Undefined 1 (2014) 1–5
IOS Press

The Open University Linked Open Data - data.open.ac.uk

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

Enrico Daga [a], Mathieu d'Aquin [a], Alessandro Adamou [a] and Stuart Brown [a]
[a] *The Open University, Walton Hall, Milton Keynes, Buckinghamshire MK7 6AA, United Kingdom*
E-mail: {enrico.daga,mathieu.daquin,alessandro.adamou,stuart.brown}@open.ac.uk

Abstract.
The article reports on the evolution of *data.open.ac.uk*, the Linked Open Data platform of the Open University, from a research experiment to a data hub for the open content of the university. Entirely based on Semantic Web technologies (RDF and the Linked Data principles), *data.open.ac.uk* is used to curate, publish and access data about academic degree qualifications, courses, research papers and open educational resources of the university. It exposes a SPARQL endpoint and several other services to support developers, including queries stored server-side and entity lookup using known identifers such as course codes and YouTube video IDs. The platform is now a key information service at the Open University, with several core systems and websites exploiting linked data through *data.open.ac.uk*. Example applications include connecting entities such as courses to media objects published in different places (YouTube, Audioboo, OpenLearn, etc.) and providing recommendations of resources based on application-specific queries. Through these applications, *data.open.ac.uk* is now fulfilling a key role in the overall data infrastructure of the university, and in establishing connections with other educational institutions and information providers.

Keywords: Linked Open Data, Dataset, University data, Education

1. Introduction

data.open.ac.uk[1] is the home of linked open data from The Open University (OU). The data available come from various institutional repositories of the university, and are collected, interlinked and openly published for reuse. Currently, the datasets relate the publications, qualifications, courses and educational resources produced at the Open University, as well as the people involved in making them. The platform also hosts the linked data output of specific research projects. All these data are available through open standards (RDF and SPARQL) and are mostly available under an open license (see Section 2.3).

1.1. Linked university data: requirements and challenges

The Open University has a vast presence on the WWW. Having distance learning in its core, the diffusion of open-access learning material is part of its student recruitement strategy as well as "brand" communication. The University publishes a number of websites that contain open-access content, most notably OpenLearn[2], and has a number of channels on social media such as YouTube[3] or Audioboo[4].

[1] http://data.open.ac.uk

[2] http://www.open.ac.uk/openlearn/

[3] https://www.youtube.com/user/TheOpenUniversity

[4] http://audioboo.fm/search?utf8=%E2%9C%93&q=the+open+university

0000-0000/14/$00.00 © 2014 – IOS Press and the authors. All rights reserved

In over 40 years, the OU developed a vast amount of educational resources, a fair deal of which are now online as freely accessible material. The creation of learning and teaching material is part of the regular job of OU faculties, and the production and reuse of media assets are key factors for the effective development of new courses. This activity is spread among several units of the University, each specialised in different types of assets. In particular, the units that have the role of disseminating this content across social media are different from the ones that produce this content. Faculties do the actual teaching and produce content, while specialised units develop media items or have the duty of dissemination and students recruitment. Knowledge sharing and reuse in this context is clearly challenging, and where Linked Open Data have an obvious role to play.

1.2. Background

Born in November 2010 as the final outcome of the LUCERO project[5], *data.open.ac.uk* was the first of its kind in the UK [1] as an initiative to expose public information from across the University in an accessible, open, integrated and Web-based format[6]. Since then, a number of British and European universities have followed, taking the experience and model generated by LUCERO to open up data for education and research [2].

In nearly 4 years, the linked data services have progressed in many ways, e.g. (1) more connections and greater engagement of OU units; (2) end-user services built on top of linked data beyond the experimental or prototypical stage; (3) constant evolution in the infrastructure.

data.open.ac.uk is today a reliable, constantly monitored service, whose data are updated on a daily basis. The quality of service offered has led to a steady increase in its usage. Applications are using *data.open.ac.uk* to obtain official information about courses and qualifications and for the discovery of and linkage to relevant content spread across the heterogeneous landscape of systems, websites and repositories of the OU, which we shall discuss in Section 5.

[5] `http://lucero-project.org`

[6] See the press release from November 2010: `http://www3.open.ac.uk/media/fullstory.aspx?id=20073`

2. The Data

data.open.ac.uk publishes Five-Star open data[7]: the dataset (1) is available on the Web (2) as structured data (3) in a standard, non-proprietary format (RDF), (4) uses (resolvable) URIs to denote things and (5) links to other datasets (see Section 2.2 for details). Moreover, it includes a number of features linked data endpoints should have in order to support discovery and reuse. Table 1 lists some key facts about *data.open.ac.uk*. The references in the table assume the `http://data.open.ac.uk` prefix, which we shall hereinafter omit for *data.open.ac.uk* URIs. The `/context/` graph represents the default union of all available named graphs, yet its URI is resolvable, as is any URI in *data.open.ac.uk* (see Section 3.5 for more details). The SPARQL service description can be obtained resolving the URI of the SPARQL endpoint, if a client requests a non-HTML format[8].

Table 1
Key facts about *data.open.ac.uk*.

Name:	Linked Data from The Open University
Domain:	`http:data.open.ac.uk`
VoID description:	`/void`
Sitemap:	`/sitemap.xml`
Union graph:	`/context/` (default graph)
Graphs:	11 officially released, 30 in total (See Section 6)
Triples:	2.865.651 officially released, 3.720.570 in total
Vocabularies:	57
Classes:	125
Properties:	785
SPARQL:	`/sparql`

2.1. Topic coverage, sources for the data

The dataset is organised in a number of graphs, containing data that are collected from various sources, which may be public websites or content management systems internal to the university. The data are exposed through a single SPARQL endpoint and can be queried all together. However, each data portion is identified by a *named graph*, reflecting its primary source. Graph names are resolvable URIs defined in the namespace `http://data.open.ac.uk/context/`. From now on, we will use the prefix `g:` to refer to graph names[9]. In this Section the con-

[7] 5-star data deployment scheme, `http://5stardata.info`

[8] The redirection can be forced using the URL: `http://data.open.ac.uk/resource/sparql`.

[9] Graph names, like any URI in *data.open.ac.uk*, are resolvable. `http://data.open.ac.uk/context/` provides the description of the union graph.

tent of the officially released datasets[10] is described. We shall group the graphs under six themes:

1. Open Educational Resources: g:openlearn, g:openlearnexplore, g:podcast, g:bbc
2. Scientific production: g:oro
3. Social media: g:audioboo, g:youtube
4. Organisational data: g:people/profiles, g:people/kmifoaf, g:qualification, g:course, g:kmiplanet, g:kis
5. Research project output: g:red, g:redperssa, g:led
6. Metadata: g:meta, g:ontology, g:about, g:catalogue

2.1.1. Open Educational Resources

A significant part of the information is metadata about educational resources produced or co-produced by the OU. Open Learn is *the home of free learning from The Open University*[11]. The Web portal includes a large number of free learning units as well as articles exploring a wide range of topics, often embedding media content. Data are collected from RSS feeds and exposed as RDF in the graphs g:openlearn and g:openlearnexplore. Other media objects are catalogued by internal content management systems and metadata are extracted and translated into RDF. It is the case of video and audio podcasts hosted at http://podcast.open.ac.uk/ and published in g:podcast. Similarly, co-productions with the BBC[12] are collected in the graph g:bbc, which also links to entities that represent BBC programmes.

2.1.2. Scientific production

Other open data that exist in another form and are transformed and linked are those of the Open Research Online repository (ORO)[13]. Open Research Online is the repository of scholarly publications and other research output of the OU. It is an Open Access resource that can be searched and browsed freely by the public. The graph has name g:oro.

2.1.3. Social media

Content is often hosted by third-party organisations, and metadata are extracted from public APIs and aggregated into RDF. The OU publishes media on YouTube – g:youtube and Audioboo – g:audioboo. Objects are often annotated with courses, qualifications or OU people they relate to. Playlists and metadata about videos and audio podcasts are extracted from Web APIs, then translated and enriched to interlink with the other entities in *data.open.ac.uk*.

2.1.4. Organisational data, courses, people, news

In other cases, data are collected form internal repositories and first made public as linked data. It is the case of reference data about courses – g:course and qualifications – g:qualification under presentation, as well as the profiles of researchers and academic staff – g:people/profiles. The Key Information Set of the OU is published by HESA[14] as part of the Unistats dataset[15]. This dataset is transformed and made available as Linked Data by the LinkedUp project[16]. The sub-graph of this dataset focusing on the OU is also published in *data.open.ac.uk* in the graph g:kis. Finally, the graphs g:kmiplanet and g:people/kmifoaf provide data about news and staff of the Knowledge Media Institute of the OU.

2.1.5. Data from research projects

data.open.ac.uk also hosts data produced by research projects. At the moment there are three datasets officially published, that come from two projects of the Arts and Humanities Faculty of the OU: the Reading Experience dataset[17] - g:red/g:redperssa and the Listening Experience dataset[18] - g:led.

2.1.6. Metadata

An important requirement of any RDF database is to specify and document its structure to support tools to discover data, automatically detect its characteristics and possibly configure their behavior accordingly. It is also necessary that open data expose and document their schema so that users can make sense of it. Four data spaces are dedicated to metadata:

- g:meta Graph metadata using mainly VoID and the SPARQL Service Description graph [3];
- g:ontology Definitions of terms used, particularly those defined in the *data.open.ac.uk* domain;
- g:about Graph metadata containing links to DBpedia entities that are topics of open educational resources and other entities of the *data.open.ac.uk* graphs;

[10] Section 6 provides further insight into the release process.
[11] http://www.open.ac.uk/openlearn
[12] http://www.bbc.co.uk
[13] http://oro.open.ac.uk
[14] Higher Education Statistics Agency, http://hesa.ac.uk
[15] Unistats data http://unistats.direct.gov.uk/open-access-data/.
[16] http://www.linkedup-project.eu
[17] http://www.open.ac.uk/Arts/RED/
[18] http://led.kmi.open.ac.uk/linkeddata/

– `g:catalogue` A description of *data.open.ac.uk* using the DCAT specification [4], describing the different ways data are distributed (mainly SPARQL endpoint and downloadable files).

2.2. Links

Links to entities in other graphs enable data integration with little effort. A major example is the usage of courses as aggregators of similar objects through the graphs. Courses are referenced by almost all sources in the university, thus enabling a good potential for use cases like recommendation (see also Section 5). This interconnection between the data from different graphs is an important feature, and most of the applications developed within the OU rely on it. The data are also linked to external datasets in the Linked Data Cloud. The `g:kis` graph includes `owl:sameAs` links of OU qualifications to the same entities in the external KIS Linked Data endpoint[19]. Courses under presentation link to `http://sws.geonames.org` to contextualise the offer, which may be subject to regional pricing[20]. The BBC coproductions graph points to BBC entities[21].

A dedicated graph includes links to DBpedia entities: they are the topics of media objects, Web pages, courses and other entities in the OU data environment. These topics are generated by DiscOU [5], an application that annotates documents of *data.open.ac.uk* entities with DBpedia entities. For example, OpenLearn Units are made of a number of web pages or video podcasts have transcripts. These are consumed by DiscOU to generate a number of `dc:subject` links to DBpedia. These links are published in the `g:about` graph.

The graph `g:led` includes time-related data, with links to the related entity in the `http://reference.data.gov.uk/id/gregorian-instant/` namespace as well as the corresponding `xsd:dateTime` literal value. Links to Web pages containing human-readable information are provided by all graphs. Sometimes these pages are also the primary source of the data. Metadata also link directly to downloadable media objects. Table 2 outlines some details about the links that exist in the different graphs.

[19] `http://data.linkedu.eu/kis/query`.
[20] Example: `http://data.open.ac.uk/course/y031`
[21] Example: `http://data.open.ac.uk/bbc/b00mfl7n`

Table 2
Graphs, links and target datasets[22]

Graph	Property	∼No	Target
course	gr:availableAtOrFrom	50000	sws.geonames.org
oro	owl:sameAs	7500	oro.open.ac.uk
redperssa	owl:sameAs	6500	dbpedia.org
bbc	owl:sameAs	120	www.bbc.co.uk
led	event:time	900	reference.data.gov.uk/
kis	owl:sameAs	160	data.linkedu.eu/kis/
about	dc:subject	10000	dbpedia.org

2.3. Licensing

Unless otherwise specified, the data under `http://data.open.ac.uk` are licensed under a Creative Commons Attribution 3.0 Unported License[23]. Specific graphs might have a different license. For instance, the `g:people/profiles` graph is distributed under the Attribution-NonCommercial-ShareAlike 4.0 International license.

2.4. The potential of linked data: a simple query

The linked data of The Open University are a dense network of interlinked entities. A typical use case is the discovery of open educational material that is related to a known media asset. If a user is interested in the OU YouTube video `https://www.youtube.com/watch?v=NcFrxXKtoXk`, the following query on *data.open.ac.uk* can help her find a number of other educational resources, as well as courses on offer:

```
prefix schema: <http://schema.org/>
prefix ou: <http://data.open.ac.uk/ontology/>
SELECT *
  FROM <http://data.open.ac.uk/context/youtube>
  FROM <http://data.open.ac.uk/context/audioboo>
  FROM <http://data.open.ac.uk/context/openlearn>
  FROM <http://data.open.ac.uk/context/course>
  WHERE {
    ?x schema:productID "NcFrxXKtoXk" .
    ?x  ou:relatesToCourse ?course .
    ?related ou:relatesToCourse ?course   }
```

3. Modelling issues

Several design choice for the *data.open.ac.uk* data models are aimed at making the approach to data as reusable as possible. In the following, we address the different aspects arising from Linked Data design.

3.1. Design of Graphs

The totality of the data in the repository is obtained from external sources. Triples in the data store are or-

[23] `https://creativecommons.org/licenses/by/3.0/`

ganised by source, so that all triples coming from a source are stored in a dedicated graph. This data management pattern is called "Graph-Per-Source" [6]. This choice allows for easier maintenance and update. See Section 6 for details.

3.2. Design of entity URIs

Identifiers (URIs[24]) have a crucial role in supporting users and developers in the usage of data. Readable and meaningful URIs help users make sense of a data snippet with a lesser effort than with opaque ones. *data.open.ac.uk* adopts a number of known patterns [6] for building nice identifiers:

a) External identifiers are reused when available. This practice follows the "Natural Keys" pattern [6]). It is the case of courses and qualifications, but also OU accounts and publications in ORO:
   `/course/a100`
   `/qualification/q46`
   `/account/sab668`
   `/oro/21166`
   This pattern is fundamental to users who are familiar with the organisational structure of the OU, as entity codes often play a key role in the communication flow of large organisations.
b) A readable type description of the entity is referenced within the path of the URI. This helps classify an item at first sight, reducing the need for additional queries, e.g.:
   `/person/0fe4...dbe76`
   `/member/0fe4...dbe76/organization/kmi`
   The second example above is slighty more complex as it reflects the statement: *This is a membership of this ID to an organisation, that is KMi*.
c) Often, a hierarchical URI is created by using the graph/source name at the beginning of the path, and then replicating the local identifier of the source, following also the Natural Keys pattern:
   `/audioboo/playlist/1252980-women-in-science`
   `/youtube/KcrtCncGAEo`

However, this structure of URIs is not enforced, as new data sources might require different strategies.

3.3. Cool URIs and nice HTTP headers

In contrast, there is a strong commitment to provide dereferenceable, "cool" URIs[25]. Any entity in the `http://data.open.ac.uk/` namespace can be resolved as an HTML page or an RDF document according to HTTP content negotiation[26]. The HTTP response supports Cross-Origin Resource Sharing[27] and a custom header pointing to the SPARQL endpoint. For example, the following request with no Accept header will be redirected to an RDF document:

```
$ curl -I http://data.open.ac.uk/account/sab668
HTTP/1.1 302 Found
Location: http://data.open.ac.uk/resource/account/
     sab668
Access-Control-Allow-Origin: *
X-SPARQL: http://data.open.ac.uk/sparql
...
```

The HTML representation of the resource is served as a different document:

```
$ curl -I http://data.open.ac.uk/account/sab668 -H
     "Accept: text/html"
HTTP/1.1 302 Found
Location: http://data.open.ac.uk/page/account/
     sab668
Access-Control-Allow-Origin: *
X-SPARQL: http://data.open.ac.uk/sparql
Content-Type: text/html; charset=UTF-8
...
```

3.4. Domain modeling: guidelines and evolutions

The dataset includes 125 classes and 785 properties from 57 public vocabularies. The choice of terms to be used is based on the following process: (1) identify the concept to be expressed; (2) search for a widespread existing vocabulary to be used; (3) if found, use it, otherwise (4) search for a less-known vocabulary to reuse; (5) if not found, create a new term; (6) in either case, if there is no well-known term to be used, try to generalise the concept and add an additional statement with a well-known term. This approach led to the adoption of a variety of vocabularies. Another consequence of this method is that sometimes information is redundant, being repeated with different properties such as a generic well-known term and a more specific less known (or proprietary) term. These are consequences of the choice to privilege the reuse of existing terms and the will to choose the best possible terms instead of being restricted to the semantics of only a few widely used ontologies. For reasons of space here we will only mention some vocabularies that are widely used across many graphs. FOAF, SKOS, SIOC, OWL,

[24] `http://www.isi.edu/in-notes/rfc2396.txt`
[25] `http://www.w3.org/TR/cooluris/`
[26] `http://tools.ietf.org/html/rfc7231#section-5.3`
[27] `http://www.w3.org/TR/cors/`

Dublin Core are used by almost all graphs. GoodRelations is used by `g:course` to specify the learning offer of the University. This vocabulary is particularly useful because the OU is a decentralised institution, and students are recruited all over the world, so prices and features of the offer may differ. Media ontologies (video, audio) are also used to describe aspects of media objects. Schema.org[28] is used in some cases, and there are plans to extend its usage to other graphs. Bibo[29] is used to describe library items and publications and XCRI[30] to describe courses and course material.

Courses and qualifications are entities with a special role in strenghtening the interlinking between graphs. Their codes are widely used within the university to annotate documents, media objects or Web pages. The requirement here is to have the possiblity to query for all content related to a given course (or qualification), or restricting the range of values to a specific graph population. There is a general property, named `http://data.open.ac.uk/ontology/relatesTo` that is widely used for this purpose. Moreover, this property is specialised in different ways:

```
| relatesTo
| - relatesToCourse
| - /bbc/ontology/relatesToCourse
| - /.../ontology/relatesToCourse
| - relatesToQualification
| - /audioboo/ontology/relatesToQualification
| - /.../ontology/relatesToQualification
```

This set of properties allows for easy querying by filtering the source of the linked entities with basic triple patterns, without the need for further constraints on the `rdf:type` or the named graph. This shortcut simplifies a number of very common queries.

3.5. Metadata modeling

VoID[31], SD[32] and DCAT[33] are the basis to describe the metalevel aspects of *data.open.ac.uk* and its graphs. Like any entity in the dataset, graph names are resolvable URIs. In the `g:meta` graph different kind of datasets are described. Named graphs are typed as `void:Dataset`, `sd:Graph` and `sd:NamedGraph`. The default graph (non named) is still described using the URI `http://data.open.ac.uk/context/` as the union of all graphs, with the only difference that it is not a `sd:NamedGraph`. Class and Property partitions are still `void:Datasets`, but are neither `sd:Graph` nor `sd:NamedGraph`.

All entities link to their named graph with the property `void:inDataset`. In this way linked data agents can reach the dataset description from any URI, for example to obtain the address of the SPARQL endpoint. This is also useful to filter the context of the entity without the need for a quad pattern, which can be hard to include in some complex SPARQL queries.

3.6. Blank nodes and other modelling issues

The general policy for blank nodes is to avoid them. The simple motivation is that entity resolution is mapped to a SPARQL query. To return complete data, the query should traverse all blank nodes and return the full sub-graph. Although certain SPARQL engines support traversal in their way of handling DESCRIBE queries, one cannot rely upon non-normative implementation features, and handling blank nodes with other query types still falls short of efficiency.

Another design choice was the use of RDF cardinal properties to list the positions of authors of publications. As described in the specificaion of RDFS: "*Container membership properties may be applied to resources other than containers*". In the `g:oro` graph, container membership properties are used alongside `dc:creator` for each author. This redundancy allows for compatibility with the widely known Dublin Core vocabulary. Users can decide to use the shallow `dc:creator` property or the fine-grained rdf membership property, depending on her query requirements, as in the following listing:

```
select ?account (count(?pub) as ?No)
from <http://data.open.ac.uk/context/
                          people/profiles>
from <http://data.open.ac.uk/context/oro>
{
  ?pub rdf:_1 ?author .
  ?author foaf:account ?account
} group by ?account
order by desc(?No)
```

Finally, all entities have a single, untyped and not lang-tagged `skos:prefLabel` literal. This is a convenience that provides a human-readable table of entities whilst still allowing multiple `rdfs:labels` to support multilinguality.

[28] http://schema.org
[29] http://bibliontology.com/
[30] http://www.xcri.co.uk/
[31] http://www.w3.org/TR/void/
[32] http://www.w3.org/TR/sparql11-service-description/
[33] http://www.w3.org/TR/vocab-dcat/

4. Services

The services provided are far more than the common Linked Data services and take into account the needs of developers, the final "customers" of a data service. Besides the SPARQL 1.1 compliant query endpoint and the resolution of all entity URIs under the `http://data.open.ac.uk/` domain using a number of RDF and non-RDF formats, some other useful funtionalities have been setup.

The `/about` service allows to resolve any URI by looking up the triple store.For example:

```
curl http://data.open.ac.uk/about/ -G --data-
    urlencode uri=http://www.bbc.co.uk/
    programmes/b021n3x1#programme -H "Accept:
    text/turtle"  -L
```

The `/lookup` service is used to retrieve entities from some well-known codes, such as the OU employee username (OUCU, e.g. ed4565), the Youtube video ID or the course code (e.g. A100)[34]

Data can be queried with SPARQL through the endpoint provided. Developers can embed the query in their code and execute it at runtime. However this practise creates a strong dependency between the application and the database. This dependency might create problems for the developers, because they do not have control of the data source, so they cannot know whether the query would continue functioning when changes on the data occur. A practical solution to this problem has been to setup a stored queries endpoint. Developers can store their queries on the server and use a plain URL to point to the data. Maintainers can then manage the evolution of the database and inform the developers of coming evolutions, when it might affect an existing query. This service is available only to applications developed internally to the university.

The content of the graphs is archived weekly, and the versions are made available for download from a section of the web site.

5. Usage

The goal of exposing interlinked data on *data.open.ac.uk* is to make existing public data more accessible, reusable and exploitable. This can only be demonstrated through applications making use of this data in innovative and/or cost effective ways. Various production systems are using *data.open.ac.uk* as source of information. For example, the OpenLearn web site queries the SPARQL endpoint to get the list of qualifications under presentation and related information. Similarly, a system from the Student Services Unit of the OU scans *data.open.ac.uk* to upgrade the list of available courses.

An application in the OU YouTube space queries *data.open.ac.uk* to get related courses and qualifications as well as other open educational content (this use case has been described in Section 2.4).

DiscOU [5] is a recommender system developed by the *data.open.ac.uk* team to support the discovery of open educational content similar to other online resources like a BBC program or a web page. This system builds an index of the open educational resources catalogued in the *data.open.ac.uk* dataset that includes a set of DBpedia entities that are representative of the resource. This index is then exploited by a similarity algorithm. This application had two nice consequences. The first is that it boosted the adoption of linked data within the university by giving an exemplary use case that is very hard to implement with legacy technologies. The second is that we used the content generated by the tool to populate the graphs of topics - `g:about`, as already described in Section 2.

These of course are only of few of the applications developed on top of the dataset, others being described on the *data.open.ac.uk* website and in [7].

Figure 5 displays the result of an analysis performed on server logs. This historical view displays the number of clients using *data.open.ac.uk*from the launch of the platform on September 2010 until today (September 2014). It can be seen that the number of clients doubled in time, particularly in the last two years. This gives a promising perspective on the adoption of linked open data.

6. Maintenance

data.open.ac.uk includes many graphs under different stage of development. The graphs considered stable are also documented on the web page of the site, and this is also reflected in the RDF description. A graph is considered stable when:
a) the process that led to the data acquisition is robust and the data provider is considered reliable (can guarantee future updates);
b) the infrastructure includes a robust update mechanism that guarantees the data to be up to date regularly;
c) or it is a dataset that does not need to be updated.

[34] Other examples can be found on the `http://data.open.ac.uk/` home page

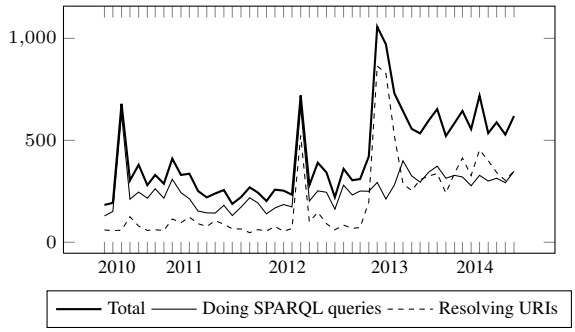

Fig. 1. The diagram above displays the progress in number of clients requesting RDF data (not HTML) on a monthly basis from the launch of *data.open.ac.uk*on September 2010 to September 2014. There are some visible peaks. The first is on 2010-11 (679 clients), then on 2012-08 (720) and 2013-05 (1057). The first is most probably related to the first launch of *data.open.ac.uk*, we presume the others to be related to the release of a new application consuming *data.open.ac.uk* data.

data.open.ac.uk publishes officially 11 stable datasets but many others are under the hood and might become "stable" as well[35].

The dataare basically a snapshot of the status of the related information at a given time. Most of the graphs are updated on daily basis. A special case is the `g:people/profile` graph, that is updated in real time from the source content management system, to immediately react to the change of policy that users might operate with respect to the privacy status of their data. Since the lifecycles of the graphs differ, the infrastructure supports three different methods:
1) graph rebuilt: the data is rebuilt entirely and a new version substitutes the previous (eg.: `g:course`);
2) incremental update: data is never deleted, and new content is added once available (eg.: `g:bbc`); and
3) synchronisation: changes in the source are reflected on the RDF graph as soon as possible (`g:people/profile`).

The maintenance of the dataset includes weekly dumps for history analysis and as backups for disaster recovery. Data is tested before the changes to be applied in order to detect some common errors. This include bad responses from external servers or corrupted downloaded files.

The repository contains more than 3.500.000 RDF triples. While this is a fairly large amount, it is far from creating scalability issues with state of the art triple stores. Indeed, the *data.open.ac.uk* platform only rarely experiences any downtime (which are mostly due to planned maintenance on the infrastructure), while being supported by a small team (amounting officially to 50% of a developer).

[35]Try out this query to see more: `SELECT DISTINCT ?G WHERE {GRAPH ?G {[] ?P []}}}`

7. Ongoing work

While *data.open.ac.uk* has evolved into a "grown-up" semantic dataset, some work is still required to make it the reference method for open data integration in the organisation. There are plans to include new data, like the upcoming course description using XCRI 2.0 as well as library data from the OUDL project[36]. Metadata are also an important asset of *data.open.ac.uk*. An investigation on the way to provide provenance information for both entity resolution and SPARQL queries is ongoing. We are observing the evolution of linked CSV specifications, and considering a service that provides predefined views over the triple store listing types of objects with their properties in this format. New data will also cover more information about people profiles and a complete graph dedicated to the organisational structure of the University.

References

[1] Fouad Zablith, Mathieu d'Aquin, Stuart Brown, and Liam Green-Hughes. Consuming linked data within a large educational organization. In *Second International Workshop on Consuming Linked Data (COLD) at 10th International Semantic Web Conference (ISWC 2011)*. Springer, 2011.

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

[3] Gregory Williams. SPARQL 1.1 service description. W3C recommendation, W3C, March 2013. http://www.w3.org/TR/2013/REC-sparql11-service-description-20130321/.

[4] John Erickson and Fadi Maali. Data catalog vocabulary (DCAT). W3C recommendation, W3C, January 2014. http://www.w3.org/TR/2014/REC-vocab-dcat-20140116/.

[5] Mathieu d'Aquin, Carlo Allocca, and Trevor Collins. Discou: A flexible discovery engine for open educational resources using semantic indexing and relationship summaries. In *International Semantic Web Conference (Posters & Demos)*, 2012.

[6] Leigh Dodds and Ian Davis. Linked data patterns. *Online: http://patterns. dataincubator. org/book*, 2011.

[7] Mathieu d'Aquin. Putting linked data to use in a large higher-education organisation. In *Interacting with Linked Data at Extended Semantic Web Conference, ESWC 2012*, 2012.

[36]`http://www.open.ac.uk/blogs/OUDL/`