





# Research on Fine-Grained Linked Data Creation for Digital Library Resources

Jing Huang<sup>1</sup> , Zhongyi Wang<sup>2</sup> , and Chunya Li<sup>3</sup> 

<sup>1</sup> Wuhan Polytechnic, Wuhan City 430074, Hu Bei Province,  
People's Republic of China  
jianmo0320@hotmail.com

<sup>2</sup> School of Information Management, Central China Normal University,  
Wuhan City 430079, Hu Bei Province, People's Republic of China  
wzywzy13579@163.com

<sup>3</sup> School of Business, Nantong Institute of Technology, Nantong 226002,  
Jiang Su Province, People's Republic of China

**Abstract.** The best practices for publishing linked data have been adopted by an increasing number of libraries, leading to the creation of a global data space—the web of digital library data. However, in library linked data publishing, most of the existing researches mainly focus on structured and semi-structured digital library resources (for example catalogue data). Researches on publishing unstructured digital library resources (for example: contents of papers) are seldom. In order to overcome this problem, this paper proposes a fine-grained linked data creation method to publish the papers stored in digital libraries into linked data. At last, in order to evaluate this method, this paper conducted an experiment on the papers on “text segmentation”. From the experiment results we find that our fine-grained linked data creation method is feasible and will promote the opening access to digital libraries resources.

**Keywords:** Linked data · Creation · Digital library · Fine granularity

## 1 Introduction

Technically, linked data refers to data published on the web in such a way that it is machine-readable. By publishing data on the web according to the linked data principles (Berners-Lee 2009), data providers can add their data to a global data space, which allows data to be discovered and used by various applications. Participants in the early stages of the research on publishing linked data were primarily researchers and developers in university research labs and small companies. Since that time the researches have grown considerably, to include significant involvement from large organizations such as the BBC, Thomson Reuters and the Library of Congress. With an imperative to support novel means of discovery, and a wealth of experience in producing high-quality structured data, libraries are natural complementers to linked data. This field has seen some significant early developments which aim at integrating library catalogs with third party information and at making library data easier accessible by relying on web standards. However, in library linked data publishing, most of the

existing researches mainly focus on structured and semi-structured digital library resources. Researches on publishing unstructured digital library resources are seldom. In order to overcome this problem, this paper proposes a fine-grained linked data creation method to publish the unstructured digital library resources into linked data to promote the opening access to digital library resources.

## 2 Related Work

Linked data is simply about using the web to create typed links between data from different sources. Since linked data was proposed, it has been adopted by an increasing number of data providers, leading to the creation of a global data space connecting data from diverse domains such as people, companies, books, scientific publications, films, music, online communities, and so on. In this paper, we concentrate on works on publishing linked data of digital libraries resources. Until now, works on library linked data publishing can be sectioned into three parts: library linked data projects, publishing methods and publishing tools.

### 2.1 Library Linked Data Projects

There are many projects on digital library linked data publishing. The American Library of Congress and the German National Library of Economics (Neubert 2009) publish their subject heading taxonomies as Linked Data. The Swedish Notional Union Catalogue is also available as Linked data. Similarly, the OpneLibrary publishes its catalogue in RDF, with incoming links from data sets such as ProductDB. Linked Data about scholarly publications is available from the L3S Research Center, which hosts a Linked Data version of the DBLP bibliography. The ReSIST project publishes and interlinks bibliographic databases such as the IEEE Digital Library, CiteSeer, and various institutional repositories. The RDF Book Mashup, a wrapper around the Amazon and Google Base APIs, provides Linked Data about books. The Open Archives Initiative has based its new Object Reuse and Exchange standard (OAI-ORE) on the Linked Data principles; this standard's deployment is likely to further accelerate the availability of Linked Data related to publications (Bizer 2009).

### 2.2 Publishing Methods

There are many practical recipes for publishing different types of information as linked data on the web. The simplest way to serve linked data is to produce static RDF files, and upload them to a web server. This approach is typically chosen in situations where the RDF files are created manually, and the RDF files are generated or exported by some piece of software that only outputs to files. However, if your data is stored in a relational database it is usually using D2R Server to publish a linked data view on your existing data base. D2R server relies on a declarative mapping between the schemata of the database and the target RDF terms. Based on this mapping, D2R Server serves a Linked Data view on your database and provides a SPARQL endpoint for the database. What's more, in view of data sources available on different kinds of Web APIs, it is

often to implement linked data wrappers to publish linked data such as RDF Book Mashup which makes information about books, their authors, reviews, and online bookstores available as RDF on the Web (Bizer et al. 2007).

A variety of linked data publishing tools has been developed. The tools either serve the content of RDF stores as linked data on the web or provide linked data views over non-rdf legacy data sources. The tools shield publishers from dealing with technical details such as content negotiation and ensure that data is published according to the linked data community best practices (Sauermann and Cyganiak 2008; Berrueta and Phipps 2008; Bizer et al. 2007). All tools such as D2R Server (Bizer and Cyganiak 2006), Triplify (Auer et al. 2009), SparqPlug (Coetzee et al. 2008), etc. support dereferencing URIs into RDF descriptions. In addition, some of the tools such as Virtuoso Universal Server, Talis Platform, Pubby (Cyganiak and Bizer 2006) etc. also provide SPARQL query access to the served data sets and support the publication of RDF dumps.

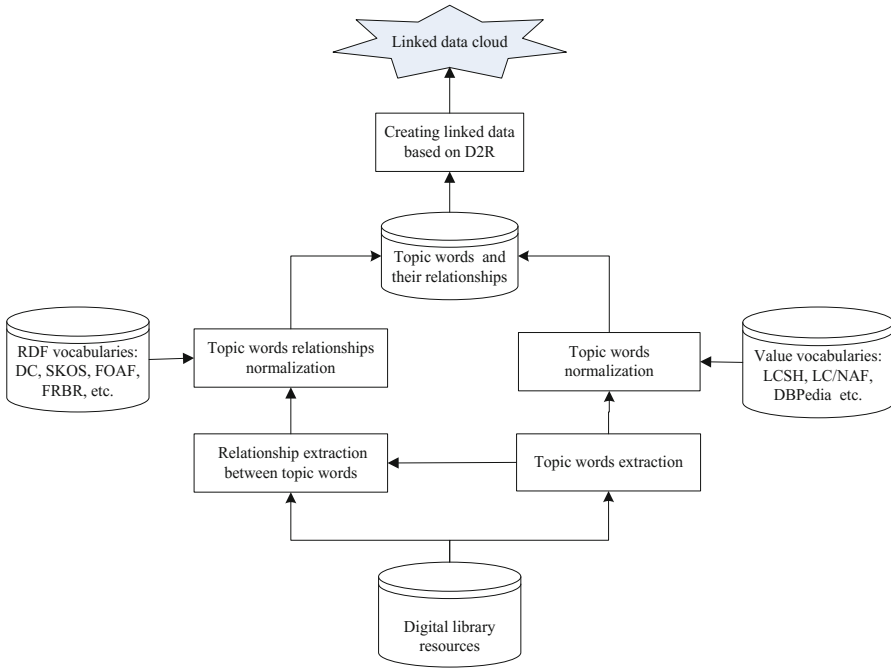
From the above discussion we can see that in library linked data publishing, most of the existing researches mainly focus on structured and semi-structured digital library resources. However, besides these two kinds of digital library resources, digital library has more unstructured resources. Although most of the digital library resources are unstructured, there are seldom research works on publishing them as linked data, which prohibits library users to fully access and explore them dramatically. In order to overcome this problem, this paper proposes a fine-grained linked data creation method to publish the unstructured digital library resources into linked data.

### 3 Fine-Grained Linked Data Creation Method

In this paper, the procedure of fine-grained linked data creation can be considered as a process of the transformation from unstructured textual information to structured data. Steps of the fine-grained linked data construction process (see Fig. 1) include: “topic words extraction”, “topic words normalization”, “Relationship extraction between topic words”, “topic words relationships normalization” and “Creating linked data based on D2R”.

#### 3.1 Topic Words Extraction

In this paper, topic words refer to nominal terms that can identify main themes in the document set. In order to realize the topic words extraction from digital libraries resources, text mining technique latent semantic analysis (LSA) is adopted. LSA is a technique in natural language processing of analyzing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. The basic idea of extracting topic words based on LSA is that: firstly topic information is gained by LSA statistical topic model; then words are graded according to this information, and at last words with high grade are selected as the topic words to identify documents. Specifically, the algorithm for extracting topic words based on LSA can be illustrated in Table 1. These extracted topic words will be used as entities in the next step of “3.3 relationship extraction between topic words”.



**Fig. 1.** The process of linked data construction

**Table 1.** The algorithm for words extraction based on LSA

Input: a set of documents

Output: topic words

1. Sentence splitting, break each document into a list of sentences using a heuristically-based approach on the basis of punctuation (“./!/?”)
2. Word segmentation and POS-tagging, use the Stanford Segmenter to segment each sentence into a list of words which are parsed with POST tags
3. Words filtering, select noun words as candidate topic words, to generate  $d = (w_1, w_2, \dots, w_n)$  where  $w_n$  is the tf-idf value of the  $n$ th word in the document  $d$
4. Generate  $D = (d_1, d_2, \dots, d_m)$  which is a collection of  $m$  documents
5. Singular value decompose, apply SVD to yield  $A$  according to the equation  $A = U\Sigma V^T$ , where  $X^T$  denotes the transposed matrix of  $X$ . The columns of  $U$  and  $V$  are the eigenvectors of  $AA^T$  and  $A^T A$ , respectively. The diagonal values of  $\Sigma$  are the corresponding singular values which are sorted in descending order
6. Extracting topic words, extract topic words from the first  $k$  columns of  $U$  according to the diagonal values of  $\Sigma$ . Because the eigenvectors in  $U$  are the principle axes for distinguishing the word feature vectors in  $AA^T$  which is a word similarity matrix where the meaning of a word  $w_i$  is expressed in terms of its dot-product with all other words  $\{w_1, \dots, w_n\}$ , these extracted  $k$  words can be viewed as topic words to reveal the themes of documents in a  $k$ -dimensional space

### 3.2 Topic Words Normalization

Since topic Words extracted through the above process are uncontrolled words, it is common for the phenomenon of synonyms and homonyms to occur. In order to overcome these problems, there is a need to translate these extracted topic words into controlled words to achieve the purpose of topic words normalization. Therefore, translating topic words into thesauri is a good solution. In this paper, different kinds of value vocabularies (for example: LCSH, LC/NAF, DBPedia etc.) are used during the process of words normalization. Each of the above extracted topic words are covert to its corresponding thesauri through the projection between words and value vocabularies. However, since it is hard and laborious to ensure the currency of the thesaurus. Some new topic words may have no their corresponding thesauri in value vocabularies. In this case, experts are invited to assign thesauruses for them.

### 3.3 Relationship Extraction Between Topic Words

After topic words normalization, the next step is extracting relationship between them. A relationship extraction task requires the detection of semantic relationship between topic words from digital libraries resources. This paper extracts relationships between the above extracted topic words by using syntactic dependency patterns. The procedure of relationship extraction (see Fig. 2) is as follows.

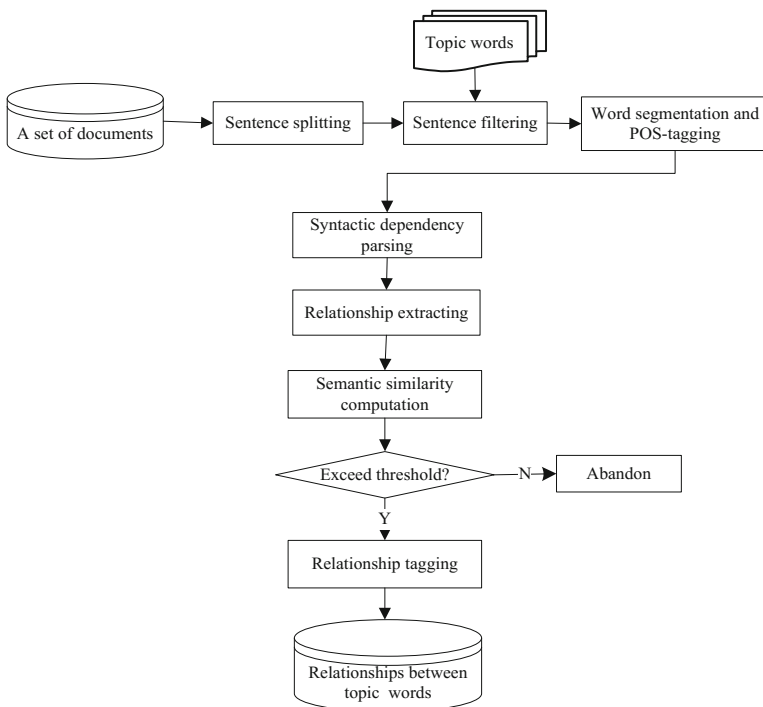


Fig. 2. The process of relationship extraction between words

**Sentence splitting.** As the task of sentence splitting has been done in the topic words extraction (see Sect. 3.1 Topic words extraction), we will not explain it in this part.

**Sentence filtering.** Irrelevant sentences which just contain one or null of the above extracted topic words are filtered out. And only sentences including two or more the above extracted topic words are selected as relevant sentences.

**Word segmentation and POS-tagging.** As the task of word segmentation and POS-tagging has been done in the topic words extraction (Sect. 3.1 Topic words extraction), we will not explain it in this part.

**Syntactic dependency parsing.** The task of syntactic dependency parsing is to encode syntactic structure with labeled directed arcs (dependencies) between the headwords of constituents and generate a dependency tree for each sentence, by using of the Stanford Parser.

**Sentence skeleton extracting.** The task of sentence skeleton extracting is to simplify dependency trees of sentences. As we focus on verb relationships, so in this paper the process of sentence skeleton extracting includes collecting for each verb its subject, object, preposition with arguments and auxiliary verb.

**Relationship extracting.** The task of relationship extracting is to extract relation triples from sentence skeleton based on their dependency relationships.

**Semantic similarity computation.** The task of semantic similarity computation is to calculate the semantic similarity between topic words  $w_i$  and  $w_j$  ( $M(w_i, w_j)$ ) by use of the formula (1), where  $\Lambda_k$  is the  $k$ -dimensional LSA space for  $D(d_1, d_2, \dots, d_m)$ , the  $i$ -th row in  $\Lambda_k$ , or  $\Lambda_k(i)$  is the LSA feature vector for word  $w_i$ .

$$M(w_i, w_j) = \cos(w_i, w_j) = \Lambda_k(i) \times \Lambda_k(j) / \sqrt{\Lambda_k(i)^2 \times \Lambda_k(j)^2} \quad (1)$$

**Relationship selecting.** If the value of semantic similarity exceeds the threshold level, the relationship between these two topic-words will be extracted as their semantic relationship, otherwise, abandon this relationship.

**Relationship tagging.** The task of relationship tagging is to label these selected relation triples with their predicates.

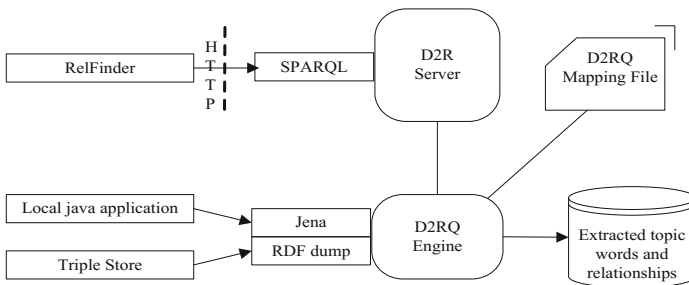
### 3.4 Topic Words Relationships Normalization

Since the relationships between topic words extracted through the above process are also uncontrolled terms, it is also very common for the phenomenon of synonyms and homonyms to occur. In order to overcome these problems, there is a need to normalize these extracted relationships. In this paper, the normalization of topic words relationships is realized through the projection between terms describing extracted relationships and RDF vocabularies (for example DC, SKOS, FOAF, FRBR, etc.). RDF vocabularies provide a controlled list of properties for describing the relationship between topic words. Since RDF vocabularies are also controlled vocabularies which

need people to maintain them, it is hard to ensure the currency of them. Therefore, in this paper, experts are invited to assign a term to normalize the extracted relationships which are not included in the RDF vocabularies.

### 3.5 Creating Linked Data Based on D2R

Through the above steps, we have obtained topic words and their relationships, next we will create linked data based on them by using of the linked data construction tool: D2R. The procedure of linked data construction based on D2R (see Fig. 3) is as follows.



**Fig. 3.** Constructing linked data based on D2R

First, entity denomination, use the D2R Server to assign an URI to each topic words, which can be used to locate and search each entity. Second, constructing RDF, use the customizable D2RQ mapping to map each topic keyword and their relationships into RDF format. Third, publishing linked data, associate them with outside linked data cloud through D2R server to make full use of outside information resources revealing topic words' meanings. Last, linked data visualization, linked data is visualized by RelFinder which can be used to extract and visualize relationships interactively explorable. RelFinder is based on the open source framework Adobe Flex, easy-to-use and works with any RDF dataset that provides standardized SPARQL access.

## 4 Experiment

### 4.1 Data Source

In order to validate the fine-grained linked data creation method proposed in this paper, we conducted an experiment on papers of the “Linked data” research field which is used to describe a recommended best practice for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using URIs and RDF. We selected the “Linked data” as the test subject mainly for two reasons. First, as the “Linked data” is a new research field, the total number of related papers is not very big. So it is relatively easy for experts to evaluate the results of our experiment on this

research field. Second, as the “Linked data” is an interdisciplinary research field, the semantic relationships between words are often very complex.

To ensure that all relevant papers can be collected as far as possible, this paper takes two mainstream databases (Web of Science, Engineering Index) as the data source. Web of Science Core Collection provides researchers, administrators, faculty, and students with quick, powerful access to the world’s leading citation databases. Engineering Index (EI) was founded in 1884 by Dr. John Butler Johnson. EI is the broadest and most complete engineering literature database available in the world. It provides a truly holistic and global view of peer reviewed and indexed publications with over 17 million records from 73 countries across 190 engineering disciplines. By using EI, engineers can be confident information is relevant, complete, accurate and of high quality.

Then, the retrieval strategy: (Title = linked data) was used to conduct the retrospective searching in these two databases. Since linked data was first proposed by Tim Berners-Lee at 2009, we set the year in the retrieval as 2006–2015. Finally, removing the unrelated and non-academic papers, we got 1045 articles.

## 4.2 Experimental Results

According to the procedure of fine-grained linked data creation method proposed in this paper, we implemented a fine-grained linked data creation system on the Windows XP platform by means of JAVA, JSP development language. The entire system can be divided into two parts: linked data creation and linked data visualization. The function of linked data creation is to translate unstructured digital libraries resources into RDFs; and the job of linked data visualization is to provide users with a customizable interface to interact with our system. As shown in Fig. 4, users only need to click “add” button to add topic words into the input box in the upper part of the sidebar. These user-given topic words are then mapped to unique objects of the knowledge base by executing an automatic or manual disambiguation, and serve as starting nodes in a graph that is drawn in the presentation area of the user interface. The links between the topic words are visualized as labeled and directed edges in accordance with their representation in the knowledge base. Besides, the sidebar offer other sophisticated functionality for the interactive exploration of the found relationships. For example, the sidebar offers four types of filters (class filter, link filter, length filter and connectivity filter) that facilitate the exploration of the graph visualization by highlighting or removing certain elements.

Through the interface of linked data, users can explore any topic words and their relationships by entering them into the input box and clicking the “find relations” button. For example, if we add some topic words on “linked data” such as “semantic web” and “linked data” into the input box, then click “Find Relations”, parts of linked data on “linked data” will be displayed (see Fig. 4).



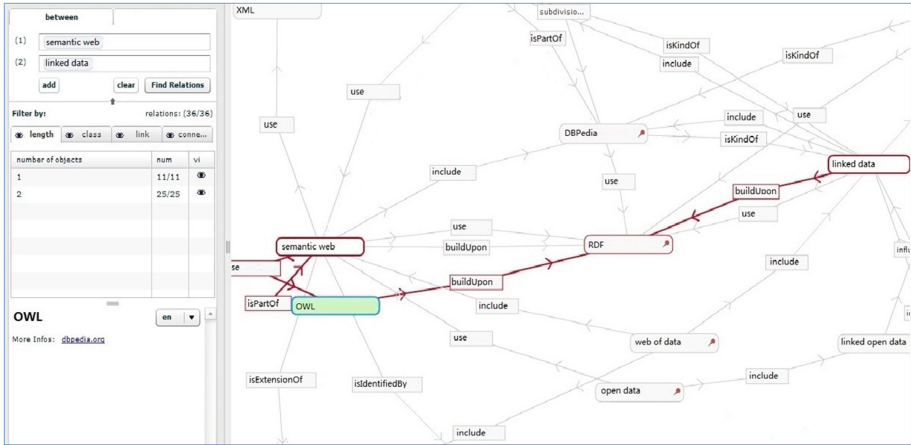


Fig. 4. Interface of linked data

## 5 Conclusion and Future Work

In order to publish the unstructured digital libraries resources as linked data based on the linked data principles and practices, this paper has done a lot of researches on it. Specifically, the main contributions of our study are threefold. First, we have analyzed limitations in library linked data publishing and pointed out that most of the existing researches mainly focus on structured and semi-structured digital library resources such as catalogue, subject heading taxonomies and so on, while researches on publishing unstructured digital libraries resources such as the text of digital books, papers are seldom. Second, this paper, in order to overcome this limitation, proposes a fine-grained linked data creation method to publish the unstructured digital library resources into linked data to promote the opening access to digital library resources. Third, using this method, the unstructured resources on the field of “linked data” is published as linked data. It is useful and meaningful to publish existing literatures in libraries as linked data. That is because most of library users have neither the time nor inclination to sift through long documents for small pieces of useful knowledge. While publishing the unstructured digital library resources into linked data can facilitate document fragment retrieval and support the delivery of the right knowledge in the right quantity.

Of course, with the application and development of library linked data, an increasing amount of libraries resources will be published and linked in the form of linked data. This trend certainly will bring new opportunities to libraries to improve their ability to serve their users.

**Acknowledgment.** This study is supported by MOE (Ministry of Education in China) Project of Humanities and Social Science: “Research on the Multi-granularity Hierarchical Topic-based Segmentation of the Digital Library Resources” (Project No. 16YJC870003).

## References

- Auer, S., Dietzold, S., Lehmann, J., Hellmann, S., Aumueller, D.: Triplify: light-weight linked data publication from relational databases. In: 18th Proceedings of International Conference on World Wide Web, pp. 621–630. IEEE, Madrid (2009)
- Berners-Lee, T.: Linked Data (2009). <http://www.w3.org/DesignIssues/LinkedData.html>. Accessed 10 June 2015
- Berrueta, D., Phipps, J.: Best Practice Recipes for Publishing RDF Vocabularies - W3C Working Group Note (2008). <http://www.w3.org/TR/swbp-vocab-pub/>. Accessed 14 June 2015
- Bizer, C.: The emerging web of linked data. *Intell. Syst.* **24**(5), 87–92 (2009)
- Bizer, C., Cyganiak, R.: D2R server-publishing relational databases on the semantic web. In: 5th Poster of International Semantic Web Conference, pp. 360–369. Georgia Center for Continuing Education, Athens (2006)
- Bizer, C., Cyganiak, R., Heath, T.: How to Publish Linked Data on the Web (2007). <http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/>. Accessed 14 June 2016
- Coetzee, P., Heath, T., Motta, E.: SparqPlug: generating linked data from legacy HTML, SPARQL and the DOM. In: 1st Proceedings of Workshop on Linked Data on the Web (LDOW 2008), Beijing, China (2008)
- Cyganiak, R., Bizer, C.: Pubby - A Linked Data Frontend for SPARQL Endpoints. <http://www4.wiwiss.fu-berlin.de/pubby/>. Accessed 14 June 2016
- Neubert, J.: Bringing the “Thesaurus for Economics” on to the web of linked data. In: Proceedings of WWW Workshop on Linked Data on the Web, Madrid, Spain (2009)