International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Ontology-based Knowledge Representation for Open Government Data

Younsi Dahbi Kawtar*[1], Lamharhar Hind[1], Chiadmi Dalila[1]

Submitted: 05/09/2022 **Accepted:** 15/12/2022

***Abstract**:* Open Government data present valuable knowledge that supports political, social, and economic value creation. The number of available OGD datasets has increased significantly with pressure being put on government administrations to open up their data. This makes the task of discovering relevant datasets difficult and time-consuming. Improving data discoverability has a significant impact in improving OGD usage. As improving discoverability relies on improving metadata representation, we explore in this paper the usage of ontologies as a knowledge representation formalism to provide a rich and semantically enhanced representation of OGD metadata that enables its processability and interpretability by machines. Our knowledge representation model covers the essential kinds of knowledge necessary for the description of OGD datasets (descriptive, technical, contextual and structural). Based on the semantic model, we propose an approach to transform OGD metadata in a semantically enhanced RDF graph. This graph will serve as a ground basis to implement a semantic search mechanism to improve data discoverability on OGD portals.

Keywords*:* Open government data, ontologies, knowledge representation, discoverability

1. Introduction

Open Government Data (OGD) is a concept that continues to prosper and evolve. It includes all data collected of produced by public administrations that is made available for the public to be freely used [1]. The publication of government data has a significant impact that could be identified through multiple aspects: fostering innovation, improving transparency, public accountability, and collaboration, and improving citizens' quality of life [2] [3]. The established method of publishing OGD is via OGD portals. These portals offer an ecosystem that enables public administrations (data providers) to publish and manage datasets (data cataloging, metadata management). However, they suffer from the lack of tools and mechanisms to support data discoverability as they provide only basic and limited search features (keywords matching, faceted search) [4]. Also, the metadata gathered by OGD portals is often represented in the format of simple key -value pairs that goes without the semantic that explicitly represents their meanings [5]. This metadata representation hinders the development of advanced search mechanisms to improve discoverability such as semantic search. Therefore, to improve the discoverability and usage of OGD datasets, metadata must be presented in a format that allows its machine processability and interpretability [10] . In the context of the Semantic Web, several vocabularies /ontologies [6] [7] [8] [9] [10] are proposed to support the representation of metadata in a semantically rich and machine-processable format. Our approach builds upon these propositions and explores the potential of ontologies [11] to give a semantically enhanced and rich description of OGD datasets.

Ontologies represent a powerful knowledge representation formalism for semantic modeling and enrichment. They present a shared understanding of a domain of interest and capture its semantic content in a way that can be machine-processable. They give a high level of expressiveness and allow reasoning and knowledge inference.

Based on ontologies, we propose a knowledge representation model aims to semantically enrich OGD metadata and covers different types of necessary knowledge for the OGD dataset’s description and semantic enrichment.

This model is structured around four main ontologies: The GovDS ontology is motivated by the need to provide four essential types of knowledge about an OGD dataset (descriptive, technical, structural, and contextual), the GovOrganization ontology which provides the necessary knowledge to represent the organizational structures and relations between the public administrations, the GovGeoEntity ontology grants the necessary knowledge to represent government administrative regions and their relations following the territory administrative subdivision, and the GovDomain ontology which is a domain ontology that guarantees a standard model across organizations to describe government concepts and relations. The model is used to enrich both data and metadata and present them in a machine-processable way. The rest of this paper is organized as follows: In section 2, we present related works, in section 3 we present our knowledge representation model, in section 4, we present its applications for the semantic enrichment of the dataset's metadata, and we finally conclude with perspectives.

2. Related works

In the context of the semantic web, several vocabularies /ontologies are proposed and used to support the publication of datasets metadata in a machine-processable format to represent several kinds of knowledge (descriptive, technical, structural) about datasets. We present some of them below.

The Dublin Core [6] is a vocabulary developed for the semantic description of digital resources. Its objective is to provide a base of

1 Mohammed V University in Rabat, Morocco
** Corresponding Author Email: younsidahbi.kawtar@email.com*

descriptive elements sufficiently structured to allow interoperability of digital resources on the Web. Since its publication in 1995, several governments and international organizations have chosen the Dublin core as a metadata repository for their administrations.

The DCAT vocabulary [7] is a W3C recommendation designed to facilitate interoperability between datasets published on the web. The goal of DCAT is to improve the discovery of datasets and enable applications to consume metadata from multiple sources quickly. It was initially developed in the context of government data catalogs.

The DCAT vocabulary allows data providers to describe datasets using the RDF model. To extend the DCAT, several application profiles have been proposed. The DCAT Application Profile for Data Portals in Europe (DCAT-AP) is a specialization of DCAT for describing public sector data sets in Europe. It defines a minimum set of properties to include in the dataset profile by specifying required and optional properties. The DCAT-AP is today the expected standard for publishing metadata in European portals. The StatDCAT Application Profile [12] extends the DCAT Application Profile for statistical data portals in Europe. Its objective is to improve the visibility and discovery of statistical datasets. StatDCAT-AP adds a set of properties particularly relevant to statistical datasets such as attributes, dimensions, statistical serial numbers, and statistical units of measure.

GeoDCAT-AP [13] is a metadata profile specifically designed to enable sharing of geospatial metadata. It aims to provide an RDF representation of geospatial metadata compliant with the DCAT Application Profile for European Data Portals (DCAT-AP).

The VoID Vocabulary [8] is an RDF vocabulary explicitly designed to describe RDF-like datasets. VoID covers four metadata domains:

• General metadata following the Dublin Core model.

• Access metadata describing how RDF data is accessed (Protocols)

• Structural metadata describing the structure and schema of datasets useful for tasks such as querying and data integration.

The DQV vocabulary [9] extends the DCAT vocabulary with additional properties and classes suitable for expressing the quality of datasets along several dimensions. The DQV is based on quality dimensions and metrics.

The PROV-O ontology [10] or Provenance Ontology is used to represent provenance information. The provenance is all the information on the entities, activities, and people involved in producing the data. This information can be used to assess its quality or reliability. Without provenance, consumers have no inherent way to trust the integrity and credibility of shared data. (W3C, Data on the Web Best Practices, 2017).

Table 1 synthesizes the coverage of those works.

Table 1. Related works coverage

	Descriptive knowledge	Technical knowledge	Structural knowledge	Government-specific Knowledge		
				Domain	*Organizational*	*Geographic*
[6]	X				X (Limited)	
[7]	X	X				
[13]	X	X				X
[12]	X	X				
[8]	X	X	X			
[9]		X				
[10]		X			X (Limited)	

2.1. Discussion

All previous works provide a rich environment for the semantic description of open datasets, however, we have found some limitations that we present below:

Specialization of vocabularies: Each vocabulary or ontology focuses on a particular aspect of metadata (description, structure, provenance, quality) without offering a broad spectrum coverage of the different knowledge necessary for describing OGD datasets.

Limitations of metadata related to the structure of datasets: all vocabularies focus mainly on the descriptive aspect without providing information about the structure of datasets; Information about the structure of datasets is provided only by the Void vocabulary whose coverage is limited to RDF datasets. The provision of knowledge related to the structure of datasets, especially in terms of government concepts described, is essential to improve data discovery and facilitate their integration.

Limitations of metadata related to the Organizational context: Organizational context refers to information about the entity/organization responsible for publishing the data. Representation of the organizational context of datasets has a significant impact on improving user trust. The organizational context is reflected by the dct: publisher relation, which links the DCAT: Dataset class with the FOAF: Agent class. However, this representation does not allow to follow the different relations between governmental organizations, their structures, and the changes and developments they are subject to over time.

Limitations of geographic context metadata: The representation of geographic context is essential for open government data because this data is related to a particular geographic entity and users are often motivated by searching for information related to a specific geographic context.

The geographical context is mainly represented in the DCAT vocabulary through the DCAT dataset class's spatial/geographical coverage property. We consider, however, that these proposals are insufficient because they do not highlight the existing relationships between the different geographical entities. Indeed, in a governmental context, the geographical entities (regions, provinces, municipalities....) are linked with relations reflected by the territorial organization of each country.

Our knowledge representation model builds on the previous works and aims to fill this gap and responds to these limitations

3. The proposed model

In this section, we present our ontology-based knowledge representation model called OGD-KM.

The model enables the formalization of the necessary knowledge for the semantic enrichment of government datasets. It is structured around four main that we describe below:

- The GovDS ontology provides the essential knowledge (descriptive, technical, contextual, and structural) necessary for the semantic description of government datasets
- The GovOrganization ontology provides the necessary knowledge to represent the organizational structures and relations between the public administrations.
- The GovGeoEntity ontology provides the necessary knowledge to represent government administrative regions and their relations following the territory's administrative subdivision.

- The GovDomain ontology is a domain that provides consensus and defines a shared conceptualization between public administrations gathering all the concepts and semantic relations necessary for the semantic representation of datasets.

3.1. The GovDS ontology

The GovDS ontology provides the essential elements for the knowledge representation necessary for the semantic description of government datasets. The structuring of this ontology is motivated by the need to provide four essential types of knowledge about government datasets (descriptive, structural, contextual, and technical) which we present below:

-Descriptive knowledge: The GovDS ontology provides the basic elements needed for the dataset's description such as title, description, and, keywords. It completes them with government-specific properties such as governmental categories, governmental themes, and the benefits expected from the publication of the dataset.

-Structural knowledge: The GovDS ontology presents the elements necessary for the description of the dataset's schema and internal structure. It provides the concepts and relationships needed to link each dataset to the government concepts it describes.

-Contextual knowledge: This knowledge is divided into three main categories:

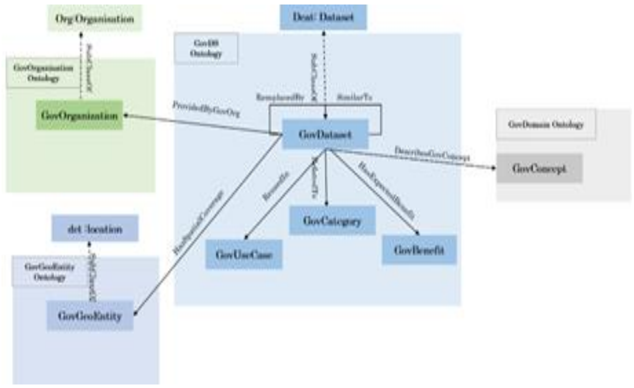

Fig.2 The main classes of the GovDS Ontology

The GovDS ontology presents information on the temporal context of the dataset through a set of properties relating to its temporal coverage, its temporal validity, its date of publication, update, and

- frequency of update.
- The GovDS ontology presents geographic context information by providing concepts and relationships that link each dataset to its geographic coverage.
- The GovDS ontology presents the information on the organizational context which refers to the information on the entity/organization responsible for publishing the dataset.

-Technical knowledge: The GovDS ontology provides the basic elements necessary for the description of the technical aspect of the dataset such as the format, the license, and the URL.

The reuse of existing standards is crucial to achieving interoperability when building an ontology. Thus, for the construction of the GovDS ontology, we reused and extended the DCAT vocabulary, a W3C standard for the description of open datasets.

For this purpose, the GovDS: Govdataset class, the main class for defining a government dataset, is a subclass of the DCAT: Dataset class, a class of the DCAT vocabulary.

Each dataset is considered as an instance of the GovDS: Govdataset class. It is described with three categories of properties shown in Figure 1:

• Descriptive properties: Title, description, keywords, category, theme.

• Technical properties: Format, license, URL, update frequency, status (valid, update)

• Temporal properties: Creation date, update date

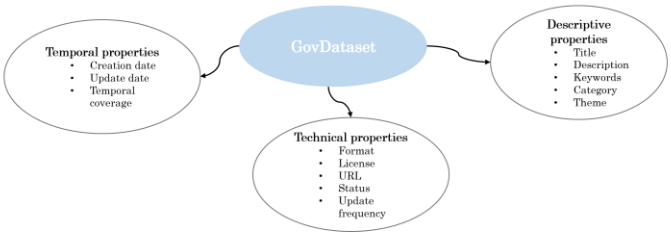

Fig.1 The main properties of the GovDataset Class

The GovDS: GovDataset class represents the core of the GovDS ontology from which relationships are established with other classes. Figure 2 presents the main classes and relationships of the GovDS ontology

The GovDS ontology also defines the GovDS: Govbenefit class GovDS: GovUseCase class which defines the possible dataset reuses. These two classes are related to the GovDS: GovDataset class with the following relations GovDS: HasExpectedBenefit and GovDS: HasPossibleUseCase.

To represent the internal structure of datasets, the GovDS: GovDataset class is linked to the GovDomain: GovConcept class of the GovDomain ontology which represents the different concepts of the domain through the GovDS: DescribesGovConcept relationship.

To represent the geographical context of the datasets, the GovDS: GovDataset class is linked to the GovGeoEntity: GovGeoEntity class of the GovGeoEntity ontology through the GovDS: HasSpacialCoverage relationship.

The GovDS ontology also makes it possible to represent relationships between datasets. For example, the GovDS: SimilarTo relationship allows you to link similar datasets. This relationship is instantiated when two datasets describe at least one concept in common.

The GovDS: ReplacedBy relation is used to represent the different versions of a dataset. This relationship is instantiated when a dataset is updated

3.2. The GovDomain Ontology

The GovDomain ontology is a domain ontology that helps provide a common model across organizations to describe government concepts and the relationships between them. The GovDomain ontology is supposed to model a consensus between Public Administrations on the concepts and relationships that exist in the published datasets. It thus forms the essential instrument for solving the problem of semantic heterogeneity. It is used to ensure the semantic enrichment of data

3.3. The Gov Organization

In the governmental context, organizations have internal structures and organizational units and are subject to many changes and developments. The GovOrganization ontology makes it possible to represent the necessary knowledge of public administrations (structures, characteristics, and evolutions).

For its construction, we reused the Organization ontology[14], a W3C standard for the description of organizations and organizational structures.

The main class of the GovOrganization ontology is the class GovOrg: GovOrganization. This latter extends the organization class and adds some properties that highlight certain characteristics of public admin as data providers.

To this end, we have added the ReliabilityScore property for the GovOrg: GovOrganization class which gives an idea of the level of trust given by users to this AP, and the EligibleToProvideInfo relationship which links the GovOrganization class to the GovConcept class . This property enables to present the different concepts and government themes for which the organization is eligible to produce and publish data.

This ontology is populated with the national public administrations.

3.4. The GovGeo Entity Ontology

In a governmental context, the geographical entities (regions, provinces, municipalities....) are linked with relations reflected by the territorial organization of each country.

The GovGeoEntity ontology is intrinsically linked to the territorial organization of each Country. For its construction, we were particularly interested in the territorial organization of Morocco without our model losing its genericity.

The GovGeoEntity: GovGeoEntity class is the main class of this ontology. It is a subclass of the dct: location class of the Dublin Core Vocabulary which refers to a spatial region or named place. We have extended it with a set of properties related to government geographic entities such as:

• Descriptive properties: name, code
• Spatial properties: area, coordinates
• Demographic properties: population

The figure 3 presents the hierarchy of the main classes and relations of the GovGeoEntity ontology.

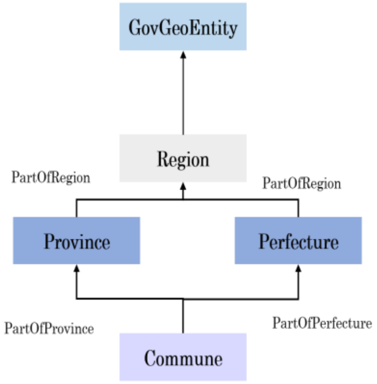

Fig.3 The main classes of the GovGeoEntity Ontology

The GovGeoEntity: Region class is a subclass of the GovGeoEntity: GovGeoEntity class, it is related to the GovGeoEntity: Province and GovGeoEntity: Prefecture classes through the GovGeoEntity: PartOfRegion relationship.

The GovGeoEntity: Commune class has a property related to its type: urban or rural. It is related to the GovGeoEntity: Province and GovGeoEntity: Prefecture class respectively with the GovGeoEntity: PartOfProvince and GovGeoEntity: PartOfPrefecture relations.

The ontology model can be reviewed and readjusted according to the particular territorial division of each country.

This ontology is populated with national geographic entities.

4. The usage of the model for the semantic enrichment of metadata

In this section, we present an approach related to the usage of our knowledge representation model for the semantic enrichment of OGD metadata.

Following the publication of a dataset, the public organization fills in the basic and organizational metadata.

Basic metadata is used to provide information on the dataset related to the following dimensions:

Descriptive dimension: This dimension is used to present a description of the dataset through a set of attributes such as title, description, associated keywords, and government category.

Temporal dimension: This dimension is used to provide information on the temporal context of the dataset through the association of a set of properties such as publication date, update date, publication frequency, and temporal coverage.

Technical dimension: This dimension is used to provide information related to the technical aspect such as format and license.

Organizational metadata is used to provide information related to the organization in charge of publishing the data.

We complete this metadata with automatic production of structural metadata and geographic metadata to bring forth a complete and adequate description of the datasets.

Structural metadata defines the schematic and the internal structure of the dataset.

Geographic metadata defines the geographic coverage of the dataset.

The different metadata are semantically enriched with our knowledge representation model and strengthened in the RDF metadata graph. This graph which serves as a ground basis to implement a semantic search mechanism on the metadata to improve data discovery on the portal.

An overview of the metadata transformation process is depicted in figure 3.

The process takes as input the dataset and its basic and organizational metadata and produces as output an RDF representation of the dataset's metadata semantically enriched with our knowledge representation model.

The process includes five steps:

- Semantic enrichment of basic metadata
- Semantic enrichment of organizational metadata
- Production and semantic enrichment of structural metadata
- Production and semantic enrichment of geographic metadata
- Metadata RDF graph consolidation and publication.

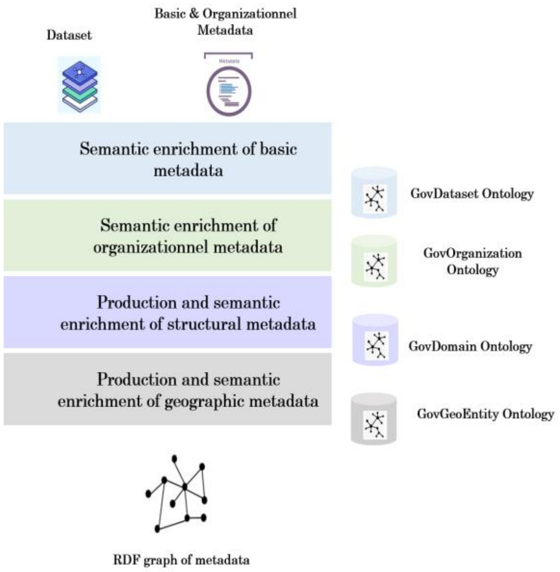
Dataset
Basic & Organizationnel Metadata
Metadata
Semantic enrichment of basic metadata
Semantic enrichment of organizationnel metadata
Production and semantic enrichment of structural metadata
Production and semantic enrichment of geographic metadata
GovDataset Ontology
GovOrganization Ontology
GovDomain Ontology
GovGeoEntity Ontology
RDF graph of metadata

Fig.5 The semantic enrichment process of metadata

4.1. Basic metadata

The basic metadata is first extracted and stored in a CSV file (MetaBase.CSV) which has a unique defined structure. Then we proceed to the definition of the RML mapping 14] which allows specifying the transformation rules from CSV to the RDF model. Later on, the CSV file (MetaBase.CSV) is transformed into the RDF model by semantically enriching it with the GovDataset ontology.

This basic metadata transformation approach is conducted in two phases:
Dataset URI production:
In this step, we associate each dataset (Ds) to a URI generated based on the attribute 'Identifier' of the GovDataset class which corresponds to the dataset identifier by respecting the following template: URIdeBase/GovDataset/{Identifier}.
Production of the RDF graph:
Based on the RML Mapping document (MetaBaseRML.ttl), the CSV file is converted to the RDF model via generating the triples that constitute the RDF graph of basic metadata.

4.2. Organizational metadata

The organizational metadata is first extracted and stored in a (MetaOrg.CSV). Then we transform the CSV file of organizational metadata into the RDF model by enriching it with the GovDataset and GovOrganization ontologies of our ontological model.
This process is conducted in two steps:

- Identification of the URI of the organization in charge of publishing the dataset:

This step consists in searching for the identifier of the organization through a URI lookup. The URI lookup is performed on the instances of the GovOrganization class of the GovOrganization ontology initially populated with all the national public organizations.
the corresponding URI is brought back to the name of the organization responsible for the data publication.

- Generation of triples:

In this step, we generate triples that constitute the organizational metadata graph. These triples allow associating the URI of the dataset with the URI of the organization through the ProvidedByGovOrganization.This relation associates instances of the GovDataset class of the GovDataset ontology with instances of the GovOrganization class of the GovOrganization ontology.

4.3. Structural Metadata

The production and transformation of structural metadata consist of associating each dataset with the governmental concepts that it describes.

This process is conducted in two steps:

- Extraction of the government concepts described in the dataset:

This step consists of extracting and identifying the different semantic types of the attributes (i.e. concepts of the ontology to which the attributes of the dataset correspond). the identification is done in a semi-automatic way. Its process is beyond the scope of this paper.

- Generation of RDF triples:

In this step, we generate triples that constitute the structural metadata graph. These triples allow associating the URI of the dataset with the URI of the concepts through the DescribesGovConcept.This relation associates instances of the GovDataset class of the GovDataset ontology with instances of the GovConcept class of the GovDomain ontology.
Note that the GovDomain ontology is an unpopulated ontology containing only domain concepts, properties, and relations. However, we instantiate the GovConcept class of the GovDomain ontology with the different concepts described in the domain ontology.

4.4. Geographic metadata

The production and transformation of geographic metadata consist of associating each dataset with its geographic coverage.
This process concerns datasets that contain information related to certain geographic entities.
Thus, for each dataset containing an attribute corresponding to a geographic entity or one of its subclasses, the process is conducted according to the following steps:

- Attribute identification and value extraction & Disambiguation of cell values:

This step consists of identifying for each value the corresponding instance in the GovGeoEntity class of the GoGeoEntity ontology or one of its subclasses and returning the canonical URI of the geographic entity.

- RDF triplet generation:

In this step, we generate triples that constitute the geographic metadata graph. These triples allow associating the URI of the dataset with the URI of the geographic entities through the GovDS:HasSpacialCoverage relation . This relation associates instances of the GovDataset class of the GovDataset ontology with instances of the GovGeoEntity class of the GovGeoEntity ontology.

4.5. Construction of the global metadata graph

The global metadata graph is built upon the union of the different metadata graphs.

The graph is published in an RDF triple store. A SPARQL endpoint is made available so machines and users can perform queries over the graph.

5. Conclusion

In this paper, we present a knowledge representation model that aims to enhance the description of OGD datasets to improve datasets 'discoverability. The model provides various kinds of knowledge about OGD datasets (descriptive, technical, contextual, and structural). We also present an approach for the usage of the model to semantically enrich OGD metadata and transform it into an RDF graph to enable its machine processability and interpretability. In our future works, we will focus on the semantic transformation of datasets and their integration to construct the Big Open government data knowledge graph.

References

[1]. Attard, J., Orlandi, F., Scerri, S., & Auer, S. (2015). A systematic review of open government data initiatives. Government information quarterly, 32(4), 399-418.

[2]. Ubaldi, B. (2013). Open government data: Towards empirical analysis of open government data initiatives.

[3]. Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. Information systems management, 29(4), 258-268.

[4]. DAHBI, K. Y., Lamharhar, H., & Chiadmi, D. (2019, October). Toward a user-centered approach to enhance Data discoverability on Open Government Data portals. In 2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS) (pp. 1-5). IEEE.

[5]. Kremen, P., & Necasky, M. (2018). Improving discoverability of open government data with rich metadata descriptions using semantic government vocabulary.

[6]. Dublin Core Metadata Initiative. (2012). Dublin core metadata element set, version 1.1.

[7]. World Wide Web Consortium. (2014). Data catalog vocabulary (DCAT).

[8]. Alexander, K., Cyganiak, R., Hausenblas, M., & Zhao, J. (2011). Describing linked datasets with the VoID vocabulary.

[9]. Albertoni, R., & Isaac, A. (2021). Introducing the data quality vocabulary (DQV). Semantic Web, 12(1), 81-97.

[10]. Garijo, D. (2013). PROV-O: The PROV Ontology Tutorial.

[11]. Gruber, T. (2018). Ontology.

[12]. Dekkers, M., Kotoglou, S., Nelson, C., Hohn, N., Pellegrino, M., & Peristeras, V. StatDCAT-AP.

[13]. Perego, A., Cetl, V., Friis-Christensen, A., & Lutz, M. (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.

[14]. The organization ontology . https://www.w3.org/TR/vocab-org/

[15]. Dimou, A., Vander Sande, M., Colpaert, P., Verborgh, R., Mannens, E., & Van de Walle, R. (2014, January). RML: a generic language for integrated RDF mappings of heterogeneous data. In Ldow.