Undefined 1 (2014) 1–5
IOS Press

The LinkedUp Data Catalogue: A Meta-Dataset of Linked Datasets in the Education Domain

Editor(s): Name Surname, University, Country
Solicited review(s): Name Surname, University, Country
Open review(s): Name Surname, University, Country

Mathieu d'Aquin [a], Alessandro Adamou [a], Stefan Dietze [b] and Besnik Fetahu [b]
[a] *The Open University, Walton Hall, Milton Keynes, Buckinghamshire MK7 6AA, United Kingdom*
E-mail: {mathieu.daquin,alessandro.adamou}@open.ac.uk
[b] *L3S, Research Center, LUH Hannover, Germany*
E-mail: {dietze,fetahu}@l3s.de

Abstract. The LinkedUp Catalogue of Web datasets for education is a meta-dataset dedicated to supporting people and applications in discovering, exploring and using Web data for the purpose of innovative, educational services. It is also an evolving dataset, with most of its content being contributed by automatically extracting relevant information from external descriptions and the included datasets themselves. In this paper, we describe the purpose and content of this dataset, as well as the way it is being created, published and maintained.

Keywords: Linked Open Data, Education, University, Education, Learning, Data Catalogue

1. Overview

LinkedUp[1] is a European Support Action dedicated to pushing forward the adoption of Web Data in education, and especially their use in innovative educational services. A core activity of the project is the organisation of a series of competitions (the LinkedUp Challenge[2]) for developers of such services. As a supporting activity for these competitions, we developed a catalogue of datasets of relevance to education, which participants to the competitions can use to identify and access data of use for their own applications. The LinkedUp Data Catalogue (also known as the 'Linked Education Cloud') is built to interact with existing data catalogues (namely CKAN[3], as deployed on `Datahub.io`), so to extract, represent and publish information about the relevant datasets published as linked data, with the catalogue itself relying on Linked Data principles. We call the resulting dataset a "meta-dataset" as it contains metadata and access information about linked datasets.

The LinkedUp Catalogue was designed as a resource for the LinkedUp Project and Challenge. However, it also represents a unique resource in the educational sector where the Linked Data movement is only starting to gain momentum [1]. It has therefore grown into a community-wide resource in its own right, which will be sustained beyond the activities of the LinkedUp Project and is being used for data discovery by developers in the education sector other than participants to the competitions. In addition to developers' use, the LinkedUp Catalogue represents a constantly evolving observatory of the practices in Linked Data publishing in the educational sector. Preliminary

[1] `http://linkedup-project.eu`
[2] `http://linkedup-challenge.org`
[3] `http://ckan.net`

studies have shown that the links that are being created between the vocabularies of the included dataset can lead to a better overall cohesion (and therefore a better reusability) of the included datasets [2], and more work is to be carried out (in particular as part of `LinkedUniversities.org` and the W3C Open and Linked Education Community Group[4]) to use the LinkedUp Data Catalogue to extract and analyse common data modelling practices in the education sector, and use this to further improve the adoption of Web Data standards for educational purposes.

Table 1 provides basic information about the LinkedUp Data Catalogue. The catalogue currently contains descriptions of 50 datasets. Each of these datasets has been included following the criteria that: 1- It had to be of explicit relevance to learning; and 2- it had to be accessible at least through a SPARQL endpoint. The second criterion is included here for several reasons. As a project, LinkedUp had for objective to encourage the use of Web Data standards. Also, SPARQL has the advantages that it supports querying for information about the dataset itself, making the automatisation of the creation of the data catalogue feasible. Finally, as all datasets are available through one, web-based, homogeneous access method, which is also the same as the access method for the catalogue itself, it facilitates the building of applications that draw from several of these datasets, exploiting the catalogue as a hub to these datasets and the links provided between their vocabularies as a way to draw consistent information from multiple datasets. However, it is important to note that to support the growth of the catalogue despite this criterion, the LinkedUp Project has carried out a number of transformations of existing, non-RDF-based datasets into RDF and linked data, so that they can be provided through a SPARQL endpoint and included in the catalogue.

The datasets currently catalogued originate from a variety of sources, and cover various aspects of education. Several of them are directly created by educational institutions, especially universities (e.g. The Open University, The University of Munster, Aristotle University), and cover aspects such as the courses they provide, the educational material they publish, their research outputs, or their teaching facilities. Other datasets cover the entire educational sector of a particular country, or even statistics about education around the world, and generally originate from open data initiatives, including governmental ones (e.g. education.gov.uk, Italian National Research Council, OECD). A large part of the data also originates from repositories of resources of relevance to education, including specifically open educational resources (e.g. OpenCourseware[5], mEducator[6]) or other types of resources, including publication repositories (e.g. DBLP[7], Nature[8]). Finally, several datasets originate from projects looking at specific aspects of education (e.g. the Terence Reading Comprehension dataset[9], or the SEEK-AT-WD ICT tools for education dataset[10]).

2. Usage

In this section, we describe the usage of the LinkedUp Catalogue as a Linked Data-based, meta-dataset of linked datasets in education. We first describe the different types of uses for which this catalogue was built, naturally focusing on data discovery. We then describe existing applications of the data catalogue, including a summary of submissions to the LinkedUp Challenge that have been using it to identify relevant datasets for their applications, and also showing how it has been used to provide an overview of the growing field of educational linked data.

2.1. Purpose of the dataset: Data discovery

As described in Section 1, the primary aim of building the LinkedUp Catalogue of datasets for education was to support participants in the LinkedUp competitions in finding and accessing relevant data for their applications. This however represents a natural use case in the broader domain of education. Indeed, data discovery and ease of access are critical in many areas. Education is a particularly interesting one, as it is naturally a heavy producer of data. However, while significantly growing [1,3,4] the use of Web Data standards and generally the adoption of open data practices is still limited in this area, as it remains distributed and

[4] `http://www.w3.org/community/opened/`

[5] `http://datahub.io/dataset/open-courseware-consortium-metadata-in-rdf`

[6] `http://datahub.io/dataset/meducator`

[7] `http://datahub.io/dataset/l3s-dblp`

[8] `http://datahub.io/dataset/npg`

[9] `http://datahub.io/dataset/terence-reading-comprehension-dataset`

[10] `http://datahub.io/dataset/seek-at-wd-ict-tools-for-education-web-share`

Table 1
Details of the LinkedUp Catalogue dataset.

Name	LinkedUp Catalogue of Dataset for Eduction
Homepage	`http://data.linkededucation.org/linkedup/catalog/`
License	CC-BY `https://creativecommons.org/licenses/by/3.0/`
SPARQL endpoint	`http://data.linkededucation.org/linkedup/catalog/sparql/`
Datahub.io URL	`http://datahub.io/dataset/linkedup-catalogue-of-educational-datasets`
Number of triples	209,806
Number of graphs	2 (one for metadata and one for mappings)
Number of datasets described	50
Number of vocabulary mappings	164

siloed, with different countries and organisations having different practices.

The LinkedUp Catalogue is therefore built to describe as much as possible both the information about a dataset and its content, to enable both human and software agents in identifying relevant data, and the channels to access them. The details of how this is achieved is described in Section 3. Here, we focus on describing the main application of the dataset: The interface of the LinkedUp Catalogue of Dataset for education.

This interface is available from the homepage of the LinkedUp Catalogue[11]. Figure 1 shows a screenshot of the page displaying and linking to the set of linked datasets included in the catalogue. This page shows basic information about each dataset, including their title, description, and (if possible) a small graph showing the main classes (types of objects) present in the dataset. The page also shows a real time indication of whether or not the SPARQL endpoint assocciated with the dataset is currently accessible, through showing their title in bold when they are accessible, and in italic when they are not. As shown in [5], SPARQL endpoints on the web are varyingly reliable and often experience downtime. This is therefore a crucial information to give the users of the datasets catalogued by LinkedUp.

Building this page relies on a very simple query: As described in Section 4, the representation of the description of datasets is based mostly on the VoID[12] and Dublin Core[13] vocabularies. The query therefore requests the title and description of every object of type void:Dataset which are attached to a SPARQL endpoint. Note that obtaining the SPARQL endpoint is crucial, as it is needed to send it probe queries checking its availability and to display the graph. It is also necessary as named graphs in each dataset are represented as sub-datasets of the overall one, and are also of type void:Dataset.

Query 1 *Listing basic information from datasets included in the catalogue.*

```
select distinct ?l ?s ?des where {
   ?d <http://www.w3.org/2000/01/rdf-schema#label> ?l.
   ?d <http://rdfs.org/ns/void#sparqlEndpoint> ?s.
   ?d <http://purl.org/dc/terms/description> ?des
}
```

Checking for the availability of each SPARQL endpoint is done through checking that we obtain from them a valid response to a simple probe query (`ASK {?s ?p ?o}`). The summary graph are obtained through an external service specifically built for the purpose of the LinkedUp Catalogue, but that can be applied on other valid SPARQL 1.1 endpoints than the ones of the datasets included. For completeness, we include below the query it uses to the endpoints to identify (as an approximation) the most common classes and their connections. This query is obviously cached to avoid overloading the external SPARQL endpoints, as it is relatively complex. Also, the summary graph is not shown for every dataset, as this query might not be supported (or might result in the endpoint timing out).

Query 2 *Query to obtain a random set of connected classes in a dataset. The most common ones are then counted.*

```
select distinct ?c1 ?c2 where {
   ?x1 a ?c1.
   ?x2 a ?c2.
   ?x1 ?p ?x2
} order by rand() limit 1000
```

From the main browsing page, each summary description of a dataset links to a more complete description, i.e. a dedicated page for the dataset, providing more information. An example of such a page is shown in Figure 2. This page provides basic in-

[11] `http://data.linkededucation.org/linkedup/catalog/browse/`
[12] `http://www.w3.org/TR/void/`
[13] `http://dublincore.org/`

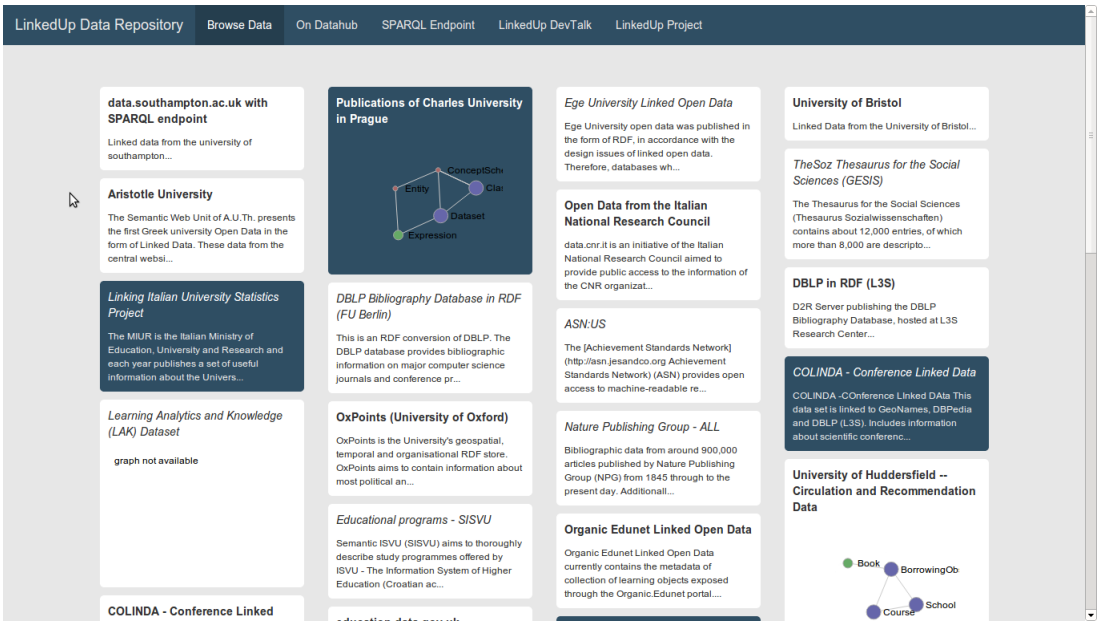

Fig. 1. Screenshot of the browsing interface of the LinkedUp Catalogue.

formation about the dataset, including title, description, the license associated with it, whether it is currently responding, and the number of graphs, classes and properties it include. It also includes an extended version of the graph shown in the browsing page (to include more classes and connections), as well as (if available) a tag cloud of the most common topics covered by the dataset, and an indication of the URI patterns used to identify objects of the different classes of the dataset, in each of its graphs. The additional information about the dataset (license, etc.) are obtained through a query very similar to the one used in the browsing page. The number of graphs, classes and properties can be obtained through querying the sub-datasets, class-partitions and property partitions of the dataset, based on their representation with the VoID vocabulary. The topic indicators for the tag cloud are obtained through the approach described in [6]. The URI patterns are obtained through an external service implementing the technique described in [7].

Finally, one of the key aspects of the LinkedUp Catalogue is that it also includes links (mappings, alignments) between elements of the vocabularies of the datasets included. These links and mappings are built according to the process described in Section 3. The impact of this in terms of usage is however relatively important since, together with the fact that the catalogue and these links are themselves available through a SPARQL 1.1 endpoint, it enables queries that can federate information from multiple other datasets, even though they might use different vocabularies. The query below shows a concrete example where, in only one SPARQL query, the LinkedUp Catalogue dataset (and SPARQL endpoint) is used as a hub to identify data about a particular concept (schools) in the included dataset, and automatically query them with their own vocabulary to retrieve instances of this concept. This usage of the LinkedUp Catalogue dataset is however still very limited, because of the lack of maturity in the implementation of the mechanisms associated with federated querying in SPARQL, and the lack of robustness of the datasets to be queried, which makes it likely that at least one will not give a valid response and therefore make the entire query fail. We however hope that with the increased development of SPARQL-related technologies, these issues will eventually disappear and make query federation a realistic use case for the LinkedUp Catalogue.

Query 3 *Example of federated query using the LinkedUp Catalogue dataset.*

```
prefix void: <http://rdfs.org/ns/void#>
prefix aiiso: <http://purl.org/vocab/aiiso/schema#>
prefix owl: <http://www.w3.org/2002/07/owl#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>

select distinct ?endpoint ?school ?cl where {
   ?ds void:sparqlEndpoint ?endpoint.
   {{?ds void:classPartition [ void:class ?cl]}
      UNION
   {?ds void:subset [ void:classPartition [
                          void:class ?cl] ]}}
```

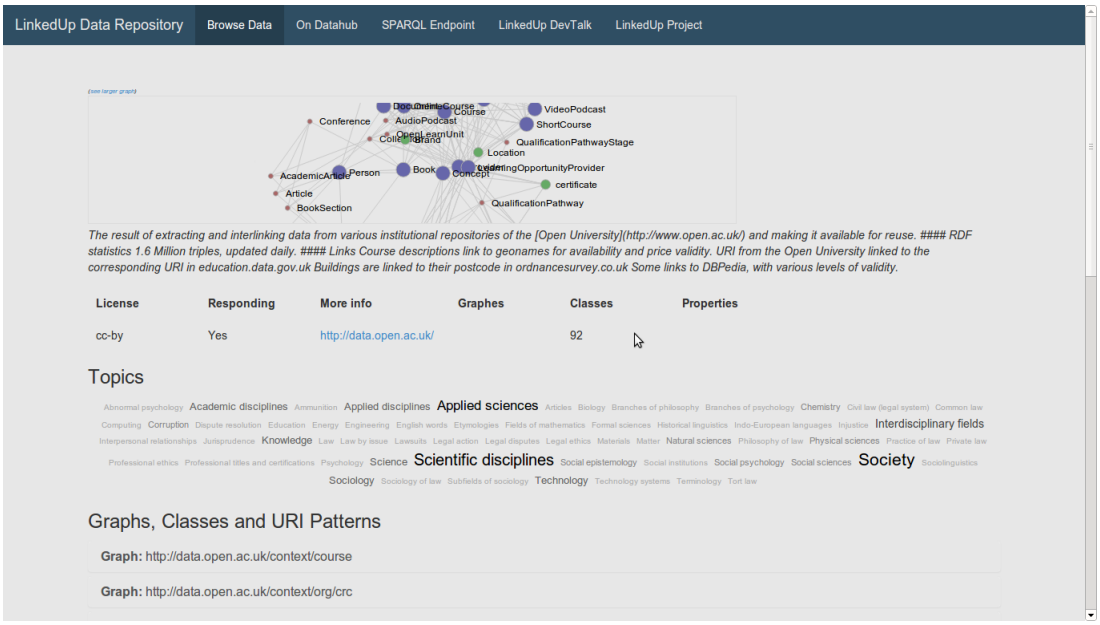

Fig. 2. Screenshot of the interface descrining a specific dataset in the LinkedUp Catalogue (here, `data.open.ac.uk`)

.

```
   {{?cl owl:equivalentClass aiiso:School}
       UNION
   {?cl rdfs:subClassOf aiiso:School}
      UNION
   {FILTER ( str(?cl) = str(aiiso:School) ) }}
   service silent ?endpoint {
      ?school a ?cl
   }
}
```

2.2. *Applications*

Most of the applications that rely, to an extent, on the LinkedUp Catalogue naturally come from the submissions to the competitions organised as part of the LinkedUp Challenge. The LinkedUp Challenge was organised in three different competitions, Veni, Vidi and Vici, which represented three phases of the challenge. The Veni competition was organized in 2013 and attracted 23 submissions. The Vidi competition which took place in the first half of 2014 attracted 14 submissions. The Vici competition which final stage is finishing at the time of writing attracted 13 submissions.

There was no requirement to use the datasets of the Catalogue in submissions, as it was only meant as support. The catalogue evolved with the competition, and was only available in a preliminary form for the first one. Nevertheless, several submissions in each of the competitions used it to different extents, mostly through identifying, at the time of development, datasets of interest and using them in their applications. For example the HyperTED application[14] submitted to the last competition used, amongst others, the TED dataset included in the catalogue, and that was created by members of the LinkedUp Project. Many other similar examples exist of either creating applications using datasets that are available through the LinkedUp Catalogue, or to create and make available new data to integrate into the catalogue. A good example of the former is the We-Share application[15] that was submitted to the first competition, and which subsequently made the data it produces available through the catalogue for others to reuse. Several other similar examples exist, from the competitions or through other channels, where the LinkedUp Catalogue played a role in the creation, publication and linkage of new linked data (e.g. the Prod dataset linking to the KIS dataset).

Another interesting type of applications of the LinkedUp catalogue, as mentioned in Section 1, is that it also enables members of the community to obtain an overview of the state of Web Data in education, and to extract from this practices that should be encourage to push forward the adoption of Web standards for data

[14] `http://linkedtv.eurecom.fr/mediafragmentplayer`

[15] `http://seek.cloud.gsic.tel.uva.es/weshare/`

dissemination in this sector. For example, a catalogue explorer[16] was created based on the descriptions of the datasets in the LinkedUp Catalogue, their mappings, and the extracted topics [8]. In [2], we proposed an initial analysis of the landscape of Web data in education, based on a preliminary version of the LinkedUp Catalogue dataset. This kind of analysis represents valuable resources for the data community, as it makes it possible to identify the common data modeling practices in this sector (e.g. identifying the most used vocabularies). It also gives a view of the cohesion of the data in this domain, including the way in which datasets can be related to each other based on their reuse of the same modeling practices. A key conclusion of [2] was indeed to show how including links between the vocabularies used in different datasets can improve this cohesion, putting closer together datasets of similar content, rather than the ones of similar origins. It is worth mentioning here that this approach – using the LinkedUp Catalogue dataset as a way to identify common data modeling practices and the way to make them more cohesive – is now proposed to be the basis of the W3C Open and Linked Education community group.

3. Creation and Update

The process to create the LinkedUp Catalogue is designed to be as automatic as possible, in order to facilitate updates especially in cases where new datasets are being included (which happens on a regular basis). We describe it in more details in this section.

3.1. Basic metadata extraction

Datasets in this process are first identified through the use of `Datahub.io`, a CKAN-based catalogue of datasets available on the Web. Initially, this relied on the use of a dedicated group (linked-education[17]), but `Datahub.io` changed the way permissions in groups are handled in a way that makes them less usable for our needs. We therefore now also maintain a list of dataset IDs from `Datahub.io` on the LinkedUp server for all relevant datasets not currently included in the `linked-education` group. From these, the CKAN API is used to extract basic metadata about each dataset (title, description, license) as

[16] `http://data-observatory.org/led-explorer`
[17] `http://datahub.io/group/linked-education`

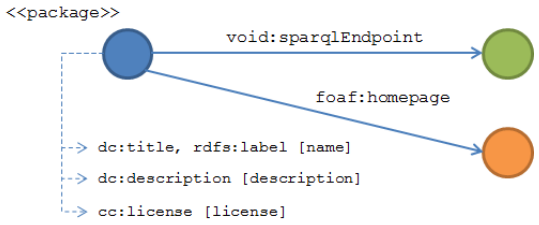

Fig. 3. Representation of the basic metadata about a dataset in the LinkedUp Catalogue.

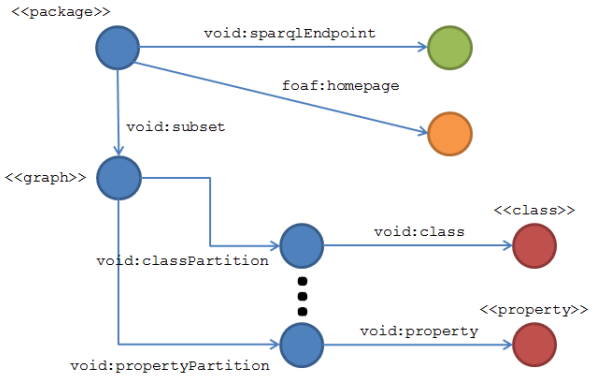

Fig. 4. Representation of the content subsets of a dataset in the LinkedUp Catalogue.

well as information about their access point (SPARQL endpoint). These initial pieces of information are represented using VoID and Dublin Core, as Shown in Figure 3. Having received the information about the SPARQL endpoint for each dataset, it is them queried to obtain the list of graphs, classes and properties included. represented as subsets, class-partitions and property-partitions in VoID, as shown in Figure 4.

3.2. Mapping vocabularies

With the growth of the catalogue by inclusion of new data endpoints and re-engineering of non-Linked Data sources, arising interoperability issues had to be addressed. Some of them dealt with the hardships and increasing complexity in formulating and issuing general-purpose queries that are valid across most, if not all, the external sources registered with the catalogue. As an addition to the initial data cataloguing workflow, we have included a vocabulary mapping phase in the workflow. This phase is implemented using basic strategies of ontology alignment [9] allowing us to formalise a degree of semantic similarity between two terms coming from distinct vocabularies. To that

Table 2
Vocabularies selected for alignment.

AIISO	http://purl.org/vocab/aiiso/schema
Bibo	http://purl.org/ontology/bibo
DERI location vocabulary	http://vocab.deri.ie/rooms
Event Ontology	http://purl.org/NET/c4dm/event.owl
FOAF	http://xmlns.com/foaf/0.1/
GoodRelations	http://purl.org/goodrelations/v1
vCard	http://www.w3.org/2006/vcard/ns
W3C Time ontology	http://www.w3.org/2006/time
WG84 Geo positioning	http://www.w3.org/2003/01/geo/wgs84_pos

end, we selected a restricted set of vocabularies to be used as the target set for aligning classes asserted in other ontologies (described in Table 2).

The process to create the mappings from the vocabularies of each new dataset in the catalogue to the selected ones is a semi-automatic process which is based on very simple principles: First, the use of the selected vocabularies is detected, and no further step is required for the vocabulary elements that are already covered. This provides a rational for the selection of these vocabularies, as the ones most commonly used for different aspects of the domain. Second, any existing link/mapping can be reused. Indeed, if not one of the selected ones, the dataset might be using a vocabulary that has already been used and mapped in another dataset. These mappings can simply be reused. Third, very simple similarity techniques are used to suggest new mappings to the catalogue manager, who has the ability to validate or invalidate them. Again here, the created mapping can be reused, but also the information about it not being valid can be kept and reused to avoid suggesting it again for future datasets. Finally, mappings can be manually created by the catalogue manager, through a simple interface showing the vocabulary elements used by the dataset and allowing her to enter the URI of corresponding elements in the selected vocabularies in a field that auto-completes the manual input with elements from these vocabularies.

3.3. Update and sustainability

Other than the processes described above, as previously mentioned, three external services are also used to create the graph summaries of each dataset, their topic clouds, and to extract the URI patterns they rely on. Besides the vocabulary mapping process described above, which is semi-automatic, and the initial input of metadata in Datahub.io, the whole workflow is automatic and runs on a regular basis, when a new dataset is included or simply to keep the catalogue dataset updated. Only the latest version of the dataset is currently being made available. It is worth mentioning here that work is ongoing on making the data management infrastructure underlying the catalogue, and some of the datasets included in it that were produced by the project, easier to maintain, with a dedicated server and additional facilities so that the data cataloguing workflow only requires minimal validation from the catalogue team for any new dataset.

4. Conclusion and Discussion

The LinkedUp Catalogue of Web datasets for education is very different, in purpose and form, to many other dataset in the Linked Data Web. It is a meta-dataset dedicated to supporting people and applications in discovering, exploring and using web data for the purpose of innovative, educational services. It is also an evolving dataset, with most of its content being contributed from automatically extracting relevant information from external descriptions and the included datasets themselves. As such, it as a limited purpose – i.e. it only supports a limited set of use cases, including data discovery and access federation. These use cases are however critical in areas such as education, where very disparate and scarce data are available from many different sources with not overall coordination. It also represents an interesting demonstration of the value of using Linked Data-related technologies, especially SPARQL, Web Vocabularies and Links/Alignments, as a way to integrate meta-information with information, and therefore enable such a Catalogue to become a hub for data access across many, potentially heterogeneous data sources. It is actually surprising to us that we could not find other examples of community-wide, Linked Data-based data catalogues.

The evolution of the dataset will naturally be aimed at addressing its current limitations. The current catalogue being restricted in scope (dataset explicitly related to education), it can obviously only address services within this particular scope. Extending it would however decrease the value of the dataset, making it less appropriate for data discovery. One approach to tackle this is reproduce the process used for building the LinkedUp Catalogue to other, connected sectors (e.g. research, health, etc.). In terms of content, other aspects of the datasets are planned to be included, such as for example the topics and the URI patterns extracted from external services, which, if represented appropriately, could be exploited by applications to better understand the appropriateness of the data to their goal. Indications of the quality of the data is of course also crucial, but not currently covered by the dataset. Finally, as described in Section 2, one of the powerful use cases of a meta-dataset such as the LinkedUp Catalogue, including information about access points to datasets and mappings between their vocabularies, is the potential it creates for query federation. This use case now needs to be further developed as the technologies for SPARQL query federation evolve to become more robust.

References

[1] Mathieu d'Aquin and Stefan Dietze. Open education: A growing, high impact area for linked open data. *ERCIM News*, (96), 2014.

[2] Mathieu d'Aquin, Alessandro Adamou, and Stefan Dietze. Assessing the educational linked data landscape. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 43–46. ACM, 2013.

[3] Mathieu d'Aquin. Linked data for open and distance learning. Commonwealth of Learning Report, 2012.

[4] S. Dietze, H. Drachsler, and D. Giordano. A survey on linked data and the social web as facilitators for tel recommender systems. In *Recommender Systems for Technology Enhanced Learning: Research Trends and Applications*, 2013.

[5] Pierre-Yves Vandenbussche, Carlos Buil Aranda, Aidan Hogan, and Jürgen Umbrich. Monitoring sparql endpoint status. In *Demo the 12th International Semantic Web Conference*, 2013.

[6] Besnik Fetahu, Stefan Dietze, Bernardo Pereira Nunes, Marco Antonio Casanova, Davide Taibi, and Wolfgang Nejdl. A scalable approach for efficiently generating structured dataset topic profiles. In *Extended Semantic Web Conference, ESWC*, pages 519–534. Springer, 2014.

[7] Mathieu d'Aquin, Alessandro Adamou, Enrico Daga, and Nicolas Jay. Extracting uri patterns from sparql endpoints. KMi Tech Report ID: kmi-14-04, 2014.

[8] Davide Taibi, Stefan Dietze, Besnik Fetahu, and Giovani Fulantelli. Exploring type-specific topic profiles of datasets: a demo for educational linked data. In *Poster and system demonstration proceedings of the International Semantic Web Conference, ISWC*, 2014.

[9] Jérôme Euzenat and Pavel Shvaiko. *Ontology matching*, volume 18. Springer, 2007.