MDR: An Ontology Vocabulary and Registry Service for Dataset Catalogs

Manuel Fiorelli[1,2][0000-0001-7079-8941], Armando Stellato[1,2][0000-0001-5374-2807], Tiziano Lorenzetti[2][0000-0001-5676-8877], Andrea Turbati[2][0000-0002-6214-4099], Willem van Gemert[3], Denis Dechandon[3], Anikó Gerencsér[3], Enrico Francesconi[4,5] [0000-0001-8397-5820]

[1] Tor Vergata University of Rome, Department of Enterprise Engineering, Rome, Italy
{manuel.fiorelli, stellato}@uniroma2.it
[2] Lore Star srl, Rome, Italy
{tiziano.lorenzetti, andrea.turbati}@lorestar.it
[3] Publications Office of the European Union, Luxembourg
{Willem.VAN-GEMERT, Denis.DECHANDON, Aniko.GERENCSER}@publications.europa.eu
[4] European Parliament, Luxembourg
[5] Italian National Research Council, Italy
enrico.francesconi@cnr.it

Abstract. The web has establshed itself as a worldwide data hub facilitating the publication and unification of data. Nonetheless, machines still seem not ready to exploit this data for independently executing complex tasks. In pursuit of fulfilling the unachieved promise of the Semantic Web to facilitate machine functionality, we have focused on one particular aspect: ensuring a comprehensive experience in any consuming application. To this end, we have investigated how the appropriate reuse and exploitation of metadata can realize this vision. We have thus defined a metadata model combining an interpretation of existing metadata vocabularies with a new lightweight ontology concerned with dataset accessibility. Then, we have developed a metadata registry and a set of associated services that complement the proposed model in satisfying our elicited requirements. As tangible evidence of our solution's effectiveness and influence, we describe and examine the implementation of the metadata registry in three distinct, open-source applications.

Keywords: Metadata, Dataset Catalogs, Findability, Machine Actionability

1 Introduction

Initiatives such as open data and open government as well as optimism about the benefits of collaboration to scientific and societal progress have fostered the publication of data on the web. Precisely aimed at creating a web of data, the Semantic Web [1] has extended the document web with the ability to associate information with well-defined meaning, while the Linked Data [2] best practices for data publication and interlinking

have evolved the web to a global data space [3]. However, dataset metadata is today mostly used for descriptive purpose and applications are not using it to improve data consumption nor presentation, let alone collaborating for solving complex tasks; in short, the Semantic Web seems not to have delivered [4]. Concerned with keeping the still unfulfilled promise of the Semantic Web to enable machine actionability, in this work we have focused on one aspect: guaranteeing a thorough experience in any data consuming application. To this end, we have investigated how the appropriate reuse and exploitation of metadata can realize this objective. We have thus defined a metadata model (section 4) combining existing metadata vocabularies with a new lightweight ontology concerned with dataset accessibility and have developed a metadata registry (described in Section 5) and a set of associated services that complement the proposed model in achieving our objective of machine actionability. The metadata registry (MDR from now on) facilitates the adoption of our model, by providing off-the-shelf and highly reusable components and services for storing, querying and even retrieving metadata from the web, leveraging on linked data publication best practices. As a ground proof of the effectiveness and significance of our solution, we expound upon the integration of the MDR into three open-source applications. Section 6 presents an evaluation of the proposed solution while Section 7 draws the conclusions.

2 Related Work

VoID[1] [5] (Vocabulary of Interlinked Datasets) is an RDF vocabulary for describing linked datasets. VOAF[2] (Vocabulary of a Friend) extends VoID to represent vocabularies, their relationships (e.g. imports, refines, etc.) and metrics. Within the OntoLex W3C Community Group[3], some of the authors of this paper contributed an extension of VoID for linguistic metadata, called LIME [6]. DCAT[4] (Data Catalog Vocabulary) is an RDF vocabulary aiming at improving the interoperability of data catalogs. Its main contribution is the differentiation between the dataset and its accessible forms such a downloadable file or a web service and the modeling of the catalog itself. Several profiles of DCAT have been proposed to meet requirements that were not covered by the core model. One such requirement is versioning, which was addressed by ADMS[5] and HCLS Community Profile[6], this latter further differentiating between the *summary*, *version* and the *distribution levels.* The PAV [7] ontology is used to link different dataset versions and introduces version identifiers. DataID [8] combines DCAT with VoID to describe complex linked datasets, retaining the focus of VoID on discoverability.

In table x, we summarize coverage of several functional aspects that we gathered as relevant for a dataset metadata vocabulary.

---

[1] http://www.w3.org/TR/void/

[2] http://purl.org/vocommons/voaf

[3] https://www.w3.org/community/ontolex

[4] http://www.w3.org/TR/vocab-dcat/, https://www.w3.org/TR/vocab-dcat-2/, https://www.w3.org/TR/2021/WD-vocab-dcat-3-20210504/

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

[6] https://www.w3.org/TR/hcls-dataset/

Table 1. A comparison of metadata models with respect to our requirements (shortened)

	VoID	VOAF	DCAT	ADMS	HCLS	DataID
represent a catalog			✓	✓	✓	✓
versioning		version = distribution.	version = dataset	version = dataset	three levels	version = dataset
SPARQL endpoint	✓		In v2	✓	✓	✓
SPARQL limitations						
express dereferenciation						
lang. / model / statistics	✓/✗/✗	✗/✗/✗	✓/✗/✗	✓/✗/✗	✓/✗/✗	✓/✗/✗
metamodel			std, format	asset type	std, format	
cover lexical resources						
linksets	✓	✓			✓	

Blank rows correspond represent lack of the related feature, e.g. the possibility of representing a dataset catalog and not just metadata about a single dataset (missing in VoID and VoAF), the possibility to represent SPARQL language incompatibilities in accessing a certain endpoint, whether HTTP dereferenciation is offered over the hosted dataset or representation of lexical resources, which are all missing from the considered vocabularies. The combination of vocabularies adopted in our case, together with the added terms, as it will be described later in the paper, aims at filling those gaps.

Moving to dataset catalogs, LOV[7] [9] (Linked Open Vocabularies) is a curated catalog of Linked Data vocabularies mainly exploiting metadata provided by the aforementioned VOAF, supporting vocabulary retrieval by search over their title, authors, description, or any of their terms, offering versioned copies of the vocabularies and integrating a few tools for inspection and analysis of their content.

OntoPortal [10] is another catalogs offering a web frontend, REST API based on JSON-LD and a SPARQL endpoint. OntoPortal goes beyond full-text search by leveraging the actual content of the dataset to support dataset browsing. OntoPortal's features include content annotation, ontology recommendation, mapping management, DOI assignment, and FAIRness assessment. The system mainly supports (OBO and OWL) ontologies and SKOS thesauri, while support for SKOS-XL was missing until recently (as we noted in [11]), thanks to the improvements brought by AgroPortal.

3 Some Relevant Use Cases

Being concerned with how metadata can help Semantic Web applications in better finding, consuming and rendering linked open data, we identified some use cases supporting our stance. The selected use cases propose scenarios in which better machine-actionability eases tasks requiring human intervention.

These use cases were motivated by the needs emerged in the adoption of open-source Semantic Web applications that we developed in the context of EU-funded programmes supporting interoperability for the digital Europe (ISA2 and DIGITAL). These

[7] https://lov.linkeddata.es/

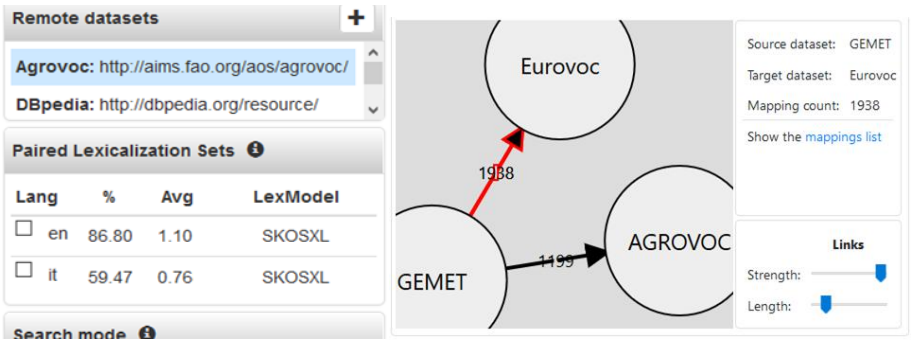

Fig. 1. Assisted search (on the left) and browsing linksets in ShowVoc (on the right)

applications include VocBench[8] [12], ShowVoc[9], MAPLE[10] (Mapping Architecture based on Linguistic Evidence) [13]. VocBench (VB henceforth) is the EU reference platform for collaborative editing of thesauri, ontologies, lexicons and RDF datasets in general. It embodies use cases related to exploration and linking of datasets published on the web. For the latter, VB leverages MAPLE, which relies on dataset metadata to instruct matching systems, possibly suggesting useful support resources. Finally, ShowVoc (or SV for short) is the data publication companion of VB, created in the context of the project PMKI (Public Multilingual Knowledge Infrastructure), combining traditional data provision following LOD policies with global activities (e.g. global search, navigation of dataset relationships, translation API). These applications are built on top of our RDF service platform Semantic Turkey[11] [14] (ST for short).

Dedicated stakeholder meetings helped lining up the roadmap for these applications and fostered the definition of these use cases. Nonetheless, the use cases elicited from them have a much broader applicability; the mentioned applications should be considered as representatives of classes of tools (editors, ontology matchers, data portals) that have a recognized role in the Semantic Web community.

In-Dataset search. Finding resources within a dataset through text search is a common functionality, which is however non-trivial when accessing an unknown dataset. VoID provides one solution (e.g. support for OpenSearch) even though a more general ground could be established on top of LOD core solutions. In **Fig. 1**, on the left, we show VB's assisted search, automating composition of a SPARQL query for finding candidate alignment matches for a resource on a remote dataset based on its labels.

Seamless, informed, navigation of local and remote datasets. While being based on a common data model (RDF), the Semantic Web offers a variety of core vocabularies (RDF, RDFS, OWL, SKOS, SKOS-XL, Ontolex) for representing content. Furthermore, distinct, standardized access modalities, such as HTTP dereferenciation and SPARQL, implement the distributed nature of the Semantic Web. Knowledge of the above should indeed be a driver for content access and visualization. In **Fig. 2** we show

[8] http://vocbench.uniroma2.it, https://ec.europa.eu/isa2/solutions/vocbench3_en
[9] http://showvoc.uniroma2.it/
[10] http://art.uniroma2.it/maple/
[11] http://semanticturkey.uniroma2.it/

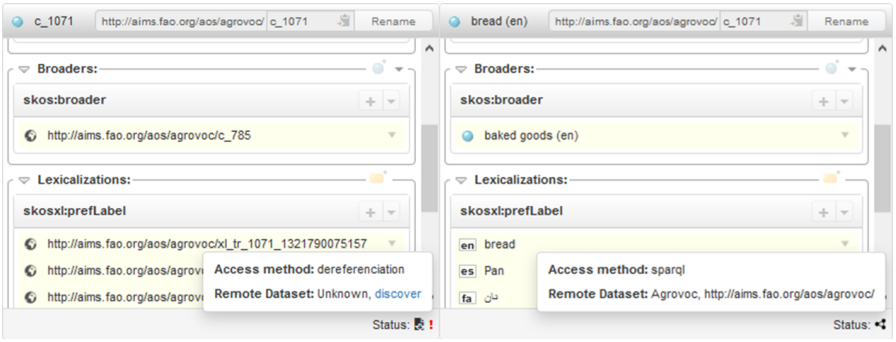

Fig. 2. Resource view on a concept from AGROVOC accessed remotely via dereferenciation (left) or using a SPARQL Endpoint (right). The latter was identified thanks to metadata.

how the exploitation of metadata properly provided by the AGROVOC [15] thesaurus dataset can drive an optimal representation of its data. On the left panel in figure, it is shown the result of an uninformed access – through HTTP dereferenciation – of VocBench to a concept of the dataset: referenced resources are not rendered (as the http dereferenciation is limited to the concise bounded description of the resource) and the SKOS-XL labels are not resolved into their literal forms. AGROVOC is accompanied by a VoID description referenced by individual resources following the suggested publication&discovery policies, which then VB (and SV alike) is able to automatically discover. This discovery prompts an informed access in which VB, aware of the available SPARQL endpoint and of the nature of the dataset queries the endpoint with a tailored query showing the resources as if they were locally hosted.

Batch Ontology Matching. Ontology Matching is a key to interoperability in presence of semantic heterogeneity. Automating this process can relieve humans in such an intensive task. However, in real world scenarios, often most of the complexity arises in setting up and fine-tuning alignment systems due to the extreme variability of the involved resources. This variability can be dealt with automatically, providing that machines are aware of it [13]. The description of a dataset should allow telling the nature of diverse datasets, e.g. ontologies, thesauri, lexicons by reporting the core vocabulary (e.g. RDFS, OWL, SKOS, OntoLex-Lemon) being used to model the content of a dataset. This knowledge can instruct a downstream matching system on how to retrieve the entities of interest and the kind of correspondences to discover. Furthermore, a matching system can benefit from some language resource to find synonyms or translations. The metadata vocabulary LIME supports representing the characteristics of such resources in terms of a common metadata model.

Dataset Cataloging. As search-engines are important gateways for the traditional Web (notwithstanding the importance of its truly distributed nature), dataset catalogs should play the same role for the Semantic Web, providing additional support for finding datasets, for traversing links between them and surfing through description of data even before diving into them. The metadata should describe linksets between datasets, allowing to extract the actual links from the dataset containing them. **Fig. 1** (right part) shows a browser of linksets between datasets within ShowVoc.

4 Metadata Model

Our metadata model (**Fig. 3**) reuses DCAT for structuring a catalog of RDF datasets and it equates void:Dataset to dcat:Distribution. The rationale is given by the different levels of abstractions intended by the two vocabularies, with DCAT representing dataset as abstractions implemented into distributions and VoID asserting ground

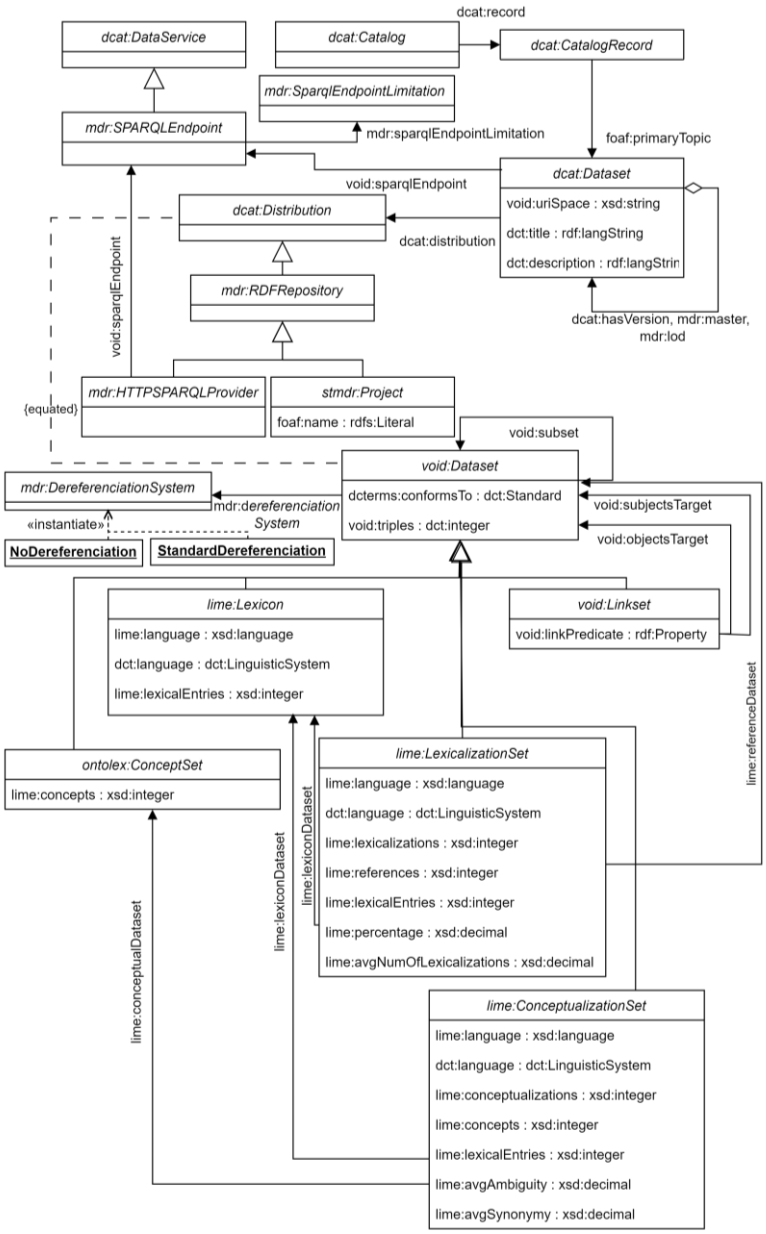

Fig. 3. The metadata model

information related to specific installments of a dataset. In addition, the model reuses VoID Dublin Core Metadata Terms for generic metadata and LIME for addressing the representation of lexical resources and, more in general, the lexical asset of any type of dataset. **Fig. 4** illustrates our model by representing metadata about the AGROVOC thesaurus. A dcat:CatalogRecord is connected to a void:Dataset representing a dataset via the property foaf:primaryTopic. The model follows DCAT 3 hierarchical approach to express versions as additional (concrete) datasets linked by a common abstract dataset through the property dcat:hasVersion. We extended this approach with the introduction of further properties mdr:lod and mdr:master, referencing respectively the evolving dataset published as linked data and the development copy (usually hosted by editing applications). As datasets are intended as abstract entities, they are associated with a dcat:Distribution representing their accessible form. The description of a distribution should include its SPARQL endpoint (void:sparqlEndpoint), while the existence of a dereferenciation system (mdr:dereferenciationSystem) is a property of a dataset, being the dataset's URI uniquely accessible

The property void:uriSpace maps URI resources to defining datasets. The modeling vocabulary of a dataset is denoted through predicate dcterms:conformsTo, representing that a dataset is a SKOS thesaurus or an OWL ontology by pointing to any of those vocabularies. As said, the model equates distributions to void:Dataset. A void:Dataset can include some subsets of particular interest, such as linksets and lexicalizations. void:Linksets represent links between two datasets, specifying the type of links and their

```
:catalog a dcat:Catalog;
   dcat:dataset <http://aims.fao.org/aos/agrovoc/void.ttl#Agrovoc> ;
   dcat:record :agrovoc_record .
:agrovoc_record a dcat:CatalogRecord;
  dcterms:issued "2019-12-08T23:05:44.439+01:00"^^xsd:dateTime;
  foaf:primaryTopic <http://aims.fao.org/aos/agrovoc/void.ttl#Agrovoc> .
<http://aims.fao.org/aos/agrovoc/void.ttl#Agrovoc> a void:Dataset;
  dcterms:title "Agrovoc";
  dcterms:conformsTo <http://www.w3.org/2004/02/skos/core> ;
  void:sparqlEndpoint <http://agrovoc.uniroma2.it/sparql>;
  void:uriSpace "http://aims.fao.org/aos/agrovoc/";
  mdr:dereferenciationSystem mdr:standardDereferenciation ;
  void:subset ex:agrovoc_it_lexicalization_set, …, ex:agrovoc_dbpedia, … .
<http://aims.fao.org/aos/agrovoc/void.ttl#ita_lex> a lime:LexicalizationSet;
  lime:avgNumOfLexicalizations 1.1945447;
  lime:language "it"^^xsd:language;
  lime:lexicalizationModel <http://www.w3.org/2008/05/skos-xl>;
  lime:lexicalizations 47341;
  lime:percentage 0.95415205;
  lime:referenceDataset <http://aims.fao.org/aos/agrovoc/void.ttl#Agrovoc>;
  lime:references 37814.
<http://aims.fao.org/aos/agrovoc/void.ttl#AGROVOC2DBPEDIA_2> a void:Linkset ;
   void:linkPredicate skos:exactMatch ;
   void:subjectsTarget <http://aims.fao.org/aos/agrovoc/void.ttl#Agrovoc>;
   void:objectsTarget <http://dbpedia.org/void/Dataset> ;
   void:triples 11001 .
```

Fig. 4. Example metadata about AGROVOC acquired from its VoID description [39]. For brevity, we showed only one lexicalization set out of 59 and one linkset out of 280.

total number. The void:uriSpace of the target dataset allows for identifying the actual links within the embedding dataset. lime:LexicalizationSet describes the lexicalization model (lime:lexicaliationModel), the natural language (lime:language), an optional lexicon (lime:lexiconDataset) (if OntoLex-Lemon is used) and relevant metrics. **Fig. 4** also illustrates one of AGROVOC's 59 lexicalization sets, stating that 95% of the concepts are covered by the 47K labels for Italian, with an average 1,19 labels per concept.

5 Implementation and Adoption of the Metadata Registry

We implemented the MDR conforming to the model described in Section 4. In its adoption within ST, the MDR acts as the central and only point to store and retrieve metadata about accessed/hosted datasets: it is consulted by other parts of ST to obtain a description of a dataset or can act as a core component for repurposing ST as a metadata portal.

The metadata module of ST (`metadata-registry`) is articulated into three submodules: `core`, `bindings` and `services`. The `core` module contains an implementation of a metadata registry that provides a Java API to create, retrieve, update or delete dataset metadata. Indeed, the `core` module is independent from ST, and the registry can be used in other Java applications. The `bindings` module integrates the metadata registry into ST, while managing a singleton of this registry, which can be consumed by other components of the system (and extensions thereof). The MDR is initialized with content originating from: i) an RDF file inside the ST data directory, containing the description of known remote datasets, ii) a settings file associated with each project managed by ST. Metadata are currently stored in different graphs inside the triple store, keeping track of their provenance. Finally, the `services` module implements a Web API that allows for integration with any kind of external application.

The MDR within ST supports different strategies to acquire dataset metadata. We should differentiate, by first, between local projects and remote datasets: the description of local projects combines information found in some configuration files with metadata generated through a profiler provided by our LIME API [16]. The options for remote datasets include manual addition of a catalog record and automatic discovery. The manual addition of a dataset prompts the user for metadata about the latest version of the dataset (top right panel in figure). The lexicalization sets can be added subsequently, optionally by lightweight profiling of a SPARQL endpoint. The discovery of dataset metadata shall be seeded by some URI identifying: i) a resource belonging to the dataset, ii) an OWL ontology published as a single document, iii) an online VoID description of the dataset. Concerning the option iii), the repository downloads the description and tries to extract relevant metadata, benefiting from our use of standard metadata vocabularies. Option ii) means that the repository can access the whole dataset, allowing for determining metadata such as the namespace and, obviously, the support for dereferenciation. Option i) is useful when the user has reached a resource belonging to an unknown dataset and operationally can activate the other options. If the resource belongs to an ontology published as a single document, then resolving the resource implicitly downloads the entire ontology, activating option ii). In other cases, the

description of the resource is scanned for standardized back-links to the VoID description of the dataset to which it belongs, possibly activating option iii).

VB offers an even smoother pattern to activate the discovery of dataset metadata from a citation of a resource. **Fig. 2** on the left shows the popup panel at the bottom of the resource view obtained by dereferencing a concept belonging to the AGROVOC thesaurus, suggesting the automatic discovery of metadata about the unknown dataset. By following a backlink to the VoID metadata of AGROVOC[12], the system finds the SPARQL endpoint of the dataset, enabling a richer resource view as discussed earlier.

6 Evaluation

We consider evaluation over a number of aspects, which we detail below:

Design & Technical quality. The proposed model combines (specific adoptions and interpretations of) widely accepted metadata vocabularies with an ontology developed by us. Following best practices, users looking up at the address of the ontology can negotiate a machine-readable or a human-friendly representation. We included, in turn in our ontology vocabulary, general metadata about the ontology required by LOV.

Availability of Both Vocabularies and Software and Reusability. The source code of ST, VB, MAPLE and SV is publicly available through Git repositories (https://bitbucket.org/art-uniroma2/, repos: semantic-turkey, vocbench3, maple, showvoc). The metadata model was published as Linked Open Data (http://semanticturkey.uniroma2.it/ns/mdr, http://semanticturkey.uniroma2.it/ns/stmdr) and the main ontology was deposited on Zenodo (https://zenodo.org/doi/10.5281/zenodo.6522179) and Linked Open Vocabularies (https://lov.linkeddata.es/dataset/lov/vocabs/mdr). The implementation can be reused (see Section 5): i) as a library in Java applications, ii) within ST extensions, iii) through the Web API of ST. Furthermore, we should mention that VB and SV are reusable applications on their own that use the metadata registry of ST.

Innovation. Our solution builds up on some of the existing works, broadening the use of existing properties from common metadata vocabularies (e.g. the above mentioned dcterms:conformsTo), combining them with LIME for richer linguistic description, a specifically developed ontology describing access mechanisms, and an extension of DCAT to address identity of datasets over time, different versions (intended as an abstract set of triples) and different accesses.

While neglecting versioning, VoID suggests to use dcterms:modified, implying that the content of a dataset may change over time. DCAT-2 addresses versioning in a non-normative section, representing versions as distinct datasets, not really enforcing a representation that should be clearly and universally understood by consumers. ADMS and DataID also treat versions as distinct datasets but prescribe specific relation URIs. The dataset per se, irrespectively of its actual versions, is not represented explicitly (and thus cannot be referred to). This avoids the need for distributing metadata between the dataset per se and its versions, which may be problematic, as it can be seen for HCLS.

[12] http://aims.fao.org/aos/agrovoc/void.ttl#Agrovoc

Firstly, the SPARQL endpoint (void:sparqlEndpoint) is centralized on the summary level, even if it makes sense for versions as well (i.e. a SPARQL endpoint providing an "archived" version of the dataset). Moreover, the data dump (void:dataDump) is decentralized to the distribution level (i.e. comparable to dcat:downloadURL), but void:dataDump (as well as other VoID properties describing different access mechanisms) should instead be considered as a shorthand for a distribution on its own and thus be linked to a version. FRBR [17] (Functional Requirements for Bibliographic Record) aimed at unifying the different practices for bibliographic archiving through a conceptual model, which is also relevant to archiving and cataloguing datasets on the web. FRBR defines the notions of work (i.e. a "distinct intellectual or artistic creation"), expression (i.e. the "intellectual or artistic realization of a work"), manifestation (i.e. the "physical embodiment of an expression of a work") and item (i.e. "single exemplar of a manifestation"). Our approach of grouping concrete datasets under a common denominator has something of FRBR: one dataset embodies something of FRBR's Work (the abstract dataset), and other versions/distributions might be connected to it, still being collected under a same umbrella. This would better allow, for instance, to say that there is a record about the "AGROVOC" dataset (dataset intended as FRBR's Work), pointing through the mdr:lod to the always evolving version of Agrovoc as a linked open dataset, a master production copy being hosted on a stmdr:Project (a particular type of distribution) and then different other versions of the dataset (e.g. past releases) linked to it. This does not preclude specific links between the versions (e.g. prior/successor) and better represents the collecting entity and the "primary" reference.

All models build on void:sparqlEndpoint to hold the address of a SPARQL endpoint, which is considered a dcat:DataService by DCAT. No other model addresses the limitations of SPARQL endpoints and the lack of standard dereferenciation. All of the discussed models use the Dublin Core property *language*, to report a language tag or a language URI. Reusing LIME, our model also represents the adopted lexicalization model and summarizing statistics about the available linguistic content (much in the same spirit of the structural metrics defined by VoID and VOAF).

Scalability. Scalability towards very large dataset catalogs is, under all practicaly considerations, not a concern. For how rich our model can be, it is still metadata and the size of even the largest dataset catalog (hosting dataset metadata) would not be considered a large dataset per se. The Publications Office (OP) of the European Union is hosting a centralized (serving all the European Commission) installation[13] of SV and of the MDR, sporting hundreds of datasets and is 1) performing well (i.e. no tangible delay for the user) and 2) has no particularly high hardware requirements. SV and the MDR have no relevant memory or processor use footprint, while a 64Gb virtual machine is being used for the triple store: this can however be considered an exogenous aspect wrt the MDR per se; additionally, the considering use of SV implies managing the dataset's data and not just their metadata, which is the real resource consuming aspect[14].

[13] https://showvoc.op.europa.eu/

[14] datasets are subjected to persistent storage in a triple store; however access optimization, caching and other services take a toll in the in-memory footprint. Additionally, for purposes of

Influence and Sustainability. In [12], we showed the impact of VB and the exploitation of its metadata registry. Recommended by the ISA² Programme (https://ec.europa.eu/isa2/) to EU public administrations for "centralizing the management of controlled vocabularies and metadata", VB is actually being broadly adopted world-wide. Besides the already mentioned instance of the European Union, FAO uses SV for its statistics open data portal Caliper[15] while an instance for the Statistics Division of the United Nations (UNSD) is being made operative. LifeWatch ERIC[16] uses SV as a data hosting and publication service, complementing its catalog – EcoPortal – of semantic resources. The MDR is being used in all the above cases to hold metadata about the datasets being aligned. MAPLE orchestrates ontology matching, and our remote alignment service API [18] interact with the connected matching systems. The DIGITAL programme of the EU (https://digital-strategy.ec.europa.eu/en/activities/digital-programme), prosecuting the above-mentioned ISA², is guaranteeing funding and progress to the project, which is in any case maintained by the company Lore Star (https://lorestar.it/) on a open-source and pay-per-support business model.

7 Conclusion and Future Work

In this work, we tackled the difficulties of data discovery, understanding and exploitation by focusing on exploitation of metadata about the datasets and Linked Open Data best practices. We thus focused on use cases provided by a semantic web platform, ST, which was used in several contexts, including a collaborative knowledge development (VB), batch ontology matching (MAPLE), open data portals (ShowVoc) and other application scenarios. We defined a metadata model addressing these use cases, by combining and extending existing metadata vocabularies. We then implemented a metadata registry driven by the defined model and evaluated it against the presented use-cases.

References

1. Berners-Lee, T., Hendler, J.A., Lassila, O.: The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American 284(5), 34-43 (2001). doi: 10.1038/scientificamerican0501-34
2. Berners-Lee, T.: Linked Data. In: Design Issues (2006). https://www.w3.org/DesignIssues/LinkedData.html. Accessed 9 November 2017
3. Heath, T., Bizer, C.: Linked data: Evolving the web into a global data space. Synthesis Lectures on the Semantic Web: Theory and Technology 1(1), 1-136 (2011). doi: 10.2200/S00334ED1V01Y201102WBE001

isolation, ShowVoc stores each dataset in an independent repository, thus having a (contained) linear increase in resource-consumption for each new dataset. This, again, applies to applications serving data other than metadata, while the resource consumption for the MDR service is minimal, as it requires a single repository.

[15] https://stats-class.fao.uniroma2.it/caliper/

[16] e-Science Infrastructure for Biodiversity and Ecosystem Research: https://www.lifewatch.eu/

4. Shadbolt, N., Berners-Lee, T., Hall, W.: The Semantic Web Revisited. IEEE Intelligent Systems 21(3), 96-101 (2006). doi: 10.1109/MIS.2006.62
5. Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: Describing Linked Datasets. In : In Linked Data on the Web Workshop (LDOW 09) colocated with WWW 09 (2009)
6. Fiorelli, M., Stellato, A., Mccrae, J.P., Cimiano, P., Pazienza, M.T.: LIME: the Metadata Module for OntoLex. In : The Semantic Web. Latest Advances and New Domains (Lecture Notes in Computer Science) vol. 9088. Springer, Cham (2015), pp.321-336. doi: 10.1007/978-3-319-18818-8_20
7. Ciccarese, P., Soiland-Reyes, S., Belhajjame, K., Gray, A.J., Goble, C., Clark, T.: PAV ontology: provenance, authoring and versioning. J. Biomed. Semantics 4(37) (2013). doi: 10.1186/2041-1480-4-37
8. Brümmer, M., Baron, C., Ermilov, I., Freudenberg, M., Kontokostas, D., Hellmann, S.: DataID: Towards Semantically Rich Metadata for Complex Datasets. In : Proc of the 10th Int. Conf. on Semantic Systems, Leipzig, Germany, 4-5 September 2014. (2014), pp.84-91. doi: 10.1145/2660517.2660538
9. 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 (December 2016). doi: 10.3233/SW-160213
10. Jonquet, C. et al.: Ontology Repositories and Semantic Artefact Catalogues with the OntoPortal Technology. In : The Semantic Web – ISWC 2023 (Lecture Notes in Computer Science). 22nd International Semantic Web Conference, Athens, Greece, November 6–10, 2023, Proceedings, Part II. vol. 14266. Springer, Cham (2023), pp.38-58. doi: 10.1007/978-3-031-47243-5_3
11. Fiorelli, M., Stellato, A., Rosati, I., Fiore, N.: Process-Level Integration for Linked Open Data Development Workflows: A Case Study. In : Metadata and Semantic Research (Communications in Computer and Information Science). 16th Research Conference, MTSR 2022, London, UK, November 7–11, 2022, Revised Selected Papers. vol. 1789. Springer, Cham (2023), pp.148-159. doi: 10.1007/978-3-031-39141-5_13
12. Stellato, A. et al.: VocBench 3: a Collaborative Semantic Web Editor for Ontologies, Thesauri and Lexicons. Semantic Web 11(5), 855-881 (2020). doi: 10.3233/SW-200370 doi: 10.3233/SW-200370.
13. Fiorelli, M. et al.: Metadata-driven Semantic Coordination. In : Metadata and Semantic Research (Communications in Computer and Information Science) vol. 1057. Springer, Cham (2019). doi: 10.1007/978-3-030-36599-8_2
14. Pazienza, M.T., Scarpato, N., Stellato, A., Turbati, A.: Semantic Turkey: A Browser-Integrated Environment for Knowledge Acquisition and Management. Semantic Web 3(3), 279-292 (2012). doi: 10.3233/SW-2011-0033
15. Caracciolo, C. et al.: The AGROVOC Linked Dataset. Semantic Web 4(3), 341–348 (2013). doi: 10.3233/SW-130106
16. Fiorelli, M., Pazienza, M.T., Stellato, A.: An API for OntoLex LIME datasets. In : OntoLex-2017 1st Workshop on the OntoLex Model (co-located with LDK-2017), Galway (2017)
17. IFLA: Functional Requirements for Bibliographic Records 19. (1998)
18. Fiorelli, M., Stellato, A.: A Lime-Flavored REST API for Alignment Services. In : Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020). European Language Resources Association, Marseille, France (2020), pp.52–60