Querying Bibliography Data Based on Linked Data

Yasemin Gültepe*
Department of Computer Engineering, University of Kastamonu, Kastamonu, Turkey.

* Corresponding author. Tel.: 05424051928; email: yasemingultepe@kastamonu.edu.tr
Manuscript submitted February 12, 2015; accepted July 3, 2015.
doi: 10.17706/jsw.10.8.1014-1020

Abstract: Usually, a common data model is used for creating and getting a résumé in web applications. Though different web applications provide the same quality résumé information, they encounter difficulties in analyzing and processing data in their different sources. Linked data technology allows overcoming these problems by integrating the data coming from different sources, linking large, voluminous, and distributed data sets with semantic sources in web of data, and forming an open linked data cloud. This study mainly aims to combine, publish, and explore the semantic information in the academic résumés of scientists/researchers working in universities and/or research establishments by use of linked data. The study deals with the use and exploration of academic résumé information through linked data approach. It was intended to conduct effective SPARQL queries on linked data network via FOAF-Academic, DBLP and Résumé/Curriculum Vitae (CV) ontologies so that different data sources would be integrated.

Key words: Linked data, querying semantic web, semantic web, RDF.

1. Introduction

Today, the results of most academic and scientific studies are available in web environment. Academic and scientific studies/research carried out from past to present provide useful data for future goals. Therefore, the academic/scientific social networks crated by academic researchers are important for scientific cooperation in academy. A need has arisen to create a semantic web where contents and concepts can be associated with one another and people so that virtual data stack that keeps snowballing due to social networks on the internet, spreading electronic journals, and personal publications does not turn into a rubbish, and monopoly in queries ends. Semantic web [1], [2] is a web environment that makes well-defined information and services easily understandable by machines. It is an extension of World Wide Web (WWW) where semantic information is provided and machines and people can work in cooperation.

Linked data is one of the approaches used for obtaining a meaningful integrity by bringing together data groups that are associated with a specific data through forming semantic links between the web pages that make up semantic web. In this way, a data layer that takes data linked to one another as basis is formed on the basis of data and relationships between them.

Linked data is based on RDF (resource description framework) technology [3]. RDF is a common data model for the reuse and sharing of information that provides field independent formal semantics regarding graph sources. Thanks to this model, it is easier for the defined information to be re-accessed when required and to be used by other systems, too. This model is based on the idea of defining the source properties and property values of the objects (sources) in the web environment. RDF data model is based

on graphs composed of triple defined as "subject", "predicate" and "object". It is provided that data be published by use of URIs in a RDF compatible way and relationships between two concepts be defined semantically again through RDF triple. Fig. 1 presents a sample RDF graph representation.

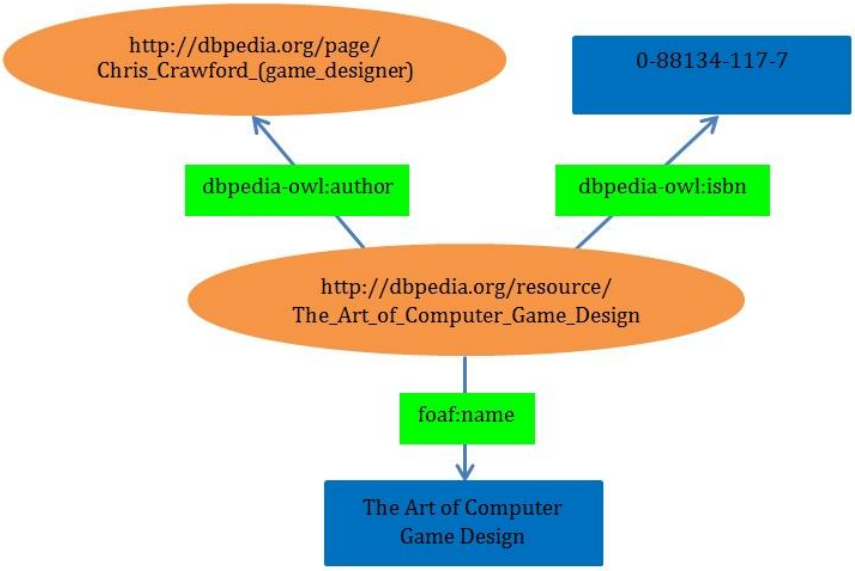

Fig. 1. Sample RDF graph.

In Fig. 1, the book entitled "The Art of Computer Game Design" is indicated with URI http://dbpedia.org/resource/The_Art_of_Computer_Game_Design. Web users can easily access the information in the web environment through browsing according to URI addresses and following relevant RDF links. "dbpedia-owl:author" property is available among book name and author sources. The properties are special sources and define the relationships between sources. The value taken by "dbpedia-owl:author" property is "Chrid Crawford" as it is single-author. This value is the value taken by the properties of sources. It is stated that the number of triple in the data cloud is expressed in billions and increases every passing day [4].

A résumé or a CV is summary information about a person's skills, experiences, educational background, work experience, and so on [5]. A résumé usually defines the work experience and skills of a person. This study mainly aims to combine, publish and explore the semantic information (e.g. information about national and international articles, papers and books) in the academic résumés of scientists/researchers working in universities and/or research establishments by use of linked data. The study deals with the use and exploration of academic résumé information which is semantically associated over linked data cloud.

This paper consists of the following chapters; the second chapter gives information about linked data technology and FOAF-Academic, DBLP and Résumé/Curriculum Vitae (CV) ontologies which are used in this paper because of the new nature of the semantic web technologies used; the third chapter presents SPARQL[1] (SPARQL Protocol and RDF Query Language) query samples used for monitoring the changes in the data in linked data cloud; and the last chapter focuses on results and future works.

2. The Publication of Linked Data

Heath *et al.* [6] showed that the publication of linked data is a three-stage process. The first stage involves providing data with meaning within themselves and making sense of the relationships between data. The second stage involves publishing linked data in a way that allows them to be processed by machines and be

[1] http://www.w3.org/TR/rdf-sparql-query/

combined by using a common standard. The third stage involves establishing RDF links for the defined data source to be linked to the other data sources on the web. RDF links allow users to explore additional data by linking data sets to one another semantically. Some stages need to be fulfilled so that data sources are defined for academic résumés in the web environment through linked data approach and RDF links are established between data sources. These stages are given below in detail.

The publication of the academic and practical works of academic staff in the web environment and the combination of other academic works play an important role in the development of science. Linked data technology can be used for the data in data sources including these works to be obtained and arranged more easily, for these works to be shared more flexibly, and for them to be offered in a way that allows making semantic inferences. For example, the résumé of an academic staff member is defined by the institution s/he works for. The same résumé is available in more than one place in different institutions. That leads to a clumsy system. Data integrity is achieved through linked data approach. By this means, the repetition of common data is prevented.

Semantic web technologies have been used in many works to model and examine bibliographic information. Some ontologies have been developed to model bibliographic information. D'Arcus and Giasson (2009) defined BIBO[2] (The Bibliographic Ontology) in order to create biographical information in the semantic web environment [7]. BIBO is also used as a RDF-format citation ontology or document classification ontology. Other ontologies such as FOAF and Dublin Core can be blended with BIBO terms as local extension. BIBO aims at the definition of concepts about authors and their works and the representation of the relationships between these concepts over a semantic network. Challenger (2012) demonstrated UniGrad ontology as an extension of FOAF in order to represent the concepts underlying university education and the research conducted in universities [8]. In that study, besides the classification of entities in the educational field within a particular context through UniGrad ontology, DBLP ontology was imported to the study in order to represent such information as coauthor relationships or user's publication types. Ontologies can be defined or extended by other ontologies in order to ensure reusability. Therefore, owl:imports tag is used as extension mechanism. This tag includes another source that is to be incorporated into the system. More than one source can be incorporated into the defined ontologies.

The present study suggests the use of the below mentioned data sources for the use and exploration of academic résumé information on linked data basis.

- Frienf of a Friend-Academic[3] *(*AFOAF): AFOAF is an OWL DL-based ontology developed for defining information about academic individuals and their relationships. This ontology is built upon FOAF ontology. FOAF ontology allows establishing a social network by defining people, their activities, and their connections with other people [9]. Kalemi and Martiri [10] developed AFOAF ontology in 2011. In AFOAF file, properties and set of classes are defined to represent the information in academic résumés. Academic staff members working for universities can be categorized in terms of some common characteristics such as fields of interest, relationships, and activities by using AFOAF files. In this way, it becomes easier to distribute and share personal data through other semantic web applications. AFOAF ontology includes 37 classes/concepts and 32 properties/relationships. Information belonging to AFOAF ontology is expressed with the namespaces "foaf" and "afoaf".
- DBLP ontology[4]: DBLP ontology contains information about the publications in the field of computer sciences in the bibliography server and their authors [11]. DBLP is a RDF-based ontology that contains information. It allows accessing the publications of an author. DBLP data can easily be obtained from a

[2] http://bibliontology.com/
[3] http://www.owl-ontologies.com/Afoaf.owl
[4] http://swat.cse.lehigh.edu/resources/onto/dblp.owl

couple of RDF graphs (e.g. author-publication graph, author-journal graph, author-conference graph, coauthor graph). Each one of these graphs is a social network example. DBLP ontology has certain pros and cons. Its pros can be summarized as follows: it is used for free and it contains many conference proceedings. Its cons are, in summary, as follows: incomplete citation information, change in the coverage of subdomains of computer sciences. DBLP RDF file contains 2 classes/concepts and 29 properties/relationships.

- Résumé RDF ontology[5]: Bojars and Breslin (2007) developed Résumé RDF ontology in order to express the semantic information in CVs [5]. This ontology contains information about staff (e.g. information about area of specialization, academic information, information about educational background, information about activities and information about works) and other information. Résumé RDF ontology includes 16 classes/concepts and 73 properties/relationships. Information belonging to Résumé RDF ontology is expressed with the namespaces "cv" and "base". Another work similar to this ontology is DOAC[6] (Description of a Career). It is an ontology proposed by Parada in order to define a CV. DOAC concepts are demographic information about people and information about people's qualifications, skills, or abilities. The advantage of Résumé RDF over DOAC is that it yields more query results.

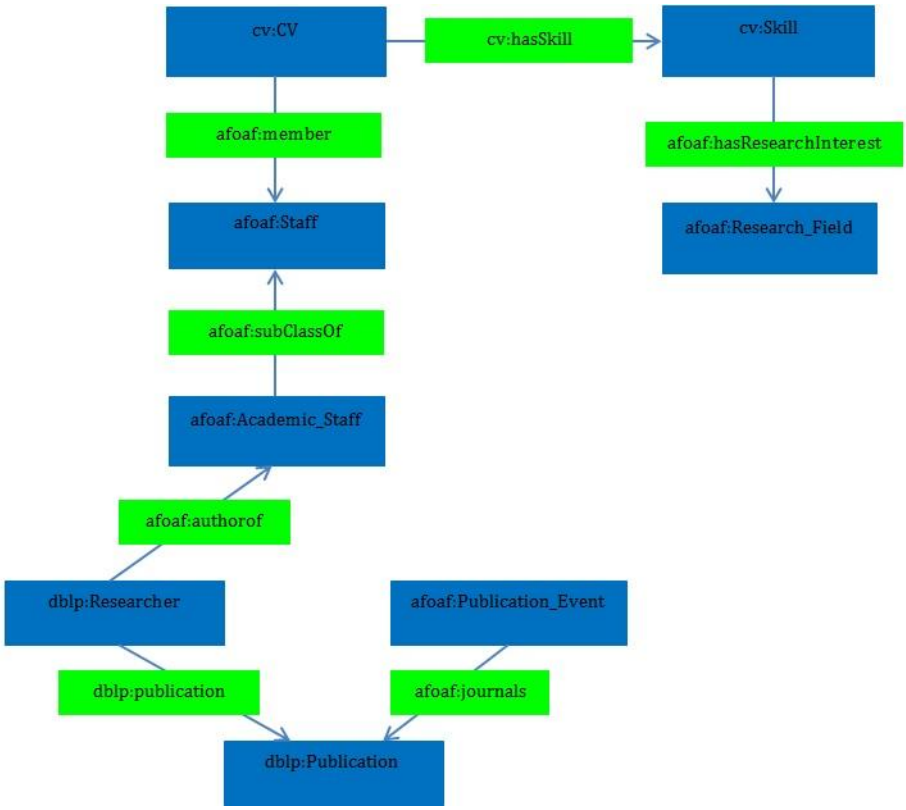

Fig. 2. The relationship among AFOAF, DBLP and RéSUMé RDF ontology.

Fig. 2 presents the concepts included in AFOAF, DBLP, and Résumé RDF ontologies that are used in the defined bibliographic field as well as the relationships between these concepts. In AFOAF ontology, *afoaf:Academic_Staff* class is used for representing academic staff sources. This class is a sub-class of the parent class *afoaf:Staff*. *afoaf:Publications_Event* class that has book, journal and magazine types appears as a sub-class of *afoaf:Event*, which is the most fundamental class for academic events. DBLP ontology can

[5] http://rdfs.org/resume-rdf/
[6] http://ramonantonio.net/doac

provide publication information from different publication events (Book, Journal, Magazine) via *afoaf:journals* relationship that is defined between *dblp:Publication* class and *afoaf:Publication_Event* class belonging to AFOAF ontology. *dblp:publication* relationship between *dblp:Publication* and *dblp:Researcher* classes gives information about publications and their authors. *afoaf:authorof* relationship is used for defining academic individuals as authors in the publications made. AFOAF ontology documents are effective not only in storing fields of interest, relationships, activities and so on belonging to academic staff. It is mainly important because it is capable of establishing relationships between people and their publications. In Résumé RDF ontology, CVs are represented with the namespace "*cv*". The objects affiliated to *cv:Skill* class and *afoaf:Research_Field* class have *afoaf:hasResearchInterest* relationship. *afoaf:member* relationship is established between *cv:CV* class and *afoaf:Staff* class for the arrangement of basic academic works in CV. In this way, many different data sources are reshaped by use of linked data technologies. This shaping exemplifies the use of linked data technologies for data integration in the bibliographic field.

3. Querying Bibliography Data Based on Linked Data

Different data sets can be used for publishing the academic backgrounds of scientists/researchers working in universities and/or research establishments. AFOAF, DBLP, and CV ontologies were used in the present study. A general résumé covers interrelated and implicit information. The use of semantic web technologies allows representing implicit information in an explicit and clear way. Ontology-based academic résumés provide more descriptive information about national and international articles, papers, books and so on. It is possible to query ontologies by using SPARQL that allows making queries flexible and global in order to establish relationships between entities in different data sources [12]. SPARQL is a query language and data access protocol for semantic web. SPARQL has an extensive usage area. A great majority of RDF and OWL query tools provide SPARQL support.

3.1. Sample Queries

SPARQL query tool of Protégé ontology editor was used for querying on the different data sources defined via SPARQL. This chapter of the paper provides two sample queries on the basis of the ontologies introduced in chapter 2.

Query 1: The list of the publications on semantic web by "Tim Berners-Lee".

```
SELECT ?title
WHERE{
    ?publication rdfs:label ?title.
    {?publication rdf:type dblp:Article.}
    UNION {?publication rdf:type dblp:InProceedings.}
    ?publication dblp:author ?person.
    ?person a afoaf:Academic_Staff.
    ?person afoaf:name ?name.
    FILTER (STR(?name)="Tim Bernes-Lee").
}
```

Query 2: The list of staff studying on semantic web in Stanford University.

```
SELECT ?name, ?email
WHERE{
    ?x afoaf:name ?name.
    ?x afoaf:mbox ?email.
    ?x rdf:type afoaf:Academic_Staff.
    ?x cv:employedIn ?z.
```

```
    ?z rdf:name Stanford University.
    ?x cv:hasSkill ?field.
    ?field cv:skillName "semantic web".
}
```

AFOAF, DBLP, and Résumé RDF ontologies were used in order to carry out sample SPARQL queries. Import to system was conducted through the import feature of Protégé 4.3 software. It allows researchers requesting similar results to use it without any query. The first sample query requests to list the "publications on semantic web by Tim Berners-Lee". Another query requests to list the "staff studying on semantic web in Stanford University".

4. Conclusion

In the interoperability of information systems, semantic heterogeneity obtained from the web environment and coming from distributed systems should be taken into consideration. Semantic web is a technology that provides new facilities for the interoperability of information. In the present study, Résumé RDF ontology was used for semantically expressing academic résumé information. Academic résumé contains the semantic information (e.g. information about national and international articles, papers and books) related to the academic backgrounds of scientists/researchers working in universities and/or research establishments. Social network design on the basis of Résumé RDF ontology can be used for integrating different data sources. In the present study, AFOAF and DBLP were used as external data sources. A common language (ontologies) of this sort which is accepted in the whole process may provide an automatic integration among all systems.

Linked data approach can be used for exploring and combining the résumé information in the semantic sources in web of data. Besides facilitating exploring and combining, linked data approach provides the reusability, heterogeneity and interoperability of data among multiple data sources. This study discussed the formation of linked data about academic staff working in universities, CVs in university academic information systems, ontologies and the experiences in data collection and data cleaning processes. In addition, the benefits of using résumé information from data web through linked data applications and using linked data technology were explained. Future works may focus on measuring and assessing the success of linked data applications.

References

[1] Daconta, M. C., Obrst, L. J., & Smith, K. T. (2003). *The Semantic Web, A Guide to the Future of XML, Web Services and Knowledge Management*. Wiley Publisher.

[2] Berners, L. T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American*, *284(5)*, 34-43.

[3] Bizer, C. (2009). The emerging web of linked data. *IEEE Intelligent Systems*, *24(5)*, 87-92.

[4] State of the LOD, Version 0.3. Retrieved 18 February, 2015 from website: http://lod-cloud.net/state/

[5] Bojars, U., & Breslin, J. G. (2007). Resume RDF: Expressing skill information on the semantic web. *Proceedings of the 1st International Expert Finder Workshop.*

[6] Heath, T., Hausenblas, M., Bizer, C., Cyganiak, R., & Hartig, O. (2008). How to publish linked data on the web. *Proceedings of the Tutorial in the 7th International Semantic Web Conference.*

[7] D'Arcus, B., & Giasson, F. (2009). Bibliographic Ontology Specification, Specification Document. Retrieved August 2014, from http://bibliontology.com/specification/

[8] Challenger, M. (2012). The ontology and architecture for an academic social network. *International Journal of Computer Science Issues, 9(2-1)*, 22-27.

[9] Yu, L. (2011). *A Developer's Guide to the Semantic Web*. Springer-Verlag Berlin Heidelberg.

[10] Kalemi, E., & Martiri, E. (2011). FOAF-Academic ontology: A vocabulary for the academic community. *Proceedings of the 2011 Third International Conference on Intelligent Networking and Collaborative Systems* (pp. 440-445).

[11] Aleman, M. B., Decker, S. L., Cameron, D., & Arpinar, I. B. (2008). Association analytics for network connectivity in a bibliographic and expertise dataset. *Semantic Web Engineering in the Knowledge Society*, 188-207.

[12] Hartig, O., Bizer, C., & Freytag, J. C. (2009). Executing SPARQL queries over the web of linked data. *Proceedings of the 8th International Semantic Web Conference* (pp. 293-309).

Yasemin Gültepe is a full-time assistant professor of Computer Engineering Deparment at Kastamonu University, Kastamonu, Turkey, in 2000. She graduated from the Engineering and Architecture Faculty-Computer Engineering Department at Çanakkale Onsekiz Mart University, Çanakkale, Turkey. Afterwards, in 2003, she completed a M.Sc. thesis study title as 'Creating of data warehouse and data mining for university information management systems' in Computer Engineering Deparment of Çanakkale Onsekiz Mart University, Çanakkale, Turkey. In 2011, she obtained her Ph.D. degree in computer engineering deparment of Ege University, Izmir, Turkey. Her Ph.D. thesis study is titled as 'Ontology based metadata management system for Turkey National Data Dictionary'. Her research interests are semantic web, metadata management, linked data, medical informatics and systems.