Open Research Online

The Open University's repository of research publications and other research outputs

Assessing the educational linked data landscape

Conference or Workshop Item

How to cite:

d'Aquin, Mathieu; Adamou, Alessandro and Dietze, Stefan (2013). Assessing the educational linked data landscape. In: ACM Web Science 2013, 02-04 May 2013, Paris, France.

For guidance on citations see [FAQs](#).

© 2013 ACM

Version: Accepted Manuscript

Link(s) to article on publisher's website:

<http://dx.doi.org/doi:10.1145/2464464.2464487>

<http://linkedu.eu/catalogue/publications/websci2013-rn.pdf>

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online's data [policy](#) on reuse of materials please consult the policies page.

oro.open.ac.uk

Assessing the Educational Linked Data Landscape*

Mathieu d'Aquin
Knowledge Media Institute,
The Open University, UK
mathieu.daquin@open.ac.uk

Alessandro Adamou
Knowledge Media Institute,
The Open University, UK
alessandro.adamou@open.ac.uk

Stefan Dietze
L3S Research Center,
Hannover, Germany
dietze@l3s.de

ABSTRACT

In this research note, we present a preliminary study of available web datasets related to education, providing an overview of this area and, more importantly, highlighting how such linked datasets form a globally addressable network of resources for education. As expected, a certain level of heterogeneity was found. We therefore also show how a minor integration effort can improve the global cohesion of such a network of educational web data.

ACM Classification Keywords

K.3.2 Computers and Education: Computer Uses in Education

General Terms

Measurement, Standardization

INTRODUCTION

There is a growing base of open educational content being made available online, both in education-specific resource repositories and in general information sources of relevance to learning and teaching (see for example [7]). Having such content accessible and discoverable on the web has a significant potential impact on the way education is delivered and received, moving from a necessarily localised, university-centric education model to a more global and open one (with the recent developments in MOOCs – Massive Open Online Courses, further illustrating this trend). Information consumed by educational environments can take many different forms, beyond the base material for teaching and learning (including for example repositories of research articles and descriptions of educational facilities), and the open exposure of such data following the principles of open, linked data is expected to make educational information easier to address, aggregate and reuse for various purposes.

Accordingly, linked data principles are increasingly being adopted by educational institutions in different contexts. However, the real potential of linked data in this area does

*This work is funded in part by the LinkedUp project (Grant Agreement: 317620) under the FP7 programme of the European Commission.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

WebSci'13, May 1 – May 5, 2013, Paris, France.
ACM 978-1-4503-1889-1.

not reside within the use of its principles and technologies by individual organisations: Its nature as a *Web of Data* highlights that significant opportunities arise from the use of such principles to create a globally usable network of information and resources to support the current trend in openness, sharing and reuse of education-related resources.

In this paper, we propose to study and assess the current state of linked data for education. We apply a general methodology that collects existing datasets explicitly related to education, extracts key information from these datasets and analyses them. Our goal is to better understand what is already available to application developers in this area, what common practices are being used and, more importantly, how the considered datasets connect with each other through common content and vocabulary reuse.

While the preliminary study presented in this research note already shows an interesting perspective, including the way minor efforts in terms of data integration and mapping can lead to significant improvements in the global cohesion (and so exploitability) of the considered network of datasets, our increasing ability to globally assess such web data in fine-grained ways is expected to contribute significantly in supporting the collaborative development, exploitation and impact of this *Web of Educational Linked Data*.

RELATED WORK

Much research work in the area of technology-enhanced learning has focused on creating standards, such as IEEE LOM and SCORM,¹ to facilitate the exchange and reuse of educational resources in different contexts. While there are many issues in relation to the adoption of such standards, their goal is now becoming more achievable through the use of linked data and semantic web technologies (see [2, 5]). Accordingly, several institutions and initiatives have started adopting these approaches to expose open data of educational relevance. These include universities, schools and research centres (see linkeduniversities.org), government agencies (see education.data.gov.uk), and projects around specific domains (see for example www.meducator.net). Usage of such data sources is slowly starting to emerge in collective resource gathering and intelligence [6] or learning analytics [4]. Considering this new, increasingly data-rich environment, understanding what is available, what the common practices are, and how sources of information connect on the Web of Data becomes critical.

¹see <http://ltsc.ieee.org/wg12/20020612-Final-LOM-Draft.html> and <http://scorm.com/>

Surveying the state and growth of available web resources has been a popular topic over the last few years in the semantic web research community. Initially, several studies have concentrated on analysing available ontologies especially relying on the crawls of semantic web search engines [3]. More recently, simple statistical analysis of datasets on the web and of their connections based on `owl:sameAs` links have been considered (see statistics on the linked data cloud²). Here, we look specifically at educational datasets, with the goal to assess how well they connect, not necessarily through explicit links between them (which are rare), but through the way they reuse common vocabularies, contain resources of similar types and therefore could be jointly used in a global educational context.

METHODOLOGY AND DATASETS

The first step to analyse education-related web data is to collect datasets that explicitly reference education and are exposed as linked data. We first query `datahub.io` to find existing (SPARQL-enabled) data endpoints that make explicit reference to education-related entities. We completed this initial step by having members of the LinkedUp support action³ add known relevant datasets. We then extracted specific sub-datasets for each endpoint, as explicit RDF graphs, as well as information about their content (the RDF types and properties employed and the corresponding vocabularies). This was achieved through a series of scripts relying on dedicated SPARQL queries sent directly to the datasets' endpoints.

From these key data about datasets, we then created a (meta)dataset, using the VoID vocabulary⁴ and made this available as our (catalogue) SPARQL endpoint⁵. Having thus obtained our basic corpus for analysis, we created further transformation scripts, translating the results of dedicated SPARQL queries to our endpoint to a format usable by the selected analysis tool. Since, in this preliminary study, we are mostly interested in the connectivity established by reuse of common vocabularies and types, we employed the network analysis and visualisation tool Gephi⁶.

The list of datasets analysed, their statistics, extracted information, and high-resolution images of the network analysis visualisations shown later in this paper are available from <http://data.linkededucation.org/linkedup/catalog/analysis/1.0/>.

The basic method described above led to 146 datasets (RDF graphs) obtained from 22 different data endpoints. Some datasets originate from universities (e.g. `data.open.ac.uk`, `data.southampton.ac.uk`, `data.aalto.fi`); others from publication repositories (e.g. `data.nature.com`,

`dblp.13s.de`), government agencies and standardisation bodies (e.g. `education.data.gov.uk`), or as the output of specific projects (e.g. `meducator.open.ac.uk`, `data.organic-edunet.eu`). All together, the datasets use 588 different types, most of them only present in one dataset each, but others being shared by up to 50 of them. Similarly, most of the 719 properties found only appear in one of the datasets, but the most common one (<http://purl.org/dc/terms/created>) is employed by 92 different datasets.

DATASET NETWORK ANALYSIS

Based on the collected corpus of datasets of educational relevance, we propose, as an initial study, to analyse how the datasets reuse vocabularies and include resources of similar types. Ultimately, this should allow us to understand and find ways to improve the overall cohesion of the network formed by these datasets, and obtain a comprehensive view on the practices and different types of data sources available.

Vocabularies Used

Figure 1. Co-occurrence network of vocabularies. Node colours indicate the connectivity (red=most connected, blue=least connected); label sizes indicate the number of datasets using each vocabulary.

There are two main mechanisms through which linked datasets can connect: directly, through explicit links between the resources they describe (mostly `owl:sameAs` relations); or indirectly by reusing vocabularies, thereby making their resources jointly usable and interoperable. As already mentioned, direct links are very rare in our datasets, so we examined how the vocabularies used determine links through which the datasets can form a network.

There are 144 different vocabularies used in our education-related datasets, many of which are specific to the datasets. Others, e.g. <http://purl.org/dc/terms/>, are widely reused, by up to 145 of the datasets. Unsurprisingly, the most popular vocabularies are not specific to education, but are used to represent general concepts and relations, such as resource metadata (Dublin Core), people (FOAF), topics (SKOS), organisations (OpenOrg), time (W3C Time Ontology), etc. Some more specific vocabularies of relevance to

²<http://wifo5-03.informatik.uni-mannheim.de/lodcloud/state/>

³<http://linkedup-project.eu>

⁴<http://rdfs.org/ns/void>

⁵<http://data.linkededucation.org/linkedup/catalog/sparql/>

⁶<https://gephi.org> - Note that, although a Gephi plugin for importing the results of SPARQL queries exists (<http://wiki.gephi.org/index.php/SemanticWebImport>), it was unavailable at the time we carried out this analysis.

education are however present, and represented in a significant number of datasets, such as the Bibliographic Ontology (BIBO), the Academic Institution Internal Structure Ontology (AIISO), or the Model of Learning Opportunities (MLO). As shown in Figure 1⁷, not only are these vocabularies popular, but they also represent hubs connecting even the most isolated datasets (i.e. those otherwise using mostly specific vocabularies). Dublin Core stands out in this network as the main connector reaching out to the most “peripheral” vocabularies used in very specific datasets (e.g. the ones coming from `data.cnr.it` – left cluster in the figure).

Types of Resources Included

While vocabularies give a general idea of the way datasets reuse common schemas, we consider here a more specific view of the way they interconnect content-wise, i.e. through the RDF types they use to categorise their resources.

As shown in Figure 2a⁸, 50 datasets describe themselves with elements including those indicating their provenance (`void:ProvenanceEvent`). As with vocabularies, also here several popular and generic types (especially, `skos:Concept`, `foaf:Organization`, `foaf:Document`) act as hubs providing a common element to many disparate, sometimes dataset-specific types. It is interesting to note that the popularity of a type does not necessarily straightforwardly correlate with its connections with others. For example, while very popular, `void:ProvenanceEvent` (50 datasets) connects much less to other types than `foaf:Organization` (9 datasets).

Figure 2b⁹ provides an insight as to how the use of common classes can help connect the considered datasets into a globally usable network. In this figure, the colours represent the classes of modularity (as calculated by Gephi using the algorithm described in [1]). Our expectation would be that this clustering method aggregates datasets based on the general domain of their content, but as can be observed here, the 3 groups of highly connected clusters contain elements that are heterogeneous in domains, but homogeneous in origin (e.g., most of the blue cluster is made of datasets from `data.southampton.ac.uk`). This separation is clearly not desirable, as one goal of applying linked data to education is to make it possible to use resources from heterogeneous sources. Clearly, this is a side-effect of the various choices of representation schemes and vocabularies made for each dataset for describing similar entities. In the next section, we show how small efforts in aligning these choices can help mitigate this artefact.

Improving Cohesion

To improve the overall connectivity of the dataset network from Figure 2b, we manually aligned a large part of the vocabularies used in these datasets with a limited subset of these

⁷high-resolution image: <http://data.linkededucation.org/linkededup/catalog/analysis/1.0/voc-network.png>

⁸high-resolution image: <http://data.linkededucation.org/linkededup/catalog/analysis/1.0/class-network.png>

⁹high-resolution image: <http://data.linkededucation.org/linkededup/catalog/analysis/1.0/dataset-class-network.png>

vocabularies. In particular, we used the DBpedia ontology¹⁰ for general concepts, BIBO¹¹ for resources and documents, XCRI¹² for courses and degrees, AIISO¹³ for institutions and, to a lesser extent, a few other vocabularies for less richly represented aspects. All together, we mapped 201 of the most frequent types in the datasets to 79 classes of these selected vocabularies.

Figure 2c¹⁴ shows the network formed by these mapped types based on their co-occurrence in datasets. This network illustrates in a clearer way how some popular types (Organisation, Place, Person, Concept, Document) act as connection points to which more specific and possibly ad-hoc ones can relate. Organisation and Person are interesting examples since, once they aggregate multiple analogous types, they appear much more prominently here than in Figure 2a.

Even more interesting, however, is how the relatively trivial data integration effort of mapping a couple of hundred types significantly improved the overall cohesion of the network and, by extension, our ability to jointly use and query our 146 datasets (e.g., to obtain a list of available resources, or learning opportunities). Indeed, as shown in Figure 2d¹⁵, the computed clusters of datasets are much less separated¹⁶ and much more driven by the domains they cover than by the sources of information and the design choices made in these sources (e.g. the red cluster focuses more on courses and educational programmes, while the green one is more about educational resources and publications).

CONCLUSION

We presented a preliminary study on 146 datasets of explicit relevance to the educational domain aimed at investigating how such datasets form a network of globally addressable resources for education. As expected, a certain degree of heterogeneity emerged, which could be partially alleviated through a relatively simple data integration effort.

A study like the one presented here is necessarily a constantly evolving effort, which is, however, important to share from the early stages. In an area like education, where dramatic (and often technology-led) changes are emerging, the availability of new datasets and the definition of their relevance are moving very quickly, with initiatives such as `LinkedUniversities.org`, `LinkedEducation.org` and `LinkedUp-project.eu` aiming at uncovering and realising the impact of such a network of educational Web data.

With an improving ability to access education-related data with standard web technologies come new possibilities for

¹⁰<http://wiki.dbpedia.org/Ontology>

¹¹<http://bibliontology.com/>

¹²<http://www.xcri.org/>

¹³<http://vocab.org/aiiso/schema>

¹⁴high-resolution image: <http://data.linkededucation.org/linkededup/catalog/analysis/1.0/class-mp-network.png>

¹⁵high-resolution image: <http://data.linkededucation.org/linkededup/catalog/analysis/1.0/dataset-class-mp-network.png>

¹⁶Gephi’s modularity measure based on [1] gives 0.323 for the new network, compared to 0.408 for the one in Figure 2b