APEX 数据和知识管理实验室
DATA & KNOWLEDGE MANAGEMENT LAB

Zhishi.me
Weaving Chinese Linking Open Data

Xing Niu[1], Xinruo Sun[1], Haofen Wang[1], Shu Rong[1], Guilin Qi[2] and Yong Yu[1]
[1]Shanghai Jiao Tong University
[2]Southeast University

2011.10.25

Agenda

- Introduction
- Semantic Data Extraction
- Web Access to Zhishi.me
- Data-level Mapping among Different Datasets
- Conclusions and Future Work

Introduction

- Why we need Chinese Linking Open Data (CLOD)?

- Linked Open Data contains very sparse Chinese knowledge at the present time.
  - *Freebase*: Chinese labels
  - *DBpedia*: Chinese labels and short abstract [1]
  - *UWN* (Universal WordNet): maps Chinese words to corresponding vocabulary entities in WordNet [2]

Introduction (con't)

- In order to attract more efforts to publish Chinese semantic data linked to the CLOD, we prefer building a hub:
  - Covers all branches of knowledge
  - Contains abundant data

- We extracted knowledge from textual articles of other independent data sources
  - DBpedia is a representative one
  - Chinese Web-based collaborative encyclopedias together contain even more articles than the largest English one: Wikipedia

Baidu Hudong C-Wikipedia

Labels
Redirects
Abstracts
Infobox Properties
Internal Links
Related Pages
External Links
Categories

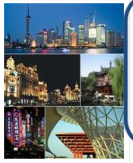
Baidu

已为你重定向至该词条，要参含义请查看上海市。

上海

编辑词条

百科名片



上海市

上海，中国大陆第一大城市；四个中央直辖市之一；是中国大陆的经济、金融、贸易和航运中心。上海创造和打破了中国世界纪录协会多项世界之最、中国之最。上海位于我国大陆海岸线中部的长江口，拥有中国最大的外贸港口、最大的工业基地。有超过2000万人居住和生活在上海地区，其中大部分属汉族江浙民系，通行吴语上海话。上海又是一座新兴的旅游目的地，具有深厚的近代城市文化底蕴和众多的历史古迹。如今上海已经发展成为一个闪耀全球的国际化大都市，并致力于在2020年建设成为国际金融中心和航运中心。上海是2010年世界博览会举办城市。

中文名称：	上海	方言：	吴语-太湖片-上海话
外文名称：	Municipality of Shanghai	气候条件：	亚热带季风气候
别名：	沪、申	著名景点：	外滩、豫园、陆家嘴、新天地、南京路、京山、枫泾古镇、淀山湖畔
行政区类别：	直辖市		
所属地区：	中国华东	机场：	上海虹桥机场、上海浦东国际机场
下辖地区：	黄浦区、徐汇区、静安区等	火车站：	上海火车站 上海南站 上海虹桥站
政府驻地：	黄浦区人民大道200号	车牌代码：	沪
电话区号：	021	市花：	白玉兰
邮政区码：	200000-202183	生产总值：	16872.42亿元(2010年)
地理位置：	长江三角洲	著名高校：	复旦、交大、同济、华师大、上大
面积：	6340.5平方公里（2009年末）	经纬度：	东经 121°4′ 北纬 31°2′
人口：	23019148人（2010年）	HDI指数：	0.908

目录

- 历史沿革
- 上海概况
- 行政区划
- 人口略述
- 政府领导
- 繁华都市

展开

历史沿革

编辑本段

古代

春秋时属吴，吴灭后入越，越王无疆大败于楚，后成为楚国春申君的封地（故上海别称"申"）。前223年秦灭楚后设会稽郡（治所在今苏州）并辖有嘤县（今昆山）、由拳县（今松江）和海盐县等。嘤县包括今上海全市大部（其余尚未成陆），前207年嘤县改名娄县。晋时松江（吴淞江）居民创造一种竹编捕鱼工具"扈"，又因江流入海处称"渎"，松江下游一带被称"扈渎"，后又改"扈"为"沪"。唐天宝十年（751年），吴郡刺史赵居贞上书，奏请划昆山南境、嘉兴东境、海盐北境，设华亭县。宋淳化二年（991年）松江淤浅，船泊支流"上海浦"。南宋咸淳三年(1267年)华亭县在上海浦西岸设"上海镇"。元朝至元十四年(1277年)，华亭县升格为府，翌年更名松江府，辖华亭县（一府一县）。元朝至元二十九年（1292年），析华亭县部分，设上海县，均隶于松江府。上海县县治在今黄浦区，这是上海建城的开始。[1]明宣德五年（1430年）应天巡抚陈规苏州，管辖南直隶（江苏和安徽全境），上海县归南直隶松江府管辖。清初改南直隶为江南省，上海县归江南省松江府。康熙六年（1667年）分江南省设江苏省和安徽省，从清朝前期开始，上海属江苏省（巡抚驻苏州）下辖的苏松太道[2]的松江府。江苏海关（简称江海关）由云台山（连云港）移来，江海关总监由驻南京的西江总督兼任。

松江府全境图（清嘉庆）

扩展阅读：

1. 上海市人民政府门户网站 http://www.shanghai.gov.cn
2. 上海市人民政府新闻办公室 http://www.shfb.gov.cn
3. 上海滩论坛 http://www.pudongba.com/
4. 上海外滩地标高清视频http://v.youku.com/v_showId_XMjUzNDAxMzMy.html

分类分类：

中国，直辖市，上海，申城，沪上

上海相关词条：

伦敦 纽约 巴黎 东京 芝加哥 洛兰克福 香港 洛杉矶 米兰 旧金山 悉尼 多伦多 苏黎世 布鲁塞尔 马德里

Hudong

上海市

分类：中国城市 中国直辖市 中国省级行政区划

编辑词条 分享

"上海"是"上海市"的同义词。词条由 小猪在线创建，主要 [illegible]，刘俊飞，多次修改，杨勇，小强啦啦[illegible]。

上海外滩

上海，中国大陆第一大城市；四个中央直辖市之一；是中国大陆的经济、金融、贸易和航运中心。上海创造和打破了中国世界纪录协会多项世界之最、中国之最。上海位于我国大陆海岸线中部的长江口，拥有中国最大的外贸港口、最大的工业基地。有超过2000万人居住和生活在上海地区，其中大部分属汉族江浙民系，通行吴语上海话。上海又是一座新兴的旅游目的地，具有深厚的近代城市文化底蕴和众多的历史古迹。如今上海已经发展成为一个闪耀全球的国际化大都市，并致力于在2020年建设成为国际金融中心和航运中心。上海是2010年世界博览会举办城市。

纠错 编辑摘要

目录

1 概述
2 历史沿革
3 行政区划
4 地质气候
5 人口

上海市 － 概述

纠错 编辑本段

上海位于北纬31度14分，东经121度29分。地处长江三角洲前缘，东濒东海，南临杭州湾，西接江苏、浙江两省，北界长江入海口，正当中国南北海岸线的中部，交通便利，腹地广阔，位置优越，是一个良好的江海港口。

点击查看详图

上海

地形 平均海拔高度为4米左右。

地形 除西南部有少数丘陵山脉外，上海境内全为坦荡低平的平原，是长江三角洲冲积平原的一部分。

能源 煤、石油、水力的储藏和生产，所需能源都需要其他省市的支援。但是，上海具有一定数量和较高质量的"二次能源"生产，产品主要是电力、石油油品、焦煤和煤气（包括液化石油气）。其他可以利用开发的能源还有沼气、风能、潮汐及太阳能。

生物 上海滨临东海，有丰富的水产资源，据统计，东海、黄海的水产资源有７００多种。此外，上海地处长江口，这里江面宽阔，海淡水交汇，是鱼类索饵、繁殖、栖息的场所，有各种鱼类２０８种，其中经济鱼类有２０多种。上海有众多的天然湖泊，螺蛳蚌等底栖生物资源比较丰富。稠密的水网，为淡水养殖提供了良好的条件。

航运 至2005年末，上海港拥有16条国际集装箱班轮航线；上海港每月集装箱航班数已达1967班，其中国际航班942班。

远洋航线：从上海始可以分别抵达香港、台湾（经第三地）、韩国、日本、东南亚、澳大利亚、以色列、地中海、西北欧、南非、南美、美国东西岸等地；

沿海航线：可抵达从北到市沿海主要港口；

长江航线：可直达长江中下游各港口；

内河航线：[illegible]

注释与参考：

[1] "上海市 http://www.china.com.cn/chinese/zhuanti/1663
[2] "上海饮食文化 http://www.dianping.com/group/communism/to
[3] "上海统计年鉴2000 http://www.stats-sh.gov.cn/2003shtj/tjnj/nj10.
[4] "宗教文化 http://[illegible]

- 上海官方网站：中国上海 http://www.shanghai.gov.cn/shanghai/node23
- 互动百科：上海博物馆百科 http://bwg.hudong.com/view/%E4%B8%8A%E6
- 2008年上海统计年鉴：主要年份各阶段教育实施情况 http://tjj.sh.gov.cn/2003shtj/tjnj/nj08.htm
- 中国新闻周刊：目睹中国电影兴衰流转：一个中国影院的百年变迁 2005-02-10 http://www.chinanews.com.cn/news/2005/200

被引用：上海市已被如下媒体引用 我来补充

分类分类：

中国城市 中国直辖市 中国省级行政区划 五角场

C-Wikipedia

上海市

维基百科，自由的百科全书

（重定向自上海）

上海市（发音（帮助·信息），吴语：Zaonhe），简称沪，别称申，中华人民共和国中央直辖市、国家中心城市。上海位于中国南北弧形海岸线中部，长江三角洲最东部。全市最北部为处于长江入海口中的崇明岛，东向东海，并隔海与日本九州岛相望，南濒杭州湾，西部与江苏、浙江两省相接。[3]

上海于新石器时代便有人类文明的遗迹[4]。先秦至元代，上海渐成为重要的粮、盐产地。明代起，上海因松江棉布而"衣被天下"[5]，成为重要的棉纺织基地。1843年开埠后，上海因地利、政策之便，在近现代科技文化与工商业领域等诸多方面领中国之先。经过百年的发展，上海形成今天的城市规模及风貌，从东南海滨仅有60万人口的县城发展成为今天拥有逾2000万居民、人口来自全球各地的国际都会。[6] 目前，上海是全球最重要的商业、经济、科教艺术城市之一。此外，上海也是中国内地的商业、经济中心以及国际航运和国际会展中心[7]，是世界公认的全球最主要快速发展的经济体的展示窗口，"为中外观瞻之所系"。[8] 根据2011年5月最新报告显示，上海市常住人口已突破2300万人。[9]

目录 [隐藏]

1 释名
2 简史
- 2.1 开埠前
- 2.2 开埠后
  - 2.2.1 开埠至1912年
  - 2.2.2 北洋时期
  - 2.2.3 国民政府时期
- 2.3 当代

3 自然地理
- 3.1 地理位置
- 3.2 气候
  - 3.2.1 气候数据
- 3.3 水文
- 3.4 岛屿

4 政治
- 4.1 行政区划
  - 4.1.1 飞地
- 4.2 外事
  - 4.2.1 领事机构
  - 4.2.2 友好城市

18 外部链接

上海市市标[注 1]

汉语拼音	Shànghǎi Shì
简称	沪
别称	申
地名出处	吴淞江支流上海浦

上海市在中华人民共和国的地理位置

概览

国家	中华人民共和国
行政区类型	直辖市
行政区划代码	31
下级行政区	17市辖区、1县
设市时间	1927年设上海特别市
现任市长	韩正
现任市委书记	俞正声
政府驻地	黄浦区人民大道200号
毗邻	江苏省、浙江省
城市精神	海纳百川 追求卓越 开明睿智 大气谦和

上海市市花白玉兰

城市（地区）

编辑

中文名：	上海市
英文名：	Shanghai
别称：	申
所属地区：	中国 上海
建立时间：	春秋战国时期
面积：	6340.5
人口：	1888.46万（2008年）
GDP：	13,698.15亿元人民币（2008年）
地理位置：	位于北纬31度14分，东经121度29分，地处太平洋西岸，亚洲大陆东沿，长江三角洲前缘
友好城市：	已与世界上52个国家的70个城市（省、州、大区、道、府、县或区）结成友好城市或建立友好交流关系
著名景点：	龙华寺、孔庙、豫园、玉佛寺、方塔、朱家角、中共一大会址、中共二大会址、孙中山故居、鲁迅故居、周公馆、毛泽东故居、外滩、东方明珠广播电视塔、人民广场、金茂大厦、上海海洋水族馆、上海新天地、上海科技馆
主要街道：	老城隍庙、云南路美食街、南京路商业街、淮海路商业街、徐家汇商业圈
地标：	上海东方明珠广播电视塔
名人：	鲁迅、巴金、茅盾、张爱玲、胡蝶、阮玲玉、周璇、陈燕燕、白杨、王人美、黎莉莉、上官云珠、徐来、陈娟娟、袁美云、谈瑛、黄宗英、赵丹、刘琼、金焰、金山、魏鹏程、韩兰根、周信勋、孙道临
特产：	沪菜，又叫本帮菜、八宝鸭、糟钵头、虾子大乌参、竹笋腌鲜
知名产业：	国际经济、金融、贸易和航运中心
机场：	虹桥国际机场，浦东国际机场
电话区码：	021
邮编区码：	200000-202183

地图

相关词条 编辑

Dict.cn

city of Shanghai

释名 [编辑]

上海一词最早始于宋代，源于一条名为上海浦的吴淞江支流。浦，原为吴语中小河的意思，且一般多指人工河。因此，上海浦（或黄浦江）因为原是吴淞江上的一条小河，而得名。当然与上海浦相对的有下海浦，上海浦位于吴淞江的南岸，下海浦位于吴淞江的北岸，地点大致位于今虹口区海门路一带黄浦江的南北岸。此外，江左地区命名河流有一个习惯，通常将靠近源头的支流谓之里或上，靠近下游河口地区的称为外或者下。而上海浦因较为靠近上游，故命名为上海浦。后由于沪上酒业发展，政府在本地设置了征收酒税的相关机构——酒务。因地近上海浦，所以称之为上海务。上海一词的影响逐渐扩大，而下海浦则伴随吴淞江的淤塞而逐渐消失，现仅有下海庙以证明下海浦的大致位置[10]。

此外，上海的简称"沪"，来源于古代上海一带的居民进行捕鱼作业的一种竹制工具。而别称"申"则是因为相传上海西部地区曾为楚国公子春申君黄歇的封地，因而得名[11]。

外部链接 [编辑]

概览

- 中国上海网：上海概览

政府

- 中国上海——上海市人民政府门户网站
- 上海人大公众网
- 上海政协
- 上海市人民政府新闻办公室

维基共享资源 中相关的多媒体资源：上海市

维基词典上的词义解释：上海

10个分类：上海铁路车站 | 上海 | 中华人民共和国直辖市 | 中国特大城市 | 中国共产党革命圣地 | 太平洋沿海城市 | 长江三角洲城市 | 国家历史文化名城 | 中国沿海开放城市 | 国家园林城市

Semantic Data Extraction

- Wikipedia uses the wikitext language, a lightweight markup language
  - database backup dumps
- Baidu Baike and Hudong Baike provide the WYSIWYG (what you see is what you get) HTML editors
  - HTML file archives

- 12 types of article content are considered
  - abstracts, aliases, categories, disambiguation, external links, images, infobox properties, internal links (pagelinks), labels, redirects, related pages and resource ids.

- Totally, 124,573,857 RDF triples are extracted

Overall Statistics on Extraction Results

Items	Baidu Baike		Hudong Baike		Chinese Wikipedia	
Resources	**3,234,950**		2,765,833		559,402	
∼ that have abstracts	393,094	12.2%	469,009	17.0%	324,627	**58.0%**
∼ that have categories	2,396,570	**74.1%**	912,627	33.0%	314,354	56.2%
∼ that have infoboxes	56,762	1.8%	197,224	**7.1%**	24,398	4.4%
Categories	**516,309**		38,446		93,191	
Properties	**13,226**		474		2,304	
		per res.		per res.		per res.
Article Categories	6,774,442	**2.09**	2,067,349	0.75	796,679	1.42
External Links	2,529,364	0.78	827,145	0.30	573,066	**1.02**
Images	2,593,856	**0.80**	1,765,592	0.64	221,171	0.40
Infobox Properties	477,957	0.14	1,908,368	**0.69**	120,509	0.22
Internal Links	15,462,699	4.78	19,141,664	6.92	9,359,108	**16.73**
Related Pages	2,397,416	0.74	17,986,888	**6.50**	—	—
Aliases	—		—		**362,495**	
Disambiguation Links	28,937		13,733		**40,015**	
Redirects	97,680		37,040		**190,714**	

Overall Statistics on Extraction Results

Items	Baidu Baike		Hudong Baike		Chinese Wikipedia	
Resources	**3,234,950**		2,765,833		559,402	
∼ that have abstracts	393,094	12.2%	469,009	17.0%	324,627	**58.0%**
∼ that have categories	2,396,570	**74.1%**	912,627	33.0%	314,354	56.2%
∼ that have infoboxes	56,762	1.8%	197,224	**7.1%**	24,398	4.4%
Categories	**516,309**		38,446		93,191	
Properties	**13,226**		474		2,304	
		per res.		per res.		per res.
Article Categories	6,774,442	**2.09**	2,067,349	0.75	796,679	1.42
External Links	2,529,364	0.78	827,145	0.30	573,066	**1.02**
Images	2,593,856	**0.80**	1,765,592	0.64	221,171	0.40
Infobox Properties	477,957	0.14	1,908,368	**0.69**	120,509	0.22
Internal Links	15,462,699	4.78	19,141,664	6.92	9,359,108	**16.73**
Related Pages	2,397,416	0.74	17,986,888	**6.50**	—	—
Aliases	—		—		**362,495**	
Disambiguation Links	28,937		13,733		**40,015**	
Redirects	97,680		37,040		**190,714**	

Web Access to Zhishi.me

- 1. Linked Data

- The Chinese characters are non-ASCII, so we choose IRIs
- IRI Patterns
  - http://zhishi.me/**[DataSource]**/resource/**[Label]**
  - e.g. http://zhishi.me/**hudongbaike**/resource/北京

- IRIs are incompatible with HTML4 [3], we have to encode non-ASCII characters in some cases
  - e.g. http://zhishi.me/**hudongbaike**/resource/**%E5%8C%97%E4%BA%AC**

Label

Abstract

Infobox

Subjects

Thumbnails

Other properties

hudong:北京

foaf:page http://www.hudong.com/wiki/%E5%8C%97%E4%...i/%E5%8C%97%E4%BA%AC
zhishi:resourceID 757341

owl:sameAs

- ☑ hudong:北京 (this)
- ☑ baidu:北京
- ☑ zhwiki:北京

MERGE PAGE

Index

- zhishi:abstract
- infobox
- dcterms:subject
- zhishi:thumbnail
- zhishi:relatedPage
- zhishi:externalLink
- zhishi:internalLink

zhishi:abstract

北京是中华人民共和国的首都，四个直辖市之一，是中国的政治、文化、科技、教育和国际交往中心，也是中国最大的陆空交通枢纽和中国北方的经济中心。北京市位于华北平原北端，东南局部地区与天津市相连，其余为河北省所环绕。以市区人口数相比，北京为次于上海的中国第二大城市。北京是一座有三千余年历史的文化名城，历史上共有五个朝代曾在此定都，北京荟萃了自元明清以来的中华文化，拥有诸多名胜古迹和人文景观。北京于2008年成功举办了第29届夏季奥林匹克运动会（2008年8月8日至8月24日）及残疾人奥林匹克运动会（2008年9月6日至9月17日）。

>>TOP

infobox

所属地区	中国
GDP	11,865.9亿元人民币(2009年)
主要街道	长安街、平安大街
面积	16800平方公里
中文名	北京
下辖地区	海淀区 西城区 顺义区 大兴区
友好城市	纽约、东京、巴黎
知名企业	首钢集团、北京科技园置地有限公司、北京京宝公司、北京珠联公司
特产	烤鸭、豌豆黄、艾窝窝、豆汁
人口	1,755万人(2009年)
电话区码	010
英文名	Municipality of Beijing

>>TOP

dcterms:subject

水浒地名 | 文化 | 首都 | 各国首都 | 植物 | 历史 | 互联网 | 地理 | 中国 | 中国城市

>>TOP

zhishi:thumbnail

>>TOP

zhishi:relatedPage

圆明园 SHOW MORE (10)

>>TOP

zhishi:externalLink

http://www.cq6.cc/index.php%3Fdoc-innerlink-%25E5%258C%2597%25E4%25BA%25AC
SHOW MORE (14)

>>TOP

zhishi:internalLink

福建省 SHOW MORE (798)

>>TOP

Web Access to Zhishi.me (con't)

- 2. Lookup Service (http://zhishi.me/lookup/)

- 3. SPARQL Endpoint (http://zhishi.me/sparql/)
  - AllegroGraph RDFStore is used to store the extracted triples and provide querying capabilities.

ZHISHI.me

hudong:北京

foaf:page http://www.hudong.com/wiki/%E5%8C%97%E4%...i/%E5%8C%97%E4%BA%AC
zhishi:resourceID 757341

Index

- zhishi:abstract
- infobox
- dcterms:subject
- zhishi:thumbnail
- zhishi:relatedPage
- zhishi:externalLink
- zhishi:internalLink

owl:sameAs

- hudong:北京 (this)
- baidu:北京
- zhwiki:北京

MERGE PAGE

zhishi:abstract

北京是中华人民共和国的首都，四个直辖市之一，是中国的政治、文化、科技、教育和国际交往中心，也是中国最大的陆空交通枢纽和中国北方的经济中心。北京市位于华北平原北端，东南局部地区与天津市相连，其余为河北省所环绕。以市区人口数相比，北京为次于上海的中国第二大城市。北京是一座有三千余年历史的文化名城，历史上共有五个朝代曾在此定都，北京荟萃了自元明清以来的中华文化，拥有诸多名胜古迹和人文景观。北京于2008年成功举办了第29届夏季奥林匹克运动会（2008年8月8日至8月24日）及残疾人奥林匹克运动会（2008年9月6日至9月17日）。

>>TOP

infobox

所属地区	中国
GDP	11,865.9亿元人民币(2009年)
主要街道	长安街、平安大街
面积	16800平方公里
中文名	北京
下辖地区	海淀区 西城区 顺义区 大兴区
友好城市	纽约、东京、巴黎
知名企业	首钢集团、北京科技园置地有限公司、北京京宝公司、北京珠联公司
特产	烤鸭、豌豆黄、艾窝窝、豆汁
人口	1,755万人(2009年)
电话区码	010
英文名	Municipality of Beijing

>>TOP

dcterms:subject

水浒地名 | 文化 | 首都 | 各国首都 | 植物 | 历史 | 互联网 | 地理 | 中国 | 中国城市

>>TOP

zhishi:thumbnail

>>TOP

zhishi:relatedPage

圆明园 SHOW MORE (10)

>>TOP

zhishi:externalLink

http://www.cq6.cc/index.php%3Fdoc-innerlink-%25E5%258C%2597%25E4%25BA%25AC
SHOW MORE (14)

>>TOP

zhishi:internalLink

福建省 SHOW MORE (798)

>>TOP

owl:sameAs
hudong:北京 (this)
baidu:北京
zhwiki:北京
MERGE PAGE

Data-level Mapping among Different Datasets

- We utilize distributed MapReduce [4] framework to accomplish the large-scale data-level mapping task.
- Map phase:
  - Resources are indexed
- Reduce phase:
  - Resources with the same index term (match candidates) gather together and further comparisons are made
- The improved version of our mapping framework participates in IM@OAEI2011
  - "Zhishi.links Results for OAEI 2011" @OM2011

Data-level Mapping (con't)

- In this paper, we focused on resolving the problem occurs in Map phase (index generating step):
  - different resources have the same label
  - the same resource have different labels

- We proposed three reasonable but not complex strategies to generate the index:
  - Using Original Labels
  - Punctuation Cleaning
  - Extending Synonyms

Data-level Mapping (con't)

- Using Original Labels
- We extract different meanings of homonyms as different resources
  - e.g. "Jupiter (mythology)" != "Jupiter Island"
  - it is impossible to find two resources that have different meanings with the same label if all homonyms are recognized

Data-level Mapping (con't)

- Punctuation Cleaning

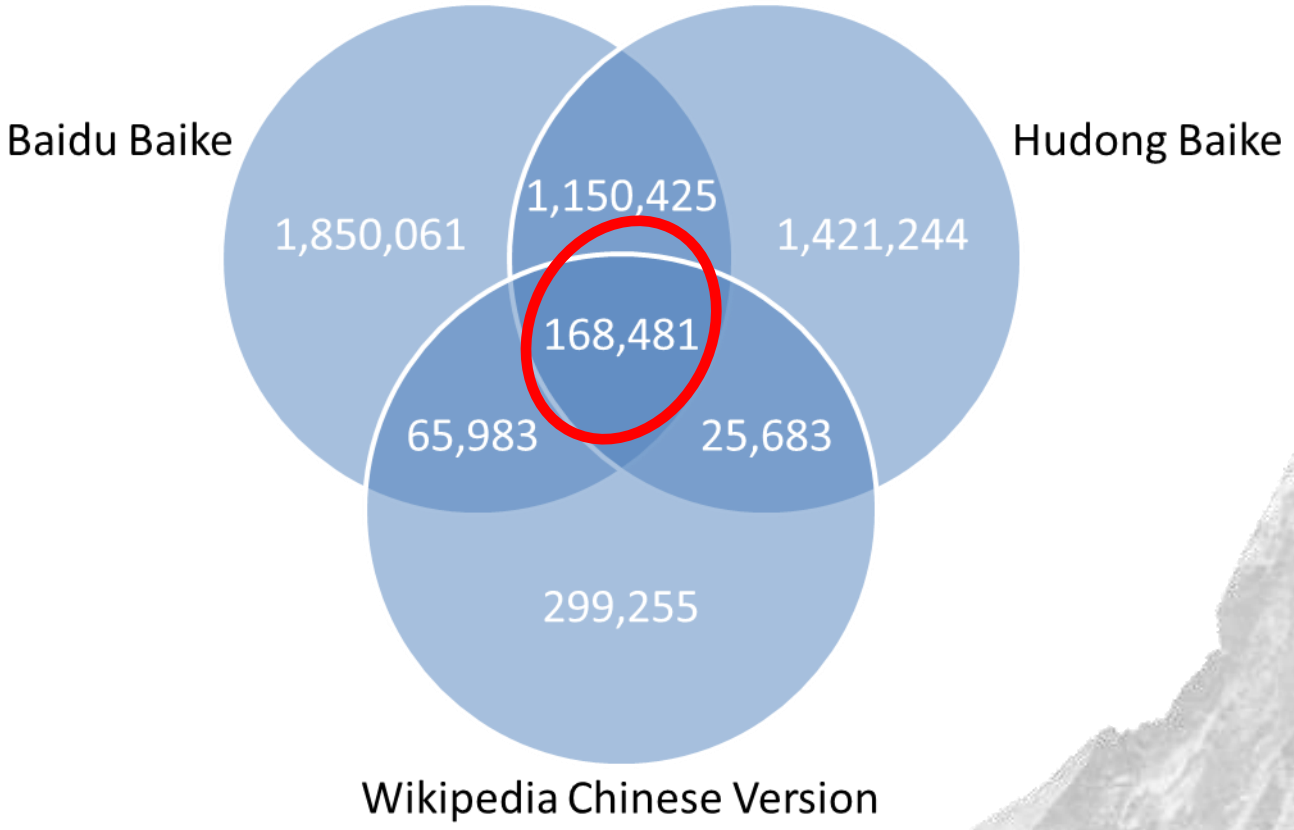
- It is one of the most efficient methods we used to discover mappings between resources with different labels
  - Just one case of Punctuation Cleaning is given here:
  - In Chinese, we often insert an interpoint (·) or a hyphen (-) between two personal name components. In some certain cases, people may just adjoin these components.

  - e.g. a comet named after two persons
  - 海尔波普彗星 = 海尔·波普彗星 = 海尔-波普彗星

Data-level Mapping (con't)

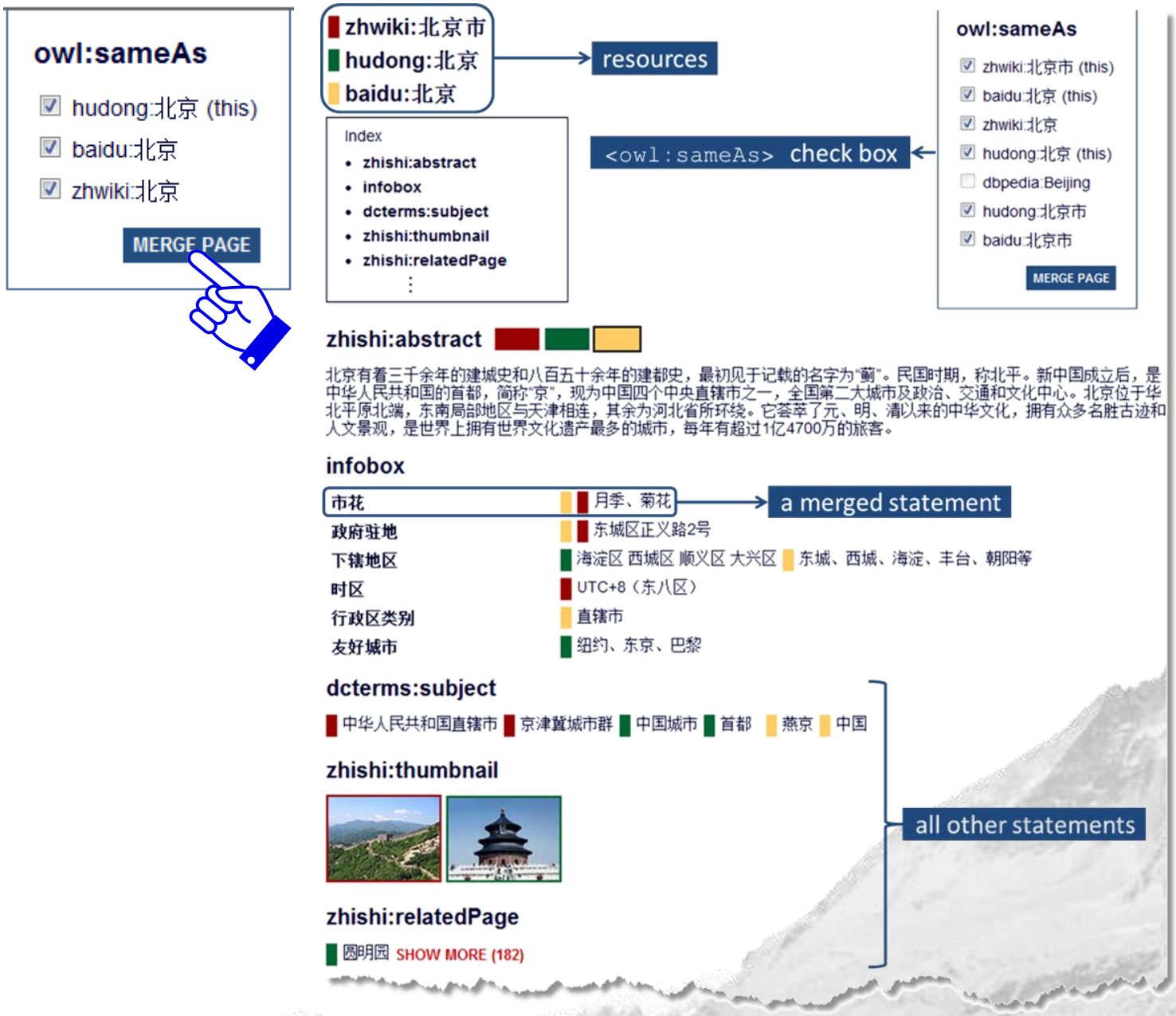
- Extending Synonyms
- Making use of high quality synonym relations obtained from redirects information
  - A redirects to B means A and B are synonyms
  - Label(A) and Label(B) are both used as index terms for a single resource

Data-level Mapping (con't)

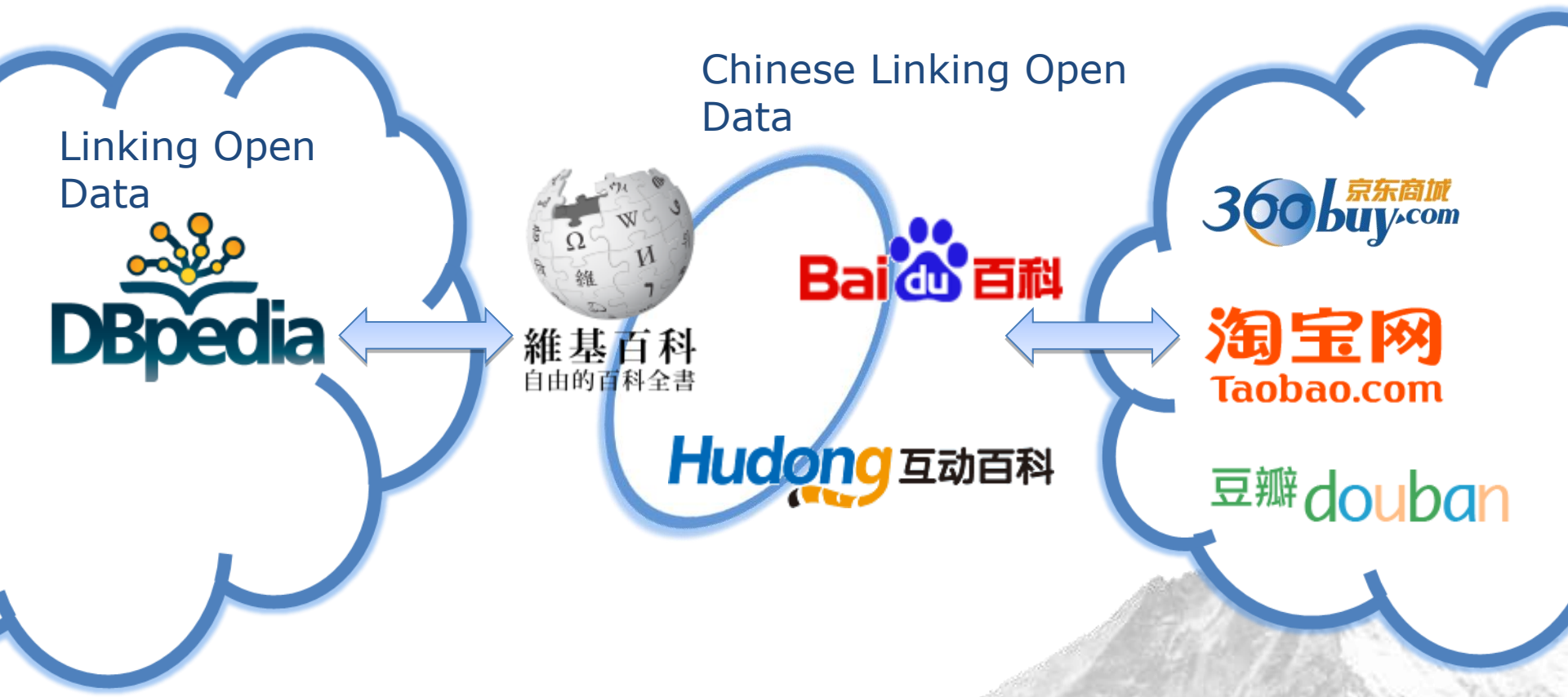


- Intersections of Our Three Data Sources Shown in Venn Diagram

Data-level Mapping (Web Access)



Data-level Mapping (con't)



Conclusions and Future Work

- Conclusions
  - Zhishi.me is the first effort to build Chinese Linking Open Data
  - We extracted semantic data from three Web-based free-editable encyclopedias
  - Three heuristic strategies were adopted to discover <owl:sameAs> links between equivalent resources
  - We provided Web access entries to our knowledge base for both professional and non Semantic Web community users.

- Future Work
  - Several Chinese non-encyclopedia data sources will be accommodated in our knowledge;
  - We will improve instance matching strategies and provide necessary evaluations of matching quality;
  - Refine extracted properties and building a general but consistent ontology automatically.

APEX 数据和知识管理实验室
DATA & KNOWLEDGE MANAGEMENT LAB

Thanks!

References

[1] Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia - A crystallization point for the Web of Data. J. Web Sem. 7(3), 154–165 (2009)

[2] de Melo, G., Weikum, G.: Towards a universal wordnet by learning from combined evidence. In: Cheung, D.W.L., Song, I.Y., Chu, W.W., Hu, X., Lin, J.J. (eds.) CIKM, pp. 513–522. ACM (2009)

[3] Raggett, D., Hors, A.L., Jacobs, I.: HTML 4.01 Specication - Appendix B: Performance, Implementation, and Design Notes. W3C Recommendation (December 1999), http://www.w3.org/TR/html4/appendix/notes.html

[4] Dean, J., Ghemawat, S.: MapReduce: Simplified Data Processing on Large Clusters. In: OSDI. pp. 137–150 (2004)