Building Linked Open University Data - Tsinghua University Open Data as a showcase

Yuanchao Ma, Bin Xu, Yin Bai, and Zonghui Li

Knowledge Engineering Group, Department of Computer Science and Technology, Tsinghua University, Beijing, China
{myc,xubin,baiy,lzh}@keg.cs.tsinghua.edu.cn
http://keg.cs.tsinghua.edu.cn

Abstract. Linked Open University Data applies semantic web and linked data technology to university data scenario, aiming at building interlinked semantic data around university information, providing possibility for unified inner- and inter- school information query and comparison. This paper proposes a general process of building linked open university data, with procedures covering choosing datasets and vocabularies, collecting and processing data, building RDF and interlink, etc. Tsinghua University Open Data is used to demonstrate the process. Tsinghua University consist of 5 well-formed, interconnected datasets, with a number of interesting applications has been built on top of them. Finally, remarkable points about data collecting and processing is discussed.

Keywords: Open Data, Linked Data, Linked Open University Data, SPARQL server, Semantic Portal

1 Introduction

As tools and standards related to the Semantic Web are becoming comprehensive and stable, how to build high-quality semantic data and how to make use of these semantic data, become two major challenges in the development of the Semantic Web. *Linked Data*[1] project describes a well-acknowledged standard method of publishing interlinked structured data based on the Semantic Web technology[2]. The *LOD*[2] project now has published hundreds of datasets using Linked Data standard, covering a large variety of domains. [4]

Linked Open University Data builds linked open data around universities and academic institutions. Linked open university data under a unified schema provides consistent data access to different universities, offering convenience for inner-university and inter-university information management.

However, building linked open university data faces some major challenges. First of all, there is no unified, well accepted vocabulary for describing university-related information. What is more, as existing well-structured data about university which is publicly accessible is very limited, collecting and organizing raw

[1] http://linkeddata.org
[2] http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

data needs extra work. Establishing interlink to connect different datasets together is also great challenge, considering complex relationship between different pieces of university data.

In this paper, we present a general process of building linked open university data. We first list common data categories about university information, then give a summary of useful vocabularies for describing university datasets. We then discuss common approaches to retrieve raw data about university, and the preprocessing and cleaning work necessary to get these raw data ready for converting, followed by naming and linking strategies to make them correspond with Linked Data standard. Then we describe techniques and tools used to convert well-organized structured data of different formats to RDF format.

While describing the general process, *Tsinghua University Open Data* (`http://data.cs.tsinghua.edu.cn/`) is demonstrated as a showcase. Unlike other open datasets, our datasets are mostly built upon data crawled from the public university website. We currently have 5 core datasets, with downloadable RDF/XML descriptions and a standard SPARQL endpoint. There is also a HTML web interface available for SPARQL query and data browsing. We have also built some applications by ourselves, including CampusAssistant, CourseFinder, etc.

Processes and techniques described in this paper serves well in building Tsinghua University Open Data, and we believe that it is helpful when building open data for other institutions, and for publishing general linked data as well.

The rest of this paper is organized as follows: In Section 2, we discuss common categories of university data, and the vocabularies used to describe these data. In Section 3, we discuss different approaches of getting raw data, and necessary procedures to clean and normalize these data, as well as converting the processed data to RDF. A detailed description of Tsinghua University Open Data is given in Section 4. Then we give a brief introduction about related work in Section 5. Finally, we discuss key points and difficulties in building linked open university data in Section 6 and conclude the paper in Section 7.

2 Data categories and Schema

2.1 Data Categories

Modern universities serve as educational institutions, academic research institutions, as well as living communities of their students and staff members. As a consequence, there are so many aspects of university data. However, several university datasets attract most attentions of people and act as core data connecting different aspects of university information, thus are of great importance in building linked open university data.

We categorize the most important university information into the following classes:

University Basic Information General properties about the institution. These mainly consist of founding year, motto, location, etc. Basic information also

includes organization structure of the institution. Organizations - schools and departments - description has organization name, size, superior organization information, location of department building, etc.

Campus Information Campus information contains geographic description of the campus, as well as building information within and around the campus.

Educational Administrative Information As facility for education, educational administrative information is important to not only students within the university, but also to researchers and students outside campus. Courses, exam schedule are examples of datasets of this category.

Faculty Information Describes staff members of the university, mostly teachers and researchers, their basic information, contact information, research related information, etc.

2.2 Vocabularies

On the Semantic Web, *vocabularies* (or *ontologies*, more strictly), define the concepts and relationships (also referred to as "terms") used to describe and represent an area of concern. Vocabularies are used to classify the terms that can be used in a particular application, characterize possible relationships, and define possible constraints on using those terms[3]. A standard set of vocabularies not only provides unified access to data consumers, but also acts as an important role in data mushup and inference.

There is a number of vocabularies that can be used in describing university information. We have summarized a set of vocabularies for university datasets, mainly considering popularity, comprehensiveness and quality. Most of the vocabularies we use are strictly defined and public accessible.

FOAF[4] (Friend Of A Friend) is a universally acknowledged vocabulary devoted to describe information about people and their relationships. FOAF is so widely used that it appears in nearly every dataset of our site.

AIISO[5] The Academic Institution Internal Structure Ontology (AIISO) is designed to describe the internal organizational structure of an academic institution. In our datasets, we use *aiiso:Course* to describe university course information, with is used along with FOAF and Org, and newly defined extended properties to link courses with organizations and staff members.

Org[6] is a widely used vocabulary devoted to describe information about organization and their relationships.ORG is the main vocabulary used in the *Organization* dataset of Tsinghua University Open Data, with each work unit as a *org:Organization*. It describes basic relationship informations such as subOrgnizationOf, hasSubOrganization, hasUnit etc.

CourseWare[7] is developed for describing courses and resources within the *ReSIST*[8] project. This vocabulary represents various information about a course, including material, pre-requirement, language, etc.

[3] `http://www.w3.org/standards/semanticweb/ontology`

[8] `http://www.resist-noe.org/`

As for those datasets which there is no suitable existing vocabulary defined, we have defined new terms, and linked our own terms to the existing vocabularies as supplements. The main vocabularies used are listed as follows.

We use OpenVocab to create our own terms. *OpenVocab*[9] is a community maintained vocabulary intended for use on the Semantic Web, ideal for properties and classes that do not warrant the effort of creating or maintaining a full schema. OpenVocab allows anyone to create and modify vocabulary terms using their web browser. We defined several properties to construct interlink between different datasets. For example, ov:deliveredBy is used for linking courses with lecturers; ov:offeredBy is used for linking courses with organizations.

3 Data Collection and Structuring

3.1 Data Collection

Considering the variety among different universities, different approaches may apply to collect raw structured data for building linked open university data. We list the most commonly used methods as follows:

Some raw data can be retrieved from the university administration facility, which are allowed to be published under certain license. These data are usually well structured, high quality with few or no noise. However, this approach need firm cooperation between data publisher and university authority, which is sometimes difficult to achieve. What's more, many kinds of data do not have structured backends.

The university publishes some of its information on the web, either as downloadable structured format or HTTP queryable. We can downloaded these data and organize them as our data sources. These data sources need slightly more preprocessing before data converting and interlinking. The organization and course schedule data sources are obtained by this approach.

Other data could come from webpage crawling, and this is often the major approach of getting raw data. We use information extraction technology to extract structured data out of description pages on the web, e.g. organizations' main page and people's homepage. Webpages are crawled following website structure of the university, we then analyze the pages and extract semantic properties to build structured data source. These data sources are usually of inferior quality, with a lot of noises and mistakes. After necessary alignment, they can be matched and linked with other data sources.

3.2 Data converting and mashup

After collection and preprocessing, we have got structured tabular data in different formats, such as SQL database, MS Excel .xls file and plain text file. These data are then marked with URI using our naming strategy.

[9] `http://open.vocab.org/`

Different data sources are then linked together using usual mapping technique. For example, in the course data source, lecturer is stored using lecturer's name. Then the lecture is mapped to a instance entry in the staff member data source who has the same name.

Finally, these data sources are converted to RDF format using RDF generating tools like *D2R Server*[10] and *ConvertToRDF*[11].

4 Tsinghua University Open Data

4.1 Datasets

Tsinghua University Open Data currently has 5 core datasets, describing the basic information and structure of the institution.

Campus Buildings and Places This dataset describes buildings, sightseeings and other geospacial entities related to Tsinghua University. Properties include name, description, geographic location, type, image, etc. This dataset currently has 784 triples, with 114 instances described.

Organizations This dataset contains description for the 87 organizations above department/school level. Organization name, homepage type and structure are described. This dataset currently has 372 triples.

Staffs We crawled all the staff members listed in all organizations' reference pages using automatic extraction along with manual selection, totally 2758 instances. Considering data deficiency and error, this may not be the actual count of all Tsinghua staff members. Name, gender, title and contact information of the staff members are described with 21631 triples.

Campus Photographs This dataset contains additional photographs about the campus. Each photograph is listed with its URL and the place it depicts. There are currently 316 photographs described with 977 triples.

Courses Schedule We extract the public course information from the university's website and form this dataset. We have done 30 recent semesters, with course name, lecturer, delivering department and open type described. Each semester has less than 2000 courses and about 10,000 triples.

Relationships of these five dataset are illustrated in Fig. 1

4.2 System Infrastructure and User interface

We have built a web portal (`http://data.cs.tsinghua.edu.cn`) for publishing our datasets using JSP and JavaScript. There is a dataset catelog page available for browsing and downloading datasets. We have also set up a standard SPARQL endpoint, with a HTML front end for query and browsing. We have also made a customized web page for every class, corresponding to HTTP request of an instance URI. Thus every URI can be dereferenced, as is required by Linked Data principle.

[10] `http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/`
[11] `http://www.mindswap.org/~mhgrove/convert/`

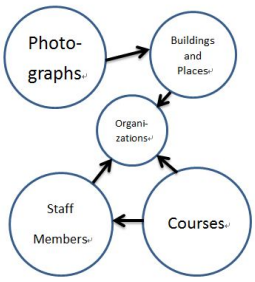

Fig. 1. Tsinghua University Open Data Cloud

4.3 Applications

We aim at providing a unified open data platform about Tsinghua University and other academic institutes, and we hope that researchers and programmers can make full use of the datasets and build interesting and useful applications upon them. Meanwhile, we have built several interesting applications by ourselves. These applications all use the SPARQL endpoint to access data, and most of them use two or more datasets.

CampusAssistant [12] provides searchable map for finding buildings, navigation, memo and paths in Tsinghua University. This application uses the *Buildings and Places dataset*, as well as the *Photographs* dataset.

CourseFinder [13] provides various ways for searching course information. A course can be found by searching its name, the lecturer's name, the facility's name, or combination of these criteria. What's more, one can easily navigate between courses, lecturers and departments; view detailed information about the lecturer, e.g. contact information(if public); and find relevant courses and teachers. All these need no extra work and is automatically achieved using interlinks between datasets.

5 Realated Work

There are numbers of projects related to building and linking open university data. *Waterloo Open Data Initiative*[14] offers information about campus buildings, classes schedule and exam information about University of Waterloo. Their data are provided with standard structured format like JSON and CSV file, but not in Linked Data format.

University of Southampton Open Data[15] provides open access to some of the university's administrative data. Currently, they provide 29 datasets, covering

[12] `http://iweb.cs.tsinghua.edu.cn/CampusAssistant`
[13] `http://data.cs.tsinghua.edu.cn/OpenData/courses.jsp`
[14] `https://opendata.uwaterloo.ca/drupal/`
[15] `http://data.southampton.ac.uk/`

from campus information to public phonebook of teachers and students, with RDF file and SPARQL endpoint provided. They have also built several applications, mostly school maps and phonebooks. They provided relatively wide-ranging datasets, but some of the datasets are not tidily bounded to the university, and some basic datasets are missing, e.g. course information and school affairs statistics.

Several other open university data portals are also on the go, *OU Linked Data*[16] currently has 6 datasets, but without university organizations and staff members description; *Open Data about the University of Oxford*[17] is under construction and not yet public. A incomplete list of European universities that have published open data can be found at *Linked Universities*[18].

6 Discussion

In this section, we will talk about some key points and difficulties me have met throughout the procedure of building open data.

First of all, "Useful" is the first principle when publishing open data, thus making data quality a great concern. We notice that several aspects of data processing act as an important role for better data quality.

For data retrieved from webpage crawling, careful strategy for handling incomplete data records is needed. Some missing fields of a record just result in missing properties of the generated instance; however, missing of some key values may leave the instance pointless and omitted. We have developed an automatic filtering system when generating RDF data from the data sources.

Problems concerning data privacy and copyright also cannot be ignored when publishing open data, especially data involving personal information. We must be sure that our collected data is either public or properly licensed; and when processing data, we do data filtering (filter bad and false data) and add non-sensitive information to data(by adding links and meta data), but we never do any modification or correction to existing data. A formal disclaimer is provided along with the data as well.

7 Conclusion and Future Work

Linked Open University Data applies Linked Data to publishing information about universities and academic institutions. In this paper, we discuss university data category, vocabularies, data collection and cleaning, as well as method of build and mushup linked data about university. Tsinghua University Open Data is presented as a showcase. We describe its datasets and applications upon them. These applications, however, is also portable for any other datasets with the same or similar schema.

[16] `http://data.open.ac.uk/`
[17] `http://data.ox.ac.uk/`
[18] `http://linkeduniversities.org/`

As future work, we will first improve our datasets. This include work in several aspects. Firstly, we will continue to find other data sources to add more dataset. Secondly, we plan to do deeper work at the data preprocessing and mapping phase, like doing name disambiguation.

Another important future work is to connect our datasets to the various existing Semantic Web data. With more raw data available, we can connect our datasets with academic publication datasets, geographic datasets, etc. This will further improve the power of our datasets and open new possibilities to applications.

We hope that more academic institutions can join the work of publishing open data, to build increasing number of datasets publish from different institutions using unified schema. Applications can then make full use of the data, creating useful and interesting tools and amazing user experience, which in turn inspire more institutions providing open data.

References

1. Berners-Lee, T., Hendler, J., Lassila, O. The Semantic Web. Scientific American, 284(5):34-43.
2. Tim Berners-Lee: Design Issues: Linked Data, `http://www.w3.org/DesignIssues/LinkedData.html`. 2009.
3. Chris Bizer, Richard Cyganiak, Tom Heath: How to Publish Linked Data on the Web (Tutorial), `http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/`. 2007.
4. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, C., Ives, Z. DBpedia: A Nucleus for a Web of Open Data. Proceedings of the 6th International Semantic WebConference (ISWC2007)
5. C. Bizer, T. Heath, and T. Berners-Lee. Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems(IJSWIS).2009
6. Auer, S., et al. Triplify - Light-Weight Linked Data Publication from Relational Databases. Proceedings of the 18th World Wide Web Conference (WWW2009)
7. Bizer, C., Cyganiak, R. D2R Server - Publishing Relational Databases on the Semantic Web. Poster at the 5th International Semantic Web Conference (ISWC2006)
8. Chen, H., Wang, Y., Wang, H., Mao, Y., Tang, J., Zhou, C., Yin, A., Wu, Z.: Towards a semantic web of relational databases: a practical semantic toolkit and an in-use case from traditional chinese medicine. In: 4th International Semantic Web Conference (ISWC), Athens, USA. LNCS, pp. 750C763. Springer, Heidelberg(2006)
9. Sören Auer and Jens Lehmann. What have innsbruck and leipzig in common? Extracting semantics from wiki content. In Enrico Franconi, Michael Kifer, and Wolfgang May, editors, ESWC, volume 4519 of Lecture Notes in Computer Science, pages 503-517. Springer, 2007.
10. Metaweb Technologies, Freebase Wikipedia extraction (wex), `http://download.freebase.com/wex/` (2009).
11. X. Li, J. Bao, J. A. Hendler, Fundamental analysis powered by semantic web, in: Proceedings of IEEE Symposium on Computational Intelligence for Financial Engineering and Economics, 2011.