



天津大学  
Tianjin University

CCF 2013 开放课题汇报  
课题编号: CCF2013-02-02

# 基于Linked Data的中文学术期刊 语义知识组织与服务关键技术研究

课题负责人: 王鑫

wangx@tju.edu.cn

天津大学 计算机科学与技术学院

2014. 12. 09 @ NLPCC 2014





天津大学



## 一、课题背景 ←

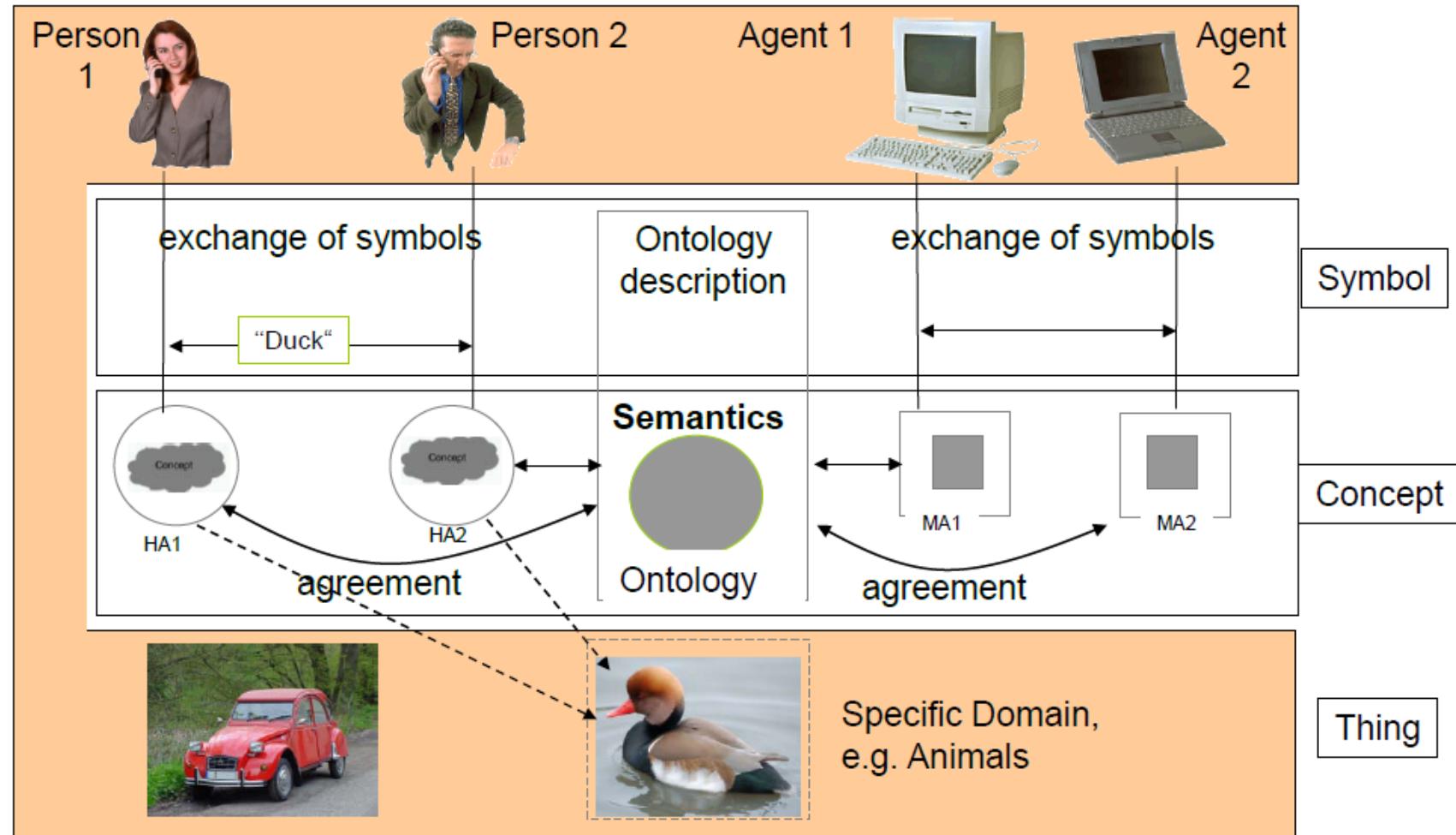
## 二、中文学术期刊本体与知识库

## 三、中文学术期刊语义知识服务

## 四、原型系统与功能演示



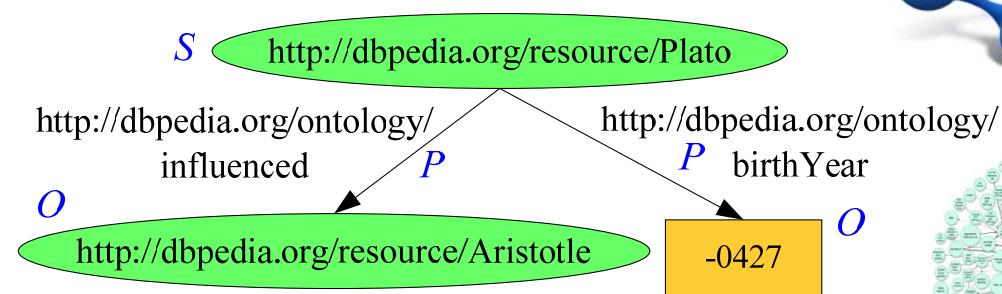
## ► 语义万维网(Semantic Web)的基本思想





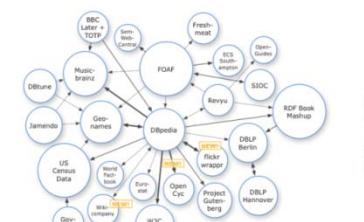
## ➤ Linked Data: 语义大数据

- RDF: 资源描述框架

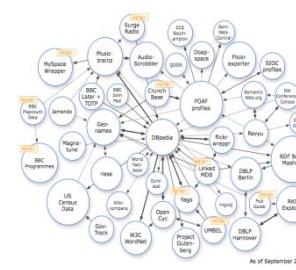


- Linked Data

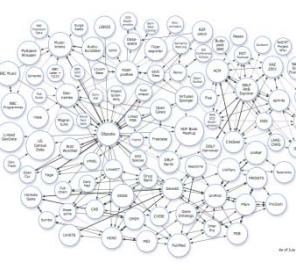
百科知识  
地理信息  
生物医学  
媒体出版  
电子政务



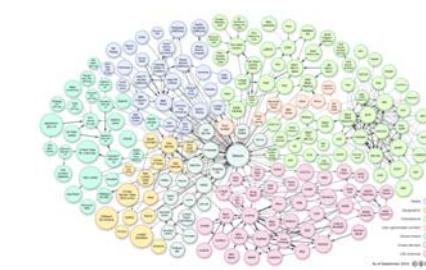
2007.10



2008.09

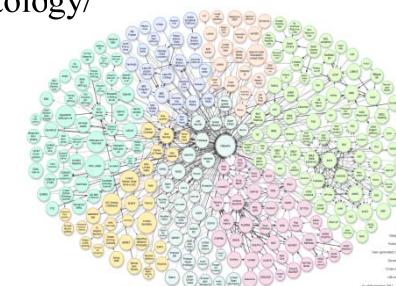


2009.07

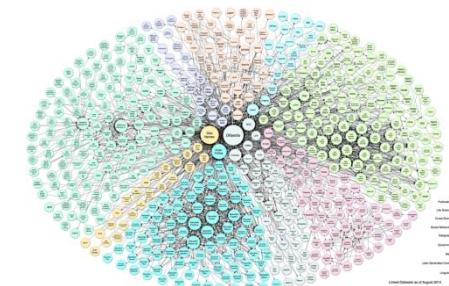


2010.09

.....



2011.09



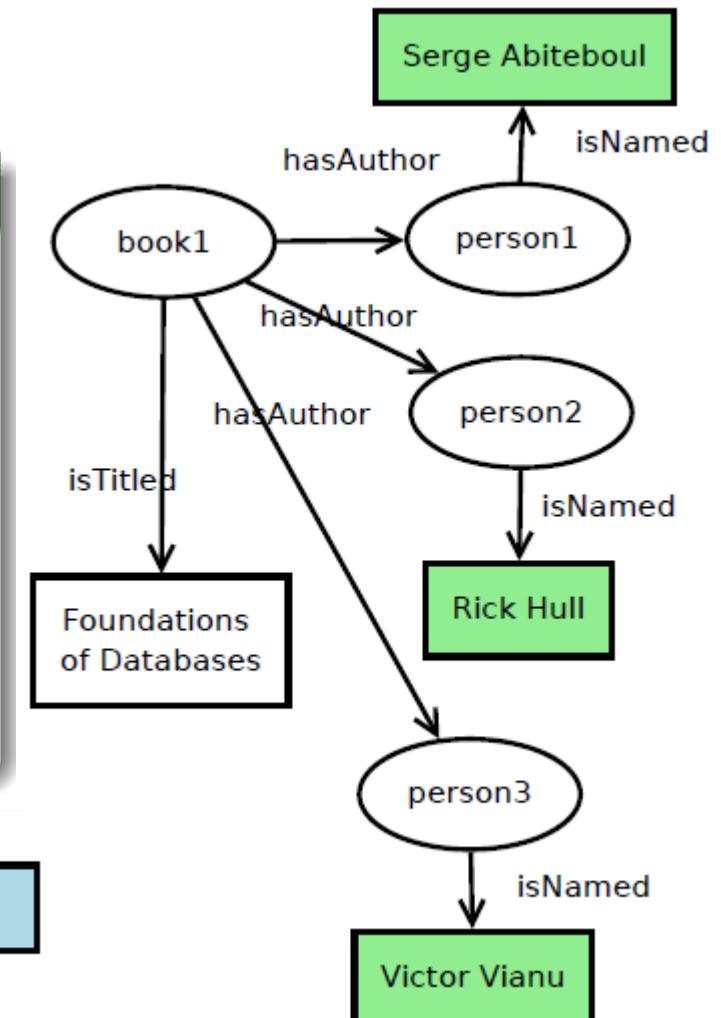
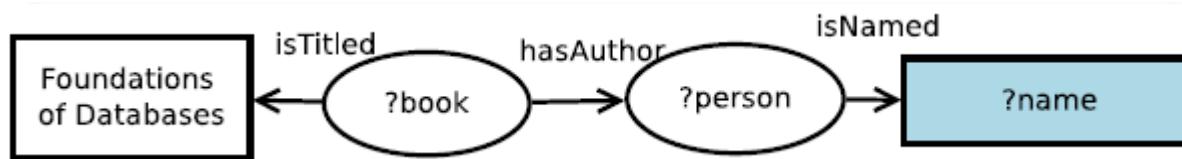
2014.08



## ➤ RDF查询语言：SPARQL

### 举例

```
SELECT ?name  
WHERE {  
    ?book isTitled "Foundations of Databases".  
    ?book hasAuthor ?person .  
    ?person isNamed ?name .  
}
```





## ➤ Linked Data与数字出版

- Linked Data四原则
  1. 用URI指代实体
  2. 用HTTP URI使实体可引用，可被人或Agent查看
  3. 查看URI时提供关于实体的有用信息
  4. 在Web上发布数据时包含实体到其他相关实体的链接

——Tim Berners-Lee

- Linked Data数字出版
  - 知识库、服务、共享

nature.com linked data



万方数据  
WANFANG DATA  
知识服务平台

VIP  
维普网  
仓储式在线出版平台

Cnki 中国知网  
cnki.net



## ► 课题研究内容

- 1. 基于Linked Data的中文学术期刊与文献知识库构建**
- 2. 基于Linked Data知识库的Web知识服务API构建**
- 3. 基于Linked Data的中文学术期刊与文献语义服务平台**



天津大学



## 一、课题背景

## 二、中文学术期刊本体与知识库



## 三、中文学术期刊语义知识服务

## 四、原型系统与功能演示



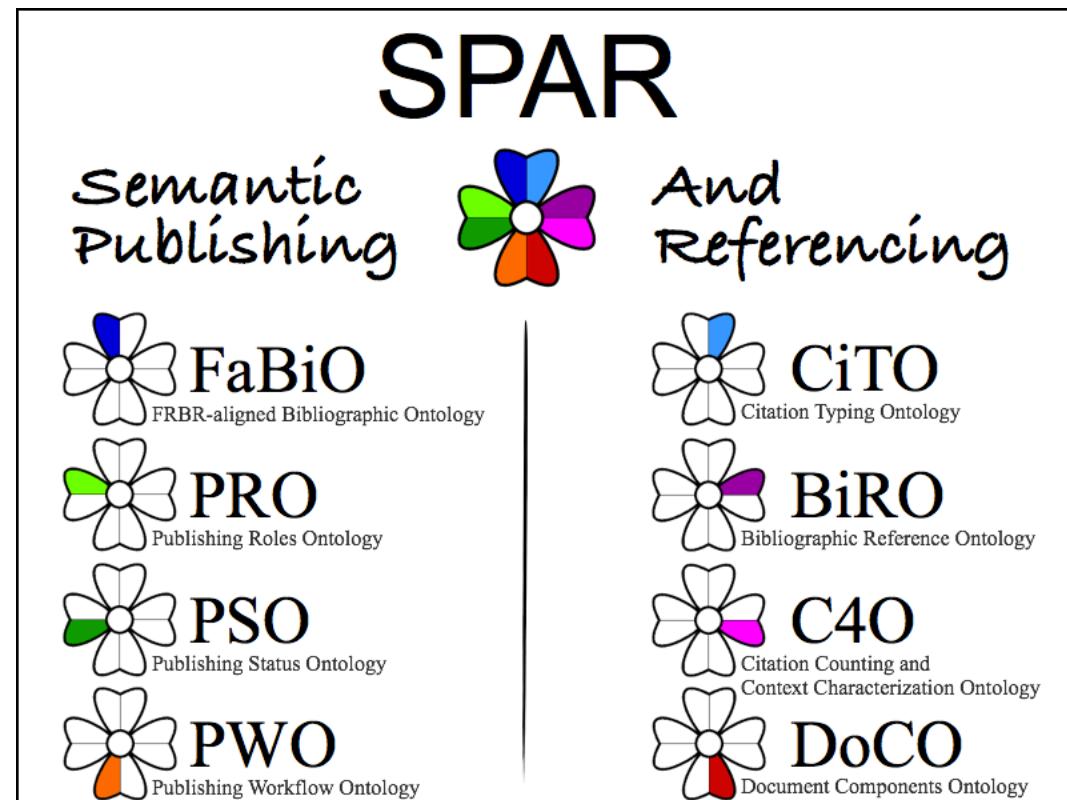
天津大学



## ➤ 中文学术期刊本体构建

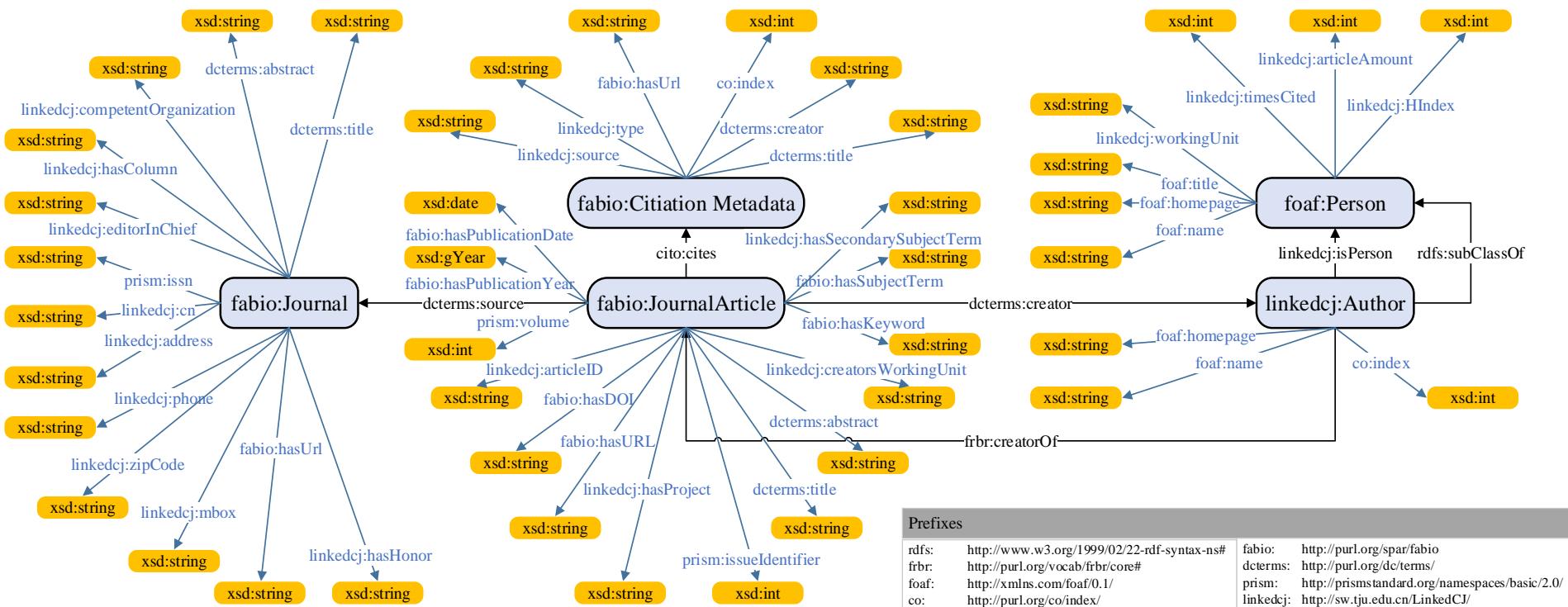
□ 原则：尽量**复用**国际上已有的语义出版本体

- FaBio
- CiTO
- Dublin Core
- FOAF
- PRISM
- ...



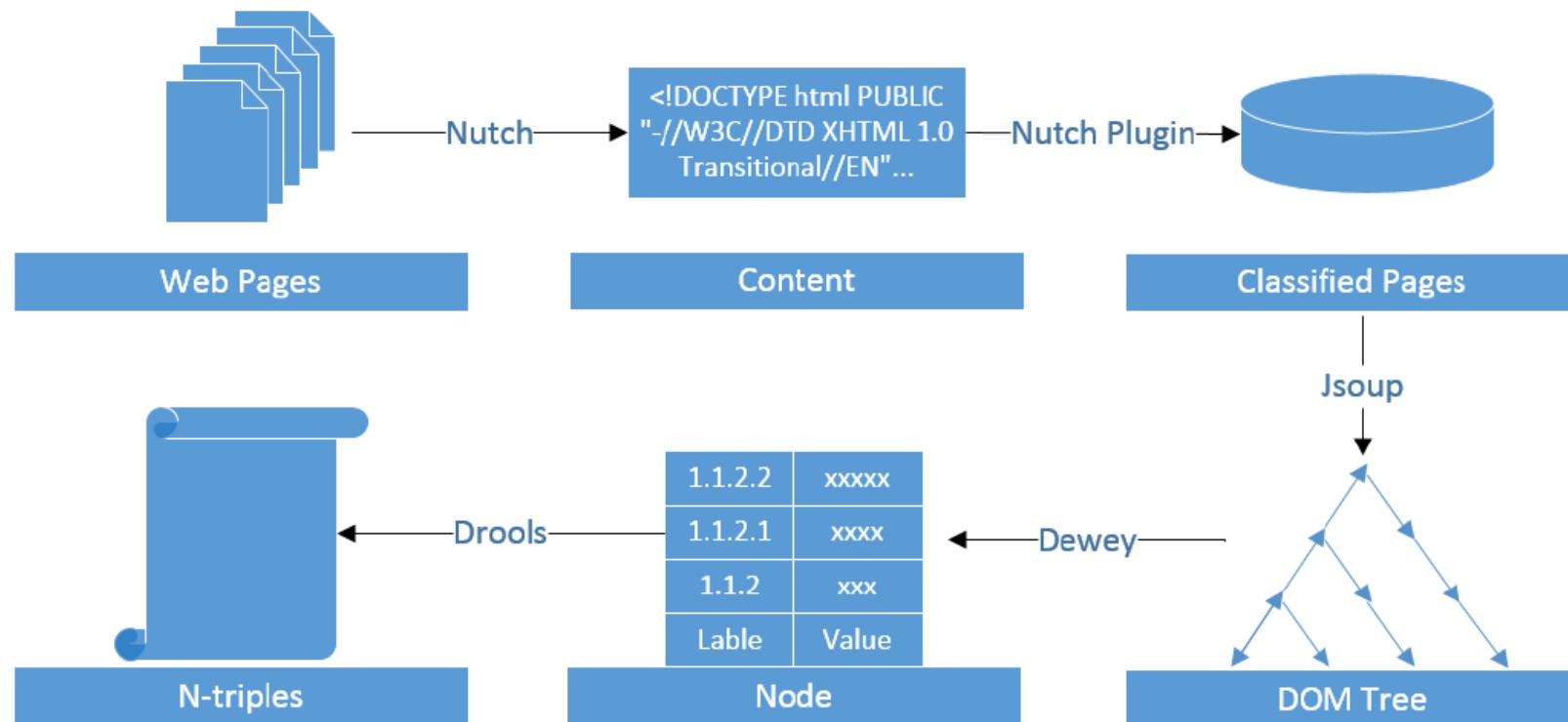


## ► 中文学术期刊本体构建 □ LinkedCJ本体：融入中文学术期刊特点





## ➤ 中文学术期刊知识库构建 □ 知识库生成





天津大学

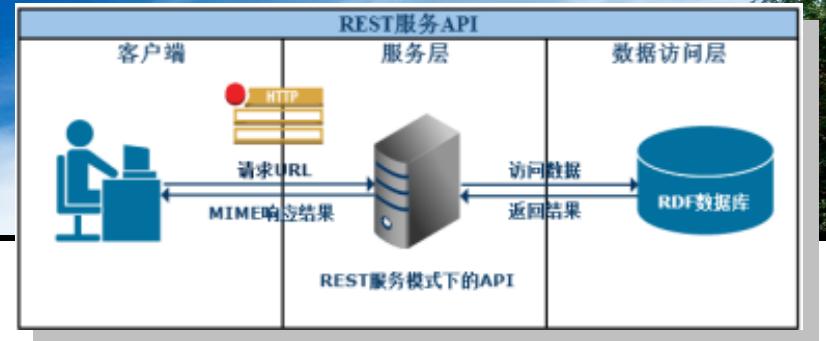


## 一、课题背景

## 二、中文学术期刊本体与知识库

## 三、中文学术期刊语义知识服务 ←

## 四、原型系统与功能演示



## ➤ 中文学术期刊知识服务API □ 基于REST服务的知识服务API

Linked Data Platform

方法名	方法描述	资源 URI
queryArticleByDoi()	根据某篇文献的 Doi, 查询并获取该文献资源。	<a href="http://localhost:8080/linkedcj/rest/journals/article?doi=">http://localhost:8080/linkedcj/rest/journals/article?doi=</a>
queryPersonByAuthorId()	根据某位作者的 ID, 查询并获取该作者的个人信息资源。	<a href="http://localhost:8080/linkedcj/rest/journals/person?authorId=">http://localhost:8080/linkedcj/rest/journals/person?authorId=</a>
queryPersonItem()	根据某个(些)人的姓名(同名不同人), 查询并获取这个(些)人的姓名、作者 ID、学历和工作单位信息资源。	<a href="http://localhost:8080/linkedcj/rest/journals/personItem?name=">http://localhost:8080/linkedcj/rest/journals/personItem?name=</a>
queryAuthor()	根据某篇文献的其中一个创作者的 ID, 查询并获取这个创作者的姓名、主页等信息资源。	<a href="http://localhost:8080/linkedcj/rest/journals/author?creatorId=">http://localhost:8080/linkedcj/rest/journals/author?creatorId=</a>
queryJournal()	根据某个期刊的 ID, 查询并获取这个期刊的刊名、主管单位、主办单位、主编等信息资源。	<a href="http://localhost:8080/linkedcj/rest/journals/journal?journalId=">http://localhost:8080/linkedcj/rest/journals/journal?journalId=</a>
queryJournalByname()	根据某个期刊的刊名, 查询并获取这个期刊的刊名、主管单位、主办单位、主编等信息资源。	<a href="http://localhost:8080/linkedcj/rest/journals/journal?name=">http://localhost:8080/linkedcj/rest/journals/journal?name=</a>
queryCitation()	根据某篇文献的某个引文 ID, 查询并获取这个引文的标题、作者、引文类型、引文来源等信息资源。	<a href="http://localhost:8080/linkedcj/rest/journals/citation?citeId=">http://localhost:8080/linkedcj/rest/journals/citation?citeId=</a>
queryResult()	根据用户输入的 SPARQL 语句, 查询并获取作者所需的信息资源。	<a href="http://localhost:8080/linkedcj/rest/journals/sparql?sparqlstr=">http://localhost:8080/linkedcj/rest/journals/sparql?sparqlstr=</a>
queryTitleDoiByKeyword()	根据用户输入字段, 查询并获取含有这些字段的文献标题和 Doi。	<a href="http://localhost:8080/linkedcj/rest/journals/title?keyword=">http://localhost:8080/linkedcj/rest/journals/title?keyword=</a>



## ▶ 中文学术期刊知识服务平台

□ Linked Data  
语义链接

dcterms:source	计算机学报
	田圭
	桂小林
	张学军



学术论文 期刊 学位 会议 外  
作者：“桂小林”

首页 > 检索结果

学科分类	标题	作者
▶ 工业技术	85篇	全部 仅全文
▶ 经济	5篇	
▶ 文化、科学…	4篇	1 大规模分布式环境下动
		期刊论文 《软件学报》 ISTIC

属性	值
10.3969/j.issn.1002-137X.2008.02.001	
LinkedCJ.articleID	jsjx200802001
fabio:hasDOI	<a href="http://dx.doi.org/10.3969/j.issn.1002-137X.2008.02.001">http://dx.doi.org/10.3969/j.issn.1002-137X.2008.02.001</a>
fabio:hasURL	<a href="http://id.wanfangdata.com.cn/Panodical_jsjx200802001.aspx">http://id.wanfangdata.com.cn/Panodical_jsjx200802001.aspx</a>
dcterms:title	数据质量研究综述
dcterms:abstract	数据质量管理是信息系统建设的首要问题.本文首先归纳了数据质量的定义和衡量指标的分类,然后对数据质量研究现状进行了两个主要方面的回顾:数据质量评估和数据质量提高.对技术的方法进行了比较和分析,并对比代表性的数据质量提高工具进行了介绍.最后提出了一一个评估数据质量提高的基础框架,并对数据质量研究方向进行了展望.
dcterms:source	<a href="http://purl.org/spar/fabio/Journal/10095">http://purl.org/spar/fabio/Journal/10095</a>
dcterms:creator	<a href="http://sw.tju.edu.cn/linkedCJ/AuthorName-%E4%9B%9C%E4%BC%A7%E7%94%A8%E6%8A%8D%200802001">http://sw.tju.edu.cn/linkedCJ/AuthorName-%E4%9B%9C%E4%BC%A7%E7%94%A8%E6%8A%8D%200802001</a> <a href="http://sw.tju.edu.cn/linkedCJ/AuthorName-%E4%9B%9C%E4%BC%A7%E7%94%A8%E6%8A%8D%200802001">http://sw.tju.edu.cn/linkedCJ/AuthorName-%E4%9B%9C%E4%BC%A7%E7%94%A8%E6%8A%8D%200802001</a>
LinkedCJ.creatorsWorkingUnit	东南大学计算机科学与工程系,南京,210095
LinkedCJ:hasSecondarySubjectTerm	TE6 TS7
fabio:hasPublicationDate	2008-05-14
LinkedCJ:hasProject	江苏省高技术研究发展计划项目
	<a href="http://sw.tju.edu.cn/linkedCJ/CitationMetadata/cites-1-jaix200802001">http://sw.tju.edu.cn/linkedCJ/CitationMetadata/cites-1-jaix200802001</a>
	<a href="http://sw.tju.edu.cn/linkedCJ/CitationMetadata/cites-2-jaix200802001">http://sw.tju.edu.cn/linkedCJ/CitationMetadata/cites-2-jaix200802001</a>

属性	值
jajx	
dcterms:title	计算机科学
dcterms:abstract	本节的读者对象是：大专院校师生，从事计算机科学与技术领域的科研、生产人员。本书宗旨是：坚持“双百方针”，活跃计算机科学与技术领域的学术气氛，重点报导国内外计算机科学与技术的发展动态，为我国的计算机科学与技术立于世界之林、达到国际先进水平奋斗而矢志不渝。
LinkedCJ:hasColumn	计算机网络与信息技术
LinkedCJ:competentOrganization	重庆西南信息有限公司（原科技部西南信息中心）
LinkedCJ:organizer	重庆西南信息有限公司（原科技部西南信息中心）
LinkedCJ:editorInChief	朱元元
prism:issn	1002-137X
LinkedCJ:cn	60-1075/TP

属性	值
name=%E6%96%87%E7%94%A8%E8%87%8A&articleId=jajx200802001	
foaf:name	徐立强
fb:creatorOf	<a href="http://purl.org/spar/fabio/JournalArticle/10.3969/j.issn.1002-137X.2008.02.001">http://purl.org/spar/fabio/JournalArticle/10.3969/j.issn.1002-137X.2008.02.001</a>
foaf:homepage	<a href="http://social.wanfangdata.com.cn/Auto/Achievement.aspx?name=%E5%BE%90%EF%AB%80%EF%AB%80%07%EF%AB%80&amp;articleId=zgdxj200604022">http://social.wanfangdata.com.cn/Auto/Achievement.aspx?name=%E5%BE%90%EF%AB%80%EF%AB%80%07%EF%AB%80&amp;articleId=zgdxj200604022</a>
LinkedCJ:isPerson	<a href="http://vmls.com/fcaff1/Person/name-%E6%96%87%E7%94%A8%87%8A&amp;articleId=zgdxj200604022">http://vmls.com/fcaff1/Person/name-%E6%96%87%E7%94%A8%87%8A&amp;articleId=zgdxj200604022</a>
co-index	2

属性	值
cites-1-jajx200802001	
co-index	1
dcterms:creator	Monge A Elkan C
dcterms:title	An efficient domain-independent algorithm for detecting approximately duplicate database records
fabio:hasURL	<a href="http://id.g.wanfangdata.com.cn/ExternalResource-jajx2008020011.aspx">http://id.g.wanfangdata.com.cn/ExternalResource-jajx2008020011.aspx</a>
LinkedCJ:type	A
LinkedCJ:source	Tucson



## ➤ 中文学术期刊知识服务平台 □ SPARQL查询服务

### SPARQL查询

```
select ?title?source where { ?article
<http://prismstandard.org/namespaces/basic/
2.0/doi>
"http://dx.doi.org/10.3724/SP.J.1016.2014.00
621"; <http://purl.org/dc/terms/title> ?title;
<http://purl.org/dc/terms/source> ?source}
```

### 查询示例

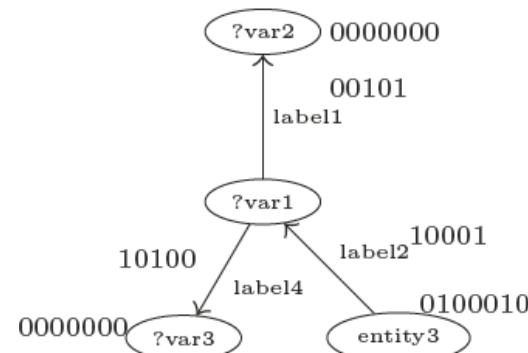
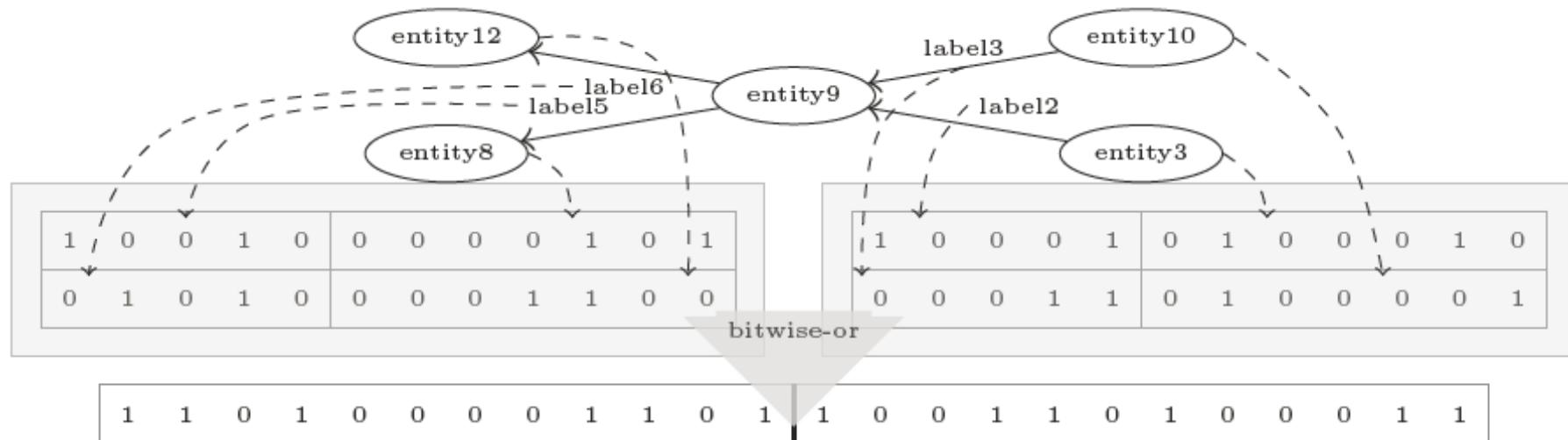
1. 如何查询DOI为"<http://dx.doi.org/10.3724/SP.J.1016.2014.00621>"的文章的题目和来源？
2. 如何查询文献《一种多变量决策树方法研究》的作者和他的工作单位？
3. 如何查询期刊《计算机科学》的主编和主办单位？

### 查询结果

title	source
基于样本的大规模人群快速创作	<a href="http://purl.org/spar/fabio/Journal/l-sjxb">http://purl.org/spar/fabio/Journal/l-sjxb</a>



➤ 大规模RDF图上的子图匹配查询  
□ 星形查询的高效执行：SPARQL关键操作



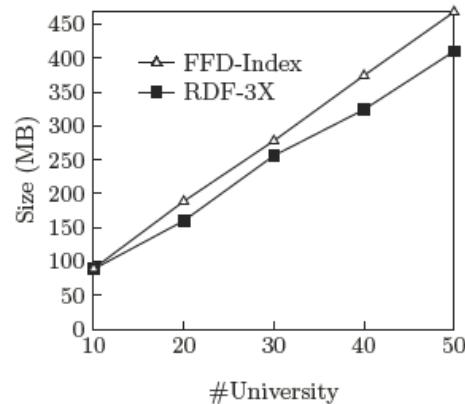
(a)  $Q_1^*$  and Bitstrings of Labels

$$\frac{10101 \ 0000000 \quad 10001 \ 0100010}{fp_{out}(Q_1^*) \quad fp_{in}(Q_1^*)}$$

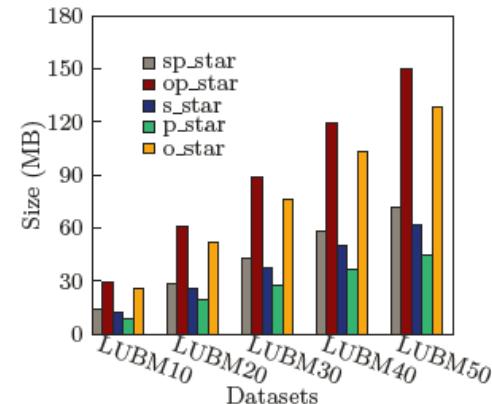
(b)  $fp(Q_1^*)$



➤ 大规模RDF图上的子图匹配查询  
□ 星形查询的高效执行：SPARQL关键操作



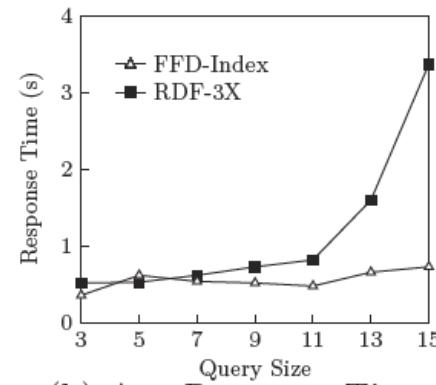
(a) Index Size



(b) FFD-Index Size

	# BEF. VLD.	# AFT. VLD.
$Q_3$	542.6	67.4
$Q_5$	477.3	45.8
$Q_7$	237.0	21.1
$Q_9$	97.3	12.7
$Q_{11}$	104.4	4.2
$Q_{13}$	85.5	1.6
$Q_{15}$	57.2	1.8

(a) Filtering Ability



(b) Avg Response Time

Fig. 8: Performance on YAGO



天津大学



## 一、课题背景

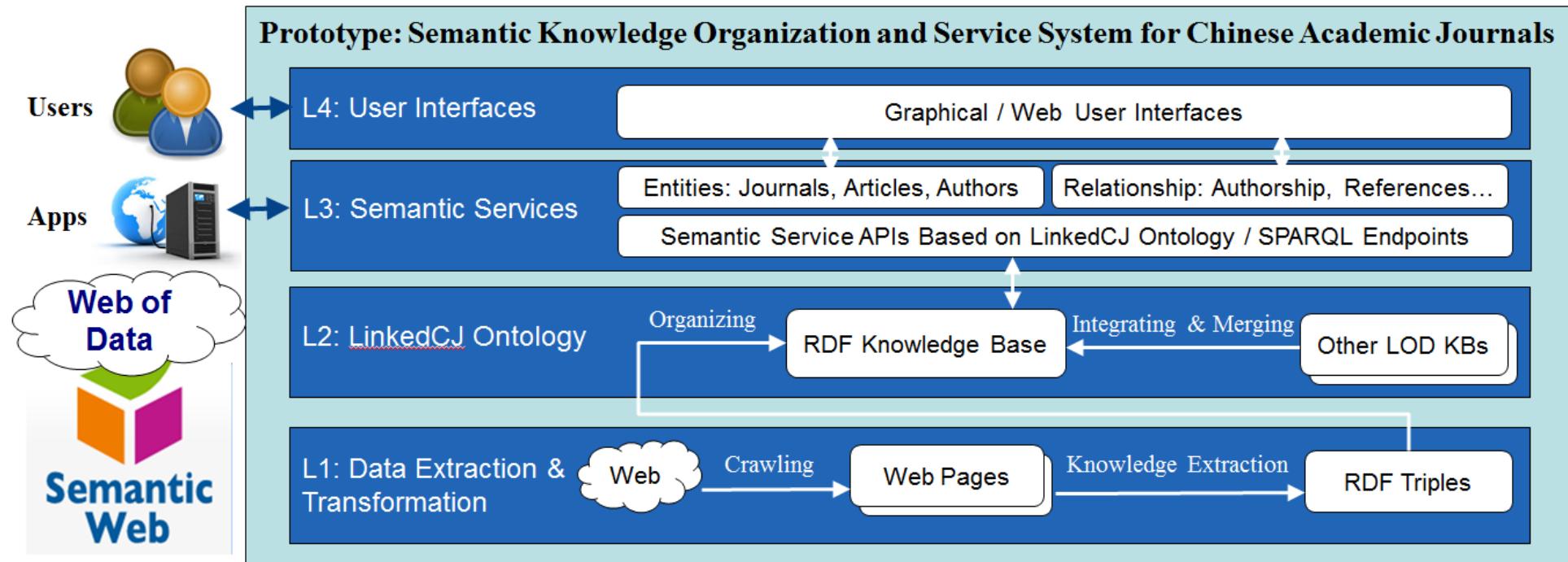
## 二、中文学术期刊本体与知识库

## 三、中文学术期刊语义知识服务

## 四、原型系统与功能演示 ←



## ► 中文学术期刊语义知识组织与服务系统 □ 系统架构





## ➤ 原型系统演示

### □ 演示1：Web页面爬取

The screenshot shows the Eclipse IDE interface with the following details:

- Project Explorer:** Shows the project structure with packages `crawl`, `LinkedCJ`, and `LinkedJournal`. The `JsoupCrawler.java` file is open.
- Code Editor:** Displays Java code for a `JsoupCrawler` class. The code uses Jsoup to connect to URLs, parse HTML, and write content to files. It handles exceptions and logs information to the console.
- Outline View:** Shows the class hierarchy and methods defined in `JsoupCrawler`.
- Problems View:** Lists terminated tasks related to the crawler.
- Console View:** Displays the output of the crawler's execution, showing a list of URLs it has processed.

```
String cguri = ele.attr("href");
doc = Jsoup.connect(cguri).timeout(100000).get();
System.out.println(cguri);

String dir_name = "F:/LinkedCJ_Project/Html/" + String.valueOf(count); // F:\LinkedCJ_Project\Html
File dir = new File(dir_name);
dir.mkdirs();
File filel = new File(dir_name + "/0.txt");
filel.createNewFile(); // 避免报错找不到文件
BufferedWriter outl = new BufferedWriter(
    new FileWriter(filel));
outl.write(doc + "\r\n");
outl.flush(); // 将写入的数据刷新到文件
outl.close(); // 关闭流

try{
    reference = "http://d.wanfangdata.com.cn/" + doc.getElementById("goref").attr("href");
    if (reference != null) {
        File file2 = new File(dir_name + "/1.txt");
        file2.createNewFile(); // 避免报错找不到文件
        doc = Jsoup.connect(reference).timeout(100000).get();
        BufferedWriter out2 = new BufferedWriter(
            new FileWriter(file2));
        out2.write(doc + "\r\n"); // \r\n转为换行符
        out2.flush(); // 将写入的数据刷新到文件
        out2.close(); // 关闭流
    }
}
```

Console Output (部分输出):

```
<terminated> JsoupCrawler [Java Application] C:\Program Files\Java\jre7\bin\javaw.exe (2014年12月6日下午4:24:10)
http://d.g.wanfangdata.com.cn/Periodical_jsxb201401001.aspx
http://d.g.wanfangdata.com.cn/Periodical_jsxb201401002.aspx
http://d.g.wanfangdata.com.cn/Periodical_jsxb201401003.aspx
http://d.g.wanfangdata.com.cn/Periodical_jsxb201401004.aspx
http://d.g.wanfangdata.com.cn/Periodical_jsxb201401005.aspx
http://d.g.wanfangdata.com.cn/Periodical_jsxb201401006.aspx
http://d.g.wanfangdata.com.cn/Periodical_jsxb201401007.aspx
http://d.g.wanfangdata.com.cn/Periodical_jsxb201401008.aspx
http://d.g.wanfangdata.com.cn/Periodical_jsxb201401009.aspx
http://d.g.wanfangdata.com.cn/Periodical_jsxb201401010.aspx
http://d.g.wanfangdata.com.cn/Periodical_jsxb201401011.aspx
http://d.g.wanfangdata.com.cn/Periodical_jsxb201401012.aspx
http://d.g.wanfangdata.com.cn/Periodical_jsxb201401013.aspx
http://d.g.wanfangdata.com.cn/Periodical_jsxb201401014.aspx
http://d.g.wanfangdata.com.cn/Periodical_jsxb201401015.aspx
http://d.g.wanfangdata.com.cn/Periodical_jsxb201401016.aspx
```



## ➤ 原型系统演示 □ 演示2：知识抽取

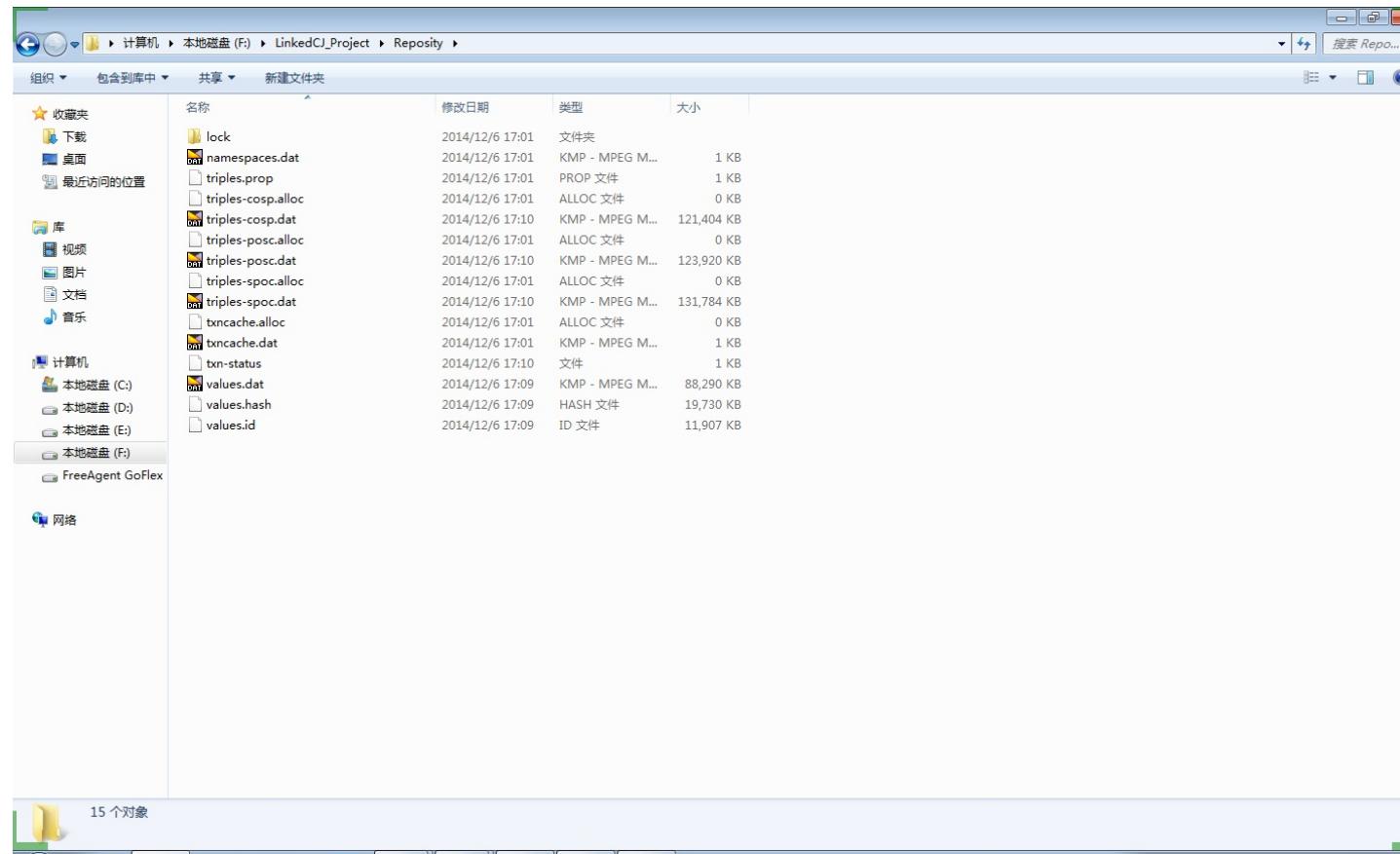
The screenshot shows the Eclipse IDE interface with the following details:

- Project Explorer:** Shows the project structure for "Java - Linked-CJ/src/main/java/com/linkedcj/LinkedCJ.java".
- Editor:** Displays the Java code for `LinkedCJ.java`. The code initializes a knowledge base, creates sessions, and writes data to files. A portion of the code is highlighted in blue.
- Outline View:** Shows the class structure with fields like `articleID`, `authorName`, `authorSubject`, `personSubject`, `journalID`, and methods like `main`.
- Task List:** A tooltip for "Connect Mylyn" is visible, prompting to connect to tasks and ALM tools.
- Console:** Shows command-line output indicating progress through 100, 200, 300, 400, 500, and 600 pages.



## ➤ 原型系统演示

### □ 演示3：知识库装载





## ➤ 原型系统演示 □ 演示4：服务平台

中文学术期刊与文献语义服务平台

欢迎使用  
中文学术期刊与文献语义服务平台

名称空间列表

- fabio:** <http://purl.org/spar/fabio/>
- Linked-CJ:** <http://sw.tju.edu.cn/Linked-CJ/>
- dcterms:** <http://purl.org/dc/terms/>
- prism:** <http://prismstandard.org/namespaces/basic/2.0/>
- cito:** <http://purl.org/spar/cito/>
- foaf:** <http://xmlns.com/foaf/0.1/>
- co:** <http://purl.org/co/>

相关链接

- 万方数据知识服务平台
- 国内一流信息资源出版、增值服务平台
- SPARQL 1.1 Query | language



天津大学



## ➤ 课题资助发表的论文

	名称	期刊/会议
1	Xu Peng, Xin Wang, and Haofen Wang. LinkedCJ: A Knowledge Base of Chinese Academic Journals Based on Linked Data. In Proceedings of the 4th Joint International Semantic Technology Conference (JIST 2014), poster, 2014.	<i>JIST 2014</i>
2	Xuedong Lyu, Xin Wang, Yuan-Fang Li, and Zhiyong Feng. FFD-Index: An Efficient Indexing Scheme for Star Subgraph Matching on Large RDF Graphs. Submitted to DASFAA 2015.	<i>DASFAA 2015 (CCF B)</i>
3	张啸野, 王鑫, 徐鹏, 冯志勇. 基于Linked Data的中文学术期刊语义知识组织与服务. 计算机工程. (待发表)	核心期刊
4	魏亚洲, 王鑫, 冯志勇, 饶国政. S-Index: 一种面向大规模 RDF 数据的高效率语义索引方案. 武汉大学学报 (理学版) (待发表) (第8届中国语义 Web和Web科学会议推荐)	核心期刊



天津大学



## ➤ 总结

□ 基于Linked Data的中文学术期刊语义知识组织与  
服务关键技术研究

知识  
抽取

本体  
知识库

语义  
服务

服务  
平台

RDF

Linked  
Data

SPARQL

GUI



天津大学



谢谢大家  
请提问题

2014.12.09@NLPCC2014