Universidad de Alcalá
Departamento de Ciencias de la Computación

DOCTORAL THESIS

Interlinking Educational Data to Web of Data

Author:
Enayat Rajabi

Supervisors:
Dr. Salvador Sanchez-Alonso
Dr. Miguel-Angel Sicilia

Alcalá de Henares, Spain
March 2015

Table of Contents

Abstract 3
Acknowledgments 4
1 State of the art 5
1.1 eLearning schemas 5
1.2 Linked Open Data datasets 6
1.3 Interlinking tools 8
2 Objectives 9
3 Methodology 11
3.1 Mapping the educational schema as RDF 11
3.2 Setting up an SPARQL endpoint or creating an RDF dump 11
3.3 Interlinking data to useful knowledge in the Web 11
4 Conclusions 13
4.1 Exposing eLearning metadata as Linked Open Data 13
4.2 Evaluating the Linked Data tools 14
4.3 Enriching the educational datasets 14
5 References 14

Abstract

With the proliferation of educational data on the Web, publishing and interlinking eLearning resources have become an important issue nowadays. Educational resources are exposed under heterogeneous Intellectual Property Rights (IPRs) in different times and formats. Some resources are implicitly related to each other or to the interest, cultural and technical environment of learners. Linking educational resources to useful knowledge on the Web improves resource seeking. This becomes crucial for moving from current isolated eLearning repositories towards an open discovery space, including distributed resources irrespective of their geographic and system boundaries. Linking resources is also useful for enriching educational content, as it provides a richer context and other related information to both educators and learners.

On the other hand, the emergence of the so-called "Linked Data" brings new opportunities for interconnecting different kinds of resources on the Web of Data. Using the Linked Data approach, data providers can publish structured data and establish typed links between them from various sources. To this aim, many tools, approaches and frameworks have been built to first expose the data as Linked Data formats and to second discover the similarities between entities in the datasets. The research carried out for this PhD thesis assesses the possibilities of applying the Linked Open Data paradigm to the enrichment of educational resources. Generally speaking, we discuss the interlinking educational objects and eLearning resources on the Web of Data focusing on existing schemas and tools.

The main goals of this thesis are thus to cover the following aspects:

- Exposing the educational (meta)data schemas and particularly IEEE LOM as Linked Data
- Evaluating currently available interlinking tools in the Linked Data context
- Analyzing datasets in the Linked Open Data cloud, to discover appropriate datasets for interlinking
- Discussing the benefits of interlinking educational (meta)data in practice

Keywords: Linked Data; Educational data; Interlinking; Tools; Schema.

Acknowledgments

Living in a foreign country with a family is hard enough as it is, specially with all the administrational issues you have to deal with. However, the problems can be overcome when you have surrounded by helpful people.

First and foremost, I offer my sincerest gratitude to Dr. **Salvador** Sanchez-Alonso, you have supported me from the beginning throughout my work with your understanding, patience, knowledge, and friendship. I am indebted to you for all you have done for me during these years.

I would also like to express my special appreciation to Professor Dr. **Miguel-Angel** Sicilia, you have been a tremendous mentor for me in my researches. Thanks for giving me this opportunity to apply in the University of Alcalá, for working on European projects, and for allowing me to grow as a researcher in your group. Your intelligent ideas on my publications have been really invaluable. I would especially like to thank Dr. **Elena** Garcia Barriocanal, you have been there to support me from the beginning.

I am also grateful to my referees, **Ivana** Marenzi and **Jad** Najjar, for revising and improving this thesis.

I would like to acknowledge my past and present colleagues in the **IERU laboratory** for their technical and non-technical helps.

Lastly, and most importantly, I would like to dedicate this thesis to my parents and family. I thank my parents, **Nasser** and **Fatemeh**, for their love, their encouragement, their faith in me and to be as ambitious as I wanted.

My special thanks go to my wife **Mahtab**, not only for her sacrifices over the years, but also for her patience, encouragement, and understanding in the most positive way during these years. She helped me a lot improve my English writing in my publications. I would also like to express my thanks to my little daughter **Bahar** for being such a kind girl always cheering me up.

1 State of the art

In the following sub-sections, we will outline the state-of-the-art of the use of digital educational resources and their context from the point of view of data interlinking. We will also indicate how our research can impact on each of those aspects.

1.1 eLearning schemas

Works related to the publication and evolution of eLearning metadata standards at the international level have been carried out by a number of organizations including the IEEE, the Dublin Core Metadata Initiative (DCMI), IMS Global, and ISO/IEC. In this context, IEEE LOM is an internationally-recognized open standard bound up with the history and development of the IMS eLearning interoperability specifications (e.g. IMS Content Packaging), and with the evolution of the ADL SCORM reference model, which supports the IEEE LOM standard alongside other specifications. Dublin Core (DC) has also been used in many systems and applications as an alternative to other metadata standards (e.g. IEEE LOM) or in combination with them to provide wider interoperability. Metadata for Learning Resources (MLR) is another recent effort in the ISO community, aimed at harmonizing LOM and Dublin Core metadata standards, as it tries to enable both the "learning object" aspects of LOM and the "entity-relationship" model of the Semantic Web associated with the Dublin Core Abstract Model. The Learning Resource Metadata Initiative (LRMI) has also developed a common metadata framework for describing learning resources on the Web. Although the goal of the last two schemas (MLR and LRMI) is to become a complement or alternative to IEEE LOM and DC, a wide variety of eLearning repositories or federations (notably the GLOBE federation) apply IEEE LOM as the base metadata schema and actively aggregate LOM records in a large scale.

Turning to the Linked Data aspect, there have been some initiatives to expose eLearning resource metadata as Linked Data. Dietze et al [1] [2] proposed an approach for linking educational resources based on the Linked Data principles using existing educational datasets and vocabularies; their aim was to exploit the wealth of existing technology-enhanced learning (TEL) data on the Web by exposing it as Linked Data within the mEducator project [3]. Other efforts and projects such as LinkedUp [4] and Linked Universities [5] have also been aimed at sharing data or metadata related to educational content based on the Linked Data principles. However, none of the mentioned studies pointed out the approach for exposing the IEEE LOM schema as Linked Data, which was initiated in 2000 in the context of the IMS Global Learning Consortium [6] (together with the ARIADNE Foundation [7]). This consortium developed both XML and RDF binding of LOM elements and, as a result, some RDF documents were produced as IMS RDF Bindings. The Dublin Core Metadata Initiative (DCMI) also proposed recommendations for expressing IEEE LOM as RDF in a

mapping document using an abstract model (DCAM) [8]. A mapping from IEEE LOM to RDF (defined by Nilsson et al. [9]) outlined the advantages of expressing learning object metadata as RDF. Nilsson also discussed some problems encountered in the process of producing the RDF binding of LOM elements and focused on some specific features of the binding, although this early work was discontinued and did not cover all the LOM elements.

Our work in this context continued and completed the Nilsson et al. approach [9] for exposing eLearning object metadata as Linked Data. To this aim, in a research [10] we examined the complete set of IEEE LOM elements along with its vocabularies and provided a complete mapping to RDF. We also implemented this approach on a large educational repository and linked its resources to the Linked Open Data cloud by following clearly established guidelines [11].

1.2 Linked Open Data datasets

Several studies have targeted the construction and analysis of the World Wide Web graph. Particularly, Deo & Gupta [12] investigated several graph models of the "traditional" Web and made use of graph-theoretic algorithms that help explain the growth of the Web. Serrano et al. [13] also analyzed the "traditional" Web by collecting content provided by Web crawlers and representing them in four different clusters, depending on the strength of their relationships. These works illustrated the data in different graphs and examined them according to size, node degree, and degree correlations. Likewise, in the LOD cloud there has been a considerable growth in the number of datasets, which are different in subject and size.

In order to provide a more precise depiction of the LOD network, we studied statistical measures from a Social Network Analysis (SNA) perspective as a way to analyze the LOD cloud from the point of view of a wide network of interactions between entities [14]. The linked datasets in such a social network, were mapped to nodes–i.e. entities–while the links between the datasets were reconsidered as the relationships between those entities. We later applied the SNA metrics to graphically highlight the LOD network, to spatially situate the datasets and to evaluate their properties (e.g., including dataset size, dataset relations to other datasets) from a mathematical perspective. Accordingly, we collected all the information about the LOD datasets from CKAN [15] by using a software component that exploited the links between the datasets. Gathering information about 337 datasets with almost 450 million links, we aligned all the collected information in a data matrix and used case-by-affiliation matrix, a general form of data matrices for social networks, in which the rows and columns refer to the LOD datasets and the values are the number of outgoing links of each dataset. The data was later imported in a SNA tool - NodeXL- to apply the SNA metrics with the aim to

recognize the central datasets in the LOD cloud, as these metrics allow enclosing the relevant correlations of a graph. Those metrics were:

- *Degree*: illustrates the number of datasets conjoined to the current node (dataset).
- *Betweenness Centrality (BC)*: when a dataset has a higher BC value, other datasets in the LOD cloud are connected to each other through it, which, in a way, implies that it plays an important role in the network.
- *Edge weight*: indicates the number of links between two datasets.

Table 1 shows the top five datasets with higher BC along with their degree values. We have regarded the LOD graph as a directed graph with the objective of better analyzing the relationships between datasets. Incoming degree values refer to the number of datasets that point to the current dataset, while the outgoing degree stands for the number of datasets pointed by the current dataset. As the table clearly illustrates, DBpedia had the highest BC, as it was part of the path in many paths between different datasets. The Geonames dataset was also the second dataset with a higher BC. It provides global place names, their location and some additional information; this is perhaps the reason why it was referred by 55 datasets and had no outgoing links.

Table 1: Top five datasets with high Betweenness Centrality

Dataset	In-Degree	Out-Degree	Betweenness Centrality
DBpedia	181	30	82,664.23573
Geonames	55	0	10,958.12104
DrugBank	8	12	7,446.525541
Bio2rdf-goa	11	8	3,751.965659
Ordance-survey	16	0	3,272.718991

Identifying the DBpedia dataset as the LOD hub in the mentioned research, we selected it as one of the main targets of our investigation for interlinking datasets.

Furthermore, in another study [11], we investigated many datasets in the LOD cloud to discover the available datasets in an educational domain. Examining the endpoints of datasets, we could select 20 datasets for our case study that responded to the queries or included an RDF dump to download. The selected collection was later used for interlinking an eLearning repository [11].

1.3 Interlinking tools

Creating links between datasets can be done manually, but this is a very time consuming task. In practice, data publishers would never perform interlinking of their resources relying only on human efforts, particularly when they maintain large amounts of data. On the other hand, they are motivated to apply interlinking tools to enrich their materials. The interlinking tools search for relationships between various datasets and discover the similarities by leveraging a number of matching techniques automatically or semi-automatically.

In a related research [16], we contrasted the existing solutions in the Linked Open Data context for interlinking, and compared them in Table 2. We considered the following measures in our work:

- Domain: The domain in which the tool is intended to be used (e.g., Multimedia, General, LOD)
- SPARQL endpoint/ RDF dump: Does the tool use an SPARQL endpoint or an RDF dump to perform the interlinking?
- Manual/Automatic: Can the user apply the tool to manually compare the entities (manual functioning) or should instead the software do the interlinking automatically (automatic mode)?
- Well-Documenting: Is the tool well documented? (e.g., user manual, etc.)
- Customization flexibility: Does the tool present facilities to customize the approach by changing some configuration parameters? Does it allow fine tuning for adapting the tool to specific purposes?

Table 2: Interlinking tools comparison

Tool	Domain	SPARQL/ RDF Dump	Manual/ Automatic	Well-documented/ frequently updated	Customization flexibility
GWAP	Multimedia	No	Manual	No	Unknown
LIMES	LOD	Yes	Automatic	Yes	Yes
LODRefine	General	Yes	Automatic	Yes	Partially
RDF-IA	LOD	RDF Dump	Automatic	No	Unknown
SAI	Multimedia	No	Automatic	No	Unknown
Silk	LOD	Yes	Automatic	Yes	Yes
UCI	LOD	Yes	Manual	No	Unknown

The word “unknown” in the table above indicates that the correspondent tool could not be tested due to the lack of documentation and application support from the original providers. As illustrated, Silk, LIMES, and LODRefine were found to be the most appropriate linking systems based on the assumed criteria. Accordingly, we examined these three interlinking tools in the mentioned research and found out that LODRefine, in comparison to other tools was not effective and scalable, particularly when the dataset is huge. A comparison of the tools also revealed that Silk and LIMES were the most promising framework in terms of finding the most amounts of valid matches between the datasets. A further explanation about LIMES follows, as this has been the principle tool applied more often in our studies:

Link Discovery Framework for Metric Spaces (LIMES) is a framework that implements a linking approach for discovering relationships between entities contained in Linked Data sources [17]. LIMES leverages several mathematical characteristics of metric spaces to compute pessimistic approximations of the similarity of instances. It then uses them to filter out a large amount of those instance pairs that do not satisfy the mapping conditions. Given a source, a target, and a link specification, LIMES processes the strings by making use of suffix-, prefix- and position filtering in a string mapper. The processing results of the string mapper (along with other types of mappers in the system) are filtered vas well as merged by using time-efficient set and filtering operations. As a result, LIMES generates links between items contained in two LD datasets via SPARQL Endpoint or RDF dump. Using a threshold in the configuration file, the user can set a value for various matching metrics. Two instances are considered as matched and are linked via a relation such as “owl:sameAs” when the similarity between the terms exceeds the defined threshold. Apart from LIMES’ diverse collection of functionalities, a recent study [17] evaluated it as a time-efficient approach, particularly when it is applied to link large data collections.

2 Objectives

A large amount of structured data on the Web is published according to the Linked Data principles [18]. This fact facilitates not only the sharing data, but also the availability of different kinds of information on the Web of Data. In an eLearning context, flagship projects such as Europeana, Linked Education, and LinkedUp have embraced the Linked Data approach to interlink educational resources in the Linked Open Data cloud. Generally speaking, they enrich their educational objects by connecting them to public or specific targets of knowledge. Such enrichment makes the objects more valuable as it connects them with useful knowledge on the Web. An enriched Web portal can thus provide suggestions to visitors and redirect a researcher to more interesting subjects in different languages in DBpedia when she explores the information in the portal.

Another relevant example can be taken from the agricultural context. A researcher might explore the contents of an organic portal (e.g., organic-edunet.eu) in order to find a specific resource. In one of the result resources, a video on the subject of organic farming catches the researcher's attention and thus follows the description in order to investigate its applied methods. The researcher has never come across a specific term (e.g., “abonos verdes”, green manure) as it has been provided in another language and does not yield any more relevant data to her. As the resources in this portal have been previously interlinked with datasets such as DBpedia, she is able to find more information on the topic including different translations.

The aforementioned discussion implies that providing links between eLearning resources and repositories plays an important role towards data enrichment, although it is a time consuming task and requires a lot of human effort. On the other side, the Linked Data applications have facilitated the enrichment by presenting several solutions for automatic and intelligent linking. Namely, the interlinking tools allow establishing links between different datasets by discovering similarities among their entities so that data publishers can leverage them to enrich their contents. This can be achieved by; a) selecting an appropriate approach for interlinking b) identifying the target dataset(s) and its entities, and c) evaluation of the outcomes. The objective of this PhD thesis was to investigate the Linked Data approach on an educational context from several perspectives including:

A. Schemas: How eLearning metadata schemas, specifically IEEE LOM, can be exposed as Linked Data? What elements in this schema are appropriate for interlinking?

To cover this objective, we carried out an analysis on the IEEE LOM elements and exposed them as Linked Open Data in practice [10]. In the mentioned study, we also evaluated the IEEE LOM elements to discover the best candidates (e.g., “keyword”) for interlinking purposes.

B. Tools: What kind of software and tools can be applied for interlinking educational datasets?

To this aim, we compared several interlinking tools in the Linked Data context and outstood two tools after evaluating their results by some human experts manually [16]. The outcome of this study was significant as it helped us to choose the most appropriate software for interlinking.

C. Datasets: What is the status of LOD datasets in terms of their use for the interlinking tasks? What educational datasets can be applied?

To do a content analysis of the LOD cloud, we examined the LOD datasets using social data analysis [14] and found out that notably the DBpedia dataset acts as a central hub of the Linked Data network and thus can be used as one of the important datasets for the interlinking purpose. We also investigated the LOD cloud to discover the most appropriate educational datasets for the sake of interlinking [11].

D. Exposing and interlinking advantages: what are the benefits of data exposing and interlinking when an eLearning collection is connected to the LOD datasets?

The advantages of publishing along with interlinking data in an educational context have been pointed out in most of our studies [10][11][16][21]. Generally speaking, demonstrating the feasibility of enrichment the eLearning resources when (meta)data is interconnected with useful knowledge on the Web, was the major achievement of the mentioned researches. Moreover, making the metadata available through an endpoint so that they can be leveraged by other datasets in the LOD cloud was another benefit that was obtained from our studies. Particularly, in [11] we indicated how an eLearning repository is enriched when it is connected to several educational datasets on the Web of Data.

3 Methodology

In practice, data publishers in the eLearning arena do not perform human-driven interlinking of their resources when they maintain large amounts of eLearning objects. On the other hand, they are motivated to leverage interlinking tools to enrich their contents. The question under discussion is how they can fulfill the process successfully. Although the methodology we aligned here focused on the educational domain, it can be applied to any context by introducing minor changes in some steps. The following statements will outline the different steps of this approach:

3.1 Mapping the educational schema as RDF

As mentioned earlier, the educational metadata are usually conformed to a given schema. To expose them as Linked Data, the data publishers map the useful metadata elements to RDF. Depending on the complexity of the metadata schema, some data providers will have to create an ontology so that the elements are properly represented in RDF. As an illustration, we performed a full conversion of IEEE LOM schema to RDF (at http://data.opendiscoveryspace.eu/ODS_LOM2LD/ODS_SecondDraft.html).

3.2 Setting up an SPARQL endpoint or creating an RDF dump

After mapping the metadata schema to RDF, the next step is setting up an SPARQL endpoint so that the data become accessible for other datasets based on the Linked Data principles [18]. In particular, an SPARQL endpoint allows the other datasets to query the data in the exposed collection. Creating a full RDF dump including all the exposed data is another approach that makes the educational materials available to the data consumers. The RDF dump creation and setting up an SPARQL endpoint can be carried out either by using some mapping services such as D2RQ or by writing a program to convert the data to RDF.

3.3 Interlinking data to useful knowledge in the Web

To enrich the eLearning resources on the Web, data publishers can carry out an interlinking process to link their data to useful information in the LOD cloud. Figure 1, illustrates the approach we proposed to interconnect an educational repository to some datasets on the Web. The most important part of an interlinking process is to proper configure a suitable tool to perform an effective matching process between the entities in the two datasets (source and target). The tool must discover the similarities between them taking into account the settings and entities that users specify in the configuration file. As a result, a set of matched resources is reported by the tools to be reviewed by the datasets stakeholders. To establish an interlinking process, the following parameters should be specified by users:

A. Source entities: which elements are going to be interlinked?

We illustrated [19] that not all the metadata elements are useful for interlinking, because some of them include customized values, controlled vocabularies, or numbers. Specifically, we recommended a set of IEEE LOM metadata elements which can be enriched in the interlinking process (e.g., title or keyword of learning objects).

B. Target entities: which target datasets and concepts are appropriate for enriching resources in an eLearning repository?

Identifying the target datasets in the LOD cloud is really important as it affects the interlinking results directly. To mention an example, we have shown [14] that DBpedia can be a good target as it has high Betweenness Centrality (BC) in the LOD network and thus the connected data to this dataset can be connected to other datasets as well. Apart from DBpedia, which includes general terms and subjects, data publishers can explore the LOD cloud datasets to discover the most appropriate and related targets to their topic. In another research [11], after exploring the LOD datasets in an educational domain and examining their endpoints, we picked up 20 educational datasets that included an active SPARQL endpoint and responded to the queries in a reasonable time. We applied the selected datasets for the interlinking purpose.

C. Other interlinking settings:

Apart from specifying the source and target datasets along with their entities, users can set a threshold for the string matching metrics by which two entities are considered as matched (called acceptance threshold). Users can also specify the type of String matching algorithm by which two text entities are compared.

After running the interlinking tool, the results are usually stored in a text file. These results include the matched concepts in the two datasets. Although the tool's output consists of a set of matched concepts according to a String similarity algorithm above the threshold, the domain experts should later review them to avoid any polysemy or ambiguity of the matched entities. In a specific research [16], we analyzed several interlinking tools and could outstand two of them (namely Silk and LIMES) as the most promising ones in the context of Linked Data.

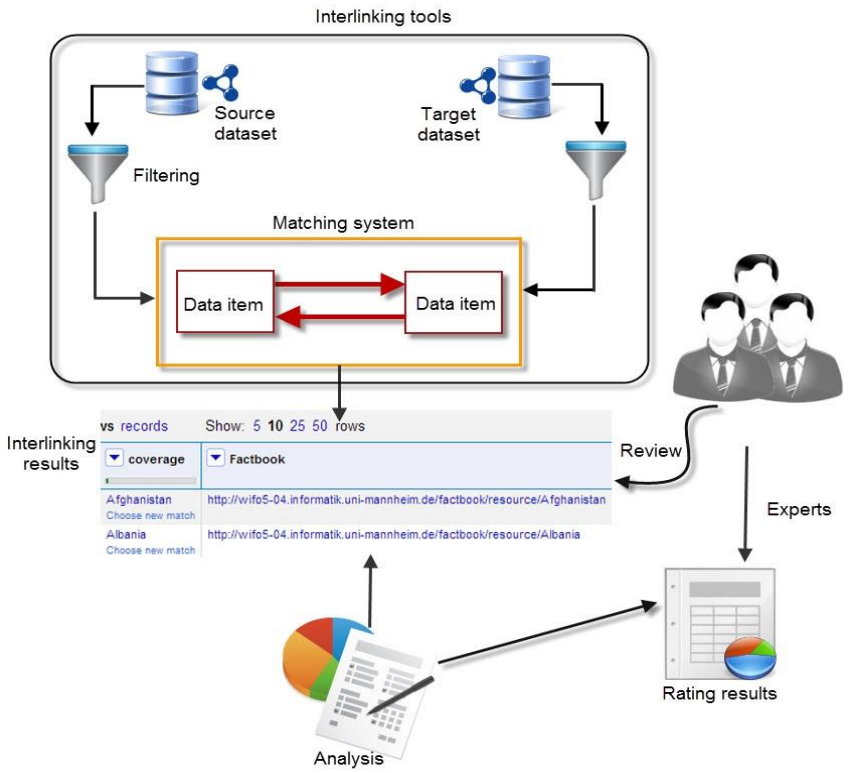


Figure 1: Interlinking approach

4 Conclusions

As mentioned earlier, interlinking the educational datasets was investigated from several perspectives. This section outlines the results from each point of view:

4.1 Exposing eLearning metadata as Linked Open Data

One of the main objectives of this work was exposing some eLearning metadata schemas as Linked Data formats. In a study [10], we presented a complete analysis on different strategies to map the IEEE LOM elements as RDF based on the Linked Data principles. From a technical point of view, we implemented the Linked Data exposure of several educational repositories (namely Organic.Edunet (http://data.organic-edunet.eu), Ariadne (http://ariadne.grnet.gr), Open Discovery Space (http://data.opendiscoveryspace.eu) as a proof of concept. Another study in this regard was performed to discover the usefulness of the IEEE LOM elements for interlinking. In [19], we showed that the following elements can be applied for linking IEEE LOM to the LOD cloud:

- Coverage: The time, culture, geography or region to which a eLearning resource applies (“General.Coverage”)
- Taxonomy: The classification term given to a eLearning resource (“Classification.Taxon”)

- Keyword: A keyword or phrase describing the topic of eLearning objects ("General.Keyword").

4.2 Evaluating Linked Data tools

As we evaluated in [16], several existing interlinking tools in the Linked Open Data cloud both theoretically and in practice can be applied to an educational context. As a result, various related tools were analyzed and consequently two tools namely, Silk and LIMES passed the presumed criteria, as the interlinking results were significant. The linking procedure in the mentioned research was carried out over a big educational repository (GLOBE [20]). The results generated by the tools were later presented to several experts for validation. The findings gathered via human evaluations revealed that applying the interlinking tools remarkably helps data publishers to connect their contents to useful datasets on the Web of Data, and thus we strongly recommend this approach. A high level of agreement among the raters of interlinking results also justified that using an interlinking tool is a reliable method for interlinking datasets when the threshold of matching concepts is high (e.g., more than 0.98).

4.3 Enriching the educational datasets

Linking an eLearning repository to several educational datasets in the LOD cloud leads to the content enrichment, as this approach links the eLearning objects to the related resources in different datasets on the Web [11]. Furthermore, one of the other benefits of an interlinking process is duplicate identification. Our examination on the interlinking results illustrated that several eLearning objects are published by different data providers on the Web, while they refer to the same learning resource. We carried out this identification by proposing a data model along with a workflow in which we contrasted the metadata elements retrieved from both datasets.

5 References

[1] S. Dietze, S. Sanchez-Alonso, H. Ebner, H. Q. Yu, D. Giordano, I. Marenzi, and B. P. Nunes, "Interlinking educational resources and the web of data: A survey of challenges and approaches," *Program*, vol. 47, no. 1, pp. 60–91, Feb. 2013.

[2] S. Dietze, H. Q. Yu, D. Giordano, E. Kaldoudi, N. Dovrolis, and D. Taibi, "Linked education: interlinking educational resources and the Web of data," in *Proceedings of the 27th Annual ACM Symposium on Applied Computing (SAC-2012).* Riva del Garda (Trento), Italy, 2012.

[3] P. D. Bamidis, E. Kaldoudi, and C. Pattichis, "mEducator: A Best Practice Network for Repurposing and Sharing Medical Educational Multi-type Content," in *Leveraging Knowledge for Innovation in Collaborative Networks*, pp. 769–776, 2009.

[4] "LinkedUp: Linking Web Data for Education," [Online]. Available: http://linkedup-project.eu/. [February 6, 2015].

[5] "Linked Universities."[Online]. Available: http://linkeduniversities.org. [February 6, 2015].

[6] “IMS Global Learning Consortium,” [Online]. Available: http://www.imsglobal.org/. [May 12, 2013].
[7] “The ARIADNE Foundation,” [Online]. Available: http://www.ariadne-eu.org/. [February 6, 2015].
[8] “Draft Recommended Practice for Expressing IEEE Learning Object Metadata Instances Using the Dublin Core Abstract Model ,” [Online]. Available: http://dublincore.org/documents/dcmi-terms/. [February 6, 2015].
[9] M. Nilsson, M. Palmer, and J. Brase, “The LOM RDF Binding-Principles and Implementation,” *Proceeding of Third Annual ARIADNE Conference.* Leuven, Belgium, 2003.
[10] E. Rajabi, M.-A. Sicilia, and S. Sanchez-Alonso, “Interlinking Educational Resources to Web of Data through IEEE LOM”. *Computer Science and Information Systems*, vol. 12, No. 1, pp. 233–255, 2015.
[11] E. Rajabi, M.-A. Sicilia, and S. Sanchez-Alonso, “Discovering Duplicate and Related Resources using Interlinking Approach: The case of Educational Datasets,” *Journal of Information Science*, first published on March 10, 2015 doi:10.1177/0165551515575922.
[12] N. Deo and P. Gupta, “Graph-Theoretic Analysis of the World Wide Web: New Directions and Challenges,” *Mathematica Contompornea, Sociedade Brasileira de Matemática*, vol. 25, pp. 49–69, 2003.
[13] M. Á. Serrano, A. Maguitman, M. Boguñá, S. Fortunato, and A. Vespignani, “Decoding the structure of the WWW: A comparative analysis of Web crawls,” *ACM Transactions on the Web*, vol. 1, no. 2, 2007.
[14] E. Rajabi, S. Sanchez-Alonso, and M.-A. Sicilia, “Analyzing Broken Links on the Web of Data: an Experiment with DBpedia,” *Journal of the Association for Information Science and Technology* (JASIST), vol. 65, no. 8, pp. 1721–1727, 2014 doi: 10.1002/asi.23109.
[15] “ckan - The open source data portal software,” [Online]. Available: http://ckan.org/. [February 6, 2015].
[16] E. Rajabi, M.-A. Sicilia, and S. Sanchez-Alonso, “An empirical study on the evaluation of interlinking tools on the Web of Data,” *Journal of Information Science*, vol 40, pp.637–648 2014, first published on June 11, 2014 doi:10.1177/0165551514538151.
[17] A. Ngonga and A. Sören, “LIMES - A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data,” *presented at the IJCAI*, 2011.
[18] C. Bizer, T. Heath, and T. Berners-Lee, “Linked Data - The Story So Far,” *International Journal of Semantic Web and Information System*, vol. 5, no. 3, pp. 1–22, 33. 2009.
[19] E. Rajabi, M.-A. Sicilia, and S. Sanchez-Alonso, “Interlinking Educational Data: an Experiment with GLOBE Resources,” presented at *the First International Conference on Technological Ecosystem for Enhancing Multiculturality*, Salamanca, Spain, 2013.
[20] “GLOBE | Connecting the World and Unlocking the Deep Web,” [Online]. Available: http://globe-info.org/. [February 6, 2015].
[21] E. Rajabi, M.-A. Sicilia, and S. Sanchez-Alonso, “Interlinking Educational Data: an Experiment with Engineering-related Resources in GLOBE,” *International Journal of Engineering and Education*, 2015. In press.

Publications:

Title: Analyzing Broken Links on the Web of Data: an Experiment with DBpedia

Publication:	Journal of the Association for Information Science and Technology (JASIST)
Authors:	Enayat Rajabi, Salvador Sanchez-Alonso, and Miguel-Angel Sicilia
Relationship with the global research objectives:	Analyzing the datasets in the Linked Open Data cloud
Impact factor:	2.230 (2013)
Date of submission:	29-May-2013
Date of acceptance:	15-Aug-2013
Date of publication:	August 2014

Analyzing Broken Links on the Web of Data: an Experiment with DBpedia

Enayat Rajabi

PhD candidate
Information Engineering Research Unit, Computer Science Department, University of Alcalá de Henares, Ctra. Barcelona km. 33.6, 28871 Alcalá de Henares, Spain. Email: *enayat.rajabi@uah.es*

Salvador Sanchez-Alonso

Associate professor
Information Engineering Research Unit, Computer Science Department, University of Alcalá de Henares, Ctra. Barcelona km. 33.6, 28871 Alcalá de Henares, Spain. Email: *salvador.sanchez@uah.es*

Miguel-Angel Sicilia

Full professor
Information Engineering Research Unit, Computer Science Department, University of Alcalá de Henares, Ctra. Barcelona km. 33.6, 28871 Alcalá de Henares, Spain. Email: *msicilia@uah.es*

Linked Open Data enables interlinking and integrating any kind of data in the Web. Links between various data sources play a key role as they allow software applications (e.g., browsers, search engines) to operate over the aggregated data space as if it was a unique local database. In this new data space, where DBpedia –a dataset including structured information from Wikipedia– seems to be the central hub, we analyzed and highlighted outgoing links from this hub in an effort to discover broken links. The paper reports on an experiment to examine the causes for broken links, and proposes some treatments for solving this problem.

Introduction

The Linked Data approach (Bizer, Heath, & Berners-Lee, 2009), as an innovative way of integrating and interlinking different kinds of information on the Web of Data, conjoins structured data in order to be utilized by machines. This approach extends sharing data via Uniform Resource Identifiers (URIs) so that institutions and data publishers can leverage it to link their data to useful external datasets (Fernandez, d'Aquin, & Motta, 2011). The Linked Open Data (LOD)[i] cloud is a set of databases from various domains that have been translated into RDF and linked to other datasets by setting RDF links between data sources. It nowadays conforms a huge collection of interlinked data aimed at improving search and discovery of related data on the Web. Specifically, Linked Data terms are served as information resources and addressed via URIs so that they become discoverable. The URI is a generic means to identify and describe not just digital entities (e.g., electronic documents, metadata), but also real world objects and abstract concepts (e.g., people, places). URIs are represented as *dereferenceable*, meaning that software agents and Linked Data client applications can look up the URI using the HTTP protocol and retrieve a description of the entity represented as RDF. URI references of links between data items are expected to be "cool"[ii] so that the target becomes accessible by the source. Hence, links between data items play an essential role in the LD approach (Popitsch & Haslhofer, 2010). In this context, broken links (Eysenbach & Trudel, 2005) become problematic, specially for data consumers which will not be able to access the desired resources. In the "traditional" Web, broken links in websites have negative effects on the search engine rankings, but in the Linked Data cloud, they may lead to splitting the linked datasets, which prevents machines to follow the URIs to retrieve further relevant data.

In this paper, we evaluate the phenomenon of broken links in the LOD cloud, focusing on the external links of DBpedia[iii], as a hub commonly used to browse and explore the Web of Data. For this purpose, we used a link checking engine, and studied the impact of broken links on the interlinked datasets from the DBpedia perspective. DBpedia extracts structured information from Wikipedia, interlinks it with other datasets and publishes the results using Linked Data conventions and SPARQL (Morsey et al., 2012).

The rest of this paper is structured as follows. Section 2 highlights the importance of link integrity in LOD and explains how the "broken link" phenomenon can negatively affect the consolidation of the Web of Data. In Section 3 we focus on the importance of the DBpedia dataset as our testing

case and we report measures of its broken links. Conclusions and outlook are finally provided in Section 4.

Problem statement and related work

Link integrity has always been a significant factor for discovering and exploring data in the Web (Davis, 1999). It aims to ensure validating a link, regardless if it points to a target inside a given dataset or to external datasets. Link integrity becomes especially important when data publishers interconnect their data to external data sources on the Web of Data. In particular, integrating links between datasets is usually established either automatically by using interlinking tools or manually by data publishers. Different types of interlinking tools e.g., User Contributed Interlinking (Hausenblas, Halb, & Raimond, 2008) build semantic links between datasets relying on user contributions such as matching two items manually, while some others e.g., RDF-IA (Scharffe, Liu, & Zhou., 2009), Silk Link Discovery Framework (Bizer, Volz, Kobilarov, & Gaedke, 2009), and LIMES (Ngonga Ngomo & Auer, 2011) work automatically according to the directions provided by the user configuration.
Link integrity between data sources becomes defective when a link target is deleted or moved to a new location, making it impossible that data browsing tools (e.g., search engines) follow the links to reach the target data. Furthermore, broken links annoy human end-users and force them to either not use the contents of data provider or to manually look up the intended target using a search engine.
Several causes can produce broken links in a dataset, including the following:

- The data source points to targets that do not exist anymore. In this case, the request is answered, but the specific resource cannot be found.
- The source that hosts the target data stops working or redirects to a new location.
- Authoring issues on the target side block the access to the linked resource, thus preventing to reach the desired data.
- The target responds but does not return the data fast enough and the browser times out.
- Human errors, e.g., misspelling the link, something that mostly occurs when interlinking is carried out manually.

Preventing broken links by data publishers is the simplest and preferable method to fix the problem. Data providers can identify the external links manually at the time of publishing datasets. This approach is applicable where data providers can maintain and monitor all internal and external links. In a decentralized Linked Open Data system, this would be impossible, as many external links are controlled by other data publishers.

Several studies have attempted to identify and remedy the broken link problem. Vesse, Hall & Carr (2010) introduced an algorithm for retrieving linked data about a URI when the URI is not resolvable. Vesse also proposed a system which allows users to monitor and preserve linked data they are interested using the expansion algorithm. Popitsch & Haslhofer (2010) presented DSNotify, a tool for detecting and fixing broken links, which can keep links between LOD datasets consistent with required user input. Lui and Li (2011) proposed an approach which relies on the metadata of data sources to track the data changes by capturing the modifications of the data in real time, adjusting notification timing with different requirements from the data consumers. However, none of these studies has been done in the LOD context, in an attempt to estimate the impact and potential causes for the problem.

In terms of Web analysis, some studies have been also conducted in the context of constructing and analyzing of World Wide Web graph. Deo & Gupta (2003) investigated several graph models of the "traditional" Web and made use of graph-theoretic algorithms that help explaining and structuring the growth of the Web. Serrano et al. (Serrano, Maguitman, Boguñá, Fortunato, & Vespignani, 2007) also analyzed the "traditional" Web by collecting content provided by web crawlers and representing them in four different clusters, depending on the strength of their relationships. They illustrated the data in different graphs and examined them according to size, node degree and degree correlations.

Experimental setting

There has been a considerable growth of datasets, which are different in subject and size. In order to provide a more precise depiction of LOD datasets, we studied statistical measures from a graph perspective. The measures are taken from the toolkit of Social Network Analysis (SNA), as a way to analyze the LOD cloud from the perspective of a wide network of interactions between entities. However, these measures are not exclusive of SNA but they were also used in previous graph analysis research of the World Wide Web. Linked datasets, in such a social network, were mapped to nodes – i.e. entities– while links between datasets were reconsidered as relationships between those entities.

There exists a wide range of SNA metrics which allow researchers to integrate and analyze mathematical and substantive dimensions of a network structure formed as a result of ties formed between persons, organizations, or other types of nodes (Wasserman, 1994) (Scott, 2000). Some of these metrics are:

- *Betweenness Centrality* (BC), which measures how often a node appears on the shortest path between two other nodes. High betweenness nodes are usually key players in a network or a bottleneck in a communication network. Thus, it is used for detecting important nodes in graphs.
- *Degree*, the count of connections a node has with other nodes including self-connections. It is the most common topological metric in networks.
- *Edge weight*, a number assigned to each edge that represents how strong the relationship between two nodes in a graph is.

We made use of SNA metrics to graphically highlight LOD network, to get insights of the arrangement of LOD datasets and to evaluate their properties from a mathematical

perspective. In consequence, we collected all the information about LOD datasets from CKAN [iv] by using a software component that exploited the links between datasets. Gathering information about 337 datasets with almost 450 million links, we aligned all the collected information in a data matrix. It should be noted that the collected data was curated by the maintainers of the datasets, and thus it can be regarded as a reliable estimation.

We used case-by-affiliation matrix (Wasserman, 1994), a general form of data matrices for social networks, in which the rows and columns refer to LOD datasets and the values are the number of outgoing links of each dataset. The data were imported in a SNA tool, in this case NodeXL [v] (Hansen, Shneiderman, & Smith, 2010), following SNA metrics applied to recognize the central datasets in the LOD, as these metrics appear to enclose the relevant correlations of a graph.

- *Betweenness Centrality (BC)*: If a dataset has a high BC value, then many datasets are connected through it to others, which implies that the dataset plays an important role in the LOD cloud.
- *Degree*: illustrates the number of datasets that are conjoined to the current node (dataset).
- *Edge weight*: illustrates the number of links between two datasets.

Table 1 illustrates the top five datasets with higher BC along with their degree values. As the table shows, LOD graph has been regarded as a directed graph. Incoming degree values refer to the number of datasets that point to the current dataset, while the outgoing degree stands for the number of datasets pointed to by the current dataset. DBpedia, as it has illustrated in the table, shows the highest BC, as it is in the middle of many paths between other datasets. *Geonames* provides global place names, their location and some additional information such as population. Hence, it includes global information that referred by 55 datasets and has no outgoing links; nevertheless, it is the second dataset with high BC.

Table 1: Top five datasets with high betweenness centrality

Dataset	In-Degree	Out-Degree	Betweenness Centrality
DBpedia	181	30	82,664.24
Geonames	55	0	10,958.12
DrugBank	8	12	7,446.53
Bio2rdf-goa	11	8	3,751.97
Ordance-survey	16	0	3,272.72

With all the information extracted, we represented the data in NodeXL, an open-source template for Microsoft Excel that facilitates to explore network graphs. Once we had all datasets in NodeXL, we filtered out those with less than 2 incoming links (226 datasets out of 337) to depict the final graph in a more understandable way. Figure 1 illustrates the generated LOD graph using a "Harel-Koren Fast Multiscale" layout. As the figure shows, DBpedia is in the centre of the LOD cloud, thus acting as a hub in the network. This was our main motivation to go deeper in the analysis of DBpedia instead of analyzing other datasets. In fact, we selected DBpedia as our case study only after a careful graph analysis and examination of its external links, which finally persuaded us that this could be properly considered the central dataset of the LOD cloud. Figure 1 also shows why some datasets such as *Geonames* and *Drugbank* have high BC among the LOD datasets, as either they have been in the middle of a path or they were pointed by other datasets.

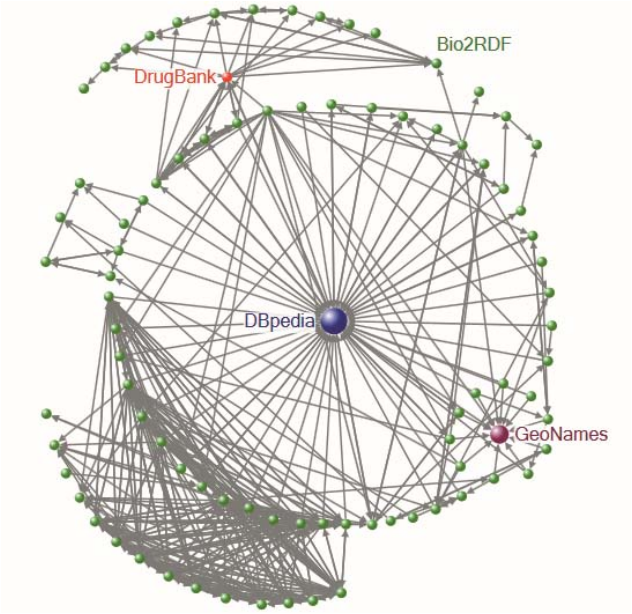


Figure 1: DBpedia as a hub in the LOD cloud graph

NodeXL also allows experts clustering a graph in different groups (Hansen et al., 2010) and we used this feature to confirm that DBpedia was indeed the central dataset in comparison to grouped datasets.

As it is well known, the DBpedia dataset contains structured information extracted from Wikipedia with the aim of making this information available on the Web of Data, as well as linking to other datasets. It includes structured data about persons, places and organizations which features labels and abstracts for 10.3 million unique things in 111 different languages (Bizer, Lehmann, et al., 2009); the full DBpedia dataset features almost 36 million data links[vi] to external RDF datasets. Links to some datasets such as the Flickr wrappr[vii] and Freebase [viii] have been automatically created by using software tools like Silk (Bizer, Volz, et al., 2009). According to the number of links between DBpedia and other datasets in the LOD cloud, the Flickr wrappr is the first dataset with more than 31 million links to DBpedia, while Freebase is the second one with 3.6 million links. Apart from these two datasets, Figure 2 illustrates the percentage of DBpedia outgoing links where "Others" in the figure (with 3%) refers to the datasets included less than 10 thousand links with DBpedia.

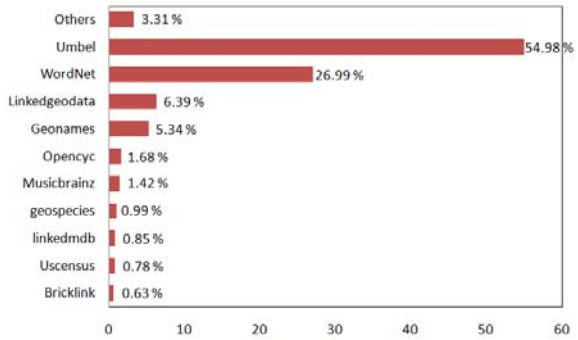


Figure 2: DBpedia outgoing links

As mentioned earlier, several problems can cause a broken link, all of which must be carefully checked. To examine the availability of the links, we programmed a link checker component that retrieved the HTTP response headers of the URLs. Particularly, a primarily broken link will return a HTTP 400 or 404 codes, indicating that there is an error with the target and thus it is unreachable. We clustered all the HTTP responses into several groups, such as "*server is unreachable*", "*time out*", and "*non existing record*". The workflow in Figure 3 shows how the tool checks every outgoing link of the DBpedia. The results afterwards were inserted into a database to be later evaluated. In addition to validating a URL, the software set the response timeout to 10 seconds, which means that the lack of any server activity of the target for this duration was considered to be an error.

The problem with this approach to link checking is that it may happen that some target data cannot be fetched at the time of the request, but the same data might be available again in future, which requires to periodically run the link checking component. We examined the availability of over 1.67 million links of DBpedia based on the schedule presented in Table 2. This schedule was started on January 2013 and followed over 4 months in order to analyze the links precisely.

Table 2: Link checking schedule

Month	Link checking dates
January	8th, 18th, and 28th
February	1st, 11th, and 28th
March	4th, 14th, and 24th
April	3rd, 13th, and 23rd

We filtered out both the *Flickr wrappr* and *Freebase* datasets as they include millions of links to DBpedia, which caused the link checking process to become too time-consuming for our computational capabilities. The target time of response was also checked, assuming that the link was live if the target responded before 10 seconds. We did not discover any authorization problem among the analyzed URLs, as some links may be unreachable due to security reasons.

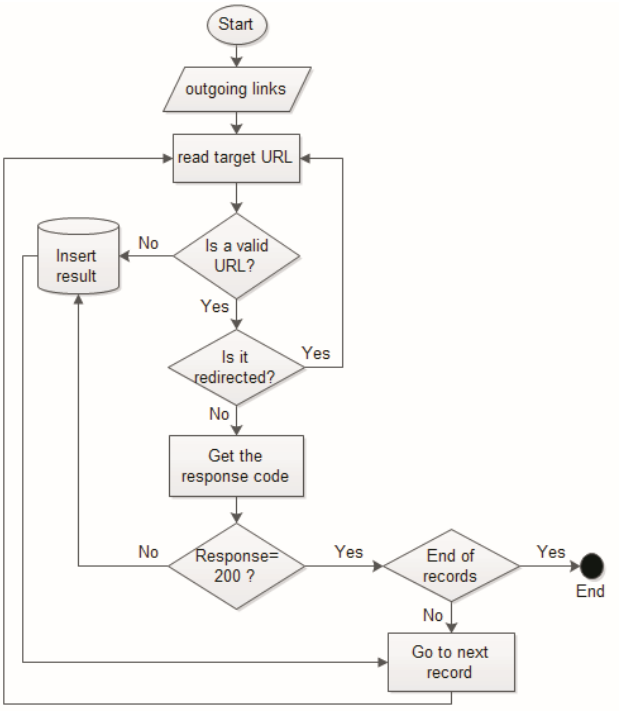


Figure 3: Link checker workflow

Table 3 illustrates the external datasets of the DBpedia along with the number of links and the average number of broken links detected during the scheduled process. The forth column in the table shows the average number of broken links relative to the dataset size (illustrated per 100,000 triples). We listed more detailed information about the datasets in Appendix 1. Table 3 shows how, for example, the *Italian_Public_school* dataset comprises a total of 169,000 triples (around 1.69 triples per 100,000) of which 5,822 were broken. The number 1,148 in the table (dividing 5,822 into 1.69) shows the status of the dataset from the availability perspective. *Diseasome*, as another example, is supposed to be very problematic as it has only 91,000 triples, 2,301 of which were not reachable through DBpedia.

Table 3: DBpedia related datasets

Dataset	Links number	Average # of broken links	Average # related to dataset size
Revyu	6	0	0
GHO	196	0	0
BBCwildlife	444	0	0
Amsterdam Museum	627	0	0
Openei	678	0	0
Dbtune - musicbrainz	838	0	0
Eunis	3,079	0	0
linkedmdb	13,758	0	0
Bricklink	10,090	1	0
Uscensus	12,592	1	0
Bookmashup	8,903	4	0
WordNet	437,796	6	0
Eurostat	490	15 (3%)	0
geospecies	15,974	41	2
Factbook	545	5	13
Nytimes	9,678	55	16
Dailymed	894	150 (17%)	91
TCM	904	151(17%)	128
Gadm	1,937	163 (8%)	Unspecified [ix]
DBLP	196	196 (196%)	1
Wikicompany	8,348	199 (2%)	Unspecified
Cordis	314	314 (100%)	4
Geonames	86,547	336	0
Umbel	891,822	1005	210
Italian_public_schools	5,822	1,940 (33%)	1,148
Gutendata	2,511	2,100 (84%)	21,000
Diseasome	2,301	2,301 (100%)	2,529
DrugBank	4,845	4,745 (98%)	619
Musicbrainz	22,980	22,980 (100%)	38
Opencyc	27,107	27,107 (100%)	1,694
Linkedgeodata	103,618	37,791 (36%)	175

Figure 4 shows the number of broken outgoing links in DBpedia by dataset (only datasets with more than 100 links).

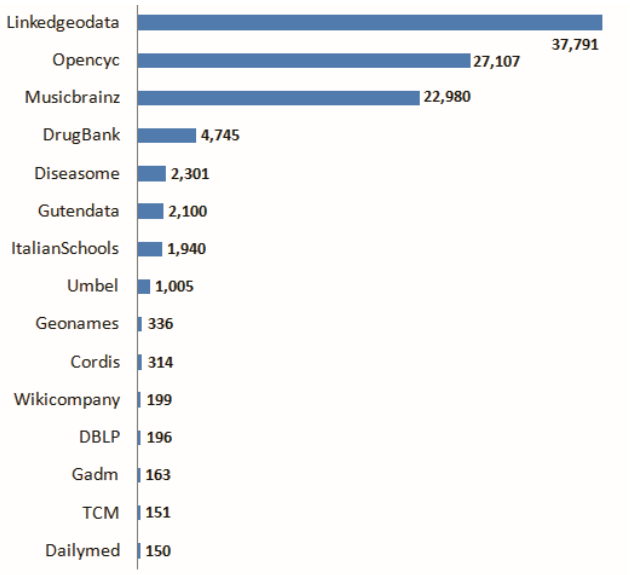


Figure 4: DBpedia broken outgoing links

An analysis was later carried out on the logs of the link checker with the aim of discovering the implications of the broken links problem. Figure 5 shows that more than 55% of the external links identified as broken were due either to the fact that the service exposing the dataset was down, or the server was not reachable. Nearly 32% of the total amount of external links referred to the targets that did not return the data in 10 seconds and the browser timed out. Furthermore, more than 10% of DBpedia links pointed to records that did not exist in the target dataset. Finally, the related services of a small percentage of the broken links (around 2%) were temporarily unavailable.

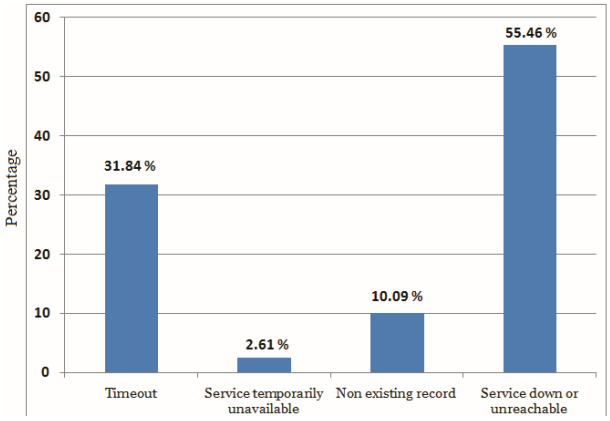


Figure 5: percentage of causes of broken links

Table 4 also illustrates each dataset along with the type of the error we faced during the links checking.

Table 4: DBpedia related datasets

Dataset	Cause of failure
Geospecies	Non existing record
Nytimes	Non existing record
Dailymed	Service temporarily unavailable
TCM	Timeout
Gadm	Service is down or unreachable
DBLP	Service temporarily unavailable
Wikicompany	Non existing record
Cordis	Timeout
Geonames	Non existing record
Umbel	Non existing record
Italian_public_schools	Service is down or unreachable
Gutendata	Service is down or unreachable
Diseasome	Service temporarily unavailable
Musicbrainz	Service is down or unreachable
Opencyc	Timeout
DrugBank	Timeout
Linkedgeodata	Non existing record / Service is down or unreachable

Conclusions and outlook

By evaluating the results of our link checking system over DBpedia, we classified the related datasets in different groups:

- Live and fully accessible datasets (through DBpedia links), such as *Openei* and *Eunis*.
- Datasets that were only partially reachable and which include links those were not accessible through DBpedia. In particular, some data did not exist in the external data sources anymore (e.g., *Wikicompany* and *Geonames*)
- Datasets which were fully broken. Our analysis shows that the hosts of these datasets were not reachable either temporarily or permanently (e.g., *Cordis* and *Musicbrainz*).
- Datasets which were not reachable during a certain period of time. Most often, the link checker could successfully access the related host(s) during a posterior checking process (e.g., *TCM, Dailymed*).
- Some datasets (e.g., *opencyc*, *Diseasome*) did not provide a linked data access including publishing dereferenceable URIs, a SPARQL endpoint, and RDF dump.
- A number of datasets (e.g., *Musicbrainz*) did not provide support, as it seems they were the result of research projects working on a voluntary or project-bound basis (e.g., individuals, and universities). Given the way they were managed, it is uncertain whether they will continue operating in the long term or at least providing free access services.

A manual evaluation of the links found that all the broken links of the *Umbel* dataset belonged to one URL. In particular, only one link in the target dataset was unreachable through around 1,000 links. In addition and with regard to what Table 1 illustrates, the *DrugBank* dataset, which had a high BC value among other datasets, even though around 98% of the outgoing links to this dataset were broken.

With respect to the results, we examined the most common current approaches for dealing with broken links in the LOD datasets. Data publishers can fix the problem automatically by using a link checking component. A link checker can be also applied to a data source to periodically detect and fix the broken links. Data consumers can also report manually the broken links to the data providers for them to resolve this problem. This solution is ineffective and slow, though. Other solutions such as the handle system (Sun, 2001) or PURL provide a number of services for unique and permanent identifiers of digital objects when web infrastructure has been changed and enable data publishers to store identifiers of arbitrary resources. Data providers can utilize those services to resolve identifiers into the information necessary to access, to later authenticate and update the current state of the resource without changing its identifier. This provides the benefit to allow the name of the item to persist over changes of e.g. location.

The research presented here in can be extended by going further in checking logs and examining the links manually. Specifically it can be analyzed for those datasets redirected to another target in terms of their availability and cause of the redirection.

There is also a wide variety of datasets published in the LOD cloud. Similar to outgoing links of the DBpedia, link checking could be extended to other datasets as well. In particular, the whole status of the LOD datasets, in terms of link availability, can be achieved by examining all the outgoing links for each dataset. As a result, the LOD cloud could be traced in the case of broken links and a central reposting system, for example as a part of LOD stat portal, can help datasets to fix their broken links.

Acknowledgements

The work presented in this paper has been part-funded by the European Commission under the ICT Policy Support Programme CIP-ICT-PSP.2011.2.4-e-learning with project No. 297229 “Open Discovery Space (ODS)”, CIP-ICT-PSP.2010.6.2- Multilingual online services with project No. 27099 “Organic.Lingua”, and INFRA-2011-1.2.2-Data infrastructures for e-Science with project No. 283770 “AGINFRA”.

References

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems, 5(3), 1–22.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., & Hellmann, S. (2009). DBpedia - A crystallization point for the Web of Data. Web Semant., 7(3), 154–165.

Bizer, C., Volz, J., Kobilarov, G., & Gaedke, M. (2009). Silk - A Link Discovery Framework for the Web of Data. In the 18th International World Wide Web Conference, Madrid, Spain.

Davis, H. C. (1999). Hypertext link integrity. ACM Comput. Surv., 31(4es).

Deo, N., & Gupta, P. (2003). Graph-Theoretic Analysis of the World Wide Web: New Directions and Challenges. Mathematica Contompornea,Sociedade Brasileira de Matemática,, 25, 49–69.

Eysenbach, G., & Trudel, M. (2005). Going, Going, Still There: Using the WebCite Service to Permanently Archive Cited Web Pages. Journal of Medical Internet Research, 7(5), e60.

Fernandez, M., d'Aquin, M., & Motta, E. (2011). Linking data across universities: an integrated video lectures dataset. In Proceedings of the 10th international conference on The semantic web - Volume Part II (pp. 49–64). Berlin, Heidelberg: Springer-Verlag.

Hansen, D., Shneiderman, B., & Smith, M. A. (2010). Analyzing Social Media Networks with NodeXL: Insights from a Connected World (1st ed.). Morgan Kaufmann.

Hausenblas, M., Halb, W., & Raimond, Y. (2008). Scripting User Contributed Interlinking. In Proceedings of the 4th workshop on Scripting for the Semantic Web (SFSW2008), co-located with ESWC2008.

Liu, F., & Li, X. (2011). Using Metadata to Maintain Link Integrity for Linked Data. In Proceedings of the 2011 International Conference on Internet of Things and 4th International Conference on Cyber, Physical and Social Computing (pp. 432–437). Washington, DC, USA: IEEE Computer Society.

Morsey,M. Lehmann, J., Auer, S., Stadler, C. and Hellmann, S. (2012) DBpedia and the live extraction of structured data from Wikipedia, Program: electronic library and information systems, 46(2), pp.157 - 181

Ngonga Ngomo, A.-C., & Auer, S. (2011). LIMES — A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data. In Twenty-Second International Joint Conference on Artificial Intelligence.

Popitsch, N. P., & Haslhofer, B. (2010). DSNotify: handling broken links in the web of data. In Proceedings of the 19th international conference on World wide web (pp. 761–770). New York, NY, USA: ACM.

Scharffe, F., Liu, Y., & Zhou., C. (2009). RDF-AI: an architecture for RDF datasets matching, fusion and interlink. In Proceedings of IJCAI 2009 IR-KR Workshop.

Scott, J. (2000). Social Network Analysis: A Handbook. SAGE publication.

Serrano, M. Á., Maguitman, A., Boguñá, M., Fortunato, S., & Vespignani, A. (2007). Decoding the structure of the WWW: A comparative analysis of Web crawls. *ACM Trans. Web*, *1*(2).

Sun, S. (2001). Establishing persistent identity using the handle system. In Proceedings of the Tenth International World Wide Web Conference.

Wasserman, S., & Faust, K. (1994). Social network analysis: Methods and applications. New York: Cambridge University Press.

Vesse, R., Hall, W., & Carr, L. (2010). Preserving Linked Data on the Semantic Web by the application of Link Integrity techniques from Hypermedia. In, Linked Data on the Web (LDOW2010), Raleigh, NC.

Appendix 1

Dataset name	URL	Subject
Revyu	http://revyu.com/	review and rate things
Gho	http://gho.aksw.org/	statistical data for health problems
BBC WildLife	http://www.bbc.co.uk/wildlifefinder/	Nature
Amsterdam Museum	http://semanticweb.cs.vu.nl/lod/am/	Culture
OpenEI	http://en.openei.org/	energy information
Dbtune	http://dbtune.org/musicbrainz/	Music
Eunis	http://eunis.eea.europa.eu	biodiversity
LinkedMDB	http://linkedmdb.org/	Movie
Bricklink	http://kasabi.com/dataset/bricklink	Marketing
Uscenus	http://www.rdfabout.com/demo/census/	Population statistics
Bookmashup	http://www4.wiwiss.fu-berlin.de/bizer/bookmashup/	Book
Factbook	http://www4.wiwiss.fu-berlin.de/factbook/	Countries
WordNet	http://www.w3.org/TR/wordnet-rdf	lexical database of English
Eurostat	http://eurostat.linked-statistics.org/	European Statistics
geospecies	http://lod.geospecies.org/	GeoSpecies
Nytimes	http://data.nytimes.com/	News
Dailymed	http://www4.wiwiss.fu-berlin.de/dailymed/	Drugs
TCM	http://code.google.com/p/junsbriefcase/wiki/TGDdataset	medicines
Gadm	http://gadm.geovocab.org/	GIS
DBLP	http://www4.wiwiss.fu-berlin.de/dblp/	Book
Wikicompany	http://wikicompany.org/	business
Cordis	http://www4.wiwiss.fu-berlin.de/cordis/	EU programmes and projects
Geonames	http://www.geonames.org/ontology/	geography
Umbel	http://umbel.org/	technology and semantics
Italian_public_schools	http://www.linkedopendata.it/datasets/scuole	schools
Gutendata	http://www4.wiwiss.fu-berlin.de/gutendata/	ebook
Diseasome	http://www4.wiwiss.fu-berlin.de/diseasome/	disease
DrugBank	http://wifo5-03.informatik.uni-mannheim.de/drugbank	Drugs
Musicbrainz	http://zitgist.com/	Music
Opencyc	http://sw.opencyc.org/	diverse collection of real-world concepts in OpenCyc
Linkedgeodata	http://linkedgeodata.org/	geography

Endnote

[i] *http://lod-cloud.net/ (Retrieved 2013-06-22)*
[ii] See *http://www.w3.org/Provider/Style/URI*
[iii] *http://dbpedia.org/About (Retrieved 2013-06-22)*
[iv] *http://ckan.org/ (Retrieved 2013-06-22)*
[v] *http://nodexl.codeplex.com/ (Retrieved 2013-06-22)*
[vi] *http://wiki.dbpedia.org/About (Retrieved 2013-06-22)*
[vii] *http://wifo5-03.informatik.uni-mannheim.de/flickrwrappr/ (Retrieved 2013-06-22)*
[viii] *http://www.freebase.com (Retrieved 2013-06-22)*
[ix] *The size of the dataset was unavailable both in the LOD database and through the provider's website*

Title: An Empirical Study on the Evaluation of Interlinking Tools on the Web of Data

Publication:	Journal of Information Science (JIS)
Authors:	Enayat Rajabi, Miguel-Angel Sicilia, and Salvador Sanchez-Alonso
Relationship with the global research objectives:	Evaluating the interlinking tools
Impact factor:	1.087 (2013)
Date of submission:	20-Jan-2014
Date of acceptance:	09-Mar-2014
Date of publication:	June 2014

Article

JIS

Journal of Information Science
2014, Vol. 40(5) 637–648

DOI: 10.1177/0165551514538151
jis.sagepub.com
SAGE

An empirical study on the evaluation of interlinking tools on the Web of Data

Enayat Rajabi
Information Engineering Research Unit, Computer Science Department, University of Alcalá, Spain

Miguel-Angel Sicilia
Information Engineering Research Unit, Computer Science Department, University of Alcalá, Spain

Salvador Sanchez-Alonso
Information Engineering Research Unit, Computer Science Department, University of Alcalá, Spain

Abstract
The rise and widespread use of Linked Data has encouraged data providers to publish and link their content in order to classify and organize information in a useful fashion. Interlinking between datasets enhances data navigation and facilitates searching. As a result, the use of interlinking tools as a way of connecting data items to the Linked Open Data cloud has become more prevalent. In this paper, we examine the results obtained by three interlinking tools used to link a large educational collection to the Linked Open Data datasets. The generated output by the interlinking tools, which was later assessed by human experts, illustrates that data publishers can rely on current interlinking approaches and thus adopt them to connect their resources to the Web of Data. Our findings also provide evidence that two of these tools, namely Silk and LIMES, can be considered as the most promising.

Keywords
Evaluation; interlinking tool; Linked Data

1. Introduction

In recent years, the Linked Data approach [1] has facilitated the availability of different kinds of information on the Web of Data. The view that information objects are discovered and shared is very much in line with the goals of the Semantic Web. Needless to say, the core of data accessibility throughout the Web is the links between items. This idea is prominent in literature on Linked Data principles [1]. Indeed, providing links between objects in a dataset, or among the elements in the Linked Open Data (LOD) cloud [2], is based on the assumption that the Web is migrating from a model of isolated data repositories to a Web of interlinked datasets. One advantage of data connectivity using RDF links [3] is the possibility of linking a resource to valuable collections on the LOD cloud. In particular, a data source is enriched when its content is connected to several datasets (e.g. geography, places, and science). A notable example can be found in e-learning. Linking educational resources from different repositories to useful knowledge on the Web enables sharing as well as navigation of learning objects. Searching becomes more effective, as many learning resources are implicitly related to the interests as well as the cultural and technical environment of learners.

Another relevant example can be taken from the agricultural context. A researcher might explore the contents of an organic portal (e.g. organic-edunet.eu) in order to find a specific resource. In one of the result resources, a video on the subject of organic farming catches the researcher's attention and she thus follows the description in order to investigate its applied methods. The researcher has never come across a specific term (e.g. abonado en verde) as it has been provided in another language and does not yield any more relevant data. As the resources in this portal have been previously

Corresponding author:
Enayat Rajabi, Information Engineering Research Unit, Computer Science Department, University of Alcalá, 28805 Alcalá de Henares, Spain.
Email: enayat.rajabi@uah.es

interlinked with datasets such as DBpedia, she is able to find more information on the topic, including various translations. Furthermore, the learning resource includes several organic keywords linked to the AGROVOC thesaurus [4], which allows her to be connected to a multilingual resource with around 40,000 terminologies.

To this end, some educational institutions (e.g. University of Muenster,[1] Open University of the UK[2]) have made their learning resources available as Linked Data, linking them to general or specific information on the LOD cloud. Moreover, several projects, such as Europeana,[3] LinkedUp [5] and Linked Education [6], have embraced the Linked Data approach and aim to link learning (meta)data to educational datasets. The DBpedia[4] dataset is now considered a central hub among the LOD datasets [7] and one of the most significant. It allows the connection of almost any type of data source to 12.6 million objects in 119 different languages [8] as well as other relevant datasets on the Web of Data. To illustrate the issue of data connectivity, one could for example point to a researcher in DBpedia and obtain her list of publications in DBLP[5] or obtain definitions and roots in DBpedia from a vocabulary in a special domain. The foregoing discussion implies that RDF links play an important role in interlinking objects from various data sources.

Creating links between datasets can be done manually, but this is a time-consuming task. Several interlinking tools address this issue by automatically or semi-automatically finding links. Most of these tools search for relationships between various datasets and discover similarities by leveraging a number of matching techniques. While it is mostly agreed amongst dataset owners that interlinking tools are useful in terms of matching concepts to the LOD cloud [6, 9], the question under discussion is how and to what extent data consumers can rely on their outcomes.

In practice, data publishers do not perform interlinking of their resources, relying on human effort when they maintain large amounts of data. On the other hand, they are motivated to leverage interlinking tools to enrich their materials. Selecting an appropriate and up-to-date tool, which meets the desired criteria, can help to achieve this goal. In this paper, some of the most relevant interlinking tools have been evaluated through inspection and analysis of the generated output. Five experts later assessed the results of these tools when we applied them to link a large educational repository to the LOD datasets. The level of consensus among raters was measured by reliability statistics. Both the outcomes and the raters' responses were reported afterwards. Figure 1 portrays the overall workflow of this study wherein various interlinking tools were used in order to link two datasets while human experts examined the results.

The remainder of the paper is organized as follows: Section 2 discusses several studies focused on the use of interlinking tools. Section 3 outlines our selected linking systems, the interlinking process and the data scope in which this study has been carried out. Section 4 presents the experimental results of interlinking and examines the reliability of experts' reviews on the outcomes. Finally, conclusions and outlook are summarized in Section 5.

2. Background and related work

In order to cover Linked Data principles [1] the web of data requires different kinds of published data sources from various domains to be linked. As manual interlinking large amounts of data is time-consuming and needs a lot of human effort, it is necessary to provide a means to automatically interlink similar concepts. These linking tools create links (e.g. owl:sameAs) among various datasets by identifying similarities between entities. Given a linking configuration in which a user specifies the settings (such as the source and target, entities that should be considered, and the criteria under which two entities are compared), the tools discover similarities and generate the outcomes (consider Figure 1). There are many approaches, which can be used in order to perform data linking, some of which are summarized below:

- User Contributed Interlinking (UCI) [10] – proposes a new way of creating semantic links between data items. This tool allows users to add, view or delete links between two data items in a dataset via a friendly user interface.
- Games With A Purpose (GWAP) [11] – provides incentives for users to interlink data using different games. The tool gathers information about some pictures and asks the user to annotate images or trace objects in the pictures.
- Semi-automatic interlinking [12] – uses an analysis technique in order to link multimedia (meta)data.
- RDF-IA [13] – performs matching and fusion of RDF datasets according to user configuration, and generates several outputs between the data items.
- Silk Link Discovery Framework [14] – finds similarities within several Linked Data sources by specifying different types of RDF links via SPARQL endpoints or data dumps.
- LIMES [15] – discovers similarities between two datasets and automatically gives users suggestions based on metrics.
- LODRefine [16] – refines, transforms, and interlinks data in a general context with the LOD datasets.

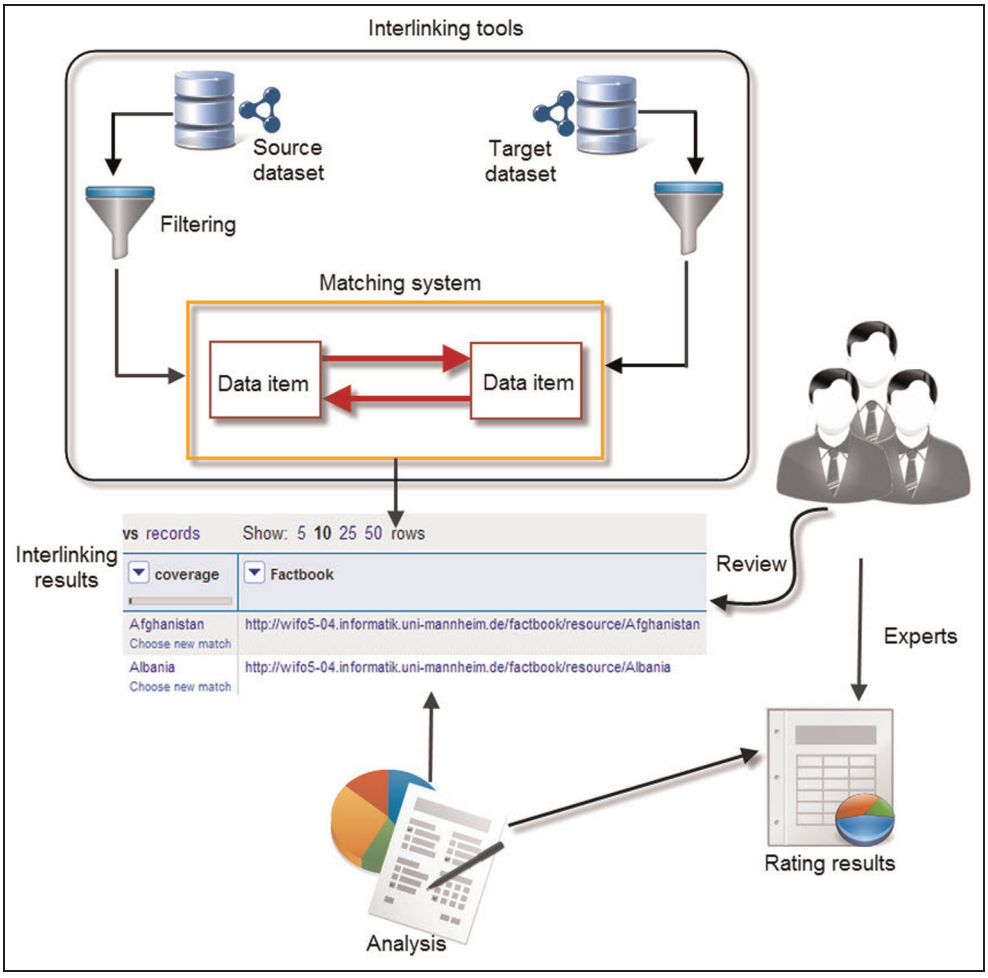


Figure 1. Interlinking and rating process.

Recently, several researches have discussed interlinking issues in the Linked Data context. A first comparative study on interlinking tools was reported by Simperl et al. [17], where various linking systems were considered from a theoretical point of view. In this paper, the authors reviewed several interlinking approaches by addressing important aspects such as required input, resulting output, considered domain and used matching techniques. The authors applied a template that included a general as well as a technical description of each tool, allowing a comparison from several perspectives: degree of automation (to what extent the tool needs human input) and human contribution (the way in which users are required to do the interlinking)

A general framework has also been proposed by Scharffe et al. [13] for data interlinking as well as an ontology alignment language which can be used in various linking techniques. After discussing a number of linking systems in the mentioned study, the authors focused on one of the tools to employ it in the proposed framework. In the context of the Ontology Alignment Evaluation Initiative, Ferrara et al. [18] evaluated several instance matching systems and reported their experimental results on a real-world benchmark task over several LOD datasets. In particular, the presented approach combined real-data and automatically generated data to provide a framework that would produce different causes of data heterogeneity. This in turn was used to verify the strengths and weakness of several data linking techniques.

In the educational context, Dietze et al. [6] proposed an approach for linking educational resources based on the Linked Data principles [1] by using existing educational datasets and exploiting the abundance of existing technology-enhanced learning data on the Web. The approach has been implemented in the context of the mEducator project [19]. Furthermore, several Linked Data projects such as LinkedUp [5], Linked Universities [20] and Linked Education [6] are prominent in the educational context and based on state-of-the-art Linked Data principles. Notably, LinkedUp aims at

advancing the exploitation of the vast amounts of public and open data available on the Web, in particular those produced by educational institutions and organizations. One of the main objectives of this ongoing project is to provide a complete framework for the evaluation of large-scale open Web data applications, taking into account educational aspects by gathering innovative scenarios of deployed tools.

Datalift [21], as another project in this context, proposes a set of tools for easing the process of dataset publication by converting raw data to Linked Data formats. Having described the data-linking task within the document, the authors divided the interlinking process into several steps, including configuration, pre-processing, matching and post-processing. Afterwards, the contributors overviewed, analysed and classified 11 linking systems in order to select the most appropriate tool for the purpose of their investigation.

In an experimental study [9] two matching techniques were employed to interlink a semi-structural data collection with the LOD datasets. The study also outlined the advantages of conjoining it to the LOD cloud by reporting the number of concepts that could be linked. The authors later discussed the results of the interlinking and the number of links found by each tool.

The foregoing studies and projects have indicated that automated interlinking tools play an important role in the emerging real Linked Data world. However, the evaluation of results generated by interlinking tools has been scarcely investigated when human experts assess the outcomes. The experimentation in this paper was carried out in order to assess if the results generated by interlinking tools were reliable and acceptable from the perspective of various experts. To achieve this goal, we conducted the entire interlinking process for each tool.

3. Experimental setting

The environment of our empirical research under which the interlinking process was carried out established two datasets (one as a source and the other as a target) in each step of our study. Accordingly, as selecting an appropriate linking tool lies at the heart of the discussion of interlinking, a set of criteria for investigating current linking tools as well as an interlinking procedure were outlined.

3.1. Data scope

Various digital repositories have exploited their data over the past 10 years as a way of tackling the problems raised by data proliferation. In particular, the GLOBE[6] collection, with around 1 million diverse learning resources [22], can be undoubtedly considered a must for interlinking purposes. GLOBE, as a federation repository, includes several data other repositories such as ARIADNE[7] and OER Commons,[8] which have manually created metadata and aggregated contents from different sources. The metadata in GLOBE is based upon IEEE LOM [23], a well-known standard for describing e-learning resources, but according to a recent study [22] only 20 elements (out of 40) have been used by the content providers. A closer look at data contained by GLOBE indicates that several metadata elements in IEEE LOM (i.e. General.Identifier, Technical.Location) are mostly customized locally by each repository and thus cannot be considered for interlinking. In a previous study by Rajabi et al. [9], the authors showed that at least the following elements can be applied for linking LOM metadata to the LOD cloud:

- coverage – the time, culture, geography or region to which a learning resource applies ('General.Coverage' element);
- taxonomy – the classification term given to a learning resource ('Classification.Taxon' element).

Given the interest in the GLOBE repository, it was possible to harvest around 500,000 metadata files from GLOBE through OAI-PMH,[9] which is a common protocol for metadata harvesting. More data could not be fetched owing to some validation errors (e.g. LOM extension errors) during the harvesting process. The gathered data was imported into a relational database to facilitate a more detailed examination from several perspectives (e.g. statistics, grouping).

Having reviewed the imported data, more than 50% of resources in GLOBE were found to belong to the Compulsory or Higher Education context and the targeted audiences were learners or teachers. More than half (around 55%) of the resources were in English and 99% of the learning objects were open and free to use. English is the most prominent language in GLOBE [24]. Taxonomy and coverage of learning objects are suitable candidates for interlinking [9]. Therefore, the linking elements used as a source in our data scope were limited to English terms of both taxonomy and coverage elements, which were represented in more than one language. On the other hand, two LOD datasets were selected as target points. At the time of this research, around 9000 datasets had been registered in the LOD cloud,[10] of which more than half were derived from global organizations. In a previous study [7], we indicated that the DBpedia

Table 1. Interlinking tools comparison.

Tool	Domain	SPARQL/ RDF dump	On Linked Data principles	Manual/ automatic	Well-documented/ frequently updated	Customization flexibility
GWAP	Multimedia	No	Yes	Manual	No	Unknown
LIMES	LOD	Yes	Yes	Automatic	Yes	Yes
LODRefine	General	Yes	Yes	Automatic	Yes	Partially
RDF-IA	LOD	RDF Dump	Yes	Automatic	No	Unknown
SAI	Multimedia	No	Yes	Automatic	No	Unknown
Silk	LOD	Yes	Yes	Automatic	Yes	Yes
UCI	LOD	Yes	Yes	Manual	No	Unknown

dataset plays a significant role in the LOD cloud, acting as a central dataset hub. This dataset features concepts for 10.3 million unique topics [8] and includes structured data about people, places and organizations. All DBpedia contents have been classified into 900,000 English concepts, and are provided according to SKOS,[11] as a common data model for linking knowledge organization systems on the Web of Data. Therefore, we selected DBpedia to link terms from the GLOBE taxonomies to DBpedia concepts. Likewise, the Factbook[12] dataset was applied for linking the coverage element of the GLOBE metadata, as it provides information such as history, people, government and transnational issues for 267 countries.

3.2. Selected tools for interlinking

As mentioned in Section 2, some of interlinking tools leverage matching algorithms to discover similarities between concepts in two datasets via SPARQL Endpoints[13] or RDF dumps.[14] For the research presented in this paper, several linking systems were investigated according to the following criteria (outlined in Table 1):

- conformity with Linked Data principles [1];
- support for SPARQL Endpoint or RDF dump file;
- the extent to which user contribution is needed (manual or automatic);
- well-documented and frequently updated tool;
- customization flexibility.

As illustrated in Table 1, Silk, LIMES and LODRefine were found to be the most appropriate linking systems based on the assumed criteria. 'Unknown' in the table indicates that authors could not test the flexibility of the tool owing to the lack of documentation and application support from the original providers. Below some of the main features of each tool are briefly discussed.

3.2.1. Silk. Silk [14] is a framework for interlinking between datasets that consists of a tool and a link specification language. When matching two datasets with Silk, the user specifies entities in a configuration file. The tool applies both string matching methods and taxonomical distance similarity in order to allow for diverse data discovery. These similarity metrics are parameterized by the user in a specific format. Silk takes two datasets as input by specifying SPARQL endpoints or RDF dumps and provides as an output 'sameAs' triples or any other predicates between the matched entities. This tool is available in three different variants, which address different use cases but use the same discovery engine. Silk Workbench, which is the case we applied for the interlinking, is a web application that guides users through the process of interlinking different data sources and offers a graphical editor to create as well as edit link specifications. As defining good linking heuristics is usually an iterative process, the Silk Workbench helps users to quickly evaluate the generated links. A number of projects, such as DataLift [21], have employed the Silk engine to carry out their interlinking purposes.

3.2.2. LIMES. Link Discovery Framework for Metric Spaces (LIMES) is a framework that implements a linking approach for discovering relationships between entities contained in Linked Data sources [15]. LIMES leverages several mathematical characteristics of metric spaces to compute pessimistic approximations of the similarity of instances. It then uses them to filter out a large amount of those instance pairs that do not satisfy the mapping conditions. Given a source, a

target and a link specification, LIMES processes the strings by making use of suffix, prefix and position filtering in a string mapper. The processing results of the string mapper (along with other types of mappers in the system) are filtered as well as merged using time-efficient set and filtering operations. As a result, LIMES generates links between items contained in two Linked Data datasets via SPARQL Endpoint or RDF dump. Using a threshold in the configuration file, the user can set a value for various matching metrics. Two instances are considered as matched and are linked via a relation such as 'owl:sameAs' when the similarity between the terms exceeds the defined threshold. Apart from LIMES' diverse collection of functionalities, a recent study [15] evaluated it as a time-efficient approach, particularly when it is applied to link large data collections.

3.2.3. LODRefine. OpenRefine[15] is a tool that allows data to be loaded, refined and reconciled. LODRefine [16] as an extension of OpenRefine provides additional functionalities particularly suited for dealing with LOD. Generally speaking, LODRefineis not only applied for cleaning and transforming data from one format to another, but it also discovers matched concepts between datasets by linking the data items to the target datasets. Matching is automatically performed in such manner that similar concepts are suggested to users for review and verification. LODRefine also allows users to expand their contents with concepts from the LOD datasets, for example, DBpedia or Freebase [25], once the data has been reconciled. The tool has a graphical user interface through which the user can import, clean and configure the target SPARQL Endpoint, or load the target RDF dumps. Users can also specify the condition under which the interlinking is to be performed. Finally, LODRefine reports the matched concepts and provides several functionalities for filtering the results. One of the advantages of LODRefine that can be highlighted is that it allows users to refine as well as manage data before starting the interlinking process. This is useful when the source includes several messy records (e.g. null, unrelated contents). Refinement of data before interlinking facilitates the process by reducing the number of source concepts.

3.2.4. Interlinking process. In an ideal scenario, a data collection would be linked to a diverse collection of datasets on the Web of Data. However, connecting each concept to an appropriate dataset one by one is too time-consuming, particularly when the number of data items is large and a domain expert has to explore the target dataset to query for the term. To minimize human contribution, data linking systems have facilitated the interlinking process by implementing a number of matching techniques. When testing an interlinking tool, several issues, such as defining the configuration for the linking process, specifying the criteria and post-processing the output, are addressed. As the GLOBE resources were not available as RDF, we had to expose the GLOBE metadata via a SPARQL endpoint. As we mentioned earlier in this study, the harvested metadata was imported into a relational database and afterwards exposed as RDF by making use of a mapping service (e.g. D2RQ[16]). We also set up a SPARQL Endpoint in order to complete the interlinking process. In the final GLOBE dataset, we discovered approximately 2342 English taxonomies ('taxon.entry' in the metadata) distributed amongst 193,000 metadata records. There were also around 5600 coverage values applied by 50,000 GLOBE resources.

When running LIMES, the user sets a configuration file in order to specify the criteria under which items are linked in the two datasets. The tool generates links between items under the specified criteria and provides output which defines whether there was a match or a similar term in order to be verified by users. In Silk, the user specifies both source and target metadata through a graphical user interface, then defines the criteria, and finally the tool generates output. Once the linking process has finished, the user can evaluate the accuracy of the generated links that are close to the similarity threshold. Specifically, the user can verify or reject each record recommended by the tool as two matching concepts.

LIMES and Silk completed the interlinking process by specifying the source and target datasets, type of matching algorithm and other filtering options. Figure 2 illustrates a possible workflow in which a data publisher could configure and run an interlinking tool to connect GLOBE and DBpedia. The threshold in the workflow was set to 0.98 for both tools, which means that two concepts are considered as matched if their similarity for the mentioned metric becomes more than 98%. The used matching algorithm was 'Levenshtein distance' [26] as a string metric for measuring the difference between two sequences.

In the case of LODRefine, the user imports the data into the tool and runs the reconciliation service after specifying the target dataset along with the type of concepts that are being linked. As a result, LODRefine reports the similarities between the concepts and the target links so that users can filter, or run another reconciliation action on the outputs.

4. Linking results evaluation

Figure 3 illustrates the results obtained by three tools (Silk, LIMES and LODRefine) employed to interlink GLOBE metadata and DBpedia categories. As can be seen, Silk and LIMES were able to match more concepts than LODRefine

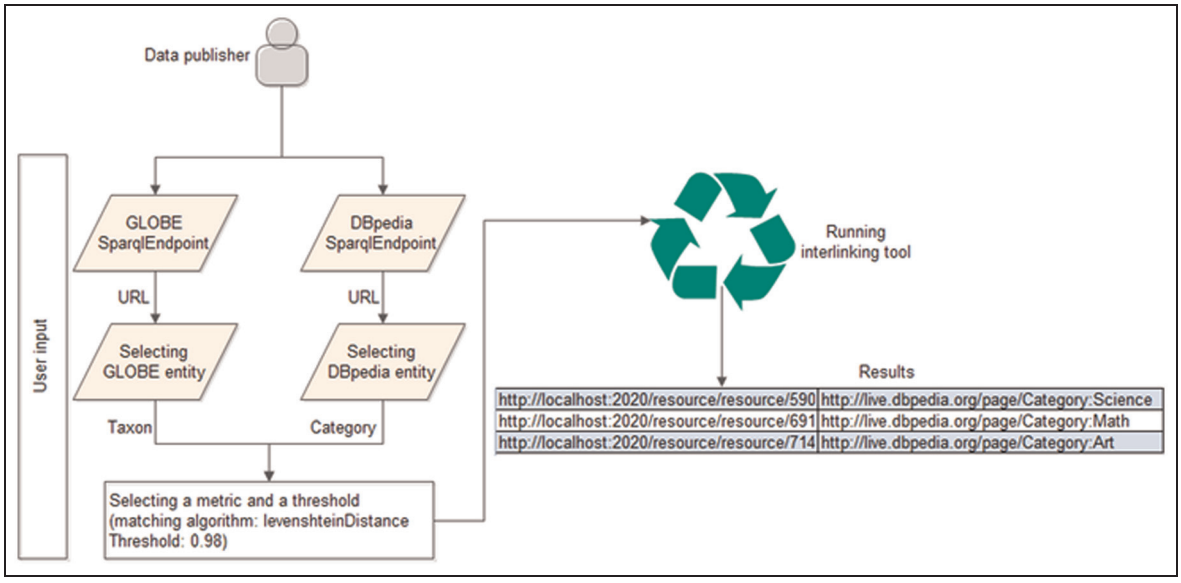


Figure 2. Sample interlinking process (GLOBE to DBpedia).

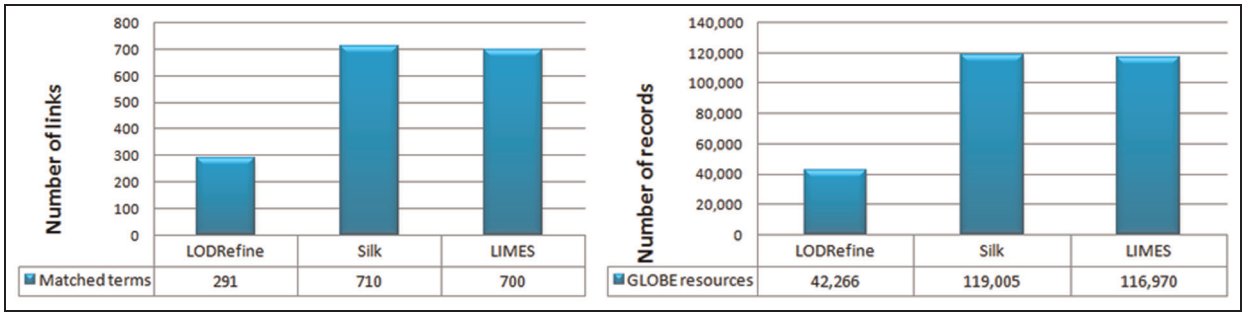


Figure 3. Taxonomy interlinking between GLOBE and the DBpedia dataset.

with around 710 and 700 terms (out of 2342), respectively, while the number of results obtained by LODRefine was considerably lower (291). The reason for the significant difference between the output of LODRefine and the other two tools may be because LODRefine does not allow users to select the matching algorithm. Instead users can refine links between the matched records via LODRefine facets (e.g. total tool judgment or according to word similarities). This means that the tool reconciles the results after finding the similarities between entities. Figure 3 illustrates the number of GLOBE records in which the terms were found. In particular, around 119,000 records included 710 terms discovered by Silk. All the DBpedia targets were de-referenceable as well as matched with more than 98% similarity to the GLOBE terms. Analysing the output, most concepts found in GLOBE belonged to various sources and were not restricted to one repository. On the other hand, the target links in DBpedia were distributed in a variety of categories ranging from Mathematics to Human Sciences, as illustrated in Table 2 (as a sample).

Indeed, the Housing category[17] was the most referenced subject in the DBpedia dataset pointed to by GLOBE concepts and discovered by both Silk and LIMES. On the other hand, LODRefine found the Asphalt category[18] to be the most commonly referenced subject.

When examining the extracted data, we identified almost 40,000 matched records (280 terms) that were common among all three tools. On logical grounds, it is apparent that both Silk and LIMES have more common links than LODRefine with around 696 concepts distributed in around 116,000 records in the GLOBE dataset, as they discovered richer results (Figure 4).

As each resource in the GLOBE repository can include more than one classification term (according to the IEEE LOM standard), each term was seen in more than one record and thus the number of repetitions (terms frequency among

Table 2. Extract of common outputs generated by all tools selected for experts' review.

Taxonomy term in GLOBE	DBpedia link found	DBpedia category
Marble	http://dbpedia.org/resource/Category:Marble	Rocks
Social Geography	http://dbpedia.org/resource/Category:Social_geography	Humans
Granite	http://dbpedia.org/resource/Category:Granite	Rocks
Magnetism	http://dbpedia.org/resource/Category:Magnetism	Physics
Lyrics	http://dbpedia.org/resource/Category:Lyrics	Songs
Business Economics	http://dbpedia.org/resource/Category:Business_economics	Business
Human Sciences	http://dbpedia.org/resource/Category:Human_sciences	Humans
Anthropology	http://dbpedia.org/resource/Category:Anthropology	Humans
Applied Mathematics	http://dbpedia.org/resource/Category:Applied_mathematics	Mathematics
Occupational Therapy	http://dbpedia.org/resource/Category:Occupational_therapy	Health_care
Linear Algebra	http://dbpedia.org/resource/Category:Linear_algebra	Mathematics
Numerical Analysis	http://dbpedia.org/resource/Category:Numerical_analysis	Mathematics
Matrices	http://dbpedia.org/resource/Category:Matrices	Mathematics
Audiology	http://dbpedia.org/resource/Category:Audiology	Health_sciences
html	http://dbpedia.org/resource/Category:HTML	Web_development
Software Engineering	http://dbpedia.org/resource/Category:Software_engineering	Computing
Detective Fiction	http://dbpedia.org/resource/Category:Detective_fiction	Crime_fiction
Comparative Literature	http://dbpedia.org/resource/Category:Comparative_literature	Literary_criticism
Peer-to-peer	http://dbpedia.org/resource/Category:Peer-to-peer	Collaboration Social_networks
Equations	http://dbpedia.org/resource/Category:Equations	Mathematics
Time Management	http://dbpedia.org/resource/Category:Time_management	Time
Semiconductors	http://dbpedia.org/resource/Category:Semiconductors	Electronic_engineering

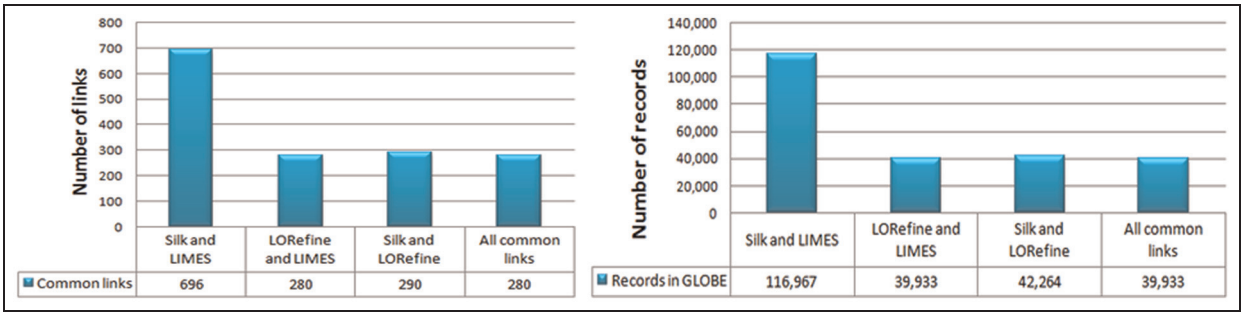


Figure 4. Common links and records among results generated by the selected linking tools.

Table 3. Interlinking results with repetition in the GLOBE repository.

	Matched terms	GLOBE resources	Repetition in GLOBE
LODRefine (A)	291	42,266	48,063
Silk (B)	710	119,005	165,292
LIMES (C)	700	116,970	162,049
A common B	290	42,264	48,061
B common C	696	116,967	162,023
A common C	280	39,933	45,334
Common results among all	280	39,933	45,334

the resources) in the final results was higher than the number of records in GLOBE (see Table 3). The expression 'B common C' in the table, for example, states the number of common records discovered by both Silk (B) and LIMES (C).

There are also many links found by one tool that were not discovered by the others. As it is apparent from Figure 5, both LIMES and Silk found almost 400 matched terms (around 92,000 records) more than LODRefine, while the difference between Silk and LIMES was only a few terms. Another noteworthy fact is that B minus C (depicted as B-C in the figure) shows that many links in B (Silk) were not discovered by C (LIMES).

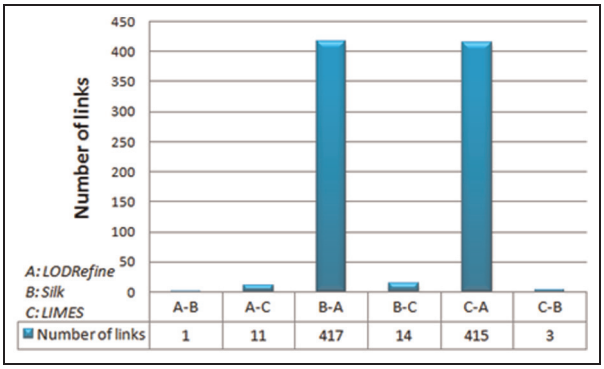


Figure 5. Number of links individually discovered by each tool.

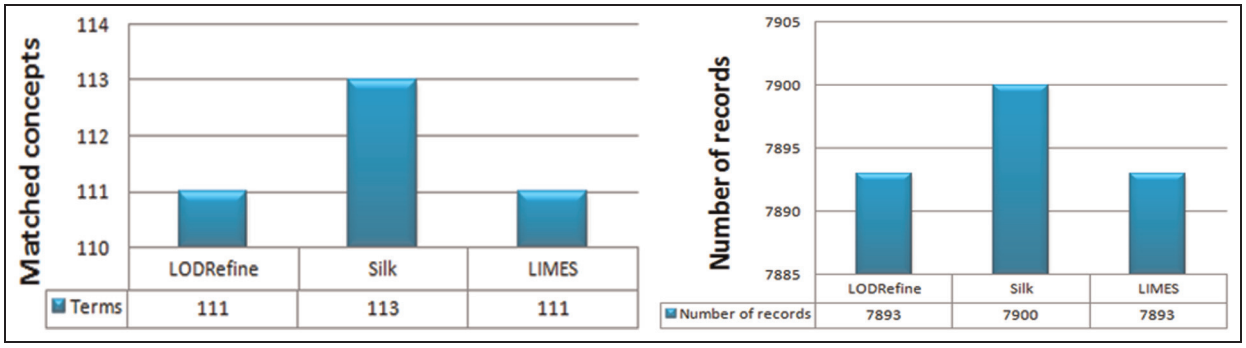


Figure 6. Coverage element interlinking between the GLOBE repository and the Factbook dataset.

There were also a few items exclusively discovered by one tool which were not identified by the others. For instance, Silk discovered five matched concepts that were observed neither by LIMES nor by LODRefine. Regarding the interlinking between the coverage element of GLOBE and countries contained in Factbook, the results obtained by the interlinking tools were almost the same. All tools could discover around 7900 matched records (out of 55,000) in the Factbook dataset, as depicted in Figure 6. All tools found at least 111 common terms except for Silk, which found an extra two concepts corresponding to seven records.

As a consequence of human and quality control, several records were selected from the final list of results and presented to five human experts. A set of criteria was defined under which the samples were chosen:

- All sample data was obtained from common matches among the tools.
- As GLOBE gathers metadata from various digital repositories, a few records were selected from each repository in order to promote diversity in the data sample collection.

The sample records which were given to the experts included GLOBE concepts and the target datasets. The metadata description related to each term was also extracted in order to help the experts detect any polysemy or ambiguity. Given that assessing each record (which includes source term, target link and description) requires some technical skills and takes time to elicit the target, we picked 25examples of taxonomy interlinking (out of 280) and 25 records corresponding to coverage interlinking (out of 111) for the evaluation (a sample taxonomy list is presented in Table 2). Finally, experts reviewed each term concerning its description and were later asked to respond whether the target link exactly matched the GLOBE term.

Focusing on reliability, all the experts approved the 25 records corresponding to coverage terms found by the tools in Factbook. This means that the coverage value and the country extracted from Factbook were undoubtedly the same from a human perspective. In the case of taxonomies matching and given that there was some disagreement among the experts' responses, we examined the degree of agreement by making use of inter-rater agreement techniques for evaluating

Table 4. Intraclass correlation coefficient

		95% Confidence interval	
	Intraclass correlation	Lower bound	Upper bound
Single measures	0.657	0.495	0.803
Average measures	0.905	0.830	0.953

Table 5. Number of accepted matched links rated by each sample (out of 25)

Experts	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5
Number of approved links	19	23	23	23	19

different raters' opinion on the same subject using the same scale or instrument. To gauge the response reliability, we applied intraclass correlation coefficient [27], one of the most popular reliability statistics to determine the internal consistency of multiple raters in a survey instrument. To this end, we imported the reviewers' data into SPSS[19] in order to analyse the responses and run the reliability statistics. In intraclass correlation statistics, the accepted value for describing internal consistency is defined by $\alpha > 0.6$ and the result is a coefficient when the value is > 0.9. Accordingly, as Table 4 illustrates, the software output for our data was valid, and the value was 0.905, which shows that the raters strongly agreed on the tools' output.

A closer look at the responses given by the experts indicates that they all accepted at least 19 samples out of 25 (76%) as matched terms. Some results were disregarded because either no related information was found or the terms did not match. For example, in two cases the raters mentioned that the metadata description in the GLOBE resources was not sufficient to rate the record as matched. Furthermore, two raters highlighted the lack of information in DBpedia (e.g. http://dbpedia.org/page/Category:Social_geography), which made the judgment difficult. Table 5 illustrates the number of approved taxonomy matching samples assessed by five different experts

5. Conclusion and outlook

The purpose of this research was to evaluate several existing interlinking tools when applied to linking a specific part of an educational collection to DBpedia and Factbook datasets. Firstly, the study set out to transform an e-learning repository into a Linked Data format and then select the most appropriate LOD dataset to specify the data scope for interlinking. Secondly, the authors investigated various related tools used for dataset interlinking and consequently selected Silk, LIMES and LODRefine, which passed a set of predefined criteria. The linking procedure was carried out over the target datasets and the results were presented to several experts for validation of the results generated by the applications. The findings gathered via human evaluation of the results obtained by the above-mentioned tools have a number of important implications for interlinking:

- As can be deduced from the linking results, applying interlinking tools helps data publishers to connect their contents to expedient datasets on the Web of Data, and thus we strongly recommend this approach.
- As almost all of the samples extracted by the tools were approved by experts in this research, it would be fair to conclude that using an interlinking tool is an effective way of linking between two datasets or from a data collection to the LOD datasets. As the paper found a high level of agreement among the raters (who mostly confirmed the results), we can confidently confirm that using an interlinking tool is a reliable method of interlinking datasets when the threshold of matching concepts is > 0.98.
- The moderately low number of results found by one of the tools reported in this study suggests that the use of several interlinking tools does not introduce significant added value for data providers.
- A comparison of the three interlinking tools reveals that Silk appears to be the most promising framework in terms of finding the most amounts of valid matches between datasets and providing diverse facilities for result verification.
- The difference between the number of results generated by LODRefine and the other tools in the case of Factbook and DBpedia also illustrates that LODRefine is not particularly effective when the scope of the target

is wide. In the case of DBpedia, the category dataset alone contains around 995,911 triples while Factbook has only 254 triples for countries. Generally speaking, all the tools fulfill the interlinking task when the scope is narrow, but when applied to broader contexts, Silk and LIMES have more stability.

- Another important conclusion that can be derived from this practical effort is the establishment of LIMES as a valid alternative for interlinking. However, LIMES still lacks some of the user facilities offered by Silk.

Further research on interlinking various data collections to other datasets on the Web of Data and the continued study of result validation are desirable to extend our knowledge of interlinking.

Acknowledgements

We would like to acknowledge the continued support of Dr Axel Ngonga and Dr Robert Isele, whose comments helped us to run the interlinking tools.

Funding

The work presented in this paper has been part-funded by the European Commission under the ICT Policy Support Programme CIP-ICT-PSP.2011.2.4-e-learning with project no. 297229, 'Open Discovery Space (ODS)', CIP-ICT-PSP.2010.6.2-Multilingual online services with project no. 27099 'Organic.Lingua', and INFRA-2011-1.2.2-Data infrastructures for e-Science with project no. 283770 'AGINFRA'.

Notes

1. University of Muenster. Available at: http://education.data.gov.uk/
2. Open University of UK. Available at: http://data.open.ac.uk/
3. Europeana, Linked Open Data. Available at: http://pro.europeana.eu/web/guest/linked-open-data
4. DBpedia dataset. Available at: http://dbpedia.org
5. DBLP dataset. Available at: http://dblp.uni-trier.de/
6. GLOBE repository. Available at: http://www.globe-info.org/
7. The ARIADNE Foundation. Available at: http://www.ariadne-eu.org/
8. OER Commons. Available at: http://www.oercommons.org/
9. Open Archives Initiative Protocol for Metadata Harvesting. Available at: http://www.openarchives.org/pmh.
10. Datahub. Available at: http://datahub.io/dataset.
11. Simple Knowledge Organization System (SKOS). Available at: http://www.w3.org/2009/08/skos-reference/skos.html
12. The World Factbook. Available at: https://www.cia.gov/library/publications/the-world-factbook/
13. SPARQL Endpoint. Available at: http://www.w3.org/wiki/SparqlEndpoints
14. RDF Dump. Available at: http://www.w3.org/wiki/DataSetRDFDumps
15. OpenRefine. Available at: http://openrefine.org/
16. D2RQ mapping service. Available at: http://d2rq.org/
17. http://dbpedia.org/page/Category:Housing
18. http://dbpedia.org/page/Category:Asphalt
19. IBM SPSS. Available at: http://www-01.ibm.com/software/analytics/spss/

References

1. Bizer C, Heath T and Berners-Lee T. Linked data – The story so far. *International Journal of Semantic Web Information Systems* 2009; 5(3: 1–22.
2. Cyganiak R. The Linking Open Data cloud diagram, http://lod-cloud.net/ (2011, accessed 13 April 2014).
3. Klyne G and Carroll JJ. Resource Description Framework (RDF): Concepts and abstract syntax, W3C Recommendation, 2004.
4. AGROVOC Vocabulary, http://aims.fao.org/standards/agrovoc/about (accessed 13 April 2014).
5. LinkedUp: Linking web data for education, http://linkedup-project.eu/ (accessed 13 April 2014).
6. Dietze S, Yu H Q, Giordano D, Kaldoudi E, Dovrolis N and Taibi D. Linked education: Interlinking educational resources and the Web of data. In: *Proceedings of the 27th annual ACM symposium on applied computing*, New York, 2012; pp. 366–371.
7. Rajabi E, Sanchez-Alonso S and Sicilia M-A. Analyzing broken links on the Web of Data: An experiment with DBpedia. *Journal of the Association for Information Science and Technology*, 2014, doi: 10.1002/asi.23109.
8. Bizer C, Lehmann J, Kobilarov G, Auer S, Becker C, Cyganiak R and Hellmann S, DBpedia – A crystallization point for the Web of Data. *Web Semant* 2009; 7(3): 154–165.
9. Rajabi E, Sicilia M-A and Sanchez-Alonso S. Interlinking educational data: An experiment with GLOBE resources. In: *First international conference on technological ecosystem for enhancing multiculturality (TEEM)*, 2013, pp. 365–374.

10. Hausenblas M, Halb W and Raimond Y. Scripting user contributed interlinking. In: *Proceedings of the 4th workshop on scripting for the Semantic Web (SFSW2008)*, co-located with *ESWC2008*, 2008.
11. Siorpaes K and Hepp M. Games with a purpose for the Semantic Web. *IEEE Intelligent Systems*, May 2008; 23(3): 50–60.
12. Hausenblas M. Interlinking multimedia: How to apply linked data principles to multimedia fragments, 2009.
13. Scharffe F, Liu Y and Zhou C. RDF-AI: An architecture for RDF datasets matching, fusion and interlink. In: *Proceedings of IJCAI*, 2009.
14. Volz J, Bizer C, Gaedke M and Kobilarov G. Silk – A link discovery framework for the Web of Data. In: *LDOW*, 2009.
15. Ngonga A and Sören A. LIMES – A time-efficient approach for large-scale link discovery on the Web of Data. *Presented at the IJCAI*, 2011.
16. Verlic M. LODGrefine – LOD-enabled Google Refine in action. Presented at the *I-SEMANTICS* (Posters & Demos), 2012.
17. Simperl E, Wölger S, Thaler S, Norton B and Bürger T. Combining human and computation intelligence: The case of data interlinking tools. *International Journal of Metadata Semantic Ontology* 2012; 7(2): 77–92.
18. Ferrara A, Nikolov A, Noessner J and Scharffe F. Evaluation of instance matching tools: The experience of OAEI. In: *Web semantics: Science, services and agents on the World Wide Web*, May 2013, vol. 21.
19. Mitsopoulou E, Woodham L, Balasubramaniam C, Poulton T, Protopsaltis A and Dietze S. mEducator: Multi type content repurposing and sharing in medical education. *The Academy Subject Centre for Medicine, Dentistry and Veterinary Medicine Newsletter*, 2010; 22: 26–28.
20. Linked Universities, http://linkeduniversities.org/lu/ (accessed 13 April 2014).
21. The Datalift project, a catalyser for the Web of data, http://datalift.org/en (accessed 13 April 2014).
22. Ochoa X, Klerkx J, Vandeputte B and Duval E. On the use of learning object metadata: the GLOBE experience. In: *Proceedings of the 6th European conference on technology enhanced learning: Towards ubiquitous learning*, Berlin, 2011, pp. 271–284.
23. IEEE P1484.12.4TM/D1, Draft recommended practice for expressing IEEE learning object metadata instances using the Dublin Core abstract model, http://dublincore.org/educationwiki/DCMIIEEELTSCTaskforce?action=AttachFile&do=get&target=LOM-DCAM-newdraft.pdf (accessed 13 April 2014).
24. Sicilia M-A, Sanchez-Alonso S, Garcia-Barriocanal E, Minguillón J and Rajabi E. Exploring the keyword space in large learning resource aggregations: the case of GLOBE. *Presented at the Workshop on learning object analytics for collections, repositories & federations*, 9 April 2013.
25. Google, Freebase Data Triples Dumps, https://developers.google.com/freebase/data (2010, accessed 16 May 2014).
26. Levenshtein V. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 1966; 10: 707–710.
27. Johnson W D and Koch G G. Intraclass correlation coefficient. In: Lovric M (ed.) *International encyclopedia of statistical science*. Berlin: Springer, 2011, pp. 685–687.

Title: Interlinking Educational Data: an Experiment with Engineering-related Resources in GLOBE

Publication:	International Journal of Engineering Education (IJEE)
Authors:	Enayat Rajabi, Miguel-Angel Sicilia, Salvador Sanchez-Alonso, and Juan Manuel Dodero
Relationship with the global research objectives:	Using the interlinking tools to interlink engineering resources of an eLearning repository (GLOBE) to Web of Data.
Impact factor:	0.360 (2013)
Date of submission:	21-Feb-2014
Date of acceptance:	30-Jul-2014
Date of publication:	In press (May/June 2015)

Interlinking Educational Data: an Experiment with Engineering-related Resources in GLOBE

ENAYAT RAJABI, MIGUEL-ANGEL SICILIA, and SALVADOR SANCHEZ-ALONSO
{enayat.rajabi, msicilia, salvador.sanchez}@uah.es
Information Engineering Research Unit, Computer Science Department,
University of Alcalá, 28805, Alcalá de Henares, Spain

JUAN MANUEL DODERO
Computer Engineering Department, University of Cádiz, Spain
juanma.dodero@uca.es

Abstract

Linking different kinds of engineering learning resources on the Web of Data enables enrichment, ease of navigation, casual discovery and improves resource seeking. This is performed by many tools and approaches built to discover similarities between the entities on the Web. In this paper we present a report primarily focused on evaluating the interlinking of engineering-related resources of a significant educational repository (GLOBE) to one of the most important datasets (DBpedia) on the Linked Open Data (LOD) cloud. After considering various interlinking approaches for link discovery, the paper focuses on the use of one of the interlinking tools (LIMES) and outlines the number of resources linked to the DBpedia dataset. In this empirical study, we report that almost 40,000 engineering resources were matched to the DBpedia concepts. Our findings are also examined as well as classified in various categories by human experts.

Keywords
Engineering resources, Linked Data, Interlinking, Educational data, GLOBE, LIMES.

1. Introduction

For years, a significant decline in the number of students graduating in Engineering fields has been observed [1]. It has been suggested that students who study science and engineering concepts experience a higher workload because this knowledge has a richer, more complex structure [2]. Of the diverse attempts to understand the problem, Felder [3] argues that more effective teaching methods in introductory courses will result in a higher retention rate. One of these methods is known as Problem-Based Learning, which requires students to identify and research, based on a poorly structured problem proposal, a set of resources useful for acquiring the knowledge needed to solve the problem [4]. However, if students lack the generic skills needed to undertake self-directed study, then it is likely that the goals of the problem-based strategy will not be achieved [5].

Making engineering students aware of the learning approach necessary can increase their involvement in courses [6]. Pedagogical approaches have been applied to initially structure open-ended problem-based learning approaches and gradually move towards open-ended problems. Nevertheless, relevant discussions on the suitability of problem-based strategies for teaching engineering conclude that it has certain limitations, which make it less suitable as an overall strategy for engineering education [2]. Another pedagogy that is usually applied in individual courses or throughout a curriculum is the project-based learning strategy [7].

In project-based learning, students access learning content when required, but the teacher prepares much of it. In problem-based learning, students control the content and delivery while the lecturer usually determines the problem. Eventually, either the students or the teacher must identify and control the content that is relevant to successfully solve the problem or achieve the project goals. The appropriate structuring of a project or an open-ended problem requires finding a handful of relevant resources, which can be found in open learning repositories as long as they are available.

From the point of view of content, various kinds of e-learning resources have motivated data providers to publish their educational documents on the Web of Data [8]. Linking engineering learning resources isolated in different repositories to valuable datasets facilitates resource seeking on the Web and pushes forward the exploitation of the large amounts of open data available on the Web [9]. Furthermore, it enriches the source information by connecting them to various targets of knowledge [8]. In particular, discovery of learning resources about engineering can be facilitated when they are interlinked with public domain datasets, statistics sources, and governmental data. Linked Data (LD) [10], as a recent approach for interlinking data, allows digital resources to be shared, reused, and accessed by students. Using LD, repository owners can publish structured data and establish categorized links between their repositories and from other sources. Furthermore, the LD approach and tools provide some solutions for intelligent linking, as well as for integration and consumption of experiment data [11]. Many educational institutions, universities and libraries have embraced LD principles and released educational resources as part of the LD cloud. DBpedia [12], one of the most used datasets [13] is an LD version of Wikipedia that makes it possible to link data items to general information on the Web. In particular, the advantages of linking of engineering content to DBpedia is to make public information usable for other datasets and to enrich datasets by linking to valuable resources on the Web of Data [14]. However, interlinking educational data is still largely unexplored. Dietze et al. [15] presented a general approach to exploit the wealth of existing technology-enhanced learning (TEL) data on the Web by allowing its exposure as LD.

In this paper, we evaluate existing approaches for interlinking objects on the Web of Data and select LIMES [16] for linking a large collection of data to the LOD cloud. As a result, we expose the GLOBE (Global Learning Objects Brokered Exchange) metadata as LD and discover the similarities between its metadata elements and DBpedia. Finally we evaluate the results and list the advantages of this interlinking process.

The rest of the paper is structured as follows. Section 2 describes briefly how educational data is nowadays exposed as LD and discusses different existing approaches for interlinking. In Section 3, we discuss the dataset used for examining the interlinking framework as our experimental setting. Section 4 provides the methods and results for our evaluation. Conclusions and outlook are provided in Section 5.

2. Background

In the last decade, the existing teaching and learning strategies in engineering education have been improved so that faculty members in academe are recommended to make enhanced design pedagogy their highest priority in future resource allocation decisions [17]. Engineering graduates also need to have a broader knowledge of fundamental engineering science and computer literacy [7]. Given that engineering students' demands are unlikely to be satisfied by a traditional engineering curriculum, they are expected to find their learning resources on the Web. On the other hand, the majority of e-learning materials, which are engineering-related, can be enriched when they are conjoined to useful information on the Web of Data. In the following sub-sections we will explain how several educational institutions have exploited their learning materials on the Web and what the current linking approaches for connecting various learning resources are. Finally, we will select our approach for interlinking.

2.1 Exposing educational resources as Linked Data

Several educational institutions e.g., the University of Muenster (DE) [18], the Open University (UK) [19], the National Research Council (CNR, Italy) [20], and the Southampton University (UK) [21] embraced the LD approach by exposing their learning resources as LD formats. Notably, we outline two educational datasets which have exploited their learning (meta)data in RDF format:

Organic.Edunet [22] is a learning portal that provides access to digital learning resources on Organic Agriculture and Agroecology – it facilitates access, usage and exploitation of such content. This collection, which currently contains the metadata of almost 11,000 resources, has exposed its content as LD [23] and published these resources as a dataset in the LOD cloud [24]. This dataset is also linked to other datasets such as DBpedia through its metadata elements.

Europeana [25], the European Union's flagship digital library project, enables search and discovery in more than 17 million items by collecting metadata from approximately 1,500 cultural data providers across Europe [25]. Europeana published a first sub-set of the Europeana dataset [26] after enriching existing metadata records via a SPARQL endpoint and data dump. It exposes data based on the Europeana Data Model (EDM), which is for publishing and linking Europeana metadata. It also links the data provider's metadata to other datasets such as DBpedia, Geonames [27] and GEMET [28].

In particular, one of the approaches for representing any kind of data as LD is mapping the collection to RDF triples [29] which has been applied for our interlinking purpose in the following steps:

A. Storing the metadata in a repository that will be accessible via the web.
B. Converting them to RDF using semantic web tools.
C. Making the educational data accessible via a SPARQL endpoint or RDF dump

2.2 Approaches for interlinking

Several tools and approaches exist for interlinking data in the LOD datasets. Simperl et al. [30] provided a comparison of interlinking tools based on a set of criteria such as use cases, annotation, input and output. Similarly, we explain some of the related tools, but focusing on their need for human contribution (to what extent users have to contribute to interlinking), their automation (to what extent the tool needs human input), and area (in which environment the tool can be applied).

From a human contribution perspective, User Contributed Interlinking (UCI) [31] is an interlinking tool that creates different types of semantic links such as "owl:sameas" and "rdf:seeAlso" between two datasets relying on user contributions. In this Wiki-style approach, users can add, view or delete links between data items in a dataset by using a UCI interface. "Games With A Purpose" (GWAP) [32] is another software which provides incentives for users to interlink datasets using games and pictures. In this tool, the user distinguishes different pictures with the same name. "Linkage Query Writer" (LinQuer) [33] is also a software for semantic link discovery between different datasets, based upon a framework that consists of APIs that allow users to write their queries in an interface.

Semi-automatic interlinking [34], as another approach for interlinking, provides a type of analysis technique to assign multimedia data to users using multimedia metadata. "Interlinking Multimedia" (iM) [35] is also a pragmatic way for applying the LD to fragments of multimedia items and presents methods for enabling a widespread use of interlinking multimedia. RDF-IA [36] is another linking application that carries out matching, fusion and interlinking of RDF datasets according to the user configuration, and generates several outputs such as interlink files including "owl:sameAs" statements between the data items. Another semi-automatic approach for interlinking is the Silk Link Discovery Framework [37] which finds the similarities by specifying the types of RDF links. Some similarity metrics are combined based on the link conditions within different LD sources. LIMES is also a link discovery software in the LOD which implements a time-efficient approach for large datasets in metric spaces [16]. This approach presents a command-line tool and a graphic user interface for finding similarities between two datasets and automatically suggests the results based on the metrics. GoogleRefine [38] is software for cleaning, transforming, and interlinking any kind of data with a web user interface. It has also the benefit of reconciling data to the LOD datasets (e.g., Freebase or DBpedia) [39]. The following table briefly summarizes the described tools and mentions the area of application for each one.

Table 1: Existing interlinking tools description

Tool	**User contribution**	**Area**
UCI	Reviewing the semantic links	General data source
GWAP	Matching of objects through playing a game	Web pages, e-commerce offerings, Flickr images, and YouTube
LinQuer	Writing LinQL queries	LOD datasets
IM	Matches multimedia by annotating and linking	Multimedia
RDF-IA	Configuring the input	LOD datasets
Silk	Configuring the input file, reviewing the result	LOD datasets
LIMES	Configuring the input file, reviewing the result	LOD datasets
GoogleRefine	Importing data, reviewing the result	General data ,LOD datasets

2.3 Selected approach for interlinking large datasets

As our approach is to interlink the datasets via a SPARQL endpoint, Silk and LIMES were selected as our final candidates. Besides that, they were well-documented, updated frequently and used rich as well as diverse matching algorithms for interlinking [37][16]. In both approaches, the user specifies the SPARQL endpoints of the datasets, comparable elements and thresholds of acceptance of output. Eventually, the tools report the results based upon the user configuration and similarities between two datasets. In a study, Ngonga et al. [16] examined both LIMES and Silk, from a time-efficiency perspective and showed that LIMES is more time-efficient than Silk for link discovery between two LOD datasets. They evaluated LIMES using synthetic as well as real data and it outperformed other approaches with respect to the number of comparisons and runtime. They also showed that the speed of this tool improves with the complexity of the mapping task and makes it especially suitable for handling large-scale matching tasks. Moreover, in a study Rajabi et al. [40] evaluated several

interlinking tools on the Web of Data and identified LIMES as one of the most promising tools for linking datasets, and thus we selected the tool to carry out the interlinking.

3. Experimental setting

GLOBE is a large repository with almost 1.2 million learning resources [41]. Including various kinds of educational data encouraged us to assess the possibility of interlinking GLOBE to the LOD datasets. GLOBE is a federated repository that consists of several other repositories, such as OER Commons [42], which has manually created metadata as well as aggregated metadata from different sources. Current research on the use of GLOBE learning resource metadata [41] shows that 20 out of 50 of the metadata elements, which are based upon the IEEE LOM schema [43], are used consistently in the repository and thus can be considered for interlinking. After analyzing the GLOBE metadata, we realized that several metadata elements such as "General.Identifier" or "Technical.Location" are mostly included local values provided by each repository and thus cannot be considered for interlinking. Additionally, constant values (e.g., dates and times) or controlled vocabularies (e.g., "Contribute.Role" and "Aggregation.Level") were not suitable for interlinking, as the user could not obtain useful information by linking these elements. In our previous study [44] we showed that the following metadata elements had the greatest possibility of interlinking to the LOD datasets:

- The time, culture, geography, or region to which the learning resource applies ("General.Coverage")
- The taxonomy given to a learning resource ("Classification.Taxon")
- A keyword or phrase describing the topic of learning objects ("General.Keyword").

As a consequence of the interlinking process, around 815,000 metadata files were harvested through the OAI-PMH [45] protocol from the GLOBE repository. Some GLOBE metadata could not be harvested due to validation errors (e.g., LOM extension errors). Particularly, several repositories in GLOBE extended the IEEE LOM by adding new elements without using namespaces, which caused a number of errors detected by the ARIADNE validation service. Analyzing the harvested records, more than half of the resources (55%) were in English language and almost all of them were free (99% without cost).From a technical point of view, around 256,898 resources (31%) with more than one million repetitions in GLOBE provided taxonomies in the metadata, of which 162,203 records (20%) with almost 524,000 repetitions were in English language. Figure 1 illustrates the top taxonomies of the metadata categorized according to their string values. The Y-axis in the diagram indicates the number of resources in thousands.

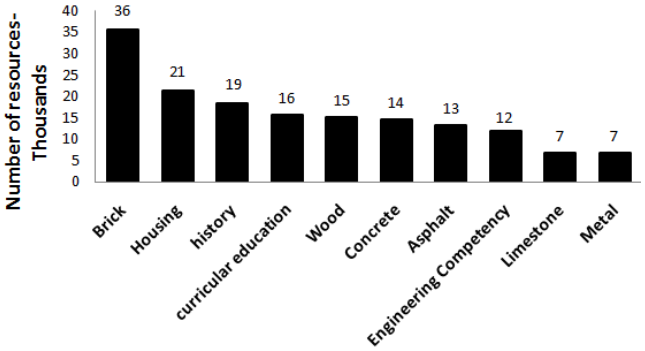


Figure 1: Top ten taxonomies in the GLOBE metadata

In order to identify the engineering resources within the GLOBE metadata, we carried out a comparative study between the "classification" category in the metadata and the latest version of the hierarchical ACM classification [46] system for some information about computing as the world's largest educational and scientific computing society. As a result, we found 39,801 records that were matched either to the ACM taxonomies or contained engineering values in the classification element. As shown in Table 2, there were almost 5,200 engineering-related resources (ERR) in GLOBE that included a "coverage" element in the metadata, while the number of ER records including the keyword element was 17,006.

Table 2: Engineering related data in GLOBE

Title	**Number**
Resources including taxonomies in English language	162,203
Engineering-related resources (ERR)	39,801(25%)
Total ERR resources provided including *coverage* values	5,221
Total ERR resources including *taxon* values	39,801
Total ERR resources including *keyword* values	17,006

To expose the former elements as RDF, we installed a D2RQ service [47], which is a mapping service for exploiting relational database as LD format (e.g., RDF, N3). To this end, we converted the harvested metadata files, which were in XML format, into a relational database. As a result, the GLOBE engineering data was accessible through a local SPARQL endpoint in order to be interlinked to the DBpedia dataset.

4. Interlinking results and discussion

As we discussed earlier in this paper, LIMES was selected for link discovery between GLOBE and DBpedia. This tool generates links between items contained in two datasets via a SPARQL endpoint or RDF dump. Users can set a threshold in LIMES for the metric above in which two entities are considered to match one another, and another threshold (e.g., 50%) for manual evaluation of the results. Interlinking can be performed either via a SPARQL endpoint or through an RDF dump. In the case of GLOBE, we set up a SPARQL endpoint for interlinking, as an RDF dump of a huge collection would have been too large and hard to parse. The SPARQL endpoints of datasets, similarity measurements, and acceptance or review conditions are set up by the user as software configurations. After running the tool, the result of interlinking obtained in two separate log files (matched concepts, and concepts for user review) is presented to the user. In the case of GLOBE, we set the matching threshold as 98% and the review threshold as 50% for manual evaluation of the results. In the following sub-sections, we will outline the interlinking results along with the human evaluation of discovered links.

4.1 Semi-Automatic Interlinking

The LOD cloud includes a wide variety of datasets that can be applied for linking entities. In this paper, we used the DBpedia dataset, which includes structured information about persons, places, and organizations. The full DBpedia dataset features labels and abstracts for 10.3 million unique topics in 111 different languages [48]. Hence, this dataset was selected for linking keywords and taxonomies of metadata. This dataset also fits the coverage element of GLOBE metadata, including places, countries and regions applicable to the learning objects. We will discuss the interlinking results to this dataset in detail.

As a result (see Table 3), values in the “Coverage” element of ERR in GLOBE have been exactly matched to 1,468 (out of 5,221) regions and places of the DBpedia dataset. Keyword and taxonomy are two elements of the LOM metadata frequently used to classify learning objects. To this end, we focused on the DBpedia classification [49] and 10,341 (out of 17,006) keywords were found by LIMES as matching the DBpedia category, while the number of matched taxonomies was around 27,099 (68%) concepts.

Table 3: Matches found between GLOBE and DBpedia

Element	Number of matches	Total records
Coverage	1,468 (28%)	5,221
Keyword	10,341 (60%)	17,006
Taxon	27,099 (68%)	39,801

As it can be seen from Table 4, there exist a wide variety of records in the GLOBE repository that had similarities (not exactly matched to the target dataset) to the DBpedia concepts and were recommended to the user for review. As the records did not fully match (with more than 50% similarity) the terms, they have been manually reviewed. Some examples of the results (matched and similar terms) are presented in the Appendix 1. Figure 2 also depicts total accepted terms between GLOBE and DBpedia (exactly and nearly matched).

Table 4: Similarity between GLOBE and DBpedia for manual review

Element	Number of similarities	Total records
Coverage	3,422 (65%)	5,221
Keyword	9,414 (55%)	17,006
Taxon	17,477 (43%)	39,801

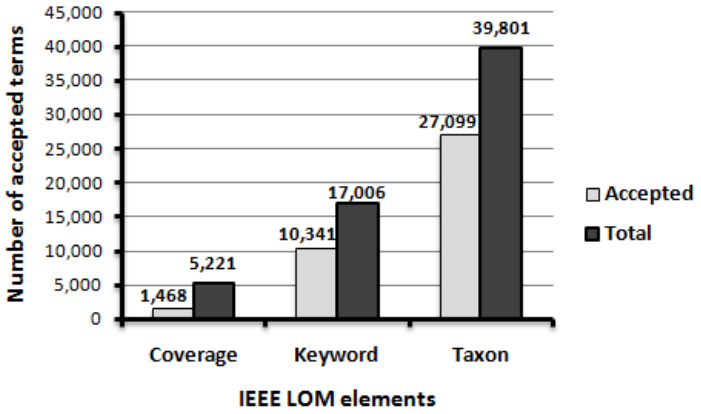


Figure 2: Total number of similarities between GLOBE and DBpedia

With respect to the similar concepts identified by the LIMES tool, the values of the "Coverage" element in both the DBpedia and GLOBE datasets had a 1-n relationship, as each country could be assigned to many regions of the GLOBE metadata. However, some of the similar concepts did not seem to be semantically accurate. For instance, "History_of_Portugal" and "History_of_Science" were two terms identified as similar concepts, but they point to different data. "Keyword" values in both the GLOBE and DBpedia datasets, as another example, had an m-n relationship, as each "Keyword" was connected to several resources of the DBpedia (and vice versa). Appendix 1 illustrates some samples of these similarities.

4.2 Manual Evaluation

As we discussed earlier, most interlinking tools present two types of results, i.e. matched and similar concepts. When analyzing the matched concepts outlined by the tools, undoubtedly both terms in GLOBE and DBpedia were the same from a string pattern-matching criterion (consider Appendix 1). As a consequence of evaluating similar terms by human expert, we presented hundred records of each result to three domain experts to assess as well as classify each one in a specific category. In the examination phase, the following possibilities might occur for each term:

- Matched: Two concepts are exactly the same (e.g., "Italy" in the GLOBE metadata and "Italy" (country) in the DBpedia dataset)
- Related: Two terms are not the same, but they have a relationship with each other as follows:
  - isPartOf: The source concept in GLOBE is a physical or logical part of the target concept. (e.g., "Mexico City" and "Mexico")
  - isParentOf: The DBpedia concept is a physical or logical part of the GLOBE concept. (e.g., "Nuclear_Energy_Companies" in DBpedia and "Nuclear energy" in GLOBE)
  - isRelatedTo: The source and target concepts have various kinds of relationships (except isPartOf and isParentOf) with each other. (e.g., "criticism" and "Criticism_of_journalism")
- isnotRelated: The similar source and target terms have string similarities, but they are not conceptually the same. (e.g., "Pacific Islands" with "Cayman_Islands")

The following table illustrates the results of the manual evaluation from the domain experts. The number in the table depicts the average number of concepts examined by the experts. Some of the similar concepts that had more than 80% string similarity were detected as "exactly matched" by experts. This was mostly correct particularly for those elements that have a close relationship to the target dataset (e.g., "Urban Studies and Planning" with "Planning_and_Urban_Research"). The "Coverage" element, as an example, had 32 concepts exactly matched to the target dataset by LIMES (e.g., "Niger, Africa" with "Niger"). In "Keyword" and "Taxon" elements, the isParentOf relationship had the most similarities among other kinds of measurements, while the isNotRelated relationship was not found among them. In particular, a majority of concepts in DBpedia were physically or logically part of the GLOBE keywords. To take an example for the Keyword element, we found around 24 concepts in DBpedia which all were part of University of Cambridge as one of the keywords in the GLOBE dataset (e.g., "University of Cambridge examinations" or "Alumni of Cambridge University") and this means that the term in GLOBE was the parent of those concepts in DBpedia (isParentOf relationship). In regard with "Taxon" element, there were found 10 concepts in DBpedia as part of martial arts (e.g., "Hybrid_martial_art" or "German_martial_arts"). On the contrary, 31 cases were found with isNotRelated relationships in the "Coverage" element in both DBpedia and GLOBE. Given that the "Coverage" element of learning resources in GLOBE mostly point to geographical places, the interlinking tool identified some concepts like "North America" and "Korea_North" as similar concepts, while conceptually they are different and thus

we categorized them as isNotRelated relationships. On the other hand, most of countries in DBpedia and GLOBE were exactly matched with each other, as the context of these two datasets on this element was very close.

Table 5: Average number of concepts (out of 100 sample records) reported by LIMES evaluated by three experts for LOM elements

Element	Similarity type	Average	%
Keyword	Exactly matched	4	4%
	GLOBE **is part of** DBpedia	8	8%
	GLOBE **is parent of** DBpedia	82	82%
	GLOBE **is related to** DBpedia	6	6%
	Is not related	0	0%
Taxon	Exactly matched	3	3%
	GLOBE **is part of** DBpedia	9	9%
	GLOBE **is parent of** DBpedia	84	84%
	GLOBE **is related to** DBpedia	4	4%
	Is not related	0	0%
Coverage	Exactly matched	32	32%
	GLOBE **is part of** DBpedia	17	17%
	GLOBE **is parent of** DBpedia	14	14%
	GLOBE **is related to** DBpedia	6	6%
	Is not related	31	31%

The foregoing discussion shows that it seems fair to conclude that when the context of the metadata element is more related to the target dataset (e.g., we consider “Coverage” element to DBpedia places), the result will include a greater frequency of matched results.

5. Conclusion

In this paper we evaluated the interlinking of engineering resources in the GLOBE repository to the DBpedia dataset. Considering various interlinking tools, we chose LIMES as our linking approach. After exposing the GLOBE metadata as LD, we analyzed the similarities of many entities in this collection and other existing datasets in the LOD cloud, such as DBpedia.

The GLOBE resources include valuable educational metadata that can be enriched when they are applied in linkable ways. By linking to the related datasets on the Web, the GLOBE users can get more valuable information about the learning resources. The “Coverage” that applies to learning resources can be linked, for example, to DBpedia places or other datasets such as Eurostat as long as it includes places (e.g. countries, cities). The more data provided in the DBpedia dataset (e.g., population, statistics data), the better help for users to obtain useful and enriched information. Furthermore, when the GLOBE resources are linked to the SKOS classification of the DBpedia, they can be discovered by any LD application, particularly those that use the SKOS classification for their search process. There exist 11 million triples in the DBpedia dataset, out of which 1.7 million triples include the SKOS category, which was conjoined with the GLOBE metadata.

Manual evaluation of the interlinking outcome by domain experts also showed that the GLOBE resources definitely have the potential to be linked to the related datasets, as we found special relationships among the results (e.g., isParentOf and isPartOf) that can be used for linking the terms in the GLOBE repository and other datasets. Based on our analysis, other datasets also exist in the LOD that can be interlinked to the GLOBE materials, when they include learning content. Linking more related and linkable datasets in the LOD cloud to huge educational repositories provides users with more flexibility to expand their knowledge regarding the source collections.

Acknowledgment

The work presented in this paper has been part-funded by the European Commission under the ICT Policy Support Programme with project No. 297229 “Open Discovery Space (ODS)” and project No. 27099 “Organic.Lingua”.

References

[1] E. Godfrey, T. Aubrey, and R. King, Who leaves and who stays? Retention and attrition in engineering education, *Engineering Education*, vol. 5, no. 2, pp. 26–40, Dec. 2010.

[2] J. C. Perrenet, P. A. J. Bouhuijs, and J. G. M. M. Smits, The Suitability of Problem-based Learning for Engineering Education: theory and practice, *Teaching in Higher Education*, vol. 5, no. 3, pp. 345–358, Jul. 2000.

[3] R. M. Felder, D. O. C. Engineering, P. H. Mohr, C. O. Engineering, and E. J. Dietz, A longitudinal study of engineering student performance and retention. I. Success and failure in the introductory course, *Journal of Engineering Education*, pp. 15–21, 1993.

[4] C. E. Hmelo-Silver, Problem-Based Learning: What and How Do Students Learn? , *Educational Psychology Review*, vol. 16, no. 3, pp. 235–266, Sep. 2004.

[5] E. Montero and M. J. Gonzalez, Student Engagement in a Structured Problem-Based Approach to Learning: A First-Year Electronic Engineering Study Module on Heat Transfer, *IEEE Transactions on Education*, vol. 52, no. 2, pp. 214–221, May 2009.

[6] D. Hernandez-Leo, V. M. Oliver, J. M. Dodero, A. Pardo, M. Romero-Ternero, Y. Dimitriadis, and J. I. Asensio-Perez, Applying Recommendations to Align Competences, Methodology, and Assessment in Telematics, Computing, and Electronic Engineering Courses, *IEEE Revista Iberoamericana de Tecnologías del Aprendizaje*, vol. 8, no. 1, pp. 15–22, Feb. 2013.

[7] J. Mills and D. Treagust, Engineering Education - is problem-based or project-based learning the answer? , *Australasian Journal of Engineering Education*, 2003.

[8] H. Q. Yu, C. Pedrinaci, S. Dietze, and J. Domingue, Using Linked Data to Annotate and Search Educational Video Resources for Supporting Distance Learning, *IEEE Transactions on Learning Technologies*, vol. 5, no. 2, pp. 130–142, 2012.

[9] M. Fernandez, M. d' Aquin, and E. Motta, Linking data across universities: an integrated video lectures dataset, *Proceedings of the 10th international conference on The semantic web* - Volume Part II, Berlin, Heidelberg, 2011, pp. 49–64.

[10] C. Bizer, T. Heath, and T. Berners-Lee, Linked Data - The Story So Far, *International Journal on Semantic Web and Information Systems*, vol. 5, no. 3, pp. 1–22, 33 2009.

[11] S. Dietze, S. Sanchez-Alonso, H. Ebner, H. Q. Yu, D. Giordano, I. Marenzi, and B. P. Nunes, Interlinking educational resources and the web of data: A survey of challenges and approaches, *Program: electronic library and information systems*, vol. 47, no. 1, pp. 60–91, Feb. 2013.

[12] "DBpedia dataset", Available at: http://dbpedia.org, [Accessed: 22-February-2014].

[13] E. Rajabi, S. Sanchez-Alonso, and M.-A. Sicilia, Analyzing Broken Links on the Web of Data: an Experiment with DBpedia, *Journal of the Association for Information Science and Technology*, Vol 65, pp. 1721–1727, doi: 10.1002/asi.23109.

[14] M. D'Aquin, Linked Data for Open and Distance Learning, Common Wealth of Learning, 2012. [Online]. Available: http://elearningeuropa.info/en/directory/Linked-Data-for-Open-and-Distance-Learning. [Accessed: 22-February-2014].

[15] S. Dietze, H. Q. Yu, D. Giordano, E. Kaldoudi, N. Dovrolis, and D. Taibi, Linked education: interlinking educational resources and the Web of data, *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, New York, NY, USA, 2012, pp. 366–371.

[16] A. Ngonga and A. Sören, LIMES - A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data, *presented at the IJCAI*, 2011.

[17] C. L. Dym, A. M. Agogino, O. Eris, D. D. Frey, and L. J. Leifer, Engineering Design Thinking, Teaching, and Learning, *Journal of Engineering Education*, vol. 94, no. 1, pp. 103–120, 2005.

[18] University of Muenster Linked Data, Available at: http://education.data.gov.uk/ , [Accessed: 22-February-2014].

[19] Open University (UK) dataset, Available at: http://data.open.ac.uk/ , [Accessed: 22-February-2014].

[20] National Research Council (CNR, Italy) dataset, Available at: http://data.cnr.it , [Accessed: 22-February-2014].

[21] Southampton University dataset, Available at: http://data.southampton.ac.uk/ , [Accessed: 22- February-2014].

[22] Organic.Edunet Project, Available at: http://wiki.organic-edunet.eu/index.php/Main_Page, [Accessed: 22-February-2014].

[23] Organic.Edunet dataset, Available at: http://data.organic-edunet.eu, [Accessed: 22-February-2014].

[24] M.-A. Sicilia, H. Ebner, S. Sanchez-Alonso, F. Alvarez, A. Abian, and E. Garcia-Barriocanal, Navigating learning resources through linked data: A preliminary report on the re-design of Organic.Edunet, *1st International Workshop on eLearning Approaches for the Linked Data Age*, 2011.

[25] B. Haslhofer and A. Isaac, data.europeana.eu - The Europeana Linked Open Data Pilot, *DCMI International Conference on Dublin Core and Metadata Applications*, The Hague, The Netherlands, 2011.

[26] Europeana dataset, Available at: http://data.europeana.eu, [Accessed: 22-February-2014].

[27] GeoNames, Available at: http://www.geonames.org/. [Accessed: 22-February-2014].

[28] GEMET Dataset, Available at: http://www.eionet.europa.eu/gemet/. [Accessed: 22-February-2014].

[29] Turtle - Terse RDF Triple Language, W3C Team Submission 28 March 2011, Available at: http://www.w3.org/TeamSubmission/turtle/. [Accessed: 22-February-2014].

[30] E. Simperl, S. Wölger, S. Thaler, B. Norton, and T. Bürger, Combining human and computation intelligence: the case of data interlinking tools, *International Journal of Metadata, Semantics and Ontologies*, vol. 7, no. 2, pp. 77–92, Oct. 2012.

[31] M. Hausenblas, W. Halb, and Y. Raimond, Scripting User Contributed Interlinking, *Proceedings of the 4th workshop on Scripting for the Semantic Web (SFSW2008)*, co-located with ESWC2008, 2008.

[32] K. Siorpaes and M. Hepp, Games with a Purpose for the Semantic Web, *IEEE Intelligent Systems*, vol. 23, no. 3, pp. 50–60, May 2008.

[33] O. Hassanzadeh, R. Xin, R. J. Miller, A. Kementsietsidis, L. Lim, and M. Wang, Linkage Query Writer , *Proceedings of the VLDB Endowment*, 2009.

[34] M. Hausenblas, Interlinking Multimedia: How to Apply Linked Data Principles to Multimedia Fragments, *LDOW*, 2009.

[35] T. Bürger and M. Hausenblas, Interlinking Multimedia – Principles and Requirements, *Proceedings of the International Workshop on Interacting with Multimedia Content in the Social Semantic Web (IMC-SSW'08)* , 2008.

[36] F. Scharffe, Y. Liu, and C. Zhou., RDF-AI: an architecture for RDF datasets matching, fusion and interlink, *Proceeding of IJCAI* IR-KR Work, 2009.

[37] O. Hassanzadeh, A. Kementsietsidis, L. Lim, R. J. Miller, and M. Wang, Semantic Link Discovery over Relational Data, *Semantic Search over the Web*, 2012, pp. 193–223.

[38] Google-refine - Google Refine, a power tool for working with messy data (formerly Freebase Gridworks), Available at: https://code.google.com/p/google-refine/. [Accessed: 22-February-2014].

[39] M. Verlic, LODGrefine-LOD-enabled Google Refine in Action, *presented at the I-SEMANTICS (Posters & Demos)*, 2012.

[40] E. Rajabi, M.-A. Sicilia, and S. Sanchez-Alonso, An empirical study on the evaluation of interlinking tools on the Web of Data, *Journal of Information Science*, first published on June 11, 2014 doi:10.1177/0165551514538151.

[41] X. Ochoa, J. Klerkx, B. Vandeputte, and E. Duval, On the use of learning object metadata: the GLOBE experience, *Proceedings of the 6th European conference on Technology enhanced learning: towards ubiquitous learning*, Berlin, Heidelberg, 2011, pp. 271–284.

[42] OER Commons, Available at: http://www.oercommons.org/. [Accessed: 22-February-2014].

[43] IEEE LTSC, IEEE Standard for Learning Object Metadata 1484.12.1-2002, Final Draft version, [Online]. Available: http://ltsc.ieee.org/wg12/20020612-Final-LOM-Draft.html. [Accessed: 22-February-2014].

[44] E. Rajabi, M.-A. Sicilia, and S. Sanchez-Alonso, Interlinking Educational Data: an Experiment with GLOBE Resources, *presented at the First International Conference on Technological Ecosystem for Enhancing Multiculturality*, Salamanca, Spain, 2013.

[45] Open Archives Initiative Protocol for Metadata Harvesting, Available at: http://www.openarchives.org/pmh. [Accessed: 22-February-2014].

[46] The 2012 ACM Computing Classification System. Available at: https://www.acm.org/about/class/2012 [Accessed: 22-February-2014].

[47] C. Bizer, D2RQ - treating non-RDF databases as virtual RDF graphs, *Proceedings of the 3rd International Semantic Web Conference (ISWC2004)*, 2004.

[48] DBpedia blog, Available at: http://blog.dbpedia.org, [Accessed: 22-February-2014].

[49] DBpedia classification, Available at: http://dbpedia.org/page/Category, [Accessed: 22-February-2014].

Biography

Enayat Rajabi is a PhD researcher in the Computer Science Department of the University of Alcalá since 2012. He holds his MSc (2004) degree in Computer Engineering from the department of Ferdowsi University of Mashhad (Mashhad, Iran). He is presently a full time researcher in the Information Engineering research unit at the University of Alcalá (Alcalá de Henares, Madrid, Spain). The aim of his PhD is to interlink educational data using Semantic Web and Linked Data.

PROF. Miguel-Angel Sicilia is currently full professor at the Computer Science Department of the University of Alcalá and the head of the Information Engineering research unit, where he leads several European and national research projects in the topics of learning technology and Semantic Web. He is the Editor-in-Chief of the International Journal of Metadata, Semantics and Ontologies, published by Inderscience, and serves as a member of editorial board of many other scientific journals in the area of Semantic Web, Computational Intelligence and Information Systems. He has been active in metadata research and was awarded the 2006 Cyc prize for the best research paper.

Salvador Sanchez-Alonso is an associate professor in the Computer Science Department of the University of Alcalá, Spain. He has participated in and coordinated several EU-funded projects over the last few years related to TEL, learning object repositories and federations, notably eContent+ Organic.Edunet, CIP-PSP Organic.Lingua and VOA3R,. His current research interests include TEL, Learning repositories, Semantic Web and Computer Science education.

Juan Manuel Dodero is currently an associate professor of the University of Cádiz, Spain and the head of the Office of Libre Software and Open Knowledge. He has a PhD in Computer Science, and his research interests include web engineering, web science and computer-assisted learning. He is the leader of a number of research projects, such as asceta.uca.es, related with his interests. He is a national coordinator of the OpenDiscoverySpace.eu project.

Appendix

Sample interlinking results

Tool	GLOBE	DBpedia	Element
LIMES matches	Montreal	http://dbpedia.org/resource/Montreal	Coverage
	Copenhagen	http://dbpedia.org/resource/Copenhagen	
	Victoria (Australia)	http://dbpedia.org/resource/Victoria_(Australia)	
LIMES review	Tibet	http://dbpedia.org/resource/Taibet	Coverage
	Rickenbach	http://dbpedia.org/resource/Krickenbach	
	Medel	http://dbpedia.org/resource/Medeo	
LIMES matches	Mechanical engineering	http://dbpedia.org/resource/Category:Mechanical_engineering	Keywords
	biometrics	http://dbpedia.org/resource/Category:Biometrics	
	Addition reactions	http://dbpedia.org/resource/Category:Addition_reactions	
LIMES review	networks	http://dbpedia.org/resource/Category:Social_networks	Keywords
	Phoenician	http://dbpedia.org/resource/Category:Phoenicia	
	revenue	http://dbpedia.org/resource/Category:Revenge	
LIMES matches	Teleconferencing	http://dbpedia.org/resource/Category:Teleconferencing	Taxonomy
	Linear algebra	http://dbpedia.org/resource/Category:Linear_algebra	
	Project Management	http://dbpedia.org/resource/Category:Project_Management	
LIMES review	economic system	http://dbpedia.org/resource/Category:Economic_Systems	Taxonomy
	Brics	http://dbpedia.org/resource/Category:Brick	
	Queueingtheory	http://dbpedia.org/resource/Category:Queueing_theory	

Title: Interlinking Educational Resources to Web of Data through IEEE LOM

Publication:	Computer Science and Information Systems (Comsis)
Authors:	Enayat Rajabi, Miguel-Angel Sicilia, and Salvador Sanchez-Alonso
Relationship with the global research objectives:	Interlinking Educational Resources to Web of Data through IEEE LOM
Impact factor:	0.575 (2013)
Date of submission:	30-Mar-2014
Date of acceptance:	01-Oct-2014
Date of publication:	January 2015

Computer Science and Information Systems 12(1):233–255 DOI: 10.2298/CSIS140330088R

Interlinking Educational Resources to Web of Data through IEEE LOM

Enayat Rajabi[1], Miguel-Angel Sicilia[1], and Salvador Sanchez-Alonso[1]

[1] Information Engineering Research Unit, Computer Science Department, University of Alcalá, 28805, Alcalá de Henares, Spain
{ enayat.rajabi, msicilia, salvador.sanchez}@uah.es

Abstract. The emergence of Web of Data enables new opportunities for relating resources identified by URIs combined with the usage of RDF as a lingua franca for describing them. There have been to date some efforts in the direction of exposing learning object metadata following the conventions of Linked Data. However, they have not addressed an analysis on the different strategies to expose Linked Data that could be used as a basis for leveraging the metadata currently curated in repositories following common conventions and established standards. This paper describes an approach for exposing IEEE LOM metadata as Linked Data and discusses alternative strategies and their tradeoffs. The recommended approach applies common principles for Linked Data to the specificities of LOM data types and elements, identifying opportunities for interlinking exhaustively. A case study and a reference implementation along with an evaluation are also presented as a proof of concept of this mapping.

Keywords: educational resources, Linked Data, IEEE LOM, learning metadata, interlinking.

1. Introduction

The purpose of learning object metadata is to support the reusability and discoverability of learning objects and facilitate their interoperability in the context of e-learning. Particularly, it is used to enable information seekers (e.g., teachers and learners) and applications such as repositories, portals and learning environments to search for, evaluate, retrieve and use learning objects. IEEE LOM [1] is a widespread standard for describing educational contents, promoting their reusability and interoperability through the use of a standardized set of descriptors and common conventions to encode descriptive metadata [2]. This standard is a commonly accepted way for describing learning resources in repositories. A recent study [3] has revealed a consistent usage of 20 out of the 50 metadata elements in the standard, considerably more elements than conventionally collected with widespread schemas such as Dublin Core.

Exposing metadata for search and discovery of resources on the Web has always been an important concern for repositories, and the use of standards is a proof of that. However, IEEE LOM does not explicitly promote relating learning objects, even though "Relation" element has been defined. Specifically, it does not recommend relations to be expressed as links, which is the universal approach in the Web of Data to express

relations among resources. The lack of a shared way of linking precludes crawlers and other applications to get the most out of the relations between resources. By analyzing some of the IEEE LOM elements (as we will discuss later), we found that linking several metadata elements (e.g., coverage) to the Linked Open Data(LOD) datasets, makes the learning object enriched and accessible to the other useful information on the Web of Data. In a sample of 815,223 IEEE LOM metadata records gathered from the GLOBE federation [4], 20% of the resources included the "Relation" element in their metadata records. We examined these resources and found that only 95,946 records (about a 12% of the total) were using URIs to express relations and others contain strings an numbers.

The Linked Data (LD) approach [5] relying on the use of RDF links, represents an alternative way of openly exposing metadata fostering interlinking. RDF links allow to interconnect any kind of resource on the Web, allowing to easily link to external datasets or repositories [6] that are already providing URIs for identifying their resources. Many institutions, universities and libraries have adopted the LD principles and have released resources and data as part of the LOD cloud [7]. Notably, DBpedia [8], one of the most used datasets, which exposes a Linked Data version of Wikipedia, makes it possible for anybody to link to general information as well as to extract relationship to other datasets.

The advantage of this new approach to express relationships between resources is making public information linkable and usable for others [9]. This has the benefit of enabling applications to exploit learning object metadata and other information available in the Web of Data. It can also be seen as an extension of open educational resources initiatives [10] in the direction of making them more readily available for discovery.

The exposure of LOM compliant metadata as Linked Data supports functionalities over RDF-defined LOM records that cannot be attained with the human-oriented version of LOM, e.g., triggering queries on SPARQL endpoints [11] no matter where the records are stored. Users can also export their educational metadata in LOD format in the same manner that libraries all over the world are doing in the library field. However, exposing LOM metadata as LOD is not straight-forward and requires a transformation of metadata plus a bootstrapping phase to identify candidate links to other datasets or educational resources, eventually with the aid of interlinking tools [12] [13]. This in turn requires the use of vocabularies to provide some shared semantics that can be exploited for the traversal of metadata across repositories. Given that some semantics are actually encoded in the IEEE LOM standard, there is a need to elaborate some RDF exposure practices for which existing proposals for mapping IEEE LOM to RDF [14] [15] can be useful but are not enough. This includes URI design and the identification of opportunities for interlinking.

This paper reports a complete analysis on the different strategies to expose IEEE LOM as Linked Data, describing how IEEE LOM elements and data types can be represented in RDF based on Linked Data principles [5] and complying with common Linked Data patterns [16]. It also reports on a case study and reference implementation and evaluates its performance. The case is based on the Organic.Edunet repository [17], a IEEE LOM-based repository of learning materials in the field of organic agriculture and agroecology.

The rest of paper is structured as follows. Section 2 briefly describes the background on exposing IEEE LOM elements and the related works in this context. In Section 3, we recommend a URI design for identification of e-learning objects in educational repositories. This section also represents a mapping of LOM elements to Linked Data

format. Section 4 provides an experimental implementation of RDF [18] binding of IEEE LOM. Section 5 presents an evaluation and performance testing over the mentioned implementation. Conclusions are provided in Section 6.

2. Background and Related Works

Work on e-learning metadata standards at the international level has been carried out by a number of organizations including the IEEE, the Dublin Core Metadata Initiative (DCMI), IMS Global [19], and ISO/IEC [20]. Achieving interoperability across these specifications has been recognized as a major challenge since 2000 [21].

IEEE LOM is an internationally-recognized open standard bound up with the history and development of the IMS e-learning interoperability specifications (e.g. IMS Content Packaging [19]), and with the evolution of the ADL SCORM [22] reference model, which supports the IEEE LOM alongside other specifications. Dublin Core (DC) has also been used in many systems and applications as an alternative to other metadata standards (e.g. IEEELOM) or in combination with them to provide wider interoperability.

A recent effort within the ISO community is Metadata for Learning Resources (MLR) [23] which aimed at harmonizing LOM and Dublin Core metadata, as it tries to enable both the "learning object" aspects of LOM and the "entity-relationship" model of the Semantic Web associated with the Dublin Core Abstract Model [24]. Moreover, it is intended to support multilingual and cultural adaptability requirements from a global perspective. The Learning Resource Metadata Initiative (LRMI) [25] has also developed a common metadata framework for describing learning resources on the web. LRMI promoted by popular search engines Google, Bing, and Yahoo, is related to schema.org and supported by Creative Commons. Although the goal of these schemas is to be a complement or alternative to IEEE LOM and DC, a wide variety of learning repositories and federations (notably the GLOBE federation [4]) use IEEE LOM as the base metadata schema and actively aggregates LOM records at a large scale.

To date, there have been some initiatives to expose learning resource metadata as Linked Data. Dietze et al [26] surveyed some high-level approaches aimed towards Linked Education by allowing its exposure as Linked Data and interlinking techniques for the educational domain. Dietze et al [27] also proposed an approach for linking educational resources based on the Linked Data principles by using existing educational datasets and vocabularies. Its aim was to exploit the wealth of existing technology-enhanced learning (TEL) data on the Web by exposing it as Linked Data. The approach has been implemented in the context of the mEducator project [28] where data from a number of open TEL data repositories has been integrated, exposed and enriched making use of the Linked Data approach.

Fernandez et al [29] presented a work on linking educational resources across universities through the use of Linked Data principles by focusing on extracting and structuring information of video lectures produced by 27 different educational institutions according to some vocabularies, e.g. FOAF. As a result of this work, a new media educational dataset was released.

There exist some other projects such as LinkedUp [30] and Linked Universities [31] which aim at sharing learning data or metadata related to educational Linked Data on state-of-the-art Linked Data principles.

In particular, Linked Data exposure of IEEE LOM is not a new subject though, as the work was initiated in 2000 in the context of the IMS Global Learning Consortium [19] (together with the ARIADNE Foundation [32]) that developed a XML binding and RDF binding of LOM elements and, as a result, some RDF documents were produced as IMS RDF Bindings. The Dublin Core Metadata Initiative (DCMI) [33] also provided recommendations for expressing DC metadata as RDF and described how the features of the DCMI are represented based on LOM to DCAM mapping document [34]. The recommended document described how to use the definitions of metadata terms defined by the IEEE LOM Standard, for RDF binding of IEEE LOM Elements together with DCMI metadata terms.

A mapping from LOM to RDF model (defined by Nilsson et al. [15]) described advantages of expressing learning object metadata as RDF. Nilsson also discussed some problems encountered in the process of producing the RDF binding for LOM elements and focused on some specific futures of the binding, although this early work was discontinued [26] and did not cover all the LOM elements.

Some other tools and IEEE LOM editors also export LOM elements as RDF. For example, ocw2rdf [35] harvests metadata from the MIT Open Course Ware web site [36] and transforms it into an RDF representation of IEEE LOM. Kunze et al [37] developed and implemented a browser-based editor in which the author can choose the type of metadata using any kind of RDF-schema available on the WWW to annotate learning resources in a specific repository (OLR3). Balog-Crisan and Roxin [38] proposed an on-line tool called RDF4LOM, to edit metadata in RDF. The proposed tool creates RDF documents according to the LOM standard.

Our work continues and completes Nilsson et al [15] approach for exposing learning object metadata as Linked Data. To this aim, we consider all the IEEE LOM elements, data types and vocabularies and provide a mapping to RDF. We also present a complete and unified solution for exposing learning object metadata and implement this approach on an educational repository so that this repository can link its (meta)data to Linked Open Data by following clear guidelines. As the RDF implementation is not straightforward and the decisions for the transformation of several items often have 2 or more possible alternatives. We tried to base our decisions and recommendations on good practices, but even so, our decisions are subject to debate and can evolve in the future.

3. Exposing IEEE LOM as Linked Data

In this section, we highlight on the exposure of the IEEE LOM elements as RDF, represented here in XML format. Initially, we discuss how e-learning objects are identified in LOM elements. The recommendation presented in this study is the outcome of a long authors' discussion with both Linked Data and e-learning experts. A complete mapping of all the IEEE LOM elements is available at http://data.organic-edunet.eu/ODS_LOM2LD/ODS_SecondDraft.html.

3.1. URI Design

In IEEE LOM, identifiers are defined as "globally unique label that identifies a learning object" and are to be provided in:

- Element 1.1: General.Identifier as the identifier of the resource
- Element 3.1: Meta-Metadata.Identifier as the identifier of the metadata record
- Element 7.2.1: Relation.Resource.Identifier as the identifier of a related resource

In a general case, the dereferenceable URIs that deliver RDF descriptions, are actually identifying metadata records and not the actual resources. In consequence, the identification in Element 3.1 is represented as the dereferenceable URI from which the RDF metadata is exposed, and there is no need to expose it again in the RDF representation. In the case of the "Relation" element, the recommended practice is using the dereferenceable URI of the resource pointed by this one, if available, in the form of a RDF link. If the URIs of learning objects are considered to form a natural hierarchy, then a patterned URI can be assigned to them [16].

In terms of technical design, World Wide Web Consortium (W3C) published some guidelines in order to define a well-formed URI [39] [40]. The document we used as a basis for our solution to define learning object identifiers, stated two approaches based upon the HTTP URI scheme and protocol which fulfils the following requirements:

Description of the identified resource should be retrievable with standard Web technologies.

A naming scheme should not confuse things and the documents representing them.

3.2. Binding Simple and Structured Elements

Two types of LOM elements exist: simple and structured (or aggregation). The following sub-sections discuss the RDF representation of each type. One metadata example of a learning resource (e.g., http://youtube.com/example_resource), represented in an XML format, is used throughout this paper and, therefore, we will avoid repeating the resource identifier in each example. As simple elements do not contain other LOM elements and mostly include one value (e.g., String) at the target, they have been represented plainly as subject, predicate and object. This RDF binding have been recently followed by many datasets in the LOD cloud (e.g., DBpedia, Factbook). As an illustration, technical format of learning objects in LOM (Technical.Format) is expressed in Turtle [41] (Consider Table 1).

Structured elements included other LOM elements (either simple or structured elements) are often realized using intermediate nodes, but there exist various options for exposing structured elements as LD depending on maximum number of entities they include (multiplicity) and their order. Several LOM elements (e.g., "General.Title") with structural format were considered with multiplicity one in the IEEE standard and given that their order is not significant in the metadata, the simplest way of representation and already compatible with a wide range of existing software, is leveraging the repeated properties in RDF. In the repeated properties, the user can assign many predicates to one subject regardless to its objects' order and thus can be applied to appropriate elements.

Table 1 illustrates how a structured element (“General.Title”) is expressed in different languages in RDF.

Table 1. RDF binding of IEEE LOM elements

Binding type	XML representation example	RDF Binding
simple element	<technical> <format> "application/x-shockwave-flash" </format> </technical>	<http://youtube.com/example_resource> dcterms:format "x-shockwave-flash".
element with multiplicity 1	<title> <string language="en"> What is organic. </string> <string language="de"> Was ist biologisch. </string> </title>	<http://youtube.com/example_resource> dcterms:title "What is organic"@en; <http://youtube.com/example_resource> dcterms:title "Was ist biologisch"@de.
structured element using blank nodes	<general><keyword> <string language="en"> Certification </string> <string language="de"> Zertifizierung </string> </keyword> <keyword> <string language="en"> Farming </string> <string language="de"> Landwirtschaft </string> </keyword></general>	<http://youtube.com/example_resource> lom:keyword _:node1,_:node2. ---------- _:node1 rdf:value "Certification" @en, "Zertifizierung" @de. _:node2 rdf:value "Farming" @en, "Landwirtschaft" @de.

Intermediate nodes, also called blank nodes due to the absence of a name (or a dereferenceable URI) to a triple, are used to indirect referencing to a element with unspecified name. Although intermediate nodes are considered as problematic approach in terms of implementation of RDF and users readability [42], their usage is unavoidable when there exists a deep hierarchy (more than two) of elements in a model or the multiplicity of structural elements is “Many”. The “Keyword” element in “General” category is a good practice, as can be expressed repetitive in more than one language and thus using repeated properties here is not applicable, as it would mix the translations of the different values. In the table above we showed that how two intermediate nodes have been used for representation of the keyword element.

Figure 1 also portrays a simple guideline for the RDF binding of simple and structured elements of IEEE LOM according to foregoing discussion.

Alternatively, RDF containers e.g., RDF:Alt and RDF:Seq [16] are applied to describe a group of values in RDF representation and they are appropriate when the element hierarchy is limited in two levels. RDF:Seq suits particularly when the order among elements is important (see Table 2). As this representation becomes more complicated in deep hierarchical structures of the IEEE LOM elements (e.g., classification.taxon), using the RDF containers for the elements that are not explicitly required to be ordered, is not recommended.

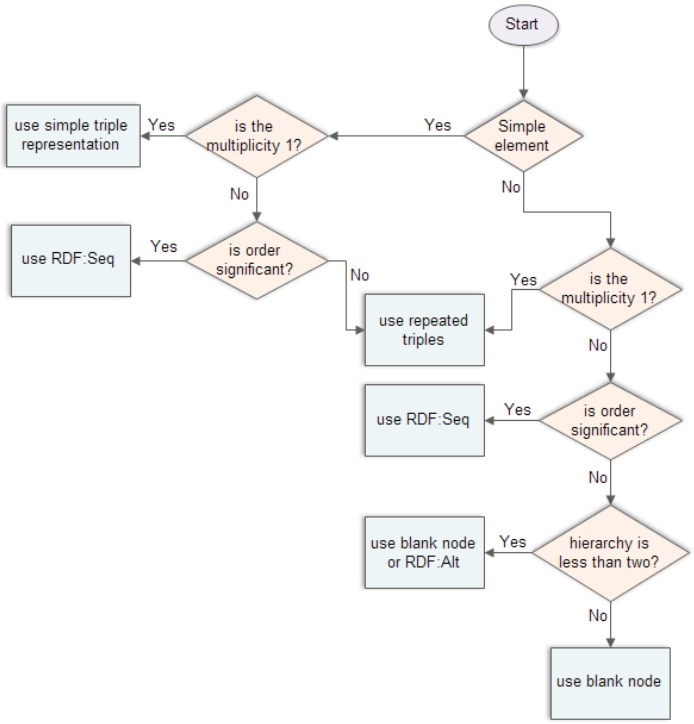


Fig. 1. The workflow of RDF binding of IEEE LOM elements

Table 2. RDF binding of a structured element using RDF containers

XML representation	RDF Binding
<classification> <keyword> <string language="en"> Organic <string> </keyword><keyword> <string language="en"> Farming <string> </keyword> </classification>	<http://youtube.com/example_resource> lom:classificationKeyword _:node1. _:node1 a rdf:Seq; rdf:_1 "Organic" @en. rdf:_2 "Farming" @en.

3.3. DataType Mapping and Reusing Vocabulary

The following sub-sections provide a description of data type mapping of the IEEE LOM elements.

3.4. CharacterString

Simple elements in String format are represented as plain literals in RDF, e.g., “Technical.Format” in the LOM standard, whose type is “CharacterString”, would be represented as follows:

```
<http://youtube.com/example_resource>
    dcterms:format "x-shockwave-flash".
```

3.5. LangString

Several IEEE LOM elements use the “LangString” data type which binds together multiple literals with equivalent expressions in different languages. The literal is expressed as a plain literal in RDF along with a language tag (e.g., en) conformed to RFC1766 [43]. The “LifeCycle.Version”, as a good practice, has multiplicity one and is therefore, as mentioned earlier, represented as a direct property pointing to a plain literal with a language tag:

```
<http://youtube.com/example_resource>
       lom:version
            "It is not available" @en,
            "No está disponible" @es.
```

3.6. DateTime

The International Standard for the representation of dates and times, ISO 8601 [20], describes a large number of “DateTime” formats. IEEE LOM standard defines at least four digits for year, two for month and two for day. For representing the time, it states two digits for hour, two for minutes, two for seconds and one or more digits representing a decimal fraction for a second. IEEE LOM elements that represent “DateTime” values can be exposed in the following format:

```
<http://youtube.com/example_resource>
   lom:contributionDate  "2011-05-17T05:53:31.00Z"
```

IEEE LOM allows “DateTime” elements to be expressed as literal with language (e.g., {“en”,“circa 1300 BCE”}). For those elements, we recommend “LangString” representation as follows.

```
<http://youtube.com/example_resource>
   lom:contributionDate "circa 1300 BCE" @en
```

3.7. Duration

Duration, as an interval data type, is recommended to be expressed as follows:

```
<http://youtube.com/example_resource>
   lom:technicalDuration  "PT0.25S"
^^<http://www.w3.org/2005/xpath-
datatypes#dayTimeDuration>.
```

In the above example, duration ("PT0.25S") shows that technical duration of the learning object is 25 seconds based upon ISO8601 [20], although the represented format is not human readable. As "DateTime" data type, elements with String value representing Duration (e.g., {"en", "Fall Semester 1999"}) are expressed as LangString.

3.8. Boolean, Integers and other Simple Data Types

In the RDF exposure, it is encouraged to reuse the XML schema data types [44]. For example for "Boolean" values, the data type of the element is indicated as true or false:

```
<http://youtube.com/example_resource>
   lom:cost false
   ^^<http://www.w3.org/2001/XMLSchema#boolean>.
```

Likewise, for expressing other simple data types such as integer, long, float, etc. using the XML schema data type is recommended.

3.9. vCard:

vCard [45] is a standard for electronic business cards. To capture a vCard, an intermediate node is recommended together with properties such as vCard:FN, vCard:ORG and vCard:Email. The entity value of contribute element in the LifeCycle category, the Table 3, is a vCard record represented in XML.

Table 3. RDF binding of vCard

XML representation	RDF Binding
BEGIN:VCARD FN:John Smith EMAIL;TYPE= INTERNET: John@example.org ORG: http://www.example.org N:John;Smith VERSION:3.0 END:VCARD	<http://youtube.com/example_resource> lom:contributionEntity _:bnode1247. _:bnode1247 vcard:FN "JohnSmith". _:bnode1247 vcard:N "John;Smith". _:bnode1247 vcard:EMAIL _:bnode1248. _:bnode1247 vcard:ORG "http://www.example.org". _:bnode1247 vcard:VERSION 3.0. _:bnode1248 rdf:value "John@example.org"; rdf:type "http://www.w3.org/2001/vcard-rdf/3.0#internet".

3.10. Undefined Data Type

The IEEE standard states "Undefined" as one of the data types of LOM elements, although most date types are expressed explicitly and can be represented in RDF. For example "xsd:dateTime", is used for "DateTime" format and "xsd:boolean" for "Booleans" and so forth. However, if an element cannot be defined in any specific date type in the LOM schema, "xsd:anyType" is recommended, which does not restrict the data content [44].

3.11. Reusing of Existing Vocabularies

Several well-known vocabularies are used in Linked Data to describe things such as people, places, and locations. By reusing known vocabularies, data publishers increase their chance of being interoperable with other parties as well as avoid the time consuming process of defining and documenting own vocabularies. In consequence, we mention a brief guide of reusing the vocabularies as an example:

- To describe simple data, use the basics of RDF and RDFS
- To name things, use "rdfs:label", "dcterms:title", and "foaf:name"
- To describe people, use FOAF and vCard
- To describe Web pages and other publications, reuse Dublin Core properties, for example "dcterms:creator" and "dcterms:description"
- To describe addresses, use vCard

4. Interlinking to other Datasets

Linked Data (LD) approach unlocks e-learning resources away from learners and enables enriching, navigation, casual discovery and improved resource seeking. Linking resources using LD also makes it easy for intelligent processing of data, as several operations e.g., integration of experiment data, consumption, and publication of experiment data are doable using the related tools. Particularly, the LD exposure of educational materials became a general approach specially for enrichment of learning resource as well as interlinking them to useful datasets on the Web of Data. To this end, some institutions have emerged their educational materials as LD. For example Europeana dataset [46], as European Union flagship digital library project, links the data providers metadata to other datasets such as DBpedia and Geonames [47] as well. As Figure 2 depicts, interlinking LOM elements to one dataset, makes the metadata global to access other valuable information over the LOD.

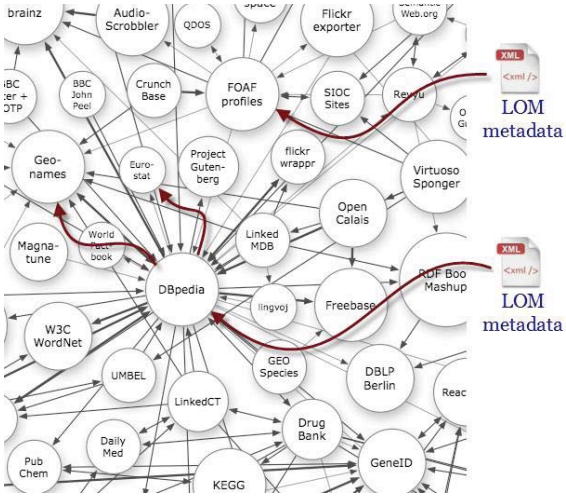


Fig. 2. Linking educational data to LOD

To this end, we examined all the IEEE LOM elements to discover the linkable elements to the LOD cloud. Figure 3 portrays the workflow we followed for the interlinking analysis. In the first step, we looked over those elements that potentially cannot be linked due to their specific values, and thus, they have been filtered out (e.g., "DateTime", controlled vocabularies). In the second step and being precise on the metadata records, we discovered that several elements (e.g., identifiers, vCard) contains local values defined by each repository according to its policy. Particularly, the values did not follow a specific rule for interlinking purpose. As a consequence of analysis, we found various elements such as "General.Title" and "Technical.Format" include string values that can be linked to the related datasets. However, after running an interlinking tool to link the data to a specific dataset, the outcomes were a few and in the most cases were not useful.

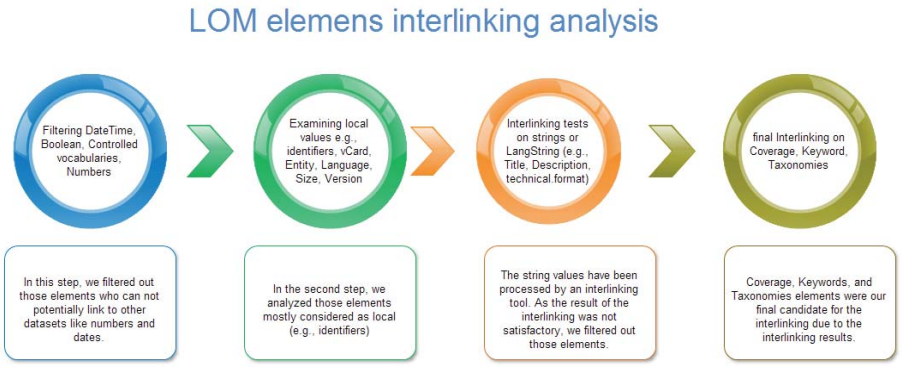


Fig. 3. Linking educational data to LOD

To be specific, linking coverage of a learning object to DBpedia, as a good practice, not only adds more geographical information about the place, but also allows metadata to be connected to other statistical sources (e.g., population, history) as well. In the

following sub-sections, we will summarize a couple of important elements of IEEE LOM, which can be linked to the LOD as well.

4.1. Linking Elements to DBpedia

The DBpedia dataset includes structured information about persons, places and organizations. It features labels and abstracts for 10.3 million unique things in 111 different languages [8]. This dataset has been recently identified as a hub in the LOD cloud [48], as it connects a wide variety datasets together with high centrality. Particularly in eLearning context, Lama et al. [49] presented an approach that automates the classification of learning objects and improves its search in repositories by annotating the learning objects with DBpedia ontology. As we will discuss later, DBpedia is also significant place for linking coverage of educational materials ("General.Coverage" in LOM) to regions, countries and cities of DBpedia Other datasets such as GEMET [50] and Eurostat [51] can be used for this purpose, as they include useful information about statistics of public places. The "Keyword" element of learning objects ("General.Keyword" in LOM) can be also linked to DBpedia, as we found a lot of similarities between the keywords of aggregated e-learning resources and the DBpedia labels.

4.2. Linking the "Classification" Category of IEEE LOM to LOD

The IEEE LOM provides an area for annotating and classifying educational resources and makes them discoverable specially when a learning resource is accessible based upon the classification it belongs. It expresses the classification of a learning object in the classification category that each classification includes purpose and taxonpath. The taxonpath states the structure of the taxonomy. One of the possibilities of classification interlinking, for example, is linking the taxonomy of a learning object to the LOD taxonomy dataset [52]. This dataset as a knowledge base provides informative LOD URIs for species concepts that improve the quality and stability of links between a species and the related data. There exist around 108,175 species concepts and 1,000 records for species occurrences [53], interlinked with the GeoNames dataset [47]. The following example illustrates part of Organic.Edunet metadata linked to the LOD taxonomy dataset through the classification category.

```
<http://youtube.com/example_resource> lom:classification
_:classification1.

_:classification1 lom:purpose lomvoc:discipline;
        lom:classificationDescription
"This classification provides many examples of Organic
Principles and Ontologies. @en";

lom:taxonPath _:taxonpath1.
```

```
_:taxonpath1 lom:taxonSource "LOD taxonomy" @en;
             lom:taxon  _:taxon1, _:taxon2.
_:taxon1 lom:id

"http://lsd.taxonconcept.org/describe/Organic_farming";
       lom:entry  "Organic farming category" @en.
_:taxon2 lom:id
"http://lsd.taxonconcept.org/describe/Certification";
lom:entry  "Certification" @en.
```

4.3. Linking the "Relation" Category of IEEE LOM to LOD

The "Relation" category of IEEE LOM groups features that establish the relationship between the learning object and other related learning objects. As learning objects may include different relations, they can be exposed in RDF in different intermediate nodes. The following example shows the relation of our sample learning object to DBpedia represented in RDF. In particular, the learning object is linked to many related resources exist in the DBpedia dataset through the Relation category.

```
<http://youtube.com/example_resource>
             lom:relation _:relation1.

_:relation1lom:relationKind  dcterms:isPartOf
          lom:relatedResource _:resource1;
            lom:resourceDescription
_:resourceDescription1.

 _:resource1 lom:relatedResourceCatalog "URI";
             lom:relatedResourceEntry
             "http://live.dbpedia.org/page/Agriculture".

_:resourceDescription1 rdf:value "Organic farming is kind
of agriculture that has been explain" @en.
```

5. Architecture and Implementation

The RDF binding of LOM elements is not sufficient for exposing educational materials as Linked Data, as Linked Data principles [5] should be covered by an educational repository in order to have the learning resources in a linkable way. To this aim, the repositories cater a service or API which users are able to make queries via SPARQL endpoint [11]. Repositories can also provide an RDF dump [53] which makes the whole dataset to be accessible through the repository website. Here, we propose an architecture along with a software prototype implemented on the Organic.Edunet [17] repository as our case study.

5.1. Exposing Organic.Edunet Resources as Linked Data

As previously mentioned, Organic.Edunet is a learning portal that provides access to digital learning resources as well as their metadata on Organic Agriculture and Agroecology and aims to facilitate access, usage and exploitation of such content. Several types of e-learning resources including reports, handbooks, presentations, experiments and lesson plans are available through the portal [17]. The LD exposure of the Organic.Edunet metadata [54] was performed by taking the following steps:

Initially, educational metadata were stored in the Organic.Edunet repository in XML format based upon an IEEE LOM Application Profile [55]. We transformed the XML files into a relational database by developing a transformer tool. In consequence, we exposed the metadata as Linked Data by using a mapping tool (e.g., D2RQ [56] as a mapping tool for mapping relational databases to RDF). In particular, we represented the educational metadata in a complete uniform dataset and made them accessible via a SPARQL endpoint and a RDF dump. The proposed architecture is presented in three layers as Figure 4 depicts.

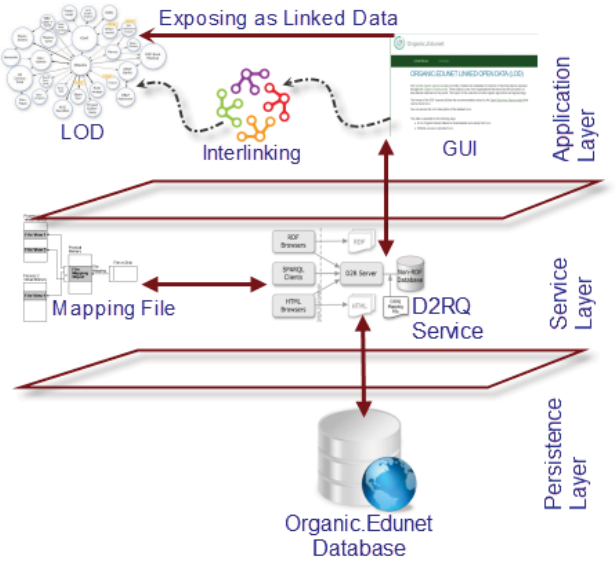


Fig. 4. The LD architecture of Organic.Edunet

In the persistence layer, the metadata are collected in the Organic.Edunet repository and converted the XML files into a relational database by developing a Java program. In the service layer, a D2rQ service mapped the relational database to the RDF format. We created a mapping file in order to express the relational structure to the RDF triples. In the application layer, we implemented an interface in front-end to depict the educational metadata in a graphical user interface (GUI) along with a search interface for users. Particularly, the SPARQL Endpoint and RDF dump of dataset made the data to be available through the GUI. We established a link between the Organic.Edunet dataset and DBpedia by mapping around 11,093 metadata records in the relational database to

RDF. This was performed by running a simple code to find the similarities between the metadata elements and DBpedia. Table 4 illustrates the matches between Organic.Edunet and the DBpedia dataset for "Keyword" and "Coverage" elements (although the interlinking analyzed have been based upon equal string match without any consideration of polysemy and lexical variants). Finally, around 73% of coverage of the learning objects (e.g., countries and cities) and 23% of keywords matched to the DBpedia concepts. Upon this finding, it is reasonable to conclude that the IEEE LOM elements include latent potential for linking to other datasets on the Web of Data.

Table 4. Interlinking Organic.Edunet to DBpedia

Metadata element	Total number	Matched number
Keyword	99,506	22,087 (23%)
Coverage	11,906	8,585 (73%)

5.2. Performance Testing over the Case Study

Regarding the performance testing of implementation, we used JMeter [57], as a testing tool for performance measurement and selected three queries to simulate the work as well. The queries became more complex from query 1 to query 3 according to Semantic Publishing Benchmark (SPB)[1], as a LDBC[2] benchmark for testing the performance of RDF engines (consider the appendix). SPB defines a set of "choke points" to evaluate the reliability of RDF database and address the complexity of queries. In particular, "join ordering" as one the choke points, tests the ability to consider cardinality constraints and decide which type of join should be used in a query, as it has been pointed out by other studies as well (e.g., [58]). We simulated as well as evaluated the queries on the same machine over D2RQ service and a triple store for 1, 5 and 10 users to compare the performance between them. Each query was repeated for five times in order to examine the results precisely. The Linked Data version of the Organic.Edunet was evaluated by making use of the query pages of D2RQ service. As Table 5 shows, the performance of queries decreases when they are run by more users. Obviously, the response time increases when they become more complex. As it can be seen from the table, there is a huge difference between response time of RDB and D2RQ services, as D2RQ performs both mapping the queries to SQL and running them over the relational database at once.

We also examined the implementation on a triple store in order to analyze the performance of executing the queries on a triple store directly. To this end, we imported the RDF dump of Organic.Edunet dataset in a Jena-Fuseki triple store [59] and evaluated the queries via its SPARQL page, as Table 6 illustrates the result of JMeter.

[1] http://ldbcouncil.org/benchmarks/spb
[2] http://ldbcouncil.org/

Table 5. Performance testing on D2RQ mapping service (5 times running of each query)

Query #	1 user	5 users	10 users
Query 1	661 ms	700 ms	746 ms
Query 2	526 ms	1773 ms	3632 ms
Query 3	1356 ms	4317 ms	9778 ms

Table 6. Performance testing on a triple store (5 times running of each query)

Query #	1 user	5 users	10 users
Query 1	8 ms	14 ms	13 ms
Query 2	10 ms	15 ms	44 ms
Query 3	67 ms	72 ms	200 ms

Executing the queries on a relational database (RDB), we realized that the analysis of executing the queries on RDB and triple store is comparable, as Figure 5 depicts the difference between these two data stores in terms of response time. The queries in both cases are shown on the x-axis (for 10 users), while the y-axis illustrates the runtime in milliseconds. Comparing these results with the mapping approach mentioned earlier, we can conclude that the D2RQ mapping service is not scalable when the number of users increases and queries become more complicated.

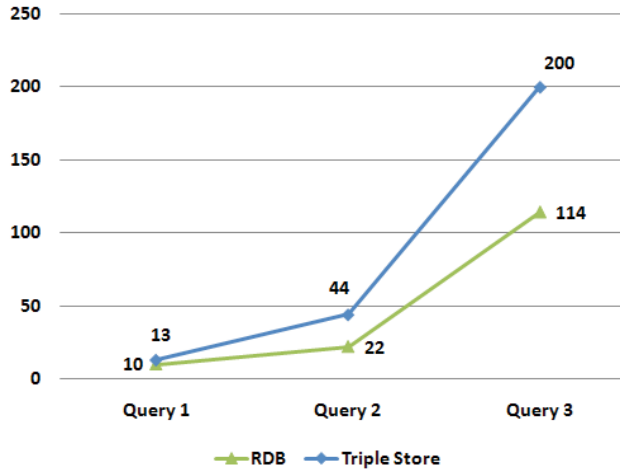


Fig. 5. Response time comparison between relational database and triple store in

6. Evaluation of the Case Study

Evaluating the interlinking results between the Organic.Edunet and DBpedia datasets, we realized that the "Coverage" element of e-learning resources in the Orgainc.Edunet repository includes information about countries and places that can be connected to the

DBpedia places. In particular, a user in Organic.Edunet can explore the dataset and obtain useful knowledge about the country or place of resources. To take a scenario about the advantage of such interlinking through the "Keyword" element, a teacher in agricultural science might explore the contents to find an article about "organic farming". In one of the results, a relevant book catches the teacher's attention and thus follows the keywords of article to find out the exact context of the learning resource. The researcher has never come across the specific terms which do not yield any more relevant data. As the learning resources in Organic.Edunet are interlinked to DBpedia, more information on topic including different translations are presented to him, when he is redirected to the DBpedia pages.

As a consequence of quality control of interlinked data in Organic.Edunet repository, we selected 20 random resources enriched by DBpedia over the "Coverage" and "Keyword" elements and presented to five end users. The Organic.Edunet resources included a full metadata information and we asked the users to explore especially the "Keyword" element linked to the DBpedia pages. In particular, the users were asked to answer 4 questions regarding the interlinked metadata elements. The questionnaire included the following statements regarding the linked items:

1. Was the link available to evaluate?

Here we asked whether the user can reach the target by clicking the provided URL or not? (As some links might not be available either the link is broken or it does not respond in a reasonable time).

2. Was the link information related to the term?

The relatedness of information to the term is evaluated by the user in the question above. It is possible that the provided information in the target semantically is not the same as source due to e.g., polysemy or ambiguity between them. For instance, there exist several abbreviations in the "Keyword" element (e.g., TOF, SDW...) which might refer to different terms.

3. Did the link information help you to find more useful data regarding the resource?

The most important question, from the authors perspective, was the usefulness of provided link. Overall, the users were asked if the link information in the target included useful knowledge about the resource and particularly could help learners to obtain usable data.

4. What do you recommend for improving the quality of interlinking?

Finally, we asked users to write their comments regarding the improvement of this experience.

Table 7. users' answers to the questions (5 times running of each query)

Question #	User 1	User 2	User 3	User 4	User 5
Q1	20	18	20	19	19
Q2	16	15	14	15	16
Q3	12	13	13	12	14

As Table 7 illustrates, almost all of the links were available for the evaluation. Also, the average number of resources that were relevant to the terms, was around 15 resources (76%). This amount of resources implies that a few number of them included

ambiguity or not informative for users to examine. To gauge the responses reliability of Question 3, we applied intra-class correlation coefficient (ICC) [60], as one of the popular reliability statistics, to determine the internal consistency of multiple raters. In this approach, the accepted value for describing internal consistency is defined by an alpha greater than 0.6 and the results is highly coefficient when value is more than 0.9. We later imported the users' answers into SPSS to analyze the responses and run the reliability statistics. Accordingly, the software output for our data was 0.726 that shows the users agreed on the results and the average number of questions determined by the users as useful was around 13 (65% of all questions).

Regarding the Question 4, one of the users commented that interlinking Organic.Edunet to the DBpedia dataset gives general information about the terms to readers, but if users want to obtain more information about the resource (e.g., relevant books or articles), they have to explore the Web. Interlinking Organic.Edunet to more educational and scientific datasets (e.g., universities) was also recommended by the user. Other users did not mention any important comments.

7. Conclusion

The widespread adoption of the Linked Data approach has led to the availability of huge amount of data ranging from public domain such as DBpedia to domain-specific space, for example Europeana which includes data about cultural heritage. Connecting e-learning resources to the LOD makes educational materials linkable to other useful datasets as well as enriches the contents as well.

To this aim, we discussed mapping and linking of the IEEE LOM elements, as an accredited metadata schema for describing educational materials, to the Linked Data based upon its principles. We developed an implementation of this approach for the Organic.Edunet repository, as our case study, so that the metadata of the e-learning resources became accessible through a graphical user interface. The metadata were also linked to other datasets in LOD (e.g., DBpedia). At the time of this research, the SPARQL endpoint of the Organic.Edunet dataset is available for users to make queries. Likewise, other educational datasets can foster their data to the released dataset. Eventually, some selected queries passed a performance testing on both relational database and triple store considering their complexity. The analysis of performance testing states that providing a powerful triple store on top of the Linked Data exposure of e-learning repositories dramatically improves the performance than using a mapping tool to convert the data as Linked Data format.

Acknowledgments. The work presented in this paper has been part-funded by the European Commission under the ICT Policy Support Programme CIP-ICT-PSP.2011.2.4 - elearning with project No. 297229 "Open Discovery Space (ODS)". The authors also would like to thank Dr. Hannes Ebner and Matthias Palmér for their assistance and helpful suggestions about the Linked Data exposure of Organic.Edunet.

References

1. IEEE LTSC, IEEE Standard for Learning Object Metadata 1484.12.1-2002, Final Draft version, [Online]. Available: http://ltsc.ieee.org/wg12/20020612-Final-LOM-Draft.html. [Accessed: 22-February-2014].
2. Ochoa, X. and Duval, E.: Towards automatic evaluation of learning object metadata quality. In Proceedings of the 2006 international conference on Advances in Conceptual Modeling: theory and practice, Berlin, Heidelberg, 372–381. (2006)
3. Ochoa, X., Klerkx, J., Vandeputte, B. and Duval, E.: On the use of learning object metadata: the GLOBE experience, Proceedings of the 6th European conference on Technology enhanced learning: towards ubiquitous learning, Berlin, Heidelberg, 271–284. (2011)
4. GLOBE | Connecting the World and Unlocking the Deep Web. [Online]. Available: http://globe-info.org/. [Accessed: 09-Jun-2013].
5. Bizer, C., Heath, T., and Berners-Lee, T.: Linked Data - The Story So Far. International Journal of Semantic Web Information System, vol. 5, no. 3, 1–22, 33. (2009)
6. Sicilia, M.-A., Ebner, H., Sanchez-Alonso, S., Alvarez, F., Abian, A., and Garcia-Barriocanal, E.: Navigating learning resources through linked data: A preliminary report on the re-design of Organic.Edunet, presented at the 1st International Workshop on eLearning Approaches for the Linked Data Age. (2011).
7. The Linking Open Data cloud diagram. [Online]. Available: http://lod-cloud.net/. [Accessed: 14-May-2013].
8. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S.: DBpedia - A crystallization point for the Web of Data, Web Semantics, vol. 7, no. 3, 154–165. (2009).
9. M. D'Aquin. Linked Data for Open and Distance Learning, Common Wealth of Learning, 2012. [Online]. Available: http://elearningeuropa.info/en/directory/Linked-Data-for-Open-and-Distance-Learning. [Accessed: 22-February-2014].
10. OER Commons. [Online]. Available: http://www.oercommons.org/. [Accessed: 17-Jun-2013].
11. SparqlEndpoint Description. W3C Wiki. [Online]. Available: http://www.w3.org/TR/sparql11-service-description/. [Accessed: 17-Jun-2013].
12. Simperl, E., Wölger, S., Thaler, S., Norton, B., and Bürger, T.: Combining human and computation intelligence: the case of data interlinking tools, International Journal of Metadata and Semantic Ontologies, vol. 7, no. 2,77–92. (2012)
13. Hausenblas, M., Halb, W., and Raimond, Y.: Scripting User Contributed Interlinking, in Proceedings of the 4th workshop on Scripting for the Semantic Web (SFSW2008), co-located with ESWC2008. (2008)
14. Dublin Core metadata initiative, Dublin Core Metadata Element SetVersion 1.1, Dublin Core metadata initiative, Dublin Core Metadata Element SetVersion 1.1. [Online]. Available: http://dublincore.org/documents/dces/. [Accessed: 22-February-2014].
15. Nilsson, M., Palmer, M., and Brase, J.: The LOM RDF Binding - Principles and Implementation, Proceeding of third annual ARIADNE Conference. (2003)
16. Dodds L., and Davis, I.: Linked Data Patterns: A pattern catalogue for modeling, publishing, and consuming Linked Data, Linked Data Patterns Pattern Cat. Model. Publ. Consum. Linked Data. (2010)
17. Organic.Edunet Project. [Online]. Available: http://wiki.organic-edunet.eu/index.php/Main_Page. [Accessed: 22-February-2014].
18. Klyne, G., and Carroll, J. J., Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C Recommendation. (2004)
19. IMS Global Learning Consortium, [Online]. Available: http://www.imsglobal.org/. [Accessed: 22-February-2014].

20. Representation of dates and times. International Organization for Standardization, ISO 8601:2004, 2004. [Online]. Available: http://www.iso.org/iso/catalogue_detail?csnumber=40874. [Accessed: 22-February-2014].
21. Flack, I., and Evan, B.: Metadata for the uninterested. [Online]. Available: http://conferences.alia.org.au/online2011/papers/paper_2011_B4.pdf. [Accessed: 22-February-2014]. (2011)
22. Advanced Distributed Learning (SCORM). [Online]. Available: http://www.adlnet.gov/scorm. [Accessed: 13-May-2013].
23. Currier, S.: Metadata for Learning Resources (MLR): An Update on Standards Activity for 2008 | Ariadne: Web Magazine for Information Professionals. 2008. [Online]. Available: http://www.ariadne.ac.uk/issue55/currier/. [Accessed: 22-February-2014].
24. Hoel T., and Mason, J.: Expanding the Scope of Metadata and the Issue of Quality, in Proceedings of the 19th International Conference on Computers in Education, Chiang Mai, Thailand. (2011)
25. Learning Resource Metadata Initiative :: About the LRMI. [Online]. Available: http://www.lrmi.net/about. [Accessed: 22-February-2014].
26. Dietze, S., Sanchez-Alonso, S., Ebner, H., Yu, H. Q., Giordano, D., Marenzi, I., and Nunes, B. P.: Interlinking educational resources and the web of data: A survey of challenges and approaches, Program Electron. Libr. Inf. Syst., vol. 47, no. 1, 60–91. (2013)
27. Dietze, S., Yu, H. Q., Giordano, D., Kaldoudi, E., Dovrolis, N., and Taibi, D.: Linked education: interlinking educational resources and the Web of data. in Proceedings of the 27th Annual ACM Symposium on Applied Computing, New York, NY, USA, 366–371. (2012)
28. Mitsopoulou, E., Woodham, L., Balasubramaniam, C., Poulton, T., Protopsaltis, A., and Dietze, S., mEducator: multi type content repurposing and sharing in medical education," Acad. Subj. Cent. Med. Dent. Vet. Med. Newsl. 01, vol. 22, 26–28. (2010)
29. Fernandez, M., Aquin, M. d', and Motta, E.: Linking data across universities: an integrated video lectures dataset, in Proceedings of the 10th international conference on The semantic web - Volume Part II, Berlin, Heidelberg, 49–64. (2011)
30. LinkedUp: Linking Web Data for Education. [Online]. Available: http://linkedup-project.eu/. [Accessed: 13-May-2013].
31. Linked Universities :: Home. [Online]. Available: http://linkeduniversities.org/lu/. [Accessed: 13-May-2013].
32. The ARIADNE Foundation, [Online]. Available: http://www.ariadne-eu.org/. [Accessed: 22-February-2014].
33. DCMI Metadata Terms. [Online]. Available: http://dublincore.org/documents/dcmi-terms/. [Accessed: 22-February-2014].
34. LOM to DCAM mapping document. [Online]. Available: http://dublincore.org/moinmoin-wiki-archive/educationwiki/attachments/LOM-DCAM-newdraft.pdf. [Accessed: 22-February-2014].
35. MARC/MODS RDFizer - SIMILE. [Online]. Available: http://simile.mit.edu/wiki/MARC/MODS_RDFizer. [Accessed: 22-February-2014].
36. MIT OpenCourseWare | Free Online Course Materials. [Online]. Available: http://ocw.mit.edu/index.htm. [Accessed: 17-Jun-2013].
37. Kunze, T., Brase, J., and Nejdl, W.: Editing Learning Object Metadata: Schema Driven Input of RDF Metadata with the OLR3-Editor. (2002)
38. Alog-Crisan, R. B., and Roxin, I.: RDF Editor for LOM. in Proceedings of International Conference e-Learning IADIS 2007, Lisbon, Portugal, 2007.
39. Cool URIs for the Semantic Web. [Online]. Available: http://www.w3.org/TR/cooluris/. [Accessed: 22-February-2014].
40. GoodURIs - W3C Wiki. [Online]. Available: http://www.w3.org/wiki/GoodURIs. [Accessed: 22-February-2014].

41. Turtle - Terse RDF Triple Language. 2011. [Online]. Available: http://www.w3.org/TeamSubmission/turtle/. [Accessed: 22-February-2014].
42. Miličić, V.: Problems of the RDF model: Blank Nodes, [Online]. Available: http://milicicvuk.com/blog/2011/07/14/problems-of-the-rdf-model-blank-nodes/. [Accessed: 22-February-2014].
43. Alvestrand, H.: Tags for the Identification of Languages, RFC 1766, Mar-1995. [Online]. Available: http://www.ietf.org/rfc/rfc1766.txt. [Accessed: 22-February-2014].
44. Fallside, D. C., and Walmsley, P.: XML Schema Part 0: Primer Second Edition, W3C Recommendation, 28-Oct-2004. [Online].
Available: http://www.w3.org/TeamSubmission/turtle/. [Accessed: 22-February-2014].
45. Representing vCard Objects in RDF, W3C Member Submission 20 January 2010. [Online]. Available: http://www.w3.org/Submission/vcard-rdf/. [Accessed: 22-February-2014].
46. Europeana Professional - Linked Open Data. [Online].
Available: http://pro.europeana.eu/web/guest/linked-open-data. [Accessed: 18-Jun-2013].
47. GeoNames. [Online]. Available: http://www.geonames.org/. [Accessed: 18-Jun-2013].
48. Rajabi, E., Sanchez-Alonso, S., and Sicilia, M.-A.: Analyzing Broken Links on the Web of Data: an Experiment with DBpedia, Journal of the Association for Information Science and Technology, Vol 65, 1721–1727, doi: 10.1002/asi.23109.
49. Lama, M., Vidal, J. C., Otero-Garcia, E., Bugarin, A., and Barro, S.: Semantic Linking of a Learning Object Repository to DBpedia. 2011. In 11th IEEE International Conference on Advanced Learning Technologies (ICALT), 460–464. (2011)
50. GEMET Dataset. [Online]. Available: http://www.eionet.europa.eu/gemet/. [Accessed: 18-Jun-2013].
51. Eurostat - Linked Data. [Online]. Available: http://eurostat.linked-statistics.org/. [Accessed: 18-Jun-2013].
52. TaxonConcept - Search Species. [Online]. Available: http://www.taxonconcept.org/search-species/. [Accessed: 18-Jun-2013].
53. DataSetRDFDumps - W3C Wiki. [Online].
Available: http://www.w3.org/wiki/DataSetRDFDumps. [Accessed: 18-Jun-2013].
54. Exposing Organic.Edunet eLearning resources as Linked Open Data. [Online]. Available: http://data.organic-edunet.eu/. [Accessed: 13-May-2013].
55. Organic.Edunet Metadata Application Profile - Organic.Edunet. [Online]. Available: http://wiki.organic-edunet.eu/index.php/Organic.Edunet_Metadata_Application_Profile.
[Accessed: 22-February-2014].
56. Bizer, C.: D2RQ - treating non-RDF databases as virtual RDF graphs. In Proceedings of the 3rd International Semantic Web Conference (ISWC2004), (2004).
57. Apache JMeter - Apache JMeterTM. [Online]. Available: http://jmeter.apache.org/. [Accessed: 22-February-2014].
58. Schmidt, M., Hornung, T., Meier, M., Pinkel, C., and Lausen, G.: SP2Bench: A SPARQL Performance Benchmark. In Semantic Web Information Management, R. de Virgilio, F. Giunchiglia, and L. Tanca, Eds. Springer Berlin Heidelberg, 371–393. (2010)
59. Apache Jena - Fuseki: serving RDF data over HTTP. [Online]. Available: http://jena.apache.org/documentation/serving_data/. [Accessed: 13-May-2013].
60. Johnson W. D., and Koch, G. G.: Intraclass Correlation Coefficient, in International Encyclopedia of Statistical Science, M. Lovric, Ed. Springer Berlin Heidelberg, 685–687. (2011)

Appendix: SPARQL Queries

Query 1:Title and description of resources for higher education for 10 resources with the following complexity: Join ordering

```
PREFIX lom:
 <http://data.organic-edunet.eu/lom_ontology.owl#>
PREFIX dcterms: <http://purl.org/dc/terms/>

SELECT ?s ?f ?desc ?r WHERE {
   {?s dcterms:title ?f.}
   {?s lom:description ?d.?d dcterms:description ?desc. }
   ?s lom:educational ?e. ?e lom:educationalContext ?r.
   FILTER regex(str(?r),"highereducation", "i").
}
limit 10
```

Query 2: Resource format category along with count of them for the resources related to organic with the following complexity: Aggregation, Ordering, Join ordering, Search.

```
SELECT ?format (count(?format) as ?count) WHERE {
   {?s dcterms:format ?format.}
   {?s lom:description ?d.
    ?d dcterms:description ?desc.
   FILTER regex(str(?desc),"organic", "i").}
}
GROUP BY (?format)
ORDER BY DESC(?count)
```

Query 3: Title and web address of courses that are in html or pdf formats with the following complexity: Search, Ordering, Join ordering, Optionals with filters, Complex filter conditions

```
Select ?title ?location
WHERE {
 {?s lom:educational ?edu.
  ?edu dcterms:type ?r. Filter Regex(str(?r),"course","i").}
OPTIONAL {?s lom:technicalLocation ?location. }
OPTIONAL {?s dcterms:title ?title. }
OPTIONAL {
       ?s dcterms:format ?format.
       Filter ((?format="application/pdf") ||
(?format="text/html")) .}
}
Order by (?title)
```

Enayat Rajabi is a PhD researcher in the Computer Science Department of the University of Alcalá since 2012. He holds his MSc (2004) degree in Computer Engineering from the department of Ferdowsi University of Mashhad (Mashhad, Iran). He is presently a full time researcher in the Information Engineering research unit at the University of Alcalá (Alcalá de Henares, Madrid, Spain). The aim of his PhD is to interlink educational data using Semantic Web and Linked Data.

Miguel-Angel Sicilia is currently full professor at the Computer Science Department of the University of Alcalá and the head of the Information Engineering research unit, where he leads several European and national research projects in the topics of learning technology and Semantic Web. He is the Editor-in-Chief of the International Journal of Metadata, Semantics and Ontologies, published by Inderscience, and serves as a member of editorial board of many other scientific journals in the area of Semantic Web, Computational Intelligence and Information Systems. He has been active in metadata research and was awarded the 2006 Cyc prize for the best research paper.

Salvador Sanchez-Alonso is an associate professor in the Computer Science Department of the University of Alcalá, Spain. He has participated in and coordinated several EU-funded projects over the last few years related to TEL, learning object repositories and federations, notably eContent+ Organic.Edunet, CIP-PSP Organic.Lingua and VOA3R,. His current research interests include TEL, Learning repositories, Semantic Web and Computer Science education.

Received: March 30, 2014; Accepted: October 01, 2014

Title: Discovering Duplicate and Related Resources using an Interlinking Approach: The case of Educational Datasets

Publication:	Journal of Information Science
Authors:	Enayat Rajabi, Miguel-Angel Sicilia, and Salvador Sanchez-Alonso
Relationship with the global research objectives:	Interlinking an educational repository to 20 educational datasets on the Web of Data
Impact factor:	1.087 (2013)
Date of submission:	3-Nov-2014
Date of acceptance:	03-Feb-2015
Date of publication:	March 2015 (online)

Article

JIS

Journal of Information Science
1–13

DOI: 10.1177/0165551515575922
jis.sagepub.com
SAGE

Discovering duplicate and related resources using an interlinking approach: The case of educational datasets

Enayat Rajabi
University of Alcalá, Spain

Miguel-Angel Sicilia
University of Alcalá, Spain

Salvador Sanchez-Alonso
University of Alcalá, Spain

Abstract
Linking a learning dataset to useful information on the Web of Data enriches its learning resources, as it enhances learners' knowledge. This enrichment is usually achieved by creating links between datasets using the interlinking tools, which facilitate connecting any kind of data in a semi-automatic manner. This paper evaluates the interlinking results between an e-learning repository and several educational datasets on the Web of Data, which leads to enrichment of the contents. Many related resources were discovered during this experimentation already matched to the GLOBE learning objects. Furthermore, this research presents a data model to find similarity between two datasets and a workflow to identify the duplicate resources by performing a semi-automatic evaluation process. A case study was also assessed by human experts.

Keywords
Duplicate; e-learning resource; GLOBE; interlinking; LIMES; linked data

1. Introduction

The Semantic Web, as a collaborative movement led by the World Wide Web Consortium (W3C), promotes common data formats for publishing data on the World Wide Web. The aim of Semantic Web is to convert the current Web, dominated by unstructured and semi-structured documents, into a 'Web of Linked Data'. It also facilitates the sharing and availability of different kinds of information on the Web. In particular, the Linked Data approach [1] has emerged as the *de-facto* standard for integrating data on the Web. It offers significant potential to tackle the interoperability issues in different contexts. In an e-learning context, for example, Linked Data enhances the discovery of Open Educational repositories contents established by the educational institutions [2] and connects the learning objects to useful knowledge on the Web [3, 4]. Data connectivity in Linked Data is performed by providing RDF links between two entities – so-called interlinking.

The Linked Data applications have also facilitated data enrichment by applying several techniques for automatic and intelligent linking. In particular, an interlinking tool establishes links between different datasets on the Web by discovering similarities among their entities. It also helps data publishers to connect their contents to useful datasets. In this paper, we evaluate the outcomes of an interlinking approach on a large learning repository, Global Learning Objects Brokered

Corresponding author:
Enayat Rajabi, Information Engineering Research Unit, Computer Science Department, University of Alcalá, 28805 Alcalá de Henares, Spain.
Email: enayat.rajabi@uah.es

Exchange (GLOBE) [5], by applying a promising interlinking tool. It connects the GLOBE resources to 20 educational datasets in the LOD cloud. We also assess the matched links to answer the following research questions:

- How are the GLOBE resources distributed in each target dataset and which datasets include more similarities with GLOBE?
- What are the benefits of this interlinking when a large learning dataset is linked to several educational datasets on the Web?
- How can the related and duplicate resources be identified by applying the interlinking approach?

The rest of the paper is structured as follows. Section 2 describes the importance of interlinking on the Web of Data and outlines the current studies in this context. In Section 3, we will present the proposed approach for interlinking and finding the duplicates. Section 4 will discuss the interlinking results and evaluate the approach using a case study. Finally, conclusions are presented in Section 5.

2. Background and related work

Interlinking tools perform the creation of links semi-automatically and connect two datasets using different kinds of links (e.g. owl:sameAs) by similarity discovery among the entities. Most of these applications follow a similar routine to carry out the interlinking process. For example, in Silk [6], a user should set the following information to run the tool:

- source and target datasets;
- source and target entities (e.g. resource title in the source dataset to book title in a library);
- criteria under which two entities are matched.

Given the criteria above, the software discovers similarities between pairs of entities and generates a set of results. Several studies have been undertaken in recent years to investigate the interlinking issues in the Linked Data context. Simperl et al. [7] have compared various linking tools by addressing the important aspects such as required input, resulting output, considered domain and matching techniques used. This comparison was applied from two specific perspectives: degree of automation (to what extent the tool needs human input) and human contribution (the way in which users are required to do the interlinking). Scharffe and Euzenat [4] also proposed a framework for data interlinking applied in different systems in which several linking tools were discussed. In a technology-enhanced learning context, Dietze et al. [3] documented an approach for interlinking educational resources based on the Linked Data principles [1] and exploiting the abundance of existing data on the Web. Several Linked Data projects such as LinkedUp [8] and Linked Education [9] have also been aimed at advancing the exploitation of the vast amounts of public, open data available in an educational context. In another empirical study, Rajabi et al. [10] applied two matching techniques to interlink a semi-structured dataset to the Web of Data and discussed the generated results in details. In the context of Open Educational Repositories, Piedra et al. [2] applied the Linked Data principles to interoperate and mash-up data from distributed and heterogeneous repositories of open educational materials. The same author, in another study [11], leveraged the principles of Linked Data to enhance the discovery of Open Course Ware (OCW) contents created and shared by the universities. The authors also developed a query method to access the OCW data using linked data techniques and linked the contents to the LOD cloud.

The aforementioned studies have demonstrated that several fundamental works have been carried out in this direction and data publishers can trust the interlinking tools to interconnect their contents to other datasets [12]. However, none of the mentioned studies investigates an interlinking approach between an educational repository and several e-learning datasets on the Web of Data. Furthermore, they do not mention duplicate identification amongst the interlinking results. Following our previous studies [10, 12, 13], we extended our approach on 20 educational datasets on the Web and scrutinized the results to discover the duplicate resources among the educational datasets.

3. Experimental setting

The GLOBE repository [5], which includes around 1 million learning object metadata [14], was selected as our source dataset. To examine the GLOBE resources and for the sake of exposing them as RDF [15], we harvested around 830,000 learning metadata from this repository and imported them into a relational database to analyse its metadata effectively and select the best possible elements for interlinking. All harvested files were in XML format based on IEEE LOM schema [16]. As we will discuss later, the candidate elements of GLOBE metadata were selected and exposed as RDF.

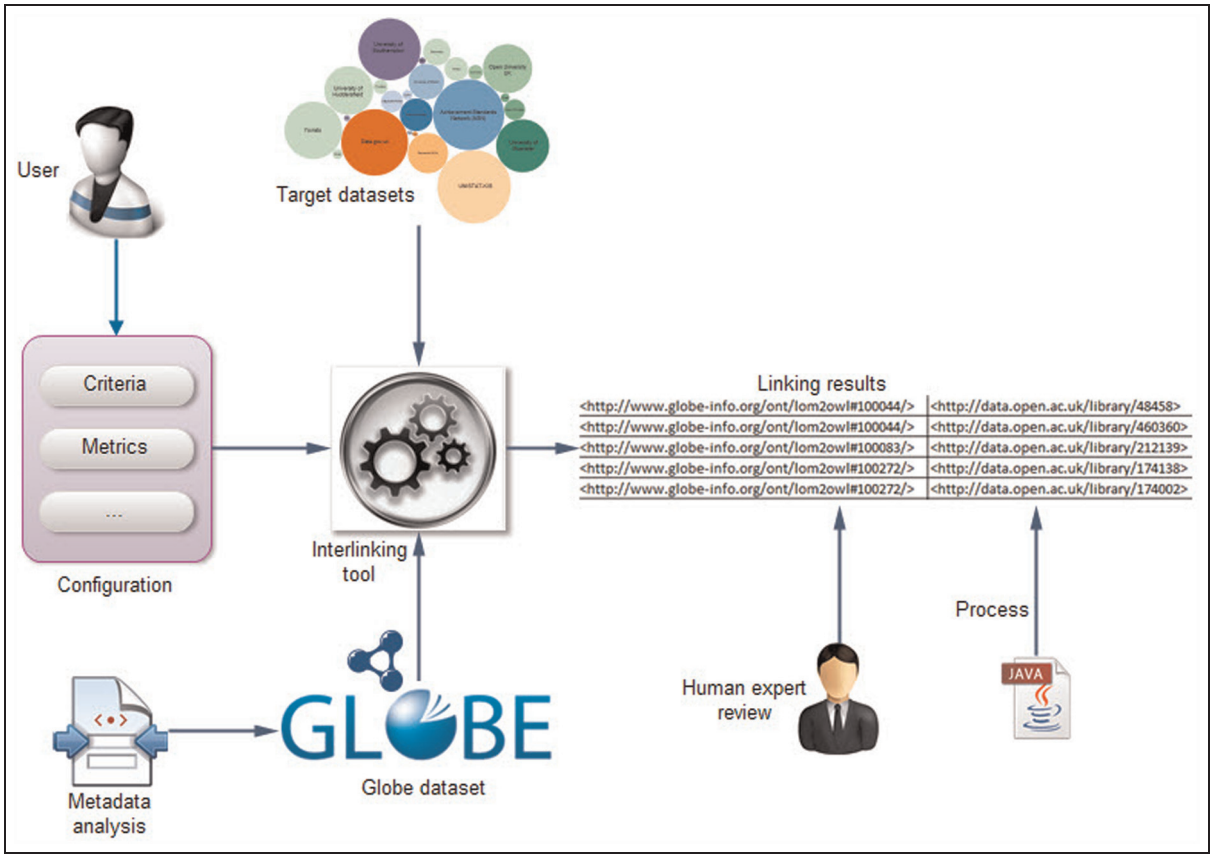


Figure 1. Workflow of proposed approach.

On the other hand, we collected a set of educational datasets on the Web of Data and prepared several queries to retrieve their available elements for interlinking. Finally, we carried out the interlinking between GLOBE and the selected educational dataset using an appropriate tool.

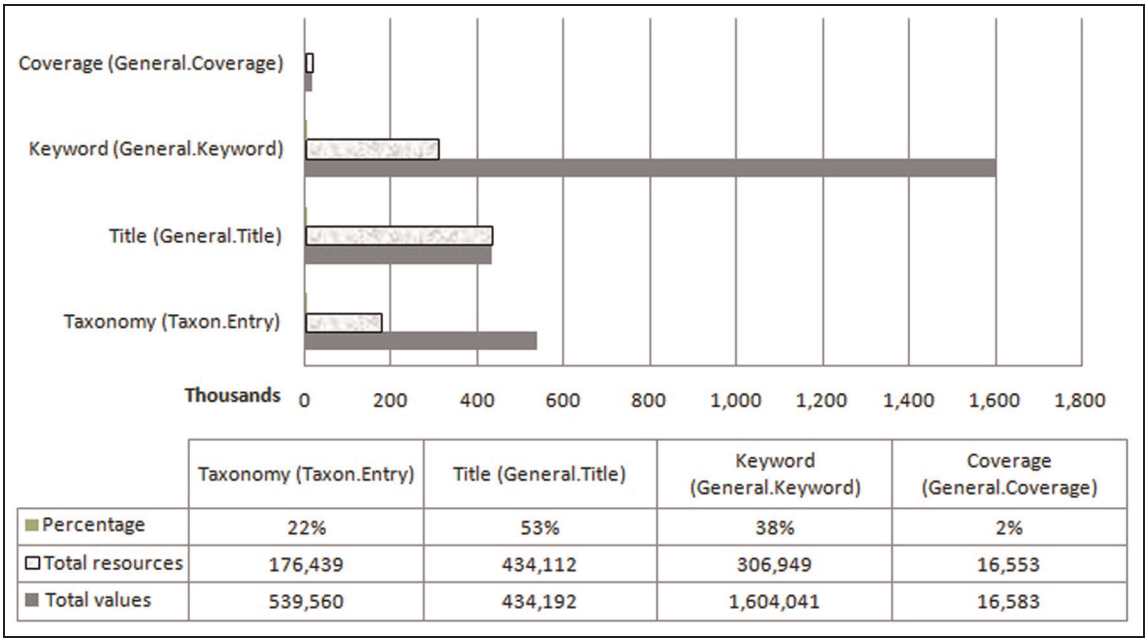
In a previous study [12], we evaluated several interlinking tools on the Web of Data and demonstrated that LIMES [17] is a promising tool in this context. In LIMES, a user specifies the endpoints of datasets, comparable entities and thresholds of acceptance of output. When a threshold is set to 0.98, for example, it means that two concepts are considered to be matched if their syntax similarity is more than 98%. The tool runs a number of matching techniques and reports the results to the user based upon the configuration and similarities between the two datasets. Figure 1 depicts the workflow we followed to perform the interlinking process, as we explained above. In brief, we used LIMES to interlink GLOBE to 20 datasets in the LOD cloud and analysed the results by writing a program. In the following subsections we will describe the analysis of the GLOBE metadata elements and target datasets.

3.1. Source dataset

We categorized the data types applied by the GLOBE metadata in Table 1 along with some examples. From the interlinking point of view, most of the elements cannot be used in the interlinking as they include 'Dates', 'Boolean' or controlled vocabularies. Focusing on the elements usage by the GLOBE resources, we realized that more than 90% of resources applied local values (e.g. identifiers), controlled vocabularies (Lifecycle.Status) and language codes, while the title of learning objects (General.Title) was highly used (97%) and more than half of the GLOBE resources included Keyword (61%) and Classification elements (59%) in their metadata. It should be noted that the Coverage element was only used by 7% of resources. As we discussed in the previous study [10, 13], four metadata elements including title ('General.Title'), coverage ('General.Coverage'), keywords ('General.Keyword') and classification taxonomy ('Classification.Taxon.Entry') were identified as candidate elements for interlinking.

Table 1. LOM elements data type and sample values

Data type	LOM element	Values examples
Boolean	Cost, Copyright	"'Yes'", "'No'"
Numeric	Technical.Size, Requirement.MinimumVersion, Requirement.MaximumVersion	"'15200'" , "'1.0'"
Local value	Identifier.Catalog, Identifier.Entry, Lifecycle.Version, Contribute.Entity, Technical.Location, Technical.InstallationRemarks, Technical.Other Platform Requirements, TypicalAgeRange, Description	"'http://localvalue.com/568545'", "'3.0'", "'No installation'", "'This is a learning object about agriculture'"
DateTimes	Contribute.Date, Duration, TypicalLearningTime	"'10P45M'"
Codes	Language codes	"'en'", "'en-US'", "'de'"
Controlled vocabularies	Structure, AggregationLevel, LifeCycle.Status, Contribute.Role, MetadataSchema, Technical.Format, Technical.Requirement.Type, Technical.Requirement.Name, InteractivityType, InteractivityLevel, Difficulty, SemanticDensity, LearningResourceType, Educational.Context, IntendedEndUserRole, Relation.Kind, ClassificationPurpose, TaxonPath.Source	"'atomic'", "'JPG'", "'low'", "'author'", "'PDF'"



	Taxonomy (Taxon.Entry)	Title (General.Title)	Keyword (General.Keyword)	Coverage (General.Coverage)
Percentage	22%	53%	38%	2%
Total resources	176,439	434,112	306,949	16,553
Total values	539,560	434,192	1,604,041	16,583

Figure 2. GLOBE elements in English language.

Given that the most prominent language of resources in GLOBE is English [14] and for the sake of manual evaluation of results by human experts, we selected those resources that provided the candidate elements in English language. Bearing this in mind, around 53% of GLOBE included English titles, while there were more than 1.6 million English keywords used by 38% of GLOBE (consider Figure 2). Regarding the other candidate elements, only 2% of GLOBE resources provided the Coverage element and 22% (around 176,000 resources) of taxonomy of learning objects were in English language. Having the selected elements for interlinking, we exposed the selected elements as RDF using a mapping service (D2RQ [18]) for mapping data to RDF and carrying out the interlinking afterwards.

3.2. Target datasets

To find appropriate targets for interlinking, we investigated several educational datasets in the LOD cloud. From a technical perspective, both source and target datasets should include either a SPARQL endpoint or an RDF dump. At first

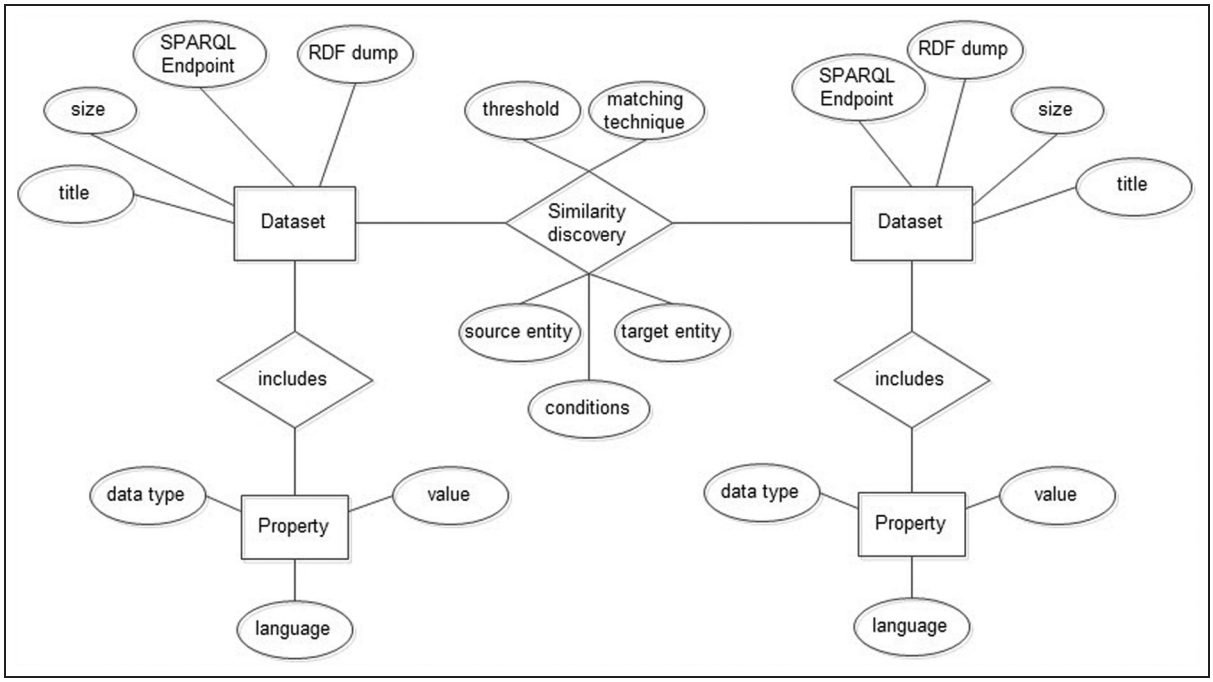


Figure 3. Similarity data model.

glance, it is obvious that most of the educational datasets lack any specific endpoint or RDF dump. Examining the datasets' endpoints illustrated that most targets were not accessible at the time of this research. Finally, we could collect 20 educational datasets who responded to the queries or included an RDF dump to download. Afterwards, we calculated the size of each dataset using SPARQL queries. Appendix 1 illustrates the size of datasets (in triples) along with their full name. Amongst candidate datasets, 'Charles University of Prague' with more than 93 million triples of publications, was the biggest dataset. 'Key Information Sets' (UNISTAT-KIS), which includes a set of information about full- or part-time undergraduate courses, was the second one found in this context, with more than 8 million triples.

3.3. Interlinking process

When running the LIMES tool, the output is a number of links in RDF (N-TRIPLE format) that connects source and target entities using the sameAs relationship. Appendix 2 illustrates a sample output generated by the tool that indicates that seven GLOBE resources were linked to 10 resources in the OpenUK dataset. As can be seen, four GLOBE resources were linked to more than one target resource, and three resources in OpenUK matched to more than one resource in GLOBE.

Figure 3 depicts the data model we followed to find similarities between two datasets. In this model, we showed each dataset along with its properties as entities and the similarities between datasets as relationships. Each dataset has a title, endpoint URI, size and other specifications as attributes. It may include many entities and have many similarities to other datasets. The similarity relationship may have different attributes itself for finding the related resources in two datasets. We will apply another workflow after the interlinking later in this paper.

It should be highlighted that throughout this study we used JAVA programming language, because of the many advantages that this programming environment provides, including various libraries (e.g., JSON, SET, Jena) in both analysis and Linked Data contexts. Figure 4 depicts the procedure we followed to perform the following tasks in our study:

- find the GLOBE resources linked to each target dataset;
- discover the total number of resources in GLOBE linked to all targets.

To address the first goal, we used a set, as a distinct list of elements, to remove the duplicates in both source and target datasets. In some of the outputs, we had to split the file into several small ones, as the size of file was more than 1 Gb (as

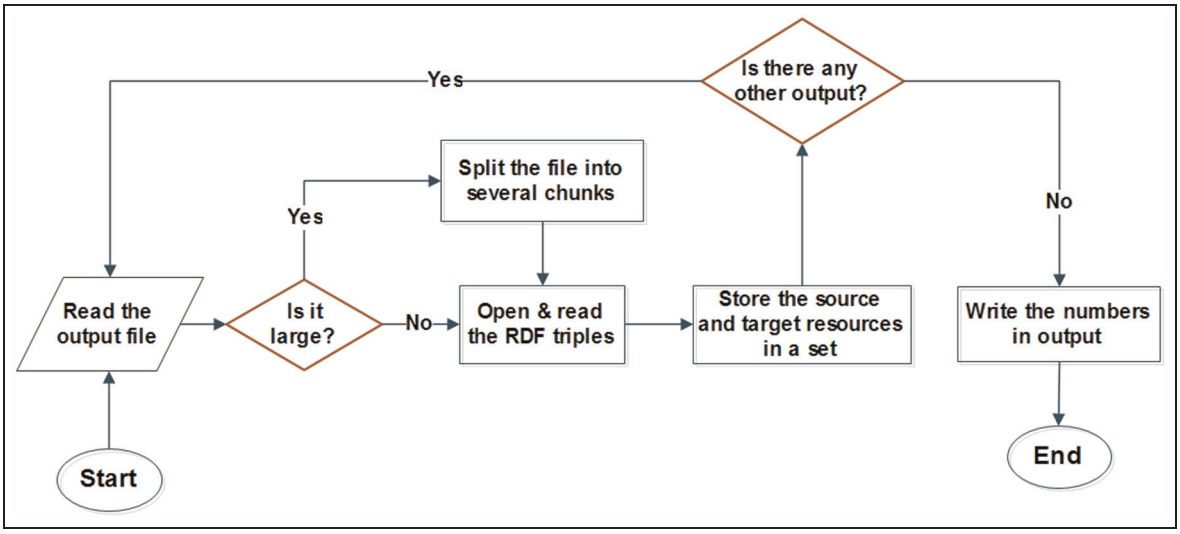


Figure 4. Workflow of finding linked resources in targets.

it included a million records) and the program could not process them with the available hardware resources. To achieve the second goal, we used the same approach extended for all datasets. In particular, the program retrieved the GLOBE resources in each output and added them to a final set to calculate the total number of resources overall. Appendix 2 illustrates the final output of LIMES result after running it against GLOBE and 20 educational datasets.

4. Discussion and results

As mentioned earlier, we used LIMES as the interlinking tool, to connect the candidate metadata elements (Title, Keyword, Taxon and Coverage) to the LOD datasets. Here, we categorize the results in four sections and present an analysis for each one.

4.1. Interlinking GLOBE elements to educational datasets

Figure 5 illustrates the interlinking results between GLOBE resources and the selected datasets. Figure 6 also depicts the GLOBE resource distribution among the target datasets. Below we report each element analysis in detail.

4.1.1. Title element. Figure 5 (a) depicts the interlinking results for the top five datasets with high similarities to the 'title' element. The *x*-axis in the figure refers to the number of GLOBE resources matched to the target dataset, while the *y*-axis shows the number of resources in the target dataset. In particular, Yovisto, with around 9117 resources, followed by OCW (the Open Course Ware consortium) and University of Bristol, had the greatest similarity to GLOBE. There were also 4127 resources in GLOBE connected to 2560 learning objects in The Open University of the UK, with around 13,600 matched links overall (see Appendix 3).

In total, five datasets (Data.gov.uk, Forge project, Semantic ISVU, MoreLab and Vergata) did not have any similarity with GLOBE. Table 2 also shows two GLOBE resources connected to several resources in the target datasets in which, for example, one resource about 'Nuclear Energy' matched three different datasets (ASN, Bristol and Huddersfield) specified with their URIs.

Analysing the interlinking outputs, we found that only a small number of GLOBE resources (around 24,000 overall) matched the target datasets through the title element. This result indicates that finding similarities for large texts is difficult for interlinking tools, as the resource titles in GLOBE mostly (around 83% of all English titles) include at least two words (e.g. 'Alternating Current Circuits'). Also, we realized that there were around 16,000 resources in GLOBE linked to at least two target datasets and 8260 resources linked to all of them. Figure 6(a) depicts the distribution of GLOBE resources among the target datasets (with more than 1% GLOBE distribution).

4.1.2. Keyword element. In the case of the Keyword element, the range of acceptance reported by LIMES was far larger than the title element (Section 4.1.1), as only one dataset (Semantic ISVU) did not have any similarity to any keywords.

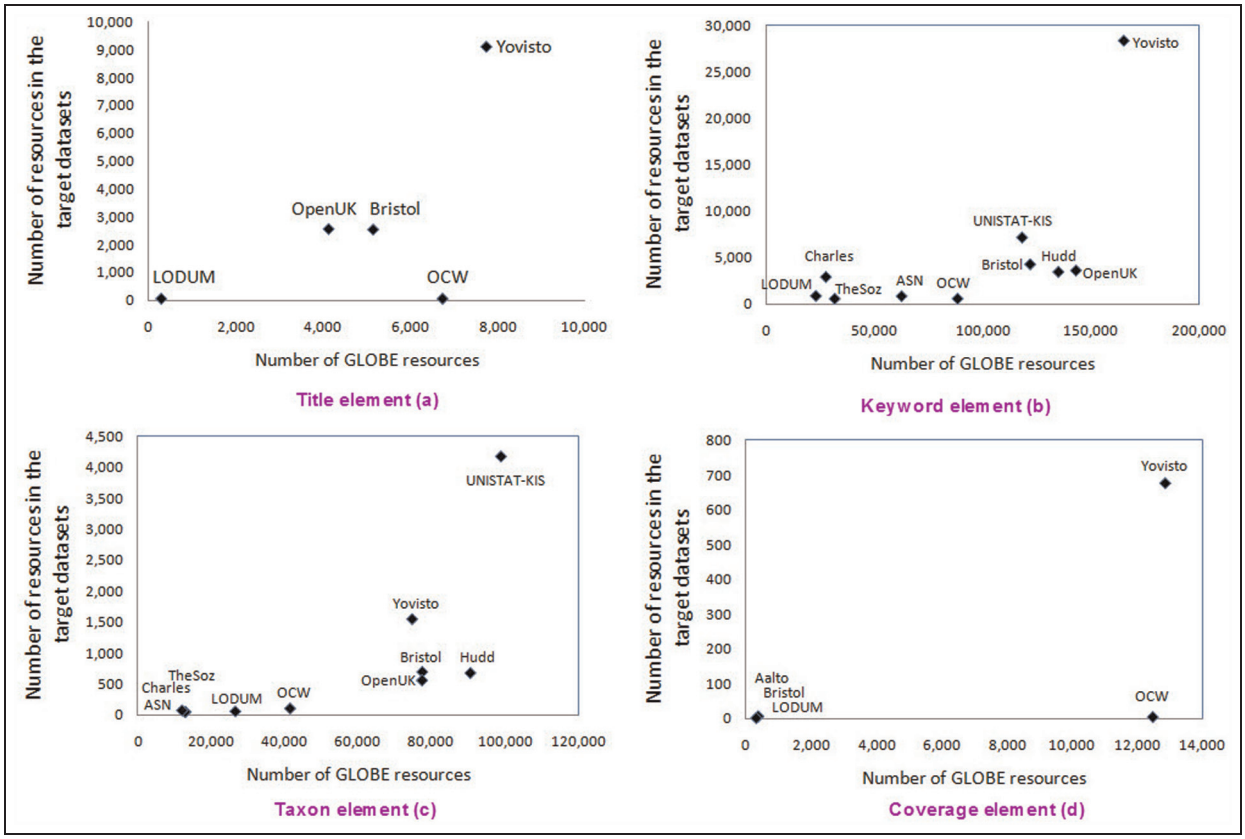


Figure 5. Interlinking results between GLOBE and target datasets based on four elements in GLOBE.

As mentioned earlier, there were more than 1.6 million English keywords in GLOBE ranging from science in education to environment literature. The large number of generated results for the Keyword element may refer to the fact that more than 50% of these keywords include exactly one term and 33% contain two words, which helps an interlinking tool to discover more similarities. As it can be seen in Figure 5(b), we observed that only one dataset (UNISTAT-KIS) had more than 6.7 million links to GLOBE (around 118,000 GLOBE metadata to 7166 resources in the target datasets).

Analysing the results, we realized that around 228,000 resources in GLOBE (74%) were matched to the target datasets and there was also a large amount of resources (almost 760,000) in common among all the results. Figure 6(b) shows the distribution of GLOBE resources among the target datasets in which Yovisto (an academic video search), the Open University of the UK and the University of Hudders field were the most referred datasets.

4.1.3. Taxon element. Most of the taxonomies of learning objects in GLOBE included terminologies in one or two words and referred to the classification of resources. In particular, around 60% of taxonomies in GLOBE contained only one word and almost 25% of them included two words ranging from science to historical concepts. As Figure 5(c) shows, around 99,000 resources in GLOBE were identified by LIMES and matched to more than 4000 resources in the UNISTAT-KIS dataset, followed by the University of Huddersfield (with 90,512 GLOBE resources) and the University of Bristol (with 77,420 GLOBE resources). Overall, only two datasets did not link to GLOBE and around 135,000 resources (76%) were connected to one or more datasets. Figure 6(c) also illustrates 13 datasets with more than 1% resource distribution in GLOBE, of which the UNISTAT-KIS dataset and the University of Huddersfield included the highest similarities to the GLOBE resources through the Taxon element.

4.1.4. Coverage element. As Figure 5(d) illustrates, eight datasets could link to GLOBE, but mostly with small numbers of results. Yovisto, as an exception, was connected to 13,000 GLOBE resources with 676 resources (with around 8 million links). There were also around 12,941 (78%) resources in GLOBE linked to all the target datasets (mostly to Yovisto and OCW). It should be highlighted that most of the matched terms referred to geographical places and countries. Figure 6(d) also shows that the references to Yovisto and OCW datasets were distributed throughout most of the GLOBE resources via the Coverage element.

Table 2. A sample of GLOBE titles matched to several resources

Title in both datasets	Globe resource	Target URI	Dataset name
Nuclear Energy	http://www.globe-info.org/ont/lom2owl# 108450	http://schools.nyc.gov/NR/rdonlyres/6C64098F-0C24-4B27-A22F-F542A2F97DA0/130926/TTS_G11_LiteracySSandScience_NuclearEnergy.pdf	ASN
		http://resrev.ilrt.bris.ac.uk/research-revealed-hub/publications/118933#pub	Bristol
		http://data.linkedu.eu/hud/book/118555	Huddersfield
Bibliography	http://www.globe-info.org/ont/lom2owl#178214	http://resrev.ilrt.bris.ac.uk/research-revealed-hub/publications/15140#pub	OpenUK
		http://data.uni-muenster.de/context/istg/allegro/6/210/T00244773	Muenster

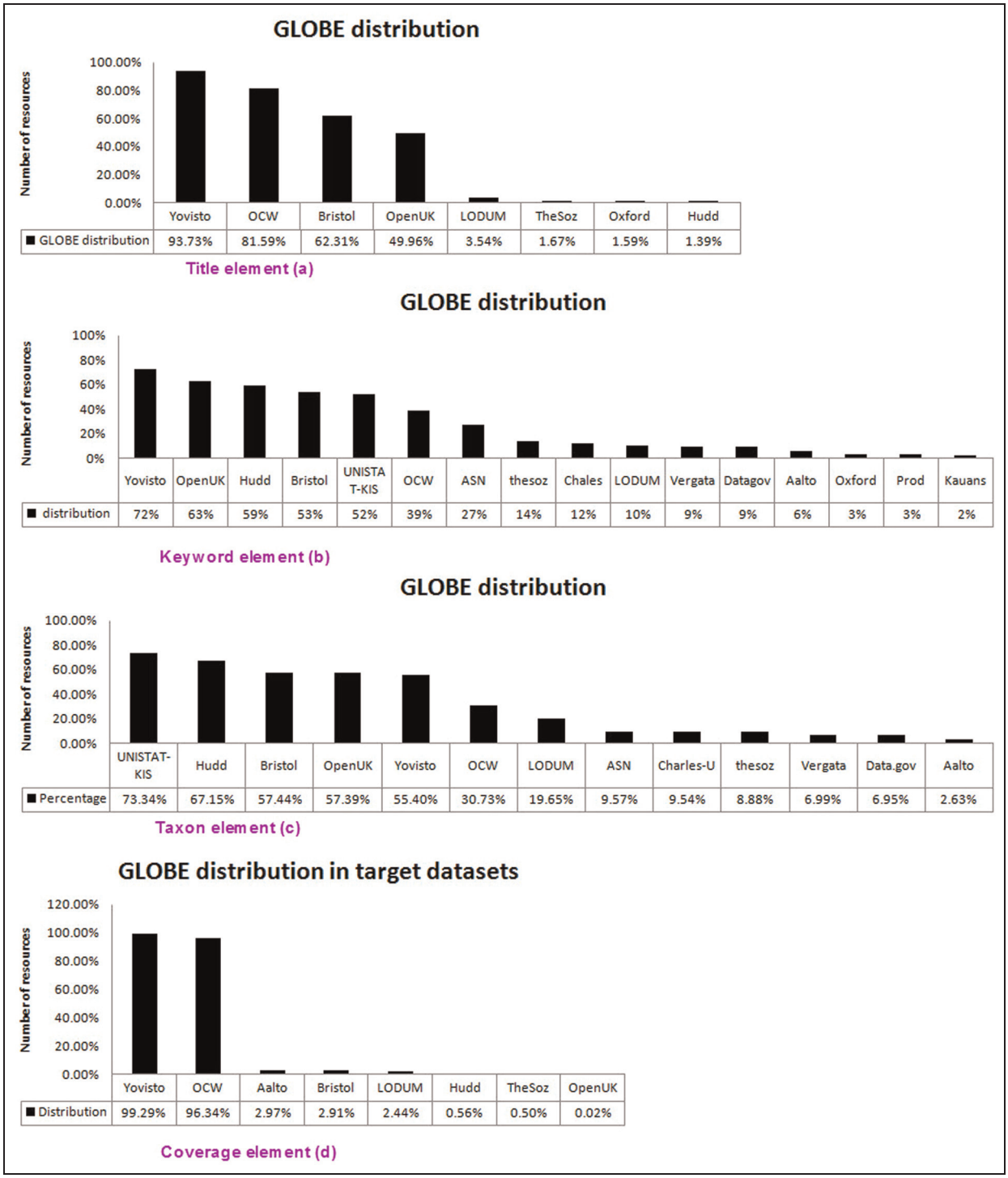


Figure 6. GLOBE resources distributions among target datasets over each element.

4.2. Human evaluation of the interlinking results

As we discussed earlier, the interlinking tool reported a set of records as an output with similar values in the source and target datasets. For example, the term 'Photosynthesis' was the title of a learning object in the ASN dataset (http://www.pbslearningmedia.org/resource/tdc02.sci.life.stru.photosynth/photosynthesis/) and the GLOBE repository (http://ariadne.cs.kuleuven.be/finder/globe/?query=Photosynthesis). However, the question under discussion is to what extent these learning resources are semantically matched or related. To this aim, we reviewed the generated results to discover

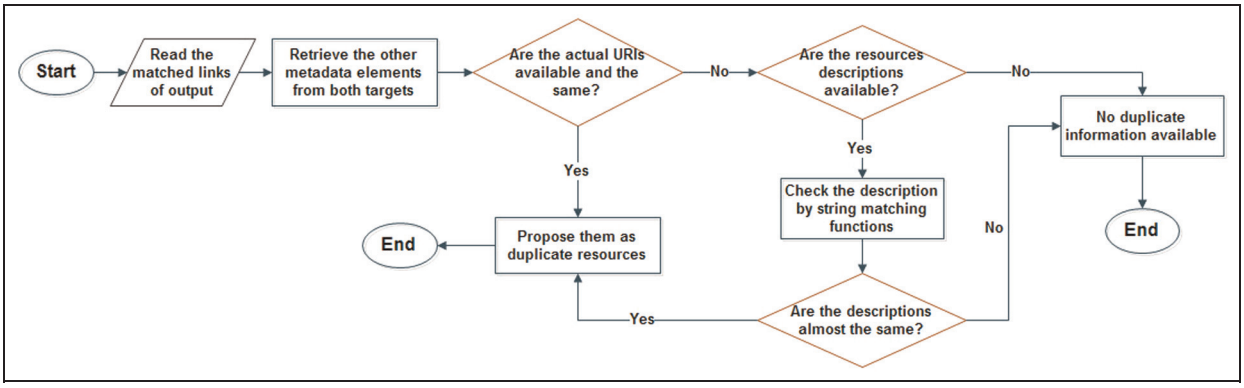


Figure 7. Duplicate finding workflow.

an appropriate target and evaluate the outputs manually. In the manual evaluation, we focused on duplicate and related resources, as we will discuss in the following sections.

4.2.1. Duplicate resources. As Figure 7 depicts and according to the data model proposed in Section 3.3, a workflow is presented for finding the duplicate resources. Having the interlinking results, we retrieved the other metadata elements including URI and description of learning objects from datasets after identifying their metadata schema (e.g. dcterms in Open UK). This task was carried out using the SPARQL queries. In the next step, we analysed the value of metadata elements. In particular, if the actual URIs of resources in both datasets point to the same internet address, they are proposed as duplicate resources. In the case of unavailability of URIs, the duplicate finding is focused on the descriptions of learning resources and analysing their values using the text-matching functions. If both resources have high similarities in their descriptions, they can also be presented as duplicates.

4.2.2. Related resources. From a technical perspective, if no similarities exist among the other metadata elements, the evaluation is continued on exploring the actual address of the resource (URI) where the learning object exists. This helps a domain expert to identify the relatedness of two resources semantically by exploring the content. Moreover, the other metadata elements of linked resources (such as description or subject) might be different in syntax, but an expert identifies them as related resources conceptually owing to their content similarity. As an example in our case study, a course about 'Latitude and Longitude' in Text/HTML format in GLOBE linked to a resource in the target dataset, but in another format (sound recording), which means that a human expert can identify their relatedness as well.

4.2.3. Case study. As a consequence of evaluating the records and given that the human evaluation of links manually requires significant effort, we selected the results between GLOBE and the Open University of the UK (OpenUK) on the title element as our case study by taking the following notes into account:

- Interlinking results usually include two identifiers which usually point to the internet addresses. Given that the resource identifiers were not implemented properly in some cases or they might be broken, we selected those that followed the good URIs [19].
- Despite providing available URIs, the metadata schema should be rich enough so that an expert can compare the information in both targets. For example, the target metadata in some cases included only three elements (title, format and subject), which is insufficient for the evaluation.

Nevertheless, the learning object metadata in OpenUK had an acceptable quality, as most of the resources included an accessible URI and a well-formed schema with a clear description. Following the proposed approach presented in Section 4.5.1, we realized that none of them pointed to the same address on the Web, but the matching analysis showed that they referred to the same learning object published by different data providers. According to our analysis, 374 resources (out of 4127) in GLOBE were identified as duplicate resources with OpenUK. Turning to the related resources, we selected 300 records of the non-duplicate results and realized that around 246 (82%) of them were semantically related to each other. Also, 48 resources (16%) were not accessible as the URLs were broken or unreachable, and the rest

(2%) were not semantically the same (false positive). Notably, we found two resources about 'functions' but in two different contexts and thus we could not categorize them as related.

To justify the proposed approach, we asked two experts to follow the workflow (Figure 7) and evaluate a set of resources that we had marked as duplicates. To this aim, we randomly selected 20 resources (out of 374) from the results and asked the experts to assess the resource URLs along with the other metadata elements by following the proposed instruction. On receiving the experts' responses, we applied the kappa measure of agreement to analyse the agreement between the observers and gauge the reliability of the responses. The maximum value of kappa is 1, which represents perfect agreement, and kappa will take the value 0 if there is only chance agreement. We later imported the experts' input to SPSS[19] and analysed the measure of agreement of results. The output of the software was valid and equal to 0.828, which demonstrates that the experts strongly agreed on the results. A closer look at the responses given by the raters indicates that most of the resources were marked as duplicates (rater 1 with 17 and rater 2 with 16 resources). The experts could not also judge the rest of resources owing to insufficient information in their metadata.

5. Conclusions

The purpose of this research was to evaluate the results of interlinking between a large learning dataset (GLOBE) to educational datasets in the Web of Data. After analysing the GLOBE metadata and selecting appropriate elements for interlinking, we applied a tool to interlink GLOBE to 20 educational datasets on the Web and evaluated the generated results. In conclusion, we outline the implications of this study as follows:

- Interlinking a learning repository to several educational datasets in the LOD cloud leads to the enrichment of content, as this approach links one e-learning resource to several other resources in different datasets on the Web.
- Evaluating the results of interlinking among the candidate elements of GLOBE demonstrates that the semantic accuracy of matched links for the Title element was higher than the Keyword and Taxon elements, although the distribution of GLOBE resources for this element was minor. Furthermore, the high percentage of GLOBE contribution in a few datasets for the Coverage element indicates that connection of this element to geographical datasets like Geonames or Factbook is appropriate. However, around 93% of GLOBE resources did not provide this element in the metadata.
- Apart from resource enrichment, one of the other benefits of an interlinking process is duplicate identification. Our examination on a set of resources illustrates that several resources are published by different data providers and point to different internet addresses on the Web, although they refer to the same learning resource. We carried out this identification by proposing a data model along with a workflow in which we compared the other metadata elements retrieved from both targets after performing the interlinking process.

Funding

The work presented in this paper has been part-funded by the European Commission under the ICT Policy Support Programme CIP-ICT-PSP.2011.2.4-e-learning with project no. 297229 'Open Discovery Space (ODS)' and INFRA-2011-1.2.2-Data infrastructures for e-Science with project no. 283770 'AGINFRA'.

References

[1] Bizer C, Heath T and Berners-Lee T. Linked data – The story so far. *International Journal of Semantic Web Information System* 2009; 5(3): 1–22.
[2] Piedra N, Chicaiza JA, López J and Tovar E. An architecture based on linked data technologies for the integration and reuse of OER in MOOCs Context. *Open Praxis* 2014; 6(2): 171–187.
[3] Dietze S, Sanchez-Alonso S, Ebner H, Yu HQ, Giordano D, Marenzi I and Nunes B. P. Interlinking educational resources and the web of data: A survey of challenges and approaches. *Program: Electronic Library and Information Systems* 2013; 47(1): 60–91.
[4] Scharffe F and Euzenat J. MeLinDa: An interlinking framework for the web of data. CoRR abs/1107.4502, 2011.
[5] GLOBE (Connecting the World and Unlocking the Deep Web),http://globe-info.org/ (accessed 12 December 2014).
[6] Volz J, Bizer C, Gaedke M and Kobilarov G. Silk – A link discovery framework for the web of Data. LDOW, 2009.
[7] Simperl E, Wölger S, Thaler S, Norton B and Bürger T. Combining human and computation intelligence: The case of data interlinking tools. *International Journal of Metadata Semantics and Ontology* 2012; 7(2): 77–92.
[8] LinkedUp Project, Linking Web data for education, http://linkedup-project.eu/ (accessed: 12 December 2014).
[9] Dietze S, Yu HQ, Giordano D, Kaldoudi E, Dovrolis N and Taibi D. Linked education: Interlinking educational resources and the Web of data. In: *Proceedings of the 27th annual ACM symposium on applied computing*, New York, 2012. pp. 366–371.

[10] Rajabi E, Sicilia M-A and Sanchez-Alonso S. Interlinking educational data: An experiment with GLOBE resources. In: *First international conference on technological ecosystem for enhancing multiculturality (TEEM)*, 2013, pp. 365–374.
[11] Piedra N, Tovar E, Colomo-Palacios R, Lopez-Vargas J and Chicaiza JA. Consuming and producing linked open data: The case of OpenCourseWare. *Program: Electronic Library and Information Systems* 2014; 48(1): 16–40.
[12] Rajabi E, Sicilia MA and Sanchez-Alonso S. An empirical study on the evaluation of interlinking tools on the Web of Data. *Journal of Information Science* 2014; 40: 637–648.
[13] Rajabi E, Sicilia MA and Sanchez-Alonso S. Interlinking educational resources to web of data through IEEE LOM. *Computer Science and Information Systems* 2015, in press.
[14] Ochoa X, Klerkx J, Vandeputte B and Duval E. On the use of learning object metadata: The GLOBE experience. In: *Proceedings of the 6th European conference on technology enhanced learning: Towards ubiquitous learning*, Berlin, 2011, pp. 271–284.
[15] Klyne G and Carroll JJ. Resource Description Framework (RDF): Concepts and abstract syntax, W3C Recommendation, 2004.
[16] IEEE LTSC. IEEE standard for learning object metadata 1484.12.1-2002, final draft version, http://grouper.ieee.org/groups/ltsc/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf (accessed 12 December 2014).
[17] Ngonga A and Sören A. LIMES – A time-efficient approach for large-scale link discovery on the web of data. Presented at the *International Joint Conference on Artificial Intelligence*, 2011.
[18] Bizer C. D2RQ – Treating non-RDF databases as virtual RDF graphs. In: *Proceedings of the 3rd international Semantic Web conference (ISWC)*, 2004.
[19] Cool URIs for the Semantic Web, http://www.w3.org/TR/cooluris/ (accessed 12 December 2014).

Appendix 1. List of selected educational datasets.

Datasets	Size (triple)	SPARQL endpoint
Charles University in Prague	93,233,661	http://linked.opendata.cz/sparql
UNISTAT-KIS	8,026,637	http://data.linkedu.eu/kis/query
Achievement Standards Network (ASN)	7,494,201	http://sparql.jesandco.org:8890/sparql
Data.gov.uk	6,619,847	http://services.data.gov.uk/education/sparql
University of Southampton	5,726,668	http://sparql.data.southampton.ac.uk/
Yovisto - academic video search http://sparql.yovisto.com/	4,932,352	http://sparql.yovisto.com/
University of Muenster (LODUM)	4,179,372	http://data.uni-muenster.de/sparql/
Open University in UK	3,588,626	http://data.open.ac.uk/sparql
University of Huddersfield	3,553,343	http://data.linkedu.eu/hud/query
Semantic ISVU (Kent)	2,421,268	http://kent.zpr.fer.hr:8080/educationalProgram/sparql
University of Bristol	1,885,124	http://resrev.ilrt.bris.ac.uk/data-server-workshop/sparql
Aalto University	1,589,122	http://data.aalto.fi/sparql
Open Courseware Consortium metadata	636,453	http://data.linkedu.eu/ocw/query
OxPoints (University of Oxford)	318,392	https://data.ox.ac.uk/sparql/
TheSoz Thesaurus for the Social Sciences (GESIS)	305,329	http://lod.gesis.org/thesoz/sparql
PROD	62,375	http://data.linkedu.eu/prod/query
Open Data @ Tor Vergata	56,968	http://opendata.ccd.uniroma2.it/LMF/sparql/select
Vytautas Magnus University, Kaunas	39,279	http://kaunas.rkbexplorer.com/sparql/
MoreLab	3,906	http://www.morelab.deusto.es/joseki/articles
Forge project	132	http://data.linkedu.eu/forge/query

Appendix 2. A sample output generated by LIMES.

Globe resource	Relationship	Target resource
< http://www.globe-info.org/101385/ >	owl:sameAs	< http://data.open.ac.uk/oro/34266 >
< http://www.globe-info.org/203273/ >	owl:sameAs	< http://data.linkedu.eu/kis/course/10007773/w14 >
< http://www.globe-info.org/203273/ >	owl:sameAs	< http://data.linkedu.eu/kis/course/10007773/x14 >
< http://www.globe-info.org/277179/ >	owl:sameAs	< http://data.linkedu.eu/kis/course/10007773/w11 >
< http://www.globe-info.org/277179/ >	owl:sameAs	< http://data.linkedu.eu/kis/course/10007773/x11 >
< http://www.globe-info.org/297474/ >	owl:sameAs	< http://data.linkedu.eu/kis/course/10007773/q52 >
< http://www.globe-info.org/ 369509/ >	owl:sameAs	< http://data.linkedu.eu/kis/course/10007773/q52 >
< http://www.globe-info.org/ 381285/ >	owl:sameAs	< http://data.linkedu.eu/kis/course/10007773/q65 >
< http://www.globe-info.org/ 381285/ >	owl:sameAs	< http://data.linkedu.eu/kis/course/10007773/w11 >
< http://www.globe-info.org/381285/ >	owl:sameAs	< http://data.linkedu.eu/kis/course/10007773/x11 >
< http://www.globe-info.org/432521/ >	owl:sameAs	< http://data.linkedu.eu/kis/course/10007773/q54 >
< http://www.globe-info.org/432521/ >	owl:sameAs	< http://data.linkedu.eu/kis/course/10007773/w06 >
< http://www.globe-info.org/432521/ >	owl:sameAs	< http://data.linkedu.eu/kis/course/10007773/x06 >

Appendix 3. Interlinking results between GLOBE and the selected educational datasets.

Dataset	Title	Keyword	Taxon	Coverage
UNISTAT-KIS	188	6,788,988	5,692,741	0
Yovisto - – academic video search	68,506	1,813,416	1,263,662	7,995,334
University of Bristol	17,858	1,872,875	657,686	733
University of Huddersfield	137	828,725	361,791	78
Open University in UK	13,644	720,023	290,837	4
Open Courseware Consortium metadata	24,657	169,737	77,493	24,933
Data.gov.uk	30	100,950	45,604	0
University of Muenster (LODUM)	333	41,522	30,148	316
Open Data @ Tor Vergata	0	61,993	28,444	0
Charles University in Prague	151	72,162	28,157	0
Achievement Standards Network (ASN)	80	131,396	16,481	0
TheSoz Thesaurus for the Social Sciences	138	36,121	13,981	65
Aalto University	14	17,110	3,843	463
Vytautas Magnus University, Kaunas	44	5,201	938	0
OxPoints (University of Oxford)	133	30,512	881	0
PROD	3	7,887	533	0
MoreLab	0	740	42	0
University of Southampton	9	55	0	0
Forge project	0	34	0	0
SUM	125,925	12,699,447	8,513,262	8,021,926
Unique resources in GLOBE	8,260	228,352	134,791	12,941
Common resources in GLOBE	16,354	760,830	413,520	13,591

Conference papers

Title: Exploring the keyword space in large learning resource aggregations: the case of GLOBE
Date of submission: 17-Mar-2013
Date of acceptance: 09-Apr-2013
Conference: Workshop on Learning Object Analytics for Collections, Repositories (LACRO 2013)
Relationship with the global research objectives: Evaluating the GLOBE resource as a big educational repository

Title: Interlinking educational data: an experiment with GLOBE resources
Date of submission: 18-Sep-2013
Date of acceptance: 14-Nov-2013
Conference: First International Conference on Technological Ecosystem for Enhancing Multiculturality (TEEM 2013)
Relationship with the global research objectives: In this paper, interlinking GLOBE resources to DBpedia and Factbook.

Title: Enriching the e-learning contents using interlinking
Date of submission: 31-Agu-2014
Date of acceptance: 15-Sep-2014
Conference: 5th conference on e-Learning 2014
Relationship with the global research objectives: In this paper, we evaluated the enriching of eLearning resources using interlinking approach.

Title: Research Objects Interlinking: The Case of Dryad Repository
Date of submission: 13-Aug-2014
Date of acceptance: 27-Nov-2014
Conference: Metadata and Semantics Research (MTSR 2014)
Relationship with the global research objectives: In this paper, we evaluated the a research data repository in the LOD cloud

Title: Exposición de metadatos de objetos de aprendizaje como datos enlazados: el caso del proyecto Open Discovery Space
Date of submission: 1-Jul-2014
Date of acceptance: 20-Oct-2014
Conference: 9th Latin American Conference on Learning Objects and Technologies
Relationship with the global research objectives: In this paper, we evaluated the eLearning resources using interlinking approach in the Open Discovery Space project.