# Linked Open Data System for Scientific Data Sets

Frederik Simon Bäumer
Heinz Nixdorf Institute, HNI
University of Paderborn
Paderborn, Germany
fbaeumer@hni.upb.de

Jangwon Gim*
Korea Institute of Science and Technology Information, KISTI
Daejeon, South Korea
jangwon@kisti.re.kr

Do-Heon Jeong
Korea Institute of Science and Technology Information, KISTI
Daejeon, South Korea
heon@kisti.re.kr

Michaela Geierhos
Heinz Nixdorf Institute, HNI
University of Paderborn
Paderborn, Germany
geierhos@hni.upb.de

Hanmin Jung
Korea Institute of Science and Technology Information, KISTI
Daejeon, South Korea
jhm@kisti.re.kr

## ABSTRACT

In this paper, we present a system which makes scientific data available following the linked open data principle using standards like RDF und URI as well as the popular D2R server (D2R) and the customizable D2RQ mapping language. Our scientific data sets include acronym data and expansions, as well as researcher data such as author name, affiliation, coauthors, and abstracts. The system can easily be extended to other records. Regarding this, a domain adaptation to patent mining seems possible. For this reason, obvious similarities and differences are presented here.

The data set is collected from several different providers like publishing houses and digital libraries, which follow different standards in data format and structure. Most of them are not supporting semantic web technologies, but the legacy HTML standard. The integration of these large amounts of scientific data into the Semantic Web is challenging and it needs flexible data structures to access this information and interlink them.

Based on these data sets, we will be able to derive a general technology trend as well as the individual research domain for each researcher. The goal of our Linked Open Data System for scientific data is to provide access to this data set for other researchers using the Web of Linked Data. Furthermore we implemented an application for visualization, which allows us to explore the relations between single data sets.

## Categories and Subject Descriptors

D.2.12 [**Interoperability**]: Data mapping
E.2 [**Data Storage Representations**]: Linked representations

## General Terms

Standardization, Languages

## Keywords

Linked Open Data, Researcher Data, Acronym Data, D2R

*: Co-responding author

At KONVENS'14, October 8–10, 2014, Hildesheim, Germany.

## 1. INTRODUCTION

The increase in the number of datasets including scholarly publications in the Linked Data cloud shows the importance of Linked Data for the scientific community. There are many different providers for scientific data available on the Web. Publishing houses, digital libraries, and resellers have popular data sources. The information types range from author information (e.g. full name, affiliation, email), over publications (e.g. title, abstract, coauthors), to specific data like acronyms and their related expansion. Each of these types has own requirements on the data presentation and storage, but they all are somehow interlinked with each other.

One common way to represent this interlinks are network models. This database model is a generalized graph structure without any hierarchical restriction, which allows storing objects with their individual relationships. This is a common way to store data, but does not fit our requirements for modern data publishing.

A more promising way to share data is the Web of Linked Data. The main idea of Linked Open Data (LOD) is to publish free accessible, structured data and to interlink it with other data. This interlinking generates more valuable information under the consequent implementation of standard Web technologies such as RDF or URI. The core benefit for further data exploration is the ability to apply complex graph queries, which allow the interlinking, combining and modifying of data.

For publishing scientific data stored in relational databases, a data bridge is needed. For that reason we present our three-component LOD system for scientific data sets, based on the popular D2R server and the customizable D2RQ mapping language. For further data exploration, we integrated RelFinder, an application that visualizes relationships between RDF objects enables the exploration of data interactively. We will demonstrate the functionalities of our system on a test set.

## 2. RELATED WORK

A lot of work regarding Linked Open Data and the Semantic Web is already done. Popular tools like the D2R server or standards like RDF, HTTP and SPARQL are well proved and used in many LOD systems [1].

A major technological advantage is the backward compatibility, for example, to relational databases (RDB) because the majority of data on the current Web is stored in this kind of databases. The process of mapping RDB to RDF is subject of current research and different approaches like D2RQ, Triplify or R2RML were created [5]. R2RML currently developed by the W3C with the main goal to define a RDB to RDF mapping standard for read-only data access. A different approach is Triplify. It is a very lightweight plugin for existing Web applications, which makes database content available as RDF and other formats.

RDB to RDF mapping can be done by applications such as D2R server. It is a java-based application for publishing the content of relational databases on the Semantic Web and for providing RDF and HTML representations of resources. The D2RQ language is used for the mapping, which is popular because of the possibility to provide access via SPARQL queries very easily [5].

The ability to interlink resources under the use of "typed relationships" allows a goal-oriented navigation trough the database content by web browsers as well as crawler applications [2]. Because these LOD systems and components are very flexible, it is possible to adapt them to different domains. Interlinked User-generated content from social networks, a Linked Open Drug Data (LODD) for pharmaceutical research and development or a LOD live database of semantically enriched sensor data, are only few successful examples, using the D2R server [3].

Latif, Afzal and Maurer [13] utilized the D2R server to publish unstructured datasets of the Journal of Universal Computer Science (J.UCS) as Linked Data in the Web. The linked Open Data project provides a new way to publish machine readable structured data on the Web and best practices for interlinking these structured datasets. Moreover, the increase in the number of datasets including scholarly publications in the Linked Data cloud shows its importance in the scientific community. In order to take advantage of benefits of LOD projects, the legacy HTML data in this journal is converted into machine-readable and structured RDF data using the D2R server. It is considered to be an appropriate for data conversion due to its good performance, scalability and the availability of SPARQL endpoint and explorer features. A RDF graph converted from the legacy HTML data has been made available in Linked Data cloud for the data reuse and interlinking. Moreover, structured journal data was interlinked with Linked Data resources and it successfully disambiguated and interlinked datasets of authors and publications with DBpedia, DBLP and Faceted DBLP as well as CiteULike.

In addition, Mitrevski, Javanovik and Stonjanov [11] identified some issues regarding the D2R server, which appear during the process of publishing the open data of the faculty of computer science and engineering in Ss. Cyril and Methodius University. One problem is that these relational databases include some confidential data about employees and students. However, the D2R server does not provide a way to convert only specific parts from the database into data in a semantic web format. Moreover, it lacks of functionality enabling the user to link existing ontologies to the tables. These issues can be solved by creating a new database called Open Data DB including only the data with no privacy infringement as well as building the mapping tool utilizing functionalities of the D2R server, but linking the data with ontologies. Although a mapping tool was created, the study presents some improvements of this system such as automatic proposal ontology annotations.

For a structured representation of semantically annotated data and a more intuitive exploration, RelFinder has been created by Heim et al. The main idea is that a structured representation of RDF data opens up new possibilieties in the way they can be accessed and queried. For that purpose, a force-directed graph layout which supports every RDF know-ledge base, that provides standardized SPARQL access, is implemented [9]. Relationships between objects can be identified much easier by exploring the data step by step. Features like highlighting, previewing, and filtering are available for further support [4]. The class filter and the link filter allow it to hide objects which contain specific classes or links. Primary these are objects, which are in this particular case not of interest. In addition to these, the length filter and the connectivity filter allow a further selection by hiding objects with a specific amount of links and relationships [9].

RelFinder is often used as stand-alone application but also as integrated visualization application. It is for example part of the DBpedia Viewer, an integrative interface for DBpedia, which combines LodLive as one more different visualization approaches next to RelFinder. The LodLive application is a web-based tool that allows the exploration of Linked Data in an intuiitive and interactive way [10]. In comparison, both applications provide a good visualization. RelFinder can convince with a higher browser compatibility. Especially when using the Internet Explorer, LodLive throws JavaScript errors drawing several relations between classes.

# 3. LOD SYSTEM FOR SCIENTIFIC DATA

The scientific data sets are collected from different resources and based on different formats. For that reasons, it is a difficult task to combine information such as researchers, affiliations or publications and to provide them as a clean, interlinked data set under the requirements of quality. To make a contribution to the existing Web of Linked Data, it is necessary to publish it in a standardized, structured way under the use of common vocabularies and over a common server framework. Furthermore, it has to be identified which information can be bundled to a data set containing relevant data and how this data can be usefully interlinked. For this reason, we developed a LOD system for scientific data sets, which can handle all related information like researcher data or publications and publish them in a structured way under the use of common semantic web technologies like RDF and SPARQL. The interlinked data becomes more useful and can be easily adapted to other research topics.

## 3.1 System architecture

The system architecture in Figure 1 can be divided in three main components. The first component contains the datasets, which include the specific data, as a result of different acquisition and preprocessing steps we applied before. Because these data sets are stored in a relational database, a Linked Data view on the existing database is needed (data bridge). A SPARQL endpoint with the ability for serving Linked Data views on relational databases is D2R Server, which is part of the second step.

We chose D2R, because it is one of the most important and most mature relevant solutions.

The second step, called "Linked Data System", is responsible for the declarative mapping between the schemata of the database and the target RDF terms, based on mapping rules. These rules are stored in mapping files and are formalized under the use of the popular D2RQ mapping language. Each rule defines in detail how resources are identified and how they have to be handled (e.g. find property values) in the SPARQL endpoint. SPARQL is a strong query language for databases, which allows it to access and to modify RDF data. The Endpoint in the third step is able to translate SPARQL based queries into SQL queries, which allows a live database access, even for non-SQL compatible applications. The "Application System" is the third step. It allows exploring the data sets visually. With this application, researchers can find new latent interconnections in the database, like for example an exceptionally usage of rare acronyms by an individual author. This exploring is based on the RDF query language SPARQL from the second step.

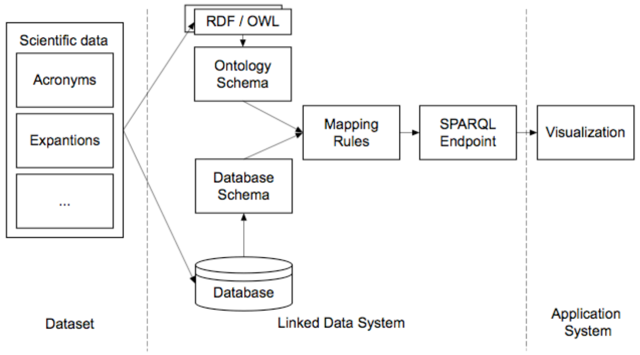


**Figure 1. LOD system architecture for scientific data sets.**

## 3.2 Data Set

Digital resources like the Digital Bibliography & Library Project (DBLP) and publishing houses like Springer, IEEE or Elsevier serve as data providers. Unfortunately, like already mentioned, the data quality is inconsistent between the data providers. For that reason, several refinement steps were applied to the data set, especially for the ambiguity detection. During the work is still in process, we plan to apply more algorithms for named entity disambiguation under the goal of an increasing data set quality. The researcher data set contains 8,370,074 publications like PhD theses, articles or online resources. These publications come with additional information, for example abstracts, author names, coauthors, year of publication and affiliations. Furthermore, we identified acronyms as well as the individual expansion based on publication's abstracts. This information is also part of the researcher data set, because they can support a future disambiguation of researchers. One of our research subjects applied on these datasets is for example the detection of technology trends and the identification of the research domain of individual researchers. Trough the data visualization, we expect to find more latent relationships between objects and classes, which allow us to disambiguate single named entities.

## 3.3 Common vocabularies

Common vocabularies are necessary to enhance the interoperability between concepts. For that reason we use them, as far they already exist and exactly describe the data field.

For the interlinking of the data and the automatic processing, the exact description of an attached common vocabulary is very important. For person related information we use for example the schema.org types and properties.

One example for a prefix is "dc", which is commonly used for the Dublin Core Meta Initiative Terms (http://www.dublincore.org). We introduced "sch" as another prefix, which describes the schemas of schema.org (http://www.schema.org). For the data field "Acronym", "Element type" and "Expansion" no fitting vocabularies are known at the moment. Because of that, we introduced our own "InSciTe" prefix and related properties. They may be replaced during the further working process and should be seen as temporary.

**Table 1. Common vocabulary used for scientific data sets**

| Data field | Property |
|---|---|
| Abstract | pmlp:hasAbstract |
| Acronym | InSciTe:Acronym |
| Affiliation | sch:affiliation |
| Author | dc:creator |
| Coauthor | sch:contributor |
| Editor | sch:editor |
| Editor Email | sch:email |
| Element type (e.g. journal) | InSciTe:ElementType |
| Authors Email | sch:email |
| Expansion | InSciTe:Expansion |
| Source URL | sch:isBasedOnUrl |
| Title | deri:hasTitle |
| Year of publication | sch:datePublished |

## 3.4 Design of URIs

Linked data is based on URIs which identify things and enable users and computer-based agents to refer to these things or look them up. In this case URIs identify entities like researchers and show their relationships to other researchers or publications.

The D2R server is managing the URIs mostly automatically. For the class overview pages, we applied the following structure:

"http://{ip}:{port}/directory/{classname}s"

For individual resources like researcher we applied:

"http://{ip}:{port}/page/{classname}/{id}"

# 4. IMPLEMENTATION

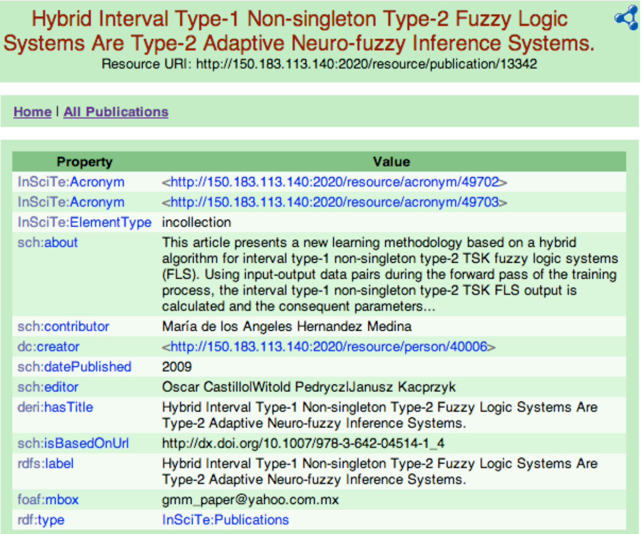
In this section, we explain the current state of the system's implementation, including the D2R server and the RelFinder visualization.

## 4.1 Test data set

For this project, we created a test set of 60 researchers and 400 related publications as well as 454 publication-related acronyms and expansions. Researchers were selected by their number of publications, which has to be at least two. KISTI has diverse scientific data sets, which are derived from papers, patents and others. In order to find more valuable relationships from those data sets, we extracted acronyms and expansions. The number of test data is 491,982. Using these acronyms and expansion we can apply these data sets to analyze technology trend and we can find specific researchers who can be an expert about these acronyms or expansions [7]. The average number of expansion about an acronym is 2.9. We observed that the distribution of expansions follows Zipf's law. We separated each acronym name based on its semantics because an acronym name can be ambiguous. Therefore, each expansion name can have its own acronym name and its own URI. In order to get more valuable analysis reports by using acronyms and their expansions, we have to create relationships between these data sets and other resources such as Linked Open Data, SNS data, Freebase, DBPedia and others. Finally, we get more information from those relationships.

## 4.2 DataHub System based on LOD

The publications view is shown in Figure 2. This view contains information concerning the publication itself, but also further information like coauthors, which have the property "sch:contributor" or acronyms. There are two detected acronyms, which are interlinked with an own class. This class contains the expansion and another related publication, which also contains this acronym in the same meaning. This allows us to find related publications based on acronyms as a first indication for the following classification.



Hybrid Interval Type-1 Non-singleton Type-2 Fuzzy Logic Systems Are Type-2 Adaptive Neuro-fuzzy Inference Systems.

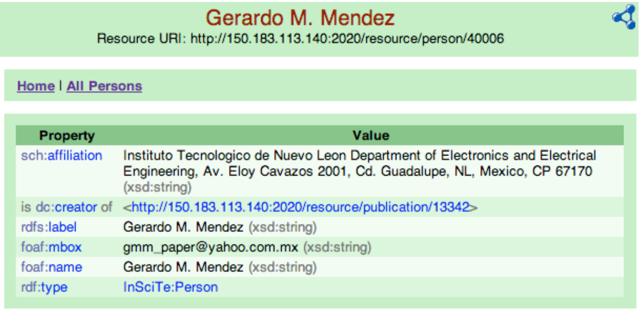
Resource URI: http://150.183.113.140:2020/resource/publication/13342

Home | All Publications

| Property | Value |
|---|---|
| InSciTe:Acronym | <http://150.183.113.140:2020/resource/acronym/49702> |
| InSciTe:Acronym | <http://150.183.113.140:2020/resource/acronym/49703> |
| InSciTe:ElementType | incollection |
| sch:about | This article presents a new learning methodology based on a hybrid algorithm for interval type-1 non-singleton type-2 TSK fuzzy logic systems (FLS). Using input-output data pairs during the forward pass of the training process, the interval type-1 non-singleton type-2 TSK FLS output is calculated and the consequent parameters... |
| sch:contributor | María de los Angeles Hernandez Medina |
| dc:creator | <http://150.183.113.140:2020/resource/person/40006> |
| sch:datePublished | 2009 |
| sch:editor | Oscar Castillo\|Witold Pedrycz\|Janusz Kacprzyk |
| deri:hasTitle | Hybrid Interval Type-1 Non-singleton Type-2 Fuzzy Logic Systems Are Type-2 Adaptive Neuro-fuzzy Inference Systems. |
| sch:isBasedOnUrl | http://dx.doi.org/10.1007/978-3-642-04514-1_4 |
| rdfs:label | Hybrid Interval Type-1 Non-singleton Type-2 Fuzzy Logic Systems Are Type-2 Adaptive Neuro-fuzzy Inference Systems. |
| foaf:mbox | gmm_paper@yahoo.com.mx |
| rdf:type | InSciTe:Publications |

**Figure 2. View on a specific publication (excerpt).**

Furthermore, the email address of the main author is shown. It is part of the publication, because email addresses can change from publication to publication. During the work is still in process, the "sch:editor" data field is not finished yet, because it contains more than one author divided by a pipe symbol. This will be part of the following work. Figure 3 shows for example the person view, which contains information about a specific researcher.



Gerardo M. Mendez

Resource URI: http://150.183.113.140:2020/resource/person/40006

Home | All Persons

| Property | Value |
|---|---|
| sch:affiliation | Instituto Tecnologico de Nuevo Leon Department of Electronics and Electrical Engineering, Av. Eloy Cavazos 2001, Cd. Guadalupe, NL, Mexico, CP 67170 (xsd:string) |
| is dc:creator of | <http://150.183.113.140:2020/resource/publication/13342> |
| rdfs:label | Gerardo M. Mendez (xsd:string) |
| foaf:mbox | gmm_paper@yahoo.com.mx (xsd:string) |
| foaf:name | Gerardo M. Mendez (xsd:string) |
| rdf:type | InSciTe:Person |

**Figure 3. View on a specific researcher**

In this case the researcher has one publication which is interlinked by the "is dc:creater of" property. Furthermore an email address and the affiliation are given. The affiliation data field is subject of our current work, because more standardization and comparability is needed in order to interlink this data field. The acronyms are interlinked within an own class, which contains the expansion and other related publication.

An example graph made by RelFinder's visualization application is shown in Figure 4. Instead of an additional HTTP server like Apache, we implemented the RelFinder application in the already existing D2R web server. That way, no difficult setup is needed in order to start our system – all components are loaded during program's initiation.

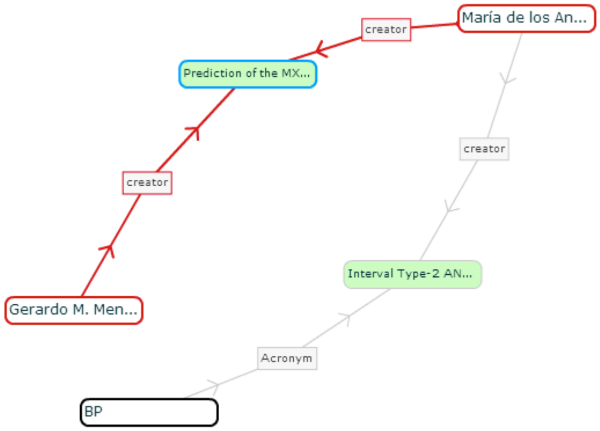


**Figure 4. Data visualization with RelFinder**

Here, we added two researcher resources and one acronym. An edge shows the relation between two RDF objects in a unidirectional way. RelFinder looks up the relations between these resources and draws a graph. In this example, the two researchers have one publication in common (red relation). Or in other words: Both researchers are creators (authors) of this publication. Furthermore the acronym is related to the second researcher trough another paper.

To build this relations, a n:m database table is needed, which contains one identifying ID for the publication and one for the author, which we call 'sequence number'. The D2R server interlinks the affiliation data with the publication data, based on this additional table. This table has to be extended for all publications and researchers as part of the further research.

# 5. TRANSFERABILITY

Because of the open architecture, this system can easily be adapted to other domains. In recent years, patent mining has gained in popularity. Considerations for data acquisition and data providing were investigated considering several aspects [14][15]. The use of RDF technology is also discussed in the area of patent mining and implemented, for instance, as information retrieval system for biomedical patents by Mukherjea and Bamba [14]. Furthermore, it visualizes the connections between patents, but does not allow any user interaction.

The presented LOD system can be applied very well to the patent mining and expand existing approaches by the factor of information integration in the semantic web (data providing). The objects of interest, for example, are inventors, assignees, titles, abstracts etc. This information can be interlinked by relations like "refers", "invented", "assigned" etc. [14]. The semantic representation of patents as well as of academic documents is similar. Both document types can be divided into two parts: The document structure and the content [15]. Common vocabularies are available for both information sources.

## 6. CONCLUSIONS

The main idea of this paper is to unify scientific data such as researcher information, publications and acronyms as well as their expansions from several original resources and to provide them as machine-readable and structured RDF graphs, which allow interlinking and automatic processing. For this reason we introduced a LOD system for scientific data sets.

Based on the data visualization trough RelFinder, the system can further help to identify latent relations between researchers based on publications, acronyms or for example co-authors and further to disambiguate single objects and classes.

As above-mentioned, this is work in progress and we want to apply further algorithms on the data sets to solve existing disambiguation problems, especially in the researcher data set. Additionally we will expand the linked properties between single classes in order to improve the identification of relations between these classes.

It could further be shown that the system is portable due to its flexible adaptation to other domains (e.g. patent mining) although the prototypical implementation was designed for other sources (academic publications).

In the near future, we will publish all data sets for research purposes. It will be made available online via our homepage (http://inscite.kisti.re.kr/ or http://semantic.kisti.re.kr).

## 7. ACKNOWLEDGMENTS

This work was supported by the KISTI [K-14-L02-C03-S03, Development of Technologies for S&T Text Big Data Analytics Application Platform] and a grant from the University of Paderborn, Germany.

## 8. REFERENCES

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

[2] Bizer, C. and Cyganiak, R. 2006. *D2r server - publishing relational databases on the semantic web*. Poster at the 5th International Semantic Web Conference.

[3] Deng, D. P., Mai, G. S., Hsu, C. H., Chang, C. L., Chuang, T. R., Shao, K. T. 2012. *Linking Open Data Resources for Semantic Enhancement of User-Generated Content.* In the book of Semantic Technology. pp. 362-367.

[4] Heim, P., Hellmann, S., Lehmann, J., Lohmann, S. and Stegemann, T. 2009. *RelFinder: Revealing relationships in RDF knowledge bases*. In Semantic Multimedia, pp. 182-187, Springer Berlin Heidelberg.

[5] Hert, M., Reif, G. and Gall, H. C. 2011. *A comparison of RDB-to-RDF mapping languages*. In Proceedings of the 7th International Conference on Semantic Systems, pp. 25-32, ACM.

[6] Jentzsch, A., Zhao, J., Hassanzadeh, O., Cheung, K. H., Samwald, M. and Andersson, B. 2009. *Linking open drug data*. In I-SEMANTICS.

[7] Kim, J., Hwang, M., Jeong, D. H. and Jung, H. 2012. *Technology trends analysis and forecasting application based on decision tree and statistical feature analysis.* In Expert Systems with Applications, 39(16), pp. 12618-12625.

[8] Le-Phuoc, D., Parreira, J. X., Hausenblas, M., Han, Y. and Hauswirth, M. 2010. *Live linked open sensor database*. In *Proceedings of the 6th International Conference on Semantic Systems*, p. 46, ACM.

[9] Lohmann, S., Heim, P., Stegemann, T. and Ziegler, J. 2010. *The RelFinder User Interface: Interactive Exploration of Relationships between Objects of Interest*. In Proceedings of the 15th international conference on Intelligent user interfaces, pp. 421-422, ACM.

[10] Lukovnikov, D., Stadler, C., Kontokostas, D., Hellmann, S., and Lehmann, J. 2014. *DBpedia Viewer-An Integrative Interface for DBpedia Leveraging the DBpedia Service Eco System*. In Proceedings of the 7th Workshop on Linked Data on the Web.

[11] Mitrevski, M., Jovanovik, M., Stojanov, R. and Trajanov, D. 2012. *Open University Data.* In Proceedings of the 9th Conference for Informatics and Information Technology.

[12] Samwald, M., Jentzsch, A., Bouton, C., Kallesøe, C. S., Willighagen, E., Hajagos, J. and Stephens, S. 2011. *Linked open drug data for pharmaceutical research and development.* In Journal of cheminformatics, 3(1), p. 19.

[13] Latif, A., Afzal, M. T. and Maurer, H. A. 2012. *Weaving Scholarly Legacy Data into Web of Data*. J. UCS, 18(16), pp. 2301-2318.

[14] Mukherjea, S. and Bamba, B. 2004. *BioPatentMiner: an information retrieval system for biomedical patents*. In Proceedings of the Thirtieth international conference on Very large data bases.

[15] Ghoula, N., Khelif, K. and Dieng-Kuntz, R. 2007. *Supporting patent mining by using ontology-based semantic annotations*. In Proceedings of the Web intelligence, IEEE/WIC/ACM international conference, pp. 435-438.