Publishing and Querying Linked Data – Case Study Technical Student Profiles

Pokhar M Jat
DA-IICT, Gandhinagar,
India

Manoj K Jain
MLS University,
Udaipur,India

ABSTRACT
In this work we demonstrate realization of semantic web by publishing profile of student on web on linked data principles. We also suggest various vocabularies that can be used for publishing profiles of a technical student in an Indian university. We take profiles of final year students at DAIICT, Gandhinagar as a case study. We have used Sesame server for publishing and querying the case data.

Keywords
Linked Data, Publishing Linked Data, Searching Semantic Web, Semantic Vocabularies.

1. INTRODUCTION
Semantic web originally made many promises that are still to be seen to taking concrete shape towards realization. Linked data, however has emerged as simple, solid, stable step forward. The term Linked Data was coined by Tim Berners-Lee in his Linked Data Web architecture note [2]. In simple terms, linked data is just a database at web scale, it lets to store and query web content as a single huge and distributed database [5].

The data model used by Semantic Web or linked data is RDF - Resource Description Framework. RDF represents data as triple statements as shown in figure-2. RDF can also be depicted as labeled directed graph, where nodes are data entities or resources, while edges are attributes called properties of entities, as shown in Figure-1 below. Label of edge is URI of property that describes it, and by referring the URI, computer programs can know more about the property term.

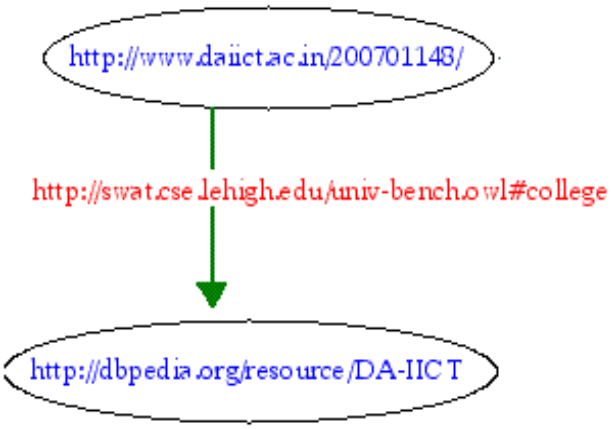

Figure 1 : Sample RDF graph

The basic assumption behind Linked Data is that the value and usefulness of data increases when there is more interlinking with other data sets. For example- when we say a resource (entity) identified by http://www.daiict.ac.in/200701148 is of type http://www.mindswap.org/mindswap#UndergraduateStudent and http://swat.cse.lehigh.edu//univ-bench.owl#degreeFrom http://dbpedia.org/resource/DA-IICT; there are various links here, from one data set to another dataset. More use of such links, more meaning we add to data, and more interoperable we are. In recent years, we have seen a phenomenon growth in the size of linked data. This motivated us to use the techniques of Semantic Web and Linked-Data and publish profiles of students, particularly related to placement and career objectives. The idea of using shared vocabularies and terms defined in them makes the data semantically understandable and interoperable. If a large number of institutes or colleges publish their data using this approach, prospective agents create various applications that can query the potential data sources and find out students that need their requirements. While, in general too, there are advantages of sharing data, this can also be seen as beneficial for students and prospective employers. Infrastructure for publishing and querying, we have used is Sesame server for its generality and popularity. [13] provides a comparison of various triple stores.

2. SEMANTIC WEB AND LINKED DATA
The vision of the semantic web is not to create a different web, but to augment the conventional syntactic web with a layer so as that the web content is represented as linked collection of concepts and resources, and are more easily interpreted by computer programs. Linked Data [1] is recent concept and is simply about using the web to create typed links between data from different Resources which may be heterogeneous systems within an entity or resources.

Typed link we mean, here is there is type associated with the link. For example in figure-1, link indicates type of property that binds the connected resources, and this type is used to infer meaning from rdf data. RDF is a data model that is extremely simple, and it represents information in the form of labeled directed graph. In this model we describe the database entities, called as Resources here as triples containing subject, predicate, and object. RDF statements/triples contain Subject, Predicate and Object which can be either URI or literals.

The Subject has a property defined by the predicate and that property has value defined by/equal to the Object. Below is set of some triples-

da:200701148	rdf:type ms:Student
da:200701148	foaf:name "Avinash Parida"
da:200701148	lubm:college dbpedia:DA-IICT

Figure 2 : Sample RDF triples (Subject, Predicate, Object)

It is interpreted as following – resource (or entity) identified by da:200701148 is of type (rdf:type) ms:Student and has name (foaf:name) "Avinash Parida", and has college (lubm:college) dbpedia:DA-IICT.

The basic principle of linked data is providing either a URI[4] that is de-reference-able or a literal for each of subject, predicate, or object. This URI describes the concerned concept so by following the given URI, one can get additional information about the concept. In fact, the meaning to the term used to describe a resource is understood by machine by dereferencing the typed linked. Linked data relies on documents containing data in RDF format. Linked data does not only simply connect these documents, it uses RDF [6] to make typed sentences that link arbitrary things in the world.

3. PUBLISHING LINKED DATA

Publishing Linked data mean making data as RDF graph accessible using HTTP protocol. This process involves following steps.

- Identify entities those needs to be described, In semantic web context, we refer them as Resources.
- Identify various properties that describe the resources
- Find out most appropriate vocabularies and terms for each identified property.
- Choose publishing format from one of given in section 3.1

3.1 Formats of publishing RDF data [4]

While RDF is logical model for representing linked data, following are some of ways of actually storing them-

3.1.1 RDF/XML files

Most simple format of storing rdf data. In this method publisher does not provide any query processing infrastructure, and the consumer should do all processing part.

3.1.2 RDF stored on triple stores

RDF triples are typically provides transparent storage of rdf triples. For querying, these provide SPARQL interface. Broadly triples stores can be categorized into DBMS based and native storage based. DBMS (mainly RDBMS) based stores use storage and retrieval functionality of existing database technologies. In this case stores may use different underlying relational schema for storing rdf triples. Stores based on native scheme implement their own strategies for storing and retrieving data, specifically tuned for rdf triples.

While relational harvests upon already available solution for storing and efficient querying methods, where as natives can be efficiently optimized for querying rdf data. Some popular infrastructure like Jena and Sesame provide both options, that is a user can store data in a RDBMS (one of many option provided RDBMS) or can store data in their native format.

3.1.3 RDF generated on the fly from relational databases:

Most enterprises still keep their data as relational databases, and some subset of this database is shared on the web in rdf format. Any of the approach discussed so far requires separate version of this subset of database – obviously leads to undesirable database inconsistency problems. The solution here is, rdf version of required data is generated on demand, RDF triples are generated from enterprise databases. For this, a popular D2R infrastructure lets to specify various mapping that can be used while generating rdf triples.

3.2 Vocabularies

Whole purpose of semantic web is to share web content in the manner that machines can process it and can understand meaning of it. As said meaning is primary by using terms that known by others (computer programs).

This objective strongly requires that we reuse the vocabularies created by others and do not create our own. Since the inception of semantic web, there has been lot of work done in this direction of defining vocabularies, however still standardization and sharing by other users is lacking, and there is no definite directory that can help for selecting vocabularies. However, following could be guiding principles for selecting vocabularies-

- Closeness with meaning of data
- Popularity of vocabulary
- Accessibility of description of vocabulary though HTTP requests.

We could not find any formal method for vocabulary selection, and therefore based on our subjective judgment using above parameters, we propose following vocabularies for describing student profiles.

- Friend-of-a-Friend (FOAF): FOAF[7] defines terms for linking people and information using the Web. It provides three kinds of term; for defining social networks of human collaboration, friendship and association.
- vCard: vCard[8] is a file format standard for electronic business cards. vCards are often attached to e-mail messages, but can be exchanged in other ways, such as on the World Wide Web or Instant Messaging. Since its early days, vCard has had an RDF vocabulary that can be used for the same purpose as the original vCard format.
- lubm: LUBM[10] is a university ontology defined by The Lehigh University as benchmark for evaluating of Semantic Web repositories
- cv: a comprehensive resume publishing ontology [9].
- Mindswap (ms): defines various academics related terms.
- Our suggested vocabulary: prefixed as da:, we could not find vocabularies defining some of terms required to be used; therefore are proposed own set of terms. We are still working to associate these terms with other related terms on the web. http://www.mindswap.org/2003/owl/mindswap

Table-1 gives list of terms used for describing various properties describing student details. Full URI of prefixes used here are given in table-2

Table 1 - Proposed Vocabularies

Property	Vocabulary:Term
Resource Type	ms:UndergraduateStudent
Student ID	da:student-id
Name	foaf:name
Gender	foaf:gender
Contact No.	foaf:phone
Email	foaf:mbox
Student Image	foaf:img
Interests	foaf:interest
Address	vcard:adr
Date of Birth	foaf:birthday
Board-12	da:board12
Marks-12	da:marks12
School-12	da:school12
UG-Degree	da:BTech
UG-Degree-Subject	da:UndergraduateDegree-Subject
College	lubm:undergraduateDegreeFrom
UG-Grade	da:grade
UG-Project	da:btp
UG-PassYear	lubm:hasGraduateYear
Skill	cv:Skill
Publications	foaf:publications
cvDocument	foaf:PersonalProfileDocument

4. QUERYING SEMANTIC WEB

Semantic web provides database perspective[5] of web, and we use term querying for semantic web rather than searching used in traditional web. After proposing various languages for querying semantic web, semantic web community appears to have come to stand of standardizing SPARQL[11] as query language for RDF data sources.

4.1 SPARQL

SPARQL stands for SPARQL Protocol and RDF Query Language. The core of SPARQL is Basic Graph Patterns (BGP). The query is defined as set of Graph patterns and matched triples from the source RDF graph are included in the result set. In order to make it more convenient to use, SPARQL inspires many features from SQL, the standard query language for relational databases. A simple example of BGP is as ?x ns:studiesAt ?y AND ?y is <DA-IICT>. Here ?x and ?y are variables, studiesAt is a property term from a vocabulary referred by ns namespace. <DA-IICT> is an individual resource on web. Most RDF data stores provide a sparql end-point(or port) [1][5]. SPARQL end-points allow programs or human beings to communicate with data-set using SPARQL language. For programs to communicate it is provided as web service, while users typical get an interface by which they can submit SPARQL query and see the result

Table 2 - URI of prefixes used in table-1

foaf	http://xmlns.com/foaf/0.1/
vcard	http://www.w3.org/2006/vcard/ns#
lubm	http://swat.cse.lehigh.edu/onto/univ-bench.owl#
da	http://intranet.daiict.ac.in/~pm_jat/student-profile.owl#
cv	http://rdfs.org/resume-rdf/
ms	http://www.mindswap.org/2003/owl/mindswap#

4.1.1 Sample Queries

We have listed down two sample queries getting answers from the dataset created for students.

Query-1: List down students (stud-id, name, and grade) who are passing out from DA-IICT in 2011 and have grade 8.0 ore more.

```
SELECT ?id, ?name, ?grade WHERE {
        ?x rdf:type ms:UndergraduateStudent
        ?x da:student-id ?id .
        ?x foaf:name ?name .
        ?x foaf:name ?grade .
        ?x lubm:hasGraduateYear
         2011^xsd:gYear .
        ?x lubm:undergraduateDegreeFrom
         <dbpedia.org/resource/DA-IICT> .
    FILTER (?grade >= 8.0 )
}
```

Query-2: List down students (name, email, grade, and college-name) who are passing in 2011 with grade 7.0 or more and have done their ug-project in the area of "semantic web".

```
SELECT ?name, ?grade, ?email, ?college WHERE {
        ?x rdf:type ms:UndergraduateStudent
        ?x foaf:mbox ?email .
        ?x foaf:name ?name .
        ?x da:grade ?grade .
        ?x da:btp ?btp .
        ?x lubm:hasGraduateYear
```

```
        2011^xsd:gYear .
        ?x lubm:undergraduateDegreeFrom
        ?college .
        FILTER (?grade >= 7.0 &&
        regex(?btp, "Semntic Web", "i") )
} ORDER BY ?college
```

Note: This query assume that we have data from more than one college. Area of Semantic Web is checked by searching this keyword using regular expression. We recommend storing abstract and key-words in this property.

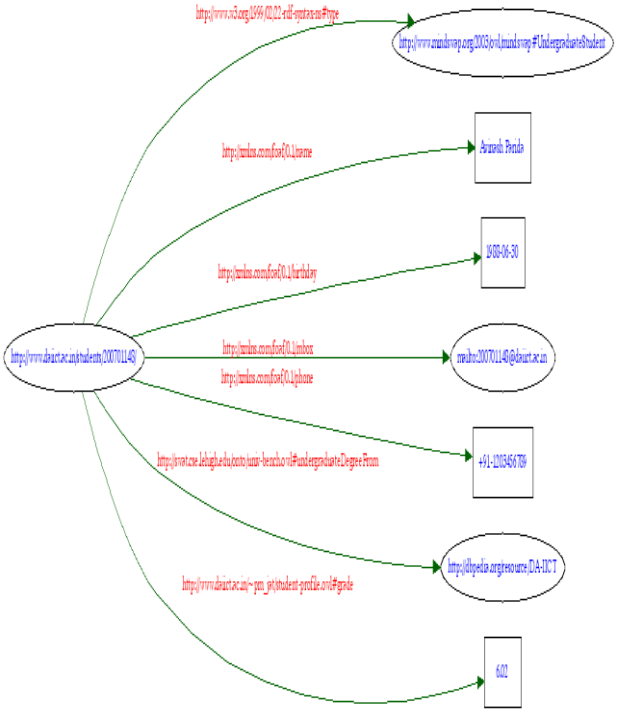

Figure 3: Partial RDF graph of information of a student.

5. CONCLUSION

We have published profiles of students of about 200 students on Sesame server using the vocabularies and terms discussed here. We believe that it should be a replicable solution for various colleges/institutes publishing profile of their students, say for placement purposes or so. For instance, hundreds of institutes are publishing profile of their students on these principles, it should be very easy to write a software agent that can answer queries like the one we have seen and run over data of all colleges. This definitely gives view that how a web of data can be created and consumed like a huge distributed database.

6. REFERENCES

[1] Chris Bizer, Tom Heath, Tim Berners-Lee. Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems, 2009

[2] Tim Berners-Lee. Linked Data - Design Issues, http://www.w3.org/DesignIssues/LinkedData.html, 2006

[3] Grigoris Antoniou, Frank van Harmelen. Semantic web Primer, 2nd edition, The MIT Press, 2008

[4] Tom Heath, Christian Bizer. Linked Data: Evolving the Web into a Global Data Space, Morgan & Claypool, 2011

[5] Olaf Hartig, Andreas Langegger. A Database Perspective on Consuming Linked Data on the Web, Datenbankspektrum, Semantic Web Special Issue, 2010

[6] Graham Klyne and Jeremy J. Carroll. Resource Description Framework (RDF): Concepts and Abstract Syntax - W3C Recommendation, http://www.w3.org/TR/rdf-concepts/, 2004

[7] Dan Brickley, Libby Miller. FOAF Vocabulary Specification 0.98, http://xmlns.com/foaf/spec/, 2010

[8] Harry Halpin, Renato Iannella, Brian Suda, Norman Walsh. Representing vCard Objects in RDF, http://www.w3.org/TR/vcard-rdf/, 2010

[9] U. Bojars, J.G. Breslin, ResumeRDF: Expressing Skill Information on the Semantic Web, The 1st International ExpertFinder Workshop (EFW 2007), 2007

[10] Yuanbo Guo, Zhengxiang Pan, and Jeff Heflin. LUBM: A Benchmark for OWL Knowledge Base Systems. Web Semantics, 2005.

[11] Eric Prud'hommeaux, Andy Seaborne. SPARQL query language for RDF. http://www.w3.org/TR/rdf-sparql-query, 2008.

[12] Chris Bizer, Richard Cyganiak, Tom Heath. How to Publish Linked Data on the Web, http://www4.wiwiss.fu-berlin.de/bizer/pub/linkeddatatutorial/, 2007.

[13] http://www.bioontology.org/wiki/images/6/6a/Triple_Stores.pdf