Enrichment of Cross-Lingual Information on Chinese Genealogical Linked Data

Hang Dong[1,2]
[1]University of Liverpool
[2]Xi'an Jiaotong-Liverpool University

Abstract
With the emergence of non-English Linked Datasets, discrepancy in language has become a major obstacle for cross-lingual access of resources in the Semantic Web. To prevent non-English monolingual Linked Datasets to form 'islands' in the Web of Data, it is suggested to enrich a further layer of multilingual information on the Linked Open Data cloud. In the domain of culture heritage, enriching cross-lingual information can enhance the multilingual retrieval of cultural heritage resources, and promote international communication in the field. In this article, methods to enrich cross-lingual information for Linked Data are summarized, with a review on the cultural heritage domain. The mobile App Demo, Learn Chinese Surnames, winning the Shanghai Library Open Data Application Development Contest on 2016, is then introduced as a case study, to present the practice of enriching English-described information on a Chinese genealogical Linked Dataset, through consuming multilingual sources in the Linked Open Data cloud. Further in the data validation and conclusion, the issues of data quality and experience of consuming Linked Data are summarized.

Keywords: Linked Data; Cross-lingual information; Chinese genealogies; Open data applications; Consuming Linked Data
Citation: Dong, H. (2017). Enrichment of Cross-Lingual Information on Chinese Genealogical Linked Data. In iConference Proceedings, Vol. 2 (pp. 31-42). https://doi.org/10.9776/17005

Acknowledgements: The author would like to thank his teammate Ilesanmi Olade from University of Liverpool for his contribution on implementing the user interface, and another teammate Kunquan Zhong from Xi'an Jiaotong-Liverpool University for his design of the App cover page. Thanks Wei Liu and Cuijuan Xia from Shanghai Library for their support and guidance on this work and research. Also acknowledgement to Shanghai Library for organizing this Open Data Application Development Contest.

Contact: hangdong@liverpool.ac.uk

1 Introduction

Since the last decade, there has been an enormous growth in the amount of information on the Semantic Web, marked by the initiative of Linked Data. The Linked Open Data cloud (LOD cloud) has increased from a dozens of dataset to a large data space containing a thousand of datasets today (Schmachtenberg, Bizer, & Paulheim, 2014), forming a Web of Data. Linked Data refers to a set of best practices to publish and connect data, organized by RDF triples, via de-referenceable URIs on the Semantic Web (Bizer, Heath, & Berners-Lee, 2009). The vast Linked Datasets cover domains in the media, government, academics, user-generated content (Schmachtenberg et al., 2014). Linked Data initiatives and projects have also begun in the libraries (Singer, 2010) and in the cultural heritage domain (Oomen, Baltussen, & Van Erp, 2012).

Although the majority of datasets in LOD cloud are described in English, leading to a language-biased Semantic web, it is witnessed that many non-English Linked datasets have been published. For example, Dogmazic[1], the French open music vocabularies, GeoLinkedData.es[2], the open initiative of creating Spanish geospatial dataset, and the Shanghai Library Genealogical Linked dataset[3], etc. These non-English datasets are rather valuable since they represent the diverse culture dependent on the languages and geographical areas. It is however that, with the issue of language discrepancies and cross-lingual access, these monolingual non-English Linked datasets could form "islands" on the Semantic Web (Gracia et al., 2012, p. 64). Similar

[1]http://www.dogmazic.net/
[2]https://datahub.io/dataset/geolinkeddata
[3]RDF structure in http://gen.library.sh.cn:8080/ontology/view, data content in http://jp.library.sh.cn/jp/home/index

issues about multi-lingual and multi-cultural are also the main challenges for cultural heritage data (Hyvonen, 2012, p. 5).

Currently, in the cultural heritage domain, multilingual Linked Data projects are still rare. It is believed in this article that, through the enrichment of cross-lingual information on current linked dataset in the cultural heritage domain, it is possible to enhance the multilingual access of cultural heritage information, boost international communication on researching these cultural heritage resources, and develop more international public data services.

Based on the principles of Linked Data and the ontologies adapted from BIBFRAME, Shanghai Library transformed and described the family books from MARC and those recorded in the *General Catalogue of Chinese Genealogy*[4] to RDF triples (Xia, Liu, Zhang, & Zhu, 2014). The new genealogy data service platform of Shanghai Library supports multiple interfaces to consume data, including a web-based portal, a SPARQL endpoint and JSON-LD (Xia, Liu, Chen, & Zhang, 2016). In 2016, Shanghai Library held an Open Data Application Development Contest[5] lasting about two months' time till the end of May, which is the first of this kind organized by Chinese libraries. A case study, an Android App Demo named Learn Chinese Surnames, is introduced to provide experience on the enrichment of information described in English from Linked Data sources, on the Chinese Genealogical dataset.

In this case study, the designer presents the methods to match entities in three famous Linked Dataset, DBpedia, Wiktionary and GeoNames, to the Genealogical dataset from Shanghai Library, to add a new layer of English-described information for cross-lingual access of Chinese genealogical information. This not only can enhance the international cooperation on research of Chinese genealogies, but also can benefit a wider user group of data services from native users to international users. From the Semantic Web point of view, the work constitutes an effort to realize the Multilingual Web of Data (Gracia et al., 2012).

In Section 2, the idea of enriching cross-lingual information for Linked Data is explained and related projects in the cultural heritage domain are reviewed. In Section 3, the Android App Demo, Learn Chinese Surnames, is presented as a case study to share the practice of consuming LOD to enrich English-described information, followed by a discussion on data quality and validation of Linked Data sources. Conclusion and future studies are in the Section 4.

2 Research Background

2.0.1 Enrichment of Cross-Lingual Information for Linked Data

The idea of enrichment of cross-lingual information for Linked Data is originally from the "multilingual Web of Data", coined by Gracia et al. (2012). The "multilingual Web of Data" has been proposed as an enhancement of the current Linked Open Data cloud with an extra layer of resources and services that boost and improve the internal linking and multilingual access of the Web of Data. Potential resources and services are listed below (Gracia et al., 2012).

1. multilingual linguistic information to describe resources in different natural languages;
2. multilingual mapping between ontologies/vocabularies that establish cross-lingual connections and between entities/instances in different natural languages;
3. (semi-)automatic services for accessing and traversing Linked Data across languages dynamically, such as multilingual Linked Data generation, ontology localization and translation, automatic cross-lingual matching.

In this proposed article, adding the first two resources (1 and 2), multilingual linguistic information and multilingual entity mapping or linking, is summarized as the task of enrichment of cross-lingual information on Linked Data. The reason for this definition is a division of micro-level and macro-level: resources 1 and 2 are on a micro-level, thus easy to be implemented in individual applications, while services such as general tools for ontology localization are on a macro-level.

[4]Wang, H., & Shanghai tu shu guan. (2008). Zhongguo jia pu zong mu. Shanghai: Shanghai gu ji chu ban she([Chinese character]). WorldCat Catalog: `http://www.worldcat.org/title/zhongguo-jia-pu-zong-mu/oclc/314020020`

[5]`http://pcrc.library.sh.cn/zt/opendata/`

It is also necessary to distinguish the level of Linked Data to enrich: on the entity level ("instance level"), and on the ontology level (vocabulary level or "conceptual level") (Gracia et al., 2012, p. 68). A group of studies in cross-lingual ontology matching deals with the mapping of ontological vocabularies in different languages (Trojahn, Fu, Zamazal, & Ritze, 2014; Mejía, Montiel-Ponsoda, de Cea, & Gómez-Pérez, 2012). Besides, it is possible to add multilingual information on the entity/instance level, for example, state that the page "Liu"[6] in the DBpedia is the same as "[chinese character]" and "[chinese character]" in the Genealogical Linked Dataset from Shanghai Library, regardless of the ontological structure used to organize these entities in each Linked Dataset.

While the two levels of data enrichment are different, they share some common types of methods (methods on the ontology level in Trojahn et al. (2014), methods on the entity level in examples from Section 2.2): First, through translation by human or by machine to derive a new vocabulary or an entity description; second, through matching or linking of vocabularies or entities in two data sources; third, through a crowdsourcing interface enabling users to add multilingual information to the Linked Dataset.

All methods have their advantages and disadvantages. Current machine translation techniques are still preparing themselves to be intelligent enough to carry the intent behind languages and cultures, while manual translation needs much human effort. Through data matching based on semantic similarity, no translation effort is required, but the precision and recall of matching largely depend on the comprehensiveness of available semantic sources; also, without adding linguistic description such as synonym and hyponym, simple string matching can produce poor results (Pazienza & Stellato, 2006), so complex metrics to measure similarity or advanced methods based on machine learning have been used in large cross-lingual matching projects such as the alignment of English WordNet to Chinese HowNet (Ngai, Carpuat, & Fung, 2002), and the semantic linking of online collaborative encyclopedia, Chinese Baidu Baike[7] and English Wikipedia (Wang et al., 2013). Through crowdsourcing, represented by Wikipedia and DBpedia (Lehmann et al., 2015), it is possible to obtain more accurate cross-lingual information based on the actual meaning that reflects culture background of resources, but the participation of a group of expert users is required. Further examples of the three methods are given in the culture heritage domain.

2.1 Enrichment of Cross-Lingual Information in the Cultural Heritage Domain

In the cultural heritage domain, due to the variety of representations and the commonality of cultures in the world, it is necessary to integrate resources of multi-culture to satisfy people??s information needs; while at the same time, resources in different cultures are sometimes originally represented in different languages, without an enrichment of cross-lingual information it would be hard to link them to each other, making these resources inaccessible to users in other languages. Currently, there are several practices for adding cross-lingual vocabularies and descriptions to cultural heritage resources, but few of them have published a Linked Data version.

An example on the enrichment of cross-lingual information on the vocabulary level is in the Digital Archives Sub-Project of Antiquities in the National Palace Museum[8] under Taiwan e-Learning and Digital Archives Program (TELDAP)[9]. To satisfy the users?? needs to retrieve Chinese art resources using English, it is necessary to align the controlled vocabularies from the National Palace Museum (NPM) in Taiwan with the Art & Architecture Thesaurus (AAT) developed by the Getty Research Institute in US (Chen & Chen, 2012). The challenge of this task is the heterogeneity in the "conceptual structures" (the manner that a concept is involved in the hierarchical or associative relationships)(Chen & Chen, 2012, p. 285) of two controlled vocabularies, due to the discrepancy in language and culture. Pure manual effort was made to map the complex conceptual structures of different degrees of similarity in the two vocabularies. This mapping of two controlled vocabularies of different language and culture can enhance their interoperability, enable multilingual search, integration and sharing of cultural heritage resources among users in Chinese and English language background.

Instead of the enriching cross-lingual information on the vocabulary level, a US academic library has led a project on enriching cross-lingual information with a focus on the entity level. English-described

[6] http://dbpedia.org/page/Liu
[7] http://baike.baidu.com/
[8] http://www.npm.gov.tw/digital/index2_2_8_en.html
[9] http://teldap.tw/en/index.html

metadata to the Collection of Chinese Scrolls and Fan Paintings has been added to ensure their retrieval in English (Matusiak, Meng, Barczyk, & Shih, 2015). First, the project group adapted the Dublin Core template to define a metadata scheme that incorporates the unique features for Chinese art resources and the bilingual fields. Then, human translation was carried out to generate the English described metadata of fields such as Main Text, Other Text, Seal Content, Subject, Coverage, etc. Compared to machine translation, human translation is suggested being more appropriate to capture the special cultural and linguistic features for the Chinese art resources (Matusiak et al., 2015).

A multilingual digital heritage project based on Linked Data structure is MOLTO[10] (Damova, Dannélls, Enache, Mateva, & Ranta, 2014). The project enables using natural language in 15 European languages to search museum artifacts. This is realized by conversions from natural language to SPARQL and from RDF triples to natural language outputs in multiple languages. For the generation of multilingual descriptions of artifacts, manual translation was carried out for the vocabularies in the ontology and the entities or instances in the classes Material and Colour. Instances of the classes Painter and painting titles was, however, untranslated due to the unavailability of lexicons and long time for human translation. The translation of museum names was derived through mapping of the article names to Wikipedia, which achieved 90% correctness for the 5 major languages French, Italian, German, Russian and Spanish.

To give a background for the case study in Section 3, Chinese traditional family books are also an important cultural heritage resource for its special value on research in history and sinology. Enriching cross-lingual information to Chinese genealogical resources, can benefit their retrieval and access in an international environment. It is however that, the only available multilingual retrieval system for family books is FamilySearch.org, operated by The Church of Jesus Christ of Latter-day Saints, previously known as the Genealogical Society of Utah. FamilySearch is the biggest genealogy project in the world, and its digital Chinese Collection of Genealogies[11] can date back from the year of 1239, cooperated with Shanghai Library in China from 2012[12]. The general FamilySearch project uses a crowdsourcing interface with a friendly tutorial webpage[13] to encourage users from the world to participate in the indexing, such as tasks to recognize texts from an old family book image, of multilingual genealogy resources. Also, FamilySearch has a Research Wiki portal[14], enabling users to collaboratively edit articles about genealogies in different countries and cultures.

To sum up, in the domain of cultural heritage, due to the relatively narrow vocabularies in the ontology, and the lack of mature, easy-to-use techniques to translate cultural-dependent texts, researchers tend to enrichment cross-lingual information manually, while for large projects, crowdsourcing methods are adopted. Multilingual Linked Data projects in the cultural heritage domain are still rare, although it is believed that they can enhance the interoperability, integration and searching of cultural heritage resources for a wider community (Hyvonen, 2012, p. 8).

3 Case study: Learn Chinese Surnames

In this study, a method is presented to enrich cross-lingual information through consuming multilingual resources in the Linked Open Data Cloud. A use case of mobile data services based on cultural heritage Linked Data published by a public library is provided. The work, an Android App Demo named Learn Chinese Surnames, winning the Shanghai Library Open Data Application Development Contest in 2016, is used as a case study.

3.1 Project Information and Data Use

Learn Chinese Surnames is an Android App designed for non-native Chinese learners to get familiar with Chinese surname culture and surname characters. The App made use of the 400 most popular modern Chinese surnames released in April 2013[15], and presented their origin, related people, metadata of early

[10] http://museum.ontotext.com/

[11] https://familysearch.org/search/collection/1787988

[12] https://familysearch.org/wiki/en/Shanghai_Library

[13] https://familysearch.org/indexing/

[14] https://familysearch.org/wiki/en/index.php

[15] https://zh.wikipedia.org/wiki/%E4%B8%AD%E5%9B%BD%E5%A7%93%E6%B0%8F%E6%8E%92%E5%90%8D#2013.E5.B9.B44.E6.9C.88 [Webpage in Chinese showing 400 surnames, originally from the book Yida, Yuan & Jiaru, Qiu. (2013). Zhong guo si bai da xing ([chinese character]). Nan chang: Jiang xi ren min chu ban she. WorldCat Catalog:

family books and information about Chinese characters, etc. Users can browse the 400 surnames by their alphabetic ranking or by popularity, learn the stroke and meaning of Chinese surname characters, the origin of surnames as well as the knowledge about family books.

The task of enriching cross-lingual information is required, since English descriptions related to Chinese characters and family books are necessary for the targeted non-native Chinese users of this App. The figure (Figure 1) listed all information demonstrated in the App. The open data from Shanghai Library already created some data described in English or as numbers, displayed in colour black, including Pinyin of surnames, metadata of family books, location where the book is stored, etc. The other information, which are acquired through consuming the external Linked Data sources or other online resources, are displayed in red, including the address where the family book is compiled, introduction, origin and notable people of a surname, meaning and stroke of Chinese surname characters, etc.

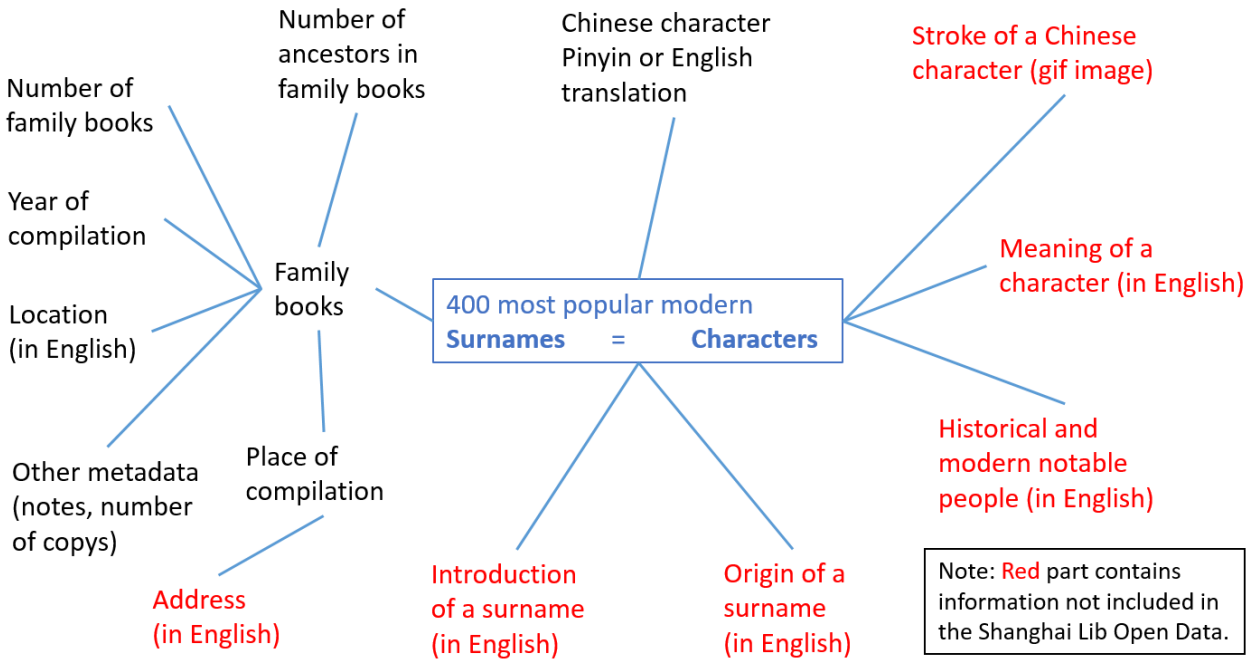


Figure 1: Information demonstrated in the App, Learn Chinese Surnames

Apart from the genealogical dataset from Shanghai Library, three other Linked Open Datasets have been used in this App, namely, DBpedia (extracting information from Wikipedia), Wiktionary and GeoNames. The table (Table 1) presents the data source, the part of information used, method of consumption, usage in the App and copyright notice for each dataset.

The interface of this App is presented in the figure (Figure 2) to show the usage of newly enriched English descriptions for Chinese genealogical data. The figure (Figure 2(a)) displays the Wiktionary and Wikipedia URL links corresponding to the surname Su ([Chinese character]), while the figure (Figure 2(b)) displays the translation of geographical place data to English in the Chinese genealogical dataset (from "[chinese character]" to "Xiuning Xian, Anhui China"). GIF Images of Simplified and Traditional Chinese Characters were called at runtime using URLs within the domain of WrittenChinese.Com[16].

3.2 Consuming Linked Data to Enrich Cross-Lingual Information

Three pieces of information in English were enriched in through consuming three Linked Open Data Sources.

- GeoNames is used to obtain the English-described information about the place of compilation of Chinese family books;

http://www.worldcat.org/title/zhong-guo-si-bai-da-xing/oclc/910234509]

[16]https://www.writtenchinese.com/

Linked Data Source	Information Used	Method of Data Consumption	Usage in the App	Copyright Notice
Shanghai Library Genealogical Linked Data	Metadata of family books (Title, year of compilation, place of compilation, location (@en)), pinyin and number of ancestors related to surnames	Real-time calling data through Restful service supported by SPARQL Endpoint	Display texts on App	(1) CC BY-NC-SA 2.0; (2) Right to use the data during contest
DBpedia and Wikipedia	Articles in English about Chinese Surnames	Offline querying data through SPARQL Endpoint	Use the whole mobile webpage within App	(1) CC BY-SA 3.0; (2) GNU Free Documentation License
Wiktionary	Articles in English about a Chinese Character	Real-time calling data from URL	Use the whole mobile webpage within App	CC BY-SA 3.0
GeoNames	Chinese Geographical location information described in English	Real-time calling data through official API	Display texts on App	CC BY-SA 3.0

Table 1: Linked Open Datasets Used in the App, Learn Chinese Surnames

- DBpedia, which extracts information from Wikipedia, is used to get the information about Chinese surnames;
- Wiktionary is used to obtain the information corresponding to a surname character.

The methods used in this project to consume these data are various, including direct access with URLs, querying the SPARQL endpoint with Restful service or calling the official API. Data can be cached locally, as it is done for DBpedia, or called online at runtime, as applied on the Chinese Genealogical data and on GeoNames. The methods are chosen for the ease of implementation, and due to access limitations in terms of speed and connection, quality of data, and data needs for the App. A brief note on the methods to consume the Linked Data sources is in the table (Table 1). The figure (Figure 3) illustrates the path of consuming Linked Open Data sources to enrich Shanghai Library Genealogical Data.

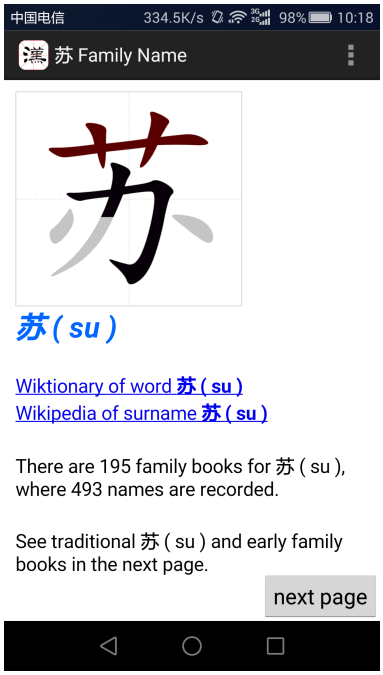
3.2.1 Enrichment of geographical data of genealogy compilation places: through API interface of GeoNames

The GeoNames dataset[17] contains over 10 million names and 9 million unique features, such as population and alternate names, of geographical places in the world. Geographical names are multilingual, thus can enable users to get the English names of places where their official languages are not English. Users may edit, correct and add new names using a wiki interface. The GeoNames ontology is available through datahub[18]. GeoNames dataset can be fully downloaded, queried using a third party SPARQL endpoint powered by FactForge[19], or requested using the official API from GeoNames.

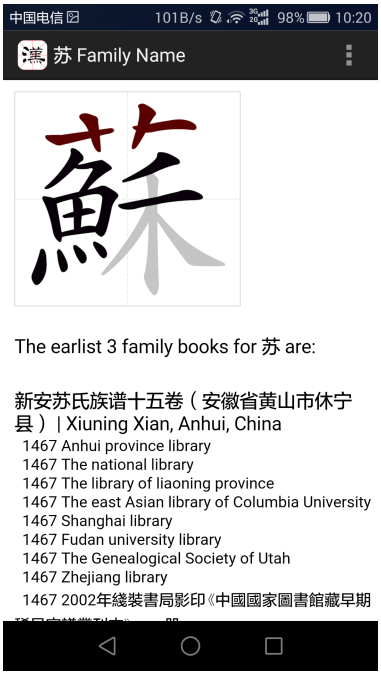
Considering the speed of connection and storage issues, the official API is used to retrieve GeoNames data at runtime. As shown in the figure (Figure 3), it is possible to use the hierarchical data (Country-Province-City-County) obtained using SPARQL endpoint from Shanghai Library to get a JSON format output using GeoNames API. Key parameters to construct the URL for API are featureCode, name and country.

To obtain GeoNames data of Hu Xian of City Xi??an in Province Shanxi in China ("[chinese character]"), which is the place where the family book "Duan Shi Shi Xi" ("[chinese character]") is compiled on 1731. The URL to query API is constructed as follow, `http://api.geonames.org/searchJSON`

[17] `http://www.geonames.org/about.html`
[18] `https://datahub.io/dataset/geonames-semantic-web`
[19] `http://factforge.net/`



(a) Wiktionary and Wikipedia Links to a Character and its Corresponding Surname



(b) Metadata in English of Earliest Chinese Family Books Corresponding to a Surname

Figure 2: Interface that Displays Data in the App, Learn Chinese Surname

?name_equals=%E6%88%B7%E5%8E%BF&featureCode=ADM3&country=CN&maxRows=10&username=XXX, where featureCode value ADM3 corresponds to the third-order administrative division; country corresponds to the abbreviation of the Country to query; username is a registered ID on GeoNames website; name_equal is a complete string match value and %E6%88%B7%E5%8E%BF is the percent-encoding in URI of the Chinese character "????". The JSON output displays the name of this place with its longitude, latitude and population.


```
{"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"}]}
```


If multiple results are returned due to the fact that some Chinese location names on the city or county level along can map to several geographical places, the province level name (adminName1 in the JSON output) can be further used to select the exact one. For places in Taiwan, the country input should be TW according to the GeoName database. Through the method above so far, no wrong or missed matches are discovered, which means that both the precision and recall nearly reach 100%.

3.2.2 Enrichment of surname culture information: through SPARQL endpoint of DBpedia

The DBpedia project[20] extracts structured information from Wikipedia based on crowdsourcing, making it a cross-domain, community-based, constant evolving and multilingual knowledge base. DBpedia has been localized to versions of 125 languages describing 38.3 million things. Serving as the core Linked Data source on the Web, the knowledge base covers 3 billion pieces of RDF triples and connects to other Linked Datasets by around 50 million RDF links.

[20] http://wiki.dbpedia.org/about

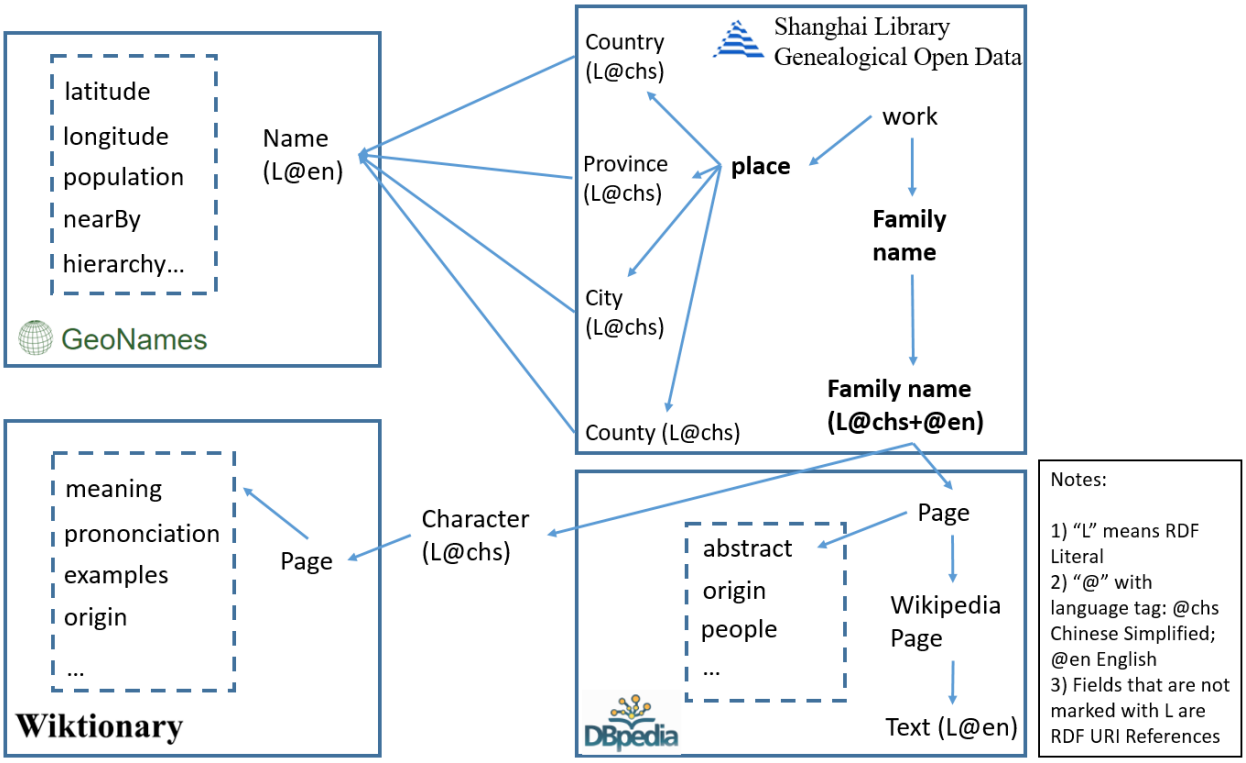


Figure 3: Consuming Three Multilingual Linked Dataset to Enrich Shanghai Library Genealogical Dataset

For this case study, DBpedia and Wikipedia are ideal sources since they contain information about Chinese surname and its culture in English. In the RDF of DBpedia, foaf:isPrimaryTopicOf links a thing in DBPedia with a page URL of Wikipedia. Although DBpedia organizes resources with standard URIs, but due to the inconsistency of vocabularies or fields among pages in DBpedia and the inconsistency among page names, it is not suggested to conduct batch processing of information in articles of Chinese surnames. In the case study, a semi-automatic approach is used, combining SPARQL queries with human decision, to achieve an acceptable data matching.

A good knowledge on the ontological vocabularies of DBpedia is vital for querying data with SPARQL. Through the dct:subject relation, the resources can be refined to a very specific domain; through dbo:abstract, the corresponding abstracts of articles in multiple languages can be extracted; by foaf:isPrimaryTopicOf, the corresponding Wikipedia URL can be obtained. The two examples below show the extraction of Wikipedia URL of Chinese surname "[chinese \character]' (Simplified\Traditional Chinese) using abstract to match Chinese characters or using URL to match pinyin. The number of output is set as the most relevant URL.

Below is an example SPARQL query of obtaining the most relevant Wikipedia URL from a Chinese character '[chinese \character]". The measure of relevance is through the number of times the Character appears in the abstract of an article; the bigger the value of number, the more relevant this abstract is. The Wikipedia URL linked to the most relevant abstract is output as the result.

```
# DBpedia SPARQL Endpoint
# Method 1: Matching strings of Chinese characters to the abstract field in DBpedia.
select distinct ?url ?ex count(?res) as ?count
where{
?res dct:subject <http://dbpedia.org/resource/Category:Chinese-language_surnames>.
?res dbo:abstract ?a.
filter(contains(str(?a), "??") || contains(str(?a), "?w")).
?res foaf:isPrimaryTopicOf ?url.
optional {?res dbo:wikiPageExternalLink ?ex}
}
order by desc(?count)
limit 1
```

Below is an example SPARQL query of obtaining the most relevant Wikipedia URL from a Chinese pinyin "Zeng" of Character "[chinese character]". The measure of relevance is through the length of matched string to the Chinese pinyin; exact matching produces the lowest length. The matched Wikipedia URL which has the lowest length is output as result.

```
# DBpedia SPARQL Endpoint
# Method 2: Matching pinyin of Chinese characters to page URL in DBpedia
select distinct ?u
where{
?m dct:subject <http://dbpedia.org/resource/Category:Chinese-language_surnames>.
filter(contains(str(?m),"Zeng")).
?m foaf:isPrimaryTopicOf ?u.
}
order by asc(fn:string-length(?u))
limit 1
```

Through a final manual matching mediated by the two methods above, it is possible to achieve 100% precision and recall of extracting a Wikipedia URL from a Chinese character.

3.2.3 Enrichment of character related information: through URIs of Wiktionary

The Wiktionary project[21] creates a multilingual dictionary of "all words in all languages" through the collaborative contribution of volunteers online; the freely open dictionary is available in 172 languages[22]. DBpedia has extracted RDF triples from Wiktionary and opened the SPARQL endpoint as a side project[23], however, currently only RDF graphs of 4 languages have been fully created, not including Chinese. Wiktionary has also its official MediaWiki API[24], but considering the inconsistency of structures among Wiktionary articles, the API is not used in the project of this case study.

In the case study, URLs of Wiktionary are directly constructed and used to get and display the HTML information on the Android App. Unlike Wikipedia, the URL of Wiktionary are more standard, thus enabling direct construction of the Wiktionary URL with any Chinese character as an input. For example, the Wiktionary URI for Chinese character "[chinese character]" is `https://en.wiktionary.org/wiki/%E5%88%98`, where "%E5%88%98" is the percent-encoding in URI of "[chinese character]". Apart from Wiktioanry, another possible multilingual dictionary Linked Data source that can be used for enriching cross-lingual information is BabelNet[25].

3.3 Data Validation

Errors and conflicts are common in datasets. When it comes to matching data in different sources, it is necessary to validate the data and solve the conflicts originated from different structures and semantics. In this case study, *comprehensiveness* and *conflicts* of data are measured.

For data comprehensiveness, over the 400 most popular modern Chinese surnames, 377 of them (94.25%) are included in the Shanghai Library Genealogical dataset, and 295 of the 400 surnames (73.75%) have a corresponding article in English Wikipedia. Based on the 377 surnames in the Shanghai Library dataset, 93 of them (24.67%) do not have a corresponding article in English Wikipedia. This result can be later used to add new entities in these data sources.

For data conflicts, through the matching of Wikipedia articles to Shanghai Library dataset, 5 conflicts on the entity level were discovered in pinyin or English translation of Chinese surnames, as listed in the table (Table 2). There are also some slight internal conflicts on the forms and suffixes of page

[21] `https://www.wiktionary.org/`
[22] `https://en.wikipedia.org/wiki/Wiktionary`
[23] `http://wiki.dbpedia.org/wiktionary-rdf-extraction`
[24] `https://en.wiktionary.org/w/api.php`
[25] `https://datahub.io/dataset/babelnet`

names in URLs of DBpedia and Wikipedia[26]. For example, surname "[Chinese character]" corresponds to Liu; "[Chinese character]" corresponds to [Chinese character] with a tone in the Pinyin[Chinese character] corresponds to Bo_(Chinese_surname) with the suffix (Chinese_surname); "[Chinese character]" corresponds to Pang_(surname) with a different suffix (surname). This inconsistency is common and caused some effort on consuming the data.

Surname Characters	Pinyin in Shanghai Lib Data	English Translation of Surnames in Wikipedia	URLs of the Wikipedia articles	Notes
??	fang	pang	`http://en.wikipedia.org/wiki/Pang_(surname)`	Both applicable in different situations
??	bai	bo	`http://en.wikipedia.org/wiki/Bo_(Chinese_surname)`	Should be "bo"
??	qu	ou	`http://en.wikipedia.org/wiki/Ou_(surname)`	Should be "ou"
??	qiang	jiang	`http://en.wikipedia.org/wiki/Jiang_(surname)`	Should be "jiang"
??	wei	ngai	`http://en.wikipedia.org/wiki/Ngai_(surname)`	Both applicable; "ngai" is special for Cantonese

Table 2: Conflicts of English Translations of Chinese Surnames between Two Data Sources

In addition, on the ontological vocabulary level, internal inconsistency exists among the pages in DBpedia: the internal structure or the list of fields among surnames does not follow a strict standard. For example, the ontological vocabularies in DBpedia pages for "[Chinese character]"[27] and "[Chinese character]"[28] are very different, due to the nature of crowdsourcing and the level of popularity among different surnames. Similar issues were also found in Wiktionary. This issue has made it hard to process the semantics of surnames on a higher granularity using the vocabularies, for example, currently it is not possible to extract all notable people related to all Chinese surnames using one SPARQL query.

4 Conclusion and Future Research

A comprehensive multilingual Web of Data is a realizable dream that could boost the international interoperability of semantic resources on the Web, based on general services on the macro-level and small or middle applications on the micro-level, as the one in the case study. The overall practice of Linked Data for many organizations is to construct a worldwide Web of Data, and offer advantages such as promoting data re-use and enhancing the discoverability and interoperbility of resources on the semantic level. It is however that, without adequate resources or services to overcome language barriers in the Semantic Web, data networking across language and culture would not be possible.

In this research, the author recapitulated the idea of enriching cross-lingual information of Linked Data to create a Multilingual Web of Data. The task of enriching cross-lingual information for Linked Data has been defined as adding a layer of information on the entity level or vocabulary level to the existing infrastructure, the Linked Open Data cloud. On the entity level, the layer of cross-lingual information can contain (a) descriptions that express the entity in a different language, or (b) URLs or links to an entity described in a different language.

Digital heritage resources in places where English is not a native language, are facing barriers of languages on the Web. Language barriers could impede the accessibility of cultural heritage resources and thus international cooperation in the field. Among recent digital heritage projects that aim at enriching cross-lingual information, manual effort including crowdsourcing, has been mostly used to derive the translations

[26]DBpedia URLs are http://dbpedia.org/page/*PAGE_NAME*, using the same page names as Wikipedia URLs https://en.wikipedia.org/wiki/*THE_SAME_PAGE_NAME*

[27]`http://dbpedia.org/page/D%C7%92ng`

[28]`http://dbpedia.org/page/Liu`

which highly depend on culture. It is also discovered that there are only few multilingual Linked Data projects in the cultural heritage domain.

For some cultural heritage resources, as the one in this case study, matching to existing multilingual data sources is a better choice, which can reduce the human effort to re-define the same entity in another language. The major part of this article, therefore, introduced the case study as a practice of enriching English-described information for Chinese Genealogical Linked Data through consuming multilingual resources in the Linked Open Data cloud.

Throughout the project, the designer consumed three multilingual semantic resources, DBpedia, Wiktionary and GeoNames, with the approaches that call the API, SPARQL, URL of these resources. The issue of data quality, including the comprehensiveness and conflicts among the three datasets, is considered to make appropriate choices for methods to consume data. The variable choices of interface is also a key for the development of open data, for example, all the four Linked Datasets used in this study (Table 1) have multiple data consumption interfaces available online, no matter official or third-party.

For future studies in terms of the project in the case study, first, it is expected to extract other cross-lingual information with higher semantic granularity, for example the relationship among all notable people of a Chinese surname and their relationships in English Wikipedia, through intelligent data matching techniques. Second, the comprehensiveness of dataset could be enhanced: it is possible to use the Genealogical data from Shanghai Library to supplement data in English Wikipedia; and use the GIS information in GeoNames to enhance the visualization of Chinese family books. Third, for copyright concerns on the use of data, such as embedding webpages and images in a mobile App through URLs, future surveys should be conducted before formally releasing the App to the public.

For a wider picture, it is worth discussing the impact of enriching cross-lingual information on digital heritage projects: how it addresses the information needs of users in the area related to a type of cultural heritage resource, and how it would benefit the cultural heritage organizations. In addition, technically, for any other multilingual digital heritage or digital library projects, the enrichment of cross-lingual information may only be a first step: Further issues are awaiting to be addressed, including the representation, query and automatic services for multilingual information on Linked Data.

References

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far. In A. Sheth (Ed.), (pp. 205–227). Hershey, PA: Information Science Reference. doi: 10.4018/978-1-60960-593-3.ch008

Chen, S., & Chen, H. (2012). Mapping multilingual lexical semantics for knowledge organization systems. *The Electronic Library*, *30*(2), 278-294. Retrieved from http://dx.doi.org/10.1108/02640471211221386 doi: 10.1108/02640471211221386

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. doi: 10.1007/978-3-662-43585-4_13

Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gomez-Perez, 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. doi: 10.1016/j.websem.2011.09.001

Hyvonen, E. (2012). Publishing and using cultural heritage linked data on the semantic web. *Synthesis Lectures on the Semantic Web: Theory and Technology*, *2*(1), 1-159. doi: 10.2200/S00452ED1V01Y201210WBE003

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., ... Bizer, C. (2015). Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, *6*(2), 167–195. doi: 10.3233/SW-140134

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. Retrieved from http://dx.doi.org/10.1108/EL-08-2013-0141 doi: 10.1108/EL-08-2013-0141

Mejía, M. E., Montiel-Ponsoda, E., de Cea, G. A., & Gómez-Pérez, A. (2012). Ontology localization. In C. M. Suárez-Figueroa, A. Gómez-Pérez, E. Motta, & A. Gangemi (Eds.), *Ontology engineering in a networked world* (pp. 171–191). Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/978-3-642-24794-1_8

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. doi: 10.3115/1072228.1072390

Oomen, J., Baltussen, L., & Van Erp, M. (2012, April 11-14). Sharing cultural heritage the linked open data way: why you should sign up. In *Museum and on the web 2012.* San Diego, CA, USA. Retrieved from http://www.museumsandtheweb.com/mw2012/papers/sharing_cultural_heritage_the_linked_open_data

Pazienza, M. T., & Stellato, A. (2006, May 24-26). Exploiting linguistic resources for building linguistically motivated ontologies in the semantic web. In *Proceedings of ontolex workshop.* Magazzini del Cotone Conference Center, Genoa, Italy. Retrieved from http://art.uniroma2.it/publications/docs/2006_OntoLex06_Exploiting%20Linguistic%20Resources%20for%20building%20linguistically%20motivated%20ontologies%20in%20the%20Semantic%20Web.pdf

Schmachtenberg, M., Bizer, C., & Paulheim, H. (2014). Adoption of the linked data best practices in different topical domains. In P. Mika et al. (Eds.), *The semantic web – iswc 2014: 13th international semantic web conference, riva del garda, italy, october 19-23, 2014. proceedings, part i* (pp. 245–260). Cham: Springer International Publishing. doi: 10.1007/978-3-319-11964-9_16

Singer, R. (2010, February 22-25). The linked library data cloud: Stop talking and start doing. In *Code4lib 2010.* Asheville, North Carolina. Retrieved from http://code4lib.org/conference/2010/singer

Trojahn, C., Fu, B., Zamazal, O., & Ritze, D. (2014). State-of-the-art in multilingual and cross-lingual ontology matching. In P. Buitelaar & P. Cimiano (Eds.), *Towards the multilingual semantic web: Principles, methods and applications* (pp. 119–135). Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/978-3-662-43585-4_8

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 & demonstrations track - volume 1035* (pp. 121–124). Aachen, Germany, Germany: CEUR-WS.org. Retrieved from http://dl.acm.org/citation.cfm?id=2874399.2874430

Xia, C., Liu, W., Chen, T., & Zhang, L. (2016, May). A genealogy data service platform implemented with linked data technology [article in chinese]. *Journal of Library Science in China*, *03*. Retrieved from http://en.cnki.com.cn/Article_en/CJFDTotal-ZGTS201603003.htm

Xia, C., Liu, W., Zhang, L., & Zhu, W. (2014). A genealogical ontology in the form of bibframe model [article in chinese]. *Library Tribune*, *11*, 002. Retrieved from http://en.cnki.com.cn/Article_en/CJFDTotal-TSGL201411002.htm