Advancing on the Linked Open University Context: a Cuban linked Open University

Yoan Antonio López Rodríguez,[1][0000−0001−5615−375X]

University of Informatics Sciences, Cuba. yalopez@uci.cu

Abstract. Linked Open Universities apply Linked Data to publish information about universities. In 2010, the Open University in the UK was launched as the first initiative to expose public information from the university in an accessible, open, integrated, and Web-based format. Since then, universities around the world have been joining that initiative by deploying their own linked open data platforms. However, during the publication of their data using the linked data principles, universities face challenges such as the lack of a unified, well-accepted vocabulary; the need of coping with the heterogeneity of datasets; the high cost to host the existing SPARQL endpoints; the performance shortcomings in federated queries over current SPARQL endpoints and the incompleteness of datasets. The aim of this Ph.D. research proposal is to advance on the Linked Open University context with a proposal for the University of Informatics Sciences from Cuba addressing some of these challenges.

Keywords: Knowledge graph · linked data · linked open university · ontologies

1 Problem Statement

Linked Open Universities apply Linked Data to publish information about universities [3,5,4]. Traditionally, universities produce large amounts of data, much of which should be publicly available [3,5]. In this context, sharing and reusing knowledge is a challenge where Linked Open Data (LOD) can play an important role[3,5]. In 2010, the Open University in the UK was launched as the first initiative to expose public information from the university in an accessible, open, integrated, and Web-based format[3]. Since then, universities around the world have been joining that initiative by deploying their own LOD platforms[6,8,12].

The process of generating linked datasets in universities consists of the following stages[10]: i) raw data collection, ii) defining the vocabulary model based on reusing existing ontologies and extend them when it is needed, iii) extracting and generating RDF datasets according to the defined vocabulary, iv) achieving interlinking among datasets internally and externally, v) storing the outcome datasets and exposing them via SPARQL endpoints, vi) exploiting datasets by developing applications and services on top, and, vii) providing optimization and quality. In this sense, during the process, institutions have faced several issues. Some of those issues are the following (without the intention of being exhaustive)[10,8,1,13,7]:

- Lack of a unified, well-accepted vocabulary that satisfies all universities' requirements. Vocabularies (or ontologies, more strictly), define the concepts and relations (also referred to as "terms") used to describe and represent an area of concern. A standard set of vocabularies provides a unified access to data consumers [5]. That is why former works about linked open universities agree with reusing existing vocabularies as much as possible. However, despite the fact that there are a lot of vocabularies/ontologies to describe entities in this context, there is no a standard vocabulary/ontology (or group of them) that meets the requirements of all universities.
- The need of coping with the heterogeneity of datasets. Open university repositories look like a sea full of wealth resources, and the problem is how we can reach the needed resources easily. Description of dataset metadata is a crucial step to cope with the heterogeneity of datasets.
- The high cost of the existing SPARQL endpoint interfaces. Due to the high expressive power of SPARQL, the processing of many kinds of queries on SPARQL endpoints is very expensive in terms of server CPU time and memory consumption. Sometimes that consumption is unpredictable and thus, SPARQL endpoints suffer from frequent downtime.
- Performance shortcomings of federated queries over current SPARQL endpoints. A 2013 survey revealed that the majority of public SPARQL endpoints had an uptime of less than 95%. This unavailability increases in federated queries where we query many knowledge graphs.
- Incomplete data is a problem that could complicate data querying and retrieving if it is not considered.

As of the previous problem statement, the aim of this Ph.D. research proposal is to advance on the linked open university context (LOU context) with a proposal for the University of Informatics Sciences from Cuba, applying novel techniques developed by the Semantic Web community and our lab to address previous challenges. The importance of this problem statement is described as follows in Section 2.

2 Importance

Universities use LOD platforms to publish and access data about staff profiles, courses, scholarly publications, and open educational resources[3]. Implementing LOD can be summarized as using the web both as a channel to access data and, as a platform for the representation and integration of data. Having university linked datasets facilitates effective reuse of data and increases the opportunities to build more effective, integrated, and innovative applications based on these datasets[6]. It is also a trend revealing the contributions and achievements of the universities and making them visible on the Web as LOD as it is a means to measure the university's reputation and standing among other international universities and institutions[10,9].

Advancing on the LOU context means resolving some existing issues with the objective of improving current and future LOD platforms. Aiming at a unified vocabulary (or group of them) entails to get better current semantic modeling processes in order to select the better vocabularies/ontologies and extend them if it is needed. Current knowledge globalization makes the standardized access to resources further relevant in the LOU context than in any other one. Besides the data itself, it is also significant to describe datasets in order to increase their discoverability and interoperability on the web via associated metadata. Related work about semantic modeling in the LOU context is described in Section 3.

On the other hand, RDF datasets access and query in the LOU context are extremely important for both, internal and external clients. A SPARQL endpoint is the common interface to triple live queryable and its shortcomings are quite well known[13,7]. To cope with the shortcomings of SPARQL endpoints, the Semantic Web Community has created some technologies such as Triple Pattern Fragment [13]. To advance on the LOU context this proposal takes into account these technologies along with common interfaces to query RDF triples. Related work about dataset´s access and query in the LOU context is described in Section 3.

Regarding incomplete data, the open word assumption of the Semantic Web assumes incomplete information by default, and thus, the possibility of finding missing facts contributes to dataset optimization. Related work about knowledge graph completion in the LOU context is described as follows in Section 3.

3 Related work

On the semantic modeling process, according to[3], the process to choose the terms is based on the following process: i) identify the concept to be expressed; ii) search for a widespread existing vocabulary to be used; iii) if found, use it, otherwise iv) search for a less-known vocabulary to reuse; v) if not found, create a new term. Moreover, according to [8] this process includes: i) listing common data categories about university information; and ii) getting a summary of useful vocabularies for describing university datasets. When it is necessary to create new ontologies, methodologies such as Methontology can be taken into consideration.

Despite there has been a recent trend towards using the same vocabularies, in the beginning, each platform used its own vocabularies/ontologies. For instance, despite the fact that a course is the same in all universities, we can find it modeled both as a Teach: Course[1] as well as an Aiiso: Course[2]. With the aim of getting closer to a unified model, our proposal aims to use the most popular vocabularies/ontologies of the state of the art, and thus, we define the first research question and hypothesis related to the semantic modeling process in Section 4.

Among the dataset metadata vocabularies, some notable ones are VoID[3] used by most university platforms to describe datasets[3,6,12], the DataCube

[1] http://linkedscience.org/teach/ns#Course

[2] http://purl.org/vocab/aiiso/schema#Course

[3] https://www.w3.org/TR/void/

vocabulary[4] to describe datasets of multidimensional data, and the DCAT[5], a feasible way to standardize metadata for catalogs, datasets and data services. In this research proposal, we define the second research question and hypothesis related to dataset metadata vocabularies in Section 4.

In regard to datasets access and query, the SPARQL language[6] is the W3C standard to express declarative queries over collections of RDF triples. There are three common interfaces to RDF triples: Data dumps, SPARQL endpoints and Linked Data documents [13]. To cope with the shortcomings of SPARQL endpoints, the Semantic Web Community has created some technologies, such as Triple Pattern Fragments (TPFs), which divides the query processing between clients and servers and allows to restrict the kinds of queries the client can send to the server[13]. Linked Connections are an example of a customized query interface for the consumption of open data in the Transportation area[2]. Linked Connections implement HTTP content negotiation[7][1]. Both, Linked Connections and TPFs use vocabularies such as Hydra[8] and Tree[9] that contribute to the automation of the client-server communication and facilitate the federated queries[1]. On the LOU platforms, to the best of our knowledge, customized query interfaces of triples have not been implemented yet, publication and consumption are mostly solved only via SPARQL endpoints. We define the third research question and hypothesis related to datasets access and query in Section 4.

Regarding incomplete data, given a Knowledge Graph $G = (E, R, T)$, where E and R denote the set of entities and relations and $T = (h, r, t)|h, t\epsilon E, r\epsilon R$ is the set of triplets (facts), the task of Knowledge Graph Completion (KGC) involves inferring missing facts based on the known facts[11]. Much research work has been devoted to KGC. A common approach to carry out KGC has been knowledge graph embeddings. Most knowledge graph embedding models only use structure information in observed triple facts, nevertheless, advanced techniques use other information besides facts such as entity types, relation paths, textual descriptions, and logical rules [14]. To advance on the LOU context, our proposal aims at defining a knowledge graph embedding model able to use other information besides facts on the output university datasets. We define the fourth research question and hypothesis related to knowledge graph completion in Section 4.

[4] https://www.w3.org/TR/vocab-data-cube/
[5] http://www.w3.org/TR/vocab-dcat-2/
[6] http://www.w3.org/tr/sparql11-query/
[7] http://tools.ietf.org/html/rfc7231#section-5.3
[8] http://www.hydra-cg.com/spec/latest/core/
[9] https://treecg.github.io/specification/

4 Research questions and hypotheses

Regarding the lack of a unified, well-accepted vocabulary that satisfies all universities' requirements, we define the first research question and hypothesis as follow:

Q1: How to contribute to having a unified, well-accepted vocabulary in the LOU context?

H1: If we use the most used vocabularies/ontologies or extend them in our LOD platform, we contribute to having a unified, well-accepted vocabulary that satisfies all universities' requirements.

Regarding dataset metadata vocabularies, we consider that advancing on the LOU context means to advance on the automation of the client-server communication, therefore we need a metadata vocabulary to allow client applications to get resources easily. In this sense, we define the second research question and hypothesis related to dataset metadata vocabularies as follows:

Q2: How to support client applications getting resources easily in the LOU context?

H2: Using the DCAT vocabulary to describe dataset metadata, client applications are able to easily get the resources in the LOU context.

With regard to datasets access and query, to advance on the LOU context means to incorporate novel approaches and technologies developed by the Semantic Web Community along with common interfaces. Query interfaces customized according to the client applications' needs can contribute to the automation of the client-server communication. We define the third research question and hypothesis related to datasets access and query as follows:

Q3: Can query interfaces customized according to the client applications' needs along with common interfaces improve datasets access and query in the LOU context?

H3: Query interfaces customized according to the client applications' needs along with common interfaces can improve the datasets access and query in the LOU context.

Regarding the incompleteness of the datasets, we define the fourth research question and hypothesis related to incomplete data in the LOU context as follows:

Q4: What kinds of information besides facts in the knowledge graph embedding models produce an acceptable trade-off in the LOU context to predict missing facts?

H4: Integrating textual descriptions and logical rules besides facts in the knowledge graph embedding models produces an acceptable trade-off in the LOU context to predict missing facts.

Considerations about how to test these previous hypotheses are described as follows in Section 5.

5 Evaluation

With the aim of finding out the most popular vocabularies/ontologies in the LOU context and testing the hypothesis H1, we carried out a state-of-the-art study. As outcomes, we found out: i) AIISO ontology[10] is applied in most works while other vocabularies such as Teach[11] and Courseware include crucial terms about university courses; ii) most works followed the principle of reusing existing ontologies as much as possible, however, all of them had to extend these vocabularies to fulfill their requirements; iii) general ontologies such as Dublin Core Metadata Element Set, FOAF, W3C Basic Geo Vocabulary, and W3C Ontology for Media Resources had a common usage along with the above vocabularies.

Unlike existing LOD university platforms that have been using the VoID vocabulary to describe their datasets, in our proposal, the DCAT vocabulary is promoted for advancing in the LOU context because of its broader coverage than other vocabularies like VoID. The explanation of the DCAT vocabulary usage (H2 evaluation) is shown below along with the customized query interface.

In order to evaluate the hypothesis H3, since all the state-of-the-art LOU platforms include a course dataset, we developed a customized query interface for the course dataset called coursesld_server[12]. We do not aim at replacing common triple interfaces, but using our proposal as a complementary query interface. University course data are a special segment of open data at the university given the public nature of the data[12,10]. With the increase of distance learning, there are a lot of course recommender applications that assist clients to find suitable courses.

Our LOD university platform publishes on its homepage the DCAT-catalog file where course recommender applications can find datasets about courses. By accessing the dataset subject property, client applications can distinguish one dataset among many other ones (as in this case, teach: Course is the wanted subject). Additionally, by getting to the dataset distribution, clients know where to retrieve the dataset (URL via distribution accessURL property) and how to retrieve it (file format via distribution mediaType property, e.g., in this case, one of the four options: JSON-LD, Turtle, N-quads, CSV). In order to support the automation of the client-server communication a Hydra API Documentation was also defined, which can be obtained via catalog-dataset-conformsTo property.

Normally, course recommender applications ask for courses about one specific topic that start on a certain date. That is why, coursesld_server interface offers the possibility of getting courses related to that topic sorted by the start date via URL templates. To deal with memory requirements and advancing on the data reuse, the interface splits the answer into fragments with no more than 100 courses per fragment. Course recommender applications are able to request the server each fragment through the Tree/Hydra vocabulary without human

[10] http://purl.org/vocab/aiiso/schema#

[11] http://linkedscience.org/teach/ns#

[12] https://github.com/yalopez84/coursesld_server

intervention. To complete the hypothesis H3 evaluation, a client application was developed called coursesld_client[13].

In regard to the evaluation of the graph completion model (H4 evaluation), a controlled experiment will be carried out. The tests will be first on at least two general knowledge graphs: DBpedia and Freebase, and then, on specific LOU datasets. Currently, the model is being developed.

6 Results

Besides the vocabularies/ontologies found in the state-of-the-art study, we put into consideration our semantic modeling process, which has benefited from previous works[3,8], and promotes the usage of the Linked Open Vocabularies Initiative (LOV[14]) to look for terms/vocabularies. Regarding H1 evaluation, we consider that it is true. If we look at the semantic model of our platform, we see AIISO vocabulary to represent the university structure and Tech vocabulary to describe the university course according to the tendency in the state of the art. Each time a platform is developed following this principle, such a university contributes to having a unified vocabulary or group of vocabularies in the LOU context.

Regarding the coursesld_server interface (H3 evaluation), its effectiveness is appreciated as an additional service for the exploitation of linked datasets. Since most developers are familiar with APIs with descriptions more than SPARQL endpoints over the RDF format, customized interfaces like the one presented in this work can serve to pave that issue. At the same time, we confirm the hypothesis H2 as the DCAT vocabulary allows to advance on the customized interfaces, federated queries, automation of client-server communication, and therefore the DCAT vocabulary allows to advance in the LOU context.

7 Reflection and future work

Future work aims at continuing with this research. Some of the previous hypotheses need further evaluation in order to reach the goal of this proposal: to advance on the LOU context. Given the wide range of topics at the university (research, academic programs, software production, internal places, energy and water consumption, etc.), adding new datasets to the LOD platform allows us to have a more complete knowledge graph. Afterward, we plan to extend the solution to other universities and evaluate the research contributions beyond the LOU context.

Acknowledgments

This research has been partially sponsored by VLIR-UOS Network University Cooperation Programme - Cuba. Special thanks go to:

[13] https://github.com/yalopez84/coursesld_client

[14] https://lov.linkeddata.es/dataset/lov/terms

– Erik Mannens. Ghent University. Research interests: Data Science. He received his Ph.D. degree in Computer Science Engineering (2011) at UGent.
– Pieter Colpaert Ghent University. Research interests: Linked Open Data, Transport. He received his Ph.D. degree in Computer Science Engineering (2017) at UGent.
– Hector R. González. University of Informatics Science. Research interests: Machine Learning. He received his Ph.D. degree in Computer Science (2019) at University of Havana.
– Yusniel Hidalgo Delgado. University of Informatics Science. Research interests: Semantic Web.

References

1. Colpaert, P.: Publishing transport data for maximum reuse. PhD Thesis, Ghent University (2017)
2. Colpaert, P., Llaves, A., Verborgh, R., Corcho, O., Mannens, E., Van de Walle, R.: Intermodal public transit routing using liked connections. In: International Semantic Web Conference: Posters and Demos. pp. 1–5 (2015)
3. Daga, E., dAquin, M., Adamou, A., Brown, S.: The open university linked data–data. open. ac. uk. Semantic Web **7**(2), 183–191 (2016), publisher: IOS Press
4. dAquin, M.: Putting Linked Data to Use in a Large Higher-Education Organisation. In: ILD@ ESWC. pp. 9–21. Citeseer (2012)
5. dAquin, M., Dietze, S.: Open education: A growing, high impact area for linked open data. ERCIM News,(96) (2014)
6. Keßler, C., Kauppinen, T.: Linked open data university of münster–infrastructure and applications. In: Extended Semantic Web Conference. pp. 447–451. Springer (2012)
7. Khan, H.: Towards more intelligent SPARQL querying interfaces. In: International Semantic Web Conference (2019)
8. Ma, Y., Xu, B., Bai, Y., Li, Z.: Building linked open university data: Tsinghua university open data as a showcase. In: Joint International Semantic Technology Conference. pp. 385–393. Springer (2011)
9. Meymandpour, R., Davis, J.G.: Ranking universities using linked open data. In: LDOW (2013)
10. Nahhas, S., Bamasag, O., Khemakhem, M., Bajnaid, N.: Added values of linked data in education: A survey and roadmap. Computers **7**(3), 45 (2018), publisher: Multidisciplinary Digital Publishing Institute
11. Sun, Z., Vashishth, S., Sanyal, S., Talukdar, P., Yang, Y.: A re-evaluation of knowledge graph completion methods. arXiv preprint arXiv:1911.03903 (2019)
12. Szász, B., Fleiner, R., Micsik, A.: A case study on Linked Data for University Courses. In: OTM Confederated International Conferences" On the Move to Meaningful Internet Systems". pp. 265–276. Springer (2016)
13. Verborgh, R., Vander Sande, M., Hartig, O., Van Herwegen, J., De Vocht, L., De Meester, B., Haesendonck, G., Colpaert, P.: Triple Pattern Fragments: a low-cost knowledge graph interface for the Web. Journal of Web Semantics **37**, 184–206 (2016), publisher: Elsevier
14. Wang, Q., Mao, Z., Wang, B., Guo, L.: Knowledge graph embedding: A survey of approaches and applications. IEEE Transactions on Knowledge and Data Engineering **29**(12), 2724–2743 (2017), publisher: IEEE