关联数据的消费与应用构建:

以上海图书馆家谱开放数据为例

第十三届数字图书馆前沿问题高级研讨班 ADLS 2016, 上海图书馆, 12月5日

董行

UNIVERSITY OF LIVERPOOL

关联数据与图书馆

- OCLC报告数据显示，2014-2015年，相比发布关联数据，图书馆更多地消费关联数据 (Smith-Yoshimura, 2016)。

发布还是消费？	2014年	2015年
只发布关联数据	4	10
只消费关联数据	25	38
既发布又消费	47	64

- “小型关联数据项目在增多，为未来的应用提供机遇。” (Smith-Yoshimura, 2016)

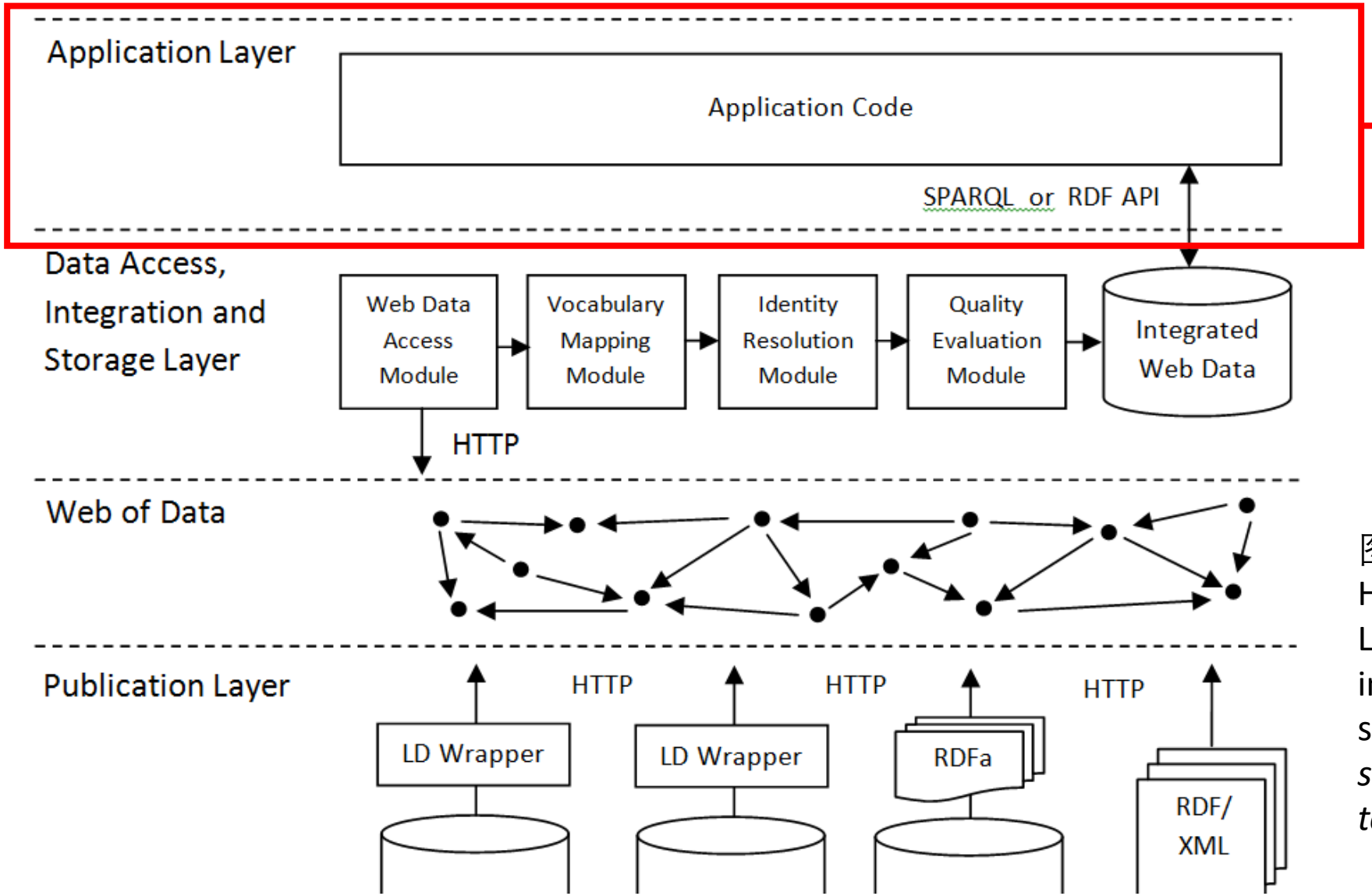
关联数据的消费和应用

- 关联数据的消费主要涉及关联数据的访问、获取、发现、查询、交换、传输、处理和利用等消费过程中所相关的各类实现方式、技术标准及工具平台（夏翠娟，刘炜，2013）。

- 关联数据的应用 (Heath & Bizer, 2011)：
  - 大型的、通用的应用，跨领域的关联数据览器和搜索引擎
  - 单种领域的应用，利用不同的消费方式来浏融合领域内的数据

- 消费关联数据是创建优秀关联数据应用的关键。

关联数据应用的宏观构架



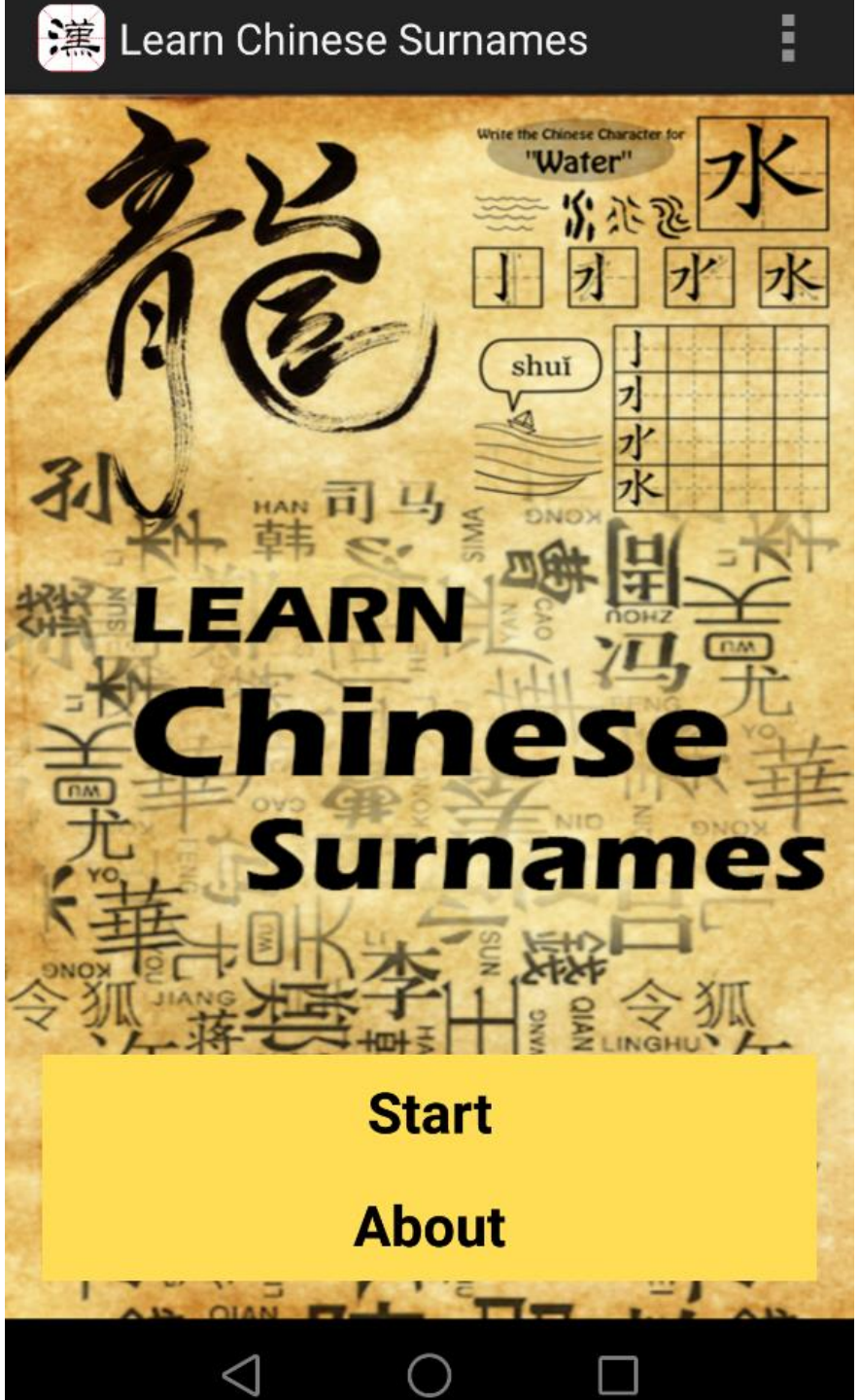
关联数据应用的微观构架

图片来源：
Heath, T. and Bizer, C., 2011. Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*, *1*(1), pp.1-136.

案例介绍

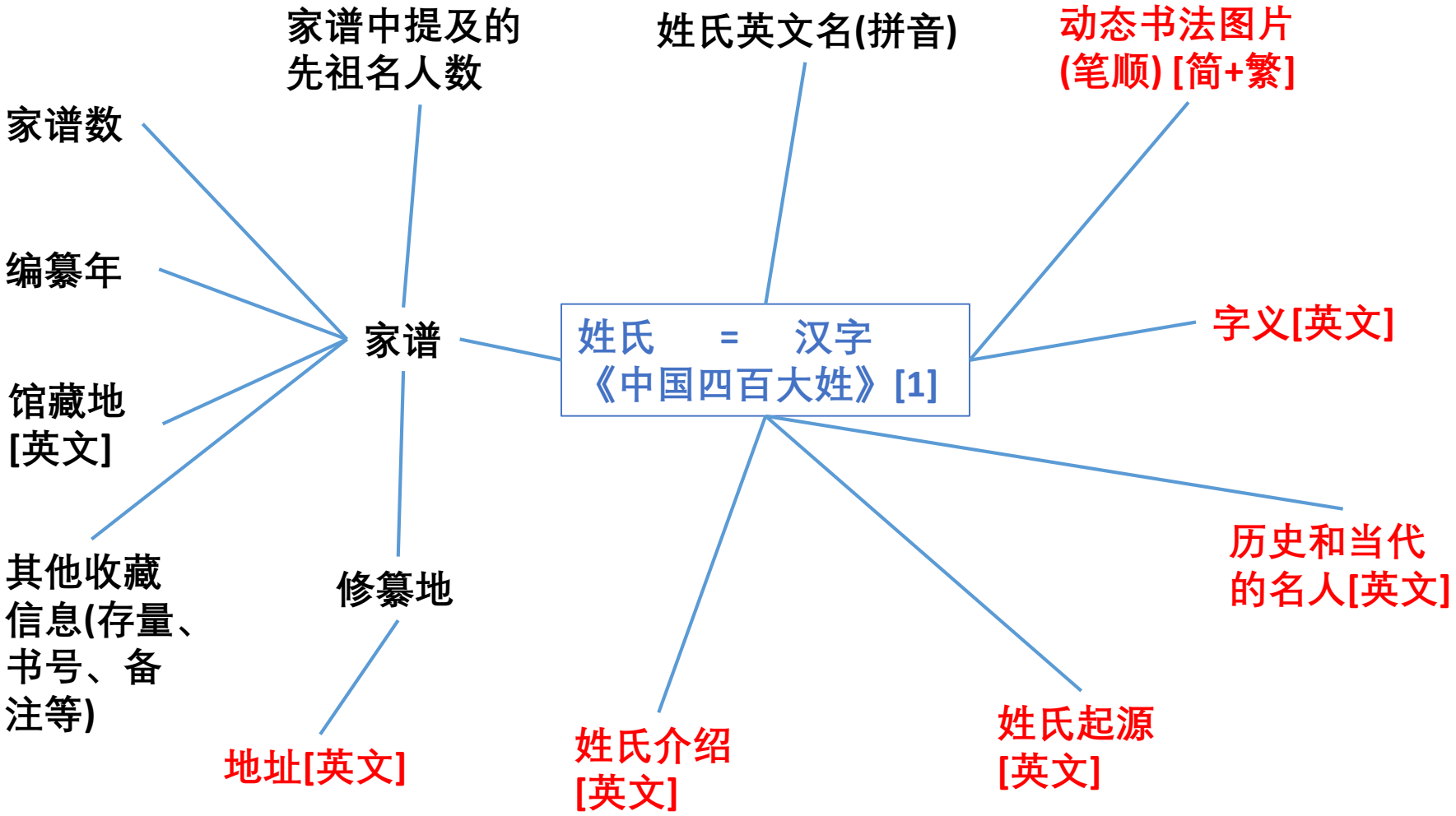
- Learn Chinese Surnames
- 一款方便外国人学习
  中国姓氏文化和汉字的
  安卓手机应用

方便外国人在生活中随时查阅和学习中国姓氏，
了解汉字的写法与含义、姓氏的来源、
姓氏的人口排名、早期的家谱等等。



用户角度: 信息

红色表示上图家谱数据中不包括的信息。

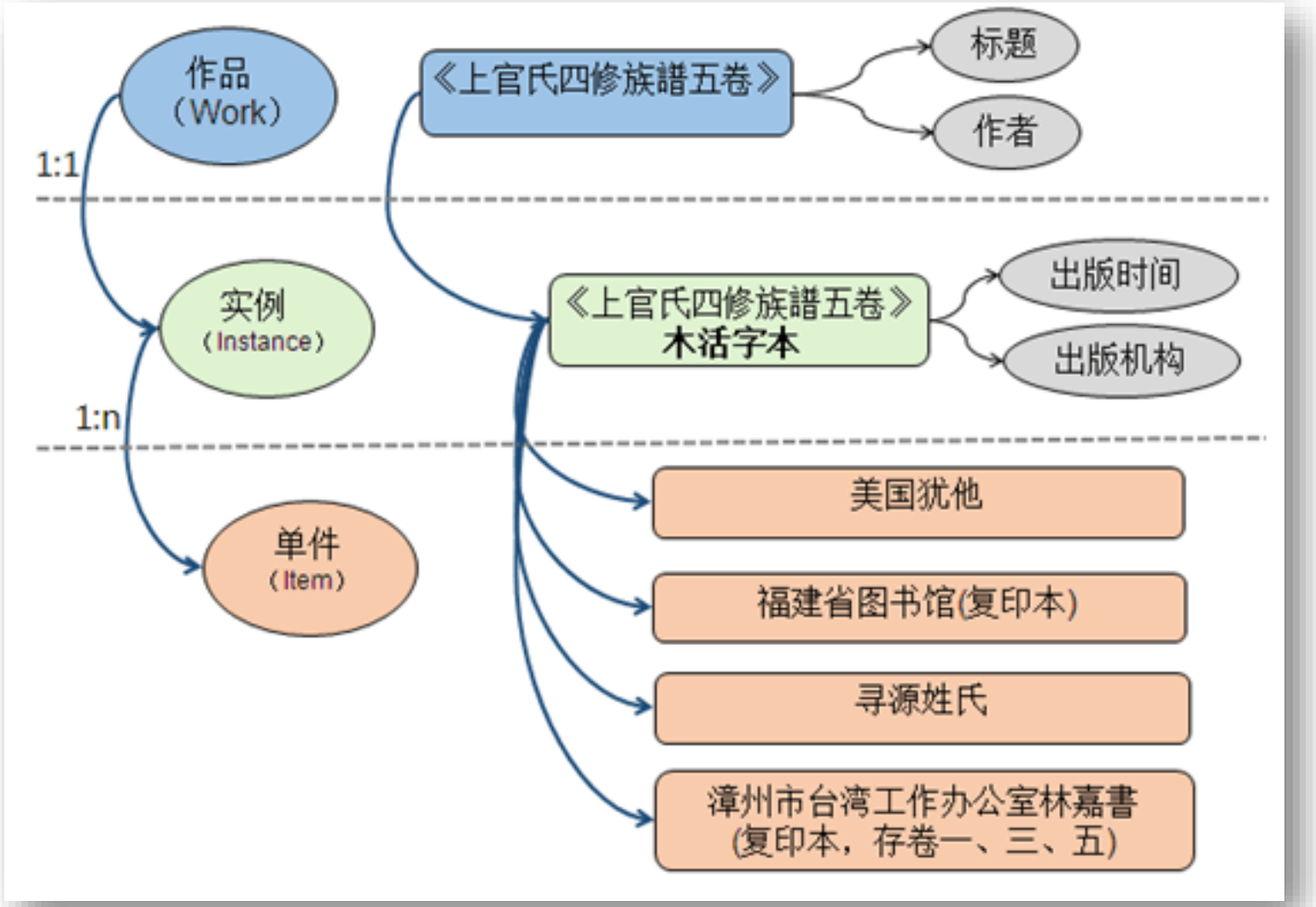


[1] 袁义达, 邱家儒. 中国四百大姓（上中下册）. 南昌: 江西人民出版社, 2013-01-01

数据源一览

数据源	信息	获取方式	使用方式	版权声明
上海图书馆家谱数据	家谱题名、纂修时间、地理位置、馆藏地英文名，姓氏英文名、姓氏人数	SPARQL在线实时调用Restful 服务获取	文字展示。	(1) CC2.0协议（署名-非商业性使用-相同方式共享） (2) 比赛授权使用
DBpedia和维基百科	姓氏的英文词条	SPARQL语句离线获取	嵌入网页。	CC BY-SA 1.0-4.0
Wiktionary	汉字的英文词条	直接从URL获取	嵌入网页。	CC BY-SA 3.0
GeoNames	中国地理位置的英文元数据	通过官方API获取	文字展示。	CC BY-SA 3.0
WrittenChinese.Com	简体和繁体汉字的动态笔顺书法图片	直接从URL获取	嵌入在线图片。	(1) Copyright (C) 2009 - 2016 v1.9.0 WrittenChinese.Com All Rights Reserved. (2) CC BY-SA 3.0
ChineseTools.eu	简体汉字的静态书法图片	直接从URL获取	嵌入在线图片。	(1) Copyright (C) 2016 - ChineseTools.eu
2d-code	繁体汉字的静态书法图片	通过官方API获取	嵌入在线图片。	版权:二维码生成，闽ICP备15012419号
(1)《中国四百大姓》 (2)中文维基百科:中国姓氏排名词条	2013年由中国伏羲文化研究会发布的中国四百大姓	引用和离线获取。	列表展示。	(1) 袁义达，邱家儒，江西人民出版社 (2) CC BY-SA 3.0 (中文维基百科)

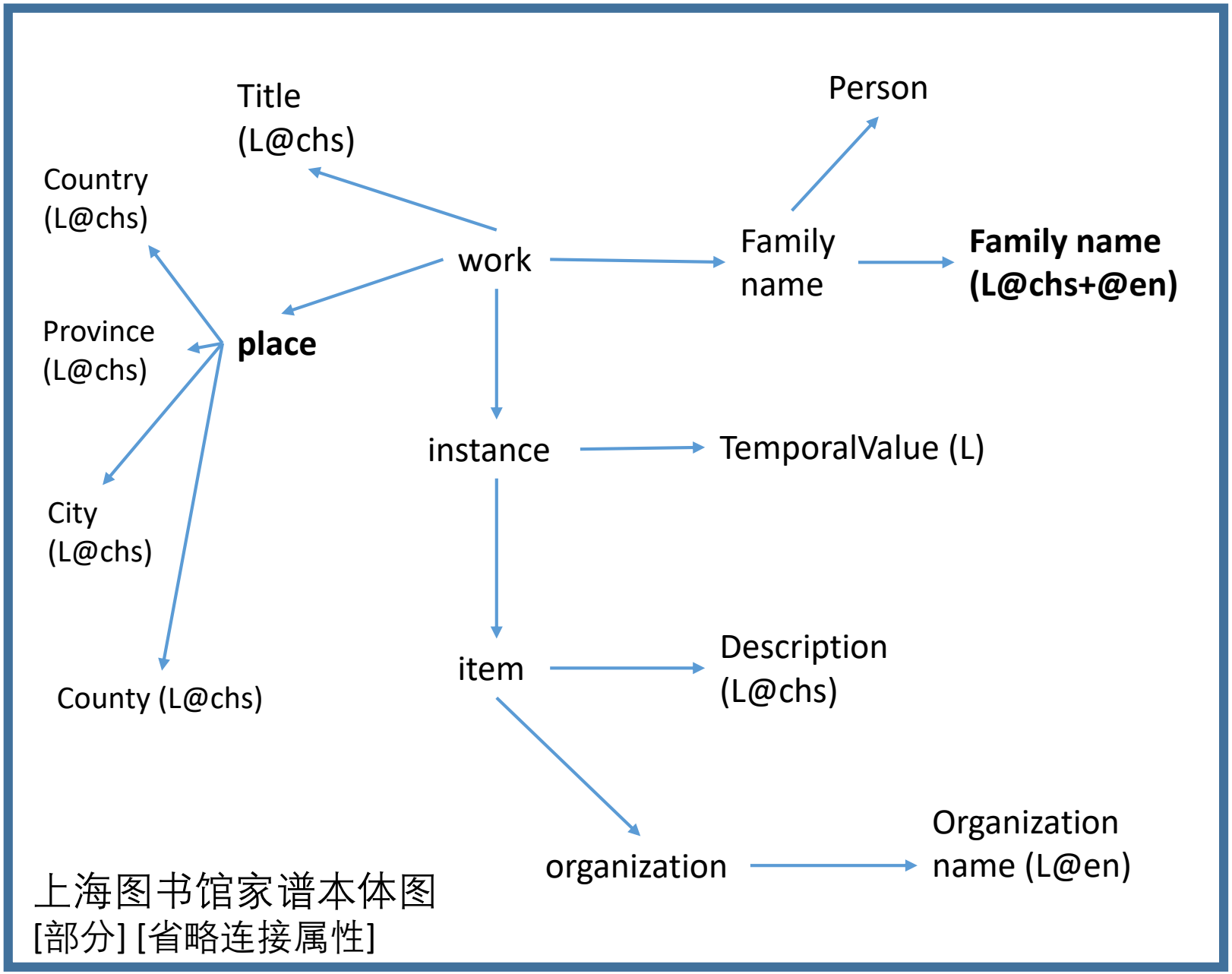
上海图书馆家谱开放数据



说明:

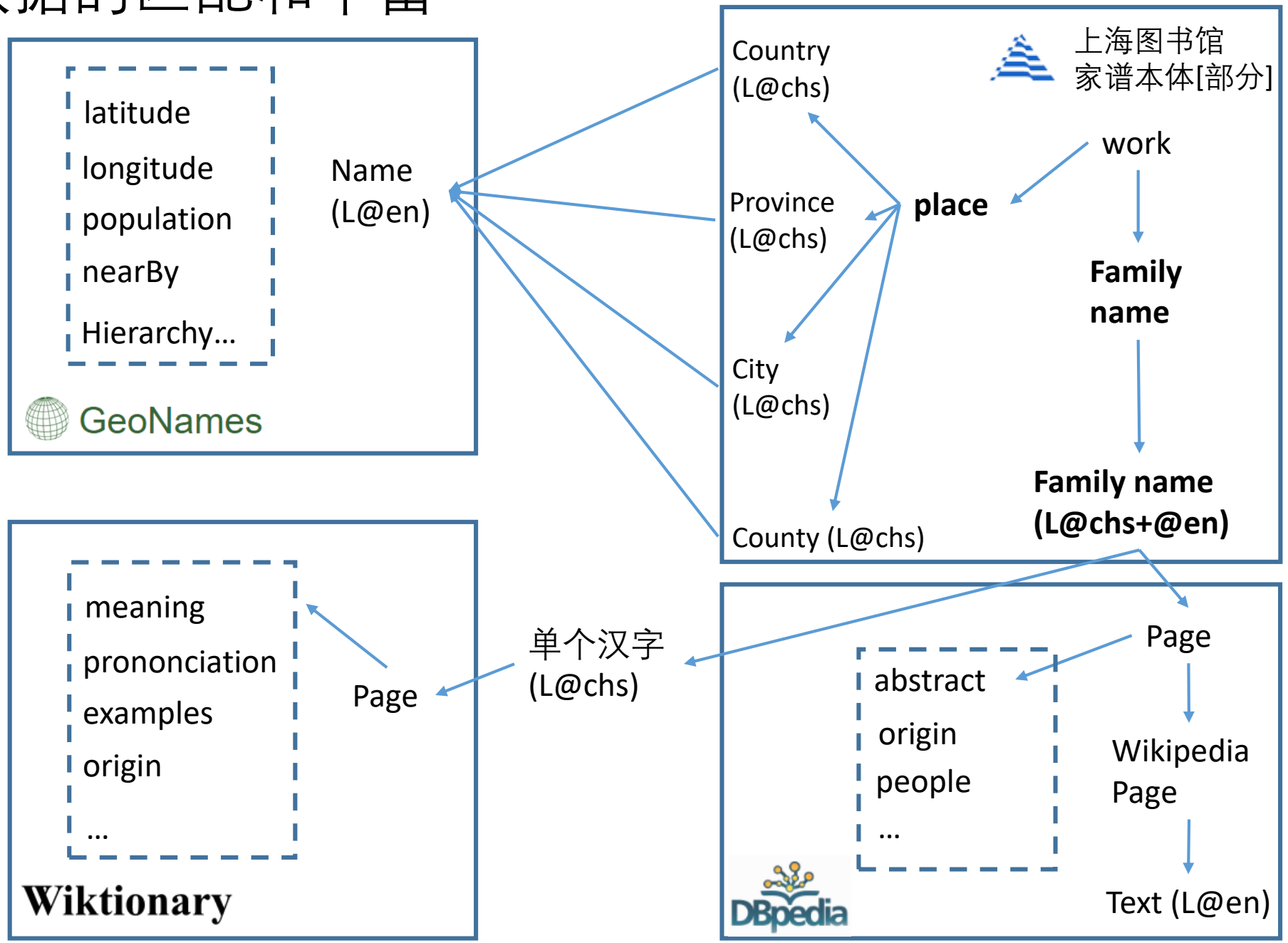
L = Literal 字符型
@chs 简体
@en 英文

未标注 L 的字段
即为URI类型



上海图书馆家谱本体图
[部分] [省略连接属性]

家谱数据的匹配和丰富


latitude
longitude
population
nearBy
Hierarchy…
Name
(L@en)
GeoNames
Country
(L@chs)
Province
(L@chs)
City
(L@chs)
County (L@chs)
上海图书馆
家谱本体[部分]
work
place
Family
name
Family name
(L@chs+@en)
meaning
prononciation
examples
origin
…
Page
Wiktionary
单个汉字
(L@chs)
Page
abstract
origin
people
…
Wikipedia
Page
Text (L@en)
DBpedia

通过消费关联数据实现跨语言

- 1 匹配姓氏汉字到维基百科词条 – 利用SPARQL端点调用
- 利用DBPedia的SPARQL Endpoint (Restful Service 访问限制，可离线使用) **(采用)**
- 人工匹配 **(采用)**
- 机器学习

- 2 匹配姓氏汉字到维基词典词条 – 构造URL直接调用
- Wiktionary API (访问限制)
- Wiktionary SPARQL Endpoint (访问限制)
- 构造URL **(采用)**

- 3 匹配中文的层级地址到GeoNames英文地点 - 利用官方API调用
- GeoNames API **(采用)**

匹配姓氏汉字到维基百科词条 – 利用SPARQL端点调用

- Dbpedia的SPARQL Endpoint:

https://dbpedia.org/sparql

- 通过SPARQL端点批量获取姓氏对应的维基百科英文词条

- 对匹配结果进行人工筛选

Query Text

```
# 通过简繁汉字匹配DBpedia资源摘要
select distinct ?url ?ex count(?url) as ?count
where{
?res dct:subject <http://dbpedia.org/resource/Category:Chinese-language_surnames>.
?res dbo:abstract ?a.
filter(contains(str(?a), "赵") || contains(str(?a), "趙")).
?res foaf:isPrimaryTopicOf ?url.
optional {?res dbo:wikiPageExternalLink ?ex}
}
order by desc(?count)
limit 1
```

(Security restrictions of this server do not allow you to retrieve remote RDF data, see details.)

Results Format: HTML

Execution timeout: 30000 milliseconds *(values less than 1000 are ignored)*

Options: Strict checking of void variables Log debug info at the end of outp

(The result can only be sent back to browser, not saved on the server, see details)

Run Query Reset

匹配中文的层级地址到GeoNames英文地点 - 利用官方API调用

- 参考： http://www.geonames.org/export/geonames-search.html

- 获取陕西省西安市户县（1731年家谱《段氏世系》的修纂地）对应的GeoNames的JSON条目
- http://api.geonames.org/searchJSON?name_equals=%E6%88%B7%E5%8E%BF&featureCode=ADM3&country=CN&maxRows=10&username=XXX

（将XXX替换为GeoNames用户名）

返回结果：

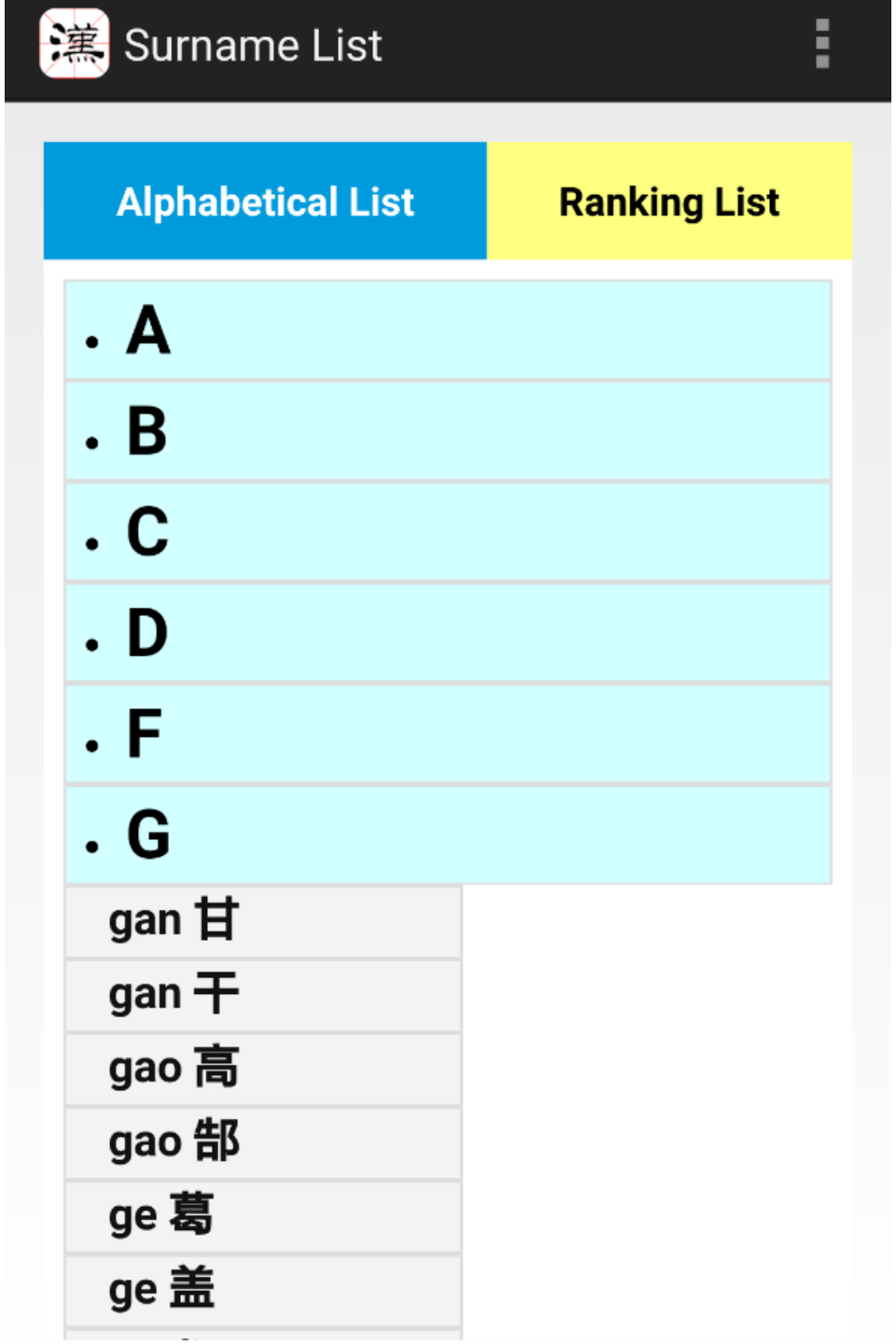

```
{
  "totalResultsCount": 1,
  "geonames": [
    {
      "adminCode1": "26",
      "lng": "108.58764",
      "geonameId": 1806562,
      "toponymName": "Hu Xian",
      "countryId": "1814991",
      "fcl": "A",
      "population": 556377,
      "countryCode": "CN",
      "name": "Hu Xian",
      "fclName": "country, state, region,...",
      "countryName": "China",
      "fcodeName": "third-order administrative division",
      "adminName1": "Shaanxi",
      "lat": "33.99969",
      "fcode": "ADM3"
    }
  ]
}
```

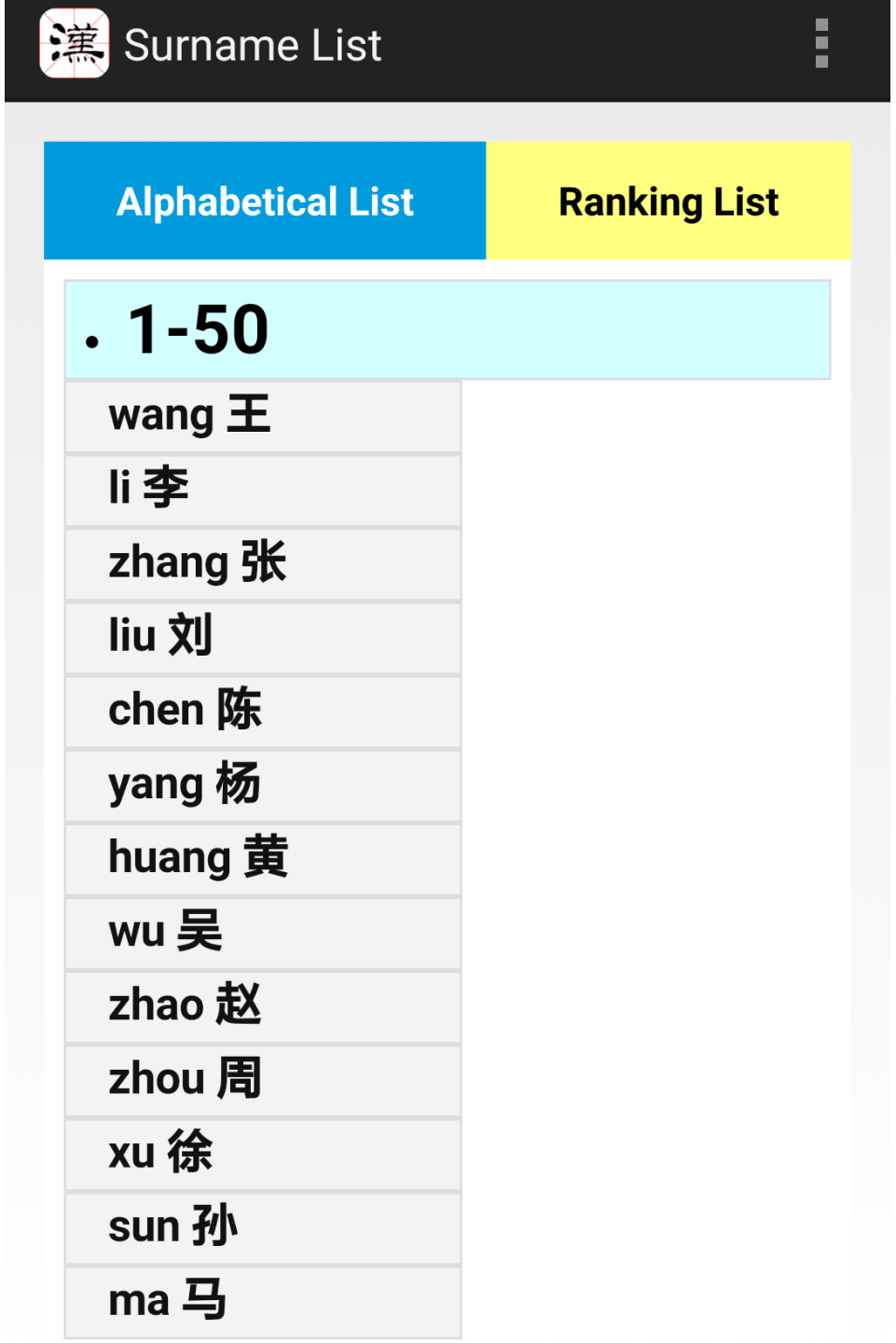
匹配姓氏汉字到维基词典词条 – 构造URL直接调用

- 构造刘姓对应的维基词典链接

https://en.wiktionary.org/wiki/%E5%88%98

在浏览器中显示为“刘”


Surname List
Alphabetical List
Ranking List
. A
. B
. C
. D
. F
. G
gan 甘
gan 干
gao 高
gao 郜
ge 葛
ge 盖


Surname List
Alphabetical List
Ranking List
. 1-50
wang 王
li 李
zhang 张
liu 刘
chen 陈
yang 杨
huang 黄
wu 吴
zhao 赵
zhou 周
xu 徐
sun 孙
ma 马

苏 (su)

Wiktionary of word **苏 (su)**
Wikipedia of surname **苏 (su)**

There are 195 family books for 苏 (su), where 493 names are recorded.

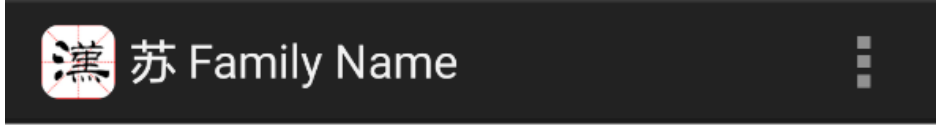
See traditional 苏 (su) and early family books in the next page.

next page

漢 苏 Family Name

The earlist 3 family books for 苏 are:

新安苏氏族谱十五卷（安徽省黄山市休宁县）| Xiuning Xian, Anhui, China
1467 Anhui province library
1467 The national library
1467 The library of liaoning province
1467 The east Asian library of Columbia University
1467 Shanghai library
1467 Fudan university library
1467 The Genealogical Society of Utah
1467 Zhejiang library
1467 2002年綫裝書局影印《中國國家圖書館藏早期



The earlist 3 family books for 苏 are:

新安苏氏族谱十五卷（安徽省黄山市休宁县）| Xiuning Xian, Anhui, China
1467 Anhui province library
1467 The national library
1467 The library of liaoning province
1467 The east Asian library of Columbia University
1467 Shanghai library
1467 Fudan university library
1467 The Genealogical Society of Utah
1467 Zhejiang library
1467 2002年綫裝書局影印《中國國家圖書館藏早期稀見家譜叢刊本》，一册

新安苏氏重修族谱五卷补遗一卷（安徽省黄山市休宁县）| Xiuning Xian, Anhui, China
1736 The national library
1736 Nanjing library
1736 The east Asian library of Columbia University
1736 The Genealogical Society of Utah

苏氏族谱六卷（安徽省）| Anhui, China
1763 Shanghai library
1763 The Genealogical Society of Utah

展现上海图书馆家谱数据的层级结构。

通过匹配到GeoNames显示一个work的英文地点名。

比拼音转换的方式更加精准。

苏

See also: **蘇** *and* **甦**

Translingual

Etymology

Simplified from 蘇 (穌 → 办)

Han character

苏 (*radical 140* 艸+4, *7 strokes*, *cangjie input* 廿大尸金 (TKSC), *composition* ⿱艹…

next page

Su (surname)

Su is the pinyin romanization of the common Chinese surname written 苏 in simplified characters and 蘇 traditionally.

It was listed 42nd among the Song-era list of the *Hundred Family Surnames*.

It is also the pinyin romanization of the very rare surname 粟.

Romanizations

next page

List of persons with the surname

Su

- Alec Su, Taiwanese singer and actor
- Su Buqing, mathematician
- Su Chin-shou, Hui chief of staff to General Ma Zhancang
- Su Daji, the beautiful concubine

So

- John So, former Lord Mayor of Melbourne
- Louisa So, Hong Kong actress
- Wesley So, Filipino chess prodigy
- William So,

next page

苏 Family Name

Su Buqing

This is a Chinese name; the family name is Su.

Su Buqing, also spelled **Su Buchin** (Chinese: 蘇步青; September 23, 1902 – March 17, 2003),[1] was a Chinese mathematician, educator, and President of Fudan University.

Biography

Notes

next page

数据的核对和补充

姓氏	姓氏英文名(上海图书馆家谱数据)	英文词条名（英文维基百科数据）	维基百科连接	注释
房	fang	pang	http://en.wikipedia.org/wiki/Pang_(surname)	fang\pang均有
柏	bai	bo	http://en.wikipedia.org/wiki/Bo_(Chinese_surname)	应为bo
区	qu	ou	http://en.wikipedia.org/wiki/Ou_(surname)	应为ou
强	qiang	jiang	http://en.wikipedia.org/wiki/Jiang_(surname)	应为jiang
危	wei	ngai	http://en.wikipedia.org/wiki/Ngai_(surname)	Ngai为广东话读音

总共400个姓氏，其中上海图书馆中包括377/400个，英文维基百科中295/400个有对应词条。

姓氏	英文名	排名	姓氏	英文名	排名
龙	long	84	岳	yue	110
康	kang	105	葛	ge	124
牛	niu	108	甘	gan	137

维基百科英文词条中未收录之上海图书馆家谱数据中的姓氏
(仅列出前6/共93个)

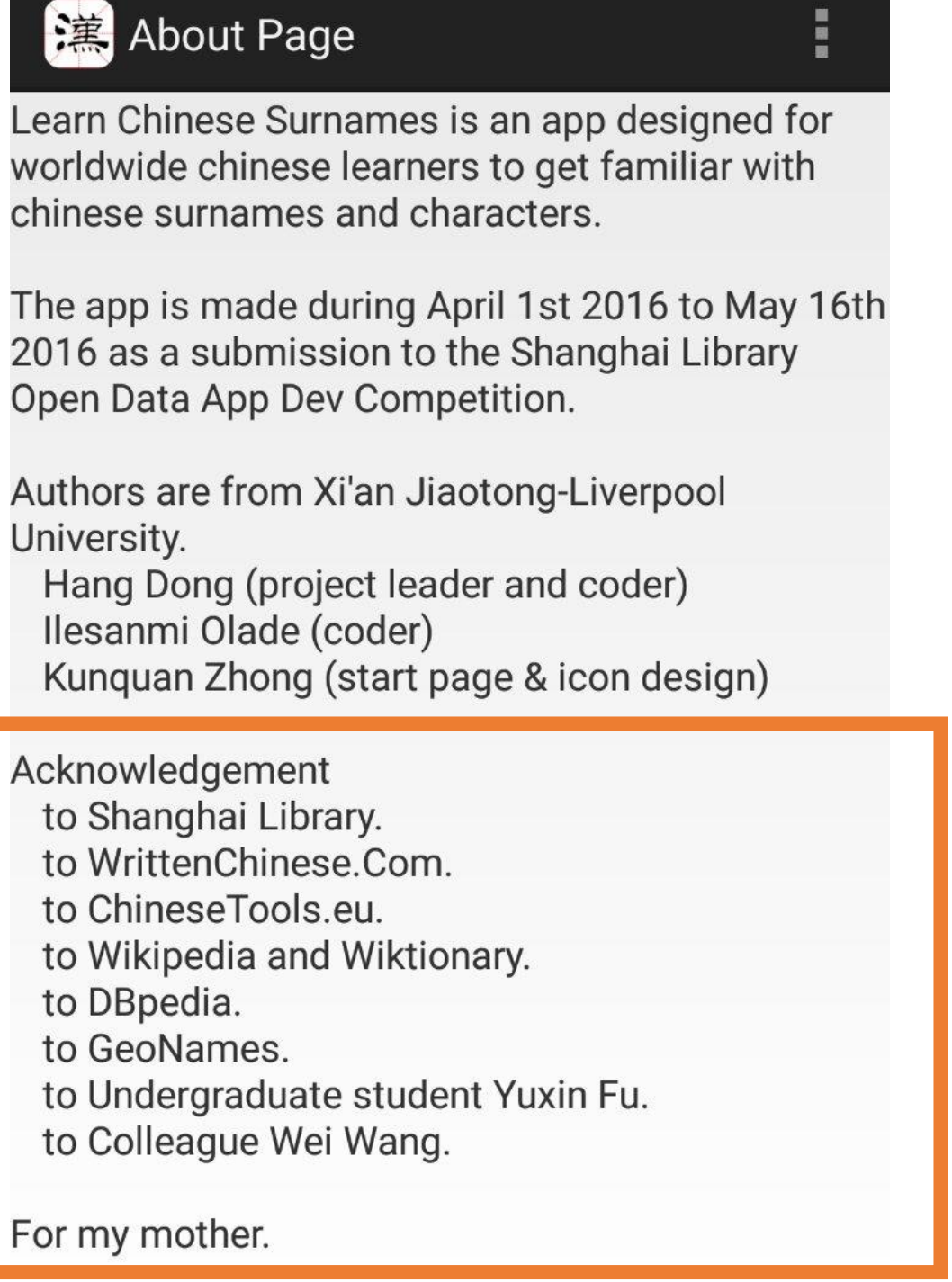
总结：对图书馆的意义

本项目探索了如何将中文的家谱本体与三种多语言关联数据集进行融合，完成跨语言的移动应用。

- **馆藏资源与关联数据的整合**
  - 关联数据使得知识得以更加细粒度化的整合
- **跨语言应用保障获取**
  - 促进馆藏资源在世界范围内的获取，有利于扩展图书馆的服务范围，促进国际交流

重要参考文献和延伸阅读

- Heath, T. and Bizer, C., 2011. Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*, *1*(1), pp.1-136.
- 编目精灵III. 关联数据应用现状：2015国际关联数据实施者调查的分析 http://catwizard.net/posts/20160904151045.html
- Smith-Yoshimura, Karen. Analysis of International Linked Data Survey for Implementers. D-Lib Magazine, 22(7/8) doi:10.1045/july2016-smith-yoshimura
- 夏翠娟, 刘炜. 关联数据的消费技术及实现. 大学图书馆学报. 2013(3):29-37.
- 夏翠娟, 刘炜, 张磊, 朱雯晶. 基于书目框架 (BIBFRAME) 的家谱本体设计. 图书馆论坛. 2014,34(11):5-19.
- 上海图书馆基于BibFrame的家谱本体 http://gen.library.sh.cn:8080/ontology/view



谢谢聆听

感谢参赛小组成员Ilesanmi Olade和钟坤权

董行
hangdong@liverpool.ac.uk