Latif, Atif; Afzal, Muhammad Tanvir; Tochtermann, Klaus; Raza, Muhammad Asif

Article

Automating Property Binding into Informational Aspects from Linked Data

International Journal of Applied Mathematics, Electronics and Computers

This Version is available at:
http://hdl.handle.net/11108/279

Automating Property Binding into Informational Aspects from Linked Data[#]

Atif Latif *[1], Muhammad Tanvir Afzal[2], Klaus Tochterman[1], Muhammad Asif Raza[3]

Abstract: Semantic Web is a collaborative effort to bootstrap the Conventional Web (Document Web) content by adding meaningful data structure to it. This transition is to make conventional web's content more valuable by providing machine understandable and semantically rich structured data. Linked Open Data project is seeding the Semantic Web Vision by publishing and interlinking related structured data. This project resulted in a heap of Linked Open Data which ranges from geographic to cross-domain datasets which provide huge opportunities for knowledge discovery and mash up application development. However to make full use of these datasets, there still lies many challenges which are needed to be trailed. One of the major bottlenecks in Linked Open Data is the extraction and then presentation of semantic data to the naive users. In past many efforts has been made to develop semantic applications which can search the required information from linked data sources accurately, further on can hide the complexities of data, organize and importantly convert the data in a readable format before presenting it to the users. "Concept Aggregation Framework" is one of these applications. According to this framework, the similar type of data is grouped into different aspects and sub-aspects before displaying it. In first step, this framework was practiced in an application known as "CAFSIAL". However, till now the grouping of the properties into aspects and sub-aspects is being done manually which need to be automated for new resource type alignment in CAFSIAL. This paper presents an automated way of grouping related properties into informational aspects using ontology structure. The evaluation has proved that the grouping of related properties can be automated with good accuracy.

Keywords: Semantic Web, Linked Open Data, Information aggregation and presentation.

1. Introduction

Semantic Web is an extension of our Conventional Web (also termed as Document Web). The existing web contains huge amount of information in form of documents. Information on the web is linked by means of hyperlinks, and using these hyperlinks any information can be linked to any other information. This information is of great diversity for example, from education to news, technology to medical science and music to religion. Most of the information and documents present on the web is for humans and not for computers to process it. Semantic Web aims to convert the data or information in a properly structured format so that it can be understood by computers and programs. This new format will make the already existing data meaningful. For the realization and bootstrapping structured data movement Linked Open Data [1][2] project was proposed in 2007. Main objective of this project is to disengage practices of well-gardening data and motivating people to publish their datasets as open and structured data. In a nutshell, it's an effort for creating a global connected data space [3] where related information is better connected. This will make information more reusable and discoverable; as well as leading towards unrestrictive data usage for better data interlinking, querying and intelligent application development. It is based on the Linked Data four principles stated by Tim Berners-Lee [4] which are:

- Use URIs as names for things
- Use HTTP URIs so that people (and machines) can look up those names (see also [5])
- When someone looks up a URI, provide useful information
- Incorporate other URI's so that more data can be exposed

Basically, these rules provide a set of guidelines for publishing data as Linked Data. Firstly this emphasizes to identify real and abstract concepts within the datasets and then assigning identified resources with unique URIs in Resource Description Framework (RDF) [6] format which are further dereference-able to present more meaningful information. One of the beauties of this new paradigm is that linked data applications can use any number of data sets for searching and connected with them at run time. Currently, more than 925 data sets consisting of over 60 billion RDF triples which are interlinked by around 673 million RDF link in Linked Data cloud are estimated[1].

The myriad of linked data is piling up by each passing day however complexity attached to this data is still retained.

[1] German National Library of Economics Leibniz Information Center for Economics, Kiel, Germany, {a.latif, k.tochtermann }@zbw.eu
[2] Muhammad Ali Jinnah University Islamabad, Pakistan, mafzal@jinnah.edu.pk,
[3] National University of Sciences and Technology, Islamabad, Pakistan, niazi_qau@hotmail.com
* Corresponding Author: Email: a.latif@zbw.eu
# This paper has been presented at the International Conference on Advanced Technology&Sciences (ICAT'14) held in Antalya (Turkey), August 12-15, 2014.

[1] http://stats.lod2.eu/ (as of March 2014)

Interestingly, when this available linked data is intended to be used for application development it poses many challenges. One of the major challenges is the lack of user friendly interfaces which can facilitates users in exploring linked data benefits by hiding the complex structure of RDF and SPARQL querying. Concept Aggregation Framework was presented by Latif et al. [7] exactly to solve this kind of challenge. According to this framework, the similar type of data retrieved from DBpedia[2] was grouped into different aspects and sub-aspects before displaying as a profile. In first step, this framework was practiced in an application known as CAFSIAL[3] [8]. This proof of concept application was tested for resource type "Persons" and was proved very helpful with its innovative data aggregation, organization and presentation algorithm. It demonstrated that this kind of application has led to bridge the gap between semantic search and end users. With this application, two aspects were concentrated on: 1) A novel Concept Aggregation Framework to present the most relevant information of LOD resources in an easy to understand way. 2) A simplified keyword search mechanism which hides the complex underlying semantic search logic. Though, at that time the grouping of the properties into aspects and sub-aspects was done manually and was only limited to resource type "Person".

The success of DBpedia and incorporation of new resource type in its knowledge base lead us to CAFSIAL extension by introducing a new approach which can automatically bind the similar semantic information into aspect and sub-aspect together using the ontology of a resource. The current paper presents an automated way of grouping related properties into informational aspects using ontology structure. This updating of CAFSIAL has been tested and evaluated for the new resource type: "Organization". The evaluation has proved that the grouping of related properties can be automated with good accuracy. Furthermore, the results were compared to traditional manual systems. It was found that the proposed system is able to achieve good results without manual effort. The application has successfully hidden the complex semantic structures from naïve user by presenting the results in user friendly and perceivable way.

This paper starts with a short overview of the state of the art and related literature. Then, the extension of application "CAFSIAL" along with new strategy and technical implementation are described in detail. Further on, the operational mechanism of new strategy and comparison of results with other application is described in detail. The paper closes with conclusion and outlook on future research.

2. Related Work

With the commencement of Linked Data project, many application started appearing which showcases the value of interlinking by consuming Linked data in different use cases. Currently there are two major kinds of applications which consume Linked Data and present the results to the users with different approaches. These applications are categorized as Linked Data consumption and Linked Data browsers applciations. With respect to this study, we have selected the applications which worked on RDF structures and SPARQL querying for providing access to the linked data content

particularly from DBpedia.

FAVIKI[4] is a tool which allows social book marking. It allows creating Wikipedia concepts as tags. It also allows creating new tags and connecting them to common universal concepts present in the knowledge world. A user interesting in any of these tags can dereference the URI of the tag to obtain the information.

DBpedia Mobile[5] is a client application meant for mobiles. It uses GPS signals from a mobile to get its current geographic position. After locating its position it renders its map indicating the nearby locations from the DBpedia datasets. Using this map user can further navigate into interlinked datasets and can obtain background knowledge about locations nearby. This is an interesting application of linked open data.

BBC music[6] is a web application for searching music. It is built on Musicbrainz metadata and identifiers. Information like name of the artist is picked from Musicbrainz and information like introduction of the artist is extracted from DBpedia.

PowerMagpie [9] is a new version of original Magpie. The primary goal of this system is that user has to make very little effort of semantic understanding of a web content. It automatically relates the major terms present in the text of the web page to semantic entities with help of dynamically discovered ontologies present on the (Semantic) Web. In order to achieve this goal, PowerMagpie has to handle four major tasks. a) Identifying relevant terms in the currently browse web page. b) Selecting online ontologies to interpret the domain terms. c) Relating the text to semantic information d) Navigating textual and semantic information together.

The Tabulator [10] is an RDF browser designed for naive users to provide interaction opportunities with the entire web of RDF data as well giving incentives to developers for posting RDF data and promoting RDF linking standards. Meanwhile it also designed for data provides to see how their data interact with rest of data on the Semantic Web. This project is an attempt to demonstrate and utilize the power of linked RDF data with a user-friendly Semantic Web browser that is able to recognize and follow RDF links to other RDF resources based on the user's exploration and analysis.

Visual Query Tools [11] allow users to form their own queries using a visual interface. Users should have knowledge of RDF and SPARQL in order to use construct queries and use these tools. Explorator [12] is a visual query tool. It allows user to construct queries visually. Explorator uses an operational model. A visual interface allows user to give criteria for query and the under lying operational model implements it. It allows information searching, exploration and visualization facilities. User can extract information without having domain knowledge.

Analysis of above mentioned systems shows us the major approaches and challenges which motivated us for developing CAFSIAL and its subsequent extension.

- In order to use most of the systems, user must possess knowledge about RDF, OWL and SPARQL.
- Most of the applications lack filtering mechanism.
- Making a visual SPARQL query requires background knowledge.
- A novice user cannot use these systems.
- Most of the systems are based on limited resource types.

3. Test Dataset and CAFSIAL Application

3.1. DBpedia

DBpedia is a semantic version of Wikipedia - a popular free Internet encyclopedia. This project is based on the extraction of

---

structured content from Wikipedia articles which is further made available as Linked Open Data. DBpedia allows querying the properties, relationship and external links of other resources which are associated with Wikipedia pages. DBpedia is considered as a nucleus and famous cross-interlinking hub within Linked Data Cloud as also described by Tim Berners-Lee [13][14]. The English version of the DBpedia knowledge base currently describes 4.0 million things, out of which 3.22 million are classified in a consistent Ontology, including 832,000 persons, 639,000 places and 209,000 organizations etc. Resource Description Framework (RDF) is used to represent these records and is accessible in form of RDF triples. For our part of study, we made use of "Organization" records present in DBPedia and accessed these records by querying DBpedia SPARQL endpoint [15].

3.2. CAFSIAL Application

CAFSIAL stands for "Concept Aggregation Framework for Structuring Informational Aspects from Linked Open Data". CAFSIAL contains data extracted from DBPedia known as cross-domain and nucleus of linked data. It contains 23 different types of resources like (Person, Place and Organization etc). Initial experiment of CAFSIAL was carried out with resource type "Person". According to this framework, relevant concepts of a resource can be aggregated from a knowledge base and the most related informational aspects can be organized in to informational aspects [8]. It addresses challenge related to data presentation. The aim is to hide the process of data extraction and processing from users and allowing users to search for information in the way user they already accustomed to. The last and most important feature in this application is the presentation of the information in a structured textual form which is easy to understand.

4. Extension of CAFSIAL

In this section strategy for automatic property binding in CAFSIAL is discussed in details.

4.1. Populating Resources

Initial version of CAFSIAL holds data about "Persons" and the process of structuring related information in different aspects was manual. The manual allocation of DBpedia properties over informational aspect in this application is illustrated in figure 1. Based on the similar mapping strategy a Concept Aggregation Framework has been applied over another DBpedia resource type "Organization" [16]. These experiments on "Persons" and "Organization" helped us in studying different issues related to information structuring and presentation and subsequently helped us in automatically mapping properties to related aspects and sub aspects.

4.2. Automating Property Binding

Data extracted from DBpedia is in semantic form and comprises of RDF tags. The aim is to structure the similar information in groups called aspects. The taxonomy of a class is defined by its ontology. Ontology gives following information of a resource.

    a.   Property Name

    b.   Label

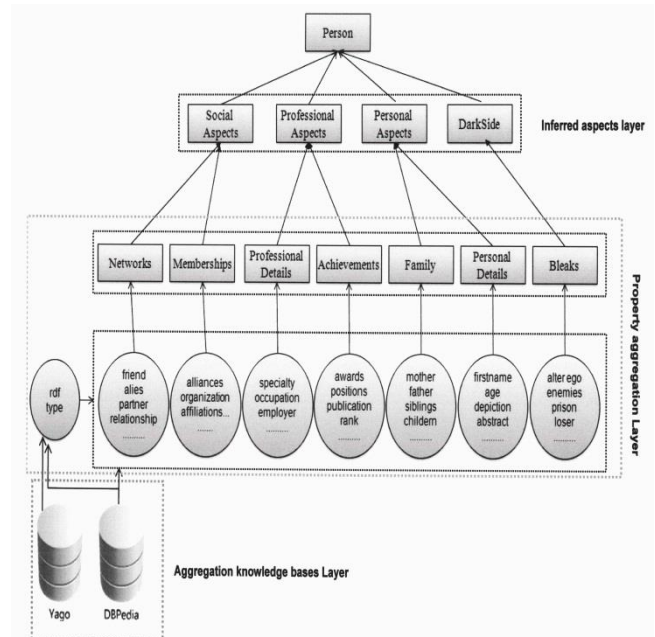   c.   Domain

    d.   Range

    e.   Comments

Figure 1. Property Mapping in CAFSIAL

"Property Name" is the name of property, "Label' is the descriptive name of the property, "Domain" describes the area or field from which a property belongs, "Range" defines the which type of data the property holds and comments gives some description about the property. Data related to any resource is extracted on the basis of the properties the resource holds. Property Names are used in SPAQL queries for retrieving the desired information. In previous experiments similar type of properties were manually placed in similar aspects for structuring the data.

4.3. Information Structuring Using Domain & Range

After studying the properties of resource type "Organization" in above mentioned experiment a hypothesis was made that the Domain and Range can be helpful in structuring the data. Similar properties appear on similar level if arranged with respect to Domain and Range. Thus the domain and range can be used in automating the processing of property binding to aspects.

5. CAFSIAL Automation

The process of property retrieval can be automated using ontology of resource i.e. Domain and Range as described in previous section. This assumption will be used as a principle for the automation of CAFSIAL application. The steps involved are described ahead.

5.1. Resource Selection

The resource for the experiment is "Organization". Since the study of this resource has already been made it will further swift up the task. Moreover, "Organization" spreads to wide range of subclasses like "Airline", "College, "Military", and "Television" etc, it will certainly allow us in testing our new rule on wide range of information.

5.2. Properties Extraction

Properties names for each sub type of Organization has been extracted and saved in a local database using SPARQL queries. An example query extracting the properties for "College" is:

```
SELECT DISTINCT ?p WHERE {
        ?s ?p ?o .
?s rdf:type <http://dbpedia.org/ontology/College> .
                    }
```

5.3. Identifying Domain & Range

Domain and Range of the properties of a resource are identified by using SPAQRL query for each property name. For example:

```
SELECT DISTINCT ?domain
            WHERE {
property_name http://www.w3.org/2000/01/rdf-schema#domain
            ?domain    }
```

The above query retrieves the domain of a property named "property_name". However, for using the above query the property must be of type owl:ObjectProperty. Unfortunately, this is not the case for majority of the DBpedia properties. Other type of properties like dbpedia2:property can't be used in above query. For the time being, this issued has been resolved by query each property for an instance of resource class, and its domain / range have been decided by analyzing the output of the query. Domain and Range of each property are also saved in a local database.

5.4. Defining a Threshold

Each DBpedia resource holds a number of properties which varies from resource to resource as described in table below. It can be noted that some of the resource types hold very large number of properties. Moreover, it is important to present only necessary of common aspects / information related to a searched content. This rule is used by "Google". All of the properties are of no use, only common, important properties holding the key data be displayed to user. In order to limit the number of properties a threshold has been defined. It was noted that the properties less than 10 records are of less importance. Thus filtering the properties with less than 10 records resolves the issue of properties selection which was manually being done in previous experiments. The filtered properties can now help is displaying the most relevant content to the user. A SPARQL query retrieving the number of records for a property named "property_name" against a resource "Company" is given below. The number of records for each property for each resource type has also been saved in the database and is also tabulated in table 1 for simplification.

```
SELECT COUNT(DISTINCT ?s) AS ?counts
            WHERE {
        ?s property_name ?o .
?s rdf:type <http://dbpedia.org/ontology/Company> .
                    }
```

5.5. Mapping of Related Properties to Aspects

Properties are grouped together to form aspects. Each aspect consists of some properties that are related to it. Maximum number of aspects displayed is 8. Name of the aspects and its details is given below:

- ➢ Abstract: Abstract describes the searched term in descriptive form. It gives a brief introduction about the subject.
- ➢ Organization Aspect: Properties whose Domain and Organization types are similar are bind aspect named similar to the organization type, for example, if the Domain type is "Airline" and the Organization is also "Airline" the name of aspect will be "Airline".
- ➢ Related Persons: Every organization has some important persons related to it. Properties related to these persons are displayed in this aspect.

Table 1. Property Count for Organization

Airline	254	Band	1007
Broadcaster	851	Broadcaster Network	279
Radio Station	462	Television Station	528
Company	2056	Law Firm	xxx
Record Label	221	Educational Institution	4113
College	133	Library	150
School	2723	University	1964
Geopolitical	Xxx	Government Agency	316
Legislature	101	Military Unit	715
Political Party	357	Sports League	330

- ➢ Financial Aspects: Financial aspect is an important aspect. It contains information related to financial issues of the Organizations.
- ➢ Important Dates: This aspect contains different dates related to the organization. For example founding data, closed date etc.
- ➢ Geographical Aspects: Information which describes geographical properties of an organization like its geographical coordinates and location etc are described in this aspect.
- ➢ Important Values: The properties describing different figures or values of the organization are displayed in this aspect. For example if the organization is a college it contains number of students, graduates and undergraduates etc.
- ➢ Web Aspect: Properties that are related to web are shown in this aspect, for example web address of the organization and its wiki page.

6. Operational Mechanism

The working of the automated CAFSIAL is explained in this section. CAFSIAL has a simple web based interface, and can be run in any of the web browser. User enters an organization's name and related names are suggested by the system automatically. User enters the name or selects one form auto-suggested terms and searched button is pressed. Names of the organizations which match the user's entered term are retrieved along the organization type and are displayed to user. For example, if user searches "Oxford", he is provided with list of possible college (organization) resources to decide on as shown in figure 2.

Figure 2. Screenshot - List of Suggested Organizations

Thus autosuggesting the organization name can help in refining

the results. When user clicks at the name of organization in the suggested list e.g. ""Christ_Church,_Oxford" properties related to the searched organization are selected from the database on the basis of organization type and are filtered by the threshold already saved. Each property is queried to DBpedia Server and data is retrieved. It is important to note that no data i.e. DBpedia dump is being stored locally. This is notable improvement since the previous version of CAFSIAL was locally storing the DBpedia dumps. As the data is retrieved, at the same time similar properties are allied together to their relevant aspects using the already stored Domain and Range.

Figure 3. Screenshot of Result Presentation.

The data retrieval and property binding is done in parallel. As soon as similar retrieval of dated of similar group of properties is completed, they are assigned an appropriate aspect name and displayed to the user. The final presentation of the results is illustrated in figure 3.

7. Evaluations

The initial evaluation of the proposed system with the existing leading similar systems is presented in this section. There are four similar systems such as: Initial CAFSIAL, Facet search, 123people, and Freebase. Facet Precision Search & Find is an online semantic system presented by Open Link Software and uses virtuoso server. 123people is an online search engine dedicated to people search and empowers users to find information about themselves and other people. Whereas Freebase is a community-curated database of well-known people, places and things working under flagship of Google. Comparisons of theses system with CAFSIAL are given below.

Table 2. Comparison between Old and new Cafsial application

Existing CAFSIAL	New CAFSIAL
Properties have been manually mapped to related aspects.	Properties are mapped to aspects using ontology i.e. Domain and Range.
Holds data in a local database	Data is retrieved from DBpedia server at runtime
Works on resource type "Person"	New CAFSIAL has used resource type "Organizations"
Output is displayed to user after all the queries are processed.	All the queries are processed in parallel and output is displayed as soon as a query is completed.

Table 3. Comparison between Facet and Cafsial

Facet	CAFSIAL
Gives option to search a key term, Label and URI semantically	Allows semantic searching by taking input in simple text form.
Gives information in text format i.e. paragraph about all the searched terms in first step	Shows the names of all organizations that matches the keyword along with its type in first step
After detailed search shows information in semantic + text form	Displays in simple text which is easily conceivable.
Uses DBpedia Data.	Takes information from DBpedia server

Table 4. Comparison between 123 People and Cafsial

123People	CAFSIAL
Gives option to search people by their name	Allows searching Organizations (like Airline, TV, Radio, Colleges).
US database is used.	DBpedia is used.
Searches data in different categories, i.e. name, organizations, telephone numbers, email addresses	Searches from Organizations Data only and returns the searched name to user for the results he/she wants
Shows names with few details on main page.	Displays common aspects and properties on the first page
Demands fee for viewing complete data of US citizens.	CAFSIAL is a Public system for research purposes.

The comparison of Freebase and CAFSIAL are given in the Table 5. The Freebase is manually created database, however, from the Table 5, it is clear that CAFSIAL is comparable in results in some cases; however, in majority of cases, the CAFSIAL has outperformed Freebase.

Table 5. Comparison Cafsial and FreeBASE

Organization Type		Airline	College	Hospital	Legislature	Library	TOTAL
No of concepts searched		39	27	52	40	24	182
CAFSIAL	Distinct Properties Searched	33	30	20	38	25	121
	Min Searched	2	18	9	32	7	61
	Max Searched	21	26	19	20	22	86
	Avg Searched	15.33	21.67	14.34	23.47	14.54	17.87
FREEBASE	Distinct Properties Searched	45	43	25	10	27	150
	Min Searched	4	12	1	4	3	24
	Max Searched	25	26	17	10	14	92
	Avg Searched	15.33	20.12	6.56	5.96	6.28	10.85

8. Conclusions

Linked Open Data provides huge opportunities for knowledge discovery and mash up application development. However to make full use of Linked Data, still there lies many challenges like extraction and presentation of semantic data. In past a Linked Data application named as: CAFSIAL which was based on "Concept Aggregation Framework" was presented. This application was aimed in providing the user with easy to use interface and presentation of the retrieved information in structured way using Concept Aggregation Framework. In this application similar type of data was grouped into different aspects and sub-aspects as a profile before presenting it to the users. However the mapping of properties to information aspect was achieved manually. In this research CAFSIAL was extended by introducing the process of automatically binding the properties to their relevant concepts by using information present in the ontology of the resource. This updating of CAFSIAL has been tested and evaluated for the new resource type: "Organization". The evaluation has proved that the grouping of related properties can be automated with good accuracy. Furthermore, the results are compared to traditional manual systems. It is found that the proposed system is able to achieve good results without manual effort. In future this proof of concept extension will be deployed along with CAFSIAL running system; moreover other approaches for generating new specific aspects and sub aspects will be targeted.

References

[1] W3C community project Linking Open Data, 2007. http://esw.w3.org/topic/SweoIG/TaskForces/CommunityPr ojects/LinkingOpenData

[2] Bizer, C., Heath, T., Ayers, D., Raimond, Y., "Interlinking Open Data on the Web", Demonstrations Track at the 4th European Semantic Web Conference, Innsbruck, Austria, 2007

[3] Heath, T. and Bizer, C. (2011). Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan and Claypool 1:1., 2011

[4] Berners-Lee, T. (2006). Linked Data -- Design Issues, 2006. http://www.w3.org/DesignIssues/LinkedData.html

[5]

[6] Sauermann, L., Cyganiak, R., Ayers, D., Völkel, M., "Cool URIs for the Semantic Web", W3C Interest Group Note, 2008. http://www.w3.org/TR/2008/NOTE-cooluris-20081203

[7] Resource Description Framework, 2004 http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/

[8] Latif, A., Afzal, M. T., Us Saeed, A., Hoefler, P., Tochtermann, K.: CAF-SIAL: Concept Aggregation Framework for Structuring Informational Aspects of Linked Open Data. In: Proceedings of NDT 2009. Ostrava (2009)

[9] Latif, A., Us Saeed, A., Hoefler, P., Tochtermann, K., Afzal, M. T.: Harvesting Pertinent Resources from Linked Open Data. In: Journal of Digital Information Management, 8(3), pp. 205{212. Digital Information Research Foundation, Chennai (2010)

[10] M. Dzbor, J. Domingue, and E. Motta. "Magpie – towards a semantic web browser".

[11] T.B. Lee, Y. Chen, L. Chilton, D. Connolly, R. Dhanaraj, J. Hollenbach, A. Lerer, and D. Sheets. "Tabulator: Exploring and Analyzing linked data on the Semantic Web".

[12] T. Catarci, M. F. Levialdi and C. S. Batini. "Visual Query Systems for Databases: A Survey".

[13] S.F.C. deAraújo and D. Schwabe. "Explorator: A tool for exploring RDF data through direct manipulation".

[14] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R.and Ives, Z., "DBpeddia: A Nucleus for a Web of Open Data", In Proceeding(s) of the 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Lecture Notes in Computer Science 4825 Springer 2007, ISBN 978–3-540–76297–3, 2007

[15] Sir Tim Berners-Lee Talks with Talis about the Semantic Web, Talis, 7 February 2008.

[16] DBpedia Sparql endpoint - http://dbpedia.org/sparql

[17] Raza, M.A.; Afzal, M.T.; Khanum, A., "Discovering, structuring and visualizing organizations from Linked Open Data," Computer Sciences and Convergence Information Technology (ICCIT) 6th International Conference , pp.738,744, Nov. 29-Dec. 1 2011