LigADOS: Interlinking Datasets in Open Data Portal Platforms on the Semantic Web

Glaucia Botelho de Figueiredo1, Kelli de Faria Cordeiro2, and Maria Luiza Machado Campos1

1 Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro, RJ 21941-916, Brazil
glaucia.botelho@ufrj.br, mluiza@ppgi.ufrj.br
2 Center of Naval System Analysis of Brazilian Navy, Rio de Janeiro, RJ 20091-000, Brazil
kelli@marinha.mil.br

Abstract. The fostering of data opening has been largely motivated by sets of laws on access to information, which establish the need to make data related to governmental activities available to citizens and society in general, as well as results from business processes or scientific research, also for accountability and transparency. There are several ways of making data available to the public, from a simple website to sophisticated applications for accessing the data. In this context, one of the options is the construction of an open data portal using platforms for data repositories and catalogs. In the last few years, there has been a rapid proliferation of this type of portals, with domain or organization specific datasets being widely disseminated in platforms like CKAN. In these platforms, datasets are organized in thematic groups and described by keywords and other attributes assigned by the publisher. Usually described by metadata with poor semantics, these datasets very often remain as “data silos”, with no explicit connection or data integration mechanism, making it difficult for the users to locate and interrelate relevant data sources. In contrast, the Semantic Web focuses on a way of modeling and representing data in an easier manner to establish interrelationships between data, accompanied by richer descriptors. Based on this scenario, this paper proposes LigADOS, an approach to create interconnections between datasets considering their content and related metadata. LigADOS is based on the principles of the Semantic Web and associated linked data solutions and technologies, to support rich access strategies to RDF data published using portal platforms like CKAN and others.

Keywords: Dataset interlinking · Semantic metadata interlinking · Open data portals platforms · FAIR Data Point

1 Introduction

The Open Data movement has been guided by the need to comply with laws to enforce transparency and accountability, which establish the need to make public sector activities data available to citizens and the society in general. Although these data can be available in different ways on the Web, the access to open data fostered the emergence of platforms
to support data portals development and deployment. Currently, it is possible to identify some possibilities of relations among published datasets, when they share some string utilized as descriptors. However, the basis of these possible relations are vulnerable, since it considers string syntax. It is not enough to know if two or more strings have the same meaning when applied to distinct situations.

The Semantic Web goes in the opposite direction. Its associated technologies provide possibilities for data access strategies capable of dealing with interrelationships between data, accompanied by richer descriptors. The core framework, coined as Resource Description Framework (RDF) extends the linking structure of the Web to name, identify and describe resources and the relationships between them. However, open data portals platforms are not able to explore the potential of RDF datasets. In addition, the machine processing capability of RDF datasets is underutilized, as there is no support, integrated or not, to access the data representation structures utilized.

Based on this scenario, this paper presents LigADOS, an approach to support rich access strategies to RDF data published on open data portal platforms. The approach contributes with richer access possibilities, exploring linked data dispositions for search, querying and navigation.

The remainder of this article is divided into seven sections. Section 2 discusses characteristics of open data portal platforms and explains the goals of an FDP. Section 3 presents the state of the art about indirect dataset interlinkage, considering metadata associated with them on open data portals. Section 4 discusses in detail the LigADOS approach. Section 5 contains an application of LigADOS prototype in The Brazilian Open Data Portal. And, finally, Sect. 6 concludes with final remarks and topics for further investigation.

2 Open Data Portals Platforms on the Semantic Web

The Semantic Web offers the fundamentals to interconnect, across the Web, information resources that are not originally associated, taking advantage of the Web infrastructure. Besides HTTP and URIs (Uniform Resource Identifier), the Semantic Web and, more specifically, linked data, use RDF (Resource Description Framework) as a data representation framework. This infrastructure is the basis of the Web of Data, which is focused on interlinking data published on the Web and machine processable semantic annotations on vocabularies and ontologies [1].

Several governmental and non-governmental organizations – at the local, regional, national and international levels – have made public some data produced by their internal processes. As examples of national level initiatives, several countries have launched their own Open Government Data Portals (OGDP) contributing to society participation and the understanding of government decisions.

Comprehensive Knowledge Archive Network (CKAN)¹, Socrata² and Opendatasoft³ are platforms for developing portals focused on publishing open government data.

¹ https://ckan.org/.
² https://www.tylertech.com/products/socrata.
³ https://www.opendatasoft.com/.
CKAN has been widely used in government portals in countries in Europe, South America, North America, Africa and Australia. Mainly local governments in the United States of America have used Socrata; and Opendatasoft is most popular with European organizations. However, Socrata and Opendatasoft are paid solutions. Dataverse⁴ and DSpace⁵ are open data portal platforms more focused on the support of the scientific community [2], that have been applied to organize data considering some scientific literature that references them, even structuring the quote.

Many open data initiatives associated with government processes use the CKAN platform as the support for their OGDPs. It is a free and open source tool, focused on creating websites for open data availability, developed and maintained by the Open Knowledge Foundation. A dataset comprises the metadata that describes it and the distributions that contain the data itself. A dataset can be composed of at least one and at most several distributions. These distributions can be of different formats, proprietary or not, such as XLSX, PDF, XML, CSV and RDF. CKAN does not impose rules on how to organize data distributions.

Natively, CKAN provides a series of descriptor options (metadata) for each dataset. Although RDF datasets have metadata in smaller grains than the set as a unit, CKAN does not exploit this advantage in the available mechanisms for assigning metadata or in its mechanisms for accessing the data. In fact, most of the available metadata options refer to structural metadata, that is, they refer to descriptive characteristics (title), storage (unique identifier) and management (use licenses). The option of metadata with semantic characteristics is limited to tags. Tags, in turn, are usually assigned by the data publishers themselves. This does not guarantee the best representativeness of the data, since there are no rules or restrictions in their assignment [3]. Furthermore, in CKAN, tags are assigned to each dataset, often referring to a thematic area, not necessarily representing the instances of the datasets. Several research have developed works to support and automate tag generation [3–5].

Currently, working groups associated with the GO FAIR initiative have recommended a kind of metadata organization perspective divided into five metadata schemes or levels [6]. FAIR is the acronym for “Findable”, “Accessible”, “Interoperable” and “Reusable”, an initiative that recommends a set of guiding principles to make data generated and/or used by scientific research easier to be discovered, accessible, interoperable and reusable by both people and machines [7].

The possibility of developing and using extensions already coded and tested, associated with the fact that most OGDPs initiatives publish their open data using the CKAN platform, led to the choice of this platform for use in the experiments of the proposed approach.

Considering metadata spread in several repositories, there is a trend to the establishment of a federation of data points to improve data discoverability by human beings and by machines. As part of the FAIR initiative, these central points, coined FAIR Data Points (FDP), actually constitute repositories of metadata. The FDP architecture includes a Web-based graphical user interface (GUI) and an application programming interface

⁴ https://dataverse.org/.
⁵ https://duraspace.org/dspace/.
The Metadata Provider is a core component responsible for the provisioning of the metadata content available in the FDP. Aligned to the FAIR principles, five metadata schemes describe complementary layers of data organization. The five layers are FDP Metadata (level 1), Data Catalog Metadata (level 2), Dataset Metadata (level 3), Distribution Metadata (level 4) and Data Record Metadata (level 5).6

The metadata organization in five layers, a hierarchy from high-level less granular to bottom-level more granular, moves towards better dataset findability, as metadata are more detailed each level to better understand the data contents without the need of accessing the data instances. In addition, metadata includes the information about dataset licenses and access protocols. This facilitates interoperability and data reuse, as proposed by FAIR initiative. Further, it opens possibilities towards data interlinkage.

3 Open Data Interlinkage

Much work has been invested in developing approaches to make linked open data easily accessible. In this context, a link makes it possible to navigate in a seamless way between resources belonging to different datasets, possibly of different domains, giving access to richer and more complete information than the data at hand. There are approaches that focus on the quality of a linkset, a special kind of dataset containing only RDF links between two datasets. For this, they define measures like the *average linkset reachability*, which is a type of metric to evaluate the number of new concepts reached by crossing a linkset [8]. There are approaches that propose supporting the discovery of links to enrich new datasets to be published, or just connect available datasets without considering metadata associated to them [9–14]. This means that they do not consider including data that play the role of descriptors. In contrast, there are approaches that propose data interlinkage, considering metadata and indirect links between the datasets. Indirect in the sense that a moderating component mediates connections. Also, most often, the links are external to the portals, because the moderating component is not embedded in the data portals [3, 5–7, 15–17].

In this regard, when indirect data interlinkages use specialized vocabulary for descriptive metadata annotation, such as DCAT, VoID, Schema.org [6, 7, 15, 16], the portals contemplate a rich set of metadata. However, they are still limited to identify how a published dataset metadata was organized, with no support to content-related interlinkage processes. Although tags are usually included in these sets of metadata, they represent only a simple string. When domain specific vocabularies are utilized, they are able to annotate only part of metadata: usually tags or the indication of the main dataset domain [3, 5, 17]. Indeed, it adds another semantic level, lifting simple strings to semantic expressions. Combining annotation using specialized vocabulary for descriptive metadata and domain specific vocabularies takes to richer descriptions allowing semantic interlinkage processes [17].

Anyway, this combination keeps metadata in dataset level or in distribution level, disregarding data instances. When data instances are considered to extract some level

---

6 [https://github.com/FAIRDataTeam/FAIRDataPoint-Spec](https://github.com/FAIRDataTeam/FAIRDataPoint-Spec)
of metadata, the semantic lifting is higher [5, 7]. Moreover, it gives the ability to data consumers to grasp the dataset content without delving into data instances.

A track of investigation that has stood out in recent years involves the establishment of guidelines and the development of solutions to deal with FAIR data, emphasizing metadata interlinkage [7]. In this context, an approach is the development of a semantic proxy to collect and adapt existing metadata to this new reality [6].

In this paper, we explore data instances as sources to obtain a more granular level of metadata. Thus, we propose an approach that considers indirect RDF dataset interconnections, by defining and exposing data record level metadata interlinkage.

4 LigADOS Approach

LigADOS proposal aims to take advantage of the benefits offered by the Web of Data to interlink datasets published on open data portals, generate higher grained semantic metadata, as well as provide support for machine processing. The focus of the approach is the interconnection of datasets already aligned with the concepts of the Web of Data, that is, datasets modeled in RDF in any type of serializations. The interconnection between datasets are established by data instances semantic classifications, that is, extracting higher level semantic descriptors from their embedded semantic annotations. Data instances classifications form a kind of semantic metadata, aligned to FAIR principles. So, they serve as inputs to compose metadata schemes proposed by FDPs, while expanding the possibilities for interconnecting datasets published in different portals. LigADOS creates an interface, aimed at the general public, to visualize dataset interlinkages, from where it is possible to easily identify and access the datasets involved in the interconnections.

The approach is composed of five macro processes: (i) metadata extraction; (ii) semantic interlinkages generation; (iii) semantic tags generation; (vi) semantic interlinkages publication; and (v) semantic metadata generation to support data portal interlinkages (see Fig. 1). A technical agent triggers the whole process, transparently to the functioning of the portal instance. This technical agent can be an artefact programmed for this purpose or a human being. The execution process occurs without external agents interactions. The processes execution beneficiaries are the portal data consumers (users), as well as the applications able to process linked data.

The “Metadata extraction” macro process starts with the scanning of the entire portal instance in order to extract metadata from all published datasets, with at least one distribution modeled in RDF. The extraction generates a new dataset, represented in RDF, containing a set of metadata annotated with the DCAT vocabulary that is then persisted in a triplestore. This choice of vocabulary follows a trend used in other works [6, 7, 16], but which is, mainly, aligned with the FAIR recommendations [7], for clearly allowing the identification of dataset distributions.

The “Semantic interlinkages generation” macro process is the core of the interlinkage processes. Firstly, LigADOS extracts the URLs of the distributions represented in RDF from the set of metadata. After that, it reads all data in order to identify the triple subjects and objects that are identified by a URI, and that are annotated with some vocabulary, ontology or thesaurus. Thus, the approach triggers federated queries to find the URI of the
class or concept that categorizes the term and/or element used to annotate a triple resource of the dataset distribution. The identification of the class or concept considers the element exactly one level above, looking for relations such as skos:broader\(^7\), rdf:type\(^8\), rdfs:subClassOf\(^9\). Once the class or concept is identified, the query also extracts the class label looking for relations as skos:prefLabel\(^10\) or rdfs:label\(^11\). These classes and/or concepts are used as inputs for assembling the triples that will be part of the Graph of Semantic Interlinkings, which is the output of this activities flow.

The Graph of Semantic Interlinkings is formed by nodes that represent: (i) the open data portal instance; (ii) the organizations that published the datasets; (iii) the datasets; (iv) the RDF datasets distributions; (v) the vocabularies, ontologies and thesaurus referenced in the datasets distributions; and (vi) the classes and/or concepts of the vocabularies, ontologies and thesaurus that classify the terms and/or elements used to annotate the triples of the datasets distributions. Note that the nodes, which represent vocabularies, ontologies and thesaurus, as well as those that represent the classes and/or concepts of these ontology resources, configure as metadata of data record level. This means they are aligned with the FAIR initiative and the metadata schema layers proposed in the specification of an FDP.

In short, the Graph of Semantic Interlinkings is a metadata graph. It aggregates in the same graph different metadata levels, when compared to the FDP metadata layered organization. Its prominence is given by the semantic descriptors of the data instances. The graph indicates the indirect interconnections between datasets and yet generates more granular semantic metadata than originally associated with datasets. Besides, aligned with the FAIR initiative, which recommends richly described metadata, this graph supports the identification of new possibilities for interoperability and reusability.

---

7 skos:broader refers to the “broader” property of the SKOS vocabulary.
8 rdf:type refers to “type” properties defined in the abstract syntax of the RDF vocabulary.
9 rdfs:subClassOf refers to the “subClassOf” property of the RDFS Schema vocabulary.
10 skos:prefLabel refers to the “prefLabel” property of the SKOS vocabulary.
11 rdfs:label refers to the “label” property of the RDFS Schema vocabulary.
To define nodes, either as a subject or as an object, as well as to establish connections between these nodes, it is necessary to use specialized vocabularies. For the Graph of Semantic Interlinkings the following vocabularies are used: Friend of A Friend (FOAF)\textsuperscript{12}, Data Catalog Vocabulary (DCAT)\textsuperscript{13}, Dublin Core™ Metadata Initiative (DCMI) Metadata Terms\textsuperscript{14}, Vocabulary of A Friend (VOAF)\textsuperscript{15}, Vocabulary for Annotating Vocabulary Descriptions (VANN)\textsuperscript{16} and, of course, the domain specific vocabularies from where class/concepts were extracted. An instance of a data portal must have only one associated graph generated.

Considering the approach defined in [5], the “Semantic tags generation” macro process generates triples using the “keyword” property of the DCAT vocabulary (\texttt{dcat:keyword}) in order to link the URIs of the datasets to the URIs of the extracted classes/concepts\textsuperscript{17}. Thus, these triples are included in the Graph of Semantic Tags, also persisted in a triplestore. As a result, it is possible to browse the labels associated with a dataset in order to explore other potential datasets that use the same tags, thus discovering indirect links between them. The difference between LigADOS and that approach is that the former obtains semantic tags at a higher level of classification than the latter one, as it contemplates classes/concepts of elements/terms used in the annotation of the triples resources. This last point is also an advantage of LigADOS, since it considers datasets represented in RDF, while in [5] consider data in tabular formats. In LigADOS, it is essential that semantic metadata are accessible on the portal platforms used as catalogs or in any repositories used as an access point for open data: governmental, institutional or research. Thus, when proposing the use of an approach for the generation and availability of semantic tags on these platforms, it corroborates to the FAIR guidelines for data discovery (Findability), also facilitating accessibility (Accessibility).

The “Semantic interlinkages publication” macro process considers data consumers on open data portals with different knowledge profiles regarding the Web of Data technologies. Due to this heterogeneity of profiles, the major factor is to provide some intuitive support for the use of RDF datasets. A requirement to meet this demand involves an interface support that uses visual and interactive resources. As portal platforms are not yet able to provide these resources natively, whereas there are software solutions that already exploit these requirements, the LigADOS approach proposes the combination of both worlds.

\textsuperscript{12} http://xmlns.com/foaf/spec/.
\textsuperscript{13} https://www.w3.org/TR/vocab-dcat-2/.
\textsuperscript{14} https://www.dublincore.org/specifications/dublin-core/dcmi-terms/.
\textsuperscript{15} https://lov.linkeddata.es/vocommons/voaf/v2.3/.
\textsuperscript{16} https://vocab.org/vann/.
\textsuperscript{17} The latest version of the DCAT vocabulary (W3C Recommendation 04 February 2020) has made the domain of the “keyword” property flexible for any type of resource (\texttt{dcat:Dataset}, \texttt{dcat:DataService}), maintaining its range as a literal. Tags are also associated with datasets on portal platforms. Therefore, both the approach defined in [5] and the approach described here consider associating the generated tags to the dataset. However, both approaches experiment to make the range of the aforementioned property more flexible, in the sense of pointing to a class/concept of a common consensus vocabulary, which has an associated label.
This macro process makes available a download option, an interactive graph exploration option and a SPARQL endpoint to the Graph of Semantic Interlinkings. In addition, it publishes an interactive graph exploration option to the Graph of Semantic Tags.

The “Semantic metadata generation to support data portal interlinkages” macro process generates a set of extended metadata, which is aligned to the metadata profile defined for a FDP. That is, a profile defined for a metadata repository that provides access to federated data on open data portals. Systematically, LigADOS reads the set of metadata extracted from the portal instance, as well as the Graph of Semantic Interlinkings to extract four of five metadata organization levels for a FDP. These metadata need to be complemented with other metadata defined for a FDP profile. Since all metadata structure depends on a FDP installation, it is essential that LigADOS receives the FDP URL installation as data input.

Among the triples to complement the FDP profile, we can cite the one that establishes the metadata schema layers. Different from other metadata levels (catalog, dataset, distribution), where it is easy to perceive the identification level, the data record level requires some adjustments, as we need first to define how to identify this level.

The data record metadata is directly associated with a class/concept extracted from a vocabulary. This class/concept may be used by different dataset distributions, from the same dataset or from others. By itself, it is not able to characterize a single data record. A combination of data is necessary for its complete characterization.

In this context, the application of the alternative approach for RDF reification – RDF* – raises as an appropriate option. The RDF* approach allows nested triples, aiming to annotate triples resources with another metadata triple. Nested (embedded) triples can naturally correspond to the concept of triple identifiers [18].

Thus, the characterization of a data record occurs by combining the identification of the data distribution with the vocabulary classes/concepts that characterize terms/elements used to annotate its triples. This association represents the core of the interconnection process proposed by LigADOS. The following fragment represents the triple structure that identifies metadata of data record level. To compose it, LigADOS extracts data from the Graph of Semantic Interlinkings (using RDF* reification on subject) and assembles the relation to the parent metadata (object).

```PREFIX dc: <http://purl.org/dc/terms/> .
PREFIX vocab: <http://namespacovocabulario.org/ns/> .

<<<FDP_URI_Distribution> dc:references vocab:URI_Class>>
dc:isPartOf <FDP_URI_ParentMetadata> .
```

LigADOS supports metadata extraction of catalog, dataset, distribution up to the level of data record metadata. The data steward must provide the metadata of the data record level, not the FDP owner. Thus, LigADOS automates this activity, also supporting portal data stewardship. At the FDP environment, the data record metadata level allows the identification of data interconnections managed by different portals without federated queries. The FDP Metadata Provider component can retrieve this graph through a SPARQL endpoint whenever required, to store in the FDP local triplestore. It should be noted that this macro process simplifies a set of activities for the inclusion of metadata that would be performed by the person in charge of the target FDP.
The data record level metadata carry the basis for indirect interconnections between datasets published in the same portal instance or between datasets published in different portals. The *Graph of Semantic Interlinkings* alone contains the interconnections within a portal. When this type of graph, belonging to different portal instances, is loaded in a FDP, it makes it possible to establish interconnections between portals. As the FDP structure is based on the cataloging of metadata to reach federated data, the interconnections between portals occur through the semantic metadata. This metadata central point makes it possible to query these interconnections in the FDP environment, without the need to use federated queries to source portals.

As LigADOS extracts instances classifications to generate data record metadata, it provides support to create indexes. As each class/concept extracted represents several resources (nodes) of the original datasets (graphs), one can detect certain kind of information without actually consulting the datasets. Although some generated benefits may be similar to those of RDF summarization [19], LigADOS does not apply any type of RDF summarization method, but creates a brand new dataset (graph) to represent non explicit data interlinkages between all datasets (graphs) published in a portal. In [19] was depicts a high-level taxonomy of the RDF summarization approaches, grouping the main methods of the algorithms into four main categories (structural, statistical, pattern-mining and hybrid), identifying subcategories whenever possible. LigADOS considers only annotated resources that is resources defined in vocabularies, ontologies or thesaurus, in any triple position, without concern about reducing datasets sizes.

## 5 LigADOS Approach Prototype and Evaluation

A prototype supporting the LigADOS approach was developed in Java, using the RDF4J framework and storing triples in GraphDB™. GraphDB™ has features for users to explore and navigate graphs generated from the result of SPARQL queries, even providing interactive features. Two modules support the approach: one to construct and publish the interlinkages, as well as the semantic tags; and another to prepare metadata for the FDP. Both work without prompting the user, running in the background of the portal instance, even to publish the interlinkages at the portal. For this, the first module uses the CKANAPI.

An experimentation of the approach and prototype was carried out in instances of open data portals based on CKAN. The Brazilian Open Data Portal is used by the Brazilian public sector to provide data on many varied topics, from governmental agencies and administrative organizations at the federal level.

The approach was applied to a subset of datasets from the environmental domain extracted from this portal. At the Brazilian Open Data Portal, there were twenty-five publications represented in RDF. However, twenty-three of them were published by the same organization, which is not enough to demonstrate the approach developed. As a result, we extracted some datasets in their original formats; next, we triplified them to evaluate the approach. As an example, we considered two datasets published by different organizations. The original publication of the “Vegetation of Brazilian Mangrove” dataset

---

18 In this paper, the original descriptions of the datasets and its distributions were translated from Portuguese to English.
is composed of three distributions. The “National Forest Information System - SNIF” dataset is composed of twenty-eight distributions. Among the associated metadata are the title, the organization that published them, a brief textual description, administrative metadata and tags. For the “Vegetation of Brazilian Mangrove” dataset, there are only three associated tags. Although there are twenty associated tags for the “National Forest Information System - SNIF” dataset, there are no intersections between the sets of tags. In general, the datasets fall within the environmental domain, but there is no evidence of relationships between the distributions of the datasets. There are neither common labels, nor coincident descriptions that lead to consider the existence of any kind of relationship between them.

After applying the LigADOS approach, which generated semantic interconnections considering the semantics embedded in the data, it was possible to verify that both datasets shared the same common Aims-Agrovoc thesaurus concepts (in Portuguese): “Vegetação” (Vegetation), “Mato” (Scrub), “Floresta de regiões húmidas” (Rain Forest) e “Floresta tropical” (Tropical Forest) (see Fig. 2). In Fig. 2, the highlighted arrows show the dataset interlink through the Aims-Agrovoc” node and distribution interlinks by common thesaurus concepts utilized by dataset distributions.

![Fig. 2. Interconnections between dataset distributions by the use of a common thesaurus](image_url)

Other interlinks were established on the Brazilian Open Data Portal. Table 1 depicts some dataset distributions interlinks for the environmental domain, including datasets published by distinct organizations, before and after the approach use. Some previous interconnections refer to the organization that published the dataset, others refer to some specific word utilized to describe the dataset or even the name of the file distribution. Just a few correspond to tags, and they do not even consider semantic aspects, because they are not linked to vocabularies, behaving as simple textual descriptions. In the “After” columns at Table 1, we disregard the original interlinkages, exposing only the brand new semantic interlinkings generated.

The numbers show that the LigADOS approach increases the number of interlinkages. However, this is not the most important aspect. The main aspect is that the new interlinkage points are considering the instances of data, namely, the semantic meaning resulting from the use of the same vocabulary, ontology or thesaurus, and not just free terms attributed at the dataset publication at the portal instance.
6 Conclusions and Future Works

This paper described LigADOS, an approach for interconnecting datasets on open data portal platforms on the Semantic Web. LigADOS explores characteristics of linked data representation and vocabulary based annotation as a strategy to support the interconnection of published RDF datasets. The main strategy is to increase the level of semantic expressiveness of the datasets, in order to establish and expose existing non-explicit interconnections among them, not previously perceived by their publishers. Also, LigADOS supplies semantic metadata to compose four of five metadata schema levels of a FDP, specially the data record metadata level, which expands the dataset interlinkages possibilities between data portals.

The results indicate that interconnections are easily established when datasets modeled in RDF make use of common vocabularies, ontologies or thesaurus. Additionally, they indicate that datasets published by different organizations and for different purposes may have hidden interconnections mainly due to the way in which data portal platforms currently handle linked data, and also because of the lack of semantic metadata. In contrast, the results of other experiments carried out indicate that datasets which, initially, seemed to deal with related issues (by observing the metadata cataloged at the time of their publications), may not actually have semantic interconnections when considering data instances. Finally, the prototype application showed that it is possible to integrate data portals platforms with triplestores like GraphDB, providing interfaces for graphical navigation in RDF.

As future work, we intend to use graph centrality measures to identify classes and/or concepts from which we can generate a subgraph. Centrality measures can be applied to quantify the nodes that most serve as hubs between datasets. Thus, semantic interconnections could be initially explored from several points and not from a single one (currently the initial points represent the “dcat:Catalog” and the “dcat:Dataset” type.

References

1. Hendler, J., Berners-Lee, T., Miller, E.: Integrating applications on the semantic web. J. Inst. Electr. Eng. Jpn. 122(10), 676–680 (2002)
2. Langer, A., Bilz, E., Gaedke, M.: Analysis of current RDM applications for the interdisciplinary publication of research data. In: Proceedings of the 1st International Workshop on Approaches for Making Data Interoperable SEM4TRA-AMAR® SEMANTICS (2019)
3. Tygel, A.A.: Semantic tags for open data portals: metadata enhancements for searchable open data. D.Sc. thesis. Federal University of Rio de Janeiro, Rio de Janeiro (2016)

4. de Lira, M.A.B.: Uma Abordagem Para Enriquecimento Semântico de Metadados Para Publicação de Dados Abertos. M.Sc. thesis. Federal University of Pernambuco, Pernambuco (2014)

5. Castro, B.P.C., Rodrigues, H.F., Lopes, G.R., Campos, M.L.M.: Semantic enrichment and exploration of open dataset tags. In: Proceedings of the 25th Brazilian Symposium on Multimedia and the Web, pp. 417–424. Association for Computing Machinery, New York (2019)

6. Moreira, J.L.R., et al.: Towards findable, accessible, interoperable and reusable (FAIR) data repositories: improving a data repository to behave as a FAIR data point. Liine em Revista 15(2), 244–258 (2019). https://doi.org/10.18617/liinc.v15i2.4817

7. Wilkinson, M., et al.: Interoperability and FAIRness through a novel combination of Web technologies. PeerJ Comput. Sci. 3, e11 (2017). https://doi.org/10.7717/peerj-cs.110

8. Albertoni, R., De Martino, M., Pouetá, P.: Linkset quality assessment for the thesaurus framework LusTRE. In: Garoufallou, E., Subirats Coll, I., Stellato, A., Greenberg, J. (eds.) MTSR 2016. CCIS, vol. 672, pp. 27–39. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49157-8_3

9. Leme, L.A.P.P., Lopes, G.R., Nunes, B.P., Casanova, M.A., Dietze, S.: Identifying candidate datasets for data interlinking. In: Daniel, F., Dolog, P., Li, Q. (eds.) ICWE 2013. LNCS, vol. 7977, pp. 354–366. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39200-9_29

10. Caraballo, A.A.M., Nunes, B.P., Lopes, G.R., Leme, L.A.P.P., Casanova, M.A.: Automatic creation and analysis of a linked data cloud diagram. In: Cellary, W., Mokbel, M.F., Wang, J., Wang, H., Zhou, R., Zhang, Y. (eds.) WISE 2016. LNCS, vol. 10041, pp. 417–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48740-3_31

11. Achichi, M., Bellahsene, Z., Ellefi, M.B., Todorov, K.: Linking and disambiguating entities across heterogeneous RDF graphs. J. Web Semant. 55, 108–121 (2019)

12. Martins, Y.C., da Mota, F.F., Cavalcanti, M.C.: DSCrank: a method for selection and ranking of datasets. In: Garoufallou, E., Subirats Coll, I., Stellato, A., Greenberg, J. (eds.) MTSR 2016. CCIS, vol. 672, pp. 333–344. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49157-8_29

13. Gomes, R.V.A., Casanova, M.A., Lopes, G.R., Paes Leme, L.A.P.: A metadata focused crawler for linked data. In: Proceedings of the 16th International Conference on Enterprise Information Systems, vol. 2, pp. 489–500. SCITEPRESS, Lda (2014)

14. Kettouch, M., Luca, C., Hobbs, M.: SemiLD: mediator-based framework for keyword search over semi-structured and linked data. J. Intell. Inf. Syst. 52(2), 311–335 (2019). https://doi.org/10.1007/s10844-018-0536-1

15. Milić, P., Veljković, N., Stoimenov, L.: Linked relations architecture for production and consumption of linksets in open government data. In: Janssen, M., et al. (eds.) I3E 2015. LNCS, vol. 9373, pp. 212–222. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-25013-7_17

16. Neumaier, S., Umbrich, J., Polleres, A.: Lifting data portals to the web of data. In: LDOW@WWW (2017)

17. Pesce, V., Maru, A., Archer, P., Malapela, T., Keizer, J.: Setting up a global linked data catalog of datasets for agriculture. In: Garoufallou, E., Hartley, R.J., Gaitanou, P. (eds.) MTSR 2015. CCIS, vol. 544, pp. 357–368. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24129-6_31

18. Hartig, O., Thompson, B.: Foundations of an alternative approach to reification in RDF. arXiv preprint arXiv:1406.3399v2 (2019)

19. Čebirić, Š., et al.: Summarizing semantic graphs: a survey. VLDB J. 28(3), 295–327 (2019)