IJICT International Journal of Information & Communication Technology

ITRC

Volume 2- Number 3- November 2010 (pp.9 - 19)

Publishing Persian Linked Data; Challenges and Lessons Learned

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

Mohsen Kahani
Web Technology Lab., Dept. of Computer Engineering
Ferdowsi University of Mashhad
Mashhad, Iran
kahani@um.ac.ir

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

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

Received: September 20, 2010 – Accepted: November 26, 2010

***Abstract*—Linked Data as an important and novel subject has attracted great attention in the realm of the Semantic Web. Many works deal with publishing existing datasets as Linked Data. This paper discusses the challenges of publishing Persian linked data, and their potential solutions, based on the experiences and lessons learned from a project focused on publishing some academic data of the Ferdowsi University of Mashhad as Linked Data.**

Keywords- Linked Data; Persian Dataset; LOD cloud; RDF; challenges and Solutions

I. INTRODUCTION

Linked Data movement has been integral to RDF publishing on the Web, emphasizing four basic principles: (i) use URIs as names for things; (ii) use HTTP URIs so that those names can be looked up; (iii) provide useful information when a look-up on that URI is made; and (iv) include links using external URIs [2].

Over the past few years, many web publishers have turned to RDF as a means of disseminating information in an open and machine-interpretable way, resulting in a "Web of Data" which now includes interlinked content exported from corporate bodies, biomedical datasets, governmental entities, and organizational data.

Data of universities and their activities is important to many web users like 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.

This paper is in continuation of the previous experience with "FUM-LD" project [6]. FUM-LD is a framework developed for publishing the data of Ferdowsi University of Mashhad (FUM) as Linked Data. Herein, we discuss some problems and challenges of Persian linked data along with possible solutions, which are mainly focused on the data

[Downloaded from journal.itrc.ac.ir on 2024-10-12]

publisher side to improve the quality of structured, machine-readable, and open data on the Web.

The main contributions of this paper are: (i) classifying the main challenges of publishing Persian linked data, (ii) proposing potential solutions for those challenges, and (iii) comparing the proposed solutions with available ones, focusing on Persian linked data.

The paper is structured as follows: After defining basic concepts in the next section, some related works are discussed and classified in Section III. The elements of the FUM-LD framework are discussed briefly in the forth section. Then, the experimental results are analyzed in Section V. The identified problems of publishing Persian linked data are discussed in Section VI. Finally, the paper is concluded and future works are presented.

II. Definition of Basic Concepts

Linked Data is a major step toward realizing the vision of the Semantic Web by generating a global web-scale data space in which entities are described in a machine understandable format. Each piece of data can be explicitly linked to other related data using different link types.

Linked Data is a rather new subject in the realm of the Semantic Web. Since its introduction in 2006 by the inventor of the Web, Tim Berners-Lee, it has attracted much attention from the Semantic Web community [2]. Linked Data is in fact a set of best practices for publishing data on the web, so that the data is machine- understandable. This machine-oriented nature of Linked Data (in comparison with the inherent human-oriented characteristic of the traditional web) is met by utilizing the Semantic Web technologies (e.g. RDF, ontology) as its main building blocks. The web-scalability characteristic of the Linked Data is achieved by the fact that it is based on the simple yet effective and mature web technologies (e.g. URI, HTTP) that have been in use for years.

Linked Data principles can be summarized as [2]:

- Using URIs as names for things.
- Use of HTTP URIs so that people can look up those names.
- Providing useful information, using the standards (RDF, SPARQL), in the URIs
- Adding links to other URIs so that more things can be discovered.

When publishing a dataset as Linked Data, a URI is assigned to each entity in the dataset. This makes the entity uniquely identifiable and accessible on the web. The access mechanism is the simple HTTP protocol. When an entity is accessed by dereferencing its URI through HTTP, appropriate representation of that entity is returned. This representation is based on the RDF, and uses different ontologies to describe different attributes of the entity. Further, it is important to note that this RDF-based representation contains RDF links to other entities in the Linked Data space. These links are typed (by the use of ontologies) and relate different entities to each other. This explicit definition of the relations between different entities is essential if the machine is expected to use the data of different entities from multiple data sources.

III. Literature Review

To investigate the state of the art in the field of linked data, 34 related woks are comparatively studied and classified in five main groups. The classification is summarized in TABLE I.

A. *Publishing and linking data*

Generally speaking, the works on publishing and linking data can be classified in three subgroups:

1) *Publishing and linking data of different domains*

In the early days of linked data, the main focus of the community was on finding good practices for publishing data. However, beside publishing data, the interlinking between datasets is also important. The links can either be set manually or generated by automated linking algorithms for large datasets. When trying to interlink data from, for instance, the geographical domain using GeoNames, it is possible to do a simple lookup using the search facility provided by GeoNames. However, when querying for the city Vienna, nearly 20 results is returned as there exist that many cities named Vienna around the world. Advanced approaches are needed to disambiguate similar matches and finally create appropriate interlinks [3, 10, 12, 13, 14, 22].

2) *Converting different formats into RDF*

At present, almost all usable ontological data is built manually or by directly transforming certain (semi-)structured data sources into certain formats of semantic data. So, some of the existing works have investigated converting existing non-RDF datasets (such as HTML, relational DB, thesaurus, etc) into RDF models [15, 16, 26, 36].

3) *Publishing non-English data*

There are few works on publishing and interlinking non-English datasets. One of them describes the conversion of a large economics thesaurus to RDF/SKOS in both German and English. The built-in multilingual features of SKOS made it easy to handle the German and English labels connected to the concepts [36].

B. *Co-referencing and data fusion*

With the growing amount of semantic data being published on the web, the problem of finding resources in different datasets that correspond to the same entity is gaining importance. Also, the diversity of ontologies used by different datasets makes it hard for data integration methods to use the semantic data structure. So, some of the researches including [7, 8, 9, 17, 18, 19, 20, 21, 22, 23] have focused on co-reference resolution and data fusion.

C. *Linked Data applications and tools*

Many researchers have tried to develop facilities that are required for different tasks in Linked Data publishing. Examples include tools, services, and

[Downloaded from journal.itrc.ac.ir on 2024-10-12]

plugins that are used for storing, exploring or querying linked data, converting different formats to RDF, generating a linked data wrapper over a database [24, 25, 26, 27, 28, 29, 30, 31].

D. *Provenance and trust*

The openness of the Web and the ease of combining linked data from different sources create new challenges, too. Systems that consume linked data must evaluate quality (such as accuracy, timeliness, reliability) and trustworthiness of the data. Most of the previous works that analyze the provenance information have focused on the voiD (Vocabulary Of Interlinked Datasets). The voiD is a vocabulary and a set of instructions that enables the discovery and use of linked datasets. It allows one to describe datasets and linksets, and enables a number of tasks to be automated in a scalable manner [32, 33, 34, 35].

E. *Data quality and link maintenance*

Datasets in the LOD cloud are far from being static in their nature and how they are exposed. As resources are added and new links are set, applications consuming the data should be able to deal with these changes. So, an important issue in maintaining the quality of data published as linked data is to update this data as well as the existing links between the data items. In order to have a successful update process, it is required to consider the type of published data, rate and frequency of data changes when adjusting the update interval [10, 32, 37, 38].

There are very few works addressing the publishing of non-English data on the Linked Open Data Web, especially for non-Latin alphabet. This research has been initiated to tackle this issue and to highlight the added values that can be concluded from the classification of challenges. It also provides some solutions for the challenges of publishing linked data, especially Persian linked data.

TABLE I. CLASSIFICATION OF PREVIOUS WOKS

publishing and linking data	publishing data of different domains	[3, 10, 11, 13, 14, 22]
	converting different formats into RDF	[15, 16, 36]
	publishing non-English data	[36]
co-referencing and data fusion		[7, 8, 9, 17, 18, 19, 20, 21, 22, 23]
linked data applications and tools		[24, 25, 26, 27, 28, 29, 30, 31]
provenance and trust		[32, 33, 34, 35]
data quality and link maintenance		[10, 32, 37, 38]

IV. FUM-LD PROJECT

In this section, the process of publishing FUM-LD is briefly described in four steps as follows.

A. *Selecting Target Data*

Different educational and organizational web-based systems are being used at Ferdowsi university of Mashhad. Currently, these systems store their data in relational databases and publish parts of this data on the Web, using traditional approaches. After studying the FUM database, five important entities are selected consisting of faculties, departments, professors, papers (published by professors) and courses. TABLE II. shows the numbers of entities in FUM database which are selected to be published as linked data.

TABLE II. NUMBER OF ENTITIES IN FUM-LD DATASET

Entity	Count
Faculty	15
Department	89
Professor	845
Paper	9777
Course	5834
Total	16560

B. *Assigning URIs*

An important step in publishing a dataset as linked data is designing a URI schema for addressing entities that are to be published. In FUM-LD, a simple schema is used for this purpose:

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

where TYPE is one of the strings 'faculties', 'departments', 'profs', 'papers' and 'courses' based on the type of the entity, and ID is the unique identifier of the entity in the database. For instance, http://wtlab.um.ac.ir/linkeddata/profs/kahani describes the resource corresponding to Mohsen Kahani.

C. *Publishing Data*

An overview of FUM-LD framework is shown in Fig. 1. It is implemented in Java and consists of a repository and three core applications briefly introduced in the following subsections:

- RDFizer for generating RDF representation of the entities
- RDF2HTML for converting RDF representation of the entities to HTML
- vioDGenerator for creating vioD specification of FUM-LD

1) *RDFizer*

RDFizer extracts data from FUM relational database and creates an RDF file for describing each entity, and stores it in the repository. Different vocabularies are used in describing resources: FOAF[1] is used for describing personal information of professors and their social network (including other professors who are members of the same faculty and department). Dublin Core[2], BibTeX[3], and

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

[Downloaded from journal.itrc.ac.ir on 2024-10-12]

MarcOnt [4] are used for describing publications of professors. SKOS [4] 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 that should be linked to the English datasets, it is required to have a mechanism for finding English equivalents of the Persian terms, when trying to search external datasets for possible links. This mechanism is provided through two simple solutions: First, for many important concepts (e.g. professor names, paper titles) there are two distinct columns in the tables of the FUM database, one column contains the Persian term (e.g. professor name in Persian), and the other column contains the English equivalent (e.g. the professor name in English). The second solution is using a local dictionary to find appropriate equivalent of the required terms. After finding the English term, the application automatically searches the external dataset for the 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.

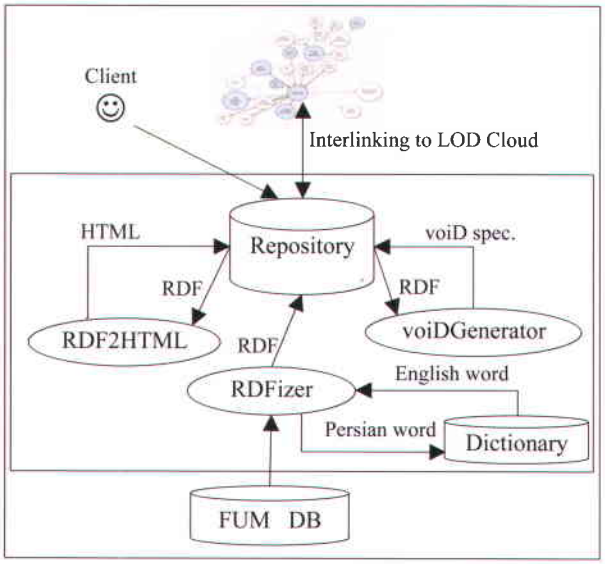

Figure 1. FUM-LD Framework

2) 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.

3) voiDGenerator

The framework uses void [1] 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, such as trust, searching, ranking and selecting datasets [1, 5].

[4] http://www.marcont.org/

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 linksets of the dataset.

D. Interlinking Data Resources

Currently, in most linked data publishing projects, interlinks between web datasets are generated entirely automatically, using heuristics to determine when two resources in two datasets identify the same object ([1, 3]). Providing links to other resources inside and outside the FUM-LD is an important issue in publishing this dataset and the RDFizer is responsible for generating such links.

1) Linking to Other Resources

Resources in FUM-LD are automatically linked to different LOD datasets. The faculty and department titles and course names are linked to related resources in DBpedia with owl:sameAs links. Countries, provinces and cities of the faculties and departments are linked to Geonames dataset by foaf:based_near predicate. Courses are linked to related terms in OpenCyc. Professors and their publications are linked to equivalent resources in DBLP and ACM. TABLE III. shows some statistics about these links.

TABLE III. SOME STATISTICS OF THE FUM-LD DATASET

Link set	Description	Count
1	Links to DBpedia Resources	4570
2	owl:sameAs links to DBpedia	1311
3	owl:sameAs links to DBLP	475
4	owl:sameAs links to ACM	38
5	skos:subject links to DBpedia	3708
6	skos:subject links to OpenCyc	449
7	Links to GeoNames resources	936

2) Interlinking to FUM-LD

In addition to links to external datasets, there are some internal links between different resources in the FUM-LD. For instance, each professor is linked to courses he/she teaches. As shown in Fig. 2 there are five different subsets in FUM-LD. This figure shows existing links between these datasets. This interlinking helps user to browse the dataset easier.

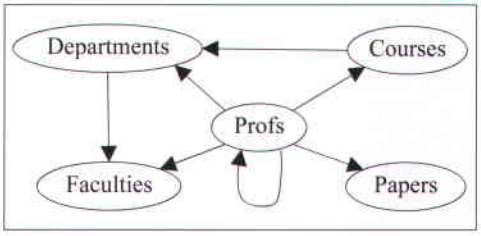

Figure 2. Interlinking of FUM-LD dataset

The total number of RDF triples in dataset is 317916, and there are 845 foaf:Person resources described. The FUM-LD consists of 5 subsets: Faculties, Department, Courses, Profs, and

[Downloaded from journal.itrc.ac.ir on 2024-10-12]

Papers. TABLE IV. shows the number of links between these subsets.

V. Analysis of the Experimental Results

In this section, some discussions about the experimental results mentioned in the previous sections are presented. The main goal of this section is to answer the questions: in which contexts, and due to which reason, the proposed framework performs successfully?

First, the success indicators for such a framework should be defined. Unfortunately, there is not yet a standard method available for assessing the quality of a linked dataset, and thereof for the evaluation of the frameworks that publish such data. However, similar works have mainly focused on measures such as the number of RDF triples, or the number of links to external datasets, to indicate the quality of their work [5, 18, 26, 29].

Here, we discuss the success of the FUM-LD project from these two points of views.

TABLE IV. Links between different subsets of FUM-LD

Source subset	Target subset	Number of links	Link type example
Profs	Profs	15150	foaf:knows
Profs	Faculties	845	foaf:member
Profs	Departments	845	foaf:member
Profs	Courses	15447	dmcNS:teaches[5]
Profs	Papers	13110	foaf:maker, dc:creator
Courses	Departments	17502	dbpprop:reference[6]
Departments	Faculties	174	dc:ispartof

It seems that the number of triples generated is not an important indicator for the quality of a linked dataset; It is only an indication of the amount of the contribution of that dataset to the whole linked data cloud. In the other words, a large dataset which includes huge number of RDF triples simply makes more data available as linked data. However, the huge number of RDF triples, necessarily does not say anything about how much these triples are precise or the links are valid. It should be noted that the number of RDF triples generated when publishing a dataset as linked data, is mainly dependent on the size of the source dataset (in terms of the number of entities and facets), and the decision that which parts of the source dataset can be published as linked data. Therefore, it is not dependent on the publishing framework or the procedure, and it cannot necessarily be concluded that the more RDF triples are generated the more successful is the publishing method.

The number of RDF triples generated in FUM-LD project is presented at the end of the Section IV. Simply based on the number of RDF triples one cannot decided on success or failure of the FUM-LD project. Obviously if it was decided to publish data of Student entities (in addition to the other 5 entity types) as linked data, it would have increased the size of the output RDF dataset.

From the point of view of the contribution to the whole linked data cloud, FUM-LD is the first linked dataset which provides academic data of an Iranian university as Persian linked data. It enables linked data consumers to easily access the publicly available data of the Ferdowsi University of Mashhad. This data, if not published as linked data, had very low accessibility since it should have been extracted from HTML pages of the FUM web pages with much more complexity and cost, and lower quality. Therefore, the success of the framework can be concluded from the improved accessibility and enhanced presence of Iranian academic data on the web.

Next we discuss the issue of success from the point of view of the number and quality of the links.

When publishing a dataset as linked data, the more links are provided from the dataset to other external dataset, the more amount of data is reachable by starting from that dataset, the dataset is more strongly linked to the data on the linked data cloud. Obviously, this is very important since it is directly related to the inherent goal and the value of the linked data. In the project of publishing FUM data as linked data, it was decided to develop a complete and specialized framework, instead of using existing tools. The reason is mainly due to the fact that having a framework specially developed for publishing academic data, provides more power and flexibility, since the framework is extensible and different heuristics and idea can be utilized to enhance the framework to use more techniques and algorithms for finding more links. If an off-the-shelf product, e.g. D2R Server, was used, we would have been limited to its limitations. Since, to the best of our knowledge, none of existing solutions deal with Persian data, it was a wise choice to develop the framework.

The success of the framework in finding links to the external datasets depends on two factors: the quality of the data in the original dataset, and the strength of the search methods employed for finding the links. If the original dataset has quality problems like missing values, inconsistent data and low-quality data, it directly affects the success of the proposed framework.

The search methods employed in the FUM-LD framework are specifically designed for publishing an academic dataset. Therefore, reusing this framework in other domains requires some modifications in terms of customizing the search algorithms and SPARQL queries used.

The number of links produced by the proposed framework are presented in TABLE III. and TABLE IV. In order to measure the quality of these links, it is reasonable to use precision and recall metrics often used in information retrieval domain.

A manual evaluation identified that the owl:sameAs links found by the framework between professor entities in FUM-LD dataset and their equivalents in the external datasets (i.e. DBLP, ACM) has the precision of 100%. The reason is the strict

[5] http://devel.patrickgmj.net/dmcNS

[6] http://dbpedia.org/property#

[Downloaded from journal.itrc.ac.ir on 2024-10-12]

search algorithm that was carefully customized so that it doesn't not have false positive.

Calculating the value of recall is much more difficult and it requires more investigation to be performed. But initial analysis indicated that some existing links have not been detected by the framework, mainly due to data quality problems in the FUM database. For instance, for some courses, the English name of the course is not entered in the database. Such missing values directly affect the recall. Another example is related to the difficulties with names of the professors. Since the search for finding papers of a professor in the external datasets requires that the URI of the professor herself be found in that external dataset by searching her name, difficulties with names affects the number of links found. For instance, some Persian names were found in the database that had a prefix of 'سید' or 'سیده'. Although their equivalent English name in the database had a similar prefix (e.g. "Seyyed" or "Sayed"), but no equivalent entity was found in external datasets by the framework, although it did exist. The reason was that the prefix was omitted from the name of the professor in the external dataset. Another example is related to multi-word names. For instance a name like "Ali Mohammad Zadeh Tehrani" from FUM database could not be matched to its equivalent "A. M. Z. Tehrani" in the external dataset[7]. Although it is possible to improve the search algorithm to consider such cases, it must be noted that number of such issues is not very small.

For datasets that data is entered under some systematic control, or by users who are aware of the data quality issues, the quality of the dataset will be acceptable and the framework will perform successfully. In contexts where data is freely generated by the end users, with no special validation and control, the success of the framework degrades. The more quality problems has the original dataset, the more data cleansing and sophisticated search methods must be employed by the framework.

Another issue is the availability of the related ontologies. When publishing a dataset as linked data, it is crucial to decide about ontologies that must be used for describing data. In contexts where appropriate ontologies do not exist, or the existing ones are not so mature and well-known, the result of applying the framework is not so promising. If the published data uses unknown and poor ontologies, its value reduces, since it is not interesting from the point of view of linked data consumers, and it is not discoverable by other publishers.

Similar issue is related to the availability of related external datasets to be linked to. When publishing a dataset of a domain that an appropriate external dataset within the same domain does not exists, the published dataset is not much linked to the linked data cloud, and it is somehow isolated. This was experienced in the FUM-LD project. For instance, for some faculties and departments (e.g. faculty of Theology, or Hadith Science department) due to their subject and domain, it was not possible to find an appropriate external dataset to link to.

[7] It must be noted that due to privacy issues, the actual name is not mentioned here, and a similar but artificial name is used.

VI. Challenges and Solutions

Different problems and challenges are identified during FUM-LD project. Here, we discuss these problems and recommend some solutions to publishers. We begin with issues relating to linking and accessing other datasets; then we discuss data problems, Persian language challenges, and maintenance challenges.

A. Linking challenges

Some problems of publishing datasets as Linked Data are aroused when linking the dataset to other datasets. Here, we discuss some of these challenges.

1) Choosing appropriate ontologies and predicates

An important issue in publishing linked data is to decide which ontologies and predicates should be used to describe the resources.

The most common solution is to select ontologies based on their popularity. Some ontologies have become the de facto standard in specific domains (for instance FOAF for personal information, or Dublin Core for information about publications).

Most of the vocabularies used for data publishing are based on known ontologies, with extensions that partly borrow from existing vocabularies and partly reside in a new vocabulary [16].

Although having good knowledge about well-known ontologies related to the domain of the dataset eases this decision making, but there are two problems in this regards: first, this popularity-based approach is not effective for all cases (e.g. for domains which there is no well-known ontology), and second, there is not any automatic approach to systematically identify and evaluate candidates.

To reach a maximum level of interoperability, a data source should aim to adhere to the commonly accepted vocabularies, as much as possible. Publishers looking for ontologies to incorporate them into their systems, just use their experiences and intuitions. This makes it difficult for them to justify their choices. Mainly, this is due to the lack of methods that help them to determine the most appropriate ontologies for describing their data.

In [40], ONTOMETRIC method is proposed, which allows the users to measure the suitability of existing ontologies, regarding the requirements of their systems.

Since the RDF semantics allows to arbitrarily mixing different, unrelated vocabularies, a method is presented in [32] which uses a custom vocabulary to model file system data and adds some information from popular vocabularies like Dublin Core and FOAF where they fit.

The approach used in the FUM-LD project is an ad-hoc one. For domains which there is a de facto standard ontology (e.g. FOAF, or Dublin Core), it is

[Downloaded from journal.itrc.ac.ir on 2024-10-12]

chosen. Otherwise, when there is no such ontology, a subjective semi-automatic approach is used to find the required ontology. First Swoogle semantic search engine is used to manually search for ontologies that contain the main concepts of the domain of interest. Then, for each of the top-5 resulting ontologies, a search is performed to estimate the popularity of that ontology on Linked Data space. To do so, a number of SPARQL queries are executed on the LOD SPARQL endpoint to see how many times the predicates of that ontology are used in the LOD cloud. The most common used ontology is then selected as the most appropriate one. As an example, when describing the data of professors in FUM-LD, a predicate was required to specify that a special course is taught by a professor. Using the approach described above, the predicate 'teaches' from dmcNS was selected.

Considering how ontologies are selected, it is not easy to evaluate quality of published data. Also using approaches which involve manual repetitive activities and subjective judgment increases the publication cost and spent-time and decreases the accuracy and quality of the results, especially for large, dynamic, and complex datasets. Therefore, one of the challenges of linked data is lack of a standard well-defined approach for choosing required ontologies and predicates.

2) Creating appropriate links between data

The task of linking data to external resources can be accomplished by tools that provide this functionality for generic Web resources, which usually apply various heuristics to detect semantically related resources (e.g., shared identifiers or object similarity [42]). These heuristics depend on the information that is available for a particular entity. For example, in [32], it depended on the raw data of the files, and also the data provided by their metadata extraction components. This information is then used as a basis for interlinking.

In another work, Silk-LSL (Link Specification Language) is used to express heuristics for deciding whether a semantic relationship exists between two entities [41]. The language is also used to specify the access parameters for the involved data sources. The <LinkCondition> section is the heart of a Silk link specification and defines how similarity metrics are combined in order to calculate a total similarity value for a pair of entities.

Heterogeneity in the schema-level makes an obstacle for automated discovery of co-reference resolution links between individuals. Although there is a multitude of existing schema matching techniques, the Linked Data environment differs from the standard scenario assumed by these tools. In particular, large volumes of data are available, and repositories are connected into a graph by instance-level mappings. In [23] authors utilized these features to produce schema-level mappings which facilitate the instance co-reference resolution process.

To overcome resource discovery issues, [43] suggested employing a number of third party services. Semantic Web search engines, such as Sindice.com, index Linked Data resources that are found by link-traversing spiders or bots. Such services may be able to return a list of entities in which a given URI exists. However, functionality varies between the services and each may require a different access mechanism. As a result, these services must be used with caution and careful attention.

In this project, through analyzing FUM database, and browsing related linked data sets like DBpedia and ACM, and performing some manual schema matching, some heuristics are found for logic of linkage. Based on these heuristics, a link discovery procedure is developed inside RDFizer which uses a string matching algorithm with an experimentally adjusted threshold. During experiments, a number of such algorithms, implemented in SimMetrics tool[8], are studied and Levenshtein, JaccardSimilarity and CosineSimilarity algorithms are selected as candidate. So, 12000 pairs are compared using three algorithms with 6 different threshold values. Then the results are evaluated by members of the team. Finally for each one, four metrics of true positive, true negative, false positive, and false negative are calculated. About 7500 pairs from 12000 pairs are related to the names of persons, and others are related to the titles of papers. Result of this experimental phase is presented in the Appendix.

After analyzing these experimental results, it was decided to use Levenshtein algorithm for the string matching phase with different thresholds: value of 0.8 for matching title of papers and value of 0.9 for names of persons.

It can be concluded that the process of link discovery, and especially determining the logic of linkage require expertise, detailed understanding of the dataset at hand, as well as familiarity with external datasets and ontologies.

B. Data Challenges

One challenge in publishing a dataset as linked data is lack of required data in the original dataset. Even in cases that such information exists in the database, incomplete and incorrect data are entered.

In [11] for linking UK government data ,authors had to deal with different sets of information about a given resource updating at different times, and also information from different sources, modified at different times, potentially overlapping with each other. For example, a school name might be recorded in five different databases, all exposed as linked data, and updated at different intervals. These considerations have led them to adopt named graphs as a mechanism for annotating sets of statements with information about their validity over time, their authoritativeness, and other named graphs in the same series. While many sources may provide information about a given resource, only one should provide authoritative information about a particular property of that resource, such as the school's name.

In our experience, we met similar problems. For instance, for some papers, data about abstract or keywords, or list of coauthors does not exist in the table of papers in the database. Different types of

[8] http://sourceforge.net/projects/simmetrics

[Downloaded from journal.itrc.ac.ir on 2024-10-12]

formats are used for entering date values (e.g. date of a conference). Also, there were Persian data in columns that should contain English data, or vice versa (e.g. there are 2 columns for storing names of professors, one for Persian, and the other for English, but English column contains Persian data). In systems such as professor portals, where data is not considered as important operational data, and it is left to the end-users to freely enter their data, such problems of low-quality or missing data lead to challenge when it comes to linking resources to related ones in external datasets.

To address this challenge, it is required to precisely analyze original data and identify existing problems, and then use the data cleansing techniques or customized ad-hoc solutions to fix the problems as much as possible. For instance, it is possible to implement algorithms to convert different formats of dates to a unique format, or to move Persian values from English columns to the corresponding Persian columns. Unfortunately, such customized solutions are specific to the dataset at hand, and have low reusability in terms of publishing datasets of a different domain. In addition to such data cleansing solutions, it is possible to use linked data itself to identify appropriate values for missing data. For instance, after linking a resource of type paper from FUM-LD to its corresponding resource in DBLP, it is possible to extract names of coauthors (or other attributes, e.g. keywords) from DBLP and add them to the specification of that resource in FUM-LD.

C. *Persian Language challenges*

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 mother tongue language 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 equivalences of a Persian name in English. Using this dictionary, it is possible to use all equivalences of a professor name, when searching external datasets. Therefore, the probability of missing a related link because of different spelling is reduced.

D. *Data and Link Maintenance*

As the Web of Linked Data expands, it will become increasingly important to preserve data and links such that the data remains useful. Updates in either of the interlinked datasets can invalidate existing links or imply the need to generate new ones.

In the other hand, an important issue in maintaining the quality of data published as linked data is to update this data as well as the existing links between the data items. When updating the dataset, information about the time of creation and modification of data is published along the dataset. Predicates like dcterms:created and dcterms:modified can be used to store such information in voiD specification of the dataset. In order to have a successful update process, it is required to consider the type of published data, rate and frequency of data changes in adjusting the update interval.

WOD-LMP (Web of Data - Link Maintenance Protocol) is a solution proposed in [42]. The WOD-LMP protocol automates the communication between two cooperating Web data sources. It assumes two basic roles: Link source and link target, where the link source is a Web data source that publishes RDF links pointing at data published by the target data source.

In [44] authors present a method for locating linked data to preserve which functions even when the URI the user wishes to preserve does not resolve (i.e. is broken/not RDF) and an application for monitoring and preserving the data. Their idea is based upon the principle of adapting ideas from hypermedia link integrity in order to apply them to the Semantic Web. They have also introduced a simple expansion algorithm which can be used to retrieve linked data about a URI even when that URI is not resolvable. This provides a tool for preserving data in the Semantic Web and recovering from data loss.

Generally speaking, there are two main situations that require updating the dataset:

1. The original dataset is changed. For instance, in case of FUM-LD project, if a new professor joines a department, new resources of types professor and paper should be added to the dataset, new internal links of type foaf:knows should be created between this professor and his colleagues (professors of the same department), new links might be available for linking these new resource to other resources in external datasets, for instance linking the new professor to a resource in ACM using owl:sameAs link.
2. A related external dataset is changed. Similar to the original dataset, external datasets might also change by introducing new resources or links. If the original dataset is linked to such an external dataset, it requires to be updated. For instance, if a new resource describing 'Computer Engineering Department of Ferdowsi University o Mashhad' is added to DBpedia, then it is a good candidate to be linked by the resource which describes the same thing in FUM-LD.

[Downloaded from journal.itrc.ac.ir on 2024-10-12]

If external datasets specify their last modification timestamp (e.g. in their voiD specification), then publishers of the original dataset are able to decide when to update their dataset. If an external dataset is updated monthly, all the links to this dataset should be updated monthly.

Therefore, from the point of view of the consumers, it is required that the times of creation and last modification of the dataset are specified to help them judge about the trustworthiness and validity of data. So, timestamps in four granularity levels can be used for this reason:

1. Original dataset level: it is possible to use a timestamp for the whole dataset to specify its creation and last modification date/time.
2. External dataset level: timestamps can be used for each of the external datasets that the original dataset is linked to. For instance, in FUM-LD project, it is possible to specify in voiD specification the last date of linking FUM-LD to ACM dataset. Therefore, a user who is following a link from a FUM-LD resource to related ACM resource knows when this link was created, and then can have a sense of validity of the link.
3. Resource-level: timestamps can be attached to each of the resources, to specify when it was created or modified.
4. Triple-level: at the lowest level, timestamps can be assigned to each triple, providing information about when it was created.

Based on the chosen granularity level, overhead of using timestamps varies. Also the possible update level varies.

At the topmost level, only 2 triples are required to specify the creation and last modification timestamp of the whole dataset, while at the lowest level, each triple is accompanied by one extra triple (if only last modification timestamp is used). Therefore, the lower the granularity level, the more space is used for the timestamps. If the dataset is finally published using a triple store, then from a query execution point of view, it is not a good idea to fill the triple store with too many timestamp triples that might have no use in query answering.

Using the topmost level, it is only possible to update the dataset as a whole, since the timestamps are used at the whole dataset, while using triple-level timestamps, it is possible to update triples independently. Therefore, the lower the granularity level, the more flexible the update process is.

In FUM-LD project, the second level is used, i.e. timestamps are used at external dataset level.

VII. Conclusion and Future Work

In this paper, some problems and challenges of publishing Persian linked data are discussed based on our experience, "FUM-LD" project. By analyzing the empirical results of this project, some publisher-oriented approaches are proposed to improve the quality of the linked data.

Since, the main focus of this project is on publishing data of Ferdowsi University of Mashhad, 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. To achieve this goal, the framework should be extended in different aspects.

It is needed to define an academic data model which includes all the important entities for publishing academic data and specifies the relationships of these entities. Most of published data on LOD are transformed from enterprise databases which contain public internal and external information including organization's assets, equipments, facilities, locations, partners, customers, and stakeholders.

An important issue in publishing linked data is deciding which data should be published. There are two considerations for selecting data:

- Data should be open
- Data should be related to the domain of interest

As for the academic data, we should extend our framework by adding an academic data model as an input of FUM-LD framework. Proper academic data sources should be studied and important entities that usually exist in any academic institute should be selected. After identifying main entities, the relationships of these entities should be defined as a metadata of framework.

A common prerequisite for 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. To improve quality of published data, a set of requirements can be defined in terms of measurable data characteristics and a Validator to perform measurements for ensuring that input data conforms to the data model and these requirements.

Evaluating data quality needs a data quality model to be defined. Therefore, it involves determining a set of data characteristics in accordance with the predefined academic data model.

By defining a data model for publishing academic data, developing a data quality model and a Validator which measures the conformance of the input data with data model and data quality characteristics, one can ensure that the published data would have a desired quality.

[Downloaded from journal.itrc.ac.ir on 2024-10-12]

APPENDIX: RESULTS OF DIFFERENT STRING MATCHING ALGORITHMS

Algorithm	Threshold	Paper titles			Professor names		
		No. of pairs	True positive (%)	False positive (%)	No. of pairs	True positive (%)	False positive (%)
CosineSimilarity	0.6	368	0.1576	0.8424	248	0.5806	0.4194
	0.7	362	0.1547	0.8453	202	0.6089	0.3911
	0.8	354	0.1469	0.8531	194	0.6340	0.3660
	0.85	343	0.1195	0.8805	110	0.8909	0.1091
	0.9	335	0.0985	0.9015	110	0.8909	0.1091
	0.95	309	0.0324	0.9676	110	0.8909	0.1091
JaccardSimilarity	0.6	57	0.9298	0.0702	183	0.6721	0.3279
	0.7	47	0.9362	0.0638	99	0.9899	0.0101
	0.8	37	0.9189	0.0811	99	0.9899	0.0101
	0.85	18	1.0000	0.0000	99	0.9899	0.0101
	0.9	11	1.0000	0.0000	99	0.9899	0.0101
	0.95	10	1.0000	0.0000	99	0.9899	0.0101
Levenshtein	0.6	71	0.8169	0.1831	1649	0.0988	0.9012
	0.7	64	0.9063	0.0938	778	0.1838	0.8162
	0.8	57	1.0000	0.0000	385	0.3221	0.6779
	0.85	56	1.0000	0.0000	317	0.3817	0.6183
	0.9	54	1.0000	0.0000	213	0.5681	0.4319
	0.95	49	1.0000	0.0000	95	0.9895	0.0105

REFERENCES

[1] K. Alexander, R. Cyganiak, et al. "Describing Linked Datasets", Linked Data on the Web workshop (LDOW2009). Madrid, Spain.

[2] T. Berners-Lee, (2006), "Linked Data. Design Issues for the World Wide Web" http://www.w3.org/DesignIssues/LinkedData.html.

[3] Y. Raimond, C. Sutton, et al. (2008). "Automatic Interlinking of Music Datasets on the Semantic Web". In International Workshop on Linked Data on the Web (LDOW 2008). Beijing, China.

[4] SKOS, Semantic Web Deployment Working Group. SKOS Simple Knowledge Organization System Reference.http://www.w3.org/TR/swbp-skos-core-spec/,2008.

[5] N. Toupikov, J. Umbrich, et al. (April 20th, 2009). "DING! Dataset Ranking using Formal Descriptions". Linked Data on the Web workshop (LDOW2009). Madrid, Spain.

[6] S. Paydar, M. Kahani, et.al, "Publishing Data of Ferdowsi University of Mashhad as Linked Data", International Conference on Computational Intelligence and Software Engineering (CiSE 2010), 2010.

[7] M. Rowe, "Interlinking distributed social graphs," in Proceedings of WWW 2009 Workshop on Linked Data on the Web, 2009

[8] Y. Liu and F. Z. Schar, C,, "Towards practical rdf datasets fusion," in Workshop on Data Integration through Semantic Technology (DIST2008), ASWC2008 Bangkok, Thailand, 2008.

[9] S. Melnik, H. Garcia-Molina, and E. Rahm, "Similarity ooding: A versatile graph matching algorithm," in In 18th International Conference on Data Engineering (ICDE), 2002, pp. 117-128.

[10] A. Hogan, A. Harth, A. Passant, pp. Decker, and A. Polleres, "Weaving the Pedantic Web," Proceedings of the Linked Data on the Web WWW2010 Workshop (LDOW 2010), Raleigh, North Carolina, USA: 2010.

[11] G. Correndo, M. Salvadores, Y. Yang, N. Gibbins, N. and Shadbolt, (2010) Geographical Service: a compass for the Web of Data. In: Linked Data on the Web (LDOW2010), 27 April 2010, Raleigh, North Carolina, USA.

[12] J. Sheridan, J. Tennison, (2010) Linking UK Government Data. In: Linked Data on the Web (LDOW2010), 27 April 2010, Raleigh, North Carolina, USA.

[13] M. Hausenblas, R. Troncy, T. Buerger, Y. Raimond, (2009) Interlinking Multimedia: How to Apply Linked Data Principles to Multimedia Fragments. In: Linked Data on the Web (LDOW2009), 20 April 2009, Madrid, Spain.

[14] O. Hassanzadeh, M. Consens, (2009) Linked Movie Data Base. In: Linked Data on the Web (LDOW2009), 20 April 2009, Madrid, Spain.

[15] R. Garcia, R. Gil, (2009) Publishing XBRL as Linked Open Data. In: Linked Data on the Web (LDOW2009), 20 April 2009, Madrid, Spain.

[16] R. Cyganiak, S. Field, A. Gregory, W. Halb, J. Tennison, (2010) Semantic Statistics: Bringing Together SDMX and SCOVO. In: Linked Data on the Web (LDOW2010), 27 April 2010, Raleigh, North Carolina, USA.

[17] A. Nikolov, V. Uren, E. Motta, (2009) Towards Data Fusion in a Multi-ontology Environment. In: Linked Data on the Web (LDOW2009), 20 April 2009, Madrid, Spain.

[18] P. Bouquet, H. Stoermer, and B. Bazzanella. An Entity Name System (ENS) for the Semantic Web. In 5th Annual European Semantic Web Conference (ESWC 2008), pp. 258-272, 2008.

[19] A. Ferrara, D. Lorusso, and S. Montanelli. Automatic identity recognition in the Semantic Web. In Workshop on Identity and Reference on the Semantic Web, ESWC 2008, Tenerife, Spain, 2008.

[20] A. Nikolov, V. Uren, E. Motta, and A. de Roeck. Integration of semantically annotated data by the KnoFuss architecture. In 16th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2008), Acitrezza, Italy, 2008.

[21] H. Glaser, I. Millard, A. Jari, T. Lewy, and B. Dowling. On coreference and the Semantic web. 7'h International Semantic Web Conference (ISWC 2008), Karlsruhe, Germany, 2008

[22] M. Rowe, (2010) Data.dcs: Converting Legacy Data into Linked Data. In: Linked Data on the Web (LDOW2010), 27 April 2010, Raleigh, North Carolina, USA.

[Downloaded from journal.itrc.ac.ir on 2024-10-12]

[23] A. Nikolov, V. Uren, E. Motta, (2010) Data linking: capturing and utilising implicit schema-level relations. In: Linked Data on the Web (LDOW2010), 27 April 2010, Raleigh, North Carolina, USA.

[24] B. Haslhofer, B. Schandl, (2008) The OAI2LOD Server: Exposing OAI-PMH Metadata as Linked Data. In: Linked Data on the Web (LDOW2008), 22 April 2008, Beijing, China.

[25] J. Li, Y. Zhao, (2008) A Case Study on Linked Data Generation and Consumption. In: Linked Data on the Web (LDOW2008), 22 April 2008, Beijing, China.

[26] P. Coetzee, T. Heath, E. Motta, (2008) SparqPlug: Generating Linked Data from Legacy HTML, SPARQL and the DOM. In: Linked Data on the Web (LDOW2008), 22 April 2008, Beijing, China.

[27] C. Zhou, C. Xu, H. Chen, K. Idehen, (2008) Browser-based Semantic Mapping Tool for Linked Data in Semantic Web. In: Linked Data on the Web (LDOW2008), 22 April 2008, Beijing, China.

[28] M. Bergman, F. Giasson, (2008) zLinks: Semantic Framework for Invoking Contextual Linked Data. In: Linked Data on the Web (LDOW2008), 22 April 2008, Beijing, China.

[29] S. Davies, J. Hatfield, Ch. Donaher, J. Zetiz, (2010) User Interface Design Considerations for Linked Data Authoring Environments. In: Linked Data on the Web (LDOW2010), 27 April 2010, Raleigh, North Carolina, USA.

[30] A. Latif, M. T. Afzal, D. Helic, K. Tochtermann, H. Maurer, (2010) Discovery and Construction of Authors' Profile from Linked Data (A case study for Open Digital Journal). In: Linked Data on the Web (LDOW2010), 27 April 2010, Raleigh, North Carolina, USA.

[31] M. Stankovic, C. Wagner, J. Jovanovic, P. Laublet, (2010) Looking for Experts? What can Linked Data do for You? In: Linked Data on the Web (LDOW2010), 27 April 2010, Raleigh, North Carolina, USA.

[32] J. Zhao, G. Klune, D. Shotton, (2008) Provenance and Linked Data in Biological Data Webs. In: Linked Data on the Web (LDOW2008), 22 April 2008, Beijing, China.

[33] O. Hartig, (2009) Provenance Information in the Web of Data. In: Linked Data on the Web (LDOW2009), 20 April 2009, Madrid, Spain.

[34] K. Alexander, R. Cyganiak, M. Hausenblas, J. Zhao, (2009) Describing Linked Datasets. In: Linked Data on the Web (LDOW2009), 20 April 2009, Madrid, Spain.

[35] N. Toupikov, J. Umbrich, R. Delbru, M. Hausenblas, G. Tummarello, (2009) DING! Dataset Ranking using Formal Descriptions. In: Linked Data on the Web (LDOW2009), 20 April 2009, Madrid, Spain.

[36] J. Neubert, (2009) Bringing the "Thesaurus for Economics" on to the Web of Linked Data. In: Linked Data on the Web (LDOW2009), 20 April 2009, Madrid, Spain.

[37] J. Umbrich, M. Hausenblas, A. Hogan, A. Polleres, S. Decker, (2010) Towards Dataset Dynamics: Change Frequency of Linked Open Data Sources. In: Linked Data on the Web (LDOW2010), 27 April 2010, Raleigh, North Carolina, USA.

[38] O. Hartig, (2009) Provenance Information in the Web of Data. In: Linked Data on the Web (LDOW2009), 20 April 2009, Madrid, Spain.

[39] J. Zhao, G. Klune, D. Shotton, (2008) Provenance and Linked Data in Biological Data Webs. In: Linked Data on the Web (LDOW2008), 22 April 2008, Beijing, China.

[40] A. Lozano-Tello, G. Perez, (2004) OntoMetric: A Method to Choose the Appropriate Ontology. In: Journal of Database Management, Vol. 15, issue 2, pp. 1-18, April-June (2004).

[41] J. Volz, C. Bizer, M. Gaedke, G. Kobilarov, (2009) Silk - A Link Discovery Framework for the Web of Data. In: Linked Data on the Web (LDOW2009), 20 April 2009, Madrid, Spain.

[42] J. Volz, C. Bizer, M. Gaedke, G. Kobilarov, (2009) Discovering and Maintaining Links on the Web of Data. In Proceedings of the 8th International Semantic Web Conference (ISWC 2009), 2009.

[43] I. Millard, H. Glaser, M. Salvadores, N. Shadbolt, (2010) Consuming multiple linked data sources: Challenges and Experiences. In: First International Workshop on Consuming Linked Data (COLD2010), 2010-11-07, Shanghai, PRC.

[44] R. Vesse, W. Hall, L, Carr, (2010) Preserving Linked Data on the Semantic Web by the application of Link Integrity techniques from Hypermedia. In: Linked Data on the Web (LDOW2010), 27 April 2010, Raleigh, North Carolina, USA.

Samad Paydar is currently a Ph.D. candidate in software engineering at computer engineering department of the Ferdowsi University of Mashhad, Iran. He received his B.Sc. in 2005, and his M.Sc. in 2007, both in software engineering from Ferdowsi University of Mashhad, Iran. His research interests include semantic web, linked data, use of the semantic web technologies in software engineering, and ontology engineering.

Mohsen Kahani is currently an associate professor and Chief Information Office (CIO) of Ferdowsi University of Mashhad, Iran. He received his B.E. in 1990, from the University of Tehran, Iran, his M.E. in 1994, and his Ph.D. in 1998 both from University of Wollongong, Australia. His research interests include information technology, software engineering, semantic web and linked data.

Behshid Behkamal is currently doing Ph.D. in Software engineering at Ferdowsi University of Mashhad, Iran. She received her B.Sc. in 1999, from the Ferdowsi University of Mashhad, Iran and her M.Sc. in 2006 from Amirkabir University of Technology, Iran. Her research interests include software quality, data quality and semantic web.

Mahboubeh Dadkhah is currently a M.Sc. Student at computer department of Ferdowsi University of Mashhad. She received her B.Sc. in computer engineering from Ferdowsi University in 2007. Her research interests include semantic web, linked data, social networks, software engineering techniques and methodologies and software testing. Her recent works focused on publishing organizational data as linked data on the web and enriching applications using published linked data.

[Downloaded from journal.itrc.ac.ir on 2024-10-12]