voting workshop. We have not yet status on the properties for the other new classes created (Service, Catalog, CatalogRecord) but we anticipate specialized properties justifying the creation of a new subclass in MOD2.

5. FAIRsFAIR Minimum metadata recommendations for semantic artefact

After the workshop, the results were collected into a spreadsheet for evaluation. The inputs were the number of votes for each endorsement level mandatory-recommended-optional. From them, we calculated the percentage of votes for each option. The option with the highest percentage was then selected. In cases where two options voted the same, the third was taken into account, e.g., 46.43% mandatory, 46.43% recommended, 7.14% optional would imply recommended, as 7.14% also voted for optional. As an auxiliary informative metric, we computed the consensus of the voting using the following formula:
consensus = (0.333 + *percentage of votes for the winning option* − *sum of percentages of the non-winning options*) / 1.333.

[10] Furthermore, we were motivated by unifying all the metadata properties *specific* to semantic artefacts.

Table 1 and Table 2 shows the result for all the properties voted.

Table 1. List of properties, and decision for the `mod:SemanticArtefact` class.

Property	# of votes	Consensus	Decision
dct:title	**23**	86.95%	mandatory
dct:license	**28**	78.57%	mandatory
dct:identifier	**28**	67.85%	mandatory
dct:accessRights	**31**	66.12%	mandatory
dct:creator	**31**	66.12%	mandatory
dct:created	**13**	53.83%	mandatory
dct:description	**28**	51.77%	mandatory
dcat:contactPoint	**31**	46.76%	mandatory
owl:versionIRI	**13**	42.29%	mandatory
dct:modified	**28**	35.70%	mandatory
dcat:keyword	**31**	32.24%	mandatory
mod:acronym	**13**	30.75%	mandatory
dcat:landingPage	**31**	22.56%	mandatory
dct:publisher	**28**	19.62%	recommended
dct:subject	**13**	19.21%	mandatory
dct:type	**22**	18.16%	mandatory
dct:issued	**27**	16.65%	mandatory
dcat:theme	**30**	9.98%	mandatory
dct:conformsTo	**30**	14.98%	recommended
dct:language	**28**	19.62%	recommended
mod:URI	**13**	19.21%	optional
dcat:distribution	**25**	15.98%	recommended
dct:contributor	**13**	30.75%	recommended
dct:rights	**23**	34.77%	recommended
dct:temporal	**24**	6.23%	recommended
dcat:qualifiedRelation	**29**	27.57%	optional
mod:status	**13**	42.29%	recommended
odrl:hasPolicy	**23**	21.72%	optional
prov:qualifiedAttribution	**23**	41.29%	optional
prov:wasGeneratedBy	**26**	13.44%	optional
dct:relation	**22**	18.16%	optional
dct:isReferencedBy	**28**	14.26%	optional
schema:includedInDataCatalog	**12**	37.48%	optional
mod:competencyQuestion	**13**	65.38%	optional
dct:accrualPeriodicity	**25**	21.98%	optional
dct:spatial	**24**	18.73%	optional
mod:usedEngineeringMethodology	**12**	24.98%	recommended
dcat:temporalResolution	**25**	39.98%	optional
mod:hasFormalityLevel	**13**	19.21%	recommended
dcat:spatialResolutionInMeters	**25**	57.99%	optional
dct:accrualMethod	**13**	42.29%	recommended

Table 2. List of properties, and decision for the `mod:SemanticArtefactDistribution` class.

Property	# of votes	Consensus	Decision
dcat:mediaType	**17**	47.05%	mandatory
dct:format	**17**	47.05%	mandatory
dct:title	**15**	39.98%	mandatory
dcat:accessURL	**17**	38.22%	mandatory
mod:hasRepresentationLanguage	**10**	24.98%	mandatory
mod:hasSyntax	**10**	24.98%	mandatory
dct:accessRights	**17**	20.57%	mandatory
dcat:downloadURL	**17**	20.57%	recommended

dct:rights	**17**	20.57%	recommended
dct:description	**17**	20.57%	recommended
dct:issued	**17**	2.92%	recommended
dct:modified	**16**	34.36%	recommended
mod:definitionProperty	**10**	39.98%	recommended
dcat:accessService	**16**	24.98%	recommended
dcat:packageFormat	**17**	20.57%	optional
dct:conformsTo	**17**	29.39%	recommended
mod:usedEngineeringTool	**10**	24.98%	optional
mod:prefLabelProperty	**10**	54.99%	recommended
mod:synonymProperty	**10**	24.98%	recommended
odrl:hasPolicy	**15**	39.98%	recommended
dcat:compressFormat	**17**	55.87%	optional
dcat:temporalResolution	**17**	64.70%	optional
dcat:byteSize	**17**	55.87%	optional
dcat:spatialResolutionInMeters	**17**	82.35%	optional

6. Results and applications

6.1. Inclusion of the metadata property requirements in MOD2

We included the requirements within MOD2 as additional information about a metadata property. With this, MOD encodes now three influential works motivating the presence of a property within the vocabulary: (i) the MIRO guidelines followed with using the property; (ii) the FAIR Principle addressed with using the property; and now (iii) the requirement in the FAIRsFAIR profile. For instance, in MOD2, the property `mod:acronym` is encoded as follow:[11]

```
### https://w3id.org/mod#acronym
mod:acronym
   rdf:type                owl:DatatypeProperty ;
   rdfs:subPropertyOf      rdfs:label ;
   rdfs:label              "acronym"@en ,
                           "acronyme"@fr ;
   rdfs:domain             mod:SemanticArtefact ;
   rdfs:range              xsd:string ;
   dcterms:description     "MOD: Short acronym label, often used as an
                           identifier within some ontology platforms such
                           as BioPortal or OBO Foundry. OMV: A short name
                           by which an ontology is formally known."@en ;
   rdfs:isDefinedBy        <http://omv.ontoware.org/2005/05/ontology> ;
   dcterms:issued          "2009-12-24"^^xsd:date ;
   dcterms:relation        <http://www.isibang.ac.in/ns/mod/1.0/acronym> ;
   pav:derivedFrom         <http://www.isibang.ac.in/ns/mod/1.0> ;
   pav:importedOn          "2015-08-05"^^xsd:dateTime ;
   skos:historyNote        "This property has been adopted from OMV
                           Ontology Metadata Vocabulary and redefined in
                           the MOD namespace."@en ;
   prov:wasInfluencedBy    "MIRO guidelines: A.1" ,
                           "FAIR principle: F2" ,
                           "FAIRsFAIR profile: MANDATORY" .
```

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

6.2. *Example of an ontology described with some mandatory metadata*

```
http://myontologyIRI.org
  rdf:type                owl:Ontology; mod:SemanticArtefact ;
  dcat:distribution       http://myontologyIRI.org/distribOWL ,
                          http://myontologyIRI.org/distribPDF ;
  mod:acronym             "MYON" ;
  dcterms:title           "My ontology" ;
  owl:versionIRI          <http://myontologyIRI.org/v1.0> ;
  dcterms:identifier      "myontologyDOI" ;
  dcterms:license         <https://creativecommons.org/licenses/by/3.0> ;
  dcat:landingPage        "myontologyWebPageURL" ;
  dcterms:creator         "http://orcid.org/0000-0002-2404-1582" ;
  dcterms:created         "2023-05-01"^^xsd:dateTime ;
  dcterms:modified        "2023-07-15"^^xsd:dateTime .


http://myontologyIRI.org/distribOWL
  mod:hasRepresentationLanguage  <http://www.w3.org/2002/07/owl>;
  mod:hasSyntax                  <http://www.w3.org/ns/formats/RDF_XML>;
  dcterms:description            "Distribution of My Ontology in OWL";
  dcat:accessURL                 "myontologycataloguerecordURL" .
```

6.3. *SemanticDCAT-AP and FAIRcat*

The minimum metadata profile described in Section 5 have been encoded into RDF/OWL (prefix `semdcat`) [4] (Figure 2). This enables retrieval via simple SPARQL queries and also adding meta-properties (annotations). These meta-properties are:

- `rdfs:definedBy` – the object is a predicate that can be used to retrieve definition of the property. This is required because individual vocabularies employ different approaches e.g., `skos:definition` or `rdfs:comment`.
- `semdcat:endorsement` – the endorsement level being `Mandatory`, `Recommended`, `Optional`. While it can be argued that "mandatory" can be alternatively expressed by OWL axioms, there is no way to express "recommended" without this extension.

At the same time, this approach leads to a more complicated and non-standard RDF/OWL representation, as Object Properties can map only OWL Classes. As such, every property needs an extra class wrapper to be defined.

We also explored an alternative way as SHACL[12] representation of the minimum metadata profile. SHACL shapes are an established way of specifying RDF graph constraints and are extensively used such as in the FAIR Data Point specification [23].[13] This approach allows elegant and straightforward specification of the set of properties, however there are two limitations:

1. Definitions cannot be linked to properties, they must be copied into `sh:description`.
2. Again, there is no way to express "mandatory". The approach taken was to extend the `sh:PropertyShape` with possibility to include `sh-e:endorsement` predicate. As SHACL is RDF, it is formally possible, however such a SHACL file will not pass the standard "SHACL of SHACL" validity, which may be a problem for some checking tools.

[12] https://www.w3.org/TR/shacl
[13] https://specs.fairdatapoint.org

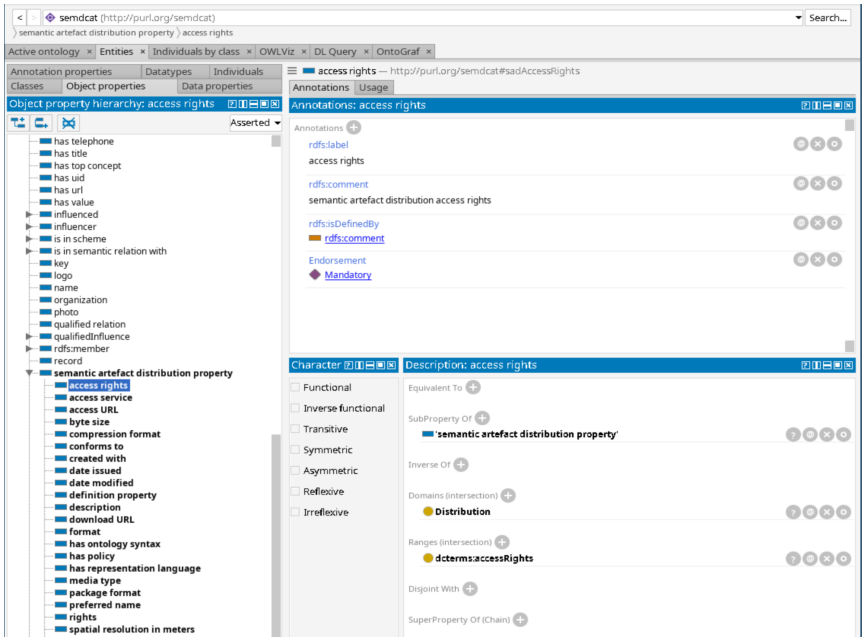


Figure 2. SemanticDCAT-AP's OWL representation in Protégé illustrating the annotations of properties.

An example of the SHACL specification is shown in Figure 3.

```
:SemanticArtefact a sh:NodeShape ;
  sh:name "Semantic Artefact" ;
  sh:description "Semantic Artefact minimum metadata" ;
  sh:targetClass mod:SemanticArtefact ;
  sh:property [
    sh:path dct:accessRights ;
    sh:class dct:RightsStatement ;
    sh:name "access rights" ;
    sh:description "Information about who access the semantic artefact or an indication of its security status." ;
    sh:minCount 1 ;
    sh:maxCount 1 ;
  ], [
    sh:path dct:accrualMethod ;
    sh:name "accrual method" ;
    sh:description "The method by which items are added to a collection." ;
  ], [
    sh:path dct:conformsTo ;
    sh:name "conforms to" ;
    sh:description "An established standard to which the semantic artefact conforms." ;
    sh:nodeKind sh:IRI ;
    sh:maxCount 1 ;
    sh-e:endorsement sh-e:Recommended ;
  ], [
```

Figure 3. SemanticDCAT-AP SHACL definition example. 3rd property illustrates the endorsement extension.

FAIRcat [4] is a proof-of-concept application, based on the federated FAIR Data Space,[14] that utilizes the described OWL-based machine-actionable SemanticDCAT-AP representation and demonstrates the potential of the common minimum metadata for FAIR semantic artefacts. At the same time, it represents a solution that can be used to increase FAIRness of semantic artefacts without any time and resource investments at

[14] https://www.eosc-pillar.eu/federated-fair-data-space-f2ds

the side of repository providers. The idea of FAIRcat is depicted in Figure 4. Semantic artefact catalogues were harvested for their items metadata. Using mappings, they are converted into the SemanticDCAT-AP and stored into a FAIR Data Point, the FAIRsFAIR Reference FAIR Data Point in our case.[15]

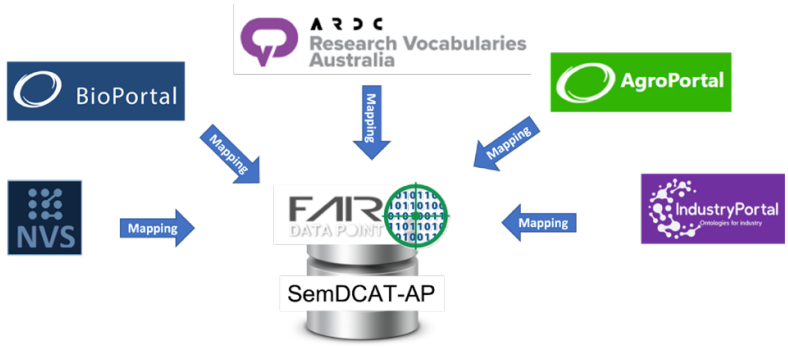


Figure 4. FAIRcat: harvest semantic artefact catalogues content and map them to a common profile.

Within FAIRcat, mappings link attributes present in the source catalogue to the equivalent ones in SemanticDCAT-AP. Those mappings are created in an editor depicted in Figure 5. The form is generated from the OWL specification, which allows populating all properties and rendering their name, definition, and level of endorsement (red "M" = mandatory). Because catalogues provide their metadata typically in JSON(-LD), the source metadata properties are identified using JSONPath.

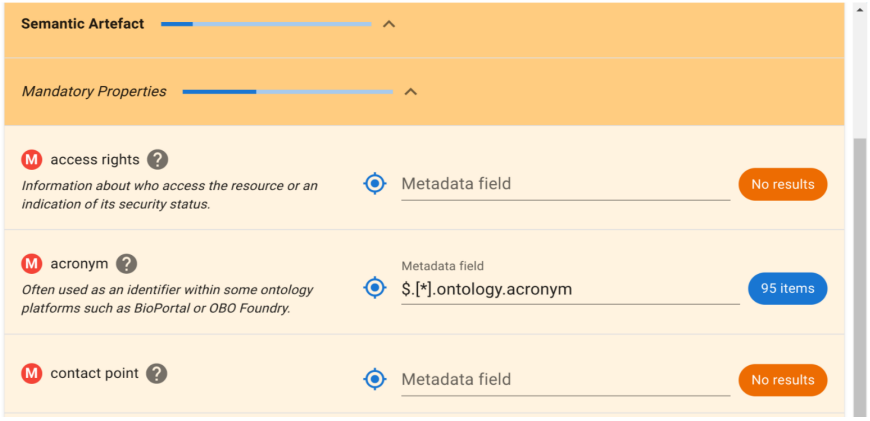


Figure 5. Illustration of mapping repository semantic artefacts to the SemanticDCAT-AP in FAIRcat.

Once the mapping step is finished, the harvested metadata can be converted into the SemanticDCAT-AP representation and stored in a FAIR Data Point [23] as illustrated in Figure 6. Apart from achieving SemanticDCAT-AP representation and FAIR-

[15] https://github.com/FAIRDataTeam/FAIRDataPoint

compliance, this approach enables harvesting multiple repositories into one FAIR Data Point that can be then searched; complex queries using SPARQL are also possible.

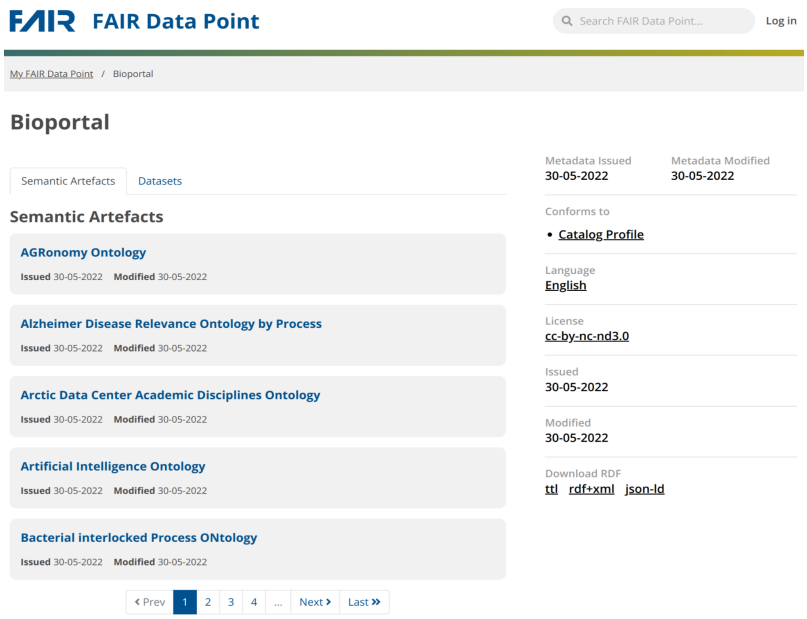


Figure 6. SemanticDCAT-AP metadata of the BioPortal semantic artefacts stored in a FAIR Data Point.

7. Conclusions and perspective

This paper presents the first attempt to define a common minimum metadata profile for FAIR semantic artefacts. This profile has been developed with the inputs from a large variety of communities. It aims to set a threshold, below which, an artefact can hardly be considered FAIR. Such a minimum metadata profile will be useful for FAIR assessment tools such as O'FAIRe and FOOPS! in their future evolutions.

With the FAIRCat prototype, three different semantic artefact catalogues have been harvested and mapped to this minimum metadata profile (via its implementation in SemanticDCAT-AP) in order to publish their content into a unique FAIR Data Point, allowing users to search across these three catalogues without copying the content or dealing with their specific APIs.

This work is now consolidated and refined in the context of FAIR-IMPACT i.e., the MOD2 proposition as well as the FAIRsFAIR profile (and its experimental implementation SemanticDCAT-AP). The aim is to reach a unified community-driven "standard to describe semantic artefacts. In the future, we also plan to investigate the W3C Profile Vocabulary (DX-PROF – https://www.w3.org/TR/dx-prof) to express the profile and eventually use it to provide a specification of a standard Application Programming Interface that semantic artefact catalogues could implement.

Finally, the SemanticDCAT-AP machine-actionable representations are an essential piece in implementing the FAIR principles as it can be used by the community to develop software tools, such as is the example of the FAIRcat prototype but also in FAIRness

assessment tools and any other relevant tools for semantic artefacts. Currently, FAIRcat is limited in its possibilities, for example it cannot harvest catalogues with complex APIs (or just in a limited way), but it demonstrates the idea. Currently, the mappings are stored into an in-house data model. As future work, these mappings will themselves be FAIRified, i.e., represented in a semantic way and stored with a persistent identifier.

Acknowledgments

We thank all the participants of the workshops and meetings organized by both the RDA VISSG and FAIRsFAIR project for time and fruitful discussions and feedback. This work has been originally supported in part by the H2020 FAIRsFAIR project (www.fairsfair.eu – grant #831558) as well as the *Data to Knowledge in Agronomy and Biodiversity* project (D2KAB – www.d2kab.org – ANR-18-CE23-0017). This work is now published –and continued– with support from the Horizon Europe FAIR-IMPACT project (https://fair-impact.eu – grant #101057344).

References

1. Batista, D., Gonzalez-Beltran, A., Sansone, S.-A., Rocca-Serra, P.: Machine actionable metadata models. Sci Data. 9, 592 (2022). https://doi.org/10.1038/s41597-022-01707-6.
2. Le Franc, Y., Coen, G., Essen, J.P., Bonino, L., Lehväslaiho, H., Staiger, C.: D2.2 FAIR Semantics: First recommendations. (2020). https://doi.org/10.5281/zenodo.3707985.
3. Hugo, W., Le Franc, Y., Coen, G., Parland-von Essen, J., Bonino, L.: D2.5 FAIR Semantics Recommendations Second Iteration. (2020). https://doi.org/10.5281/ZENODO.5362010.
4. Franc, Y. Le, Bonino, L., Koivula, H., Essen, J.P., Pergl, R.: D2.8 FAIR Semantics Recommendations Third Iteration. (2022). https://doi.org/10.5281/ZENODO.6675295.
5. Dutta, B., Toulet, A., Emonet, V., Jonquet, C.: New Generation Metadata vocabulary for Ontology Description and Publication. In: Garoufallou, E., Virkus, S., and Alemu, G. (eds.) 11th Metadata and Semantics Research Conference, MTSR'17. , Tallinn, Estonia (2017). https://doi.org/10.1007/978-3-319-70863-8_17.
6. Amdouni, E., Jonquet, C.: FAIR or FAIRer? An integrated quantitative FAIRness assessment grid for semantic resources and ontologies. In: Emmanouel Garoufallou and María-Antonia Ovalle-PerandonesAndreas Vlachidis (eds.) 15th International Conference on Metadata and Semantics Research, MTSR'21. pp. 67–80. Springer, Madrid, Spain (2021). https://doi.org/10.1007/978-3-030-98876-0_6.
7. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. Semantic Web and Information Systems. 5, 1–22 (2009). https://doi.org/10.4018/jswis.2009081901.
8. Janowicz, K., Hitzler, P., Adams, B., Kolas, D., Vardeman, C.: Five stars of Linked Data vocabulary use. Semant Web. 5, 173–176 (2014). https://doi.org/10.3233/SW-140135.
9. Hasnain, A., Rebholz-Schuhmann, D.: Assessing FAIR data principles against the 5-star open data principles. In: The Semantic Web: ESWC 2018 Satellite

Events. pp. 469–477. Springer, Heraklion, Greece (2018). https://doi.org/10.1007/978-3-319-98192-5_60/TABLES/1.
10. Garijo, D., Poveda-Villalón, M.: Best Practices for Implementing FAIR Vocabularies and Ontologies on the Web. In: Cota, G., Daquino, M., and Pozzato, G.L. (eds.) Applications and Practices in Ontology Design, Extraction, and Reasoning. IOS Press (2020). https://doi.org/10.3233/SSW200034.
11. Matentzoglu, N., Malone, J., Mungall, C., Stevens, R.: MIRO: guidelines for minimum information for the reporting of an ontology. J Biomed Semantics. 9, 6 (2018). https://doi.org/10.1186/s13326-017-0172-7.
12. Jonquet, C., Toulet, A., Dutta, B., Emonet, V.: Harnessing the power of unified metadata in an ontology repository: the case of AgroPortal. Data Semantics. 7, 191–221 (2018). https://doi.org/10.1007/s13740-018-0091-5.
13. Garijo, D., Poveda-Villalón, M.: Best Practices for Implementing FAIR Vocabularies and Ontologies on the Web. In: Cota, G., Daquino, M., and Pozzato, G.L. (eds.) Applications and Practices in Ontology Design, Extraction, and Reasoning. IOS Press (2020). https://doi.org/10.3233/SSW200034.
14. Poveda-Villalón, M., Espinoza-Arias, P., Garijo, D., Corcho, O.: Coming to Terms with FAIR Ontologies. In: Keet, M.C. and Dumontier, M. (eds.) 22nd International Conference on Knowledge Engineering and Knowledge Management, EKAW'20. pp. 255–270. Springer, Bolzano, Italy (2020). https://doi.org/10.1007/978-3-030-61244-3_18.
15. Garijo, D., Corcho, O., María Poveda-Villalón: FOOPS!: An Ontology Pitfall Scanner for the FAIR principles. In: 20th International Semantic Web Conference, ISWC'21: Posters, Demos, and Industry Tracks. CEUR Workshop Proceedings, 2980 (2021).
16. Coxid, S.J.D., Gonzalez-Beltranid, A.N., Magagna, B., Marinescu, M.-C.: Ten simple rules for making a vocabulary FAIR. (2021). https://doi.org/10.1371/journal.pcbi.1009041.
17. Zeng, M.L., Clunis, J.: FAIR + FIT: Guiding Principles and Functional Metrics for Linked Open Data (LOD) KOS Products. Journal of Data and Information Science. 5, 93–118 (2020). https://doi.org/10.2478/JDIS-2020-0008.
18. Frey, J., Streitmatter, D., Götz, F., Hellmann, S., Arndt, N.: DBpedia Archivo: A Web-Scale Interface for Ontology Archiving Under Consumer-Oriented Aspects. In: International Conference on Semantic Systems, SEMANTICS'20. pp. 19–35. Springer (2020). https://doi.org/10.1007/978-3-030-59833-4_2/TABLES/2.
19. Amdouni, E., Bouazzouni, S., Jonquet, C., O'faire, C.J.: O'FAIRe makes you an offer: Metadata-based Automatic FAIRness Assessment for Ontologies and Semantic Resources. Int J Metadata Semant Ontol. 16, 16–46 (2022). https://doi.org/10.13039/501100001665.
20. Amdouni, E., Bouazzouni, S., Jonquet, C.: O'FAIRe: Ontology FAIRness Evaluator in the AgroPortal Semantic Resource Repository. In: 19th Extended Semantic Web Conference, ESWC'22, Demo and Poster session. pp. 89–94. LNCS, Springer, Hersonissos, Crete, Greece (2022). https://doi.org/10.1007/978-3-031-11609-4_17.
21. Noy, N.F., Shah, N.H., Whetzel, P.L., Dai, B., Dorf, M., Griffith, N.B., Jonquet, C., Rubin, D.L., Storey, M.-A., Chute, C.G., Musen, M.A.: BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res. 37, 170–173 (2009). https://doi.org/10.1093/nar/gkp440.

22. Côté, R.G., Jones, P., Apweiler, R., Hermjakob, H.: The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. BMC Bioinformatics. 7, 97 (2006). https://doi.org/10.1186/1471-2105-7-97.
23. Suarez-Figueroa, Hartmann, J., Sure, Y., Haase, P., Suarez-Figueroa, M.: OMV–ontology metadata vocabulary. In: Welty, C. (ed.) Workshop on Ontology Patterns for the Semantic Web, WOP'05. p. 9. Springer, Galway, Irland (2005).
24. Allocca, Carlo and d'Aquin, Mathieu and Motta, E.: DOOR - Towards a Formalization of Ontology Relations. In: International Conference on Knowledge Engineering and Ontology Development, KEOD'09. pp. 13–20. , Madera, Portugal (2009).

September 2023

The EarthPortal, towards an ontology repository for the Earth System semantic artefacts

Guillaume ALVISET [a], Christelle PIERKOT [a] and Marine VERNET [a]
[a] *IR Data Terra-CNRS*
ORCiD ID: Guillaume Alviset https://orcid.org/0009-0004-4295-6593, Christelle Pierkot https://orcid.org/0000-0002-2591-3311, Marine Vernet https://orcid.org/0000-0002-39906-5579

Abstract. In order to describe research data in a standard and FAIR way, Earth and Environmental research communities have developed and manage heterogeneously separate vocabularies, thesauri and ontologies which, when conducting interdisciplinary studies, lead to discovery, interoperability and semantic search challenges. To address this, this article proposes the use of an ontology repository to store, expose, catalog and map semantic artefacts. The chosen approach is based on the mature and flexible OntoPortal technology to create the EarthPortal, a dedicated prototype for Earth and environmental sciences within the framework of the FAIR-IMPACT project and in strong collaboration with user-communities. It will therefore make sure to include all functionalities required to address issues mentioned above to ensure semantic artefacts quality.

Keywords. Ontology repository, Earth sciences, Ontoportal, FAIR, semantic artefacts catalogue

1. Introduction

As in other scientific fields, the communities involved in the Earth system need access to semantic artefacts [1] for different purposes (data integration, metadata management, semantic search, etc.). However, vocabularies have been developed independently for each of the different compartments of the Earth System (atmosphere, land surfaces, solid earth and sea), and have not been defined in a transdisciplinary way. Today, these domains have to work together to conduct their scientific research, and they need to evolve their practices and find better ways of combining and matching the different semantic artefacts in an effective manner and in a FAIR way[2]. To achieve this, one approach is to use an ontology repository to store and catalogue the various semantic artefacts (from vocabularies to ontologies) for Earth systems. It provides a centralized catalogue which allows to create and store mappings between concepts from different domains. Assessing the FAIRness of the stored semantic artefacts also becomes easier. The FAIR-IMPACT project supports the implementation of FAIR-enabling practices, tools and services across scientific communities at a a European, national and institutional level. The work-package 4 “Metadata and Ontologies” deals with greater

and more harmonised use of semantic artefacts, leading to semantic interoperability between disciplines. In this perspective, implementation of semantic artefacts catalogues by new communities is supported. The approach described in this paper fits into this context.

After reviewing the different practices and tools [3] used by our communities to share and expose their semantic artefacts, we will focus on the solution chosen, the reasons for choosing this solution and the improvements planned to ensure that this ontology repository meets the requirements of the Earth System and Environmental Communities.

2. Semantic Artefact Management in Earth and Environmental Sciences

The two semantic artefacts commonly used in the Earth and Environmental Sciences are the SWEET ontology[1] (Semantic Web for Earth and Environment Technology Ontology) defined by the Semantic Technologies Group of ESIP[2], and the GCMD Thesaurus (Global Change Master Directory) defined by the NASA Earth Science Data Systems[3]. However, since these semantic artefacts are high-level and do not cover all the features of the Earth System, some initiatives have been done by the communities to define dedicated semantic artefacts. In France, the Data Terra Research Infrastructure which covers the four Earth System compartments, has specified the vocabularies for each of them (Theia/Ozcar[4], Aeris[5], Odatis[6] and Formater[7]), and is in the process of developing an ontology, based on the SOSA ontology [4], including different type of concepts (variables, platforms, sensors, feature types).

The climate community has also defined the Climate Analysis (CA) ontology [5] based on the SOSA ontology with the addition of meteorological and geographic elements associated terms. At the European level, Actris[8], the research infrastructure supporting research into climate and air quality has built a controlled vocabulary[9]. EPOS, the infrastructure for the solid earth compartment is in the process of specifying its own[10]. The NERC Vocabulary server gives access to standardised and hierarchically organized vocabularies, for the oceanographic and associated domains [6]. For the environmental domain, the ENVO[11] ontology which represents knowledge about environments, environmental processes, ecosystems, habitats, and related entities has been set-up.

All these semantic artefacts, from vocabularies and thesauri to ontologies, need to be managed into specialized repositories in order to be used and shared by communities:

[1] https://github.com/ESIPFed/sweet
[2] https://wiki.esipfed.org/Semantic_Technologies
[3] https://www.earthdata.nasa.gov/esds
[4] https://in-situ.theia-land.fr\/skosmos\/theia_ozcar_thesaurus/en/
[5] https://skosmos.aeris-data.fr/
[6] https://vocab.ifremer.fr/
[7] https://catalogue-terresolide.ipgp.fr/voc/discipline/en/
[8] https://www.actris.eu/
[9] https://vocabulary.actris.nilu.no/skosmos/actris_vocab/en/index
[10] https://registry.epos-eu.org/ncl
[11] https://www.ebi.ac.uk/ols/ontologies/envo

- OLS [7] provided by the EliXir community is a repository for biomedical ontologies that aims to provide a single point of access to the latest ontology versions. OLS has been chosen by the ENVO ontology to share the concept for the Environmental domain. However, the current version of OLS (OLS3) is no longer updated but it will be replaced by a newer on the end of October 2023. Unfortunately, it is not capable of hosting something else than ontologies in OBO and OWL formats.
- Skosmos [8] is an open source web-based SKOS browser and publishing tool, mainly used to access controlled vocabularies for indexing, information retrieval and vocabulary development. Skosmos has been adopted by some communities to specify their vocabularies, such as the atmospheric one (e.g. AERIS and ACTRIS research infrastructures have both used Skosmos). Nevertheless, it is strictly restricted to the usage of SKOS and do not support ontologies.
- VocPrez is a tool (API and web front-end) for the read-only delivery system of SKOS vocabularies, compliant with Content Negotiation by Profile. It has been adopted by the NERC Vocabulary Server from the BODC [9], the OGC Definition server or the British Geological Survey. However, VocPrez is restricted to a custom implementation of SKOS which can require adjustments if used with other tools. Its read-only structure also not suitable for users to be able to submit their own semantic artefacts.
- UKGovLD is a Linked Data Registry developed by the UK Government, to manage code lists or controlled vocabularies, which support content negotiation. It provides a service to create and manage controlled, authoritative lists of identifiers as URIs. It is used by the French geological survey (BRGM) to define and expose their vocabularies and by the French Data Terra RI, as a FAIR incubator for the concepts used by its data clusters. Similarly to other vocabulary browsers, it does not support ontologies.
- LOV (Linked Open Vocabularies) is a catalogue of reusable vocabularies for the description of data on the Web. LOV supports Vocabulary Search, Vocabulary Assessment, Vocabulary Mapping [10]. Anyway, it does not provide any service beyond cataloguing vocabularies.
- Ontoportal is an open source and generic technology to build ontology repositories or semantic artefact catalogues which has been set-up by Stanford University in 2005, for the biomedical ontologies with the BioPortal appliance [11]. The technology was first reused by the Agrifood communities with the release of AgroPortal [12]. More communities adopted then this technology, with the addition of EcoPortal for the ecology part [13], MatPortal for materials science, IndustryPortal for industry and related domains.

3. Ontoportal and Ontoportal Alliance

The overview of the panel of technologies that we have seen previously makes one of the solutions stand out of the others: Ontoportal. This open source ontology repository and semantic artefact catalogue has several advantages over the others[3]:

- It regroups the features of some other technologies and combine them with new ones, such as mappings creations and storage, FAIRness assessment, text

annotation, version management, a REST API to read and write content. It has the advantage of being domain agnostic and support a wide range of semantic artefacts from SKOS thesauri to complex ontologies in different formats.
- It is actively maintained, the Ontoportal alliance is regularly pushing fixes and new features through the centralization of the code repository
- The technology is well proven today. OntoPortal comes from the NCBO BioPortal that is available since 2005 and is widely used across biomedical communities but also with AgroPortal and the agrifood community.
- It can be easily deployed in a pre-configured Open Virtual Machine Format (OVF) appliance[12].

The OntoPortal initiative comes back from 2005, where the BioPortal code is first re-used for agrifood communities leading to the release of AgroPortal. More communities adopted then this technology, with the addition of EcoPortal, MedPortal, SIFR Bioportal, IndustryPortal and MatPortal, publicly available. All of these are part of the consortium known as the OntoPortal Alliance. The source code is located in a common repository which is then forked by each alliance member, making all public OntoPortal instances connected to it. This makes the feature development and integration easier: a community can merge code back to the shared repository, which can be then pulled to other communities in a streamlined workflow.

The choice of OntoPortal answers most of the issues currently encountered in the Earth systems community concerning semantic artefacts. There is a strong need of unifying multidisciplinary workflows to ensure semantic artefacts are up to FAIR standards, which can be answered with the tools available in OntoPortal. Moreover, the deployment of additional instances of OntoPortal for different thematics opens some opportunities, such as an inter-portal federation. While each domain has their own set of resources, there are still some of them that can be found overlapping with others (as in Earth sciences can overlap with Ecology, so EarthPortal with EcoPortal). While this is still at the stage of an idea, portals federation could enrich existing resources with even more links to related material to push multidisciplinary workflows and avoid duplication.

4. The EarthPortal

Earth sciences being multidisciplinary by nature, each sub-domain follows its own conventions in the creation and usage of semantic artefacts. However, when it comes to working together, it is common to witness similarities between standards, which then require to be manually curated to provide the necessary data and metadata crosswalks for joint efforts. For example, volcanologists and atmospheric specialists have to rely on space-based observations to track the evolution of eruptive activity in near real-time. To achieve this, scientists need to find and aggregate relevant datasets from Solid Earth and Atmospheric communities. This can be done by using an Interdisciplinary Discovery and Access Service [14] which can provide users with an easy and FAIR service for discovery and access to multidisciplinary and aggregated data sets. This service has to rely on a semantic component to ensure that the discovery process is as efficient as possible. The EarthPortal, the Ontoportal instance that will be dedicated to Earth systems, will

[12] `https://ontoportal.github.io/documentation/administration/steps/getting_started`

centralize and catalog semantic artefacts from all sub-domains, to meet the following needs:

- Better discovery of the semantic artefacts used in Earth sciences. They are currently scattered across a significant amount of repositories even at sub-domain level, if available on the web at all.
- Providing the necessary tools for interdisciplinary work. This will be mainly done by storing and creating mappings between vocabularies and ontologies from different sub-domains.
- Assess the FAIRness of semantic artefacts. While their usage is strongly recomm--ended for data and metadata to be FAIR, vocabularies and ontologies also need their own metadata to reach FAIR standards. Although depositing artefacts in the repository already gives a certain level of FAIRness, the user must provide additional informations about their submissions. The O'FAIRE [15] tool initially developed by the AgroPortal team will be used in that regard.

The EarthPortal will allow users to submit their own semantic artefacts, but will also contain links to existing vocabularies and ontologies. Since it is not possible to ask all actors from Earth science communities (or even broader communities in the case of multidisciplinary standards) to store all their resources into the same repository, the OntoPortal technology also allows external resources to be referenced and work with as if they were included. This initiative being driven at a general level, the usage of EarthPortal can see the emergence of requests from the different Earth systems communities, leading to the development of new features. One of the leads towards the evolution of this technology concerns semantic artefacts mapping. Similarly to data and semantic artefacts, mappings can also embed metadata. While most of the information is similar, mappings are also subject to automation (i.e. lexical matching) which then deliver additional metadata about the results of this automation process. SSSOM [16] might be a good candidate to issue standards concerning mapping sharing. Since it is one of the main features of the EarthPortal, integration of SSSOM is being considered.

5. Conclusions

The usage of the OntoPortal technology to manage Earth system semantic artefacts is motivated by the panel of tools and flexibility it provides. Its dedicated instance for Earth science called EarthPortal will be part of the OntoPortal alliance, which brings potential for collaborations with other members. Not only it offers additional technical insight but also a way to support the EarthPortal in its early stages, to manage user feedback and develop new features. The first step will be to collect and reference the semantic artefacts used in each sub-domain of Earth system, then enrich them with mappings and metadata to ensure compliance with the FAIR principles. Of course, this is not without a cost. Users might not understand the entire scope of the thesauri, vocabularies and ontologies they use nor do they have the knowledge to do so. The EarthPortal will bring all the necessary features to lower the entry bar for the creation and usage of semantic artefacts by listening to feedback from each community and developing the corresponding features. Furthermore, in the context of a collaboration between FAIR IMPACT, in charge of the EarthPortal development, and FAIR-EASE, another European

project, the content of the EarthPortal repository could be used to enhance the semantic analyser planned by the FAIR-EASE project.

References

[1] Le Franc Y, Parland-von Essen J, Bonino L, Lehväslaiho H, Coen G, Staiger C. D2.2 FAIR Semantics: First recommendations. FAIRsFAIR; 2020 Mar. doi: https://doi.org/10.5281/zenodo.3707985

[2] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PA, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3:160018. doi: 10.1038/sdata.2016.18. Erratum in: Sci Data. 2019 Mar 19;6(1):6. PMID: 26978244; PMCID: PMC4792175.

[3] Jonquet C, Graybeal J, Bouazzouni S, Dorf M, Fiore N, ... Ontology Repositories and Semantic Artefact Catalogues with the OntoPortal Technology. 2023. ⟨hal-04088537⟩

[4] Janowicz K, Haller A, J.D. Cox S, Le Phuoc D, Lefrançois M, SOSA: A lightweight ontology for sensors, observations, samples, and actuators, Journal of Web Semantics, Volume 56, 2019, Pages 1-10, ISSN 1570-8268, doi: https://doi.org/10.1016/j.websem.2018.06.003.

[5] Wu J, Orlandi F, O'Sullivan D, Dev S, LinkClimate: An interoperable knowledge graph platform for climate data, Computers & Geosciences, Volume 169, 2022, 105215, ISSN 0098-3004, doi: https://doi.org/10.1016/j.cageo.2022.105215.

[6] Intergovernmental Oceanographic Commission of UNESCO. 2019. Ocean Data Standards, Vol.4: Technology for SeaDataNet Controlled Vocabularies for describing Marine and Oceanographic Datasets - A joint Proposal by SeaDataNet and ODIP projects. Ostend, IODE/UNESCO. (IOC Manuals and Guides, 54, Vol. 4.) 31 pp. (IOC/2019/MG/54 Vol.4).

[7] Jupp S. ... (2015) A new Ontology Lookup Service at EMBL-EBI. In: Malone, J. et al. (eds.) Proceedings of SWAT4LS International Conference 2015

[8] Suominen O, Ylikotila H, Pessala S, Lappalainen M, Frosterus M, Tuominen J, Baker T, Caracciolo C, Retterath A. Publishing SKOS vocabularies with Skosmos. Manuscript submitted for review, June 2015.

[9] British Oceanographic Data Centre (2023). The NERC Vocabulary Server, Natural Environment Research Council, https://vocab.nerc.ac.uk.

[10] Vandenbussche PY, Atemezing GA, Poveda-Villalón M, & Vatant B. (2017) Linked Open Vocabularies (LOV): a gateway to reusable semantic vocabularies on the Web. Semantic Web, 8(3), 437-452.

[11] Natalya FN, ... BioPortal: ontologies and integrated data resources at the click of a mouse, Nucleic Acids Research, Volume 37, Issue suppl_2, 1 July 2009, Pages W170–W173, doi: https://doi.org/10.1093/nar/gkp440

[12] Jonquet C, Toulet A, Arnaud Z, Aubin S, Dzalé Yeumo E, ... AgroPortal: A vocabulary and ontology repository for agronomy, doi: https://doi.org/10.1016/j.compag.2017.10.012.

[13] Kechagioglou X, ... EcoPortal: An Environment for FAIR Semantic Resources in the Ecological Domain. Proceedings. Vol. 1613. 2021.

[14] Krijger T, Boldrini E, Roncella R, Papeschi F, Kokkinaki A, Moncoiffe G, Chaffard V, Thijsse P. (2023). FAIR-EASE_D2.1_Environmental Data Infrastructures_Services Description Report (1.0). Zenodo, doi: https://doi.org/10.5281/zenodo.7920551

[15] Amdouni E, Bouazzouni S, Jonquet C. O'FAIRe: Ontology FAIRness Evaluator in the AgroPortal semantic resource repository. ESWC 2022 - 19th Extended Semantic Web Conference, Poster and demonstration, May 2022, Hersonissos, Greece. pp.89-94, ⟨10.1007/978-3-031-11609-4_17⟩. ⟨lirmm-03630543v3⟩

[16] Matentzoglu N, ... A Simple Standard for Sharing Ontological Mappings (SSSOM), Database, Volume 2022, 2022, baac035, doi: https://doi.org/10.1093/database/baac035

Toward FAIR Semantic Publishing of Research Dataset Metadata in the Open Research Knowledge Graph

Raia ABU AHMAD [a,1], Jennifer D'SOUZA [b], Matthäus ZLOCH [c], Wolfgang OTTO [c], Georg REHM [a], Allard OELEN [b], Stefan DIETZE [c] Sören AUER [b]

[a] *DFKI GmbH – Deutsches Fourschungszentrum für Künstliche Intelligenz, Berlin, Germany*

[b] *TIB – Leibniz Information Centre for Science and Technology, Hannover, Germany*

[c] *GESIS – Leibniz Institute for the Social Sciences, Köln, Germany*

ORCiD ID: Raia Abu Ahmad https://orcid.org/0009-0004-8720-0116, Jennifer D'Souza https://orcid.org/0000-0002-6616-9509, Matthäus Zloch https://orcid.org/0009-0007-1692-7053, Wolfgang Otto https://orcid.org/0000-0002-9530-3631, Georg Rehm https://orcid.org/0000-0002-7800-1893, Allard Oelen https://orcid.org/0000-0001-9924-9153, Stefan Dietze https://orcid.org/0009-0001-4364-9243, Sören Auer https://orcid.org/0000-0002-0698-2864

Abstract. Search engines these days can serve datasets as search results. Datasets get picked up by search technologies based on structured descriptions on their official web pages, informed by metadata ontologies such as the *Dataset* content type of *schema.org*. Despite this promotion of the content type *dataset* as a first-class citizen of search results, a vast proportion of datasets, particularly research datasets, still need to be made discoverable and, therefore, largely remain unused. This is due to the sheer volume of datasets released every day and the inability of metadata to reflect a dataset's content and context accurately. This work seeks to improve this situation for a specific class of datasets, namely research datasets, which are the result of research endeavors and are accompanied by a scholarly publication. We propose the *ORKG-Dataset* content type, a specialized branch of the Open Research Knowledge Graoh (ORKG) platform, which provides descriptive information and a semantic model for research datasets, integrating them with their accompanying scholarly publications. This work aims to establish a standardized framework for recording and reporting research datasets within the *ORKG-Dataset* content type. This, in turn, increases research dataset transparency on the web for their improved discoverability and applied use. In this paper, we present a proposal – the minimum FAIR, comparable, semantic description of research datasets in terms of salient properties of their supporting publication. We design a specific application of the *ORKG-Dataset* semantic model based on 40 diverse research datasets on scientific information extraction.

Keywords. semantic publishing, digital libraries, scholarly infrastructure, FAIR data principles, open science

[1]Corresponding Author: Raia Abu Ahmad, raia.abu_ahmad@dfki.de.

1. Introduction

Scientific research has long been conducted on datasets that measure and model aspects of the world. This phenomenon is becoming more acute given the vast amounts of datasets [1,2] released in the present age of data science [3]. Datasets are as diverse as science is [4,5]. For instance, datasets in medicine can reflect patient disease histories, datasets in earth science can reflect geological or climatic features of the world, while datasets in artificial intelligence model phenomena as machine learning objectives for a computer. The large volume of datasets released on the Web opens up new avenues for their search and discovery. Search engines now offer dedicated search platforms such as Google Dataset Search [6] for discovering datasets from various public repositories such as OpenAIRE [7] and Zenodo[2]. To make datasets discoverable, dataset publishers must offer the dataset metadata information per certain prescribed formats. For example, Google advocates the *Dataset*[3] type. However, many research datasets[4] do not come with a sufficient amount of structured metadata that fully describe their usage potentials. Therefore, discovery mechanisms used by search engines nowadays cannot detect and expose them to users. We have observed that metadata such as title, publisher, etc., contingent on their accuracy and maintenance [8], are insufficient to describe a research dataset's full context and content. Therefore, such limited metadata proves fairly uninformative to guarantee that relevant research datasets will be discovered on the Web.

Studies show that, in academia, the predominant search pattern for research datasets is either a serendipitous event of finding a dataset when reading scholarly publications or actively searching for datasets in publications [4]. This emphasizes that the content of scholarly publications, which gives insights into datasets, is necessary for dataset discovery [4]. Prior work highlights three criteria based on which users select datasets: *relevance*, *usability*, and *quality* [9]. These factors, resulting from human data interaction studies [10,11,12,13,14,15], necessitate the supply of content and context information of a dataset (e.g. domain(s) the dataset covers, source(s) it was gathered from, and metrics to evaluate it) to support informed decisions of its use for a task. We observe that such information is found in the scholarly publications that describe research datasets [4].

Therefore, the way forward toward improved research dataset discovery is to complement its metadata with a structured representation of the contributions of its accompanying scholarly article. To adhere to current standards of semantic descriptions, the representation of these contributions should be findable, accessible, interoperable, and reusable, i.e. it should adhere to the FAIR principles [16]. To this end, semantic publishing models [17] of scholarly contributions such as the Open Research Knowledge Graph (ORKG) can be directly leveraged [18]. The ORKG publishing model presents a next-generation skimming device of scholarly contributions, that permits viewing their semantic representations in a similar way to comparisons of products on e-commerce websites [19]. Thus, to model research datasets, we utilize the ORKG content type, which is a typed resource with a class from a predefined set of classes.

[2] `https://zenodo.org`

[3] `https://schema.org/Dataset`

[4] Note in this work we draw a distinction between the generic concept of a dataset on the Web and research datasets, in particular. Research datasets are those that are outcomes of a research endeavor and are thereby accompanied by a scholarly publication.

In this paper, we present the ORKG-Dataset content type – a specialized branch of the ORKG. The design of ORKG-Dataset was driven by three main research questions (RQs). **RQ1:** How to present structured research dataset descriptions within the semantic web scholarly publishing model as knowledge graphs (KGs)?; **RQ2:** Which salient features can be extracted from scholarly article descriptions that serve the dataset selection criteria of *relevance*, *usability*, and *quality*?; and **RQ3:** How can such a representation benefit others in terms of creating customizable snapshots of specific information?

The rest of this paper is structured as follows: Section 2 presents the design principles of ORKG-Dataset and how they were met within the ORKG, addressing RQ1. Section 3 demonstrates an application example of ORKG-Dataset on the scholarly publications of 40 research datasets used for scientific information extraction (IE), addressing RQ2 and RQ3. Finally, Section 4 concludes our paper.

2. Design Principles of ORKG-Dataset

Although previous work has been conducted to describe the content and context of datasets, it is not fine-grained enough in terms of the properties it offers. For example, the Data Source Description Vocabulary [5] and the Data Catalog Vocabulary (DCAT) [6] hav no possibility to describe models trained on the dataset and their evaluation scores. Additional resources for describing datasets such as releasing datasheets [20] are not modeled using semantic web technologies that enhance the usability and discoverability of datasets.

With this outlook, we outline the principles of ORKG-Dataset design using the following scenario. Imagine a researcher looking for research datasets in a particular field. Her search can be characterized based on two activities elicited in prior work [15], (1) linking (i.e. "finding commonalities and differences between two or more datasets" [15]) and (2) time series analysis (i.e. ordering datasets on a timeline). In the present status quo of scholarly communication, a large share of the information discussed above is already available, but it is hidden within the unstructured text of scholarly articles accompanying the research datasets. However, searching for a dataset by examining unstructured text descriptions involves significant cognitive tie-ups which boils down to finding a needle in a haystack. The researcher would need to sift through millions of results from academic search engines, identify those that actually contribute a dataset, and tediously search through the articles for key information before finding a suitable dataset.

The ORKG-Dataset content type introduced in this paper is one step in a long-term research agenda of the ORKG to bring about a paradigm shift from document-based to structured knowledge-based scholarly communication [18]. Specifically, the salient aspects of research are encoded as structured property-value pairs within the ORKG. Given the salient structured format applied on scholarly communication, the two search characteristics, i. e., linking and time series analysis, are directly supported in the interaction mechanisms of the ORKG front-end interface. Several structured papers with similar properties can be combined and placed next to each other within a comparison view [19]. The ORKG platform combines semantic web technologies with front-end design components and back-end storage and query systems [21]. It utilizes Resource Descrip-

[5] `https://dqm.faw.jku.at/ontologies/dsd/4.0.0/index.html`
[6] `https://www.w3.org/TR/vocab-dcat-3/`

tion Framework (RDF) as its default graph data representation language[7], connecting ontologies through subject-predicate-object triples. The front-end interfaces are built with ReactJS, fetching data from back-end APIs, while the Neo4J storage software[8] enables effective data querying using SPARQL[9].

Our proposed design requires each paper contribution to be typed as both the default `https://orkg.org/class/Contribution` class and the `https://orkg.org/class/Dataset` class. This is critical to separate other kinds of research contributions from research dataset contributions in the ORKG. We identified the following design requirements in order to generate comparable and wholesome dataset structured representations as the ORKG-Dataset content type.

- **Standardized Nomenclature:** We established a standard nomenclature for research datasets, starting with a controlled vocabulary that can later evolve into an ontology. This is achieved by reusing concepts from existing metadata ontologies such as `https://schema.org/Dataset`. We then added predicates specific to research datasets in the ORKG web namespace to enhance the vocabulary. To establish further equivalences between predicates in different ontologies, the RDF *same-as* relation was utilized.
- **Use of Templates**: to maintain consistent formatting when recording new research datasets, it was essential to define a form-based template comprising a set of predetermined relevant predicates. The ORKG facilitates this requirement by implementing a template system that consists of recurring subgraph property patterns [10]. These templates enable the specification of commonly applicable properties across multiple research contributions within a KG.
- **FAIR Standards Compliance**: the third and final requirement is that the ORKG-Dataset model should be compliant as much as possible with the FAIR (Findable, Accessible, Interoperable, and Reusable) guiding principles laid out for scientific information [16]. Addressing *findability*, the ORKG has a system to assign digital object identifiers (DOIs) to aggregate component parts of its graph data, making them available in global scholarly infrastructures such as DataCite and Crossref [22]. ORKG resources are also findable via regular search engines due to being published on the web. In terms of *accessibility*, all ORKG components are accessible by HTTP protocol via REST or its user interface. Additionally, the actual graph data is separated from metadata making it possible to access the latter without the former [19]. The ORKG satisfies *interoperability* by using RDF, the recommended format of the W3C for representing knowledge on the web in a machine-accessible format. To ensure *reusability*, provenance metadata information is created automatically when publishing structured contributions in the ORKG, and the graph data is published as CC BY-SA.[11]

[7] `https://www.w3.org/RDF/`
[8] `https://neo4j.com`
[9] `https://www.w3.org/TR/rdf-sparql-query/`
[10] For more information about templates in the ORKG: `https://orkg.org/about/19/Templates`
[11] `https://creativecommons.org/licenses/by-sa/2.0/`

3. The ORKG-Dataset Application

In this section, we demonstrate a use-case application of the ORKG-Dataset content type on research datasets in the field of natural language processing (NLP) on the research problem of scientific information extraction.

3.1. Datasets Curation

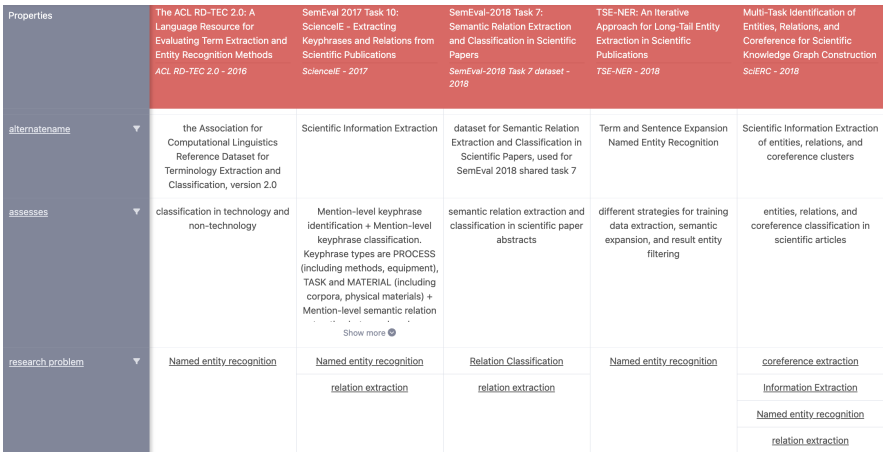
As a first step, we manually curated a collection of research datasets relevant to the problem at hand. We searched through benchmarking catalogs such as PapersWithCode[12], competition websites such as Kaggle[13], academic search engines such as Google Scholar[14], and systematic review papers [23]. From these sources, we finally arrived at a representative list of 40 research datasets spanning the years 2011 to 2022.

3.2. Datasets Semantic Representation

The next step was to create a structured representation for each of the research datasets based on their accompanying paper contributions. We, as a team of four annotators specialized in the field of scientific information extraction, followed an iterative methodology to identify the main contribution components. Our discussions resulted in the following main facets of information.

- **Research Problems**: our initial search in scientific IE identified research datasets addressing various sub-problems. Some examples include citation classification, sentence classification, rhetorics annotation, relation extraction, coreference resolution, automated leaderboard construction, knowledge graph construction, scientific claim verification, text summarization, and text generation. This was modeled with the ORKG predicate *research problem* (`https://orkg.org/property/P32`) and thus offers dataset consumers a clear indicator of *relevance* to their specific tasks.
- **Statistical attributes**: often when using machine learning methods, developers require additional statistics. E.g., for sentence classification datasets, how many sentences were annotated, and for how many documents. As such, we bundled nine relevant properties within a statistics template (`https://orkg.org/template/R220250`) to facilitate uniform modeling of this information across research datasets. Statistics information offers a direct *usability* indicator to the dataset consumers.
- **Quality**: one way to reflect a dataset's annotation quality is by inter-annotator agreement (IAA) scores, e.g. by using Cohen's kappa [24]. This quality indicator can be specific to different information scopes, such as entities, relations, or sentences. To represent this information, we created the *Data-centric result* template (`https://orkg.org/template/R220939`), which records the evaluation score and linked metric using the QUDT standardized methodology for evaluations (`https://qudt.org/schema/qudt/Quantity`). Additionally, the *has*

[12] `https://paperswithcode.com/datasets`
[13] `https://www.kaggle.com/datasets`
[14] `https://scholar.google.de`



Properties	The ACL RD-TEC 2.0: A Language Resource for Evaluating Term Extraction and Entity Recognition Methods *ACL RD-TEC 2.0 - 2016*	SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications *ScienceIE - 2017*	SemEval-2018 Task 7: Semantic Relation Extraction and Classification in Scientific Papers *SemEval-2018 Task 7 dataset - 2018*	TSE-NER: An Iterative Approach for Long-Tail Entity Extraction in Scientific Publications *TSE-NER - 2018*	Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction *SciERC - 2018*
alternatename	the Association for Computational Linguistics Reference Dataset for Terminology Extraction and Classification, version 2.0	Scientific Information Extraction	dataset for Semantic Relation Extraction and Classification in Scientific Papers, used for SemEval 2018 shared task 7	Term and Sentence Expansion Named Entity Recognition	Scientific Information Extraction of entities, relations, and coreference clusters
assesses	classification in technology and non-technology	Mention-level keyphrase identification + Mention-level keyphrase classification. Keyphrase types are PROCESS (including methods, equipment), TASK and MATERIAL (including corpora, physical materials) + Mention-level semantic relation Show more	semantic relation extraction and classification in scientific paper abstracts	different strategies for training data extraction, semantic expansion, and result entity filtering	entities, relations, and coreference classification in scientific articles
research problem	Named entity recognition	Named entity recognition relation extraction	Relation Classification relation extraction	Named entity recognition	coreference extraction Information Extraction Named entity recognition relation extraction

Figure 1. Excerpt of a screenshot of research datasets addressing scientific IE in the ORKG comparison view with structured metadata descriptions based on a set of properties defined as the ORKG-Dataset content type. The full comparison of 40 research datasets is accessible at `https://orkg.org/comparison/R280270/`.

evaluation item property (`https://orkg.org/property/P71154`), modeled by the nested template *Evaluation item* (`https://orkg.org/template/R221194`), allows specifying the granularity of annotations. This comprehensive unit of information serves as one direct indicator of dataset *quality*.

- **Performance Benchmarks**: Scholarly articles of research datasets report performance benchmarks. We identified this as an indirect *quality* indicator for dataset consumers. We modeled this aspect with the help of the existing *Leaderboard* template (`https://orkg.org/template/R107801`) which includes properties that allow specification of model name, model code URL, and allows specifying the score and metric per QUDT standards.
- **Metadata**: technologies for efficient and effective reuse of ontological knowledge are one of the key success factors for developing ontology-based systems [25]. In this vein, we reused 19 relevant properties from the `https://schema.org/Dataset` content type. Some of the properties are *name*, *alternatename*, *asseses*, *description*, and *URL*. Of particular interest is the URL property which is used to record the URL source where the dataset can be downloaded. These properties were modeled as the generic Dataset template (`https://orkg.org/template/R178304`) and could be uniformly applied across the 40 papers.

Putting all the pieces together, each of the 40 structured papers was finally typed with the ORKG-Dataset content type class `https://orkg.org/class/Dataset` which links to the generic Dataset metadata template. The result of our annotations is publicly accessible as an ORKG Comparison view `https://orkg.org/comparison/R280270/`. Figure 1 shows a partial screenshot.

3.3. Customizable Querying

A unique benefit of the semantic representation of research datasets in the ORKG is the ability to create customizable snapshots of the interconnected data graph in the context

of the larger graph data capturing various other kinds of scholarly contributions. This presents users with advanced search and selection options that can be implemented via SPARQL queries. We elicit this in the following three scenarios.

3.3.1. Bibliometric view.

With regard to bibliometrics, one could obtain detailed metadata about the authors, publication dates, and citation numbers. For example, researchers could get citation statistics and obtain the most cited dataset for a particular task during a specific period, in order to get some idea of the design and specificities of datasets that had a high impact in the community.

3.3.2. Dataset view.

Another example query is finding ground-truth datasets that could be used to train a model to solve a particular task. One such example query can be found in Figure 2. In addition, it is possible to filter for datasets that label particular entities, such as "methods" or "materials" (Figure 3), and which match particular inter-annotator agreement scores if available. Most of the papers come with the actual URLs under which the published ground-truth datasets could be found. Thus, one could fetch the URL for all datasets of interest.

```
SELECT ?task (GROUP_CONCAT(?dataset;separator=',') AS ?dataset)
WHERE {
  res:R280270 pred:compareContribution ?contribution .
  ?contribution a class:Dataset ;
     rdfs:label ?dataset .
  ?contribution pred:P32/rdfs:label ?task
}
GROUP BY ?task
```

Figure 2. Example query to obtain a list of ground truth datasets and the tasks they address. Full query: `https://tinyurl.com/query-example-1`.

```
SELECT DISTINCT ?concept GROUP_CONCAT(?dataset;separator=',')
WHERE {
  res:R280270 pred:compareContribution ?contribution .
  ?contribution a class:Dataset ;
     rdfs:label ?dataset .
  ?contribution pred:P34062/rdfs:label ?concept .
  FILTER( ?concept = "Method"^^xsd:string
    OR ?concept = "Research problem"^^xsd:string)
}
GROUP BY ?concept
```

Figure 3. Example query to filter for datasets that label "Method" and "Research problem" as labeled entity types in the ground truth. Full query: `https://tinyurl.com/query-example-2`.

3.3.3. SOTA view.

An integrated comparison table as ours also allows to uncover implicit relations between entities, for example, to search for competing, state-of-the-art machine learning models. Competing models are models that have been used to solve a similar NLP task, whereas a non-competing model would be a model that has not been used directly to solve the same task [26]. We could also filter out models that score above a particular evaluation metric, for example, models which score about 0.7 value in F1-score. Querying for the mostly used evaluation metrics could give researchers an idea of the metrics most widely used by the community.

4. Conclusion

In this article, we presented the ORKG-Dataset content type as an approach for the semantic publishing of research datasets. The ORKG-Dataset content type breaks new ground in two main aspects: 1) in the studies of the transition from document-based to structured knowledge-based scholarly communication; and 2) of moving away from just metadata-based semantic descriptions of research datasets to including salient features of their accompanying scholarly publications as an enriched and more informative representation. Future developments of the ORKG-Dataset will apply it to more fields and thus make it more generalizable, as well as refine the properties to describe the quality of datasets using metrics beyond IAA.

Funding Statement

This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number: NFDI4DataScience (460234259).

References

[1] Cafarella MJ, Halevy A, Madhavan J. Structured data on the web. Communications of the ACM. 2011;54(2):72-9.
[2] Manyika J, Chui M, Farrell D, Van Kuiken S, Groves P, Almasi Doshi E. Open Data: Unlocking Innovation and Performance with Liquid Information— McKinsey & Company; 2014.
[3] Verhulst S, Young A. Open data impact when demand and supply meet key findings of the open data impact case studies. Available at SSRN 3141474. 2016.
[4] Gregory K, Cousijn H, Groth P, Scharnhorst A, Wyatt S. Understanding Data Retrieval Practices: A Social Informatics Perspective. 2018.
[5] Mayer-Schönberger V, Cukier K. Big data: A revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt; 2013.
[6] Brickley D, Burgess M, Noy N. Google Dataset Search: Building a search engine for datasets in an open Web ecosystem. In: The World Wide Web Conference; 2019. p. 1365-75.
[7] Manghi P, Bardi A, Atzori C, Baglioni M, Manola N, Schirrwagen J, et al. The OpenAIRE research graph data model. Zenodo. 2019.
[8] Chapman A, Simperl E, Koesten L, Konstantinidis G, Ibáñez LD, Kacprzak E, et al. Dataset search: a survey. The VLDB Journal. 2020;29(1):251-72.
[9] Koesten L, Simperl E, Blount T, Kacprzak E, Tennison J. Everything you always wanted to know about a dataset: Studies in data summarisation. International Journal of Human-Computer Studies. 2020;135:102367.

[10] Piwowar HA, Chapman WW. Public sharing of research datasets: a pilot study of associations. Journal of informetrics. 2010;4(2):148-56.

[11] Boukhelifa N, Perrin ME, Huron S, Eagan J. How data workers cope with uncertainty: A task characterisation study. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems; 2017. p. 3645-56.

[12] Gregory K, Groth P, Cousijn H, Scharnhorst A, Wyatt S. Searching data: a review of observational data retrieval practices in selected disciplines. Journal of the Association for Information Science and Technology. 2019;70(5):419-32.

[13] Kern D, Mathiak B. Are there any differences in data set retrieval compared to well-known literature retrieval? In: Research and Advanced Technology for Digital Libraries: 19th International Conference on Theory and Practice of Digital Libraries, TPDL 2015, Poznań, Poland, September 14-18, 2015, Proceedings 19. Springer; 2015. p. 197-208.

[14] Thoegersen JL, Borlund P. Researcher attitudes toward data sharing in public data repositories: a meta-evaluation of studies on researcher data sharing. Journal of Documentation. 2021;78(7):1-17.

[15] Koesten LM, Kacprzak E, Tennison JF, Simperl E. The Trials and Tribulations of Working with Structured Data: -a Study on Information Seeking Behaviour. In: Proceedings of the 2017 CHI conference on human factors in computing systems; 2017. p. 1277-89.

[16] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data. 2016;3(1):1-9.

[17] Shotton D. Semantic publishing: the coming revolution in scientific journal publishing. Learned Publishing. 2009;22(2):85-94.

[18] Auer S, Oelen A, Haris M, Stocker M, D'Souza J, Farfar KE, et al. Improving access to scientific literature with knowledge graphs. Bibliothek Forschung und Praxis. 2020;44(3):516-29.

[19] Oelen A, Jaradeh MY, Stocker M, Auer S. Generate FAIR Literature Surveys with Scholarly Knowledge Graphs. In: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020. JCDL '20. New York, NY, USA: Association for Computing Machinery; 2020. p. 97–106. Available from: `https://doi.org/10.1145/3383583.3398520`.

[20] Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Iii HD, et al. Datasheets for datasets. Communications of the ACM. 2021;64(12):86-92.

[21] Jaradeh MY, Oelen A, Farfar KE, Prinz M, D'Souza J, Kismihók G, et al. Open research knowledge graph: next generation infrastructure for semantic scholarly knowledge. In: Proceedings of the 10th International Conference on Knowledge Capture; 2019. p. 243-6.

[22] Haris M, Stocker M, Auer S. Persistent Identification and Interlinking of FAIR Scholarly Knowledge. arXiv preprint arXiv:220908789. 2022.

[23] Nasar Z, Jaffry SW, Malik MK. Information extraction from scientific articles: a survey. Scientometrics. 2018;117(3):1931-90.

[24] McHugh ML. Interrater reliability: the kappa statistic. Biochemia medica. 2012;22(3):276-82.

[25] Simperl E. Reusing ontologies on the Semantic Web: A feasibility study. Data & Knowledge Engineering. 2009;68(10):905-25.

[26] Daw S, Pudi V. Extraction of Competing Models using Distant Supervision and Graph Ranking. In: SDU@AAAI; 2022. .