Extending FUM-LD Framework by Including an Academic Data Model

Mahboubeh Dadkhah, Mohsen Kahani
Web Technology Lab., Dept. of Computer Engineering
Ferdowsi University of Mashhad
Mashhad, Iran
mb.dadkhah@stu-mail.um.ac.ir
kahani@um.ac.ir

Samad Paydar and Behshid Behkamal
Web Technology Lab., Dept. of Computer Engineering
Ferdowsi University of Mashhad
Mashhad, Iran
samad.paydar@stu-mail.um.ac.ir
behkamal@stu-mail.um.ac.ir

***Abstract*— Linked Data techniques have become interesting to organizations, especially for those which have some open data to publish, such as universities. In this paper an extended framework is presented for publishing a subset of data of Ferdowsi University of Mashhad as linked data. Since, academic data of various universities are modeled differently, a common data model for academic institutes is proposed, too.**

Keywords- Publishing Framework; Academic Data Model; Linked Data

I. INTRODUCTION

The intention of early Web was interlinking information from various sources to solve organizational problems, such as the high turnover of people and the restriction of information to data stores. The focus of this Web is mainly on humans as information consumers, and therefore, information is presented in a way which is readable and understandable for humans. With the enormous growth of the Web, the size of this information space has become so huge that the effective use of available information is beyond human capabilities. Semantic Web introduced the idea of extending this document-centric and human-oriented web with a semantic layer which enables machines to understand process and consume web information.

Linking Open Data (LOD) project[1] is the most important project initiated by Linked Data community to provide simple guidelines for presenting structured data in uniform machine-understandable format. It also puts forward mechanisms for publishing, interlinking and accessing this data.

Linked Data techniques have became interesting to organizations of every shape and size, especially for those which have some open data to publish, such as universities.

The data of universities and their activities is important to many web users such as students, researchers and teachers. Such data, if published as Linked Data and linked to appropriate datasets (e.g. general datasets like DBpedia, or special datasets like DBLP or ACM), can provide valuable benefits by enabling different scenarios of fulfilling users' information need. For instance, it can help students to search for professors or departments to apply, based on the professor's attributes or the properties of the department.

In this paper, a framework for publishing academic linked open data will be presented. The first version of this framework, named FUM-LD, has been discussed in [1]. FUM-LD stands for Ferdowsi University of Mashhad Linked Data.

The aim of this paper is to present the extended parts of FUM-LD; so, we use "extended FUM-LD framework" term to refer to our project.

The paper is structured as follows: At first, a brief overview of the literature is presented. In section 3 by focusing on extended parts, the main components of FUM-LD framework are discussed. The experimental result of this project will be presented in Section 4. Finally, the paper is concluded and our future works is given in Section V.

II. RELATED WORKS

With regard to Linked Data, there are some works in the literature, dedicated to the experiences of publishing datasets. Here, we shortly discuss some of these works.

In [2], authors discussed common errors in RDF publishing, their consequences for applications, along with possible publisher-oriented approaches to improve the quality of structured, machine-readable and open data on the Web. They provided discussions for some issues related to how data is found and accessed, parsing and syntax issues, reasoning issues, inconsistent data, and ontology hijacking, both from the perspectives of publishers and data consumers.

LinkedMDB [3] deals with publishing data of major movie datasets (e.g. IMDb, OMDB) as linked data. LinkedMDB includes links to several external datasets, like FreeBase, DBpedia, RottenTomotoes, YAGO, Geonames and lingvoj.

In [4] challenges of publishing Persian linked data are discussed then by analyzing the experimental results of the project and classifying the problems, some publisher-oriented solutions are proposed.

A common prerequisite to publishing data is the quality of data. Data quality is often defined as the ability of a collection of data to meet desired requirements. It is therefore important to ensure the data is going to be published as linked data have a high data quality.

ISO/IEC 25012 [5] is one of the SQuaRE series of International Standards, under the general title of Software

[1] http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

engineering Software product Quality Requirements and Evaluation (SQuaRE) which defines a data quality model. This International Standard can be used with other standards in the SQuaRE series to establish data quality requirements, define data quality measures, or plan and perform data quality evaluations.

In [6], authors identified the need to evaluate data quality in Web portals and defined a data quality model, denominated as SPDQM (SQuaRE Portal Data Quality Model), which is based on a previous model, PDQM (Portal Data Quality Model) and the ISO/IEC 25012.

In [7], authors discussed the experiences gained in the process of academic bibliography of the University of Economics in Prague transformation from its current data format (MARC) to RDF-based data. The data model consists of concepts that its description or meaning matches closely as possible to MARC bibliographic data element meaning.

It should be noted that there is too much little data represented in Farsi language on the LOD cloud (e.g. there is little Farsi data on DBpedia). As a result, there is no work regarding publication of Farsi datasets as Linked Data.

III. FUM-LD FRAMEWORK

In this section, the extended FUM-LD framework is described. FUM-LD is a project for publishing academic linked data and stands for Ferdowsi University of Mashhad – Linked Data.

Since the first version of this framework is discussed in [1], here we focus on the extended parts of FUM-LD.

As Shown in Fig. 1, the extended FUM-LD framework consists of a repository and six core applications, which will be introduced briefly in the following subsections.

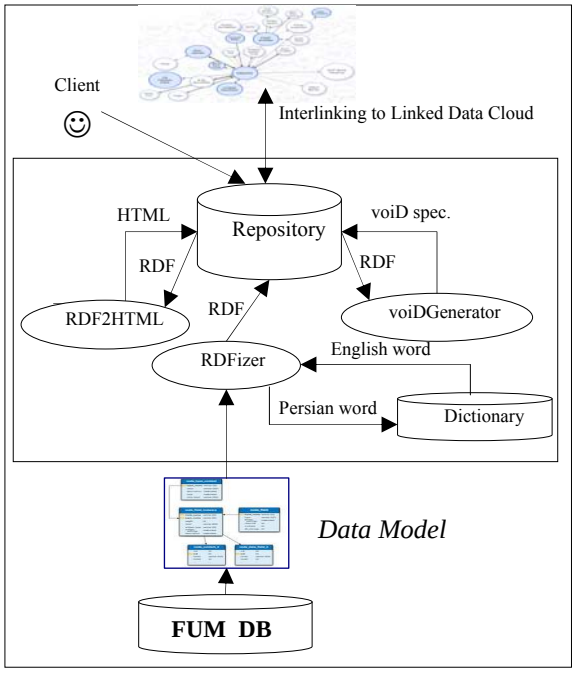

Figure 1. FUM-LD framework

A. RDFizer

RDFizer extracts data from FUM relational database and creates an RDF file for describing each entity, and stores it in the repository. Before publishing the dataset, quality of data being published is evaluated. Data quality evaluation involves specifying data quality characteristics. The ISO/IEC 25012 [5] standard determines a set of generic characteristics for data quality. Three quality characteristics from this set are adapted for the framework in accordance with the inherent viewpoint of data quality and predefined academic data model. These characteristics are Completeness, Accuracy and Consistency. Completeness determines for each field the degree to which significant data values exist (i.e., not spaces, nulls, or defaults). The measurement function is: A/B. Where 'A' is the number of records with required value and 'B' is total number of records. Accuracy determines for each data element if those values are from the valid value set. Based on the valid value set, three categories are defined: string values, values which have a template like email address, date and URLs and fields that their value should be in an acceptable range (e.g. publish date, birth date and values with predefined minimum and maximum). The measurement function is: A/B. Where 'A' is the number of records with the specified field accurate and 'B' is number of records.

Different vocabularies are used in describing resources: FOAF [2] is used for describing personal information of professors and department). Dublin Core[3], BibTeX[4], and MarcOnt[5] are used for describing publications of professors. SKOS [8] subjects are used in describing courses, departments and faculties.

Linking FUM dataset to other external datasets consist of two steps. Since FUM is a Persian dataset, at first each term should be translated to English. For this purpose, a local dictionary is used to find the appropriate equivalents of terms. Then the application automatically searches the external dataset for each English term using its SPARQL endpoint. This search is based on a number of empirical heuristics and simple SPARQL templates defined in RDFizer which are instantiated in runtime to perform the search.

B. RDF2HTML

In addition to RDF representation, FUM-LD framework generates human-friendly HTML representation of resources. RDF2HTML processes the RDF files in the repository and generates corresponding HTML files and stores them in the same repository.

C. voiDGenerator

The framework uses voiD [9] vocabulary to describe the published dataset. It is a vocabulary for describing RDF datasets in terms of their provenance, statistical, structural and licensing information. Using voiD to describe published datasets provides advantages from different points of view,

[2] http://xmlns.com/foaf/spec/
[3] http://dublincore.org/
[4] http://www.bibtex.org/
[5] http://www.marcont.org/

such as trust, searching, ranking and selecting datasets [9], [10].

voiDGenerator processes RDF files in the repository and generates the voiD specification of the whole dataset as a single RDF file. In addition to some basic information about the dataset (e.g. its subject, definition, publication date, contributors, example resources), this specification declares the main vocabularies used in describing the resources, number of resources of type foaf:Person, total number of RDF triples, different subsets and link sets of the dataset.

D. Dictionary

Since most data on LOD cloud is published in English, it is hard to link a Persian dataset to the related external datasets. To the best of our knowledge, there is no work in the literature discussing this problem, even for other non-English datasets. In multi-language systems where data is generated freely by ordinary end-users, it is possible that some users choose their native languages while others use English for entering their data, whether for their convenience, or because of their field of activity. For instance, in the FUM database, for the engineering faculty members, data mostly contains English data, while for the theology faculty members, Persian and Arabic data is dominant. As another example, identical Persian terms exist in different English forms in the database, e.g. a single Persian name "سعید" is entered both as "saeed" and "saeid". Such problems caused by multi-lingual data, introduce challenges when searching external datasets for related resources to be linked, and decrease the quality of the published dataset.

One way of addressing such problems is to use a dictionary to identify different equivalences of a word from one language to another. For instance, in FUM-LD framework, the dictionary element provides access to different equivalents of a Persian name in English. Using this dictionary, it is possible to use all equivalents of a professor name, when searching external datasets. As a result, the probability of missing a related link because of different spelling is reduced.

E. Data Model

Most of published data on LOD are transformed from enterprise databases. These databases contain public internal and external information including organization's assets, equipments, facilities, locations, partners, customers and stakeholders as well as some confidential information.

An important issue in publishing linked data is deciding which data should be published. Two factors must be considered when selecting data. First, data should be open and second, data should be related to the domain of interest.

Since our current works focus on publishing academic data, we propose an academic data model as an input of FUM-LD.

After studying some academic data sources, nine important entities usually exist in any academic institute are selected as follows:

University,
Faculty,
Department,
Research Group,
Professor,
Student,
Course,
Project,
Publication.

Identifying main entities, the relationships of these entities are defined as a metadata of FUM-LD framework. Proposed data model as well as relationships between entities is represented in Fig. 2. The required properties of each entity is specifically shown in TABLE I.

Specifying required data entities for publishing academic data as linked data, three levels of different priority categories to different data elements of each entity are defined, based on their perceived importance. Priorities are used to ensure that the most important data elements are provided with high quality. Three different categories are used to classify priority of data elements in each entity of the data model. Some of data elements have significant impact on the quality of published data; therefore it is essential to provide this data with complete, accurate and consistent values. These data elements are specified as high priority elements and are included English and Persian names and titles for all entities, university and faculty URLs. There are data elements with medium priority which have some impacts on the quality of published data. Some of the medium priority data elements are contact information like address, phone number and fax number, homepage or email address of a project or research group. The last category consists of Low priority data elements. These elements will have little impact on the quality of published data. Some of low priority data elements are English and Persian descriptions, establish date of a university, faculty or research group.

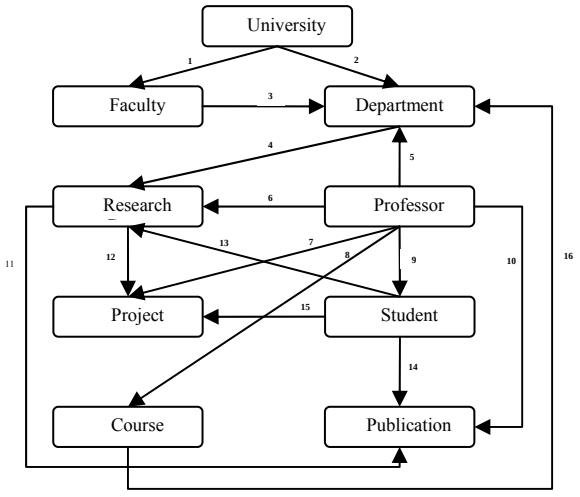

Figure 2. Proposed Academic Data Model

TABLE I. RELATIONSHIPS BETWEEN ENTITIES OF DATA MODEL

No	From	To	Relationship
1	University	Department	Includes
2	University	Faculty	Includes
3	Faculty	Department	Includes
4	Department	Research Group	Includes
5	Professor	Department	Membership
6	Professor	Research Group	Leadership
7	Professor	Project	Management
8	Professor	Course	Present
9	Professor	Student	Supervise
10	Professor	Publication	Has
11	Research Group	Publication	Has
12	Research Group	Project	Has
13	Student	Research Group	Membership
14	Student	Publication	Has
15	Student	Project	Work on
16	Course	Department	Present at

IV. EXPERIMENTAL RESULTS

A subset of data of Ferdowsi University of Mashhad has been published as linked data with FUM-LD framework. The process of publishing these data based on the FUM-LD framework has four steps. The first step of the process is providing target data. Different educational and organizational web-based systems are being used at Ferdowsi University of Mashhad. After studying the FUM database, required data for creating entities correspond to the data model entities were extracted from the FUM DB. In the second step, appropriate URIs has been assigned to entities. Different approaches are possible for assigning URIs to entities which are to be published. In FUM-LD, a simple URI schema is used for this reason:

http://wtlab.um.ac.ir/LD/TYPE/ID

Where TYPE is name of the entity in data model, and ID is the unique identifier of the entity in the database. In the third step the data has been published by FUM-LD framework. As interlinking data sources is an important issue in publishing this dataset, the last step is providing links to other resources inside and outside the FUM-LD.

To evaluate the quality of data being published, these three data quality characteristics are measured for data elements of each category. For each category a desirable average value of data quality model characteristics is suggested. After verifying structure of the given data base (number of tables, name of each table, number of fields, name and type of each field), the average values of data quality model characteristics for all the data elements in each category are measured and compared with the suggested values. Suggested values of data quality characteristics and measured value for data elements of FUM DB are presented in TABLE II.

TABLE II. COMPARISON OF MEASURED CHARACTERISTICS FOR DATA ELEMENTS OF FUM DB WITH SUGGESTED VALUES

		Completeness	Accuracy	Consistency
high priority	Suggested value	100%	100%	100%
	Measured value	97.237 %	86.315 %	96.427 %
medium priority	Suggested value	80%	80%	80%
	Measured value	78.461 %	67.329 %	82.019 %
low priority	Suggested value	20%	20%	20%
	Measured value	44.954 %	62.781 %	73.964 %

V. CONCLUSION

This paper describes the FUM-LD framework, a framework for publishing academic linked open data on the web. Six core applications composing the framework are discussed. Entities usually exist in academic institutes are studied and important ones are selected to form an academic data model for publishing academic data. The proposed data model is generic and can be used for publishing data of any academic institute.

Experimental results from publishing data of Ferdowsi University of Mashhad as linked data with this framework are reported. The results showed that tackling data-related problems is really important in publishing high-quality datasets.

Since, the main focus of this project is on publishing academic data, we are going to improve FUM-LD framework. So, our future works include developing a comprehensive framework to publish academic linked data and improving the quality of the published dataset.

REFERENCES

[1] S. Paydar, M. Kahani, B. Behkamal, M. Dadkhah, and E. Sekhavaty, "Publishing Data of Ferdowsi University of Mashhad as Linked Data," in International Conference on Computational Intelligence and Software Engineering (CiSE), 2010, pp. 1-4, doi: 10.1109/CISE.2010.5676872.

[2] A. Hogan, A. Harth, A. Passant, S. Decker, and A. Polleres, "Weaving the Pedantic Web, Linked Data on the Web workshop (LDOW2010). Raleigh, North Carolina, USA, 2010.

[3] O. Hassanzadeh, M. Consens, "Linked Movie Data Base," Proceedings of the 2nd Workshop on Linked Data on the Web (LDOW2009), Madrid, Spain, April 2009.

[4] B. Behkamal, M. Kahani, S. Paydar, M. Dadkhah, and E. Sekhavaty, "Publishing Persian linked data; challenges and lessons learned," in 5th International Symposium on Telecommunications (IST), 2010, pp. 732-737, doi: 10.1109/ISTEL.2010.5734119.

[5] [ISO/IEC-FDIS-25012], Software Engineering - Software product Quality Requirements and Evaluation (SQuaRE) - Data Quality model, 2008.

[6] C. Moraga, M. Moraga, C. Calero, and A. Caro, "SQuaRE-Aligned Data Quality Model for Web Portals," in Ninth International Conference on Quality Software, 2009, pp. 117-122, doi: 10.1109.

[7] J. Hladka and J. Mynarz, "Exposing the University of Economics' academic bibliography database as linked data," in 3rd International Conference of Semantic Web in Bibliotheken (SWIB10), Cologne, Germany, 2010.

[8] A. Miles and S. Becchofer, "SKOS simple knowledge organization system reference. W3C working draft," World-Wide Web Consortium, January 2008.

[9] K. Alexander, R. Cyganiak, M.Hausenblas and J. Zhao, "Describing Linked Datasets," Proceedings of the 2nd Workshop on Linked Data on the Web (LDOW2009), Madrid, Spain, April 2009, doi: 10.1.1.178.2369.

[10] N. Toupikov, J. Umbrich, R. Delbru, M. Hausenblas, and G. Tummarello, "DING! Dataset ranking using formal descriptions," Proceedings of the 2nd Workshop on Linked Data on the Web (LDOW2009), Madrid, Spain, April 2009.