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] APEX Data & Knowledge Management Lab, Shanghai Jiao Tong University
{xingniu, xrsun, whfcarter, rongshu, yyu}@apex.sjtu.edu.cn
[2] Southeast University
gqi@seu.edu.cn

Abstract. Linking Open Data (LOD) has become one of the most important community efforts to publish high-quality interconnected semantic data. Such data has been widely used in many applications to provide intelligent services like entity search, personalized recommendation and so on. While DBpedia, one of the LOD core data sources, contains resources described in multilingual versions and semantic data in English is proliferating, there is very few work on publishing Chinese semantic data. In this paper, we present Zhishi.me, the first effort to publish large scale Chinese semantic data and link them together as a Chinese LOD (CLOD). More precisely, we identify important structural features in three largest Chinese encyclopedia sites (i.e., Baidu Baike, Hudong Baike, and Chinese Wikipedia) for extraction and propose several data-level mapping strategies for automatic link discovery. As a result, the CLOD has more than 5 million distinct entities and we simply link CLOD with the existing LOD based on the multilingual characteristic of Wikipedia. Finally, we also introduce three Web access entries namely SPARQL endpoint, lookup interface and detailed data view, which conform to the principles of publishing data sources to LOD.

1 Introduction

With the development of Semantic Web, a growing amount of open structured (RDF) data has been published on the Web. Linked Data[3] initiates the effort to connect the distributed data across the Web and there have been over 200 datasets within Linking Open Data (LOD) community project[3]. But LOD contains very sparse Chinese knowledge at the present time. To our knowledge, only Zhao[20] published some Chinese medicine knowledge as Linked Data. However, all data is represented in English thus Chinese language users can hardly use it directly. Some multilingual datasets exist in LOD. Freebase[4], a collection of structured data, contains a certain number of lemmas with Chinese labels. DBpedia only extracts labels and short abstract in Chinese[4]. UWN[11] is another effort in constructing multilingual lexical knowledge base, which maps Chinese

[3] http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
[4] http://www.freebase.com/

words to corresponding vocabulary entries in WordNet. In other words, all of these datasets attach English descriptions to Chinese lemmas. Lack of real-world useful Chinese semantic data cramps the development of semantic applications as well as the Semantic Web itself in the Chinese language community.

When building Chinese Linking Open Data, we have to face some challenging problems that exist in building LOD. In our work, we focus on dealing with two of these challenging problems, that is, managing the heterogeneity of knowledge in different data sources and efficiently discovering `<owl:sameAs>` relations between millions of instances.

It is worthwhile to note that many knowledge bases are not original data providers but extracted from textual articles of other independent data sources. DBpedia[1], which structures Wikipedia knowledge, is a representative one. Fortunately, Chinese textual knowledge is abundant. Chinese Web-based collaborative encyclopedias together contain even more articles than the largest English one: Wikipedia. Besides Wikipedia Chinese version, two Web-based local encyclopedias, Baidu Baike[5] and Hudong Baike[6] has about 3.2 million and 2.8 million[7] articles respectively.

While there is plenty of Chinese textual knowledge, there is very few Chinese semantic data extracted from these sources of knowledge. We are making efforts to build the first Chinese LOD. A fusion of three largest Chinese encyclopedias is our initial achievement which has practical significances: we wish it can help in attracting more efforts to publish Chinese semantic data linked to Zhishi.me. The potential applications of this effort include Chinese natural language processing (entity recognition, word disambiguation, relation extraction), Chinese semantic search, etc.

The work we present in this paper includes:

- We take the three largest Chinese encyclopedias mentioned above as our original data and extract structured information from them. In total, about 6.6 million lemmas as well as corresponding detailed descriptions such as abstracts, infoboxes, page links, categories, etc. are parsed and presented as RDF triples. Procedure of structured data extraction will be introduced in Section 2.
- Since these three encyclopedias are operated independently and have overlaps, we integrate them as a whole by constructing `<owl:sameAs>` relations between every two encyclopedias. Some parallel instance-level matching techniques are employed to achieve this goal. Detailed methods' descriptions can be found in Section 3.1.
- In order to make connections with existing linked data and build a bridge between the English knowledge base and the Chinese knowledge base, we also use `<owl:sameAs>` to link resources in CLOD to the ones in DBpedia, a central hub in LOD. We will discuss it in Section 3.2.

[5] `http://baike.baidu.com/`
[6] `http://www.hudong.com/`
[7] Statistics collected in March, 2011

– The first Chinese Linking Open Data around the world has been published as RDF triples on the Web via Zhishi.me. Access mechanisms are presented in Section 4. Finally, we will make some conclusion and outline future work in Section 5.

2 Semantic Data Extraction

We do not generate knowledge data from scratch, but structure existing Web-based encyclopedia information. In this section, we will first introduce the strategies we used to extract semantic data in the form of RDF triples and then give a general view of our knowledge base.

2.1 Extraction Approach

We have three original data sources: Wikipedia Chinese version, Baidu Baike, and Hudong Baike. They provide different ways to edit articles and publish them. Thus there is no one-plan-fits-all extraction strategies.

Wikipedia provides database backup dumps[8], which embed all wiki articles in the form of wikitext source and meta data in XML. The techniques for extracting information from Wikipedia dumps are rather mature. DBpedia Extraction Framework[4] is the most typical efforts. We employ a similar extraction algorithm to reveal structured content from infobox templates as well as their instances.

Wikipedia uses the wikitext language, a lightweight markup language, while both Baidu Baike and Hudong Baike provide the WYSIWYG (what you see is what you get) HTML editors. So all information should be extracted from HTML file archives. Article pages come from these three encyclopedias are alike with minor differences in layout as shown in Figure 1. Currently, we extract 12 types of article content: abstracts, aliases, categories, disambiguation, external links, images, infobox properties, internal links (pagelinks), labels, redirects, related pages and resource ids. They will be explained in detail as follows:

Abstracts. All of these three encyclopedias have separate abstract or summary sections and they are used as values of `zhishi:abstract` property.

Aliases. In Wikipedia, editors can customize the title of an internal link. For example, `[[People's Republic of China|China]]` will produce a link to "People's Republic of China" while the displayed anchor is "China". We call the displayed anchors as the aliases of the virtual article and represent them using `zhishi:alias`. Users cannot rename internal links in Hudong Baike and Baidu Baike, so aliases are not included in these two sources.

Categories. Categories describe the subjects of a given article, `dcterms:subject` is used to present them for the corresponding resources in Zhishi.me. Categories have hyponymy relations between themselves which are represented using `skos:broader`.

[8] Dumps of zhWiki: `http://dumps.wikimedia.org/zhwiki/`

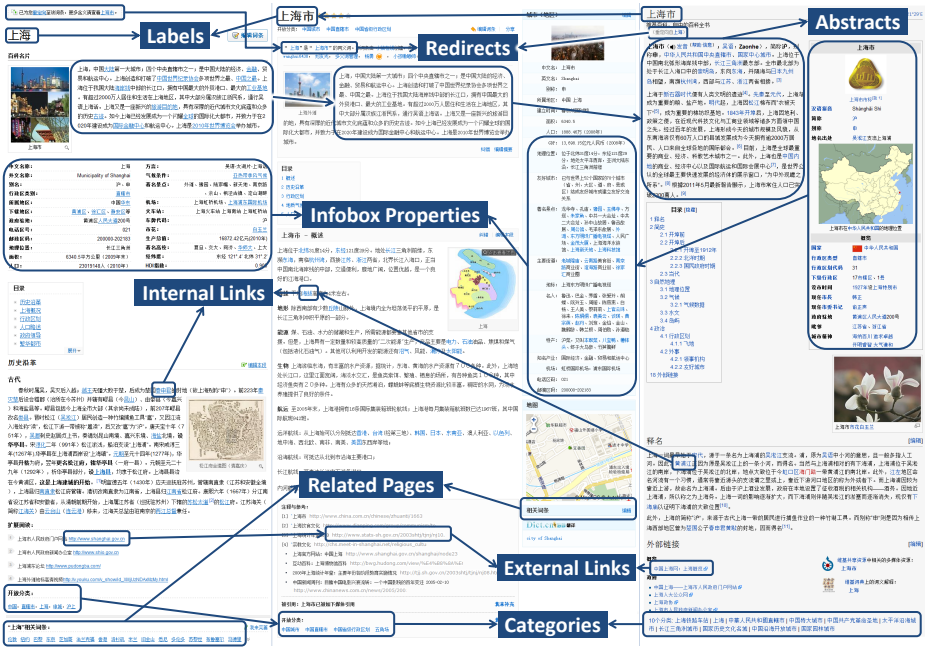

Fig. 1: Sample Encyclopedia Article Pages from Baidu Baike, Hudong Baike, Chinese Wikipedia Respectively

Disambiguation. We say a single term is ambiguous when it refers to more than one topic.

- Hudong Baike and Wikipedia (mostly) use disambiguation pages to resolve the conflicts. However, disambiguation pages are written by natural language, it is not convenient to identify every meaning of homonyms accurately. Thus we only consider a topic, that linked from one disambiguation page and which label shares the same primary term, as a valid meaning. For example, "Jupiter (mythology)" and "Jupiter Island" share the same primary term: "Jupiter".
- Baidu Baike puts all meanings as well as their descriptions of a homonym in a single page. We put every meaning in a pair of square brackets and add it to the primary term as the final Zhishi.me resource name, such as "Term[meaning-1]".

All disambiguation links are described using `zhishi:pageDisambiguates`.

External Links. Encyclopedia articles may include links to web pages outside the original website, these links are represented using `zhishi:externalLink`.

Images. All thumbnails' information includes image URLs and their labels are extracted and represented using `zhishi:thumbnail`.

Infobox Properties. An infobox is a table and presents some featured properties of the given article. These property names are assigned IRIs as the form `http://zhishi.me/[sourceName]/property/[propertyName]`.

Internal Links. An internal link is a hyperlink that is a reference in an encyclopedia page to another encyclopedia page in the same website. We represent these links using predicate `zhishi:internalLink`.

Labels. We use Hudong Baike and Wikipedia article's titles as labels for the corresponding Zhishi.me resources directly. Most Baidu Baike articles adopt this rule, but as explained above, meanings of homonyms are renamed. Labels are represented by `rdfs:label`.

Redirects. All of these three encyclopedias use redirects to solve the synonymous problem. Since Wikipedia is a global encyclopedia while the other two are encyclopedias from Mainland China, Wikipedia contains redirects between Simplified and Traditional Chinese articles. There are rules of word-to-word correspondence between Simplified and Traditional Chinese, so we just convert all Traditional Chinese into Simplified Chinese by these rules and omit redirects of this kind. Redirect relations are described by `zhishi:pageRedirects` to connect two Zhishi.me resources.

Related pages. "Related pages" sections in Baidu Baike and Hudong Baike articles are similar to "see also" section of an article in Wikipedia but they always have fixed positions and belong to fixed HTML classes. Predicate `zhishi:relatedPage` is used to represent this kind of relation between two related resources.

Resource IDs. Resource IDs for Wikipedia articles and most Baidu Baike articles are just the page IDs. Due to the reason that every Zhishi.me resource of homonyms in Baidu Baike is newly generated, they are assigned to special values (negative integers). Articles from Hudong Baike have no page IDs, so we assign them to private numbers. `zhishi:resourceID` is used here.

2.2 Extraction Results and Discussions

Encyclopedia articles from Baidu Baike and Hudong Baike are crawled in March, 2011 and Wikipedia Dump version is 20110412. From the original encyclopedia articles (approximately 15 GB compressed raw data), we totally extracted 124,573,857 RDF triples. Table 1 shows statistics on every extracted content in detail.

Baidu Baike accounts for the most resources, the most categories, as well as the most resources that have categories. It indicates that this data source has a wide coverage of Chinese subjects. The other two data sources, Hudong Baike and Chinese Wikipedia, are superior in relative number of infobox and abstract information respectively. In Table 1, the decimals follows the absolute number is the fractions that divided by the total resources number in each data source.

Comparing other types of content extraction results, each data source has it's own advantage. Baidu Baike has 0.80 images per resource, which means 0.80 picture resources are used on average in one article. Analogously, Hudong Baike has relative more infobox properties and related pages than others. While Wikipedia Chinese version articles contain more external and internal links.

Table 1: 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**	

Table 2: Most Used Properties in Each Data Source

Baidu Baike		**Hudong Baike**		**Chinese Wikipedia**	
Chinese Name	37,445	Chinese Name	152,447	Full Name	3,659
Nationality	22,709	Sex	74,374	Population	3,500
Date of Birth	21,833	Occupation	71,647	Area	3,272
Birthplace	19,086	Nationality	70,260	Website	3,061
Occupation	18,596	Era	61,610	Language	2,875
Foreign Name	16,824	Date of Birth	57,850	Height	2,710
Alma Mater	10,709	Home Town	52,518	Kana	2,577
Representative Works	9,751	English Name	52,504	Hiragana	2,203
Nationality[1]	9,621	Kingdom[2]	41,126	Director	2,116
Achievements	7,791	Scientific Name	41,039	Romanization	2,100
Kingdom[2]	7,749	Achievements	40,751	Prefectures	2,099
Category	7,732	Category	40,709	Japanese Name	2,096
Alias	7,725	Family[2]	39,858	Starring	2,015
Family[2]	7,715	Phylum	39,637	Scenarist	1,949
Scientific Name	7,355	Class[2]	39,412	Address	1,949

[1] An ethnic group.
[2] A category in the biological taxonomy.

Alias, disambiguation links and redirects constitute a valuable thesaurus that can help people to search out most relevant knowledge. Wikipedia Chinese ver-

sion performs better in this aspect, for the overwhelming superiority it achieves in number of these attributes, even if it has a narrower resource coverage.

Infobox information is the most worthy knowledge in encyclopedias, so we carry out further discussions on this issue here. In Table 2, we list some most frequently used infobox properties. The original properties are written in Chinese but we translated them into English.

The types of resources that are more likely to use infobox can be easily inferred from these frequently used properties. Hudong Baike, which has abundant infobox information, has a large quantity of persons and organisms described in minute detail. Similarly, most listed Baidu Baike properties manifest different facets of somebodies or living things. Chinese Wikipedia also describes lots of people, but in a little different perspective. In addition, featured properties for geographical regions (population, area, etc.) and films (director, starring, etc.) can be seen.

All data sources have their own characteristics, nevertheless they represent subjects in a similar manner, which makes it possible to integrate these attributes. We will give specific examples in Section 4.

Resources that have infobox information are much less than ones with categories. Unfortunately, the quality of these categories are not very high due to the reason that encyclopedia editors usually choose category names casually and many of them are not used frequently. Thus we adopt some Chinese words segmentation techniques to refine these categories, and then choose some common categories to map them to YAGO categories[18] manually.

Top 5 categories in each data source are listed in Table 3. Total number of instances of these categories accounts for over one third resources that have category information. Also notice the top categoires have many overlaps in the three data sources. This suggests the integration of these knowledge bases that we will discuss in the next section is based on good sense.

Table 3: Top 5 Categories with the Number of Their Instances in Each Data Source

Baidu Baike		**Hudong Baike**		**Chinese Wikipedia**	
Persons	376,509	Persons	93,258	Persons	50,250
Works	266,687	Works	81,609	Places	28,432
Places	109,044	Words and Expressions	70,101	Organisms	15,317
Words and Expressions	69,814	Places	40,664	Organizations	12,285
Organisms	55,831	Pharmaceuticals	22,723	Works	8,572
Subtotal	877,885	Subtotal	308,355	Subtotal	114,856
Account for	36.6%	Account for	33.8%	Account for	36.5%

3 Data-level Mapping among Different Datasets

Baidu Baike, Hudong Baike and Wikipedia have their own adherents. Most of the time, users edit a certain article by their personal knowledge and that lead to heterogeneous descriptions. Mapping these articles with various description styles can help to integrate these separated data sources as a whole. At the same time, we try to bridge the gap between our Chinese knowledge base and English one (the Linking Open Data). Descriptions in different languages of a same subject can supplement each other. We will introduce the methods we use to achieve these goals in next two sub-sections.

3.1 Mappings within CLOD

Finding mappings between datasets in Linking Open Data is usually done in two levels. One is the practice of schema-level ontology matching, as Jain *et al.*[9] and Raimond *et al.*[14] did. The other one aims at matching instances and we mainly focus on this kind of mapping discovery. Not all existing instance matching algorithms are suitable for finding `<owl:sameAs>` links between large-scale and heterogeneous encyclopedias. For example, KnoFuss[13] need instance data represented as consistent OWL ontologies, however, our extracted semantic data does not meet this requirement. Raimond *et al.*[17] proposed an interlinking algorithm which took into account both the similarities of web resources and of their neighbors but had been proved to be operative in a really small test set.

Silk[19] is a well-known link discovery framework, which indexes resources before detailed comparisons are performed. Pre-matching by indexes can dramatically reduce the time complexity on large datasets, thus we also match our encyclopedia articles based on this principle.

Simply indexing resources by their labels has some potential problems. One is that the same labels may not represent the same subject: different subjects having the same label is quite common. The other one is opposite: same subject may have different labels in some cases. These two possible situation would affect the precision and recall in matching respectively.

We will introduce how we deal with this problem in practice by proposing three reasonable but not complex strategies to generate the index.

Using Original Labels The first index generation strategy is just using original labels. This strategy normally has a high precision except it comes with the problem of homonyms. Fortunately, we extract different meanings of homonyms as different resources, which has been introduced in Section 2.1. In other words, it is impossible to find two resources that have different meanings with the same label if all homonyms are recognized. This fact ensures the correctness of this strategy.

There is no denying that the performance of this method depends on the quality of existing encyclopedia articles: whether the titles are ambiguous.

Punctuation Cleaning When it comes to the second problem: discovering mappings between resources with different labels, one of the most efficient methods we used is punctuation cleaning. Figure 2 shows some examples of same entity having different labels due to the different usage of Chinese punctuation marks. These cases can be handled by the punctuation cleaning method.

(1) 肖申克的救赎 = 《肖申克的救赎》
(2) 海尔波普彗星 = 海尔·波普彗星 = 海尔-波普彗星
(3) 奋进号航天飞机 = “奋进号”航天飞机

Fig. 2: Examples of Same Entity Has Different Labels

1. Some Chinese encyclopedias encourage editors to use guillemets (≪≫) to indicate the title of a book, film or album etc. However, guillemets are not imperative to be part of titles. Example (1) in Figure 2 illustrates a same film with/without guillemets.
2. In Chinese, we often insert an interpunct (·) between two personal name components. In some certain cases, people may insert a hyphen instead or just adjoin these components. Example (2) in Figure 2 shows three different labels to indicate a comet named after two persons.
3. According to the usage of Chinese punctuation marks, it is a good practice to quote a cited name by double styling quotation marks (“”). However, it is not a mandatory requirement. Example (3) in the figure indicates a space shuttle with its name *Endeavour* quoted and not quoted.

Punctuation marks may have special meanings when they constitute a high proportion of the whole label string. So we calculate the similarity between two labels using Levenshtein distance[10] and attach penalty if strings are too short.

Extending Synonyms The third strategy we use in index generation also deals with the problem of linking resources with different labels. This one is making use of high quality synonym relations obtained from redirects information (A redirects to B means A and B are synonyms). We can treat redirects relations as approximate `<owl:sameAs>` relations temporarily and thereupon find more links based on the transitive properties of `<owl:sameAs>`.

Usually, the title of a redirected page is the standard name. So we just link two resources with standard names to avoid redundancy. Resources with aliases can still connect to other data source via `pageRedirects`.

Since our dataset is very large, it still has a great time and space complexity even we adopt the pre-matching by index method. We utilize distributed MapReduce[5] framework to accomplish this work. All resources are sorted by their index term in a map procedure, and naturally, similar resources will gather

together and wait for detailed comparisons. In practice, totally approximately 24 million index terms are generated from our data sources. This distributed algorithm makes it easier to discovering links within more datasets because pairwise comparisons between every two datasets are avoided.

We say two resources are integrated if they are linked by `<owl:sameAs>` property. Thus two data sources have intersections when they provide descriptions for mutual things. The number of links found by our mapping strategies is reflected in Figure 3. It confirms the nature of heterogeneity in these three data sources. Original 6.6 million resources are merged into nearly 5 million distinct resources, while only a small proportion (168,481, 3.4%) of them are shared by all.

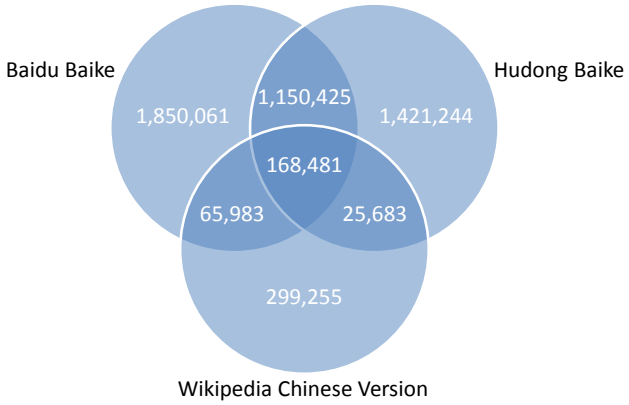

Fig. 3: Intersections of Our Three Data Sources Shown in Venn Diagram

3.2 Linking CLOD with LOD

DBpedia, a structured Wikipedia, is now one of the central data sources in Linking Open Data. Mapping resources between Zhishi.me and DBpedia is a cross-lingual instance matching problem, which is remaining to be solved. Ngai *et al.*[12] tried to use a bilingual corpus to align WordNet[9] and HowNet[10]. However, their experimental results showed that the mapping accuracy, 66.3%, is still not satisfiable.

Fu *et al.*[7] emphasized that translations played a critical role in discovering cross-lingual ontology mapping. Therefore, we make use of the high-quality Chinese-English mapping table at hand: Wikipedia interlanguage links. Wikipedia interlanguage links are used to link the same concepts between two Wikipedia language editions and this information can be extracted directly from wikitext.

[9] A large lexical database of English. `http://wordnet.princeton.edu/`
[10] A Chinese common-sense knowledge base. `http://www.keenage.com/`

DBpedia's resource names are just taken from the URLs of Wikipedia articles, linking DBpedia and wikipedia dataset in Zhishi.me is straightforward. Likewise, resources in DBpedia and the whole Zhishi.me can be connected based on the transitive properties of `<owl:sameAs>`. In total, 192,840 links are found between CLOD and LOD.

4 Web Access to Zhishi.me

Since Zhishi.me is an open knowledge base, we provide several mechanisms for Web accessing. Not only client applications can access it freely, but also human users can lookup and get the integrated knowledge conveniently.

4.1 Linked Data

According to the Linked Data principles[11], Zhishi.me create URIs for all resources and provide sufficient information when someone looks up a URI by the HTTP protocol. We have to mention that in practice, we use IRIs (Internationalized Resource Identifiers) instead of URIs. This will be discussed in detail next.

Since Zhishi.me contains three different data sources, we design three IRI patterns to indicate where a resource comes from (Table 4). The Chinese characters are non-ASCII, so we choose IRIs, the complement to the URIs[6], for Web access. But IRIs are incompatible with HTML 4[16], we have to encode non-ASCII characters with the URI escaping mechanism to generate legal URIs as "href" values for common Web browsing.

Table 4: IRI Patterns

Sources	IRI Patterns
Baidu Baike	`http://zhishi.me/baidubaike/resource/[Label]`
Hudong Baike	`http://zhishi.me/hudongbaike/resource/[Label]`
Wikipedia Chinese version	`http://zhishi.me/zhwiki/resource/[Label]`

Data is published according to the best practice recipes for publishing RDF vocabularies[2]. When Semantic Web agents that accept "application/rdf+xml" content type access our server, resource descriptions in RDF format will be returned. We try our best to avoid common errors in RDF publishing to improve the quality of the open data published on the Web. This issue has been discussed in [8].

[11] `http://www.w3.org/DesignIssues/LinkedData.html`

Fig. 4: An Example of Integrated Resources Page

In order to encourage non Semantic Web community users to browse our integrated data, we merge descriptions of the same instance and design an easy-to-read layout template to provide all corresponding contents. Figure 4 shows a sample page that displays the integrated data.

There are two ways to browse the integrated resources: via the lookup service, which will be introduced in the next section, or via "`<owl:sameAs>` box" at the upper right corner of a resource displaying page. In "`<owl:sameAs>` box", all

resources have `<owl:sameAs>` relation with currently displaying resources are listed and users can tick resources they want to merge.

Detailed descriptions listed in the view page are all statements with subjects being currently displaying resources, so all distinct subjects are assigned different colors which can help users to recognize where a description comes from.

Triples extracted from infoboxes are still presented together. If some statements are sharing the same predicate (property), we will merge the objects (values) but remain the identity colors. The comprehensive property-value pairs rely upon linking heterogeneous data sources.

Other statements grouped by their predicates, such as subjects, thumbnails, relatedPages, etc. are listed subsequently. Users may click on "SHOW MORE" if they want to know more information.

4.2 Lookup Service

For users who do not know the exact IRI of a given resource name, we provide a lookup service to help them. This service is available at `http://zhishi.me/lookup/`. Our index is constructed using four matching strategies presently:

- Returning all resources whose labels exactly match the user's query.
- Using `<owl:sameAs>` links to provide the co-references.
- Some known synonymous are used to provide as many as possible similar resources.
- If a given name has several different meanings (which reflected in a disambiguation page), all corresponding resources of these meanings will also be returned.

Figure 5 gives a sample query. If we search for "Pacific Ocean", it returns not only resources whose labels are exactly "Pacific Ocean" but also a TV miniseries named "The Pacific" (quoted by guillemets) and two disambiguation resources.

As mentioned above, the lookup service is also an entrance to integrate interrelate resources.

Fig. 5: A Sample Query to Lookup Service

4.3 SPARQL Endpoint

We also provide a simple SPARQL Endpoint for professional users to customize queries at `http://zhishi.me/sparql/`. We use AllegroGraph RDFStore[12] to store the extracted triples and provide querying capabilities.

5 Conclusions and Future Work

Zhishi.me, the first effort to build Chinese Linking Open Data, currently covers three largest Chinese encyclopedias: Baidu Baike, Hudong Baike and Chinese Wikipedia. We extracted semantic data from these Web-based free-editable encyclopedias and integrate them as a whole so that Zhishi.me has a quite wide coverage of many domains. Observations on these independent data sources reveal their heterogeneity and their preferences for describing entities. Then, three heuristic strategies were adopted to discover `<owl:sameAs>` links between equivalent resources. The equivalence relation leads to about 1.6 million original resources being merged finally.

We provided Web access entries to our knowledge base for both professional and non Semantic Web community users. For people who are familiar with Linked Data, Zhishi.me supports standard URIs (IRIs) de-referencing and provides useful information written in RDF/XML format. Advanced users can also build customized queries by SPARQL endpoint. Casual users are recommended to get well-designed views on the data when they use Web browsers. Both lookup service and data integration operations are visualized.

It is the first crack at building pure Chinese LOD and several specific difficulties (Chinese characters comparing and Web accessing for example) have been bridged over. Furthermore, we have a long-term plan on improving and expanding present CLOD:

- Firstly, several Chinese non-encyclopedia data sources will be accommodated in our knowledge. Wide domain coverage is the advantage of encyclopedia, but some domain-specific knowledge base, such as 360buy[13], Taobao[14] and Douban[15], can supplement more accurate descriptions. A blueprint of Chinese Linking Open Data is illustrated in Figure 6[16].
- The second direction we are considering is improving instance matching strategies. Not only boosting precision and recall of mapping discovering within CLOD, but also augmenting the high-quality entity dictionary to link more Chinese resources to the English ones within Linking Open Data. Meanwhile, necessary evaluations of matching quality will be provided.

[12] `http://www.franz.com/agraph/allegrograph/`
[13] A Chinese language B2C e-Business site, http://www.360buy.com/
[14] A Chinese language C2C e-Business site, http://www.taobao.com/
[15] The largest Chinese online movie and book database, http://www.douban.com/
[16] All sites mentioned in this figure have rights and marks held by their respective owners.

When matching quality is satisfactory enough, we will use a single constant identifier scheme instead of current source-oriented ones.

– Another challenge is refining extracted properties and building a general but consistent ontology automatically. This is an iterative process: initial refined properties are used for ontology learning, and the learned preliminary ontology can help abandon inaccurate properties in return. This iteration will reach the termination condition if results are convergent.

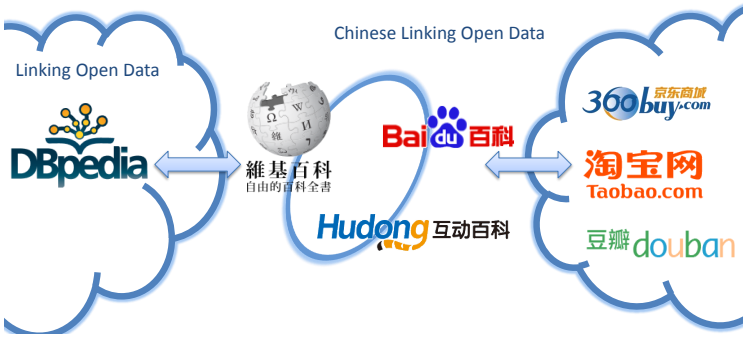

Fig. 6: A Blueprint for Chinese Linking Open Data

References

1. Auer, S., Lehmann, J.: What Have Innsbruck and Leipzig in Common? Extracting Semantics from Wiki Content. In: Franconi, E., Kifer, M., May, W. (eds.) ESWC. Lecture Notes in Computer Science, vol. 4519, pp. 503–517. Springer (2007)
2. Berrueta, D., Phipps, J.: Best Practice Recipes for Publishing RDF Vocabularies. W3C Working Group Note (August 2008), `http://www.w3.org/TR/2008/NOTE-swbp-vocab-pub-20080828/`
3. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. Int. J. Semantic Web Inf. Syst. 5(3), 1–22 (2009)
4. 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)
5. Dean, J., Ghemawat, S.: MapReduce: Simplified Data Processing on Large Clusters. In: OSDI. pp. 137–150 (2004)
6. Duerst, M., Suignard, M.: Internationalized Resource Identifiers (IRIs). proposed standard 3987 (January 2005)
7. Fu, B., Brennan, R., O'Sullivan, D.: Cross-Lingual Ontology Mapping - An Investigation of the Impact of Machine Translation. In: Gómez-Pérez, A., Yu, Y., Ding, Y. (eds.) ASWC. Lecture Notes in Computer Science, vol. 5926, pp. 1–15. Springer (2009)

8. Hogan, A., Harth, A., Passant, A., Decker, S., Polleres, A.: Weaving the pedantic web. In: 3rd International Workshop on Linked Data on the Web (LDOW2010) (2010)
9. Jain, P., Hitzler, P., Sheth, A.P., Verma, K., Yeh, P.Z.: Ontology Alignment for Linked Open Data. In: Patel-Schneider et al. [15], pp. 402–417
10. Levenshtein, V.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10(8), 707–710 (1966)
11. 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)
12. Ngai, G., Carpuat, M., Fung, P.: Identifying Concepts Across Languages: A First Step towards a Corpus-based Approach to Automatic Ontology Alignment. In: COLING (2002)
13. Nikolov, A., Uren, V.S., Motta, E., Roeck, A.N.D.: Integration of semantically annotated data by the knofuss architecture. In: Gangemi, A., Euzenat, J. (eds.) EKAW. Lecture Notes in Computer Science, vol. 5268, pp. 265–274. Springer (2008)
14. Parundekar, R., Knoblock, C.A., Ambite, J.L.: Linking and building ontologies of linked data. In: Patel-Schneider et al. [15], pp. 598–614
15. Patel-Schneider, P.F., Pan, Y., Hitzler, P., Mika, P., 0007, L.Z., Pan, J.Z., Horrocks, I., Glimm, B. (eds.): The Semantic Web - ISWC 2010 - 9th International Semantic Web Conference, ISWC 2010, Shanghai, China, November 7-11, 2010, Revised Selected Papers, Part I, Lecture Notes in Computer Science, vol. 6496. Springer (2010)
16. Raggett, D., Hors, A.L., Jacobs, I.: HTML 4.01 Specification - Appendix B: Performance, Implementation, and Design Notes. W3C Recommendation (December 1999), `http://www.w3.org/TR/html4/appendix/notes.html`
17. Raimond, Y., Sutton, C., Sandler, M.: Automatic interlinking of music datasets on the semantic web. In: Proceedings of the 1st Workshop about Linked Data on the Web (LDOW2008) (2008)
18. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. In: Williamson, C.L., Zurko, M.E., Patel-Schneider, P.F., Shenoy, P.J. (eds.) WWW. pp. 697–706. ACM (2007)
19. Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Discovering and Maintaining Links on the Web of Data. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) International Semantic Web Conference. Lecture Notes in Computer Science, vol. 5823, pp. 650–665. Springer (2009)
20. Zhao, J.: Publishing Chinese medicine knowledge as Linked Data on the Web. Chinese Medicine 5(1), 1–12 (2010)