LOD.CS.UNIPA Project: an experience of LOD at the University of Palermo

Davide Taibi
Institute for Educational Technologies
National Research Council of Italy
Palermo, Italy
davide.taibi@itd.cnr.it

Giovanni Panascì
Department of Mathematics and Computer Science
University of Palermo
Palermo, Italy
giovipanasci@hotmail.com

Biagio Lenzitti
Department of Mathematics and Computer Science
University of Palermo
Palermo, Italy
biagio.lenzitti@unipa.it

ABSTRACT

This paper describes the LOD.CS.UNI.PA Project and its main goal, the transformation process of data already available on the web site of the Computer Science curricula web site at the University of Palermo into data ready to be connected to the LOD. Since 1997 information about bachelor and master degrees in Computer Science at the University of Palermo has been published on the web, and provides a reference point for students, teachers and researchers who have easy access to the information they require. However, the users of the web are now changing; data cannot be published only for human comprehension but intelligent devices also need access to web data and above all they need to understand them. In 2006 Tim Berners Lee presented a star rating system for the data available on the web. Following his five star classification, the aim of the work described in this paper is to raise the level of data concerning degrees in Computer Science at Palermo University, from one star (data available on the web in whatever format), to five stars (data connected to other LOD datasets). So far the data has been transformed to a four star level in which each item has a URL that can be dereferenced, but the final objective of this project is to reach the five star level.

Categories and Subject Descriptors

E.1 [**Data**]: *distributed data structures.* K.3 [**Computers and Education**].

General Terms

Keywords

Linked Open Data, Learning, Academic web data.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Linked Learning 2012, 17 April 2012, Lyon, France.

1. INTRODUCTION

Universities and research centers produce a huge quantity of data every year, related to courses, teachers and students. Moreover, the information provided by this type of organization changes constantly, often due to teachers moving from one institution to another, or visiting professors or researchers staying only for a short period of time. Usually, all these data are published on the web, and the websites of universities and research centers are a reference point for students, particularly Erasmus (EuRopean Community Action Scheme for the Mobility of University Students)[1] students, teachers and general users interested in information about the subjects syllabus, expertise and contact details of the teachers, the lecture timetable, exam dates, and so on. The Web has provided a platform for making data accessible to a huge number of people. The evolution of the Web is strictly connected to the way users interact with it. Nowadays, the potential users of web data are not only human beings but also software services and software agents. For this reason data should be published on the web using standards and technologies which can be understood and elaborated automatically. At present, the most popular Web applications, such as Facebook and Youtube, offer the Application Program Interface (API) that allows software agents to access the information they host. Semantic Web technologies [1] provide an adequate technological substrate for supporting the representation of concepts and the relationships between them through ontologies, and the recent evolution of Linked Open Data is the natural way to publish, integrate, and link data semantically described. The information available on the web uses different typologies and is published in heterogeneous formats. Linked Open Data aims to provide a technological substrate for publishing structural data in a standardized format. The advantages of such an approach are tangible and it is increasingly common for data on the web to be published following the LOD principles [2]. While the linking of pages has marked the success of the Web, at the same time LOD aims to connect datasets and the concepts they host, providing information not only for humans but also for software agents.

This paper describes the transformation process of the data hosted on the University of Palermo Computer Science department web site. In section 2, experiences of similar processes in other

[1] http://ec.europa.eu/education/lifelong-learning-programme/erasmus_en.htm

institutions are analyzed, taking into consideration not only the most well-known project in the LOD scenario but also two Italian projects that deal with university and research center data. Section 3 explains the process of transformation, providing a description of the ontology designed, the tools used and the endpoint developed to access the data. In section 4 future developments are proposed, as at present the transformed data are not ready to be included in the Linked Data cloud. In fact to satisfy all the four linked data principles, it is necessary to provide more links with the dataset in the cloud, and some suggestions on how to achieve this aim are presented.

2. RELATED EXPERIENCES

Many universities and research centers have embarked on projects which aim to expose semantic representations of their data. The most significant of these are already represented by a node in the Linked Open Data cloud[2]. The advantages of the Linked Data approach, in which the concepts expressed in the datasets are semantically linked, have a considerable impact in the academic world, as witnessed by the numbers of experiences already undertaken. The Linked Data provides a way of connecting data that by their very nature are already linked. Universities are connected via publications and common research interests, and these data are public and available on the web, in most cases on institutional web sites using traditional web technologies. It is important to note that the majority of Linked Data initiatives at universities have been carried out in the UK. An analysis of the Italian scenario shows that even though many projects using LD have been developed for government data, only a few have been implemented for university or research center data (CNR and LOIUS are the most notable).

The following list highlights some of the most significant projects in this area:

- The Lucero project[3]: this project is promoted by The Open University of Milton Keynes (UK). It allows access to a dataset of public data concerning The Open University publications, course materials, podcasts and video lessons. Many applications have been built upon the datasets to demonstrate the usefulness of structured data in a semantic format.

- The open data platform of the National Research Council of Italy (CNR) [3] uses a Linked Open Data compliant format to publish information related to structural organization, administrative data and personnel data for the whole CNR. Applications to browse data and facilitate visualization and research have also been provided.

- LODUM[4] project: this project aims to publish the University of Munster data on the LOD. One of the applications developed on the semantically structured data is the productivity map on Google Earth, in which the height of the university buildings is directly proportional to the number of publications produced by the researchers working in the building.

- LOIUS[5] project: the data published by the Statistical Office of the Italian Ministry of Education, University and Research (MIUR), are represented following the principles of the Linked Open Data, using semantic technologies. This data mainly concerns the number of graduate students from each university,

- The University of Sheffield's Department of Computer Science has developed a project to provide access to data concerning members of staff, research groups, and publications [4]. This information is represented using semantic web technologies following the principle of Linked Data.

In the United Kingdom the use of Linked Data to publish university data is becoming increasingly common. Besides the experiences cited in the list above, the universities of Lincoln[6], Oxford[7] and Southampton[8] have already exposed their data using the LOD approach.

3. LOD.CS.UNI.PA Project

3.1 The data available at the degrees in Computer Science of Palermo University

Since 1997 the University of Palermo has published information on the web about university staff, research groups, didactic groups, faculties and departments. Today, each department provides information about its own staff and courses. And for each course, more detailed information concerning timetables and course syllabuses is directly managed. However, this information is generally structured in a format that cannot be interpreted semantically and is generally published on the web in pdf or doc format, with the exception of a few engineering, science and economics websites which provide the information in html format. For example, the web page for computer science degrees is called CISI (in Italian "Consiglio Interclasse Scienze Informatiche") and has been on the web since 2003.

The successful experiences reported in section 2, highlight the need for data to be properly revised and provided with a semantic representation according to Semantic Web and LOD principles. The aim of the LOD.CS.UNI.PA Project is to raise the level of the data in line with the star rating system proposed by Tim-Berners-Lee in 2006 [5]:

1. make your stuff available on the web (whatever format)
2. make it available as structured data (e.g. excel instead of image scan of a table)
3. use non-proprietary format (e.g. csv instead of excel)
4. use URLs to identify things, so that people can point at your stuff
5. link your data to other people's data to provide context

At present the CISI data are available on the web in an unstructured format so they have only one star, but to satisfy the

[2] http://richard.cyganiak.de/2007/10/lod

[3] http://lucero-project.info

[4] http://lodum.de/

[5] http://sw.unime.it/loius/info.html

[6] http://data.lincoln.ac.uk/

[7] http://data.ox.ac.uk/

[8] http://data.southampton.ac.uk/

five star criteria they must become Linked Data, linked with other datasets in the cloud.

3.1.1 *Existing data*

This paper analyzes the part of the CISI database containing information about users (teaching staff and administrative staff), courses (with their modules) and subjects.

- The user table stores all the users belonging to the department and their roles. Besides general information such as name, surname and phone number, more specific data are also stored: role, home page and location of the office in the department. Teachers are classified into professors, associates, researchers and visiting professors.
- The subject table contains specific information about the subjects taught at the department. Each subject is identified by its name, identifier code and the number of modules (in fact a subject can be divided into two or more modules). Further information is provided about its credit value according to the ECTS (European Credit Transfer System) and indications are included as to whether Erasmus students are eligible to take exams in the subject.
- The module table contains information about its name, code and module home page; each module is associated to a subject and to a professor.

3.2 Data transformation process

The data transformation process adopted by the project consisted of three phases: ontology building, data extraction and storing, and access interface creation, as shown in figure 1.

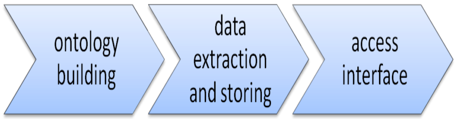


Figure 1: Transformation process phases

In the ontology building phase, a specific ontology was created in accordance with Semantic Web and Linked Open Data principles and guidelines, using existing ontologies. AIISO and AIISO Roles and FOAF provided some classes and properties for academic information and user information respectively. Considering that some information is related specifically to the Italian University system and some entities such as the ECTS values and Erasmus eligibility is not included in these ontologies, it was necessary to develop a more extensive ontology .

The starting point of the data extraction phase was the database described above. In this phase, extraction technologies (i.e. dump-based extraction) and instruments (i.e. Triplify, ARC2) were opportunely configured and used. In particular, in this phase, Triplify was used to query the database and extract RDF triples and ARC2 was used for the triple store.

In the third phase, a SPARQL Endpoint was created to access the triples contained in the triple store. More detailed information about ontology, extraction and storing processes and access interfaces are provided respectively in 3.2.1, 3.2.2 and 3.2.3.

3.2.1 *Ontology building*

The reuse of existing data and in particular of existing ontologies is crucial in the logic of Linked Open Data and the ontology built follows this guideline. AIISO (Academic Institution Internal Structure Ontology) [6] and AIISO Roles [7] ontologies were partially reused. AIISO provides classes and properties to describe internal university organizations, while AIISO Roles provides information about academic roles. The following tables show the main classes and properties reused from the two ontologies.

	AIISO	**AIISO Roles**
Classes	Department Faculty Module Subject	Professor Associate Researcher Visiting Professor
Properties	code name part_of	

Table 1 : Concepts and properties from AIISO

It was necessary to extend these ontologies and adapt them to suit the Italian university structure because these ontologies were built for the United Kingdom universities organization and the Italian scenario is quite different. For this reason, certain properties have been added to Subject and Module: 'cfu' defines how many hours of lectures and study are necessary, 'period' indicates teaching periods, 'type' defines whether a Subject or a Module teaching period refers to a bachelor or master's degree and 'sector' defines the scientific field. In particular, the cfu property falls within the European Credit Transfer System which has been developed to facilitate the recognition of study periods abroad, thus improving the quality and volume of student mobility in Europe under the Erasmus program, and providing a credit accumulation system at European level. Some properties were also added to the teacher class: 'teacherOf' specifies which subject or module is taught by a teacher (in AIISO ontology this property describes which subjects or modules are taught at the university and not the professor who teaches them), 'room' shows the location of the teacher's room in the Department and 'receipt' defines the teacher's office hours for receiving students. Personal teacher data are indicated using FOAF ontology [8].

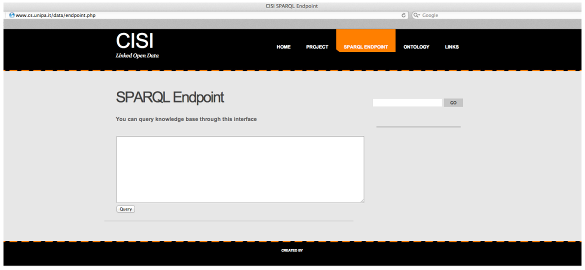
3.2.2 *Data extraction and storing*

Information contained in the CISI SQL database was extracted using a dump-based technology that takes advantage of creating periodic archives containing database dumps; the same technology was also successfully used in the DBpedia project in conjunction with other Wikipedia extraction techniques [9]. Triplify was opportunely configured to extract dump information [10] and generate an RDF file containing N3 RDF triples for unambiguous resource identification through URIs, in agreement with Linked Open Data principles. ARC2[9], that natively supports PHP, allowed the creation of a triple store; it was necessary in the CISI LAMP architecture.

[9] https://github.com/semsol/arc2/wiki

3.2.3 *Access interface*

Figure 2 shows a screenshot of the SPARQL Endpoint which is available at the following address: http://cs.unipa.it/data/csunipa. It makes possible to query triple store and shows the query results. Triplify is used to provide an unambiguous resource identifier allowing correct URI dereferencing: in fact the creation of unambiguous identifiers allows users to access the URI and use information contained in a single resource.



SPARQL Endpoint

You can query knowledge base through this interface

Figure 2 : Screenshot of the endpoint

4. CONCLUCIONS AND ONGOING WORKS

The project presented in this paper aims to create a semantic representation of the data extracted from the CISI database, following the principle that Linked Open Data must be accessible not only by human users but also by machines. Many universities and research centers all over the world are following a LOD approach to publish their information. Data previously available on the web in html format, have been transformed into RDF triple format as part of the Web Of Data, where they can be linked with similar data types stored in datasets. The main advantages of such an approach lie in the capability of linking data silos to each other in the Web of data thus creating information that can be accessed not only by humans but by also from automatic consumption and elaboration.

So far, according to the Berners-Lee rating system, the CISI data have been transformed into four star data. To achieve the fifth star it is necessary to create links to other datasets that are already in the cloud and to increase the number of triples (about 1000 triples, excluding FOAF triples, is the minimum requirement). This target can easily be achieved by involving other departments (e.g. the math department, and other science departments), or even the entire faculty or university. Moreover, the CISI data can be expanded by inserting information into the triple store which regards teaching, such as research topics, professors' interests and publications, topics, or information about classroom and laboratory location, seating capacity or more generally about department buildings. It would even be possible to create references to the geographical position of a department, like latitude and longitude, linking the CISI dataset with the Geonames ontology.

The linking of the CISI dataset to other datasets already in the Linked Open Data cloud could be achieved, creating connections with data from other universities already in the cloud, or linking information about professors to datasets containing their scientific publications. Future work will progress in this direction with the aim of publishing the CISI data in the Linked Data cloud, providing access not only to humans but also to machine elaboration.

5. REFERENCES

[1] T. Berners-Lee, J. Henler e O. Lassila, The Semantic Web, A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American 284(5):34-43 (2001).

[2] Auer, S., The emerging web of linked data. Proceedings of the 2011 International Conference on Intelligent Semantic Web-Services and Applications, (2011).

[3] C. Baldassarre, E. Daga, A. Gangemi, A.M. Gliozzo, Salvati A., Troiani G.: Semantic Scout: Making Sense of Organizational Knowledge. Proceedings of Knowledge Engineering and Knowledge Management by the Masses EKAW, (2010).

[4] M. Rowe, Data.dcs: Converting Legacy Data into Linked Data. In proceedings of Linked Data on the Web Workshop, WWW 2010. Raleigh, USA (2010).

[5] T. Berners-Lee, N. Shadbolt and W. Hall, The Semantic Web Revisited, IEEE Intelligent Systems, v.21 n.3, p.96-101, (2006).

[6] R. Styles, N. Shabir, Academic Institution Internal Structure Ontology (AIISO), Available: http://vocab.org/aiiso/schema (2008).

[7] R. Styles, C. Wallace, Academic Institution Internal Structure Ontology Roles (AIISO Roles) Available at: http://vocab.org/aiiso-roles/schema (2008).

[8] L. Miller e D. Brickely, FOAF Vocabulary Specification, (2010).

[9] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak e S. Hellman, DBpedia - A Crtystallization Point for the Web of Data, (2009).

[10] S. e. a. Auer, Triplify - Light-Weight Linked Data Publication from Relational Databases, (2009).