# LinkedCJ: A Knowledge Base of Chinese Academic Journals Based on Linked Data

Peng Xu[1], Xin Wang[1], and Haofen Wang[2]

[1] School of Computer Science and Technology, Tianjin University, Tianjin, China
{wxlyf,wangx}@tju.edu.cn
[2] Department of Computer Science and Engineering, East China University of Science and Technology, Shanghai, China
whfcarter@ecust.edu.cn

**Abstract.** Nowadays the amount of academic articles published in Chinese is growing rapidly, however, existing methods of managing and querying Chinese academic journals and articles are not semantic-based. Our work consists of creating an ontology for represents and organizing bibliographic information of Chinese academic journals and articles. Moreover, we develop software applications based on Nutch, jsoup, and Drools for transforming millions of Web pages from website of Wanfang into approximately 15 million triples stored in a triple store. Finally, the knowledge base is evaluated using the Semantic Service Platform of Chinese Academic Journals and Articles (SSPCAJA). Results of the functional test show that information of Chinese academic journals and articles is effectively represented on the platform.

**Keywords:** linked data, knowledge base, Chinese academic journals, ontology

## 1 Introduction

The three major Chinese Web publishers, VIP, Wanfang, and CNKI have embodied over 38 million, 20 million, and 36 million in articles respectively, as well as 12000, 7000, and 8000 in journals, until July 2013. However, existing methods on managing and querying Chinese academic journals and articles are not semantic-based.

The Semantic Web, which stores all information in the form of Linked Data[1] instead of hyperlinked Web pages, focuses on semantic interpretation of the data on the Web. At present, leading publishers, such as NPG, DBLP, and CrossRef, have been organizing semantic data from journals on the basis of linked data. But most of the vocabularies in NPG, DBLP, and CrossRef cannot adapt to Chinese academic journals because the structure of knowledge organization cannot be reused directly. However in China, semantic-based knowledge organizations for Chinese academic journals have not yet been constructed, while the popular alternatives are data centralizing platforms, such as C-DBLP and Not Old academic search.

In this paper, we seek to organize data from several Chinese academic journals on the basis of linked data. The main contributions of this paper are:

– we build the LinkedCJ ontology for representing information of Chinese academic journals;
– we develop a method for extracting RDF triples from the Web pages of Wanfang using a series of tools including Nutch plugins, jsoup, and Drools;
– we conduct a set of functional tests on the SSPCAJA using LinkedCJ knowledge base and compare SSPCAJA with linked data platforms of NPG and DBLP to evaluate the LinkedCJ knowledge base.

The rest of the paper is organized as follows. Section 2 describes the construction of the LinkedCJ ontology. Section 3 gives the method for extracting RDF triples from the crawled Web pages of Wanfang. The functional test and evaluation of the knowledge base are presented in Section 4. Finally, Section 5 concludes the paper and gives the future work.

## 2 LinkedCJ Ontology

In order to organize and construct the knowledge base of Chinese academic journals, it is necessary to create an ontology according to the characteristics of Chinese academic journals.

LinkedCJ, our new ontology for Chinese academic journals, inherits a list of classes, object properties and data properties from existing ontologies of semantic publishing, such as Functional Requirements for Bibliographic Records (FRBR)[2], FRBR-aligned Bibliographic Ontology (FaBiO)[3], Citation Typing Ontology (CiTO)[4], Dublin Core Metadata (DC)[5] and The Friend of a Friend (FOAF). LinkedCJ adds several new classes, object properties and data properties in order to indicate exclusively information of Chinese academic journals. Fig.1 shows the top level classes and object properties of LinkedCJ.

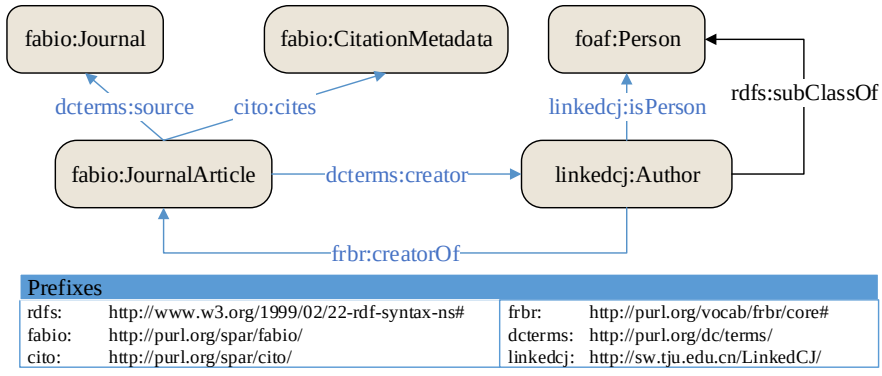


**Fig. 1.** Top level classes and object properties of LinkedCJ

LinkedCJ ontology contains 5 classes, 5 object properties, and 35 data properties. A list of original classes, object properties, and data properties, such as linkedcj:cn, linkedcj:hasProject, and linkedcj:hasSecondarySubjectTerm, take part in the ontology

for representing items with Chinese characteristics that does not include in existing ontologies, such as FRBR, FaBiO, CiTO, DC, and FOAF.

In summary, the vast majority of the information from Chinese academic journals and articles could be represented by LinkedCJ normatively.

## 3 Triple Extraction

In order to construct the knowledge base, we develop software applications based on Nutch, jsoup and Drools for transforming HTML pages from website of Wanfang into triples stored in a triple store. Fig. 2 shows the whole process of triple extraction.

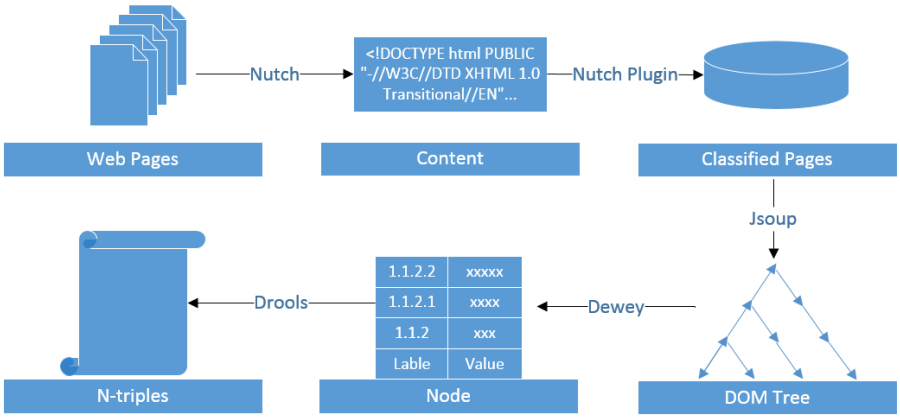


**Fig. 2.** The process of triple extraction

Wanfang is the data resource of our knowledge base. Triples have been gathered from the 13 journals of China Computer Federation (CCF), such as the Chinese Journal of Computers. After approximately 300000 pages fetched in this way, we have acquired a copy of HTML raw data including approximately 48000 journal articles, 590000 citations and 180000 authors. Finally, an N-triples (.nt) file containing about 15 million triples has been generated.

## 4 Evaluation and Comparison

In order to evaluation the LinkedCJ knowledge base, we have performed several test cases on the Semantic Service Platform of Chinese Academic Journals and Articles (SSPCAJA). For the purpose of comparison, two leading linked data service platforms, NPG and DBLP, have been tested at the same time.

Table 1 shows the summary of the functional tests. Symbol "○" means the function or information is provided, while symbol "×" means the function or information is not provided. Maximum kinds of supported queries, as well as the personal name disambiguation mechanism, are the main advantages of the LinkedCJ knowledge base.

**Table 1.** The results of the functional test on query types.

| | Query type | SSPCAJA | NPG | DBLP |
|---|---|---|---|---|
| | SPARQL | ○ | ○ | × |
| By title | Show start page | × | ○ | × |
| | Show end page | × | ○ | × |
| | Show secondary subject term | ○ | × | × |
| | Show project | ○ | × | × |
| | Show creator‘s working unit | ○ | × | × |
| By author | Search author | ○ | ○ | ○ |
| | Personal name disambiguation | ○ | × | × |
| By journal | Search journal | ○ | × | ○ |
| | Sorted by volume and number | × | × | ○ |
| | By keywords | ○ | × | × |

Other test cases have revealed that LinkedCJ is superior to NPG and DBLP in representing information of Chinese academic journals.

## 5 Conclusions

In this paper, we present the design and implementation of LinkedCJ, the knowledge base of Chinese academic journals. LinkedCJ knowledge base is evaluated using the SSPCAJA. Results of the test cases show that information of Chinese academic journals and articles is correctly represented by the knowledge base.

**Acknowledgements.** This work is supported by CCF Opening Project of Chinese Information Processing (Grant No. CCF2013-02-02) and the National Natural Science Foundation of China (Grant No. 61100049).

## References

1. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems, 5(3), 1-22 (2009)
2. IFLA Study Group on the Functional Requirements for Bibliographic Records.: Functional Requirements for Bibliographic Records: Final Report. UBCIM publications, München (1998)
3. Peroni, S., Shotton D.: FaBiO and CiTO: Ontologies for describing bibliographic resources and citations. Web Semantics: Science, Services and Agents on the World Wide Web. 17, 33-43 (2012)
4. Shotton, D.: CiTO, the Citation Typing Ontology. Biomedical Semantics. 1(S-1) S6 (2010)
5. Kurtz, M.: Dublin Core, DSpace, and a brief analysis of three university repositories. Information Technology and Libraries. 29(1), 40-46 (2013)