

# Consuming Linked Data within a Large Educational Organization

Fouad Zabalth<sup>1</sup>, Mathieu d'Aquin<sup>1</sup>, Stuart Brown<sup>2</sup> and Liam Green-Hughes<sup>3</sup>

<sup>1</sup> Knowledge Media Institute (KMi)

<sup>2</sup> Communications

<sup>3</sup> Information Technology

The Open University, Walton Hall, Milton Keynes, MK7 6AA, United Kingdom  
{f.zabalth, m.daquin, s.a.brown, l.e.green-hughes}@open.ac.uk

**Abstract.** Large universities tend to spread their services across several departments to serve their substantial student base. It is very common for this to result in developing different systems, which end up in creating many disconnected data silos within the organization. Data isolation is one of the main bottlenecks that prevent unlocking the full potential behind exploiting such data, to provide a better experience at the level of application deployment and data analysis. The Open University is in the process of connecting their data silos by relying on the Linked Data principles within the LUCERO project. We discuss in this paper three use-cases through which we consume Linked Data produced at the Open University: (1) a student services use-case showing how we exploit data connections to deliver learning material related to courses through the university's main course information website; (2) a mobile course application that enables students to easily explore courses by subject, qualification or research topic; and (3) a Leanback TV service that provides students the ability to watch, with a degree of control, a set of podcasts grouped in different channels. Through these use cases, we highlight in this paper the advantages and effects of consuming Linked Data within an organization.

## 1 Introduction

Today's organizations generate a significant amount of information as a result of their daily operational activities. The Open University is a large higher and further education institution in the UK which is dedicated to distance learning. It delivers courses and qualifications to more than 200,000 students across the UK, as well as in other countries. To realise this, the OU's information repositories are among its core assets. However, as it is the case of most large organizations, different systems tend to be developed throughout the years to perform specific tasks, which result in different disconnected data backbones. The absence of connections and links at the data level often makes data interchangeability between systems difficult and costly, which often leads to preventing the organization in fully exploiting their data.

Through the LUCERO project<sup>4</sup>, the OU has put in place a workflow to connect their various data repositories following the Linked Data principles [6, 7]. The current published data include for example course information (among others), with links to items internal to the OU, such as the library catalogue and other learning materials, as well as to external entities, such as countries in GeoNames<sup>5</sup> or other entities in DBpedia<sup>6</sup>.

We aim in this paper to give an overview of the LUCERO project, and the workflow to generate Linked Data (Section 3). Then we discuss the consumption of Linked Data that we illustrate through three use-cases: (1) the student services use-case (Section 4), where we show how Linked Data helped discovering related learning material to courses, and displaying them on mainstream courses' web pages of the OU; (2) the mobile course browser (Section 5), through which we show how links between academic qualifications, subjects and courses helped delivering a better course browsing application to students; and (3) the Leanback TV webapp (Section 6) which provides a Leanback TV experience by relying on video material from the OU's Linked Data set.

## 2 Related Work

In today's push towards creating semantics behind the data exposed over the web, we are witnessing an increase in the publishing side of Linked Data from educational bodies, following the original effort realised at the OU<sup>7</sup>. For example, the University of Southampton have made their data available as Linked Data<sup>8</sup>. There are also plans from the University of Oxford to do the same<sup>9</sup>, as well as the University of Lincoln<sup>10</sup>, among others.

While more and more data is being exposed (as can be inferred from the Linking Open Data cloud [1]), one aspect that has not yet been fully explored is the consumption side of Linked Data. We believe that the effect of Linked Data on the development effort, as well as on the potential applications that can be done with the data is a long term process. We argue that there is no one killer app to be discovered, but many small applications that will introduce substantial benefits at various levels to an organization and its data.

## 3 LUCERO Overview

The LUCERO project aims to transform and expose institutional repositories within the Open University as Linked Data. We present in this section the work-

---

<sup>4</sup> <http://lucero-project.info>

<sup>5</sup> <http://www.geonames.org>

<sup>6</sup> <http://www.dbpedia.org>

<sup>7</sup> <http://www.jisc.ac.uk/news/stories/2010/11/linkeddadata.aspx>

<sup>8</sup> <http://data.southampton.ac.uk/>

<sup>9</sup> <http://data.ox.ac.uk/>

<sup>10</sup> <http://data.lincoln.ac.uk/>

flow involved in generating the OU Linked Data, in addition to the data already available.

### 3.1 The Workflow for Generating Linked Data

As part of the LUCERO's main contributions, a workflow was put in place to handle the extraction and exposure of RDF elements from existing repositories. Figure 1 highlights the steps involved in this workflow, which we briefly describe in this section.

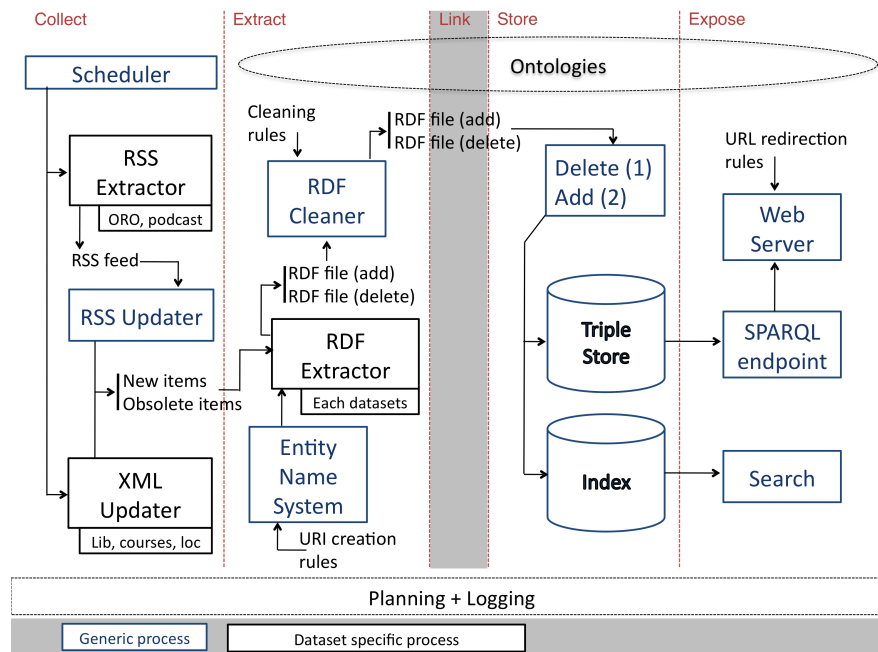


Fig. 1. The LUCERO workflow.

In the first step, we *Collect* the information that resides in the original repositories across the various departments of the Open University. We set up different extractors to handle the extraction of the required elements from the sources. For example, the RSS extractor is customized to process video or audio Podcasts<sup>11</sup> and publication data (i.e., Open Research Online<sup>12</sup>). The XML updater is used where the data is available in XML formal such as course descriptions. A scheduler handles an automatic daily process to check the changes that occurred at the data level, and run the workflow accordingly.

<sup>11</sup> <http://podcast.open.ac.uk>

<sup>12</sup> <http://oro.open.ac.uk>

Based on the identified items, we *Extract* the relevant RDF elements with the appropriate vocabulary to use. For example, in the case of a course item, we identify the need to model the course title, availability, cost, etc. In terms of the choice of vocabularies, our approach was to mostly reuse available ontologies, resulting in having in some cases more than one to apply. For example, AIISO<sup>13</sup> and Courseware<sup>14</sup> are both used, among others, to represent courses. An important component of the workflow is the *Entity Name System*, which ensures that each entity extracted gets a unique identifier across all our datasets, independently from its repository of origin. This guarantees for example that, when a book coming from the library catalogue dataset is part of course M366<sup>15</sup> and a podcast is also related to course M366, both the book and podcast are pointing to the same course with a unique identifier. To achieve this, we design our code to apply a set of specific patterns per entity type to follow predefined URI prefixes (e.g. “<http://data.open.ac.uk/course/>” for Courses, or “<http://data.open.ac.uk/library/>” for library material), coupled with a unique identifier originating from the source.

Once the RDF data to add (or delete) are identified, we store them in a triple store, and expose them through a web portal accessible at <http://data.open.ac.uk>. This portal provides a SPARQL endpoint, as well as information about the datasets available, sample queries and applications.

### 3.2 The Data Available

The OU has already extracted and made available Linked Data from various sources across the university. The process is continuous, with more data made available whenever new resources are made available. The current set of data that can be used include:

- **Courses Information:** In addition to the ability of getting the information through the OU website, courses information were also accessible in an XML format, which made the transformation into Linked Data easier. This dataset provides access to courses information including title, description, availability by date and location, prices along with related books and other material.
- **Podcasts:** This dataset provides access to the OU podcast items, with their corresponding title, related courses, subjects, video duration, etc. This set has been extracted from the OU podcast RSS feeds<sup>16</sup>.
- **Library Catalogue:** The OU library catalogue has been transformed from MARC records [2] into Linked Data. This dataset provides access to library items such as books, AV materials, reports among others. Such items were linked back to OU courses, and held information related to subjects, authors, publisher, Library of Congress Subject Headings [3], etc.

<sup>13</sup> <http://vocab.org/aiiso/schema>

<sup>14</sup> <http://courseware.rkbexplorer.com/ontologies/courseware>

<sup>15</sup> <http://data.open.ac.uk/page/course/m366.html>

<sup>16</sup> see for example <http://podcast.open.ac.uk/feeds/l314-spanish/rss2.xml>

- **Research Publications:** This dataset provides OU research publications with information about the authors, publication type (e.g., conference papers, theses, books, etc.), date of publication, among others.
- **OU YouTube Channel:** This dataset exposes the OU YouTube Channel<sup>17</sup> as Linked Data, with information about the title, tags, download URL and thumbnails.
- **OpenLearn:** The free online courses provided by the OU through its OpenLearn system<sup>18</sup> are also available now as Linked Data. This includes the title of OpenLearn units, topics, related courses, tags, among others.
- **Reading Experience Database:** Reading experiences from the Reading Experiences Database<sup>19</sup> are also available as Linked Data. This dataset contains information about readers, their gender, their reading objects and their experience details.
- **The Open Arts Archive Events:** Events from the Open Arts Archive website<sup>20</sup> are made accessible through data.open.ac.uk. This includes information about the events’ collaborators, artists, speakers and related podcasts among others.
- **KMi News and People:** These datasets present information about research staff<sup>21</sup> at the Knowledge Media Institute along with related news articles<sup>22</sup>.
- **Estates Buildings:** This dataset contains information about the OU buildings located across the UK. The buildings’ address, floor details and depiction are all accessible as Linked Data.

This wealth of data has been exploited in various scenarios (see for example [7] and [4]). A list of available applications is available online<sup>23</sup>. We present next our three use-cases through which we show the potentials of consuming Linked Data at the OU.

## 4 Student Services Use-Case

Students are among the main consumers of data at the Open University. Providing them with a good experience in reaching the required information with the least effort needed is key to the university’s success. The first use-case we present here is in the context of student services. Students are constantly looking for appropriate courses to enrol in at the Open University. Currently this is done through the main “Study at the OU” website<sup>24</sup>, where students can browse courses by subject, and get further information related to the courses and qualifications they are counted towards. The Open University produces a substantial

<sup>17</sup> <http://www.youtube.com/user/TheOpenUniversity>

<sup>18</sup> <http://openlearn.open.ac.uk>

<sup>19</sup> <http://www.open.ac.uk/Arts/RED/index.html>

<sup>20</sup> <http://www.openartsarchive.org>

<sup>21</sup> <http://kmi.open.ac.uk/people/>

<sup>22</sup> <http://kmi.open.ac.uk/news/>

<sup>23</sup> <http://data.open.ac.uk/applications/>

<sup>24</sup> <http://www3.open.ac.uk/study>

amount of courses related material, for example through the OU YouTube Channel or podcast website. Such material can be of great benefit for students to form a better understanding of what to expect from a course, and helping them make a decision about what course to take. However the problem is that traditionally, explicit links between courses and related material are not available. Moreover, modifying the underlying data infrastructure (e.g., databases) or application to expose such links can be very costly in terms of development time and resources.

With the availability of the Linked Data portal at the Open University through data.open.ac.uk where items and their corresponding links are explicitly defined, it became possible to easily get the entities related to specific courses. Having courses, podcasts and OU YouTube material (among others) in a coherent representation following well defined vocabularies, enabled a seamless extraction of needed entities through the available SPARQL endpoint. With this in place, we are in the process of extending the current courses' pages at the Open University, to offer students the functionality to explore learning material when visiting a course page. Figure 2 shows an example of related media, displayed on the right of the screen.

The image shows a screenshot of the Open University website for the course "The arts past and present". The page is divided into several sections: "About this course", "Course work includes", "Study explained", "Related media", and "Student reviews".

Annotations with blue arrows point to various elements:

- A box labeled "Related media" highlights the "Related media" section on the right side of the page.
- A box labeled "OpenLearn" highlights the "OpenLearn" link and the "try AA100 study materials" link.
- A box labeled "Podcasts" highlights the "AA100 podcast" link and the "[Audio - 10:26]" text.
- A box labeled "iTunesU" highlights the "AA100 introduction" link and the "[Video - 3:56]" text.
- A box labeled "More media..." highlights the "More media..." link at the bottom of the related media section.

The "About this course" section includes details such as Course code (AA100), Credits (60), OU Level (1), SCQF level (7), and RNEQ level (4). The "Course work includes" section lists 7 Tutor-marked assignments (TMAs) and End-of-module assessment. The "Study explained" section provides a breakdown of the course structure, including modules, credits, and assessment methods.

The "Related media" section lists the following items:

- OpenLearn: try AA100 study materials (Description can go here...)
- Podcasts: AA100 podcast (Description can go here... [Audio - 10:26])
- iTunesU: AA100 introduction (Description can go here... [Video - 3:56])

The "Student reviews" section includes a review from a user named "Being a lover of literature..." who mentions that the course is "greatly worth mandatory for those wishing to study the arts and humanities".

Fig. 2. Related media to the “AA100 – Arts Past and Present” course.

The formal “Study at the OU” site currently only describes Open University modules in text and yet there are other opportunities to demonstrate module content to prospective students through multimedia (podcast/iTunesU content and YouTube) or through the OU’s Open Educational Resources project OpenLearn. Before the metadata associated with this content was represented as Linked Data the inherent links between the OU’s OER, multimedia content and modules were not exploitable by machines. This meant that any attempt to supplement the OU’s online prospectus with related content represented an unrealistic requirement in terms of man hours, especially considering that such content quickly becomes out-of-date.

The data used through this application are the Podcast and iTunesU datasets, in addition to the OpenLearn and YouTube data. Data is provided to this application through SPARQL queries returning all multimedia / OpenLearn content related to a module is made available as XML, and will be incorporated into the XML files that draw together data from other institutional OU sources before being rendered as HTML.

## 5 Mobile Course Application

The Open University currently supports an iPhone and iPad application making content from its online prospectus available as an app. This application works well but there was a desire to improve it in two ways: (1) to make the application available across more platforms, and (2) to increase interactivity by making available material related to OU subject areas (e.g., Arts and Humanities, Maths and Computing, etc.) available on description pages as well as making more specific related content available at module description level (e.g., content related more specifically to “A101 - Introduction to Humanities”). Before LUCERO made these repositories available in a Linked Data format it was not possible for the OU to programmatically combine the data that creates the online prospectus with the content that it publishes in iTunesU, YouTube and OpenLearn. Although at a human level it was known that these resources related to modules (indeed frequently they are derived from module production process) this relationship was not expressed in a machine-readable way.

The newly-developed Study at the OU application is an HTML5 build made available in Android and iOS through thin client applications which serve to supplement a view onto the site through the devices browser (e.g., some of the navigation is handled through native application functionality whilst the content is a real-time web view). A similar approach is taken to the way the content will be surfaced via a Chrome add-on. Figure 3 shows a page through which students can explore courses and related content by qualifications, subjects, research, etc.

The data consumed in this app are from the Podcast, iTunesU, OpenLearn and YouTube datasets. The application reads (selected) data from the XML that is used to build the desktop online prospectus and combines this with data from data.open.ac.uk. When using the application a user can (whether they are at subject area level or module level) opt to view related resources. Podcast

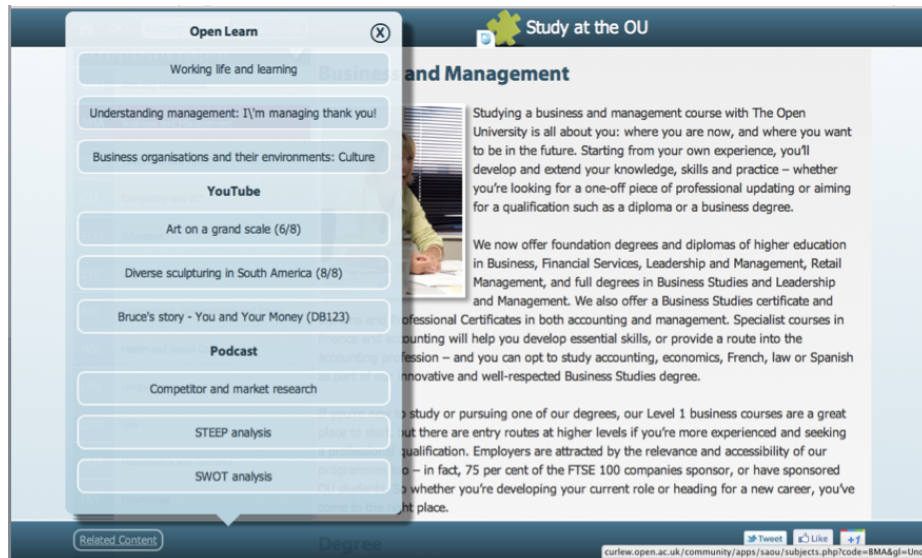


Fig. 3. Mobile course application.

and YouTube content are handled through the devices native functionality and links to related OpenLearn content (webpages) are made available through the browser. The content read from [data.open.ac.uk](http://data.open.ac.uk) is dynamic and up-to-date so the application shows only the latest 3 resources from each area. This serves to keep the pages relevant and fresh and also maintains a good level of usability.

## 6 Leanback TV Webapp

The OU has a plethora of video material spanning over various topics and subjects. With these hundreds of hours of interesting recordings, the question is: how do we engage the students in exploiting this vast resource of information, and exploring new areas of interests that might trigger new learning needs? With the current podcast portal, the student has to know what he/she is looking for in advance, before searching or going through the topics for the podcast of interest.

The purpose of Leanback TV applications is to provide viewers with a series of automatically selected set of video material that can be played one after the other, coupled with a degree of control over what to watch. We are witnessing an increase in Leanback TV services, provided for example by YouTube Leanback<sup>25</sup>, Redux TV<sup>26</sup> and Vimeo Couchmode<sup>27</sup>. In these applications, a set of videos are preselected for viewers, with the ability to skip or choose among the set of videos,

<sup>25</sup> <http://www.youtube.com/leanback>

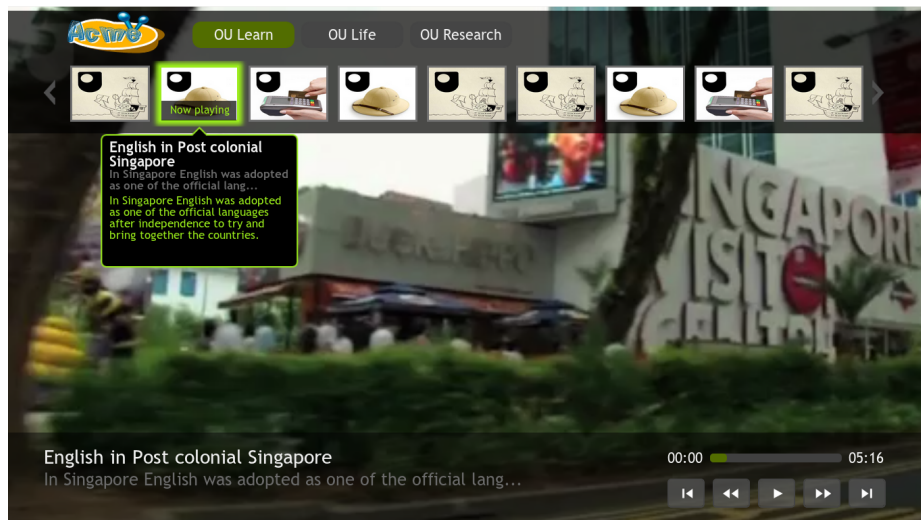
<sup>26</sup> <http://redux.com/tv>

<sup>27</sup> <http://vimeo.com/couchmode>



as well as choosing different channels. Such channels group the videos to watch into themes, such as YouTube's Trends and Best of YouTube.

In our context, we provide a Leanback TV experience over the set of learning podcast material. However, podcast information can be extracted from the current systems through its RSS feed, which doesn't provide the flexibility we needed to manipulate and extract the data. We rely on the Linked Data provided through [data.open.ac.uk](http://data.open.ac.uk), which we access through SPARQL queries to generate the set of videos to play through an HTML5 webapp interface. We used different queries to populate the channels with podcasts related to: OU research, OU learn and OU life. This is highlighted in Figure 4. Videos are also filtered based on their length, and only the ones that are below 10 minutes are shown to the users. Short videos have a higher chance to be watched by students, who know that something else will come up soon, in case such podcasts are of little interest to them. This will encourage students to explore new subjects, while relaxing in front of their TV or computer without having to choose what to watch next.



**Fig. 4.** Screenshot of the Leanback TV webapp.

With the presence of SPARQL, it is possible to create a fine grained selection of videos. For example, one can create a channel about videos related to a specific course, or a channel about the renaissance architecture. This is much harder to achieve through RSS feeds without further data customization at the source. Once the system is put in place, the options of creating new channels is endless, and very easy to extend. In the current demo<sup>28</sup>, the selected podcasts are mainly

<sup>28</sup> <http://labs.greenhughes.com/ldleanback/> – this application currently only works with the Google Chrome browser

of general interest, where very specialized videos such as one hour lectures on mathematics are not needed. However, in other contexts, this could exactly be what's needed for a mathematics course webpage. Such different channels can now be easily created through just creating a different SPARQL channel. This is a very useful for scenarios where university teachers can have custom made channels for their classrooms. For example, this is the SPARQL query used to generate the OU Life podcast channel:

```
PREFIX dc: <http://purl.org/dc/terms/>
PREFIX wm: <http://www.w3.org/TR/2010/WD-mediaont-10-20100608/>
PREFIX rdfs: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?thumb ?desc ?title ?source
WHERE {
?podcast <http://digitalbazaar.com/media/depiction> ?thumb .
?podcast dc:title ?title .
?podcast wm:description ?desc .
?podcast <http://digitalbazaar.com/media/download> ?source .
?podcast wm:createDate ?createDate .
?podcast dc:subject <http://data.open.ac.uk/topic/ou_life> .
?podcast rdfs:type <http://data.open.ac.uk/podcast/ontology/
VideoPodcast> .
}
ORDER BY DESC(?createDate)
LIMIT 25
```

Through this query, we are getting the *thumbnail*, *description*, *title* and *source* of the podcasts limited to the OU Life subject (i.e.,  $\langle http://data.open.ac.uk/topic/ou\_life \rangle$ ), ordered by their creation date. Figure 5 displays a snippet of the results that can be processed through data.open.ac.uk. The application is based on Google TV Resources<sup>29</sup>, and the full implementation details are available online<sup>30</sup>.

## 7 Discussion

Linked Data is transforming the way we access and consume data. From our three use-cases, we observe that the effects on the organizations and the way they consume and provide their data can be at the following levels:

**Application development effort:** Linked Data reduces the cost involved to pull data out of repositories. In all the three use-cases, data was retrieved from one common endpoint using common tools and standards. Once the developer masters the way to extract Linked Data, and apply them in his/her

<sup>29</sup> <http://code.google.com/p/gtv-resources/>

<sup>30</sup> <http://www.greenhughes.com/content/html5-leanback-tv-webapp-brings-sparql-living-room>

## Query Result (25)

thumb	desc	title	source
<a href="http://podcast.open.ac.uk/feeds/platform/juliette_00230_15052_std.jpg">http://podcast.open.ac.uk/feeds/platform/juliette_00230_15052_std.jpg</a>	"Young student talks about OU experience"	"Juliette part 5"	<a href="http://podcast.open.ac.uk/feeds/r/20110801T171938_5.m4v">http://podcast.open.ac.uk/feeds/r/20110801T171938_5.m4v</a>
<a href="http://podcast.open.ac.uk/feeds/platform/juliette_00230_15051_std.jpg">http://podcast.open.ac.uk/feeds/platform/juliette_00230_15051_std.jpg</a>	"Young student talks about her OU campus"	"Juliette part 4"	<a href="http://podcast.open.ac.uk/feeds/r/20110801T171755_4.m4v">http://podcast.open.ac.uk/feeds/r/20110801T171755_4.m4v</a>
<a href="http://podcast.open.ac.uk/feeds/platform/juliette_00230_15050_std.jpg">http://podcast.open.ac.uk/feeds/platform/juliette_00230_15050_std.jpg</a>	"Young student talks about OU experience"	"Juliette part 1"	<a href="http://podcast.open.ac.uk/feeds/r/20110801T170618_3.m4v">http://podcast.open.ac.uk/feeds/r/20110801T170618_3.m4v</a>

Fig. 5. SPARQL query results.

development environment, it will be very easy to apply the same methods in other applications. This will encourage developers to focus on the data usage, rather than on the technical difficulties involved. In the case of the three applications described in this paper, the development has been performed without having to go back to the original data maintainers. This, by itself, is considered a substantial decrease in the cost involved within the organization.

**User experience:** We observe that Linked Data enabled the development of applications that will enhance user experience. This observation surely needs to be backed by some kind of survey once the applications are fully deployed and accessed by a substantial amount of users. However we foresee that, by for example simply aiding the users by decreasing the required effort in finding relevant material while looking for courses, or interesting podcasts, our applications will be positively embraced by our end-users.

**Connecting to others:** In this paper, we have mostly focused on the benefit of connecting data silos internal to the organization. However, we also provide links to external datasets, including for example to the countries in which courses are available in GeoNames, to the description of the Open University in <http://education.data.gov.uk> and to the postcode units of OU buildings in <http://data.ordnancesurvey.co.uk>. Providing such external links not only supports the enrichment of our own data, but also provides users with additional entry points into it, which are not dictated by the university's view on the provided information. We are for example investigating scenarios in which resources at the OU could be linked to TV programs from the BBC, providing ways to discover study material relating to the topics treated in a TV program.

## 8 Conclusion

In this paper, we have presented three use-cases that demonstrate how Linked Data can be consumed in a large organization such as the Open University. The goal was, through these examples, to discuss the benefits, issues and effects of integrating Linked Data practices in such an organization. Indeed, based on our experience in building the Linked Data platform of the OU and supporting other departments in consuming such data, we believe that the success of Linked Data will not be based on the development of “killer apps”, but on solving in a cost effective way numerous problems that organizations, data managers and developers are facing. We also wanted to emphasize the importance of organizational aspects in the adoption of Linked Data in the context of a large organization. We selected the three presented use cases as they represent clear examples where Linked Data is consumed in projects led by non-academic services of the University (communications/online services, student services and IT), being clearly driven by the “business needs” of the organization, rather than by the application of a particular technological solution.

As we mentioned at the beginning of the paper, many other universities are now following the same approach as we have developed within the LUCERO project, to expose public university resources as Linked Data. While this trend is still at an early stage, interesting cases and ways of consuming linked data are expected to emerge from it. Being able to discover educational resources related to a topic or a course independently from the way and place in which they were produced seems an obvious scenario. It however generates new challenges related to the aggregation of data from different institutions<sup>31</sup>, as well as to the availability of common classification schemes for these resources [5].

## References

1. C. Bizer, T. Heath, and T. Berners-Lee. Linked data-the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009.
2. D. J. Byrne. *MARC Manual: Understanding and Using Marc Records*. Libraries Unlimited, December 1997.
3. L. M Chan. *Library of Congress subject headings: principles and application*. Libraries Unlimited, Inc., PO Box 6633, Englewood, CO 80155-6633, clothbound: ISBN-1-56308-195-4., 1995.
4. M. d’Aquin, F. Zablith, and E. Motta. wayOU – Linked Data-Based social location tracking in a large, distributed organisation. In *Proceedings of the 8th Extended Semantic Web Conference (Demo)*, page 461465, 2011.
5. M. Fernandez, M. d’Aquin, and E. Motta. Linking data across universities: an integrated video lectures dataset. In *International Semantic Web Conference, ISWC, In-Use track*, 2011.
6. T. Heath and C. Bizer. *Linked Data: Evolving the Web into a Global Data Space*. 2011.
7. F. Zablith, M. Fernandez, and M. Rowe. The OU linked open data: production and consumption. In *Proceedings of the 1st International Workshop on eLearning Approaches for the Linked Data Age (Linked Learning) at ESWC*, Heraklion, Greece, 2011.

---

<sup>31</sup> See <http://linkeduniversities.org>