Enrichment of Cross-Lingual Information on Chinese Genealogical Linked Data

iConference 2017, Wuhan, China, March 22-25, 2017

Hang Dong

University of Liverpool, Department of Computer Science

Linked Datasets as of August 2014

UNIVERSITY OF LIVERPOOL

About the research

- This study aims to address language barrier issues for Chinese cultural heritage resources in Linked Open Data, based on a project wining the Shanghai Library Open Data Competition in 2016.

- Background of this research: Language barrier in the LOD cloud
- Methods for Cross-Lingual Data/Ontology Matching, in cultural heritage domain
- Case study on Chinese Genealogical Linked Data

A Cross-Lingual IR Scenario

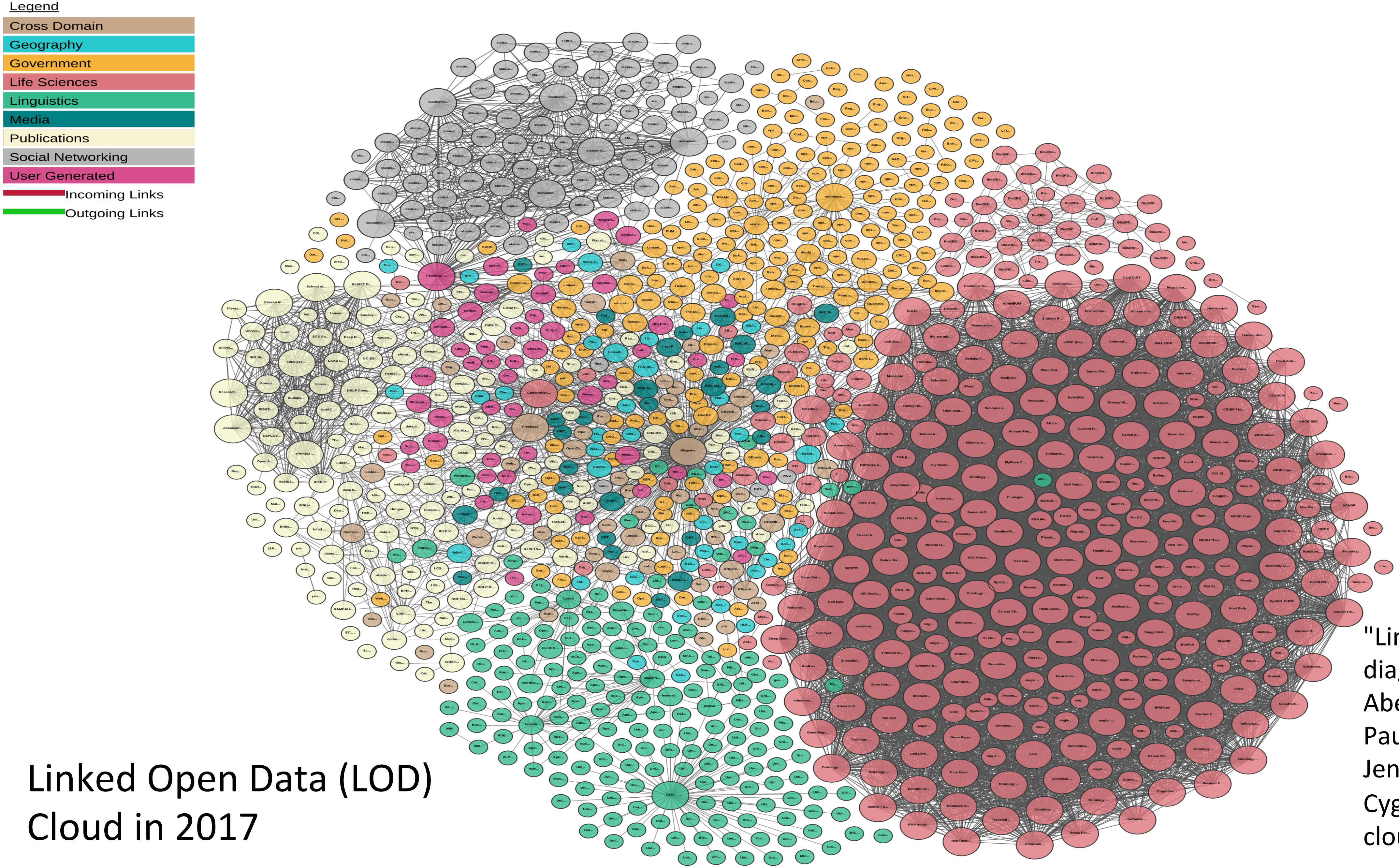
A people from the US wants to know the history of a Chinese surname "Liu", genealogy (family books) of this Surname, the Chinese character, stroke of the character, and he tries to search it online. He is not very familiar with Chinese. What would he do?

What is special here?

Multiple type of information

Cultural specific information

A Cross-Lingual searching in the LOD Cloud?



Linked Open Data (LOD) Cloud in 2017

"Linking Open Data cloud diagram 2017, by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/"

Towards a Multi-lingual Linked Open Data (LOD)?

- Low multi-linguality: (Gómez-Pérez et al., 2012)
    - Most of LOD data are in English (around 85% of all tagged literals).
    - Monolingual dataset 78.9%.
    - Multilingual datasets doubled during 2012.

- The increasing amount of non-English monolingual data are forming **"islands"** in the LOD cloud.

Addressing the Language barrier in the LOD Cloud

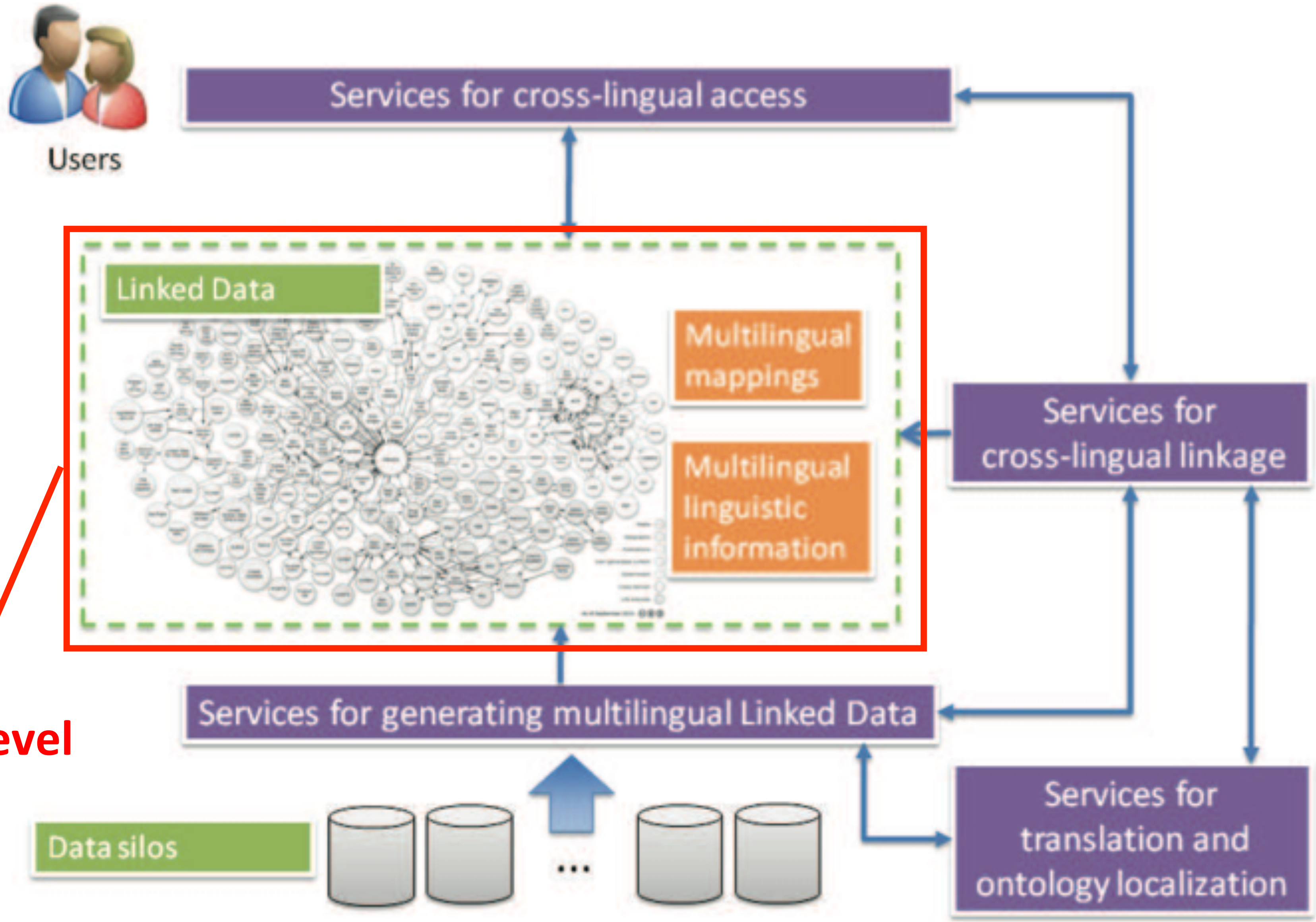
Research Question:

- 1 How to link Chinese LOD data to English described LOD data to prevent “island”?
- 2 Any practice in the cultural heritage domain and how to enable cross-lingual access of cultural heritage data?

Macro-level

Enrichment of Cross-Lingual Information from the micro-level perspective

Micro-level


Users
Services for cross-lingual access
Linked Data
Multilingual mappings
Multilingual linguistic information
Services for cross-lingual linkage
Services for generating multilingual Linked Data
Data silos
...
Services for translation and ontology localization

A proposed architecture of a multilingual Web of Data by Gracia et al. (Image from Gracia et al. 2012)

Realizing a Multi-lingual Web of Data

- A macro-level perspective:
  - (semi-) automatic services and models built on top of the LOD infrastructure to enable multi-lingual data generation, representation and cross-lingual access.
- A micro-level perspective: a layer of additional information,
  - multilingual linguistic information
  - multilingual mapping between ontologies/vocabularies
  - multilingual mapping between instances/data

Methods for Cross-lingual Ontology/Data Matching

- 1. Translation:
  - Machine Translation for ontology matching (Trojahn et al. 2008)
  - Manual Translation

- 2. Ontology Alignment and Instance Linking
  - Natural Language Processing and Machine Learning: for large alignment of data and vocabularies (Ngai, Carpuat, & Fung, 2002; Wang et al., 2013)
  - String matching to linking to a multi-lingual source (to Wikipedia in Damova et al., 2014).

- 3. Crowdsourcing: users' power (e.g. Dbpedia)

Cross-lingual cultural heritage information

- Direction 1： on the vocabulary level
  - Manual translation to align a Chinese museum controlled vocabulary to a US art controlled vocabulary, AAT (Chen & Chen, 2012)

- Direction 2： on the entity/data level
  - Manual translation of metadata of Chinese painting resources to English (Matusiak et al., 2015)

For Linked Open Data: MOLTO (Damova et al., 2014) and Europeana (Stiller et al., 2014)

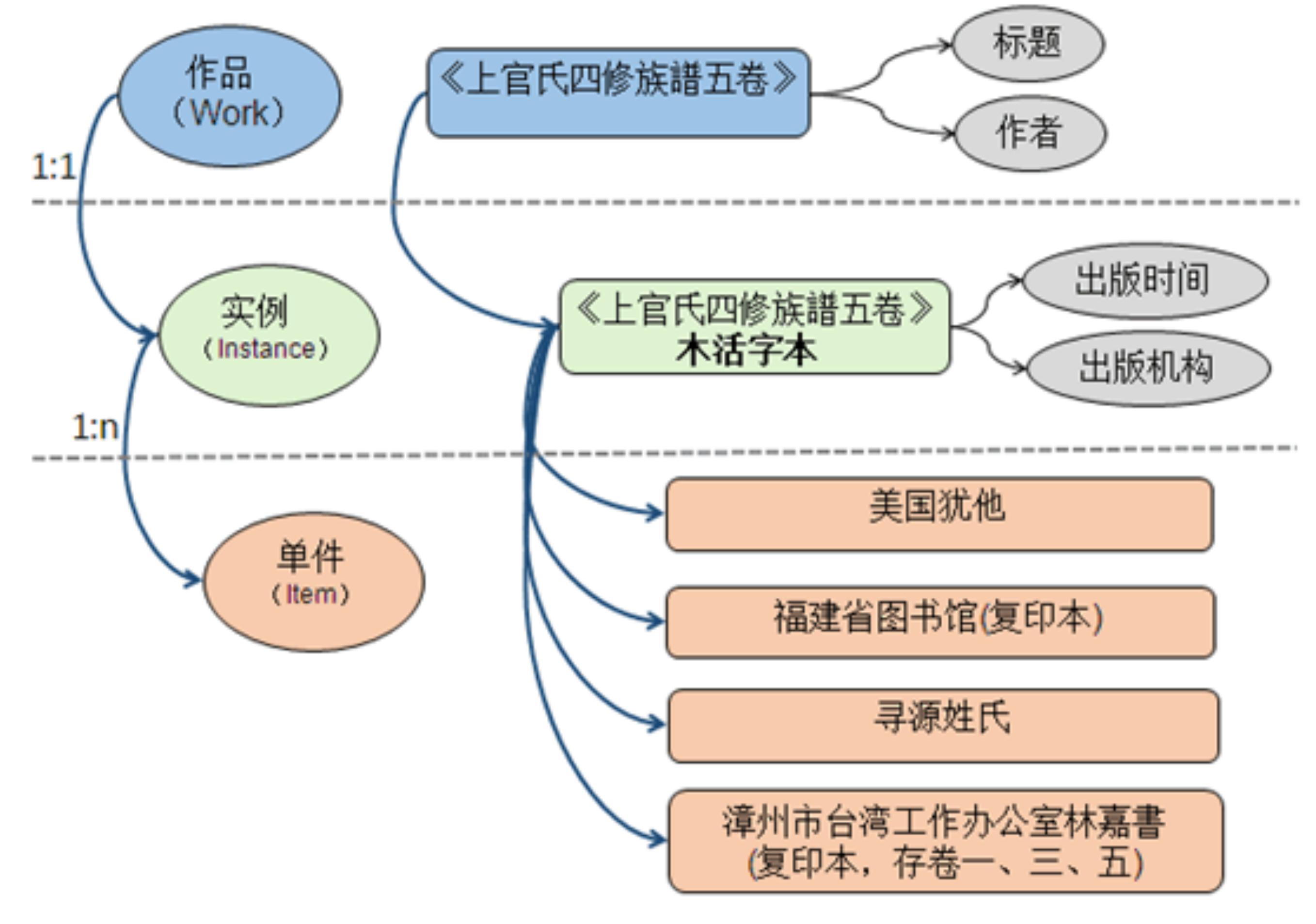
- Concerning the both directions on a large scale, manual + machine.
- The Europeana project has a structured approach with thorough validation.
- Among European languages.

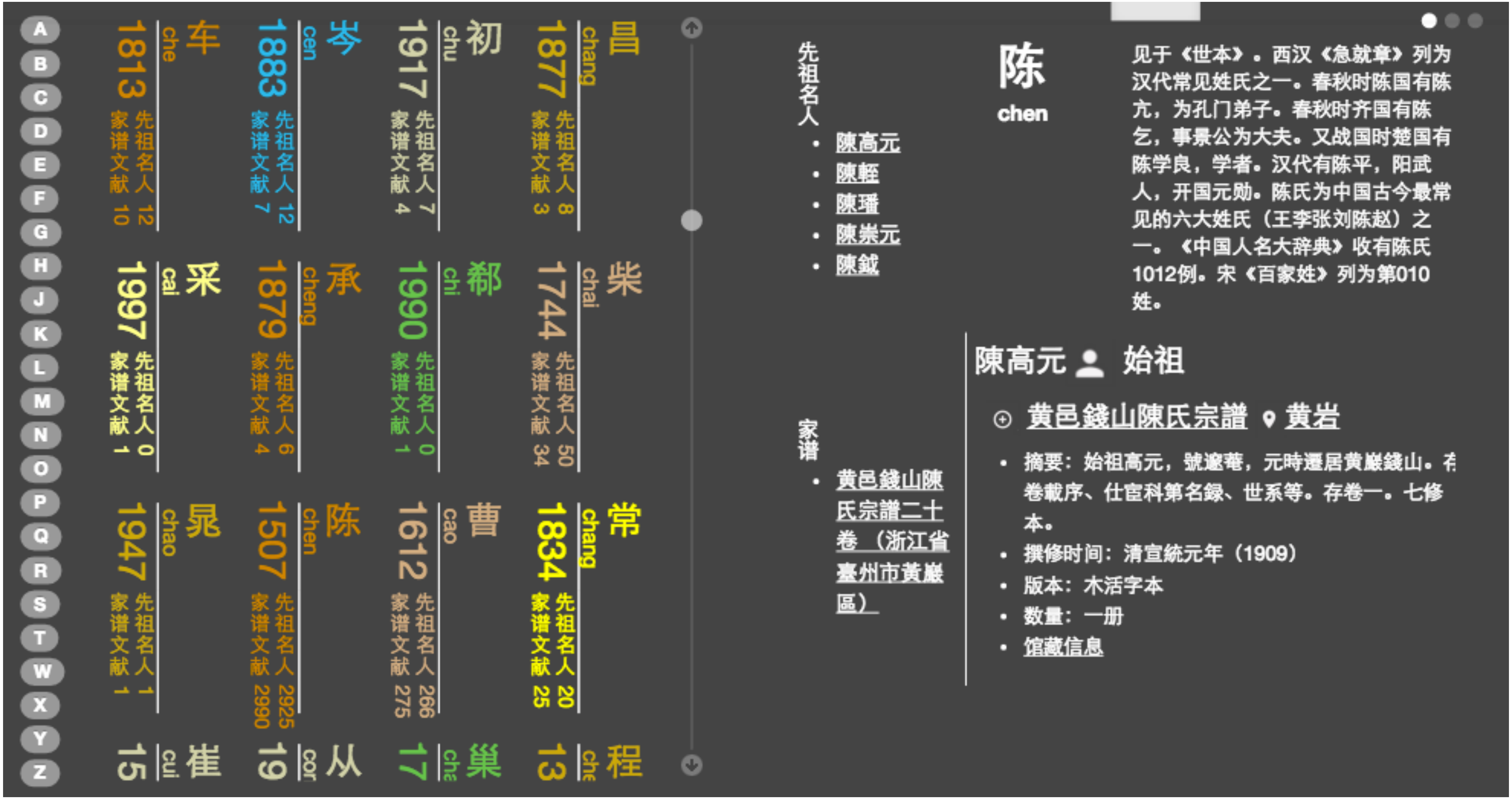
Enrichment of C-L information on Chinese Genealogical Data

- Chinese Genealogical Data from Shanghai Library
  - Published as Linked Data
  - Almost Mono-Lingual

- Method:

Match to multi-lingual data in LOD





昌 chang 1877 先祖名人 8 家谱文献 3

初 chu 1917 先祖名人 7 家谱文献 4

岑 cen 1883 先祖名人 12 家谱文献 7

车 che 1813 先祖名人 12 家谱文献 10

柴 chai 1744 先祖名人 50 家谱文献 34

郗 chi 1990 先祖名人 0 家谱文献 1

承 cheng 1879 先祖名人 6 家谱文献 4

采 cai 1997 先祖名人 0 家谱文献 1

常 chang 1834 先祖名人 20 家谱文献 25

曹 cao 1612 先祖名人 266 家谱文献 275

陈 chen 1507 先祖名人 2925 家谱文献 2990

晁 chao 1947 先祖名人 1 家谱文献 1

程 che 13

巢 cha 17

从 cor 19

崔 cui 15

先祖名人

- 陳高元
- 陳輊
- 陳璠
- 陳崇元
- 陳鉞

家谱

- 黄邑錢山陳氏宗譜二十卷（浙江省臺州市黄巖區）

陈

chen

见于《世本》。西汉《急就章》列为汉代常见姓氏之一。春秋时陈国有陈亢，为孔门弟子。春秋时齐国有陈乞，事景公为大夫。又战国时楚国有陈学良，学者。汉代有陈平，阳武人，开国元勋。陈氏为中国古今最常见的六大姓氏（王李张刘陈赵）之一。《中国人名大辞典》收有陈氏1012例。宋《百家姓》列为第010姓。

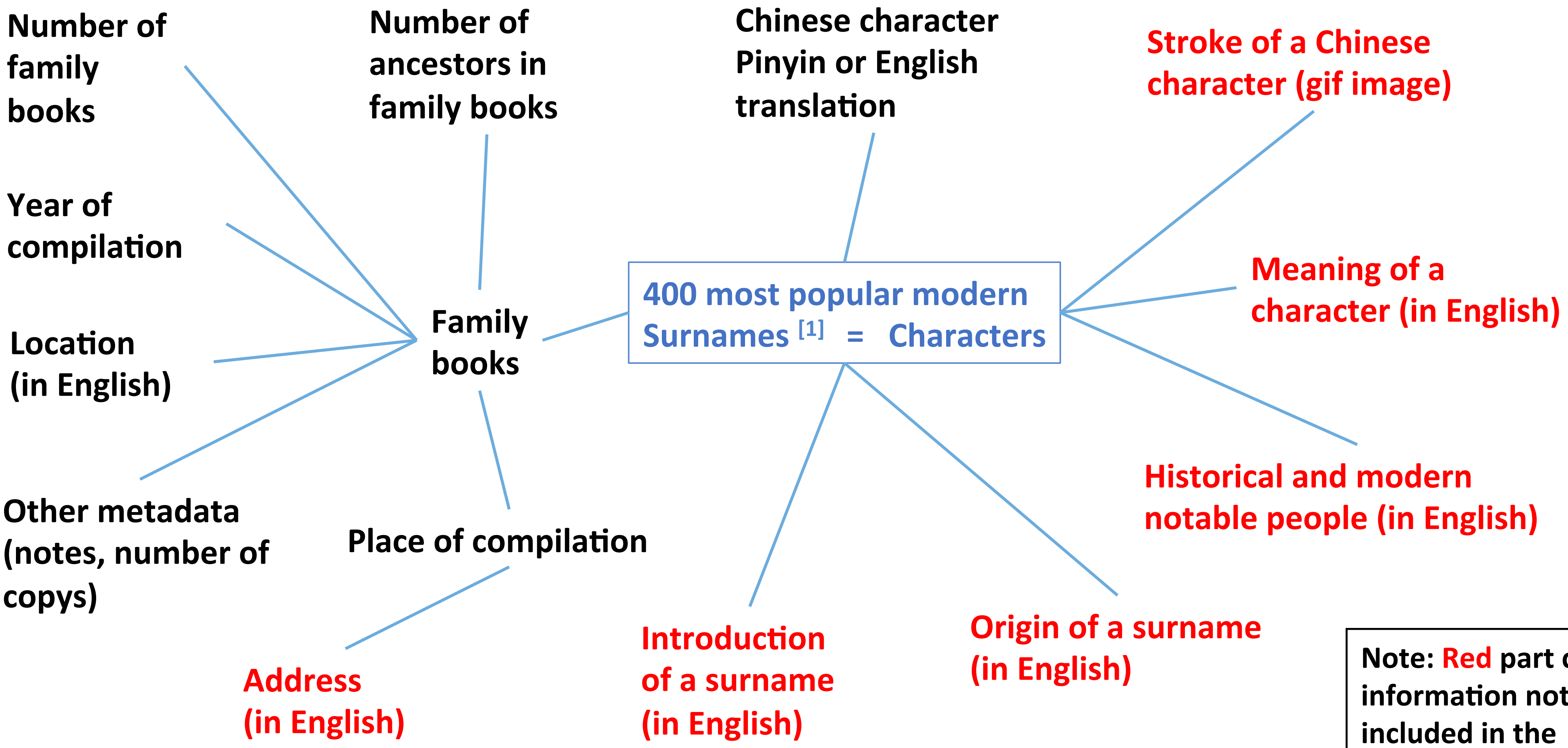
陳高元 始祖

⊕ 黄邑錢山陳氏宗譜 黄岩

- 摘要：始祖高元，號遼菴，元時遷居黄巖錢山。有卷載序、仕宦科第名錄、世系等。存卷一。七修本。
- 撰修时间：清宣統元年（1909）
- 版本：木活字本
- 数量：一册
- 馆藏信息

Shanghai Library Genealogical Knowledge Service Platform Beta. http://jp.library.sh.cn/jp/home/index

Original vs. Enriched Information



Number of family books

Number of ancestors in family books

Year of compilation

Location (in English)

Other metadata (notes, number of copys)

Family books

Place of compilation

Address (in English)

400 most popular modern Surnames [1] = Characters

Chinese character Pinyin or English translation

Stroke of a Chinese character (gif image)

Meaning of a character (in English)

Historical and modern notable people (in English)

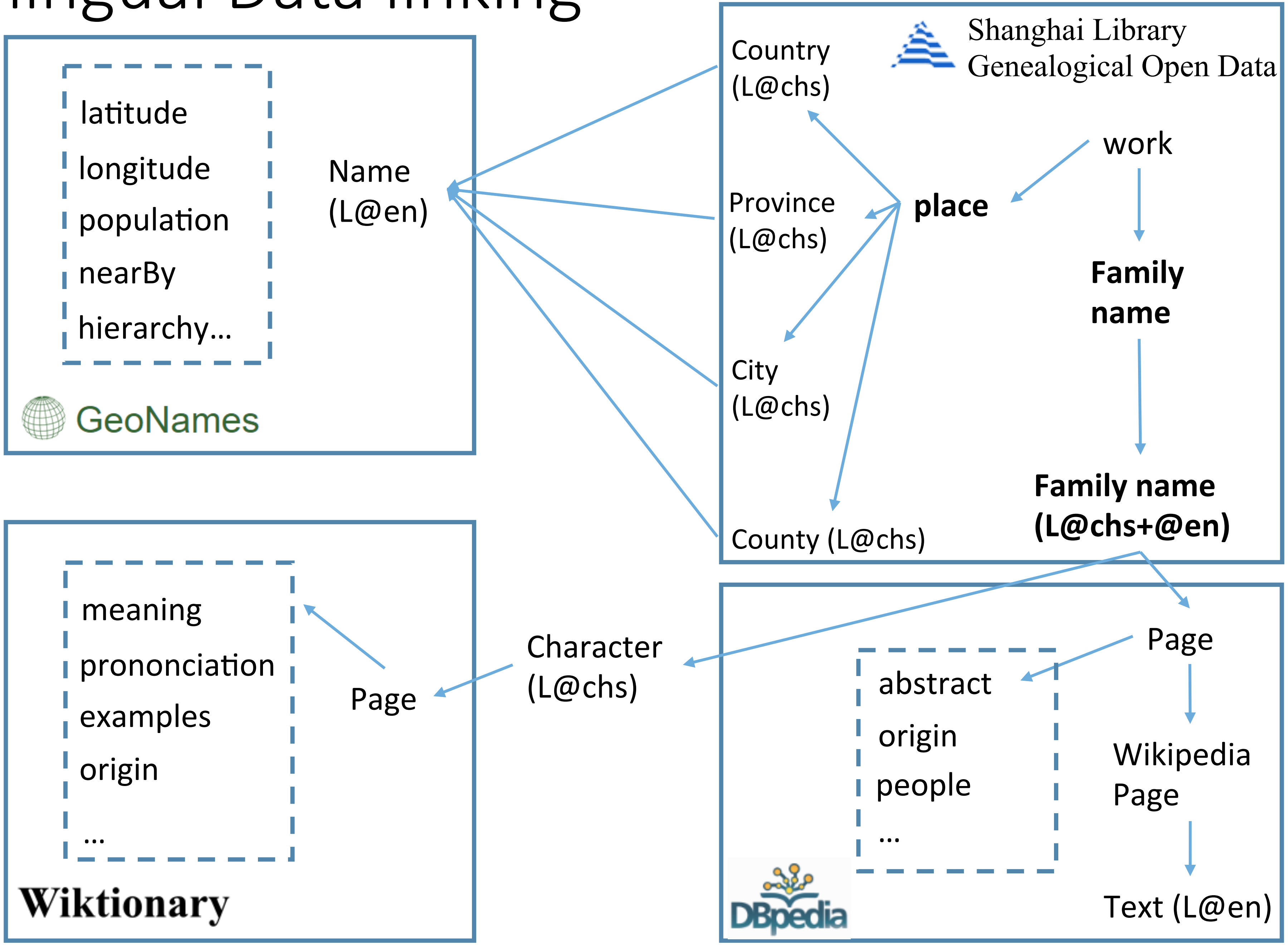
Origin of a surname (in English)

Introduction of a surname (in English)

Note: Red part contains information not included in the Shanghai Lib Open Data.

[1] Yida, Yuan & Jiaru, Qiu. (2013). Zhong guo si bai da xing. Nan chang: Jiang xi ren min chu ban she.

Cross-lingual Data linking



latitude
longitude
population
nearBy
hierarchy…

Name
(L@en)

GeoNames

Shanghai Library
Genealogical Open Data

Country
(L@chs)

work

Province
(L@chs)

place

Family
name

City
(L@chs)

Family name
(L@chs+@en)

County (L@chs)

meaning
prononciation
examples
origin
…

Page

Wiktionary

Character
(L@chs)

Page

abstract
origin
people
…

Wikipedia
Page

Text (L@en)

DBpedia

Notes:

1) “L” means RDF Literal
2) “@” with language tag: @chs Chinese Simplified; @en English
3) Fields that are not marked with L are RDF URI References

Consuming linked data for enriching Cross-lingual data

- 1 Enriching surname culture information – Querying RDF using SPARQL
- Using DBpedia SPARQL Endpoint
- Using DBpedia Data Dump
- Human filtering

- 2 Enriching character related information – Through Wiktionary URLs
- Wiktionary API (access limit)
- Wiktionary SPARQL Endpoint (access limit)
- Construct URLs

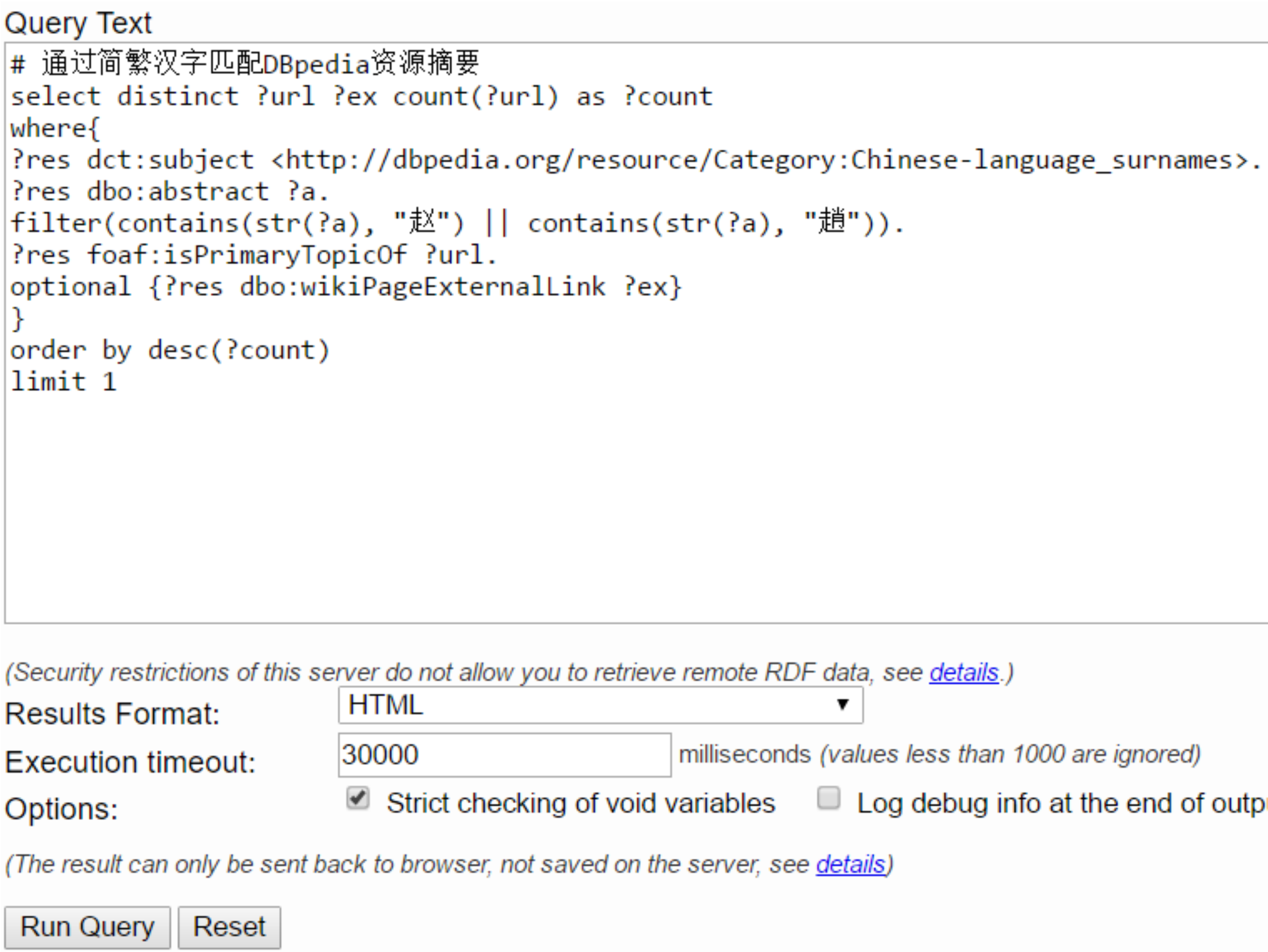
- 3 Matching Genealogical Names to GeoNames – Using API interface
- GeoNames API

Enriching surname culture information – Querying RDF using SPARQL

- Dbpedia的SPARQL Endpoint: https://dbpedia.org/sparql

- Data dump available

- Human filtering



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)

Matching Genealogical Names to GeoNames – Using API interface

- API for searching GeoNames：http://www.geonames.org/export/geonames-search.html

- Obtain GeoNames data of Hu Xian of City Xi’an in Province Shanxi in China, which is the place where the family book “Duan Shi Shi Xi” is compiled on 1731.

- http://api.geonames.org/searchJSON?name_equals=%E6%88%B7%E5%8E%BF&featureCode=ADM3&country=CN&maxRows=10&username=XXX

  (replace XXX to your GeoNames username)

返回结果:

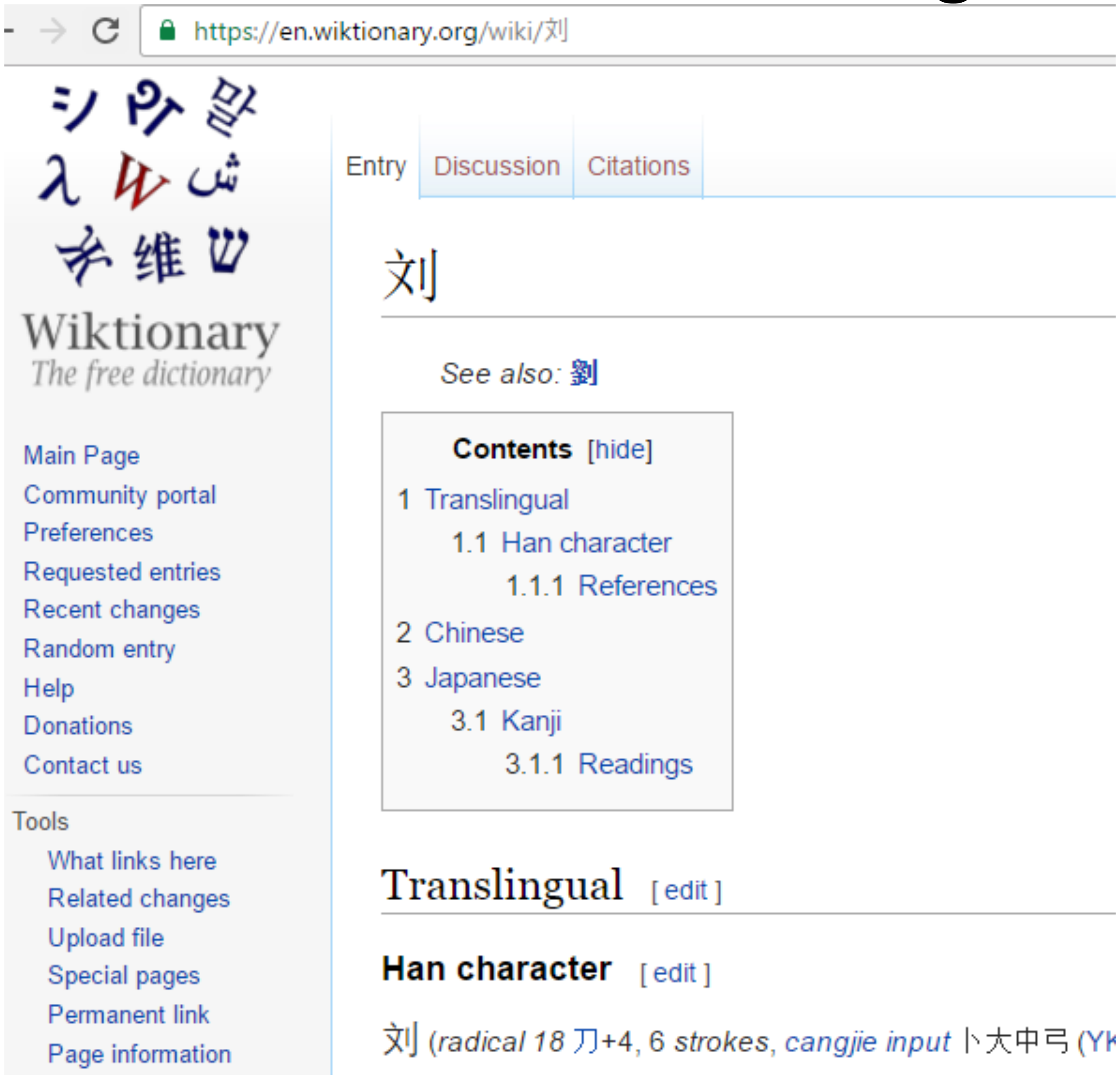
```
{
  "totalResultsCount": 1,
  "geonames": [
    {
      "adminCode1": "26",
      "lng": "108.58764",
      "geonameId": 1806562,
      "toponymName": "Hu Xian",
      "countryId": "1814991",
      "fcl": "A",
      "population": 556377,
      "countryCode": "CN",
      "name": "Hu Xian",
      "fclName":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

Enriching character related information – Through Wiktionary URLs

- Construct a corresponding URL

from a Chinese Character

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



https://en.wiktionary.org/wiki/刘

刘

See also: 劉

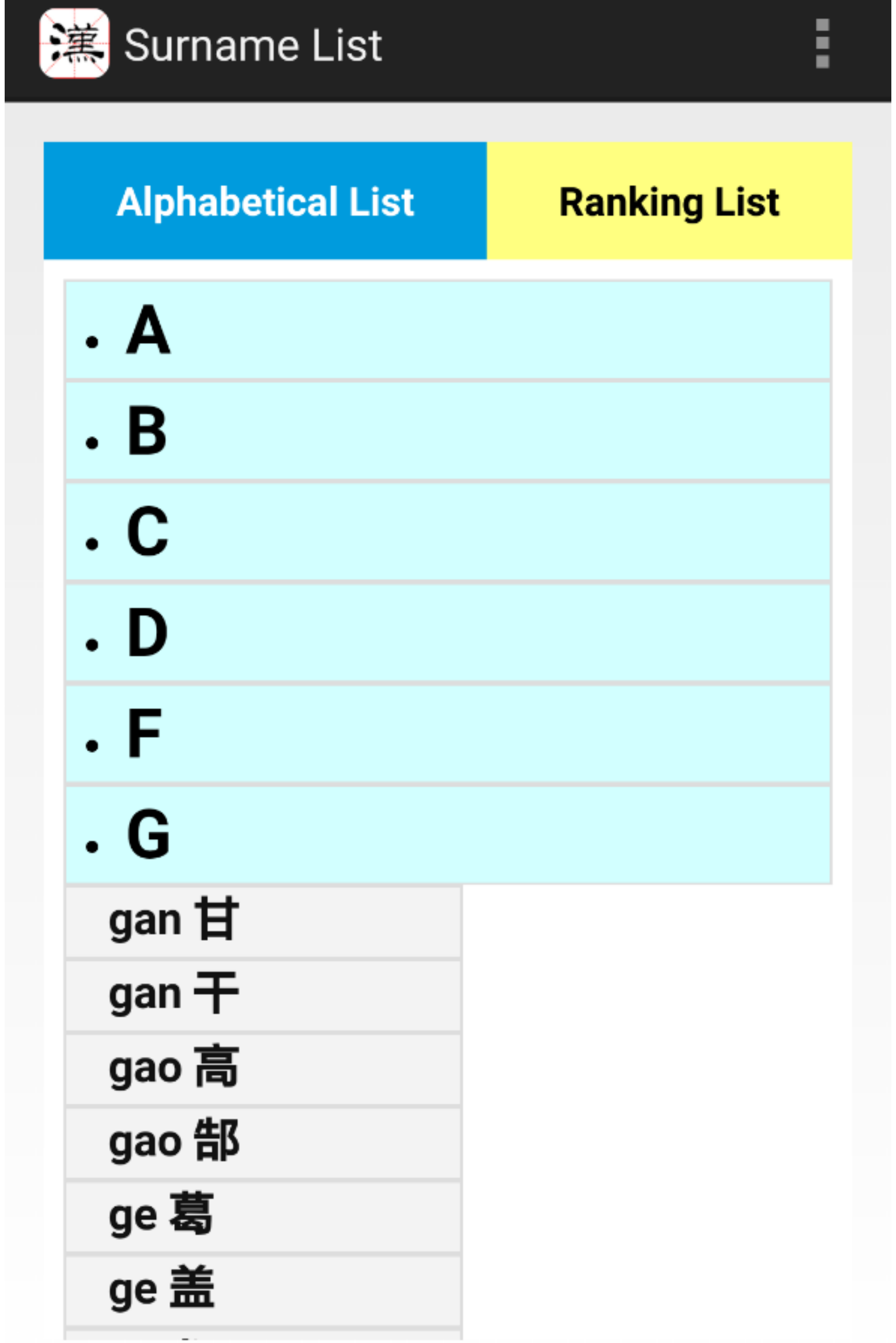
Contents [hide]

1 Translingual
1.1 Han character
1.1.1 References
2 Chinese
3 Japanese
3.1 Kanji
3.1.1 Readings

Translingual [edit]

Han character [edit]

刘 (*radical 18* 刀+4, 6 *strokes*, *cangjie input* 卜大中弓 (YK

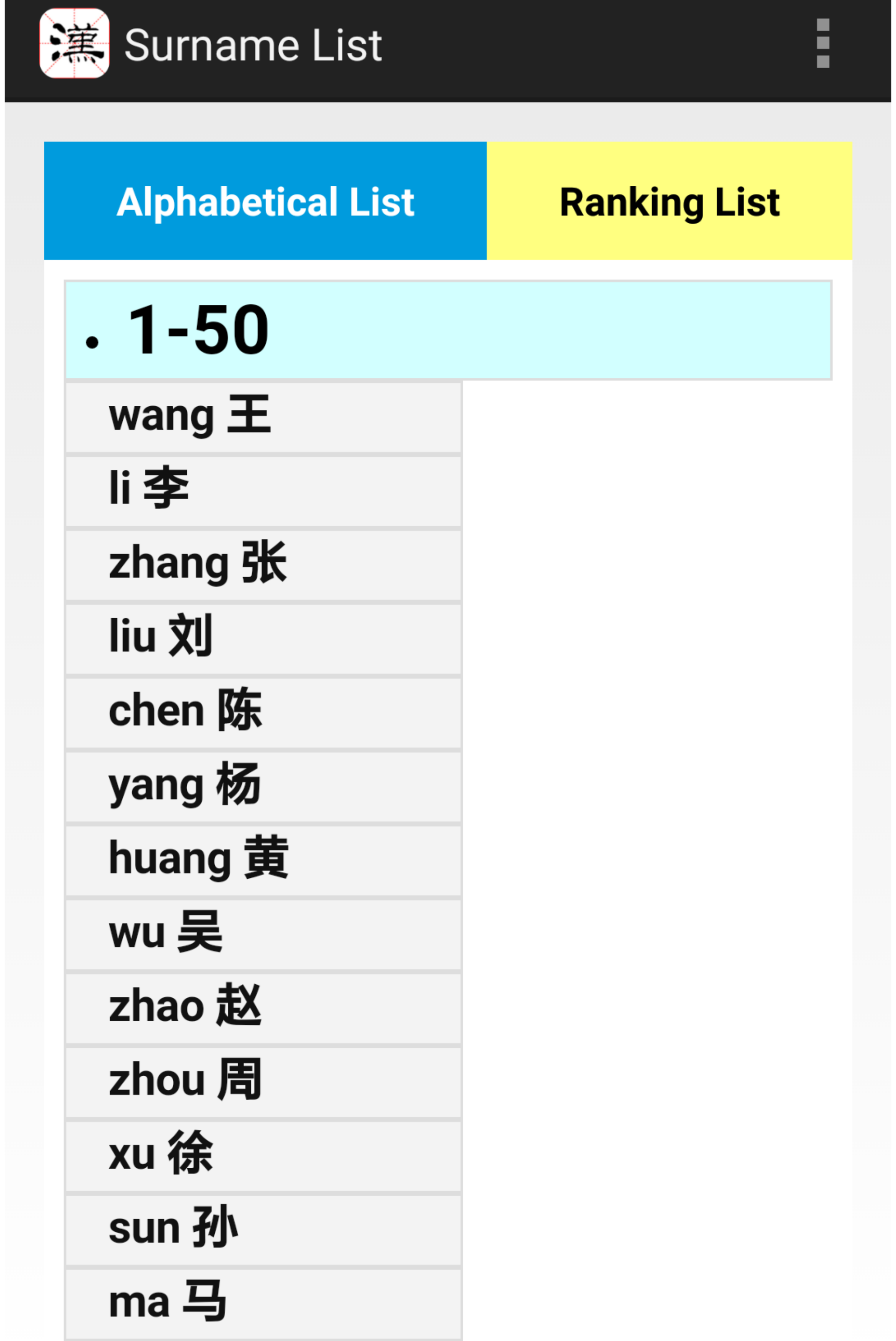


Surname List

Alphabetical List | Ranking List

- A
- B
- C
- D
- F
- G

gan 甘
gan 干
gao 高
gao 郜
ge 葛
ge 盖

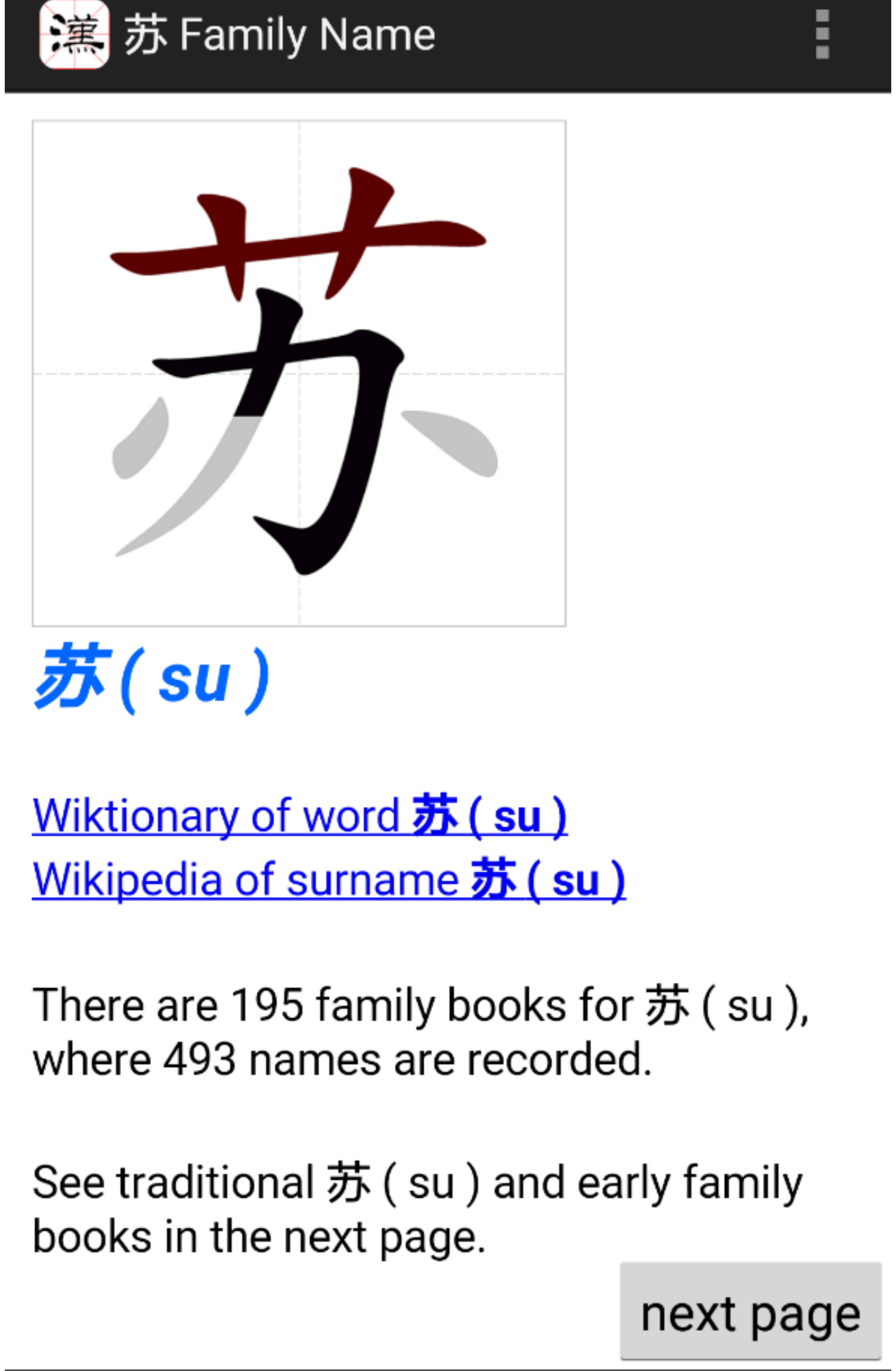


Surname List

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



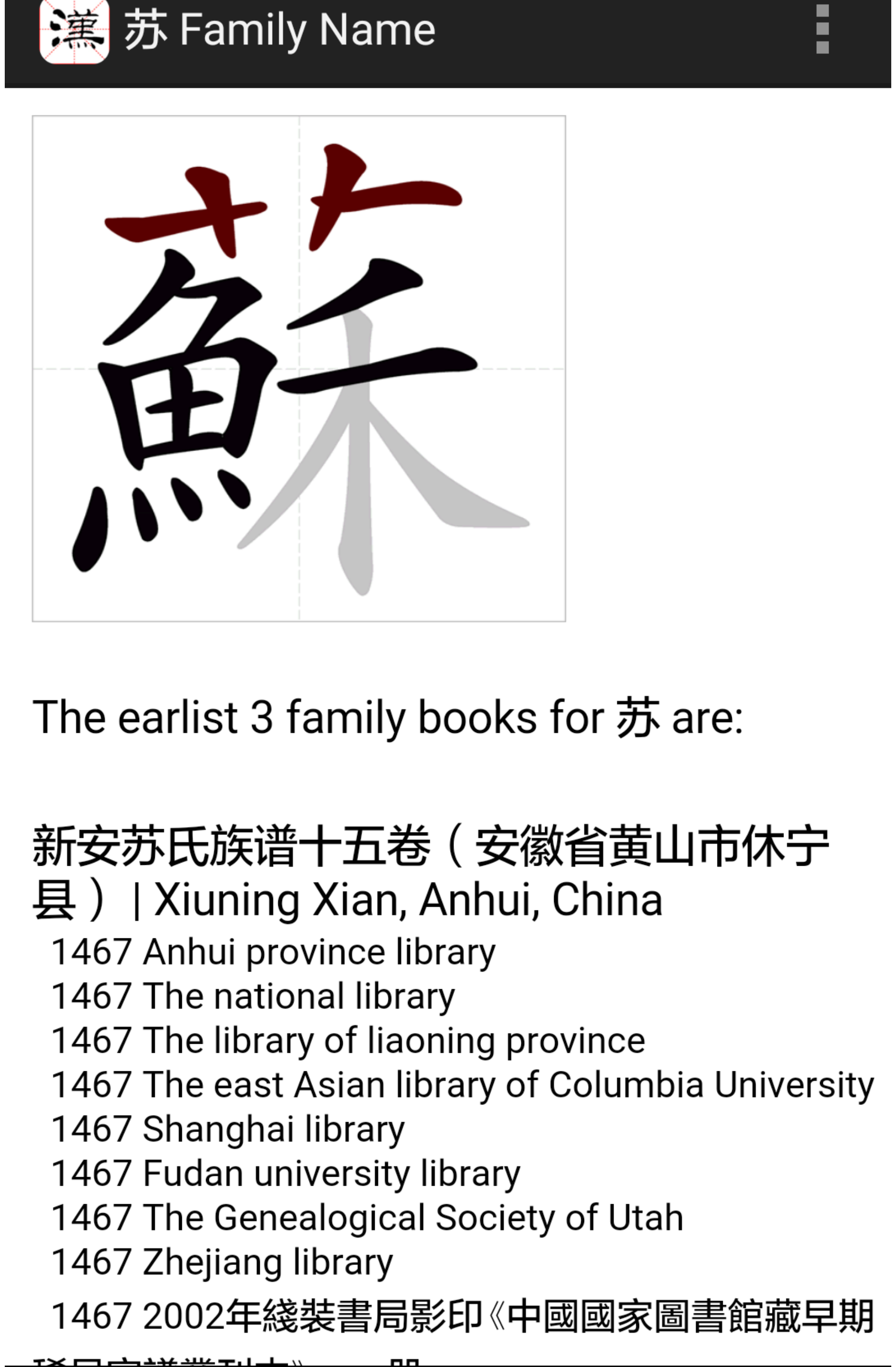
苏 Family Name

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.

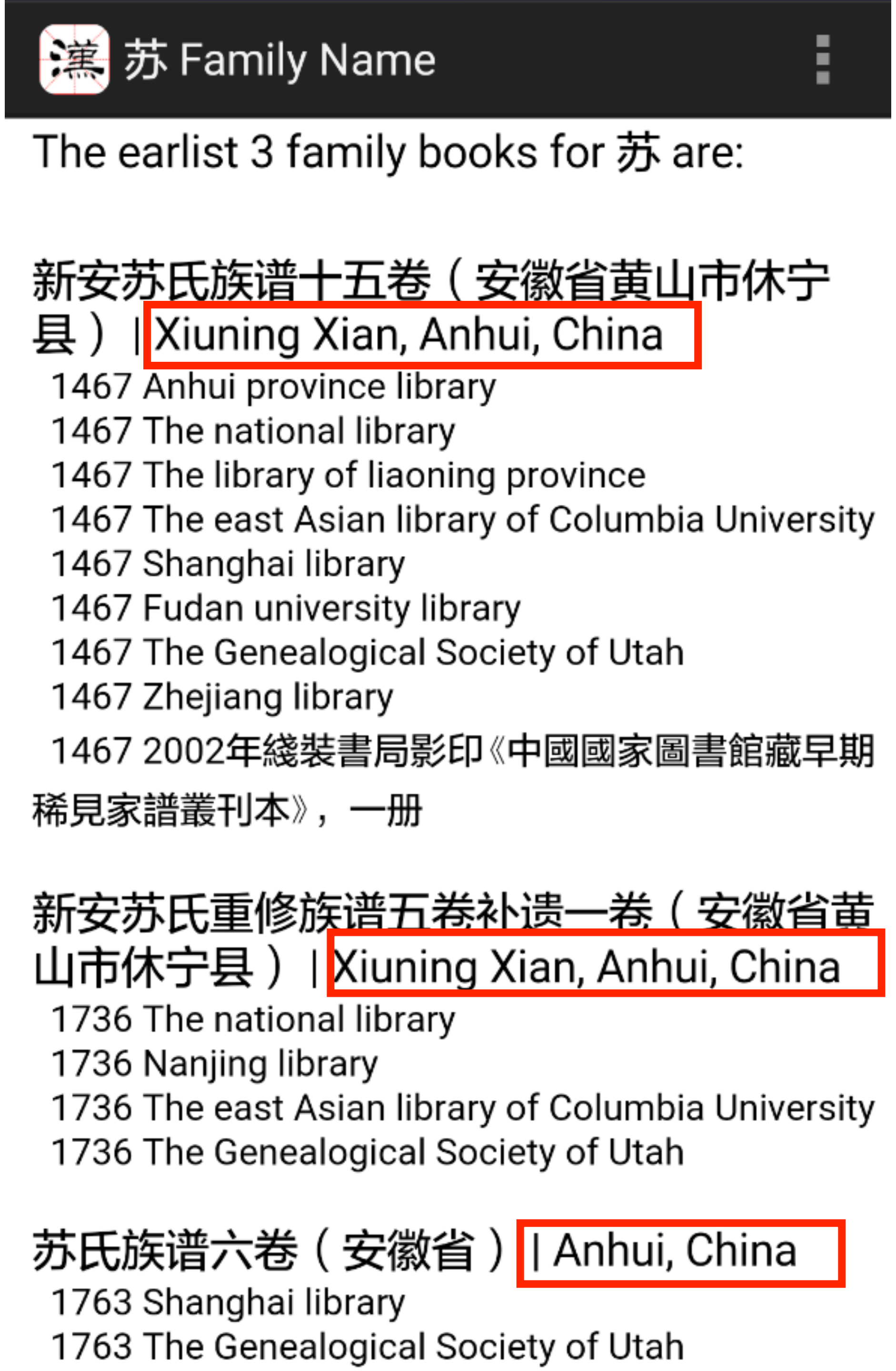


苏 Family Name

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

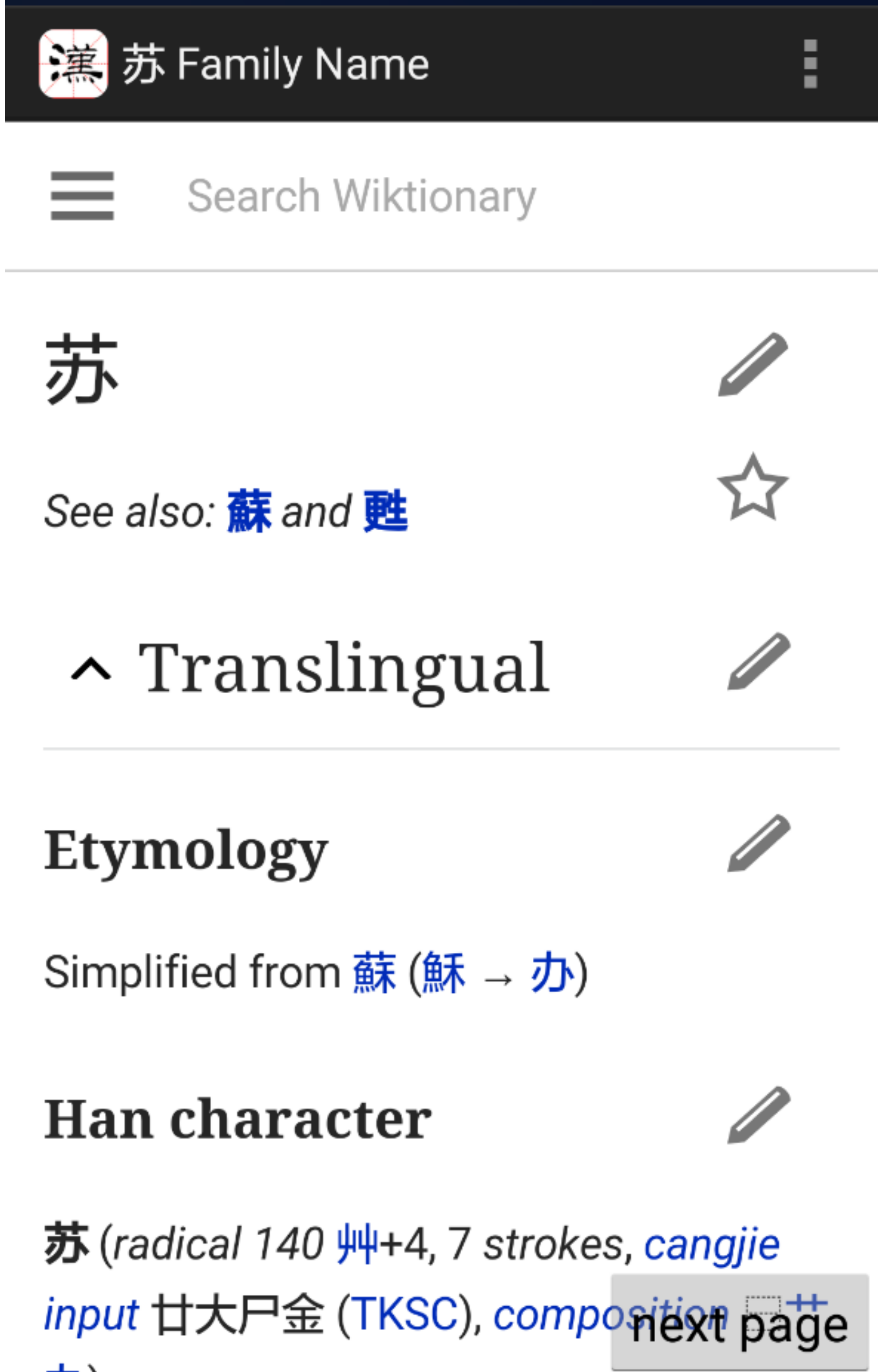
新安苏氏重修族谱五卷补遗一卷（安徽省黄山市休宁县）| Xiuning Xian, Anhui, China
1736 The national library
1736 Nanjing library
1736 The east Asian library of Columbia University
1736 The Genealogical Society of Utah

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

Display of levels of geographical data in a mobile App.

Matching GeoNames data to the genealogy compilation places.

The English part (marked by red boxes) are the enrichment information after data matching.



苏

See also: **蘇** *and* **甦**

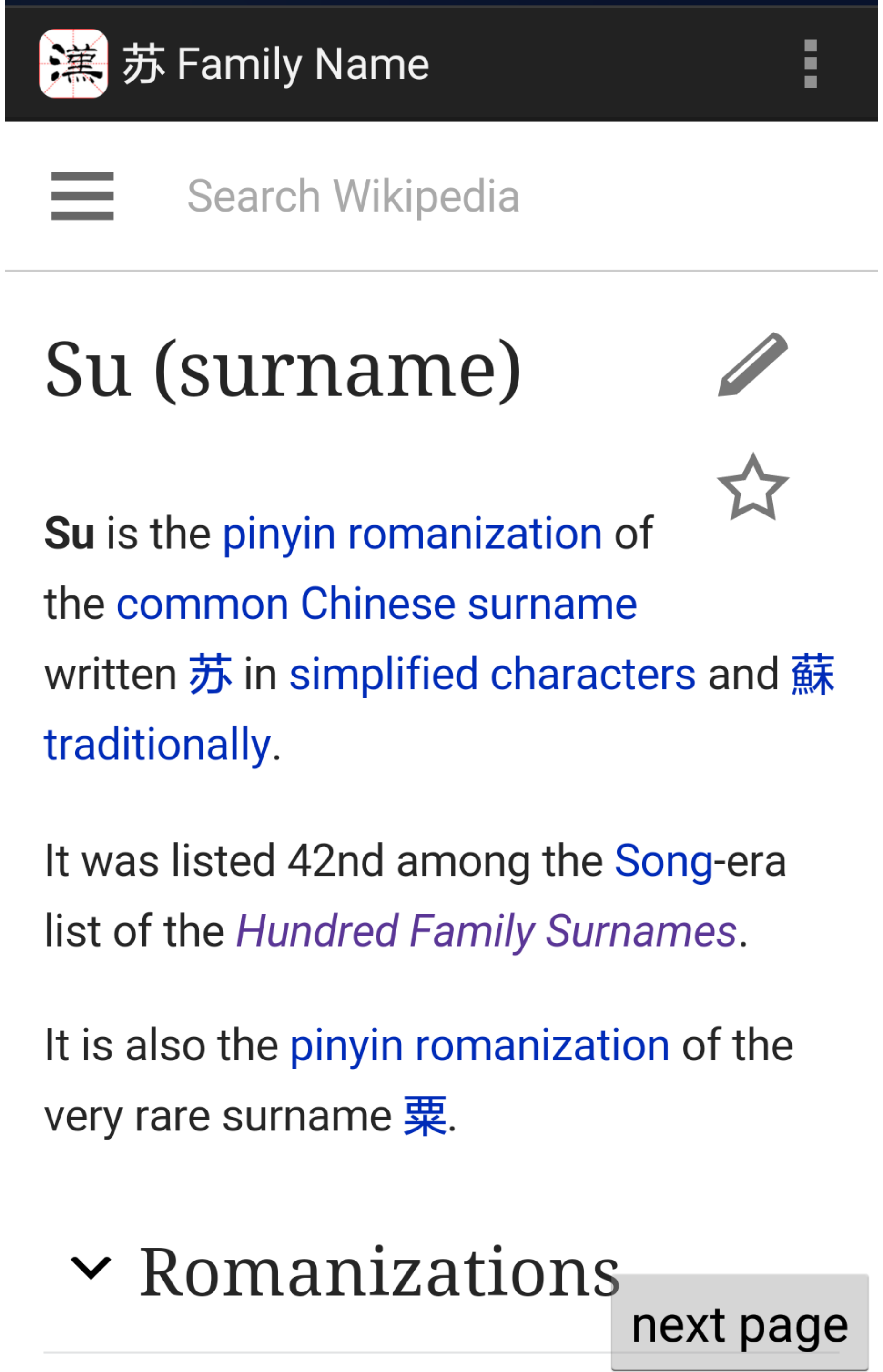
Translingual

Etymology

Simplified from 蘇 (穌 → 办)

Han character

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



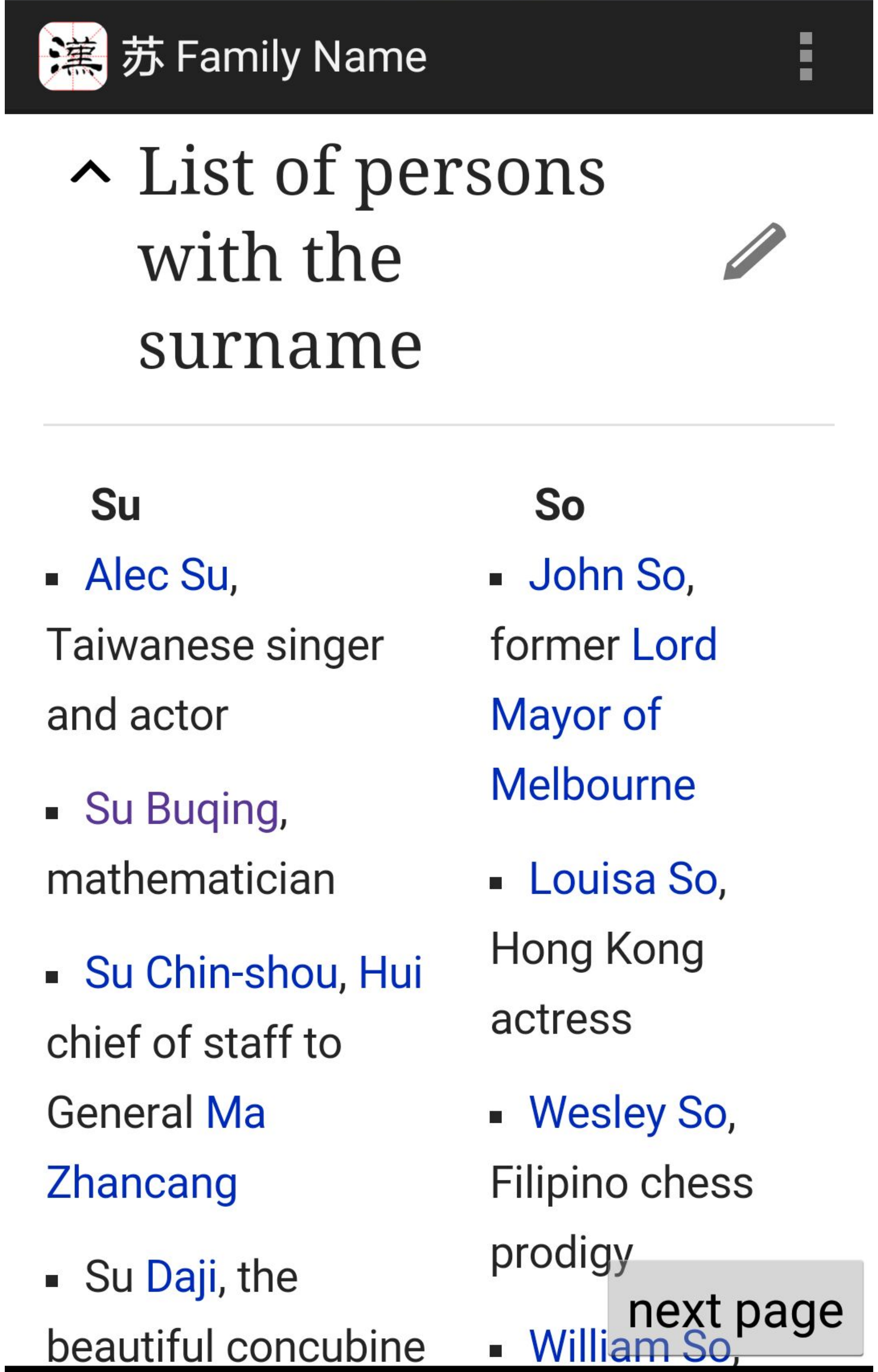
Su (surname)

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



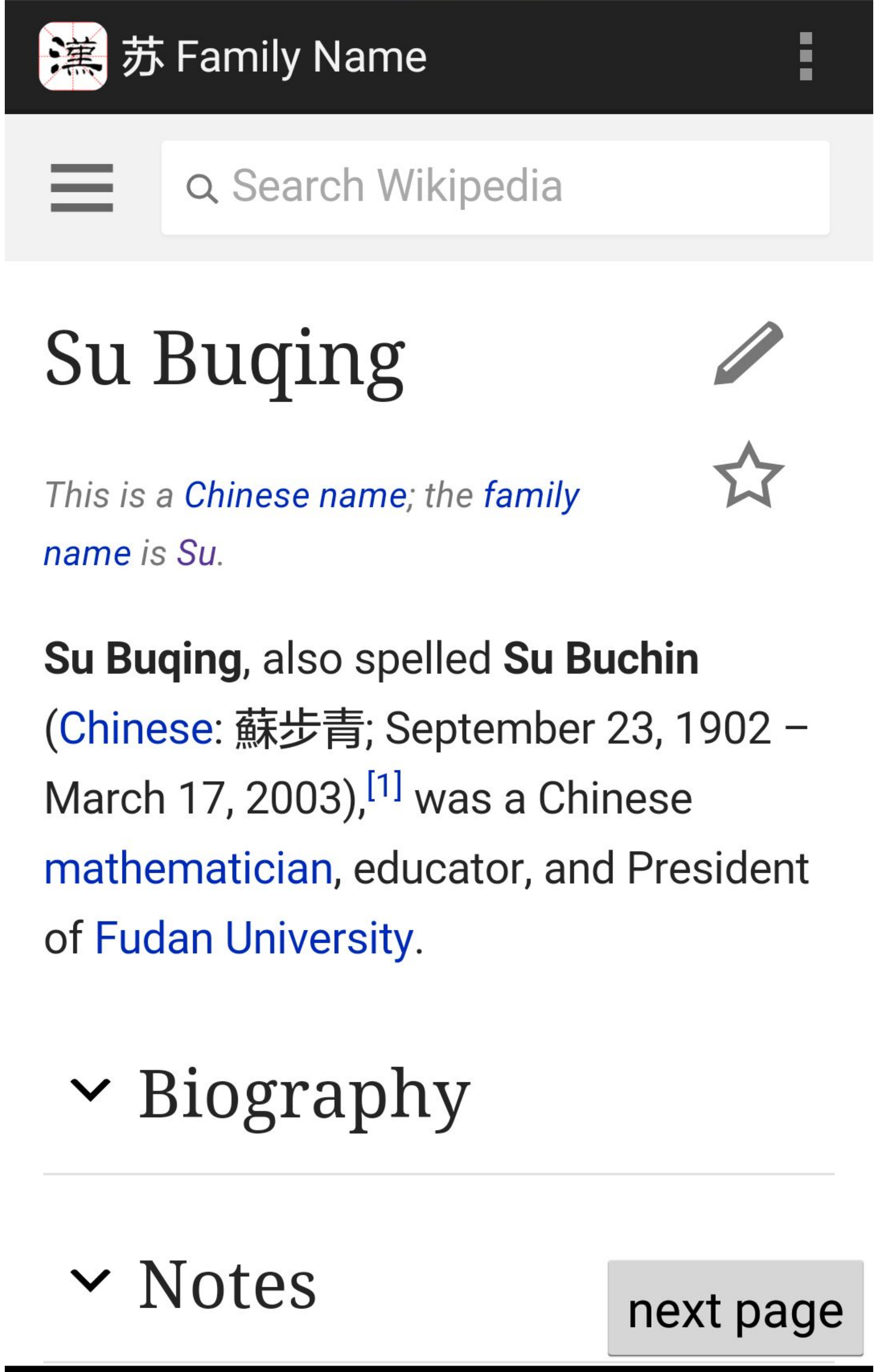
List of persons with the surname

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,



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.

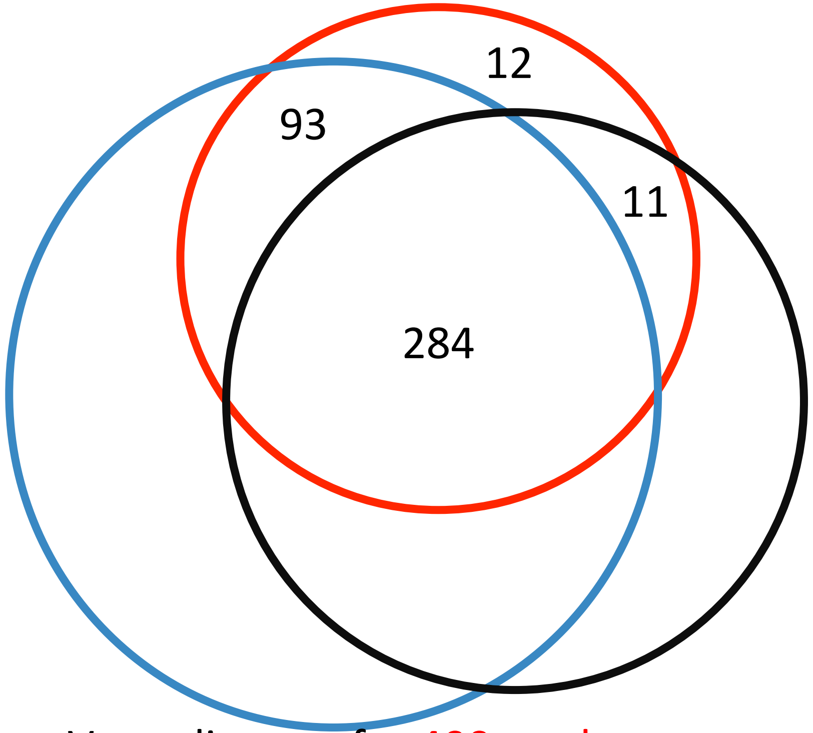
Biography

Notes

Data Validation and Statistics

Surname	Pinyin (Shanghai Lib)	English Name in Wikipedia	Wikipedia URL	Notes
房	fang	pang	http://en.wikipedia.org/wiki/Pang_(surname)	Both applicable in different situations
柏	bai	bo	http://en.wikipedia.org/wiki/Bo_(Chinese_surname)	Should be “bo”
区	qu	ou	http://en.wikipedia.org/wiki/Ou_(surname)	Should be “ou”
强	qiang	jiang	http://en.wikipedia.org/wiki/Jiang_(surname)	Should be “jiang”
危	wei	ngai	http://en.wikipedia.org/wiki/Ngai_(surname)	Both application: “ngai” is Cantonese

Data Conflicts between Shanghai Library Dataset and Wikipedia



Venn diagram for 400 modern surnames, Shanghai Lib dataset (608 surnames) and English Wikipedia

Data comprehensiveness

Conclusion & Future Studies

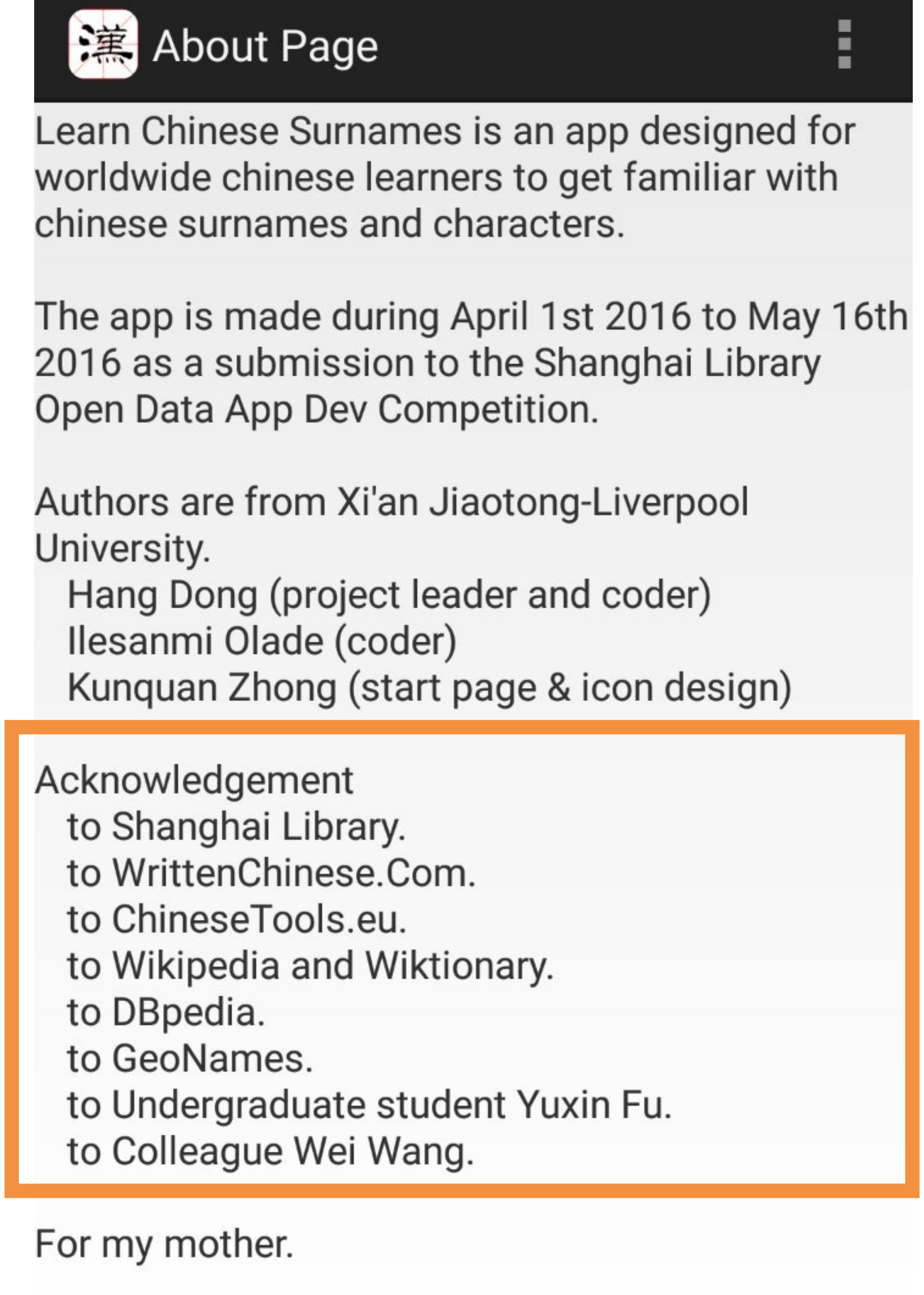
- Conclusion: Through consuming the LOD data, it is viable to enrich cross-lingual information for cultural heritage data.

- Deeper information extraction and mapping of C-L Data
  - *extract the unstructured information from Wikipedia, e.g. notable people linked to a surname, and match them with the people in Family Books?*

- Representation of enriched C-L Data
  - *represent the enriched multi-lingual information as RDF triples accurately to the Shanghai Library Genealogical Data?*

References

- Genealogical Ontology Structure based on BibFrame by Shanghai Library http://gen.library.sh.cn:8080/ontology/view
- Asunción, Gómez-Pérez, . . . Aguado-de-Cea, G. (2013). *Guidelines for multilingual linked data*. Paper presented at the Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics, Madrid, Spain.
- Damova, M., Dannélls, D., Enache, R., Mateva, M., & Ranta, A. (2014). Multilingual natural language interaction with semantic web knowledge bases and linked open data. In P. Buitelaar & P. Cimiano (Eds.), *Towards the multilingual semantic web: Principles, methods and applications* (pp. 211–226). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gómez-Pérez, A., Buitelaar, P., & McCrae, J. (2012). Challenges for the multilingual web of data. Web Semantics: Science, Services and Agents on the World Wide Web, 11, 63-71.
- Matusiak, K. K., Meng, L., Barczyk, E., & Shih, C.-J. (2015). Multilingual metadata for cultural heritage materials: The case of the tse-tsung chow collection of chinese scrolls and fan paintings. *The Electronic Library*, *33*(1), 136-151.
- Ngai, G., Carpuat, M., & Fung, P. (2002). Identifying concepts across languages: A first step towards a corpus-based approach to automatic ontology alignment. In *Proceedings of the 19th international conference on computational linguistics - volume 1* (pp. 1–7). Stroudsburg, PA, USA: Association for Computational Linguistics.
- Stiller, J., Petras, V., Gäde, M., & Isaac, A. (2014). Automatic Enrichments with Controlled Vocabularies in Europeana: Challenges and Consequences. In M. Ioannides, N. Magnenat-Thalmann, E. Fink, R. Žarnić, A.-Y. Yen, & E. Quak (Eds.), *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection: 5th International Conference, EuroMed 2014, Limassol, Cyprus, November 3-8, 2014. Proceedings* (pp. 238-247). Cham: Springer International Publishing.
- Wang, Z., Li, J., Wang, Z., Li, S., Li, M., Zhang, D., . . . Tang, J. (2013). Xlore: A large-scale english-chinese bilingual knowledge graph. In *Proceedings of the 2013th international conference on posters &#38; demonstrations track - volume 1035* (pp. 121–124). Aachen, Germany, Germany: CEUR-WS.org.



About Page

Learn Chinese Surnames is an app designed for worldwide chinese learners to get familiar with chinese surnames and characters.

The app is made during April 1st 2016 to May 16th 2016 as a submission to the Shanghai Library Open Data App Dev Competition.

Authors are from Xi'an Jiaotong-Liverpool University.
  Hang Dong (project leader and coder)
  Ilesanmi Olade (coder)
  Kunquan Zhong (start page & icon design)

Acknowledgement
  to Shanghai Library.
  to WrittenChinese.Com.
  to ChineseTools.eu.
  to Wikipedia and Wiktionary.
  to DBpedia.
  to GeoNames.
  to Undergraduate student Yuxin Fu.
  to Colleague Wei Wang.

For my mother.

Thank you for your attention

Acknowledgement to my teammates
Ilesanmi Olade, Kunquan Zhong;
and to Wei Liu, Cuijuan Xia from Shanghai Library.

Hang Dong | 董行
hangdong@liverpool.ac.uk

Android App file download:
http://pcrc.library.sh.cn/zt/opendata/apk/Learn%20Chinese%20Surnames%20.apk

Shanghai Library Open Data Comp 2017.
http://pcrc.library.sh.cn/zt/opendata/2017/ [page in Chinese]