Latif, Atif; Tochtermann, Klaus

Conference Paper

Finding Resources in Scholarly Communication and Cross-domain Linked Dataset

Suggested Citation: Latif, Atif; Tochtermann, Klaus (2013) : Finding Resources in Scholarly Communication and Cross-domain Linked Dataset, In: Proceedings of ACM International Conference Proceeding Series (ICPS) ICIPM2013, 8th International Conference on Information Processing and Management, 1-3th April, 2013, Seoul, Korea, ACM, New York

Finding Resources in Scholarly Communication and Cross-domain Linked Dataset

Atif Latif
ZBW - German National Library of Economics Leibniz Information Center for Economics,
Kiel, Germany
a.latif@zbw.eu

Klaus Tochtermann
ZBW - German National Library of Economics Leibniz Information Center for Economics,
Kiel, Germany
k.tochtermann@zbw.eu

ABSTRACT
Linked Open Data project motivates people to publish their data as structured data for enabling machine understanding, better information linking and knowledge discovery. This effort has succeeded in bringing up variety of Linked Open Data; ranging from domain specific to cross-domain structured datasets. Currently, these published linked datasets provide huge opportunities for information linking and intelligent application development. Scientific scholarly communication datasets are one of the major contributors in steering today research work and has a big share in Linked Data Cloud as well. On the other hand, cross-domain datasets are proved to be good data sources for various content type enrichments because of its crowd sourced knowledge bases. By keeping in mind the importance of scholarly communication and cross-domain datasets it will be great to know; what these datasets has to offer each other in Linked Data sphere. We are of a view; if these datasets are linked can offer comprehensive information regarding artifacts of scholarly communication datasets e.g. authors and publications. However, currently interlinking with appropriate and quality data in linked data cloud is still quite a challenge. In this paper we presented a case study by interlinking author from scholarly communication dataset (DBLP) with person record from cross-domain dataset (DBpedia). Moreover, we have investigated: a) how much author information is there in DBpedia for indexed DBLP scientific authors and b) validated our assumption that there is meaningful data between these datasets.

Categories and Subject Descriptors
H.3.3 [**Information Systems**]: *Information Storage and Retrieval Information Search and Retrieval*. I.2.8 [**Problem Solving, Control Methods, and Search**]: *Heuristic methods*.

General Terms
Algorithms, Management, Design

Keywords
Data fusion, data linking, linked data, semantic web, dblp, dbpedia.

1. INTRODUCTION
Linked Open Data project was envisioned for producing and sharing structured data that is understandable and process-able by machines. Main objective of this project is to disengage the practice of well-gardening key datasets and motivating people to publish their datasets as open structured data, which further can be helpful both for the people and machines for constructing intelligent services. In a nutshell, it's an effort for bootstrapping Semantic Web Vision at global web scale by creating a global connected data space [1] where related information is better connected. It will make information more discoverable leading us to issuance of simple and complex queries; as well as development of intelligent web services. The W3C community project *Linking Open Data* [2][3] was initiated in 2007. 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
- Include links to other URIs so that they can discover more things

These rules give us a set of guidelines for publishing data as Linked Data. In a nutshell, these rules are first to identify the real and abstract concepts in the datasets and then assigning these identified resources with a unique URI in Resource Description Framework (RDF) [6] which is further dereference-able to present more meaningful information. By following these set of rules individuals, researchers and organizations can publish their datasets as structured data and can interlink them with other datasets to bring more value in. A large number of linked datasets, supporting tools and techniques have emerged after this project conception. Currently more than 300 data sets consisting of over 31 billion RDF triples which are interlinked by around 504 million RDF links are recorded in September 2011[1].

[1] http://lod-cloud.net/

Scientific Scholarly communication data-sets are one of the core contributors in Linked Data sphere. With the commencement of Library Linked Data Movement [19] many of the distinguished digital libraries i.e. LIBRIS – Swedish National Bibliography and authority database[2], ZBW – German National Library of Economics[3,4], Europeana[5] and GESIS[6] has started publishing their data as Linked Open Data. One of the prominent digital libraries is DBLP [7] which has been covering computer science literature from last two decades. It provides access to huge metadata of scientific literature published in the well-known workshops, conference proceedings and Journals. DBLP L3S dataset[7] is a semantic version of DBLP legacy dataset and provides scientific literature metadata as linked data. This metadata of papers usually consists of author names, paper titles, paper keywords and venue information. To date, many research studies has been performed to access and interlink DBLP scientific literature with other scholarly communication datasets. One of the recent conducted studies [8] has successfully showcased the potentials and benefits in using DBLP dataset for content enrichment. On the other hand, a cross-domain datasets e.g. DBpedia[8] and Freebase[9] provides a rich crowd sourced knowledge base for discovering, enriching and management of semantic information. DBpedia is one of the popular and important cross-domain dataset which covers real world concepts in form of structured pages. Recent completed case study [9] around DBpedia has shown its utility for interlinking with domain specific datasets i.e. person, places, movies etc.

By keeping in mind the claimed benefits in recent studies we are of view that if DBLP and DBpedia are interlinked can provide additional and important information about the artifacts of scientific publications i.e. author, papers, keywords. Further on, interlinking results can help developers to develop a linked data application which can facilitate students and researcher to find information about certain author or paper in these mentioned datasets. However, for that to happen we need a comprehensive interlinking strategy which can automatically find and validate the relevant results. Currently, the task of finding relevant external data to interlink is still a challenge as compared to publishing in RDF. The challenges involved in interlinking are: i) understanding of the underlying semantic structures and ontologies ii) availability of live SPARQL endpoints with adequate knowledge of sparql querying and semantic technologies for data retrieval and further processing and iii) most importantly validation of the retrieved data relevance and it's quality.

In this paper we firstly focused on these challenges and investigated the interlinking possibilities with a concrete use case where a scholarly communication dataset DBLP is interlinked with profile of scientific authors in DBpedia. For that purpose we have searched, identified, queried and disambiguate the records with various heuristics. Secondly, the goal of this study is to investigate the interlinking scores between DBLP and DBpedia datasets. We assume that we will find good interlinking scores between these datasets.

This paper starts with a short overview of the related literature. Then, the test datasets which were investigated for this study are discussed in the dataset section. In the next section, the actual use case and the technical implementation are described in detail. The paper closes with an observation over interlinking results and an outlook on future research.

2. RELATED WORK

With the recent research developments, concepts of Linked Data are maturing but still there are some unsolved challenges. Regarding automatic generation of similar RDF links; Heath et al. [10] has surveyed many key-based as well as similarity-based approaches for interlinking. They summed up that problem of making links with external data sources is still a challenge and required more case studies and tools for better interlinking in Linked Data cloud. Related work in this paper is divided into two parts 1) case studies and 2) tools and services.

2.1 Interlinking Digital Journal with DBpedia

In this study [8], authors of open digital journal dataset were interlinked with the DBpedia by extending the CAF-SIAL (Concept Aggregation Framework for Structuring Informational Aspects of Linked Open Data)[10]. CAF-SIAL is a proof of concept linked data application which finds and harvests person's relevant information from the Linked Data cloud. Moreover, this information is organized and presented in informational aspect of person (personal, professional, educational and social) in a profile. This system is currently used by the Journal of Universal Computer Science[11] where authors were successfully interlinked with DBpedia. Moreover, author's information is presented in a well-crafted profile which has proved very beneficial for the journal administration and journal users. This work has motivated us to linkup the DBLP datasets with DBpedia.

2.2 Interlinking Scholarly Communication Datasets with DBLP

This study [9] depicts a concrete use case where a relational database of scientific authors and publication from Know-Center (Austria's competence center for knowledge management and knowledge technologies) was "RDFized" and were further semi-automatically linked with DBLP. In this study, first manual selection of the suitable scholarly linked data datasets was performed by querying popular author of Know Center in available scholarly linked datasets, i.e. DBLP D2R L3S, CITESEER and ACM RKB Explorer[12], and Semantic Web Dog Food [11] which resulted in selection of DBLP for interlinking. Secondly, after interlinking data was arranged and presented in an automatic generated profile. This study has shown the potentials and added value of DBLP as a dataset for content (publications and authors) enrichment and resource discovery.

[2] http://libris.kb.se

[3] http://linkeddata.econstor.eu/

[4] http://zbw.eu/stw

[5] http://data.europeana.eu/

[6] http://lod.gesis.org/thesoz/

[7] http://dblp.l3s.de/d2r/

[8] http://dbpedia.org

[9] http://freebase.com

[10] http://cafsial.dratiflatif.com/

[11] http://www.jucs.org/

[12] http://rkbexplorer.com/

One of the famous tool and linked data search web service which can be used for external interlinking are introduced next.

2.3 SILK

SILK framework [12] provides an interlinking tool which is used to discover relationships of resources within different linked dataset. By using Link Specification Language (LSL) and its provided conditioning features, data publisher can set criteria for matching RDF resources for interlinking. SILK framework works on data sources that are interlinked with the SPARQL specification and are already present in RDF format. This tool is very handy where data publisher already know the certain criteria and conditions which needs to be fulfilled. In its simplest form a restriction just selects all entities of a specific type inside the data source. For instance, in order to interlink cities in DBpedia, a valid restriction may select all entities with the type dbpedia:City. For more complex restrictions, arbitrary SPARQL triple patterns are allowed to be specified. However, in the dataset where inconsistencies are present and we need instance data level matching, this tool might not be very helpful. We are also facing similar situation in our case study, hence we have to come up with technique where we can add some heuristic to cater disambiguation problem.

2.4 SINDICE

Sindice [16] is a state of the art semantic search engine and web index. It provides searching facility to process, consolidate and query the Web of Data. Through it's offered searching interfaces, public API access and sparql endpoint it provides access to the huge pool of indexed RDF resources. Users are allowed for forming query with various triple patterns through its API to search. Sindice results often need to be analyzed and refined before they can be directly used for a particular use case. Similar kinds of services are provided by semantic search engines like Falcon [17] or Swoogle [18]. We selected Sindice for our study because of its larger indexing infrastructure and easy to use search API.

3. DATASET

For our interlinking study, we have accessed and made use of two linked data repositories i.e. DBLP and DBpedia. These datasets are provided openly and accessible by sparql endpoint. Given below are the details of the datasets:

3.1 DBpedia

DBpedia is a semantic version of Wikipedia[13] - a popular collaboratively edited free internet encyclopedia. DBpedia project is based on the extraction of structured content from Wikipedia articles which is further made available on World Wide Web 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]. There has been valuable work done on studying the reliability of Wikipedia URI's [15] that is a backbone of the DBpedia. This study suggests that the meaning of a URI stays stable approximately 93\% of the time. The English version of the DBpedia knowledge base currently describes 3.77 million things, out of which 2.35 million are classified in a consistent Ontology, including 764,000 persons, 573,000 places, 333,000 creative works, 192,000 organizations, 202,000 species and 5,500 diseases[14]. Resource Description Framework (RDF) is used to represent these records and is accessible in form of RDF triples. For our part of study, we have concentrated on the persons records present in DBpedia and accessed these records by querying DBpedia sparql endpoint.

3.2 DBLP D2R L3S Server

The DBLP D2R L3S server[15] is based on the XML dump of the DBLP database. The DBLP [7] database provides bibliographic information on major computer science journals and conference proceedings. The database contains more than 1 million articles and 14,663 authors. To query the DBLP L3S data set, the D2R Server, a semantified version of DBLP bibliography, was accessed via its sparql endpoint.

4. INTERLINKING FRAMEWORK

For the interlinking of the DBLP and DBPEDIA dataset, a multi-step strategy was devised to find similar resources in form of authors. These discovered resources were further processed for interlinking of resources with ***owl:sameAs*** relationship to produce a final mapping file to the community. The framework for this strategy is illustrated in Figure 1. The strategy looked as follows:

4.1 Acquisition Service for Author Metadata

With the help of a web service, metadata of the authors from the DBLP were retrieved via a SPARQL query from D2R server sparql endpoint[16]. Afterwards, all retrieved author records containing full names and URIs were stored in a relational database for further processing.

4.2 Author URI Acquisition from DBpedia

The inconsistencies in names e.g. umulate and special characters were firstly processed for the acquisition of the DBLP author URI from DBpedia. Afterwards, DBpedia dataset was queried with author names in two settings for producing maximum matched results. The two approaches used for this purpose is explained further:

Firstly, Sindice Search API[17] was used to search authors for their DBpedia URI's. For that we wrote a web service which took the authors name iteratively as an input and automatically called the API with formulated search queries. The resulting URIs was then filtered automatically on the basis of heuristics to make sure that they belonged to the DBpedia dataset. In this process, URIs for 1097 out of the 14663 authors in question was found. Due to recall and precision problems which every search engine displays we decided to use DBpedia sparql endpoint for improving our results.

For that, we employed a string-matching algorithm that compared names of the authors in question with author names from the DBPedia dataset. To accomplish this, we queried sparql endpoint

[13] http://en.wikipedia.org/

[14] http://dbpedia.org/About

[15] http://dblp.l3s.de/

[16] http://dblp.l3s.de/d2r/sparql

[17] http://sindice.com/developers/searchapiv3

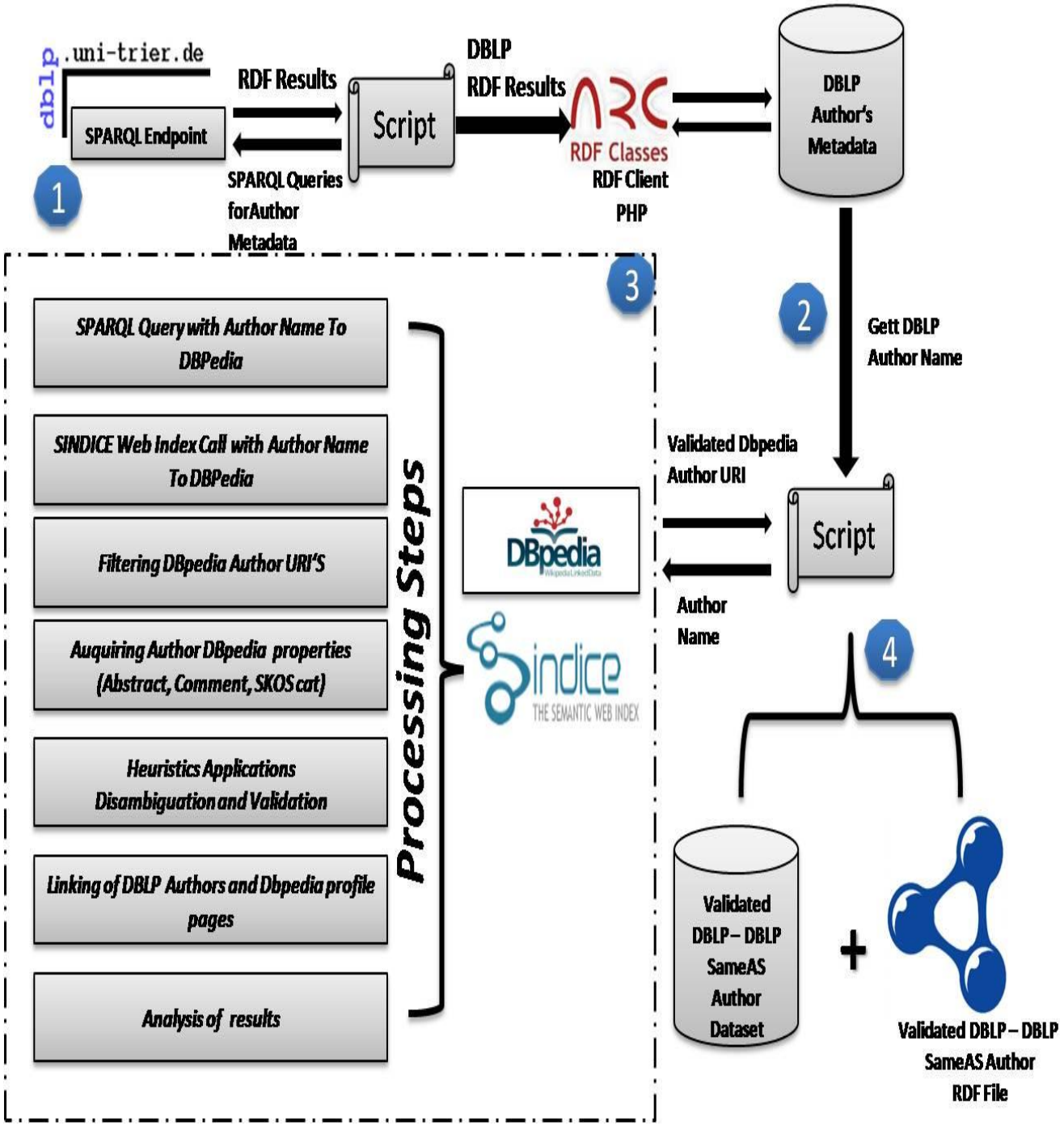

Figure 1.Interlinking framework

of DBpedia with our web service. This step provided us with the URI's of author whom full name was matched with the person name in DBpedia dataset. This way we found a few additional URIs which was not retrieved by Sindice in the previous step. At the end, we combined the results from step1 and step2 and constructed a relational database which stored the names and URIs of each author.

4.3 Validation Service

In order to ensure the validity of the matched URIs, a validation service was written which will help us to validate DBpedia person and DBLP author matches. This service iteratively took the author name with matched DBpedia URI and constructed a sparql query for retrieving the associated properties in DBpedia. By manual inspection and re-referencing of the acquired URI's, we discovered some irregularities which are discussed below:

- First, a large portion of person have similar name like author name in DBpedia but actually belonging to other professions i.e. sports, celebrity, actor etc. which may lead us to the wrong person interlinking. For that we have to think about heuristics which can help us to identify persons who belongs to academic or education field.
- Secondly, URI's of some authors are present in the DBpedia but providing no additional triples i.e. no abstract and categories. This particular limitation may make it impossible for predicting the correct match.

Hence to disambiguate authors, a set of heuristics were written. After manual inspection, it was noted that there are certain kind of properties in DBpedia for type person who can be exploited to predict correct person. These properties are dbpedia:abstract/ dbpedia:comment and SKOS categories. For example SKOS categories and keywords being used to represent the persons belonging to education profession are: "computer science, computer scientist, professor, informatics, researcher ....." etc. All of these mentioned properties represent a person belonging to scientific community. Thus the persons having same names and belonging to different professions can be filtered out by applying these keywords. For that an automated script was written to check the keywords in the abstract, comments and SKOS categories of the URI. After applying this script on 1097 authors, 395 URIs was left. The remaining URI's were either bad links (showing no data) or representing non-scientific persons.

5. OBSERVATIONS

The analysis of results which we obtained after this study leads us to the following observations.

- Not many of computer scientist's profiles present in DBpedia dataset. By keeping in mind the huge usage of social media it came out as a surprise. However, we are of a view if more profiles can be maintained by the editors it can help in: a) more valuable interlinking and b) scientist to increase their visibility. This practice may also be valuable in increasing interlinking scores between DBpedia and DBLP and further on interlinking results can be used in development of various mash-up applications.
- In interlinking process we have observed many person data articles / pages which lead to ambiguous results due to same name issues. We propose that some additional measure can be taken by DBpedia to cater this issue. For instance, editors can introduce a special category or profession property text added to the URI for a person page.
- Over all, success of Linked Open Data depends upon the heavy interlinking between external datasets so that we can have a single connected global data space where simple and sophisticated queries are possible. If it happens can lead us to more knowledge discoveries. However currently, the lack of adequate tool, quality data and few interlinking case studies impeding the way to achieve basic vision of Linked Open Data. Still there are need of more efforts for data creation, publishing tools and interlinking mechanisms. We hope this study can be useful for future interlinking studies.

6. CONCLUSIONS

The Linked Open Data movement provides a heap of Linked Data and motivates people to open up their dataset for interlinking. Scientific scholarly communication datasets and cross-domain datasets are the major contributors in Linked Data cloud. By keeping in mind the importance of scholarly communication and cross-domain datasets their interlinking can result in enrichment of content and resource discovery. However, the process of interlinking data with external datasets stills quite a challenge. In this paper, first we investigated the interlinking challenge and proposed a multi-step strategy for interlinking. We have applied and showcased our case study by interlinking scholarly communication dataset DBLP and cross-domain DBpedia dataset. Results of this study has shown that there is not many of computer scientist profile exists which can be interlinked with DBLP. However, we found that biographical information of a person with other relevant properties in DBpedia can be very useful first in disambiguation of persons and secondly in improving author information with in scientific communication dataset. At the end we have generated a mapping file of validated interlinking links as of owl:sameAs relationship which can be used for research purpose. We are of this view that if more information of authors can be created by editors or authors in DBpedia can lead us further to development of interesting mashup applications. This can also be a key factor in increasing authors' visibility to outside datasets as well.

7. REFERENCES

[1] 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) 1--136 http://linkeddatabook.com/editions/1.0/#htoc56

[2] W3C community project Linking Open Data. (2007). http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

[3] Bizer, C., Heath, T., Ayers, D., Raimond, Y. (2007). Interlinking Open Data on the Web. Demonstrations Track at the 4th European Semantic Web Conference, Innsbruck, Austria. (May 2007) http://www.eswc2007.org/pdf/demo-pdf/LinkingOpenData.pdf

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

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

[6] Resource Description Framework. (2004) http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/

[7] Michael, L. (2009). DBLP - Some Lessons Learned. PVLDB 2(2). (2009) 1493—1500

[8] Latif, A., Afzal, M.T., Helic, D., Tochtermann, K., Maurer, H. (2010). Discovery and Construction of Authors' Profile from Linked Data (A case study for Open Digital Journal). Linked Data on the Web (LDOW 2010).

[9] Latif, A. and Tochtermann, K. (2012). Webbing Semantified Scholarly Communication Datasets for Improved Resource Discovery. JDIM 10 (4) , 245-253

[10] Bizer, C., Heath, T., Berners-Lee, T. (2009). Linked data-the story so far. International Journal on Semantic Web and Information Systems (IJSWIS) 5 (3) , 1--22 .

[11] Semantic Web Dog Food. (2009) http://data.semanticweb.org/

[12] Volz, J., Bizer, C., Gaedke, M., Kobilarov, G. (2009). Silk– A Link Discovery Framework for the Web of Data. In: Proceedings of CEUR-WS Vol-538 of 2nd Linked Data on the Web Workshop (LDOW2009), Madrid, Spain. (2009) http://events.linkeddata.org/ldow2009/papers/ldow2009_paper13.pdf

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

[14] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R.and Ives, Z. (2007). DBpeddia: A Nucleus for a Web of Open Data. In Proceedings of the 6th International Semantic Web Conference (Busan, Korea 2007). Springer.

[15] Hepp, M., Siorpaes, K., Bachlechner, D. (2007). Harvesting Wiki Consensus Using Wikipedia Entries as Vocabulary for Knowledge Management. IEEE Internet Computing. 11(5) pp.54-65 Sep 2007

[16] Oren, E., Delbru, R., Catasta, M., Cyganiak, R., Stenzhorn, H., Tummarello, G. (2008). Sindice.com: A Document-oriented Lookup Index for Open Linked Data. International Journal of Metadata, Semantics and Ontologies. 3(1) (2008) 37-52

[17] Cheng, G., Ge, W., Qu, Y. (2008). Falcons: Searching and Browsing Entities on the Semantic Web. In: Proceedings of 17th International World Wide Web Conference, Beijing, China. (2008) 21--25

[18] Ding, L., Finin, T., Joshi, A., Pan, R., S. Cost, R., Peng, Y., Reddivari, P., C. Doshi, V., Sachs, J. (2004). Swoogle: A Search and Metadata Engine for the Semantic Web. In: Proc. Thirteenth ACM Conference on Information and Knowledge Management, Washington, D.C., USA. (2004) 8—13

[19] Library Linked Data Incubator Group: Use Cases. http://www.w3.org/2005/Incubator/lld/XGR-lld-usecase-20111025/