A maturity model for catalogues of semantic artefacts

Oscar Corcho, Fajar Ekaputra, Ivan Heibi, Clement Jonquet, Andras Micsik, Silvio Peroni, Emanuele Storti

A maturity model for catalogues of semantic artefacts

Oscar Corcho[1][0000−0002−9260−0753], Fajar J. Ekaputra[2,3][0000−0003−4569−2496], Ivan Heibi[4,5][0000−0001−5366−5194], Clement Jonquet[6,7][0000−0002−2404−1582], Andras Micsik[8][0000−0001−9859−9186], Silvio Peroni[4,5][0000−0003−0530−4305], and Emanuele Storti[9,10][0000−0001−5966−6921]

[1] Ontology Engineering Group (OEG), Computer Science School, Universidad Politécnica de Madrid, Madrid, Spain
ocorcho@fi.upm.es
[2] DPKM, Vienna University of Economic and Business, Vienna, Austria
[3] Data Science Research Group, TU Wien, Vienna, Austria
fajar.ekaputra@wu.ac.at
[4] Digital Humanities Advanced Research Centre (/DH.arc), Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy
[5] Research Centre for Open Scholarly Metadata, Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy
{ivan.heibi2,silvio.peroni}@unibo.it
[6] MISTEA, University of Montpellier, INRAE & Institut Agro, France
[7] LIRMM, University of Montpellier & CNRS, France
jonquet@lirmm.fr
[8] Department of Distributed Systems (DSD), Institute for Computer Science and Control (SZTAKI), Eötvös Loránd Research Network (ELKH), Budapest, Hungary
andras.micsik@sztaki.hu
[9] Department of Information Engineering, Polytechnic University of Marche, Ancona, Italy
[10] European Council of Doctoral Candidates and Junior Researchers (Eurodoc), Brussels, Belgium
e.storti@univpm.it

Abstract. The work presented in this paper is twofold. On the one hand, we aim to define the concept of *semantic artefact catalogue* (SAC) by over viewing various definitions used to clarify the meaning of our target of observation, including the meaning of the focal item: semantic artefacts. On the other hand, we aim to identify metrics and dimensions that can be used to assess the maturity of such catalogues. In particular, we define a maturity model to measure, compare and evaluate available semantic artefact catalogues. By setting these dimensions and their metrics, catalogues can be classified by each dimension. So the maturity of both the catalogues and the dimensions as a whole can be expressed. Such a maturity model and its application to 26 semantic artefacts catalogues —from various disciplines and relying on various technologies— are available to be later enriched.

Keywords: semantic artefacts · ontologies · semantic artefact catalogues · ontology repositories · ontology libraries · semantic interoperability · EOSC.

1 Introduction

With the advent of Open Data [41], the Open Science movement [53], and the FAIR Principles [57] in the scholarly ecosystem, the role and need for storing, managing and sharing data grew significantly in academia. In the past years, The General Data Protection Regulation (GDPR) was an important step around data management in Europe and was the main responsible, at least in the beginning, for scientists' fears of making their work impossible as data scientists. Indeed, one of the reasons for the introduction of the European Open Science Cloud (EOSC) has been for providing a safe European environment for data management compliant with the GDPR, to avoid the risk European scientists start to entrust all their data to foreign owned/registered data servers to bypass European laws [10].

In the EOSC, a strategic relevance has been given, since the beginning, to the issues to address for implementing a real interoperability among all the infrastructures, services and, of course, data that are shared by researchers in this European cloud. Indeed, one of the most cited and used document produced in the past years was the EU report about the EOSC Interoperability Framework. The goal of this document was to identify "the general principles that should drive the creation of the EOSC Interoperability Framework (EOSC IF), and organises them into the four layers [...]: technical, semantic, organisational and legal interoperability" [16].

Two years ago, moved by the principles highlighted in the EOSC IF report, the EOSC Association (`https://www.eosc.eu`) has promoted the creation of several task forces (EOSC TFs) dedicated to specific aspects to address for enabling the implementation of the EOSC. One of these task forces has been entirely dedicated to *semantic interoperability* (`https://www.eosc.eu/advisory-groups/semantic-interoperability`), that is the means to ensure "that the precise format and meaning of exchanged data and information is preserved and understood throughout exchanges between parties" [16]. The charter of this task force [4], and the subsequent working activities held in this context, has concerned several aspects of semantic interoperability. One of these was entirely focused on the management and sharing of the main tool for enabling semantic interoperability, i.e. *semantic artefacts*. A semantic artefact is a machine-actionable and machine-readable formalisation of a conceptualisation, enabling sharing and reuse by humans and machines, that may have a broad range of formalisation, from loose sets of terms, taxonomies, thesauri to higher-order logic constructs, vocabularies and ontologies. Often, these semantic artefacts are stored and shared by means of specific services called registries, libraries, repositories, catalogues, or simply terminology/vocabulary servers, each providing a mixture of functionality — ranging from simple metadata descriptions to

advanced content-based services — in order to facilitate finding, accessing, understanding and re-using of such semantic artefacts and enabling their long-term preservation.

Considering this context, the paper presents the outcomes of an extensive analysis done by the EOSC TF on Semantic Interoperability aiming at answering the following research questions:

1. What is a catalogue of semantic artefacts?
2. Which metrics and dimensions can be used to assess the maturity of such catalogues?

To answer these research questions, we have gathered various definitions concerning the concept of semantic artefact and of catalogues storing and serving them (either at the metadata or data level or both). Then, we have defined, by analysing the current literature on the topic, a model to measure, compare and evaluate available semantic artefact catalogues. We present this *maturity model* [17], i.e. the main resource introduced in this paper, as composed by several *dimensions* in which catalogues could be compliant and/or improved. Catalogues can be classified by each dimension, and so the maturity of both the catalogues and the dimensions as a whole can be expressed. In addition, we analysed a collection of 26 semantic artefacts catalogues, aiming, on the one hand, at completing the maturity model by adding additional features (or sub-criteria) for each of the dimensions identified and, on the other hand, at showing how existing catalogues comply with such dimensions and sub-criteria.

The structure of the paper follows our overall approach of investigation. In Section 2, we introduce the details of collecting definitions and maturity dimensions. Section 3 describes the selected material and our methods for analysis. Section 4 presents the analysis results also giving an overview of current state, while Section 5 discusses the outcomes and outlines lessons learned. Finally, Section 6 concludes the paper sketching out possible future developments.

2 Literature review

2.1 Process for collecting relevant works

We involved all the members of the EOSC Task Force on Semantic Interoperability (EOSC-TF-SI) to provide us relevant material related to that includes at least one of the two aspects of interest for our study, i.e., (1) definitions of *semantic artefact* catalogues and (2) dimensions that can be used to measure the maturity of such catalogues.

After having gathered all such relevant documents, we asked to some members of the EOSC-TF-SI to read them and highlight any passage in the text referring either to definitions related to catalogues of semantic artefacts or maturity measures. Overall, we found that fifteen of the gathered documents contained relevant texts, as shown in Table 1. The raw data of this overview are available in [11].

Table 1. Documents introducing some definitions of catalogues of semantic artefacts and highlighting at least one maturity dimension identified by the analysis (Me: Metadata, Op: Openess, Qu: Quality, Av: Availability, St: Statistics, Pi: PID, Go: Governance, Co: Community, Su: Sustainability, Te: Technology, Tr: Transparency, As: Assessment).

Document	Definition	Me	Op	Qu	Av	St	Pi	Go	Co	Su	Te	Tr	As
[2]		x											
[5]		x	x	x	x								
[6]			x		x			x		x		x	
[8]		x				x	x						
[15]			x		x			x			x	x	
[16]	x	x	x		x		x	x			x		
[18]		x		x			x		x				
[23]	x		x					x	x	x			
[25]		x	x	x	x		x	x		x		x	x
[28]				x					x		x		
[31]	x	x	x		x		x			x	x	x	
[34]	x	x	x		x	x	x		x	x	x	x	
[38]	x							x	x	x	x	x	
[47]			x		x			x	x	x	x	x	
[48]													x

2.2 Definitions of catalogues of semantic artefacts

Before defining a catalogue of semantic artefacts, we need to first agree on a definition to adopt for a semantic artefact.

Previous studies used terms such as Knowledge Organization Systems (KOS) [61] or knowledge artefact [39] to address semantic artefacts. A KOS has been adopted as a general term to encompass all types of schemes used to organise information and promote knowledge management, such as classification schemes, gazetteers, lexical databases, taxonomies, thesauri, and ontologies. The aim of these schemes is to underline the semantic structure of a domain, which needs to be embodied as web services to facilitate a resource discovery and retrieval (by acting as semantic road maps) for either humans or machines.

Considering other more recent works, a semantic artefact has been defined as a machine-actionable and -readable formalisation of a conceptualisation enabling sharing and reuse by humans and machines [16][31]. Semantic artefacts may have a broad range of formalisation, which include ontologies, terminologies, taxonomies, thesauri, vocabularies, metadata schemas and standards [16][31]. The term semantic artefact was also strongly advised as an overarching term in the conetxt of the H2020 FAIRsFAIR project's task on "FAIR semantics". Despite the different forms of a semantic artefact, some works used blanket term such as "ontologies" or "vocabularies and ontologies" [33]. Moreover, semantic artefacts are serialised using a variety of digital representation formats, e.g., RDF Turtle, and OWL, using XML (RDF) and JSON-LD [31].

The notion of ontology library was introduced in [20], defined as "A library system that offers various functions for managing, adapting and standardizing groups of ontologies". In addition, [20] highlighted the importance of making such libraries easily accessible and offer efficient support for re-using existing relevant ontologies and standardizing them based on upper-level ontologies and ontology representation languages.

The terms "collection", "listing" or "registry" are also used to describe ontology libraries. All correspond to systems that help reuse or find ontologies by simply listing them (e.g., DAML or DERI listings) or by offering structured metadata to describe them (e.g., FAIRSharing, BARTOC, Agrisemantics Map). Yet, those systems do not support additional services that go beyond the description of the items, e.g., a content analysis of the ontologies or a search index on the ontology content [34]. A new concept introduced by [30] to cover these aspects is an ontology repository with advanced features which enables search, browsing, managing metadata, customizing, and mapping an application to query the contents of the ontologies. [21] and [42] provide the latest reviews of ontology repositories.

By the end of the 2000's, the topic was of high interest as illustrated by the 2010 ORES workshop [1] or the 2008 Ontology Summit (`http://ontolog.cim3.net/wiki/OntologySummit2008`). The Open Ontology Repository Initiative [3] aimed to create a joint infrastructure of ontology repositories through collaboration. At the time, the initiative utilized the NCBO BioPortal technology [56], which was the most advanced open-source technology for ontology management, but it was not yet available as a "virtual appliance" as it is today. Later, the initiative considered using the OntoHub [40] technology for broader application, but it has since been discontinued.

Other recent works used to refer to catalogue of semantic artefacts with terms such as "repository" [16] [38] and "registry" [16], as well as hypernyms such as "infrastructure" and "service" [16][23]. Additionally, the FAIRsFAIR project employed the term "semantic registry," which is defined as a "catalogue that contains metadata about semantic artefacts" [31].

Two of the works we have considered, while talking about catalogues, do not focus specifically on semantic artefacts. In particular, [23] presents the generic term open science infrastructures and clarifies that they are "services, protocols, standards and software that the academic ecosystem needs to perform its functions during the research lifecycle". Instead, [38] introduce the concept of trustworthy digital repositories and provides an operational definition for them that should "remit to actively preserve data in response to changes in both technology and stakeholder requirements". The only work referring explicitly to semantic artefacts as the kind of items contained in the catalogue is [16].

Considering the status we have presented, there is a clear need to adopt an inclusive definition that, in principle, enables us to consider as a catalogue also web pages (e.g. https://w3id.org/mobility) with descriptive metadata of the semantic artefacts included in the catalogue in human-readable form.

2.3 Identification of maturity dimensions for catalogues

From the document analysed, we have extracted 12 dimensions that can be used to measure the maturity of the catalogues of semantic artefacts. These dimensions, summarised in Table 1 indicating the documents that describe or refer to them, are listed as follows:

- **Metadata (Me)**. The identification of the minimal set of metadata to describe the catalogue and its semantic artefacts. Huge importance is also given to the use of metadata standards and schemas (e.g., DCAT or Schema.org), the adoption of machine-readable formats, the documentation associated, and the licenses used to release the metadata.
- **Openness (Op)**. The concept of being open from different perspectives. On the one hand, it concerns technical openness, referring to the metadata handled in the catalogue, the software used to run the catalogue, and the services and protocols used to access the metadata. On the other hand, openness also refers to the social attitude of enabling anyone interested in depositing and also helping govern the catalogue.
- **Quality (Qu)**. The possibility of having mechanisms to check the quality of the metadata provided and, thus, of the catalogue itself. In particular, if processes and workflow are in place for peer reviewing new entities and curating the catalogue.
- **Availability (Av)**. It refers to the availability of the metadata and if there are methods in place for guaranteeing privacy and access only to certain data due to legal or other contextual issues.
- **Statistics (St)**. The availability of statistics referred to the catalogue (number of semantic artefacts handled, number of users, etc.) in time to measure the usage of the catalogue and its growth.
- **PID (Pi)**. The use of persistent identifiers (PIDs) referring to the metadata of the various semantic artefacts described in the catalogue and their contextual entities (author, curator, etc.).
- **Governance (Go)**. The rules to define the governance of the catalogue and its goals and purpose, which should allow community input and responsibility for the integrity of the metadata.
- **Community (Co)**. The mechanism in place to involve the community in the catalogue, identifying and reaching target users' expectations and attracting stakeholders from diverse lived experiences and viewpoints.
- **Sustainability (Su)**. The models in place to sustain services financially and preserve the catalogue in the long run.
- **Technology (Te)**. The tools that the catalogue should provide to enable users to have a better experience in exploring the data, such as REST APIs, Web search interfaces, SPARQL endpoints, etc.
- **Transparency (Tr)**. The processes behind the catalogue, from the elections of new members of the various governing boards, curators, etc., to the clarity in exposing fees for the services offered by the catalogue itself and its revenue model.

– **Assessment (As)**. The presence of some practice in place for assessing the catalogue against all these dimensions, e.g. by adopting self-assessment exercises and/or by asking third parties to run an independent assessment of the catalogue.

3 Methods and material

This section presents the methodology followed for the analysis of semantic artefact catalogues. The process is divided into three steps: 1) selection of catalogues, 2) setup of the assessment process, 3) analysis of the catalogues, and 4) harmonization and summarization.

3.1 Collection of catalogues

To collect potential catalogues of semantic artefacts, a preliminary search was conducted. Potential catalogues of semantic artefacts have been identified by direct knowledge of the co-authors or members of the EOSC-TF-SI-TF. The resulting list of potential catalogues was then screened to remove duplicates and those that were clearly irrelevant to the study. In particular, we decided to keep in the analysis only those potential catalogues that refer mainly to semantic artefacts. This exclusion criteria has been made to filter out (i) generic repositories that may also contain semantic artefacts, even if it is not the primary resource types they refer to (e.g., Zenodo), and (ii) generic repositories (e.g., Google or other general-purpose search engines). The resulting set included 26 selected catalogues. With this list, our goal was not to be exhaustive but rather to cover well multiple application domains and also get a good representation in terms of underlying technology used to build the catalogues (e.g., OntoPortal, OLS, SKOSMOS, etc.).

3.2 Setup of the assessment

The identified catalogues were evaluated based on their relevance to semantic artefacts. A spreadsheet was created with the selected catalogues listed on rows and the 12 maturity dimensions on columns. Additional columns were dedicated to the names of the reviewers and comments. Each reviewer was assigned a number of catalogues to review.

Twelve separate tabs in the spreadsheet were devoted to describe the possible values to use for each maturity dimension. These tabs contained the name of the dimension and a set of particular features, each with a number, a description, the name of the reviewer that proposed it, and whether it had been validated by the group. In addition, there was a column for possible comments.

The main table in the first tab was extended with two additional columns. The first column allowed us to specify whether the catalogue store a copy of the semantic artefacts it describes. The second column referred to the open software or tool used to implement the catalogue in order to distinguish if the software is generic (i.e., can be used to deploy multiple catalogues) and open-source.

3.3 Analysis

Each reviewer had the responsibility to evaluate a first small set of 2-3 catalogues. For each catalog and each dimension, the reviewer had to select which features of the dimensions applied for the catalogue at hand. If, for a given dimension, a certain feature was not already present, the reviewer could add such a new feature, making it available for the next reviewers. A number from 3 to 7 features for each dimension have been added during this step for a total of 63 features.

After the first analysis, early results and issues were discussed by all reviewers. The analysis was then extended by assessing other potential catalogues, ideally 5 catalogues each, to have a clear view of other aspects that did not arise from the former analysis. In addition, we decided to invite other interested people to such a study afterwards, ideally after we have assessed all the potential catalogues. Our perceived risk here is to be biased by a particular point of view while we are still in a phase of preliminary analysis.

3.4 Harmonization and summarization

Following the completion of the catalogue descriptions, a final review and summarization process was conducted. Each reviewer was tasked with analyzing 2 to 4 dimensions, where they reviewed the corresponding features provided by other reviewers and suggested potential edits such as merging similar features, removing irrelevant ones, or splitting them into separate aspects. As a result, 16 out of 63 identified features have been removed from the analysis, keeping 47 features. The main table was updated accordingly to reflect any changes made by the reviewer. A final meeting was held to review and harmonize the catalogue dimensions and features, in order to ensure a consistent set of dimensions and features that could be used to compare and analyze the different catalogues.

During this meeting, any remaining inconsistencies or ambiguities in the dimension features have been discussed and made final decisions on how to harmonize them across all catalogues.

4 Results

The section reports our assessment of the selected metadata catalogues. As a result, we identified a set of distinct features for each dimension (Section 4.1), and provide the characterization of the selected semantic artefact catalogues according to the dimensions and their features (Section 4.2).

4.1 Dimension Features

In this section, we provide a brief description of the identified features (or subcriteria) from each dimension as the basis of our assessment. These features resulted from the harmonization effort (cf. Section 3.4).

Metadata. (a) *custom vocabulary* - custom metadata is used to describe semantic artefacts, (b) *standard vocabulary* - a well-known, widely shared or standard metadata vocabulary is used, (c) *primary metadata* - the original semantic artefact metadata are preserved in the catalog, (d) *version metadata* - metadata for each distribution/version of semantic artefact is available, (e) *human readable* - metadata is visible in the user interface in an harmonised manner, (f) *machine readable* - metadata is accessible via API or machine supported formats.

Openness. (a) *fully oss* - based on open source software, (b) *customised oss* - the catalog is based on an open source software but the customised instance is not available for public, (c) *open model* - the metadata model / ontology used to document the semantic artefacts is openly available, and (d) *open contribution* - external or registered users can add/propose new semantic artefacts for inclusion.

Quality. (a) *curation by owner* - changes (or new submissions) to the semantic artefact can only be conducted by the catalog owner, (b) *curation by maintainer* - changes (or new submissions) to the semantic artefact can be made by the maintainers/curators of the artefact, (c) *certified maintainer* - the maintainers/curators of the semantic artefacts are certified and assigned by the catalogue owner, (d) *metadata by editor* - metadata is curated by a group of editors, and (e) *metadata by system* - metadata is generated/curated by an assessment system.

Availability. (a) *no restriction* - no authentication methods provided; contents are freely available without restrictions, (b) *multilinguality* - items are translated and available in several languages, and (c) *moderated services* - some functionalities for access and modification of semantic artefact are available only to registered users and content creators.

Statistics. (a) *catalog statistics* - basic metrics about the metadata catalog, (b) *resource statistics* - metrics on each semantic artefact, and (c) *social metrics* - social metrics for semantic artefacts, e.g., stars, likes, and number of contributors.

PID. (a) *metadata record PID* - PID metadata might be specified in the semantic artefact record (e.g. ORCIDs for curators), and (b) *resource PID* - PID used to identify the semantic artefact object.

Governance. (a) *3rd party* - items are managed via 3rd party tool, e.g. GitHub, (b) *description* - governance is described as part of the catalog, and (c) *rules* - rules for proposing new items to the catalogue are introduced.

Community. (a) *read only* - no direct involvement is possible; users can only communicate to the catalog via read-only API and email, (b) *read write* - curators and developers can use services to get information from the catalogue and be directly involved through the creation of their own records to be added to the catalogue and increase their visibility, (c) *3rd party* - community features delegated to 3rd party tool, e.g. GitHub issues, and (d) *suggestion* a dedicated page for content suggestion is available.

Sustainability. (a) *organization* - the catalog is a service provided by an organization (university, institute or one of its research units), (b) *community* - the catalogue is maintained by a community with members from various organizations or infrastructure, (c) *management board* - a multidisciplinary community-driven service, strongly sustained by an operational team, and (d) *(research) project* - sustained by funds coming from one or more projects.

Technology. (a) *REST API* - a service to access semantic artefact information and/or metadata via a REST interface, (b) *web search GUI* - a service to access semantic artefact information and/or metadata via a web search GUI, (c) *SPARQL endpoint* - a service to access semantic artefact information and/or metadata via a SPARQL endpoint, and (d) *alignment* - a service to align (part of) semantic artefacts that might be used within a catalog.

Transparency. (a) *documented curation* - data flow of curation is documented, (b) *automatic curation* - curation process happened automatically based on a documented process flow, and (c) *resource versioning* - records on previous version of items are available.

Assessment. (a) *shared metrics* - assessment in terms of FAIRness is provided, and (b) *custom metrics* - assessment against catalog's own assessment metrics.

4.2 Assessment Result

We provide the result of our assessment as Table 2. In addition to the assessment dimensions and dimension features, we also observe the type of the catalogue. Specifically, we look into catalogues containing both data and metadata, which is classified as a *repository* according to the classification from Jonquet [34]. These catalogs are marked with an asterisk (*) in the table header alongside the catalogue names. The raw data of Table 2 are available in [11].

Table 2. An overview of assessment results. Asterisk in the name means that the catalogue also store a copy of the semantic artefacts it describes.

	Dimension & Features	RDA Registry [45]	ARCO* [12]	BioPortal* [43]	TS4TIB* [50]	Archivo* [26]	LOV* [54]	Prefix.cc [19]	EU-Vocabularies* [52]	ODP* [27]	SPAR* [44]	FAIR Sharing [46]	AgroPortal* [33]	JFOW* [51]	OBO Foundry [49]	Bartoc* [55]	EBI-OLS [35]	IVOA* [32]	MatPortal* [24]	EcoPortal* [36]	Loterre* [13]	MedPortal* [7]	ESIP Portal* [22]	NERC NVS [9]	Ontobee [60]	HeTOP [29]	Ontohub* [14]
Me	custom vocabulary		x	x	x	x	x				x	x	x		x		x	x	x	x	x	x	x	x	x	x	x
	standard vocabulary								x				x				x				x							
	primary metadata	x		x	x		x		x			x	x						x	x		x						
	version metadata			x		x							x						x	x		x						
	human readable			x						x	x		x		x	x		x	x	x		x	x	x	x		x
	machine readable			x			x		x				x		x	x			x	x		x	x	x				
Op	fully oss			x		x	x	x		x	x		x		x	x			x	x				x	x		x
	customised oss				x												x						x				
	open model		x			x						x	x								x			x		x	x
	open contribution			x		x	x	x	x	x	x	x	x		x			x	x	x	x		x	x	x			x
Qu	curation by owner only		x		x												x					x						
	curation by maintainer			x			x		x	x	x	x	x	x	x			x	x	x	x							
	certified maintainer	x										x																
	metadata by editor												x			x				x								
	metadata by system					x							x															
Av	no restrictions	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x			x
	multilinguality	x							x												x					x		
	moderated services			x								x	x						x	x		x		x		x		
St	catalog statistics	x		x	x	x	x					x	x						x	x		x						x
	resource statistics	x		x	x	x	x		x				x						x	x	x	x			x	x		
	social metrics			x									x		x				x	x			x					
Pi	metadata record PID											x			x			x										
	resource PID		x	x					x		x		x		x		x		x	x		x						
Go	3rd party	x												x	x													
	description		x	x		x	x		x	x			x		x		x	x	x	x	x				x			
	rules			x		x	x		x	x	x		x		x				x	x	x		x					
Co	read only		x		x		x		x							x	x				x					x	
	read and write			x				x		x		x	x						x	x		x						
	3rd party	x								x	x			x	x		x											x
	suggestion			x		x							x					x	x	x		x	x	x	x			
Su	organization	x	x	x	x	x		x	x	x		x	x			x	x		x	x	x	x		x	x	x		
	community						x				x				x	x												x
	management board									x		x			x	x							x					x
	(research) projects			x								x	x						x	x			x					
Te	REST API			x	x	x	x	x	x			x	x			x	x		x	x	x	x	x	x		x		
	web search GUI	x	x	x	x	x	x	x	x	x		x	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x
	SPARQL endpoint		x						x				x						x	x	x	x	x	x	x			
	alignment			x									x						x	x		x						
Tr	documented curation	x				x				x	x	x	x		x			x										x
	automatic curation					x							x															
	resource versioning			x		x							x		x				x	x		x	x					x
As	shared metrics					x							x															
	custom metrics														x													

5 Discussion

All the catalogues maintained by a community with members from different organizations/infrastructures (5 out of 26), are open-source and provide no authentication methods/restrictions to their contents. Furthermore, catalogues that enable external/registered users to add/propose new semantic artefacts to be included in the catalogue (65%), delegate the quality control of these changes to the maintainers/curators of the artefacts. Indeed, generally, more than half of the analysed catalogues (65%) permit the curation of their data by either the owners of the catalogue or by the maintainers of the semantic artefacts: no catalogue provides both strategies. This aspect is a positive sign which shows that the majority of the catalogues are concerned with guaranteeing a good quality to their data.

In two cases the sustainability could not be assessed. In 5 cases projects are the only source of funding, which may be temporal. In the remaining, the sustainability seems sufficiently stable. Furthermore, all the catalogues maintained by a community with members from different organizations/infrastructures (19%), are open-source and provide no authentication methods/restrictions to their contents.

The wide majority of catalogues (92%) provide at least a search web GUI. The non-SPARQL catalogues represent 61% (16) of the total, around 88% of these use a web search GUI and 56% combine it also with REST APIs. Therefore, only 10 out of 26 provide a SPARQL endpoint. 70% of these provide machine-readable metadata, in fact, the preferred technology used by all the catalogues integrating machine-readable metadata is either REST APIs, web search GUI, or both. A relatively large number of catalogues incorporate all three basic technologies (30%), i.e. REST API, web search GUI, and SPARQL endpoint. These catalogues provide data with no authentications/restrictions and contents are freely available without restrictions, in addition, both external/registered users can add/propose new semantic artefacts to be included.

All catalogues, with only one exception, provide access with no restriction. However, 8 of them (i.e., around 31%) include functionalities, mostly as APIs, that are accessible only to registered users. Only a small percentage of catalogues (i.e., 11.5%) provide a PID (Persistent IDentifier) for the metadata of a given record, while a larger share (i.e., 38%) use PIDs for identifying resources, 88% of these latest group use custom metadata to describe semantic artefacts. In other words, the attribution of new custom metadata occurs with care toward the use of PIDs in such a process.

Almost half of the catalogues provide statistics on resources, typically including the number of classes, properties and axioms. Nine of them (i.e., around 35%) also provide general statistics which aggregate information across all resources. In a few case, also social statistics are included, e.g., in the form of metrics taken from GitHub metadata where the original resources are stored (e.g., number of received stars and contributors).

Governance processes and/or structure is described by more than a half of the catalogues, 71% of which also explicitly provide rules for contributors willing

to propose new resources. In 8 cases, external 3rd-party solutions, particularly GitHub, are used as management tools for the resources. Part of the catalogues are more open than others to the contribution of the community. In particular, while almost one-third only provide the capabilities to communicate with the catalogue through read-only APIs, only 8 out of 26 provide also the possibility for resource creation.

Only 9 catalogues (i.e., around 35%) explicitly document a workflow for data curation, which in two cases is mostly automated, 8 of these are open for modifications to be made to the semantic artefacts by external/registered users. In 9 cases, the catalogue provides also previous versions of the resource, enabling a versioning system useful for backward compatibility and documentation.

The vast majority of the catalogues do not provide information on (self) assessment against quality criteria. Among the few exceptions, AgroPortal [33] includes an assessment in terms of FAIR score, an evaluation on the satisfaction of each aspect of the 15 FAIR principles. FAIR score includes a number of questions specific for ontologies and semantic resources, with the capability to compute the score for each resource and for the whole catalogue. On the other hand, Archivo [26] proposes a rating based on a set of automatically assessed criteria (whether the ontology is retrievable and parsed correctly, is provided with a clear and proper license statement, and is logically consistent).

Adopting standard vocabulary for metadata is an essential aspect to join, compare, and curate the semantic artefacts, in addition, it fosters the interoperability of these items, yet, only 15% (4) use standard vocabularies. Regardless, further analysis is needed to definitely affirm that all the other catalogues, do not use standard vocabularies. Indeed, it is possible that these catalogues use custom vocabularies made by extending standard vocabularies.

6 Conclusions

Overall, this paper contributes to the ongoing effort by the EOSC Task Force on Semantic Interoperability to address interoperability challenges towards the vision for a Europe-wide shared data infrastructure based on FAIR ecosystem of data and services. In particular, this work addresses the need for defining the notion of semantic artefact catalogues and identifying metrics and dimensions to assess their maturity. By analyzing current literature, a model to measure, compare, and evaluate available catalogue services for semantic artefacts has been defined, which can classify catalogues by each dimension and express the maturity of both the catalogues and dimensions as a whole.

The analysis done for the maturity model shows the current state of the catalogue, which is a requirement for guiding and shaping future developments. In particular, the analysis will be integrated with the ongoing work on minimum (meta)data sets and interoperability indicators which is currently being carried out within the EOSC Task Force on Semantic Interoperability. By combining these efforts, the aim is to provide recommendations for governance and processes for the preservation and maintenance of semantic artefacts. This will involve

identifying gaps and areas for improvement in current approaches to semantic interoperability, and developing strategies to address these challenges.

In the future, we aims at interlinking aspects of the dimensions identified within the maturity model presented in this paper with recommendations of other EOSC Task Forces. For instance, the ESOC Task Force on FAIR Metrics and Data Quality have produced some guidelines, such as [37] [59] [58], that may be used in the context of the *Quality* maturity dimension for providing even more in-depth specifications for measuring the FAIRness of semantic artefacts catalogues. Similarly, the work underdevelopment in the EOSC Task Force on PID Policy and Implementation (`https://www.eosc.eu/advisory-groups/pid-policy-implementation`) may provide additional insights related to the *PID* maturity dimension highlighted in this paper.

Acknowledgements

We thank all members of the EOSC Task Force on Semantic Interoperability[11] for the fruitful discussions, suggestions, and the joint work. The work of SP has been partially funded by the European Union's Horizon 2020 research and innovation program under grant agreement No 101017452 (OpenAIRE-Nexus) and the European Union's Horizon Europe research and innovation program under grant agreement No 101095129 (GraspOS). The work of CJ has been partially funded by the European Union's Horizon Europe research and innovation program under grant agreement No 101057344 (FAIR-IMPACT). The work of AM has been partially funded by the Data Repository Platform project of the Eötvös Loránd Research Network (ELKH ARP), with SZTAKI as its Leading Partner, project work taking place in SZTAKI DSD.

Authors' contribution

Authors' contribution according to CRediT (`https://credit.niso.org/`): Data Curation, Investigation, Methodology, Visualization (all authors); Conceptualization, Supervision (OC, SP); Validation (FE, AM, SP, ES); Writing – original draft, Writing – review & editing (FE, IH, CJ, AM, SP, ES).

References

1. ORES-2010 Ontology Repositories and Editors for the Semantic Web. CEUR (2010)
2. Alrashed, T., Paparas, D., Benjelloun, O., Sheng, Y., Noy, N.: Dataset or not? a study on the veracity of semantic markup for dataset pages. In: Hotho, E.B., Dietze, S., Fokoue, A., Ding, Y., Barnaghi, P., Haller, A., Dragoni, M., Alani, H. (eds.) A, pp. 338–356. The Semantic Web – ISWC 2021 (Vol. 12922, Springer International Publishing (2021). https://doi.org/10.1007/978-3-030-88361-4_20

[11] `https://www.eosc.eu/advisory-groups/semantic-interoperability`

3. Baclawski, K., Schneider, T.: The open ontology repository initiative: Requirements and research challenges. In: Proceedings of workshop on collaborative construction, management and linking of structured knowledge at the ISWC (**18** (2009)
4. Baumann, K., Corcho, O., Horsch, M.T., Jouneau, T., Molinaro, M., Peroni, S., Scharnhorst, A., Vancauwenbergh, S., Vogt, L.: Task Force Charter: Semantic Interoperability. Charter (Dec 2021), https://www.eosc.eu/sites/default/files/2021-12/eosca_tfsemanticinteroperability_draftcharter_20210614.pdf
5. Benjelloun, O., Chen, S., Noy, N.: Google dataset search by the numbers. In: Pan, J.Z., Tamma, V., d'Amato, C., Janowicz, K., Fu, B., Polleres, A., Seneviratne, O., Kagal, L. (eds.) The Semantic Web – ISWC 2020 (Vol, pp. 667–682. 12507, Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-62466-8_41
6. Bilder, G., Lin, J., Neylon, C.: The Principles of Open Scholarly Infrastructure. The Principles of Open Scholarly Infrastructure (2020), https://doi.org/10.24343/C34W2H
7. BMICC: MedPortal, http://medportal.bmicc.cn/, Last accessed on 2023-05-07
8. Brickley, D., Burgess, M., Noy, N.: Google dataset search: Building a search engine for datasets in an open web ecosystem. The World Wide Web Conference pp. 1365–1375 (2019). https://doi.org/10.1145/3308558.3313685
9. British Oceanographic Data Centre: The NERC Vocabulary Server, https://vocab.nerc.ac.uk, Last accessed on 2023-05-07
10. Burgelman, J.C.: Politics and Open Science: How the European Open Science Cloud Became Reality (the Untold Story). Data Intelligence **3**(1), 5–19 (Feb 2021). https://doi.org/10.1162/dint_a_00069
11. Busse, C., Corcho, O., Ekaputra, F.J., Goble, C., Heibi, I., Jonquet, C., Lange, C., Le Franc, Y., Micsik, A., Palma, R., Peroni, S., Storti, E., Widmann, H.: Raw data for the creation of a maturity model for Catalogues of Semantic Artefacts (May 2023). https://doi.org/10.5281/ZENODO.7916746
12. Carriero, V.A., Gangemi, A., Mancinelli, M.L., Marinucci, L., Nuzzolese, A.G., Presutti, V., Veninata, C.: Arco: The italian cultural heritage knowledge graph. In: The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part II 18. pp. 36–52. Springer (2019)
13. CNRS and INIST: Linked Open TERminology REsources, https://www.loterre.fr/, Last accessed on 2023-05-07
14. Codescu, M., Kuksa, E., Kutz, O., Mossakowski, T., Neuhaus, F.: Ontohub: A semantic repository engine for heterogeneous ontologies. Applied Ontology **12**(3-4), 275–298 (2017)
15. Confederation of Open Access Repositories, SPARC*: Good Practice Principles for Scholarly Communication Services (2019), https://sparcopen.org/our-work/good-practice-principles-for-scholarly-communication-services/
16. Corcho, O., Kurowski, K., Ojsteršek, M., Choirat, C., van de Sanden, M., Coppens, F.: EOSC interoperability framework. Publications Office of the European Union (2021). https://doi.org/10.2777/620649
17. Corcho, O., Ekaputra, F.J., Heibi, I., Jonquet, C., Micsik, A., Peroni, S., Storti, E.: Catalogues of Semantic Artefacts - Maturity Dimensions and Sub-Criteria (May 2023). https://doi.org/10.5281/zenodo.7916686
18. Cox, S.J.D., Gonzalez-Beltran, A.N., Magagna, B., Marinescu, M.C.: Ten simple rules for making a vocabulary fair. PLOS Computational Biology **17**, 6 (2021). https://doi.org/10.1371/journal.pcbi.1009041

19. Cyganiak, R.: Namespace lookup for rdf developers, https://prefix.cc, Last accessed on 2023-05-07
20. Ding, Y., Fensel, D.O.L.S.: The key to successful ontology reuse. SWWS pp. 93–112 (2001)
21. d'Aquin, M., Noy, N.F.: Where to publish and find ontologies? A survey of ontology libraries. Journal of Web Semantics **11**, 96–111 (2012). https://doi.org/10.1016/j.websem.2011.08.005
22. ESIP: ESIP Community Ontology Repository, http://cor.esipfed.org/, Last accessed on 2023-05-07
23. Ficarra, V., Fosci, M., Chiarelli, A., Kramer, B., Proudman, V.: Scoping the open science infrastructure landscape in europe **10**, 5281 (2020). https://doi.org/10.5281/ZENODO.4159838
24. Fraunhofer Materials and BAM: the ontology repository for materials science, https://matportal.org/, Last accessed on 2023-05-07
25. French Open Science Steering Committee: Exemplarity criteria for funding from the National Open Science Fund through platforms, infrastructures and editorial content (2019), https://www.ouvrirlascience.fr/examplarity-criteria-for-funding-from-the-national-open-science-fund/
26. Frey, J., Streitmatter, D., Götz, F., Hellmann, S., Arndt, N.: Dbpedia archivo: a web-scale interface for ontology archiving under consumer-oriented aspects. In: Semantic Systems. In the Era of Knowledge Graphs: 16th International Conference on Semantic Systems, SEMANTiCS 2020, Amsterdam, The Netherlands, September 7–10, 2020, Proceedings 16. pp. 19–35. Springer International Publishing (2020)
27. Gangemi, A., Presutti, V.: A Semantic Web portal dedicated to ontology design patterns (ODPs), http://ontologydesignpatterns.org/, Last accessed on 2023-05-07
28. Gregory, K.M., Cousijn, H., Groth, P., Scharnhorst, A., Wyatt, S.: Understanding data search as a socio-technical practice. Journal of Information Science **46**(4), 459–475 (2020). https://doi.org/10.1177/0165551519837182
29. Grosjean, J., Merabti, T., Dahamna, B., Kergourlay, I., Thirion, B., Soualmia, L.F., Darmoni, S.J.: Health multi-terminology portal: a semantic added-value for patient safety. In: Patient Safety Informatics, pp. 129–138. IOS Press (2011)
30. Hartmann, J., Palma, R., Gómez-Pérez, A.O.R.: Ontology repositories. In: Handbook on Ontologies, pp. 551–571. Springer Berlin Heidelberg (2009)
31. Hugo, W., Le Franc, Y., Coen, G., Parland-von Essen, J., Bonino, L.: D2.5 fair semantics recommendations second iteration. Tech. rep. (2020). https://doi.org/10.5281/ZENODO.5362010
32. International Virtual Observatory Alliance (IVOA): IVOA Vocabularies, https://ivoa.net/rdf/, Last accessed on 2023-05-07
33. Jonquet, C., Toulet, A., Arnaud, E., Aubin, S., Dzale-Yeumo, E., Emonet, V., Graybeal, J., Laporte, M.A., Musen, M., Pesce, V., Larmande, P.: AgroPortal: A vocabulary and ontology repository for agronomy. Computers and Electronics in Agriculture **10**(1016), 126–143 (2018). https://doi.org/10.1016/j.compag.2017.10.012
34. Jonquet, C.: Ontology Repository and Ontology-Based Services – Challenges, contributions and applications to biomedicine & agronomy. Université de Montpellier. (2019), https://theses.hal.science/tel-02133335
35. Jupp, S., Burdett, T., Leroy, C., Parkinson, H.E.: A new ontology lookup service at embl-ebi. SWAT4LS **2**, 118–119 (2015)

36. Kechagioglou, X., Vaira, L., Tomassino, P., Fiore, N., Basset, A.: Ecoportal: An environment for fair semantic resources in the ecological domain. In: Proceedings. vol. 1613, p. 0073 (2021)
37. Lacagnina, C., David, R., Nikiforova, A., Kuusniemi, M.E., Cappiello, C., Biehlmaier, O., Wright, L., Schubert, C., Bertino, A., Thiemann, H., Dennis, R.: Towards a Data Quality Framework for EOSC. Tech. rep. (Jan 2023). https://doi.org/10.5281/ZENODO.7515816
38. Lin, D., Crabtree, J., Dillo, I., Downs, R.R., Edmunds, R., Giaretta, D., De Giusti, M., L'Hours, H., Hugo, W., Jenkyns, R., Khodiyar, V., Martone, M.E., Mokrane, M., Navale, V., Petters, J., Sierman, B., Sokolova, D.V., Stockhause, M., Westbrook, J.: The TRUST Principles for digital repositories. Scientific Data **7**(1), 144 (Dec 2020). https://doi.org/10.1038/s41597-020-0486-7
39. McGuinness, D.L.: In: Fensel D, Hendler J, Lieberman H, Wahlster W (eds) Spinning the semantic web: bringing the World Wide Web to its full potential, Chapter 6. MIT Press, Cambridge, MA, pp 171–194 (2003)
40. Mossakowski, T., Kutz, O., Codescu, M..O.A.: semantic repository for heterogeneous ontologies. In: Proceedings of the Theory Day in Computer Science (DACS-2014). Satellite workshop of ICTAC-2014 (2014)
41. Murray-Rust, P.: Open Data in Science. Nature Precedings (2008). https://doi.org/10.1038/npre.2008.1526.1
42. Naskar, D., Dutta, B.: Ontology Libraries: A Study from an Ontofier and an Ontologist Perspectives. In: Proceedings of the 19th International Symposium on Electronic Theses and Dissertations (ETD 2016). Lille, France (2016), `http://docs.ndltd.org/collection/etd2016/etd16-hal-01398427-2.pdf`
43. Noy, N.F., Shah, N.H., Whetzel, P.L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D.L., Storey, M.A., Chute, C.G., et al.: Bioportal: ontologies and integrated data resources at the click of a mouse. Nucleic acids research **37**(suppl_2), W170–W173 (2009)
44. Peroni, S., Shotton, D.: The spar ontologies. In: The Semantic Web–ISWC 2018: 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part II 17. pp. 119–136. Springer (2018)
45. Phipps, J., Hillmann, D.: The rda registry: supporting rda in a multilingual world (2017)
46. Sansone, S.A., McQuilton, P., Rocca-Serra, P., Gonzalez-Beltran, A., Izzo, M., Lister, A.L., Thurston, M., Community, F.: Fairsharing as a community approach to standards, repositories and policies. Nature biotechnology **37**(4), 358–367 (2019)
47. Skinner, K., Lippincott, S.: Assessment Checklist (Commonplace). Knowledge Futures Group (2020). https://doi.org/10.21428/6ffd8432.5175bab1/00710d8a
48. Skinner, K., Lippincott, S.: Values and Principles Framework and Assessment Checklist (Commonplace). Knowledge Futures Group (2020), `https://doi.org/10.21428/6ffd8432.5175bab1`
49. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., et al.: The obo foundry: coordinated evolution of ontologies to support biomedical data integration. Nature biotechnology **25**(11), 1251–1255 (2007)
50. Technische Informationsbibliothek (TIB): Terminology service for tib, `https://service.tib.eu/ts4tib/`, Last accessed on 2023-05-07
51. The Joint Food Ontology Workgroup (JFOW): Joint-food-ontology-WG, `https://github.com/FoodOntology/joint-food-ontology-wg`, Last accessed on 2023-05-07

52. The Publications Office of the European Union: Eu vocabularies, https://op.europa.eu/en/web/eu-vocabularies/, Last accessed on 2023-05-07
53. UNESCO: UNESCO Recommendation on Open Science. Programme and meeting document SC-PCB-SPP/2021/OS/UROS (2021), https://unesdoc.unesco.org/ark:/48223/pf0000379949
54. Vandenbussche, P.Y., Atemezing, G.A., Poveda-Villalón, M., Vatant, B.: Linked open vocabularies (lov): a gateway to reusable semantic vocabularies on the web. Semantic Web **8**(3), 437–452 (2017)
55. Verbundzentrale des GBV (VZG): Basic Register of Thesauri, Ontologies & Classifications (BARTOC), https://bartoc.org/, Last accessed on 2023-05-07
56. Whetzel, P.L.N.T.: Powering semantically aware applications. Journal of biomedical semantics **4**(1), 1–10 (2013)
57. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., 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. Scientific Data **3** (Mar 2016). https://doi.org/10.1038/sdata.2016.18
58. Wilkinson, M.D., Sansone, S.A., Marjan, G., Nordling, J., Dennis, R., Hecker, D.: FAIR Assessment Tools: Towards an "Apples to Apples" Comparisons. Tech. rep. (Dec 2022). https://doi.org/10.5281/ZENODO.7463421
59. Wilkinson, M.D., Sansone, S.A., Méndez, E., David, R., Dennis, R., Hecker, D., Kleemola, M., Lacagnina, C., Nikiforova, A., Castro, L.J.: Community-driven Governance of FAIRness Assessment: An Open Issue, an Open Discussion. Tech. rep., Zenodo (Dec 2022). https://doi.org/10.5281/zenodo.7390482
60. Xiang, Z., Mungall, C., Ruttenberg, A., He, Y.: Ontobee: A linked data server and browser for ontology terms. In: ICBO (2011)
61. Zeng, M.L.: Knowledge organization systems (kos). KNOWLEDGE ORGANIZATION **35**(2-3), 160–182 (2008). https://doi.org/10.5771/0943-7444-2008-2-3-160