Metadatamatic: A Web application to Create a Dataset Description

Pierre Maillot, Olivier Corby, Catherine Faron, Fabien Gandon, Franck Michel

Metadatamatic: A Web application to Create a Dataset Description

Pierre Maillot
Univ. Cote d'Azur, Inria, CNRS, I3S
Sophia Antpolis, France
pierre.maillot@inria.fr

Olivier Corby
Univ. Cote d'Azur, Inria, CNRS, I3S
Sophia Antpolis, France
olivier.corby@inria.fr

Catherine Faron
Univ. Cote d'Azur, Inria, CNRS, I3S
Sophia Antpolis, France
faron@i3s.unice.fr

Fabien Gandon
Univ. Cote d'Azur, Inria, CNRS, I3S
Sophia Antpolis, France
fabien.gandon@inria.fr

Franck Michel
Univ. Cote d'Azur, CNRS, Inria, I3S
Sophia Antpolis, France
franck.michel@inria.fr

ABSTRACT

This article introduces Metadatamatic, an open-source, online, user-friendly tool for generating the description of a knowledge base. It supports the description of any RDF dataset via a user-friendly web form that does not require prior knowledge of the vocabularies begin used, and can enrich the description with automatically generated statistics if the dataset is accessible from a public SPARQL endpoint. We discuss the models and methods behind the tool, and present some initial results suggesting that Metadatamatic can help in increasing the visibility of public knowledge bases.

CCS CONCEPTS

• **Information systems** → **Information integration**; **Resource Description Framework (RDF)**; *Web applications.*

KEYWORDS

metadata, provenance, datasets, linked data, RDF, semantic web

ACM Reference Format:
Pierre Maillot, Olivier Corby, Catherine Faron, Fabien Gandon, and Franck Michel. 2023. Metadatamatic: A Web application to Create a Dataset Description. In *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion), April 30-May 4, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3543873.3587328

1 INTRODUCTION

The deployment of semantic web technologies [1] is driven by the desire, among others, to improve the findability, accessibility, interoperability, reusability, and reliability of data. In multiple domains, a growing number of knowledge bases (KBs) are made publicly available either through public SPARQL endpoints or downloadable RDF dumps. For instance, DBpedia is generated by turning information from Wikipedia pages into RDF.

To ensure the FAIR properties, KGs should provide rich metadata typically encompassing authors, licensing, provenance information, access interfaces, and a succinct overview of their content. There are several vocabularies available for writing such descriptions in RDF, with VoID [3], DCAT [9] and SPARQL-SD [8] being recommended by the W3C. However, during an experiment on a large number of publicly accessible KBs, we observed that less than 10% of them contain some form of provenance information [5], which is detrimental to their traceability and reusability.

In practice, the task of writing a KB description requires a solid understanding of multiple vocabularies in addition to VoID, DCAT, and SPARQL-SD, depending on the domain of interest. To alleviate this burden, this article presents a tool that supports the generation of a KG description without requiring prior knowledge of the vocabularies.

2 MOTIVATION

In a previous study [5], we attempted to retrieve the descriptions of 339 active KBs whose SPARQL endpoints were collected from public catalogues such as the LOD-Cloud, LinkedWiki and Wikidata. The results of this experimentation are represented graphically in KartoGraphi [4]. These results revealed that only 33 out of 339 KBs contained any form of description. Among those with a description, 26 KBs contained information about their creators, contributors or publishers, 18 contained temporal information about their creation or publication, and 16 contained information about their licenses. We also attempted to retrieve the description hosted as `.well-known/void` files, as recommended in the VoID vocabulary, but found only 10 such files for the 339 KBs.

We attribute the lack of self-description in publicly available KBs to the lack of incentive and the difficulty of writing and maintaining these descriptions, starting with identifying which of the many vocabularies are appropriate. We assessed at least 11 vocabularies that are classically used for the general description of a KB (listed in Table 1), yet some authors found up to 32 possible vocabularies for that task [2]. These vocabularies are both complementary and overlapping. More domain-specific vocabularies can also be used, such as the extension of DCAT for geographic datasets GeoDCAT [7]. Many articles have recommended "description profiles" in various use-cases and using various vocabularies, as shown in the survey by Benelli et al. [2].

While many parts of a description, such as statistics, can be extracted automatically, it remains necessary for KB providers to manually enter important parts of the description such as licensing

WWW '23 Companion, April 30-May 4, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9419-2/23/04.
https://doi.org/10.1145/3543873.3587328

and authoring information. Besides, identifying the appropriate classes and properties from the various vocabularies can be challenging. Initiatives such as DCAT-AP [6] or the COST Action for Distributed Knowledge Graphs have proposed lists of classes and properties for each description feature, but effectively writing the description remains a burdensome task.

The aim of Metadatamatic is to tackle the manual description challenge by providing a user-friendly method for creating a KB description via a web form, enriched with statistics automatically generated from the KB, and represented in RDF using the vocabularies recommended by the community.

3 PRINCIPLES

Describing a KB is not a straightforward task, as several elements must be considered. For example, one must consider whether each named graph is a dataset of its own, or all the named graphs are part of the same dataset. One must also consider the level of detail required in the description, such as a precise description of each step of the process that generated the KB. To guide those decisions, we have studied existing vocabularies and proposals, and selected a KB description model that is as close as possible to the consensus that is currently emerging from the various initiatives. Moreover, the need for exhaustiveness can make the description process tedious, particularly in the case of large and complex KBs. To make the process simpler, a description tool must be able to extract as much information from the KB as possible. The following subsections explain how we support these principles.

3.1 A comprehensive description model

The survey [2] indicates that there are numerous vocabularies that can be used to represent the same features of a KB description. Using the results of our experiments with KartoGraphi [4], we studied the vocabularies used in the descriptions that we were able to gather, as well as the elements proposed in VoID, DCAT, DCAT-AP and DataID. From this study, we carefully selected 11 candidate vocabularies, listed in Table 1, that we could potentially used in the descriptions generated by Metadatamatic.

We retained three widely used and specialised vocabularies: VoID, DCAT, and SPARQL-SD, in addition to Dublin Core Terms for general properties such as the title and description of a KB. Other vocabularies were not used for various reasons. Some vocabularies are somehow restricted to specific communities, such as DataID. In our experiments, some well-known vocabularies like Schema.org and SKOS showed limited usage in this specific context of describing KBs. PROV-O and PAV are used to represent provenance information, yet different patterns are used to account for this information. As a consequence, a static web form would need to opt for a specific pattern and, arguably, would lack the flexibility required to be really helpful.

3.2 Automated description extraction

Metadatamatic converts the content of the web form into RDF, and utilises the previously proposed IndeGx framework [5] to extract KB descriptions from their content. We reused the SPARQL queries from this framework to extract various data from the content of the KB: list of vocabularies, language tags and named graphs, as well as the numbers of classes, properties and triples. Figure 1 shows the number of KBs that returned a result when IndeGx tried to retrieve the different description features from 339 endpoints. In particular, we can see that less than 15% of the endpoints managed to answer to the queries designed to extract the languages and the number of triples. This can be explained by the fact that those queries require to process respectively all the literals and all the triples of a KB, and therefore frequently time out.

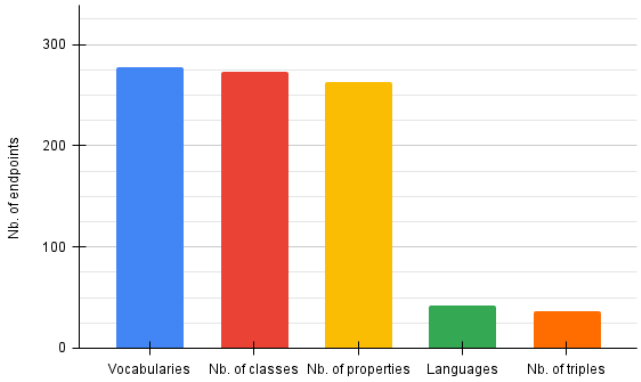

Figure 1: Number of endpoints, out of a total of 339, for which IndeGx was able to process a given characteristic.

Although some of these queries do not succeed, this automated extraction greatly reduces the effort for the user, particularly when dealing with KBs that contain a large number of named graphs and vocabularies. For example, the Scholarly Data base [10] contains more than thirty named graphs.

Some features such as the KB license and languages have a limited set of most frequent values. To make the form more user-friendly, Metadatamatic suggests such frequent values with auto-completion. The values suggested for the license come from the RDFLicense dataset, while the values suggested for the languages come from the LexVo dataset.

3.3 Description augmentation with aligned vocabularies

As mentioned above, multiple vocabularies can be used in a KB description and some of them have overlaps. Although using all available vocabularies simultaneously would make the description more tedious, in particular the part that is automatically generated, it would make the information more easily accessible and findable since users would not need to figure out which vocabularies were used. Therefore, for certain features, Metadatamatic can add equivalent triples using properties from other recognized vocabularies. Table 2 show the properties and types used in the description initially generated, and lists equivalent properties and types added once the use hits the *Apply equivalences* button.

Vocabulary	Short description	Vocabulary	Short description
DCAT	Description of a catalog of KBs	*Schema.org*	General description
DCTerms	General description	SKOS	Concept description
DataID	Kb description, extends DCAT	**SPARQL-SD**	Endpoint description
FOAF	Description of persons	VCard	People and organisations description
PAV	Provenance description, extends PROV-O	**VoID**	KB description
PROV-O	Provenance description		

Table 1: Most well-known vocabularies usable to describe a KB. The vocabularies in bold are those used in the description generated by Metadatamatic. Those is italic are those added after augmentation with aligned vocabularies.

Description feature	Initial property or type	Properties/types added after augmentation
Dataset rdf:type	dcat:Dataset	dcmitype:Dataset, void:Dataset, sd:Dataset, schema:Dataset, prov:Entity
Title	dct:title	schema:name
Description	dct:description	schema:description
Creator	dct:creator	prov:wasAttributedTo, pav:authoredBy, pav:createdBy
Publication date	dct:issued	schema:datePublished
Version	dcat:version	schema:version, dct:hasVersion, pav:version
Language	dct:language	schema:inLanguage
DataService rdf:type	dcat:DataService	sd:Service
DataService endpoint	dcat:endpointURL	sd:endpoint

Table 2: Augmentation of the KB description with aligned vocabularies.

4 OPEN SOURCE AND ONLINE APPLICATION

Metadatamatic is accessible online[1] and its code is available on GitHub under the Apache 2.0 open source license. Figure 2 shows the web interface and main sections of Metadatamatic. As can be seen, the form consists of multiple sections, each dedicated to a feature of the KB description: title, description, endpoint URL, etc. For each description feature, one or multiple values can be added depending on the feature, e.g. there can be multiple titles but only one publication date. The translation of each feature to RDF is displayed in Turtle beneath the feature, and the complete description is displayed at the bottom of the page and can be downloaded.

The user may also import a KB description which Metadatamatic uses to populate the form, thus making it possible to improve an existing description.

Let us now consider the example of the description of the French chapter of DBpedia. Metadatamatic offers the possibility to enter different titles and descriptions in different languages. They use respectively the `dct:title` and `dct:description` properties. Entering a SPARQL endpoint URL allows Metadatamatic to extract features directly from the endpoint. The endpoint is also added to the KB description with property `void:sparqlEndpoint`, and entails the creation of a resource of type `dcat:DataService`. In the license section, a list of suggestions is proposed by name and URI of well-known licenses. Upon selection, the URI of the license is entered as value of the `dct:license` property. The vocabularies, languages, triples, classes and properties features are extracted from the endpoint, they are respectively described using `dct:language`, `void:triples`, `void:vocabulary`, `void:triples`, `void:classes`, `void:properties`. The named graphs are also extracted atomatically and described using a data structure from SPARQL-SD. At the time we did the extraction, the form yielded a description that consists of over 160 triples, whereas only 5 values were entered manually. The result of our description of the French chapter of DBpedia appears in Turtle as shown in Listing 1.

[1]https://wimmics.github.io/voidmatic/

Listing 1: Extract of the description of the French chapter of DBpedia, in Turtle format.

```
ex:dataset1 a dcat:Dataset;
    dct:creator wimmics:, culturegouvfr:, inria:;
    dct:description "DBpedia Fr is the french chapter of DBpedia, it
      is part of the DBpedia internationalization effort whose
      purpose is to maintain structured data extracted (...)"@en;
    dct:license cc4:;
    dct:title "DBpedia French chapter"@en;
    void:sparqlEndpoint fr:sparql;
    void:classes 806;
    void:properties 27461;
    void:triples 20572176;
    void:vocabulary bibo:, cc:, dc:, dcat:, dct:, foaf:, geo:,
        georss:, goodrelations:, oa:, openvocab:, owl:, powders:,
        pro:, prov:, schema:, rdf:, rdfs:, skos:, sd:, vann:, voaf:,
        vcard:, void:, wde:;
    sd:namedGraph
        [a sd:NamedGraph; sd:name frgraph:metadata ],
        [a sd:NamedGraph; sd:name frgraph:process_tags ],
        [a sd:NamedGraph; sd:name frgraph:statistics ],
        ...
ex:dataset1-service
    a dcat:DataService;
    dcat:endpointURL fr:sparql;
    dcat:servesDataset ex:dataset1.
```

5 RESULTS

Since the launch of Metadatamatic on 12/08/2022, we started to collect descriptions of datasets that were not in the public catalogues, such as Wikidata, LinkedWiki, and the LOD-Cloud. Among them, the following datasets were never publicly described before:

Figure 2: Example of Metadatamatic's web interface with a subset of the sections

- The knowledge graph SemJoconde describing the French museums: in addition to the description features manually entered, Metadatamatic computed that it contains 16459098 triples, 24 classes, 474 properties and 101 named graphs.
- The UNESCO vocabularies endpoint created by the UNESCO under the Creative Commons 4.0 license: in addition to the description features manually entered, Metadatamatic automatically retrieved the 5 different languages of its literals.
- The SMS ontology, by the Smart Morphing and Sensing consortium: in addition to the description features manually entered, Metadatamatic computed the number of triples in the dataset.

Listing 2 presents the RDF description of SemJoconde generated by Metadatamatic.

Listing 2: Extract of the description of the SemJoconde dataset, in Turtle format.

```
ex:dataset2 a dcat:Dataset, sd:Dataset;
    dct:title "SemJoconde"@en;
    dct:description "Representation semantique, basee
        sur CIDOC-CRM, de la base Joconde"@fr;
    dct:creator "Jean-Claude Moissinac";
    dct:issued "2018-02-12"^^xsd:dateTime;
    dcat:keyword "artworks", "cultural heritage",
        "creators", "joconde database";
    dct:language "fr", "en";
    void:sparqlEndpoint semjoconde:query;
    void:classes 23;
    void:properties 470;
    void:triples 16459098;
    void:vocabulary <http://erlangen-crm.org/current/>,
                    <http://datamusee.givingsense.eu/onto/>;
    sd:namedGraph semjograph:dataid, semjograph:dmcreators.
semjograph:dataid a sd:NamedGraph; sd:name semjograph:dataid.
semjograph:dmcreators a sd:NamedGraph; sd:name semjograph:dmcreators.
```

6 CONCLUSION

Metadatamatic was designed in the context of the DeKaLoG and D2KAB ANR projects. This service was motivated by the lack of dataset descriptions observed in the IndeGx framework experimentation and its visualization in the KartoGraphI website. We have shown how this open-source and online tool assists users in generating a description of their datasets without requiring prior knowledge of any vocabularies, and how it offers automated extraction and augmentation features to simplify the description process. We also showed that, as soon as it was released, this tool allowed us to make visible datasets that were unknown to major linked datasets indexes so far. Future works include the generation of FAIRness and accountability indicators to further improve the metadata available for linked open data sources.

ACKNOWLEDGMENTS

This work is supported by the ANR DeKaloG (Decentralized Knowledge Graphs) project, ANR-19-CE23-0014, and by the ANR D2KAB (Data to Knowledge in Agriculture and Biodiversity) project, ANR-18-CE23-0017, both projects from CE23 - Intelligence artificielle and by the 3IA Côte d'Azur ANR-19-P3IA-0002.

REFERENCES

[1] Dean Allemang, James A Hendler, and Fabien Gandon. 2020. *Semantic web for the working ontologist.* ACM Press.
[2] Mohamed Ben Ellefi, Zohra Bellahsene, John Breslin, Elena Demidova, Stefan Dietze, Julian Szymanski, and Konstantin Todorov. 2017. RDF Dataset Profiling - a Survey of Features, Methods, Vocabularies and Applications. *Semantic Web* 9 (08 2017). https://doi.org/10.3233/SW-180294
[3] Richard Cyganiak, Jun Zhao, Michael Hausenblas, and Keith Alexander. 2011. *Describing Linked Datasets with the VoID Vocabulary.* W3C Note. W3C. https://www.w3.org/TR/void/
[4] Pierre Maillot, Olivier Corby, Catherine Faron, Fabien Gandon, and Franck Michel. 2022. KartoGraphI: Drawing a Map of Linked Data. In *Extended Semantic Web Conference.* Springer.
[5] Pierre Maillot, Olivier Corby, Catherine Faron, Fabien Gandon, and Franck Michel. 2023. IndeGx: A Model and a Framework for Indexing RDF Knowledge Graphs with SPARQL-based Test Suits. *Journal of Web Semantics* (Jan. 2023), 100775. https://doi.org/10.1016/j.websem.2023.100775
[6] Interoperable Europe Programme of the European Commission. 2022. *DCAT Application Profile for data portals in Europe.* https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe/release/211
[7] Andrea Perego, Vlado Cetl, Anders Friis-Christensen, and Michael Lutz. 2017. GeoDCAT-AP: Representing geographic metadata by using the "DCAT application profile for data portals in Europe". In *Joint UNECE/UNGGIM Europe Workshop on Integrating Geospatial and Statistical Standards, Stockholm, Sweden.*
[8] Gregory Williams. 2013. *SPARQL 1.1 Service Description.* W3C Recommendation. W3C. https://www.w3.org/TR/2013/REC-sparql11-service-description-20130321/.
[9] Peter Winstanley, David Browning, Riccardo Albertoni, Alejandra Gonzalez Beltran, Simon Cox, and Andrea Perego. 2020. *Data Catalog Vocabulary (DCAT) - Version 2.* W3C Recommendation. W3C. https://www.w3.org/TR/2020/REC-vocab-dcat-2-20200204/.
[10] Ziqi Zhang, Andrea Giovanni Nuzzolese, and Anna Lisa Gentile. 2017. Entity Deduplication on ScholarlyData. In *Proceedings of ESWC 2017 (Lecture Notes in Computer Science).* Springer.