# Assembling and Applying an Education Graph based on Learning Resources in Universities

Tom Heath
Talis Education Ltd
43 Temple Row, Birmingham
B2 5LS, UK
tom.heath@talis.com

Ross Singer
Talis Education Ltd
43 Temple Row, Birmingham
B2 5LS, UK
ross.singer@talis.com

Nadeem Shabir
Talis Education Ltd
43 Temple Row, Birmingham
B2 5LS, UK
nadeem.shabir@talis.com

Chris Clarke
Talis Education Ltd
43 Temple Row, Birmingham
B2 5LS, UK
chris.clarke@talis.com

Justin Leavesley
Talis Education Ltd
43 Temple Row, Birmingham
B2 5LS, UK
justin.leavesley@talis.com

## ABSTRACT

This paper introduces the notion of the education graph, a conceptual representation of the resources and interconnections at the heart of the learning process. We present our latest work on the *Talis Aspire* family of products that, through the use of Linked Data principles and technologies, enables the assembly and application of a rich education graph based on learning resources used in tens of UK universities. Techniques for entity extraction and reconciliation across data sources are presented, in addition to descriptions of recommendation generation from portions of this education graph.

## Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

## Keywords

education, linked data, semantic web, learning, teaching, education graph

## 1. INTRODUCTION

Recent years have seen explosive growth in the so-called 'Web of Data', exemplified by initiatives such as the Linking Open Data project [2]. This grassroots effort has stimulated the publication on the Web of many billions of data points in Linked Data form, covering countless topical domains from pharmaceuticals to films [7]. The result is a huge, cross-domain *data graph*, upon which new forms of applications and interactions may be delivered to users.

Despite the breadth and diversity of this graph, and the availability of some generic applications upon it (e.g. [16], [1]), we believe the majority of user-facing applications built on this Web of Data, or on Linked Data in general, will be domain specific. Furthermore, as machine-readable data becomes ever more readily available, and ever greater in volume, the need for trustworthy, authoritative data of verifiable provenance becomes increasingly significant [6].

Against this backdrop, and as we describe in Section 2, a number of universities have begun to make portions of their data available on the Web in Linked Data form. These initiatives, as with those of the BBC [9] and New York Times[1], are particularly noteworthy because the data publication happens *at source*, rather than via a third party republishing openly licensed data. While the diversity of these data sets is impressive, with few exceptions the resulting Linked Data does not capture the heart of what we call the *education graph* – a Linked Data representation of the learning resources (e.g. textbooks, papers, videos) and connections (e.g. between student and lecturer, module, topic, *etc.*) that lie at the heart of the learning process. It is by harnessing the connections in this learning-oriented portion of the graph that, we believe, Linked Data can fundamentally enhance the educational process.

In this paper we present our latest work on the suite of applications in the *Talis Aspire* family[2]. Specific contributions of the paper include: 1) the presentation of a Linked Data application deployed in the higher education sector and in use by thousands of university students daily; 2) the description of how the data created by these deployments is assembled into a broader education graph (providing a unique insight into the resources to teach at universities in the UK), including methods for entity extraction and reconciliation from citation data; 3) the description of methods used to generate recommendations from this education graph.

In addition to these modest research contributions, we believe the work presented here represents a major contribution as perhaps the most comprehensive, mature, and widely used application of Linked Data in the education sector.

The remainder of this paper is structured as follows. After reviewing related work in more detail, we describe one

*Linked Learning 2012*, 17 April 2012, Lyon, France.


[1]`http://data.nytimes.com/`
[2]`http://www.talisaspire.com/`

of the applications in the *Talis Aspire* family, and detail the mechanisms used to build a substantial graph describing the resources used to teach higher education courses, from deployments of this application at a significant proportion of UK universities. In the latter sections of the paper we describe how we use this graph to develop novel applications for customers, and close by discussing avenues for ongoing and future research.

## 2. RELATED WORK

Perhaps the most widely cited graph-based application at present is *Facebook*[3], which has become synonymous with the notion of a 'social graph', where interpersonal connections are used to enable or enhance online interactions between friends and acquaintances. Similar ideas have been applied to the field of research, through sites such as *IAmResearcher*[4] and *Google Scholar*[5] which capture and utilize connections between researchers, such as co-authorship relationships.

In the field of learning and teaching specifically, the desire to discover and reuse educational resources (or 'learning objects') has produced a substantial body of work, and technical standards, devoted to describing these resources in the form of metadata records. The background to this trend, plus relevant standards (e.g. *Dublin Core*, *IEEE LOM*, *SCORM*), have been described in detail elsewhere [4][11][14].

More recently, local initiatives [12] and broader collaborations (e.g. *Linked Universities*[6]) have become a focal point and stimulus for higher education institutions publishing their data online according to Linked Data principles. The resulting data sets are broad in nature, ranging from courses to learning resources, bus stops and buildings, and are complemented by publication of cross-institutional datasets describing video lectures [5].

As introduced in Section 1, our approach differs from the majority of other work in the field by placing its primary emphasis on how the resource is situated within a specific learning context, such as its use on a particular course as recommended by a particular tutor. It is this rich collection of resources and interconnections, coupled with organising structures such as topic hierarchies, that we view as a prototypical *education graph*. In particular, we believe that the connections in such a graph, by virtue of being created by practising lecturers in the field, provide valuable indicators of, for example, the relevance and quality of learning resources. The next section introduces the *Talis Aspire Campus Edition* application from which this education graph is compiled. In later sections we discuss how this graph is applied to support additional applications.

## 3. TALIS ASPIRE CAMPUS EDITION

In a previous paper [3] we introduced the *Talis Aspire* resource list management application, now known as *Talis Aspire Campus Edition*. This *software-as-a-service (SaaS)* application enables university lecturers to create and manage lists of learning resources associated with a particular module/course, and share these with students studying those modules/courses. The benefits for students include a single point of access for multiple lists, and integration of the lists with library catalogues and institutional systems for electronic journal access.

From a technical perspective, *Campus Edition* is a native RDF application implemented in PHP and backed by a SaaS triplestore. The application exposes all lists and related information in RDF according to the Linked Data principles; where customers have chosen to do so, this data is publicly available on the Web[7]. Further implementation details are given in [3].

At present, *Campus Edition* is deployed to around 30 institutions in the UK and beyond; this translates to a market penetration of around one quarter of UK universities and a potential audience of several hundred thousand university students.

## 4. THE 'EDUCATION GRAPH'

Each deployment of *Talis Aspire Campus Edition* results in the creation of a significant data set representing courses and/or modules available at that institution and the resources used to teach them. The data that makes up each of these *institutional graphs* remains in the ownership of the university, but is used in aggregate form to create a broader education graph, upon which we have developed additional services that can further enhance and complement the *Campus Edition* application (see Section 6).

## 5. ASSEMBLING THE EDUCATION GRAPH

In this section we will detail some of the processes involved in building the education graph from lists of learning resources created in *Talis Aspire Campus Edition*. Broadly speaking, this is a two-stage process: the first consists of creating the institutional graph of resources and courses; the second of identifying and materialising points of intersection between these graphs to produce a coherent, interconnected whole.

### 5.1 The Institutional Graph

The institutional graph has a number of components, such as: 1) the hierarchy of organisational units that offer courses (i.e. faculties, departments, schools), 2) the directory of modules/courses offered, and 3) the list of resources associated with each course. Items 1) and 2) are typically imported automatically into the *Campus Edition* system using data provided by the institution[8], while the latter may be created by teaching staff or librarians through the Web interface, or imported from legacy systems.

In both the latter cases, recognition or provision of an identifier (e.g. ISBN, DOI) enables retrieval of data about the resource from other sources (such as an institutional library catalogue) in order to construct a resource description of acceptable quality. In other cases, such as where legacy resource lists have been imported, a degree of data cleaning may be required. Specifically, legacy data sources often refer

[3]http://www.facebook.com/

[4]http://www.iamresearcher.com/

[5]http://scholar.google.co.uk/

[6]http://linkeduniversities.org/lu/

[7]Example: http://www.readinglists.manchester.ac.uk/lists/40EB7382-C58E-9DF5-FE36-D6CD64ABDA83.rdf

[8]We note that if more institutions adopted the approaches described in Section 2 this data could be sourced from the institutional Web site directly, with no manual intervention required by university staff to prepare this for import.

to resources using bibliography-style citations in plain text, with no consistent style or additional structure. These references must be converted to structured data for integration into the institutional education graph. For such cases we have developed a comprehensive citation extraction framework.

#### 5.1.1 Extracting Bibliographic Data from Plain Text Citations

We conceive of the citation extraction process as converting a textual citation to a small graph connecting authors to works, publication venues (journals, conferences, etc.) and publishers. These fragments naturally form leaves on the 'resource list tree' but also create the potential for *cross-list* clusters of resources that, for example, share the same publisher or topic.

This first stage of lifting plain text citations into graph form is achieved in the following way:

1. Citation strings are pre-processed using regular expressions to resolve simple syntactic quirks that cause issues in the next stage.

2. The cleaned strings are passed through a version of FreeCite[9] (itself based on the CRF++ library[10]) that has been heavily modified for our needs[11].

3. The output from this Web service is cleaned again using regular expressions and returned as a JSON object.

Having obtained a clean, structured representation of a bibliographic record from plain text data, it is then passed through an entity reconciliation process. By taking advantage of data characteristics specific to the scholarly domain we believe we can implement a more effective approach than domain-neutral data linking frameworks, such as *Silk* [17].

#### 5.1.2 Entity Reconciliation across Data Sources

The goal of this process is to match the record against high-quality reference data sources, thereby validating (or not) the accuracy of the record. These data sources are as follows:

- **Open Library**, the openly editable and openly licensed library catalogue, stored in a *Talis Platform*[12] hosted RDF store.

- **OpenKB**, a data set describing serials/journals, assembled in-house from sources such as *CUFTS*[13], *National Library of Medicine*[14] and *SAO/NASA Astrophysics Data System*[15], and stored in a triplestore based on *Apache Jena TDB/Fuseki*[16]. It is our intention to supplement these with additional sources, such as *PubMed*[17] and the *British National Bibliography*[18].

- **CrossRef**[19], a data set of tens of millions of scholarly articles, stored in a MySQL database.

Records that describe books are compared against the Open Library data, where we attempt to match on a precise edition of a work. Scholarly articles are matched first against the OpenKB data set to retrieve additional data such as an ISSN of the periodical in which it appears. The enriched record is then used to search CrossRef for a fuller citation.

Through this process we build up an initial graph-like description of the resource, which in turn is used to reconcile the item description with those of other instances of the same *work* or *expression of a work* (to borrow *FRBR* terminology [15]) that already exist within *Talis Aspire Campus Edition*. Addressing this issue at an institutional level is, in fact, just one instance of a broader challenge that manifests when we unify multiple graphs. Therefore the same approach, described in Section 5.2, is used in both cases.

### 5.2 Unifying the Institutional Graphs

To enable reconciliation of resources at an institutional and super-institutional level, the system has a notion of *canonical resources*, to which instances of that work or expression on different lists can be mapped. This is achieved (at both institional level and above) by querying the existing data for canonical resources that share a unique identifier, such as an ISBN or DOI. On this basis, in a process akin to reasoning using *inverse functional properties* in OWL, it is possible to infer that two resources are in fact the same expression (of the same work). Where an identifier (and associated resource description) is not already present in the system, we use external data sources (such as Open Library) which maintain similar records relating works to the identifiers associated with their various manifestations.

Before the entity reconciliation process described above can take place, resource descriptions must be retrieved from the stores underlying each instance of *Talis Aspire Campus Edition*. This is achieved via the *OAI-PMH*[20] service provided by the Talis Platform, which returns lists of new or updated records to be passed to the entity reconciliation workflow. Once each resource has been mapped to a corresponding canonical resource in the unified data set, pointers are added back to its institution-level representation, thereby allowing all cross-institutional descriptions of a particular resource to be retrieved from the graph.

## 6. APPLYING THE EDUCATION GRAPH

Having assembled a broader education graph from the set of institutional graphs, we apply this data set to a number of additional application domains: 1) a directory of high quality learning and teaching resources, 2) recommendations of related lists and resources. These applications are described in the sections below.

### 6.1 Talis Aspire Community Edition

*Talis Aspire Community Edition*[21] is a publicly accessible view of portions of the broader education graph. It serves primarily as a searchable directory of high-quality learning and teaching resources, organised by discipline. Both

[9] http://freecite.library.brown.edu/
[10] http://crfpp.sourceforge.net/
[11] https://github.com/rsinger/free_cite
[12] http://www.talis.com/platform/
[13] http://researcher.sfu.ca/cufts
[14] http://www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.pubmedhelptable45/
[15] http://adsabs.harvard.edu/abs_doc/journals.html
[16] http://incubator.apache.org/jena/
[17] http://www.ncbi.nlm.nih.gov/pubmed/
[18] http://www.bl.uk/bibliographic/datafree.html
[19] http://www.crossref.org/
[20] http://www.openarchives.org/OAI/openarchivesprotocol.html
[21] http://community.talisaspire.com/

the discipline-centric and resource-centric pages also link to resource lists (from participating universities) in that discipline or that include the particular resource, respectively. By surfacing the list–resource and list–institution relationships, the system provides indicators of provenance and therefore authority, while the compilation of the directory from lists of resources hand-picked by academic staff at universities ensures a quality baseline is achieved.

An additional feature of *Talis Aspire Community Edition* is the suggestion of recommended resources and related courses, as described below.

## 6.2 Recommendations from the Education Graph

As would be expected, significant patterns are manifest in the assembled education graph, such which resources typically feature on particular types of courses and those resources which typically co-occur on the same lists. As originally envisioned in [13], we use this data and these patterns to generate two forms of recommendations, resource-to-resource and list-to-list, that are then surfaced in various ways within the *Talis Aspire* family of products.

### 6.2.1 Resource-to-Resource Recommendations

In the style of [10], we generate item-to-item recommendations based on item co-occurrence on resource lists. Specifically, each resource is associated with a set of all lists on which it occurs. The similarity between two resources is defined by the *Jaccard* coefficient [8] of their corresponding list sets, weighted by list length – i.e. co-occurrence of resources on longer lists is given less weight then co-occurrence on short lists.

Recommendations generated in this fashion are surfaced in two ways: directly on resource pages in *Talis Aspire Community Edition* (see Figure 6.2.1) and in *Talis Aspire Campus Edition* via a *Community Edition* widget. One interface design challenge we have faced is in how to communicate the source of these recommendations to users. Specifically, in a system based on lists selected by authoritative sources (i.e. university lecturers), the label "recommended resources" can be misleading and an alternative must be chosen carefully.

### 6.2.2 List-to-List Recommendations

Using the same approach as for resource-to-resource recommendations we also generate list-to-list recommendations. In this case, the Jaccard coefficient of similarity between two lists is computed based on the items on those lists. List-to-list recommendations are exposed through a 'related courses' feature in *Talis Aspire Community Edition*, as shown in Figure 6.2.2.

### 6.2.3 Recommendation Implementation

Given that our item-to-item and resource-to-resource recommendations are, in effect, the mirror image of each other, both can in principle be computed from the same input matrix. In practice, the relevant portion of the matrix is selected and keyed off the list or resource URI, as appropriate, and passed as input to the first of multiple, sequential *map-reduce* jobs run on *Apache Hadoop MapReduce*[22]. As part of the recommendation generation process, we prune sparse portions of the matrix (of which there are many), to reduce the overall number of comparisons.

[22] http://hadoop.apache.org/

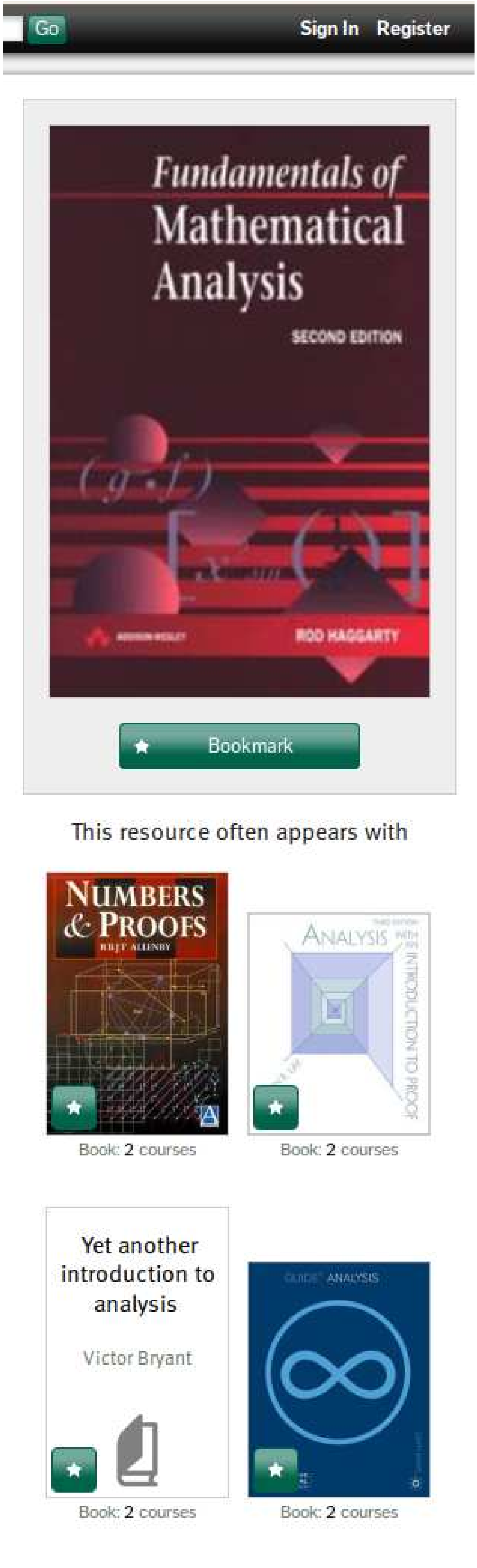


**Figure 1: Related resource recommendations in *Talis Aspire Community Edition*.**

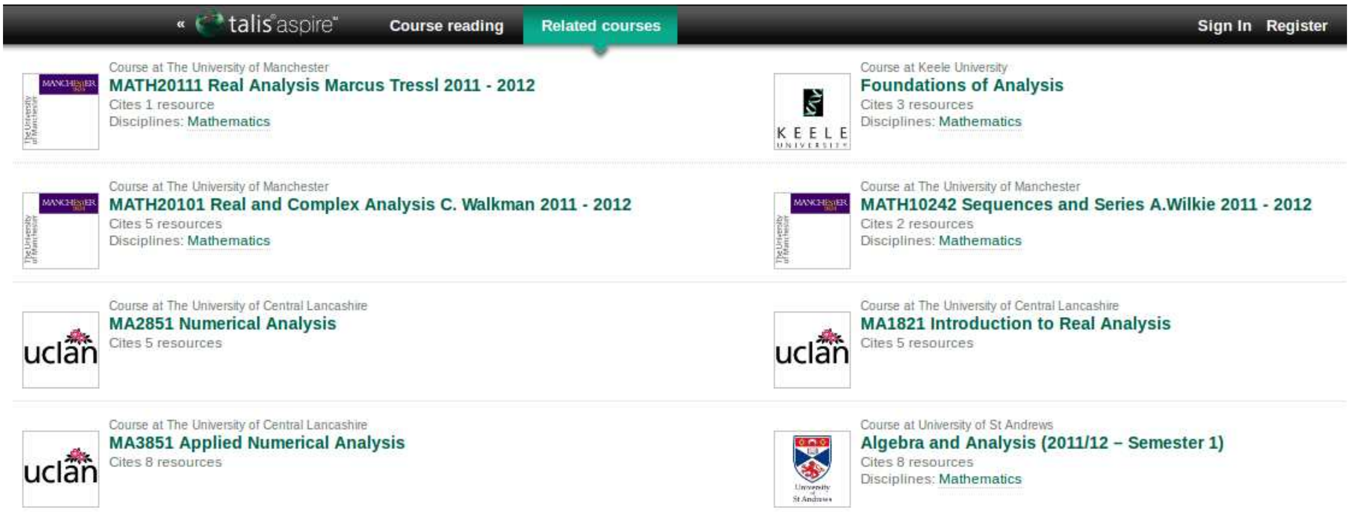


**Figure 2: Related course recommendations in *Talis Aspire Community Edition*.**

It is also pertinent to mention that when discussing resources/items in the context of recommendations, we are referring to works rather than expressions of works. This approach was a direct response to our finding that different expressions of the same work were being recommended, due to their co-occurrence with the same items on different lists. Different expressions are mapped to the same work using the approach described in Section 5.2.

## 7. DISCUSSION AND FUTURE WORK

In this paper we have presented the latest developments related to *Talis Aspire*, a family of products that enable the creation of institutional graphs describing the resources, and resource lists, at the heart of the learning and teaching process. The foundation of the *Talis Aspire* product family on Linked Data principles and technologies enables the assembly of these institutional graphs into a broader education graph, using the entity extraction and reconciliation processes described above. This assembled education graph is able to deliver richer insights that are not feasible with data from one university alone. Two examples of such have been presented in this paper: resource and list recommendations, and the *Community Edition* directory of learning and teaching resources, both of which reveal sector-level patterns in the usage of learning and teaching resources.

In ongoing research we are exploring the potential for additional services based on this education graph that can further enhance the *Campus Edition* product for customers, with the aim of helping universities fundamentally enhance the learning experience for students. Underpinning this strand of research is a new data warehousing infrastructure that enables us to easily gain insight into, and investigate novel applications of, the education graph. Again, the use of Linked Data principles and technologies, particularly the inherently flexible RDF data model, has made the construction of this data warehouse rather trivial.

From a technical perspective, our ongoing research also involves development of a high-performance triplestore based on *MongoDB*[23], *without* support for ad-hoc SPARQL queries. The rationale for this approach stems from our experience of developing applications that use a limited number of 'canned' queries, each of which is of a predictable form. This brings two benefits: firstly, *a priori* knowledge of queries allows pre-computation of all possible result sets for a given data set; secondly, the underlying infrastructure does not need to be capable of answering arbitrary queries, which can bring performance benefits such as reduced indexing requirements at load/update time. As a further benefit, the availability of native map-reduce capabilities in MongoDB can reduce the complexity inherent in exporting, transferring, and pre-processing RDF data for use within e.g. Hadoop MapReduce.

An additional insight from our experiences of developing graph-based applications is as follows: the internal representation and storage mechanisms for graph-centric data within applications is of rather less significance than the ability to process, transform, and merge graphs based on a common data model. It is at this inter-application layer that RDF and Linked Data can bring greatest benefits, but this says nothing about how the data underlying an application should be stored and queried.

In future research we would like to revisit the existing applications of the education graph, to more deeply understand their impact on users. For example, it is our goal to formally evaluate the quality of our current recommendations and investigate potential enhancements to the algorithms adopted and developed in the first iteration of the system. Specific issues worthy of investigation include the value (or otherwise) of weighting based on list length, and whether an additional *list position* function we employ could be enhanced by exploiting list section structure and section sequences into the algorithm. In addition, a deeper understanding of the various functions that can be served by these recommendations, e.g. assisting learners with resource discovery vs. assisting teachers with list creation, would be of great benefit in informing future development of the products.

[23] `http://www.mongodb.org/`


## 8. ACKNOWLEDGMENTS

We would like to acknowledge all customers of *Talis As-*

*pire Campus Edition*[24] for their contributions to shaping the product and to the broader education graph described in this paper.

## 9. REFERENCES


[1] C. Becker and C. Bizer. Exploring the geospatial semantic web with dbpedia mobile. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(4):278 – 286, 2009.

[2] 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.

[3] C. Clarke. A resource list management tool for undergraduate students based on linked open data principles. In *Proceedings of the 6th European Semantic Web Conference*, Heraklion, Greece, 2009.

[4] S. Downes. Learning objects: Resources for distance education worldwide. *The International Review of Research in Open and Distance Learning*, 2(1), 2001.

[5] M. Fernandez, M. d'Aquin, and E. Motta. Linking data across universities: An integrated video lectures dataset. In *10th International Semantic Web Conference (ISWC 2011)*, 2011.

[6] T. Heath. Linked data – welcome to the data network. *IEEE Intelligent Systems*, 15(6):70–73, 2011.

[7] T. Heath and C. Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool, 1st edition, 2011.

[8] P. Jaccard. The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50, 1912.

[9] G. Kobilarov, T. Scott, Y. Raimond, S. Oliver, C. Sizemore, M. Smethurst, C. Bizer, and R. Lee. Media meets semantic web — how the bbc uses dbpedia and linked data to make connections. In *Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications*, ESWC 2009 Heraklion, pages 723–737, Berlin, Heidelberg, 2009. Springer-Verlag.

[10] G. Linden, B. Smith, and J. York. Amazon.com recommendations: item-to-item collaborative filtering. *Internet Computing, IEEE*, 7(1):76–80, Jan. 2003.

[11] M. McClelland. Metadata standards for educational resources. *Computer*, 36:107–109, November 2003.

[12] M. Rowe. Data.dcs: Converting legacy data into linked data. In C. Bizer, T. Heath, T. Berners-Lee, and M. Hausenblas, editors, *LDOW*, volume 628 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2010.

[13] N. Shabir and C. Clarke. Using linked data as a basis for a learning resource recommendation system. In *1st International Workshop on Semantic Web Applications for Learning and Teaching Support in Higher Education (SemHE'09)*, September 2009.

[14] S. A. Sutton. Conceptual design and deployment of a metadata framework for educational resources on the internet. *J. Am. Soc. Inf. Sci.*, 50:1182–1192, November 1999.

[15] B. Tillett. What is frbr? a conceptual model for the bibliographic universe. *Technicalities*, 25(5), 2003.

[16] G. Tummarello, R. Cyganiak, M. Catasta, S. Danielczyk, R. Delbru, and S. Decker. Sig. ma: live views on the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 8(4):355–364, 2010.

[17] J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov. Discovering and maintaining links on the web of data. In *Proceedings of the 8th International Semantic Web Conference*, ISWC '09, pages 650–665, Berlin, Heidelberg, 2009. Springer-Verlag.


[24] `http://campus.talisaspire.com/our-customers/`