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

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

Editor(s): Philippe Cudre-Mauroux, University of Fribourg, Switzerland
Solicited review(s): Amrapali Zaveri, University of Leipzig, Germany; Michael Luggen, Bern University of Applied Sciences, Switzerland; Alex Olivieri, University of Applied Sciences and Arts Western Switzerland Valais (HES-SO Valais-Wallis), Switzerland

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, scholarly publications 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*. 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

The Open University (OU) has a vast presence on the WWW. Having distance learning in its core, the dissemination 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[1], and has a number of channels on social media such as YouTube[2] or Audioboo[3].

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 across several units of the University, each specialised in different types of assets. In particular, the units that have the role of disseminating 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 are in charge of dissemination and student recruitment. Knowledge sharing and reuse in this context is a challenge where Linked Open Data have an obvious role to play.

`http://data.open.ac.uk` is the home of linked open data from the OU. The data available come from various institutional repositories of the University, and are collected, interlinked and openly published for reuse. Born in November 2010 as the final

[1] `http://www.open.ac.uk/openlearn/`
[2] `https://www.youtube.com/user/TheOpenUniversity`
[3] `http://audioboo.fm/search?utf8=%E2%9C%93&q=the+open+university`

outcome of the LUCERO project[4], *data.open.ac.uk* was the first initiative to expose public information from across the University in an accessible, open, integrated and Web-based format[5]. 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 four 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 that progressed beyond the experimental or prototypical stage; (3) constant evolution in the infrastructure.

In Section 2 we describe the data included in *data.open.ac.uk*, while Section 3 focuses on the design aspects and the modelling choices that have been made. The services offered by *data.open.ac.uk* are described in Section 4, and their usage in applications in Section 5. Technical aspects and maintenance issues are illustrated in Section 6, before concluding the article with future work.

2. The Data

data.open.ac.uk publishes Five-Star open data[6]: 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. Moreover, it includes a number of features to support discovery and reuse of its linked data endpoint. 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*. The SPARQL service description can be obtained resolving the URI of the SPARQL endpoint, if a client requests a non-HTML format[7].

data.open.ac.uk includes many graphs at different stages of development. The graphs that are considered stable are officially released, documented on the website, and marked as such in the RDF description. A graph is considered to be stable when:
a) the process that led to the data acquisition is robust and the data provider is considered reliable (i.e. can guarantee future updates);
b) the infrastructure includes a robust update mechanism that guarantees the data are updated regularly, unless they are static data that do not need to be updated.
11 stable graphs have been released so far through *data.open.ac.uk*, but many others are also available with lower degrees of support and warranty[8]. In this article we focus on the graphs officially released.

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

Name:	Linked Data from The Open University
Address:	`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 and data sources

The dataset is collected from various sources, which may be public websites or content management systems internal to the University. Data from each source at hand are completely remodelled to include as much information as in their original form. Remodelled data are then exposed through a single SPARQL endpoint and can be queried as a whole. 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.

We shall group the graphs under six themes:

1. Open Educational Resources
2. Scientific production
3. Social media
4. Organisational data
5. Research project output
6. Metadata

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

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

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

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

[8] The following query lists all graphs: `SELECT DISTINCT ?G WHERE {GRAPH ?G {}}`

2.1.1. Open Educational Resources

A significant part of the information in this category is metadata about educational resources produced or co-produced by the OU. Open Learn is *the home of free learning from The Open University*[9]. 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[10] 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)[11]. 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 corresponding graph is `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 from 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[12] as part of the Unistats dataset[13]. This dataset is transformed and made available as Linked Data by the LinkedUp project[14]. The sub-graph of this dataset focusing on the OU is also published in *data.open.ac.uk* as `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 OU Faculty of Arts, namely the Reading Experience dataset[15] (`g:red` and `g:redperssa`) and the Listening Experience dataset[16] (`g:led`).

2.1.6. Metadata

An important requirement of any RDF database is to specify and document its structure to support external agents in automatically discovering data, detecting the characteristics of the data and possibly configuring their behavior accordingly. It is also useful for open data to expose and document their schema so that users can make sense of it. Three data spaces are dedicated to metadata: 1) `g:meta` - Graph metadata using mainly VoID and the SPARQL Service Description; 2) `g:ontology` - Definitions of terms used, particularly those defined in the *data.open.ac.uk* domain; 3) `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.

2.2. Links

Links between entities of different graphs enable data integration with little effort. A key example in *data.open.ac.uk* is the use of courses as aggregators for similar, related objects through the graphs. Courses are referenced by almost all sources in the University, thus enabling use cases such as content recommendation[17]. The data are also linked to external datasets in the Linked Data Cloud. The `g:kis` graph includes `owl:sameAs` links from OU qualifications to the same entities in the external KIS Linked Data end-

[9] `http://www.open.ac.uk/openlearn`

[10] `http://www.bbc.co.uk`

[11] `http://oro.open.ac.uk`

[12] Higher Education Statistics Agency, `http://hesa.ac.uk`

[13] Unistats data `http://unistats.direct.gov.uk/open-access-data/`.

[14] `http://www.linkedup-project.eu`

[15] `http://www.open.ac.uk/Arts/RED/`

[16] `http://led.kmi.open.ac.uk/linkeddata/`

[17] While this use case is one of the most interesting, and applications using *data.open.ac.uk* do implement it in different ways, evaluating it is out of the scope of this article.

point[18]. Courses under presentation link to `http://sws.geonames.org` to contextualise the offer, which may be subject to regional pricing[19]. The BBC coproductions graph points to BBC entities[20].

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 [1], 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 collected by DiscOU and analysed using DBpedia Spotlight[21] 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 related entities in the `http://reference.data.gov.uk/id/gregorian-instant/` namespace as well as the corresponding `xsd:dateTime` literal values. Links to Web pages containing human-readable information are provided by all the graphs. Sometimes these pages are also the primary source of the data. Metadata also link directly to downloadable media objects. Links (especially internal ones) in most datasets are generated from the source information and are 100% accurate. Others, such as DiscOU are automatically generated and, although evaluating them is out of our scope, some known inaccuracies are being recorded. Table 2 outlines some details about the external links that exist in the different graphs.

Table 2

Graphs, links and target datasets.

Graph	Property	∼No	Target
about	dc:subject	490000	dbpedia.org
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	500	reference.data.gov.uk
kis	owl:sameAs	80	data.linkedu.eu

2.3. Licensing

Unless otherwise specified, the data released through *data.open.ac.uk* are licensed under a Creative Commons Attribution 3.0 Unported License[22]. Specific graphs might have a different license. For instance, the `g:people/profiles` graph is distributed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license.

3. Modelling issues

Several design choices for the *data.open.ac.uk* data models are aimed at making the 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 organised 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” [3]. One drawback of this approach is that if the same information is contributed by several sources, it will be replicated in the two graphs and might lead to inconsistencies between graphs. However, assessing the provenance of each statement (graph) is a core requirement for a service that publishes integrated data, and facilitates maintenance.

3.2. Design of entity URIs

Identifiers (URIs[23]) 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 [3] for building such identifiers:

a) External identifiers are reused when available. This practice follows the "Natural Keys" pattern [3]. It is the case of courses and qualifications, but also of 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.

[18] `http://data.linkedu.eu/kis/query`.

[19] Example: `http://data.open.ac.uk/course/y031`

[20] Example: `http://data.open.ac.uk/bbc/b00mfl7n`

[21] `http://spotlight.dbpedia.org/`

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

[23] `http://www.isi.edu/in-notes/rfc2396.txt`

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:

```
/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. Domain modeling

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. 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, 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[24] is used in some cases, and there are plans to extend its usage. Bibo[25] is used to describe library items and publications and XCRI[26] 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 opportunity here is 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 ...
 relatesToQualification
  /audioboo/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.

3.4. Metadata modeling

VoID[27] and SD[28] 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 not instances of `sd:Graph` or `sd:NamedGraph`.

All entities link to their named graphs 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.

[24] http://schema.org

[25] http://bibliontology.com/

[26] http://www.xcri.co.uk/

[27] http://www.w3.org/TR/void/

[28] http://www.w3.org/TR/sparql11-service-description/

3.5. Blank nodes and other modeling issues

Naming resources globally enables important facilities for data consumption and maintenance. This is why we avoid blank nodes. It is of great usefulness to be able to identify a resource in a graph-independent way: a) any entity from a previously selected result set can be inspected in the current endpoint by only knowing the identifier, and b) it is easier to compare dataset dumps, reducing the operation to a *diff* on two sorted triple collections. A negative consequence of blank nodes is that the data become redundant on incremental updates, as updating the same information twice will add the information twice because of blank nodes being local identifiers. Similarly, users who download the same data might end up with different RDF graphs, making it harder to maintain consistency in their applications. Also, we have not been able to so far identify any practical advantage on using blank nodes instead of portable identifiers in *data.open.ac.uk*.

Another design choice was the use of RDF cardinal properties to list the positions of authors of publications. As described in the specification 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 query:

```
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 name for entities whilst still allowing multiple `rdfs:labels` to support multilinguality.

4. Services

There is a strong commitment to provide dereferenceable, "cool" URIs[29]. 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[30]. The HTTP response supports Cross-Origin Resource Sharing[31] and a custom header pointing to the SPARQL endpoint. However, *data.open.ac.uk* provides more services than the common Linked Data ones and take into account the needs of developers, who are typically the end "customers" of a data service. Indeed, besides the SPARQL 1.1 compliant query endpoint, some other useful funtionalities have been setup.

The `/about` service allows to resolve any URI by looking up the triple store. For example, the following request will show information available about a URI outside the *data.open.ac.uk* domain:

```
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)[32].

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 practice 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. One practical solution to this problem was to setup an endpoint for stored queries. 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 on a weekly basis, and the versions are made available for download from a section of the website.

[29] `http://www.w3.org/TR/cooluris/`

[30] See `http://tools.ietf.org/html/rfc7231#section-5.3`. Sometimes content negotiation is hard to use for developers. For this reason the *data.open.ac.uk* server also support the `output` query parameter.

[31] `http://www.w3.org/TR/cors/`

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

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 that make 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 website queries the SPARQL endpoint to get the list of qualifications under presentation, along with 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. If a user is interested in, for instance, the OU YouTube video `https://www.youtube.com/watch?v=NcFrxXKtoXk`, the following query to *data.open.ac.uk* can retrieve 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 }
```

DiscOU [1] 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 processed by a similarity algorithm. Two positive outcomes of this application have been recorded. First, it boosted the adoption of linked data within the University by giving an exemplary use case that is otherwise very hard to implement using legacy technologies. Second, we used the content generated by the tool to populate the graphs of topics `g:about`, as already described in Section 2.

These are but a few of the applications developed on top of the dataset. Others are described on the *data.open.ac.uk* website and in an earlier paper [4].

Figure 1 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 shows that the number of clients has since doubled over time, particularly in the last two years. This gives a promising perspective on the adoption of linked open data in this context.

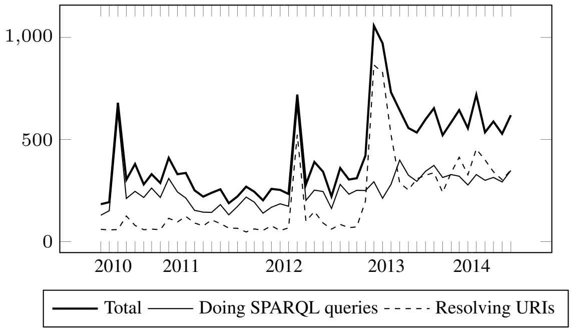

Fig. 1. The diagram above displays the progress in number of clients requesting RDF data (not HTML) on a monthly basis from the launch of *data.open.ac.uk* in September 2010 to September 2014. There are some visible peaks. The first is on 2010-11 (679 clients), then on 2012-08 (720) and 2013-05 (1057).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

6. Technical notes and maintenance

The data are basically a snapshot of the status of the related information at a given time. Most of the graphs are updated on a daily basis.

Since the lifecycles of the graphs differ, the infrastructure supports three different update policies:
1) graph rebuild: the data are rebuilt entirely and a new version substitutes the previous (e.g. `g:course`);
2) incremental update: data are never deleted, and new content is added once available (e.g. `g:bbc`); and
3) synchronisation: changes in the source are reflected on the RDF graph as soon as possible (e.g. `g:people/profile`).

The data import activity is performed with a number of dedicated procedures, orchestrated according to the specific case. We can summarise the process as follows, abstracting from the specific cases:

1. *data collection*: an item is collected from a data source;
2. *transformation*: the item is inspected, the data translated into RDF and enriched, by materialising some inferences and eventually inspecting other data sources;
3. *update plan*: depending on the update policy, the commands to perform the change are prepared (for instance to replace the whole data in a graph, or to perform a delete query and then add the content of the file);

4. *update execution*: the related graph is updated.

The datasets updated daily (both as full replacement or incremental additions) might have misalignments for up to 24 hours with respect to the sources. However, the change rate of that information is slow and full timeliness not really important for existing applications. This process requires less than one hour to be completed, and it does not affect the running system until data replacement occurs, which runs in less than 2 minutes. The information published tends to reflect the sources as much as possible, and inaccurate information is identified and corrected following feedback from users. A special case is the `g:people/profile` graph, which 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. When a profile is updated, the change is notified to the updating procedure that adds the profile to a queue, which is then inspected regularly. Relevant triples are deleted with an ad-hoc query and the new version is loaded. The method guarantees full accuracy and good timeliness despite loading the live system with write transactions[33].

data.open.ac.uk code is mostly written in PHP (front-end services) and Java (data importing and remodeling). The system relies on existing, open source software, especially the Fuseki server from Apache Jena[34]. The maintenance of the dataset includes weekly dumps for history analysis and as backups for disaster recovery. Data are tested before applying changes in order to detect common error patterns. This includes 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 causing scalability issues with state-of-the-art triple stores. Indeed, the *data.open.ac.uk* platform only rarely experiences any downtime, and even that is mostly due to planned maintenance on the infrastructure, given that it is supported by a small team (officially amounting to 50% of a single developer).

7. Ongoing work

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 heterogeneus landscape of systems, websites and repositories of the Open University. While *data.open.ac.uk* has evolved into a full-grown semantic dataset, some work is still required to make it the reference method for open data integration in the organisation. In particular, there is a need for new tools and services that can make the data we offer easier to explore, understand, query and embed in applications efficiently. Metadata are also an important asset of *data.open.ac.uk*. An investigation into the ways 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. There are plans to include new data, such as the upcoming course description using XCRI 2.0, as well as library data from the OUDL project[35]. Other sources of data will also cover more information about people profiles and a complete graph dedicated to the organisational structure of the University.

References

[1] M. d'Aquin, C. Allocca, and T. Collins. Discou: A flexible discovery engine for open educational resources using semantic indexing and relationship summaries. In B. Glimm and D. Huynh, editors, *Proceedings of the ISWC 2012 Posters & Demonstrations Track*, Boston, USA, November 2012.

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

[3] L. Dodds and I. Davis. *Linked Data Patterns*. Online: `http://patterns.dataincubator.org/book`, (accessed: December, 2014), 2011.

[4] M. d'Aquin. Putting linked data to use in a large higher-education organisation. In C. Unger, P. Cimiano, V. Lopez, E. Motta, P. Buitelaar, and R. Cyganiak, editors, *Proceedings of the Interacting with Linked Data (ILD) workshop at the 9th Extended Semantic Web Conference (ESWC)*, volume 913, Heraklion, Greece, May 2012. CEUR-WS.

[5] F. Zablith, M. d'Aquin, S. Brown, and L. Green-Hughes. Consuming linked data within a large educational organization. In O. Hartig, A. Harth, and J. Sequeda, editors, *Second International Workshop on Consuming Linked Data (COLD) at 10th International Semantic Web Conference (ISWC 2011)*, volume 782, Bonn, Germany, October 2011. CEUR-WS.

[33] In general, write transactions are always sequential and we did not experience significant issues with the stability or efficiency of the live system during updates.

[34] `https://jena.apache.org/`

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