

大数据时代的图书馆服务创新

上海图书馆开放数据服务

张奇 上海图书馆



四个部分




- 大数据
- 开放数据
- 上图案案例
- 收获与启示




大数据

快速扫描



Big Data, it's more than
what you think

- **基础性资源**
- **战略性资源**
- **重要生产力**



Big Data, it's more than
what you think

Volume

大量

Velocity

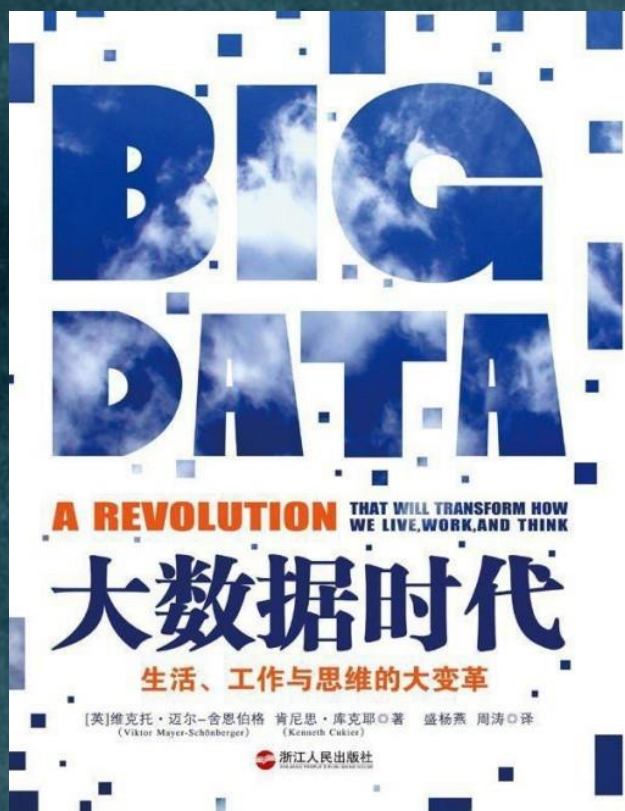
高速

Variety

多样

Value

价值



数据化

世间万物的数据化

- 当文字变成数据
- 当方位变成数据
- 当沟通变成数据

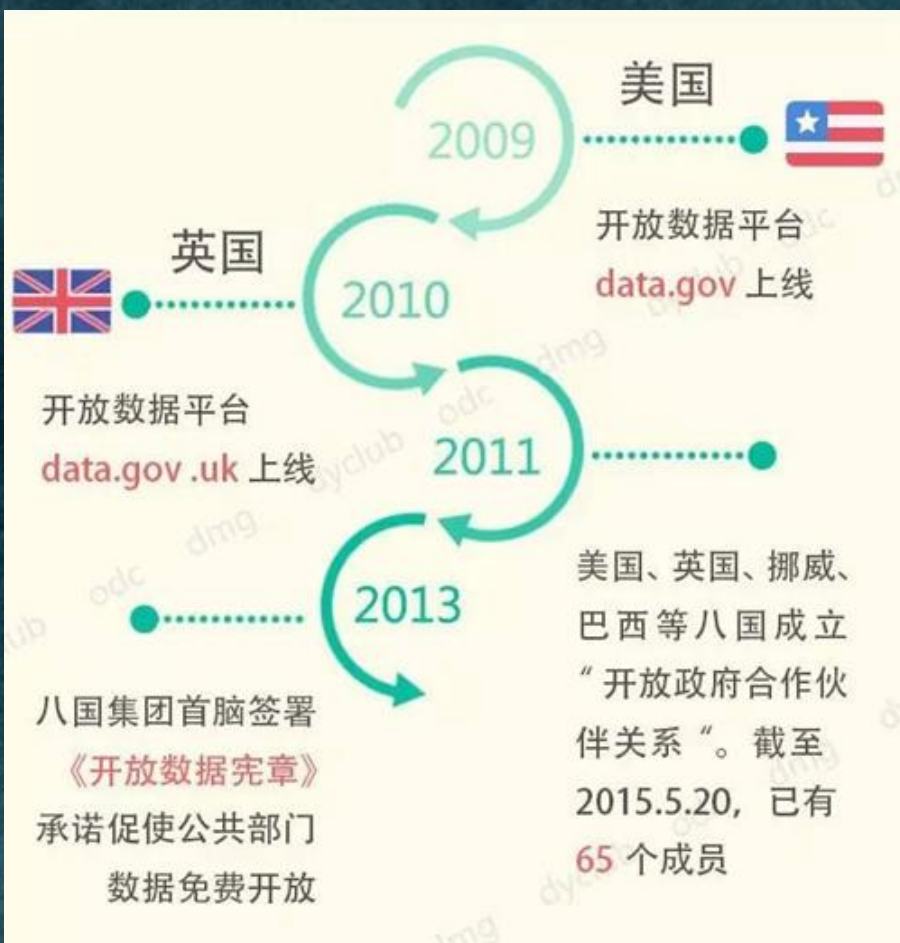
.....

2

开放数据

常速扫描

开放数据进程



“优易数据杯”全国开放大数据创新创意大赛

暨国家气象信息中心首届气象数据开放大赛

60年中国历史气象数据 10万奖金寻找最懂气象数据的人

[用户
反馈](#)



SODA

Shanghai Open Data Apps

上海开放数据创新应用大赛

开放数据：3项基本元素

- 1 非歧视性**：数据若开放，则其对任何人都开放。
- 2 机器可读性**：数据若开放，则其应是机器可读格式。
- 3 开放授权性**：数据若开放，则其对应授权条款应确保使用者自由免费访问、获取、使用、增值、演绎、拷贝、传播的权利。

-- 英国开放知识基金会（Open Knowledge Foundation）

几个概念

概念区分



√



不一定



×

所有人 免费 机器 无限制
可获取 使用 可读 使用

共享数据



公开数据



开放数据



开放数据种类

- 传统的统计数据
- 元数据、对象数据、规范数据
- 实时、连续发生的数据，如交易数据、行为数据、传感数据.....

开放数据领域

序号	数据分类(按字母顺序)	数据集实例
1	公司	公司/企业登记
2	犯罪与司法	犯罪统计、安全
3	地球观测	气象/天气、农业、林业、渔业和狩猎
4	教育	学校名单、学校表现、数字技能
5	能源与环境	污染程度、能源消耗
6	财政与合同	交易费用、合约、招标、地方预算、国家预算(计划和支出)
7	地理空间	地形、邮政编码、国家地图、本地地图
8	全球发展	援助、粮食安全、采掘业、土地
9	政府问责与民主	政府联络点、选举结果、法律法规、薪金(薪级)、招待/礼品
10	健康	处方数据、效果数据
11	科学与研究	基因组数据、研究和教育活动、实验结果
12	统计	国家统计局、人口普查、基础设施、财产、从业人员
13	社会流动性与福利	住房、医疗保险和失业救济
14	交通运输与基础设施	公共交通时间表、宽带接入点及普及率

高价值的
数据
14个重点领域

G8数据开放宪章的14个重点开放领域

开放数据格式

- 机器可读文件格式

JSON , XML , RDF , CSV , 电子表格 (如Excel等) , 文本文件 (如TXT等) , 专有格式 , HTML等

开放数据格式

- W3C 《开放政府数据指南》

鼓励政府以XML, RDF, CSV格式发布原始数据并形成结构化数据, 采用语义, 元数据, 标识等技术处理数据, 并用URI, RDF使之形成人和机器都可读的关联数据。

3

上图案例

家谱开放数据服务

- A
- B
- C
- D
- E
- F
- G
- H
- J
- K
- L
- M
- N
- O
- P
- Q
- R
- S
- T
- W
- X
- Y

X

玄
xuan
先祖名人 1
家谱文献 1

辛
xin
先祖名人 23
家谱文献 28

相
xiang
先祖名人 3
家谱文献 4

邢
xing
先祖名人 26
家谱文献 35

忻
xin
先祖名人 2
家谱文献 3

荀
xun
先祖名人 4
家谱文献 2

向
xiang
先祖名人 58
家谱文献 52

幸
xing
先祖名人 18
家谱文献 7

星
xing
先祖名人 3
家谱文献 1

肖
xiao
先祖名人 16
家谱文献 13

洗
xian
先祖名人 6
家谱文献 4

修
xiu
先祖名人 3
家谱文献 2

先祖名人

- 蕭邦寧
- 肖玉振
- 肖漢三
- 肖奎
- 肖華九

家谱

- [西昌社溪蕭氏族譜不分卷（江西省吉安市泰和縣）](#)

肖

xiao

现行较常见姓氏。汉、满、蒙、回、壮、苗、瑶、佤、水、侗、拉祜、普米、土家、锡伯等多个民族皆有此姓。《续通志》、《清通志》之《氏族略》俱收载。汉代有肖安国；明代有肖靖，宣德中解元；清代有肖雅奇，镶黄旗汉军，任三等护卫。

蕭邦寧 始遷祖

西昌社溪蕭氏族譜 泰和

- 责任者：康振春
- 摘要：始遷祖邦寧，字安中，宋代人。書名據版心題。
- 撰修时间：1912年
- 版本：木活字本
- 数量：二册
- 馆藏信息

开放数据的内容

姓氏

608种

谱籍地名

1609种

596家收藏机构

家谱文献

53987种

先祖名人

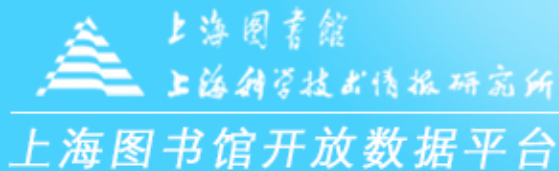
130867人

历史朝代

880种

<http://jp.library.sh.cn>

http://data.library.sh.cn



本网站是上海图书馆的开放数据平台，提供各种数据表、规范档、书目数据等。

中国历史纪年表

More...

Copyright@上海

朝代列表

开发者指南

在此输入朝代，年号，帝王姓名，公元纪年检索（起止年份请使用~分隔，如：1621~1849）

查询

- 夏 商 周 西周 东周 春秋 战国 秦 汉 西汉 东汉
- 三国 晋 西晋 东晋 南北朝 隋 唐 五代 宋 北宋 南宋
- 元 明 南明 清 民国

朝代	年号	帝王	帝王姓名	开始年份	终止年份	RDF
唐				618	907	RDF
	安乐		李轨	618	623	RDF
	天兴		刘武周	618	620	RDF
	昌达		朱粲	618	619	RDF
唐	武德	高祖	李渊	618	626	RDF
楚	太平		林士弘	618	622	RDF
梁	鸣凤		萧铣	618	621	RDF
燕	始兴		高开道	618	624	RDF
梁	永隆		梁师都	618	628	RDF
许	天寿		宇文文化及	618	619	RDF
高昌	义和		鞠訶	618	619	RDF

《上川明经胡氏世系表》

先祖名人
共8915人



关联数据 (Linked Data)

知识单元的细粒度化

知识组织的语义化

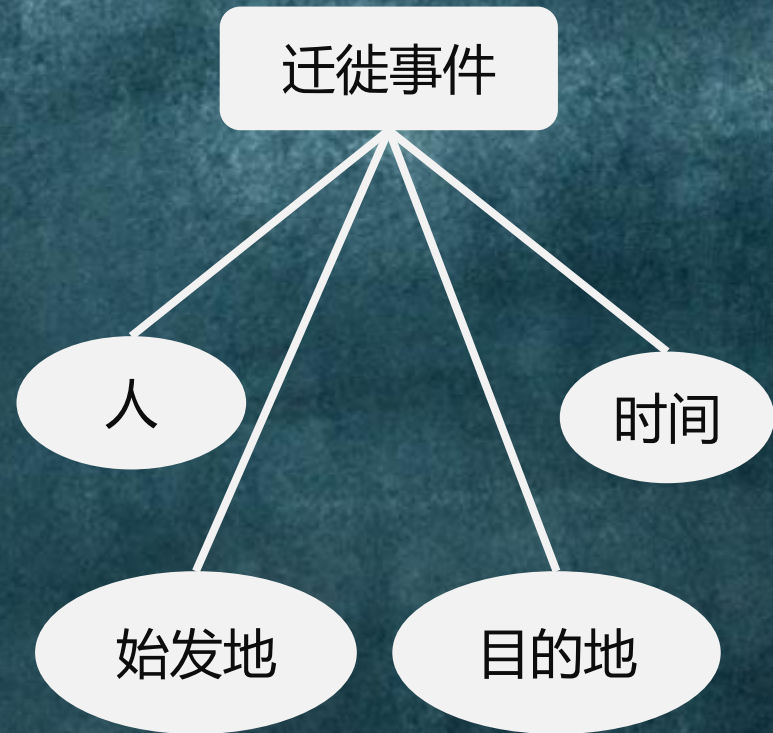
数据格式的标准化的

知识单元的的细粒度化

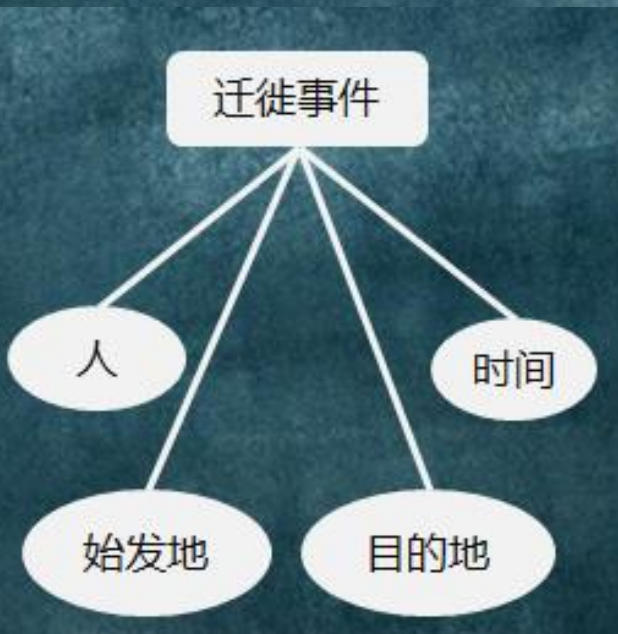
人物：胡昌翼
时间：唐昭宗天佑元年(公元904年)
始发地：长安
目的地：婺源考川

人物：胡延政
时间：北宋开宝八年(975年)
始发地：婺源考川
目的地：绩溪胡里

概念
抽象



基于本体的知识组织



从概念(Concept)
到本体(Ontology)

```
shl:Migration a rdfs:Class ;  
  rdfs:label "迁徙" ;  
  rdfs:subClassOf shl:Resource ;  
  rdfs:comment "一次迁徙事件。" .  
shl:originalLocality a rdf:Property ;  
  rdfs:label "原居地" ;  
  rdfs:domain shl:Migration ;  
  rdfs:range bf:Place ;  
  rdfs:comment "迁徙之前的居地。" .  
shl:locality a rdf:Property ;  
  rdfs:label "迁居地" ;  
  rdfs:domain shl:Migration ;  
  rdfs:range bf:Place ;  
  rdfs:comment "迁徙完成后的居地。"  
.
```

从结构化数据到机器可读的知识

描述胡适先生的RDF/Turtle数据

```
<http://data.library.sh.cn/jp/entity/person/uqwn21pa7iah9p87> a shl:Person ;
  shl:name "洪骅" ;
  foaf:name "胡适" ;
  shl:courtesyName "适之" ;
  foaf:familyName
<http://data.library.sh.cn/authority/familyname/rvmgzfsec8os93mv> ;
  shl:orderOfSeniority "3" ;
  shl:birthday "光绪辛卯十一月十七未时" ;
  shl:description "考派美国留学生名适字适之事迹见学了生光绪辛卯十一月十七
未时娶江氏生光绪庚寅十一月初八辰时" ;
  shl:relatedWork <http://data.library.sh.cn/jp/resource/work/jklhb5c3ga1rvxe3>;
  shl:roleOfFamily<http://data.library.sh.cn/jp/vocab/ancestor/xian-zu>;
  rel:spouseOf <http://data.library.sh.cn/jp/entity/person/kvop3qdif1okxhoa>;
  rel:childOf <http://data.library.sh.cn/jp/entity/person/mc7khkqgql3rpceb> .
```

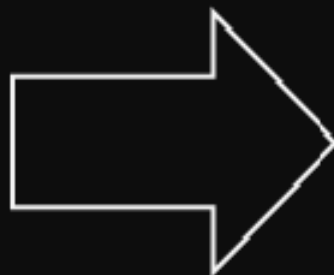

数据开放方式

SPARQL
Endpoint

Restful API

HTTP URI

数据消费接口



JSON-LD

返回数据



<http://pcrc.library.sh.cn/zt/opendata/>

SHANGHAI LIBRARY

开放数据 应用开发竞赛

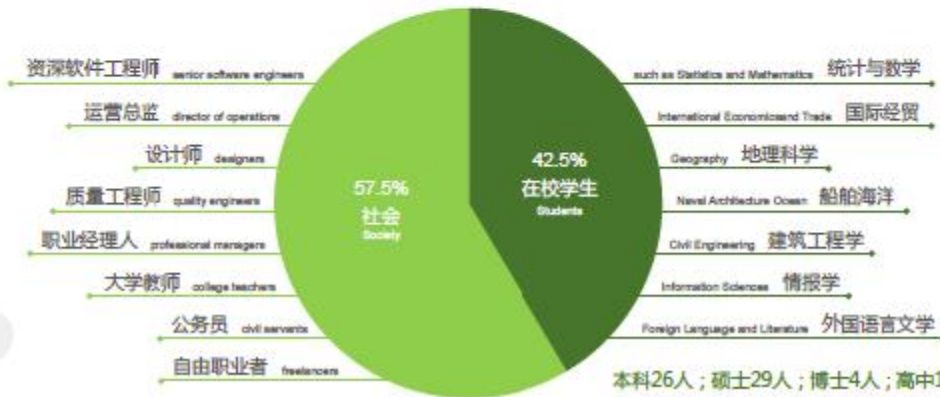
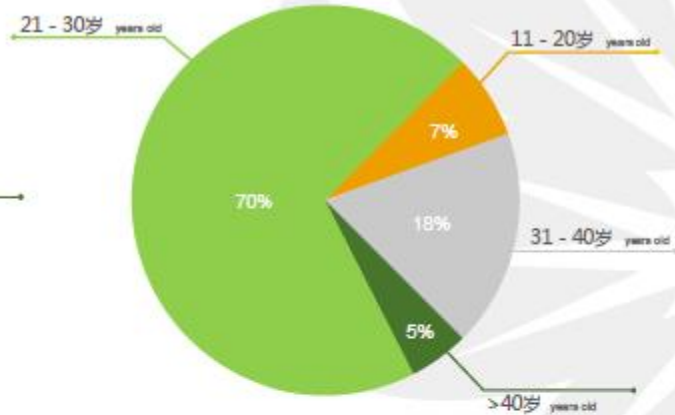
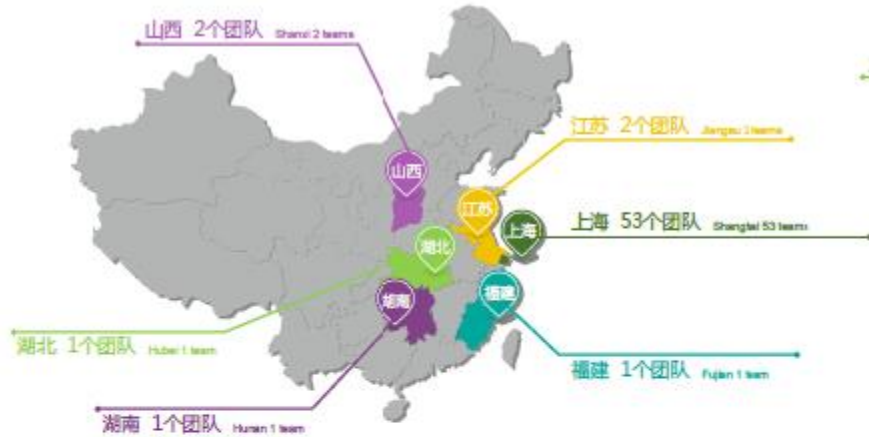
分享会通知

N

上图开放数据大赛



上图开放数据大赛



本科26人; 硕士29人; 博士4人; 高中1人; 小学1人
 26 undergraduates, 29 master degree candidates, 4 doctoral degree candidates, 1 high school student, 1 pupil

上图开放数据大赛

获奖名单 *Winners*

一等奖 1名

1st Prize

团队名：西交利物浦队

Team: XJTLU Team

作品名称：学习中国姓氏

Learn Chinese Surnames

二等奖 2名

2nd Prize

团队名：橄澜科技

Team: Ganlan

作品名称：靠谱网

KaoPu

团队名：ECNUGISer

Team: ECNUGISer

作品名称：家谱信息服务平台

Genealogy knowledge service platform

入围奖 5名

3rd Prize

团队名：时光寻根

Team: Qingtime

作品名称：家谱互动平台

Genealogy Edit and Multimedia
Publish Platform

团队名：南京谱牒研究

Team: Xungen

作品名称：上图家谱寻根系统

团队名：拥抱开放数据

Team: datahug

作品名称：“谱渡”家谱续修微网站

PuDu

团队名：海翼知图谱团队

Team: Hiekn-KG

作品名称：家谱社交平台

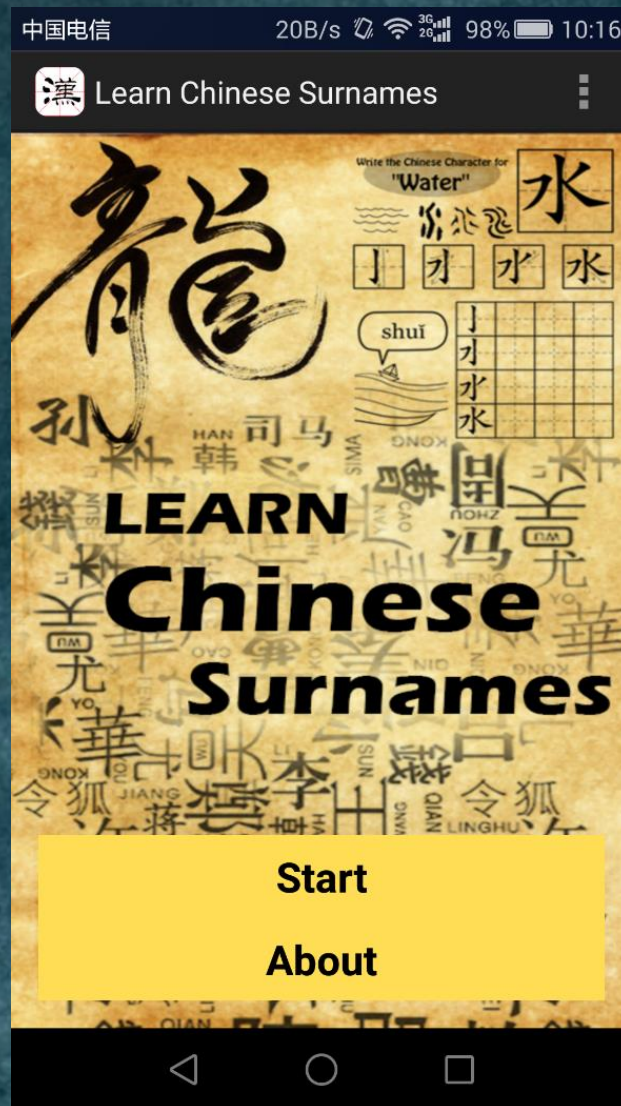
Genealogy Social Networking Platform

团队名：Alpha Force

Team: Alpha Force

作品名称：华夏家谱

上图开放数据大赛





The earliest 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（安徽省黄山主体宁邑） | Xiuning Xian, Anhui

上图开放数据大赛

主要用途 / Major usage

方便非中文母语人士了解中国姓氏文化和汉字文化

Enable non-native Chinese speakers to learn Chinese surname culture and Chinese character culture.

主要内容与功能 / Contents and functions

以2013年出版的《中国四百大姓》（袁义达，邱家儒主编）为依据，用英文描述从姓氏的汉语拼音字母顺序、姓氏的人口排行两个维度，展示了用英文描述的排行前400位*的中国姓氏及其起源，姓氏的简繁体汉字及其含义，与姓氏相关的名人、华人家谱及其收藏情况等。

Based on the "400 Most Common Chinese Sumames" (edited by Yuan Yida & Qiu Jiaru) published in 2013 and from the dimensions of pinyin alphabetical order and population ranking of Chinese surnames, the project presents in English the top 400* Chinese surnames and their origins, their corresponding Chinese characters in simplified and traditional forms, surname-related celebrities, family trees of the Chinese people and their collection status, etc.

*占全国人口总数97%、单姓人口在8万以上、位列中国姓氏排行榜前400位的400个姓氏

Top 400 Chinese surnames that account for 97% of the total Chinese population jointly and cover more than 80,000 persons each.

创意与应用价值 / Creativity and application value:

- 将中国家谱、姓氏文化、汉字文化三者巧妙地结合起来
Combining skillfully the Chinese genealogy, surname culture and Chinese character culture
- 实现上图家谱知识本体与DBpedia、Wikionary和GeoNames等知名数据源的关联
Realizing the linkage of Shanghai Library Genealogical Knowledge Database and other famous data sources, such as DBpedia, Wikionary and GeoNames.
- 展现出上图家谱开放数据在世界关联数据云中的潜在语义价值
Presenting the potential semantic value of Shanghai Library Open Genealogical Data in the world linked data cloud
- 为图书馆实现多元化和国际化的公共数据服务作了探索
Exploring the possibility of libraries to offer plural and international public data services

上图开放数据大赛



在近两个月的比赛过程中，项目成员通过共同讨论和远程协作(三个成员最后分散在三个不同的地区)，克服各类困难，付出了最大之努力，也加深了彼此的友谊。我们学习了上海图书馆的家谱知识本体、关联数据集的使用方法以及基本的安卓开发。在答辩会前后的交流中，小组成员也认识了图书情报界的老师和其它优秀的参赛队伍成员，受益良多。项目本身也具备丰富的学术价值，对于我们日后研究课题的延伸也有帮助。总之，这是一次特别难忘的比赛；它对于我们队有独特的意义。



Over the past 2 months, through discussion and remote collaboration, we overcame lots of difficulties, made the greatest efforts and deepened our friendship. We've learnt how to use Shanghai Library Genealogy Knowledge Database and the linked databases as well as the basic Android development. Through defense meeting, we have also got the opportunity to know some teachers in the Library Information field and other excellent rivals and have derived much benefit. To sum up, this is an impressive contest. It has a unique meaning for us.



4

收获与启示

简短小结

收获与启示

- 勇敢的踏浪者
- 潜在的数据价值
- 创新的智慧火花
- 志趣相投的朋友
- 大数据时代的角色定位

谢谢！

欢迎批评指正

qzhang@libnet.sh.cn