

# Extraction of RDF Dataset from Wikipedia Infobox Data

Jimmy K. Chiu, Thomas Y. Lee, Sau Dan Lee, Hailey H. Zhu, David W. Cheung

*The University of Hong Kong, Hong Kong*

*{khchiu,ytlee,sdlee,hzyhu,dcheung}@cs.hku.hk*

## ABSTRACT

This paper outlines the cleansing and extraction process of infobox data from Wikipedia data dump into Resource Description Framework (RDF) triplets. The numbers of the extracted triplets, resources, and predicates are substantially large enough for many research purposes such as semantic web search. Our software tool will be open-sourced for researchers to produce up-to-date RDF datasets from routine Wikipedia data dumps.

## 1. INTRODUCTION

### 1.1 Motivation

Resource Description Framework (RDF) is recommended by W3C as a solution for representing Internet resources in semantic web [2]. Many research studies have been done on semantic web searching or efficient retrieval of RDF data. However, we lack good RDF datasets for research purposes because many existing RDF datasets have some of the following problems. First, the some datasets do not contain a large number of triplets, resources or predicates for experiments of datamining techniques on very large databases. For example, the Jamendo RDF dataset[8] contains only around 1M triplets and the Barton dataset[7] contains only 221 predicates. Second, some datasets require specific domain knowledge to comprehend, which makes human interpretation of experimental results difficult. For example, the Uniprot dataset[3] requires researchers to have life science knowledge. Third, many datasets are static, which cannot represent continuous changes of web resources and their relationships.

Wikipedia is an online encyclopedia edited by the public. Although the content format of a Wikipedia page (wiki page) is essentially unstructured, many pages still have pieces of structured information called *infobox*, which can be transformed into meaningful RDF data. While the Wikipedia data are easy to understand by common sense, they are also regularly archived into database dumps for making evolving datasets. Therefore, it is useful to transform Wikipedia infobox data into an RDF dataset for benchmarking and experimenting database techniques. This paper describes the tool we have developed to extract infobox data from Wikipedia database dumps into RDF datasets.

### 1.2 Wikipedia Data

Wikipedia is regularly backed up into dump files[4]. One format of these dump files is in XML, as shown in Fig. 1. The content of a single page is put inside a `<page>` tag. For


```
<page>
  <title>Tsing Ma Bridge</title>
  <id>91180</id>
  <revision>
    <id>160476218</id>
    <timestamp>2007-09-26T14:40:42Z</timestamp>
    <contributor>
      <username>Sameboat</username>
      <id>953247</id>
    </contributor>
    <text>...wikitext here...</text>
  </revision>
</page>
```

Figure 1: XML fragment segment from Wikipedia dump for the “Tsing Ma Bridge” wiki page

a wiki page, the title is given by the `<title>` element and the source code (wikitext) is given by the `<text>` element.

Although wiki pages are written largely in the loosely-structured WikiMedia markup language[5], many wiki pages have a section of structured data called *infobox*. An infobox is a collection of name-value pairs, each containing an infobox field and a field value. An infobox field describes an attribute about the wiki page. Each infobox has a name, which is associated with one infobox template. The infobox template defines a *suggested* list of fields (attributes) that can be defined on the wiki pages describing the same class of objects or concepts. For example, Fig. 2 shows a wikitext fragment that specifies the infobox data for the “Tsing Ma Bridge” wiki page. This infobox has a name **Bridge** and contains different fields, e.g., **bridge\_name**, **carries**, **width**, etc. These fields are defined in the **bridge** infobox template and shared by all infoboxes about bridges. Each field specifies an attribute of the object or concept described by the wiki page. For example, the **bridge\_name** is “Tsing Ma Bridge”. The value of a field may contain double-square-bracketed *interwiki links* to other pages. For example, the field **locale** contains two interwiki links `[[Ma Wan Channel]]` and `[[Ma Wan|Ma Wan Island]]` pointing to wiki pages titled “Ma Wan Channel” and “Ma Wan” respectively. Note that “Ma Wan Island” is the label of the latter link for display.

The aim of our software tool is to convert all infobox data in a Wikipedia dump into an RDF dataset. In each RDF triplet, the subject represents a wiki page, the predicate represents an infobox field, and the object represents a literal value or a wiki page. We plan to open-source our tool for researchers to prepare RDF datasets from Wikipedia dumps.

<div> <div>Tsing Ma Bridge</div>  </div>	
Tsing Ma Bridge at night	
Official name	Tsing Ma Bridge
Carries	6 lanes of roadway (upper) 2 MTR rail tracks, 2 lanes of roadway (lower)
Crosses	Ma Wan Channel
Locale	Ma Wan Island and Tsing Yi Island
Design	Double-decked suspension bridge
Width	41 metres (135 ft)
Longest span	1,377 metres (4,518 ft)
Vertical clearance	62 metres (203 ft)
Opening date	April 27, 1997
Toll	HK\$30 (cars)
Coordinates	<span><span><span><span><span>22°21′05″N</span> <span>114°04′27″E</span></span></span><span><span>﻿</span> / <span>﻿</span></span><span><span>22.35139°N 114.07417°E</span><span><span>﻿</span> / <span>22.35139; 114.07417</span></span></span></span></span>

**Figure 2: The infobox data for the “Tsing Ma Bridge” wikipedia**

## 2. EXTRACTION PROCEDURE

Our proposed extraction procedure consists of three parts: *infobox data extraction*, *data cleansing*, and *conversion to RDF triplets*.

### 2.1 Infobox Data Extraction

The structured infobox data in a wikipedia are embedded in the wikitext of the wikipedia. Thus, it is better to extract only infobox data from the dump so that the data can be processed more efficiently in the subsequent operations. The original dump is parsed to produce another XML file that contains mainly the page name and infobox data. Fig. 3 shows the structure of the extracted XML output file for the page shown in Fig. 1. Again, each `<page>` element contains the contents of a page. The page title is inside the `<title>` tag. The infobox data are structured inside the `<infobox>` tag. The `<name>` tag inside contains the infobox name. Each name-value pair is enclosed with the `<entry>` tag where the infobox field name is tagged `<property>` and the field value is tagged `<value>`. The size of the extracted XML file is just 3.9GB, which is much smaller compared to the dump (21.7GB uncompressed for the snapshot dated 2009-06-18). The size reduction is 82%.

### 2.2 Data Cleansing

Some WikiMedia markups may appear inside an infobox; most of them are used to formatting infobox data to enhance presentation to human readers, e.g., `'''bold'''` is used to embolden the enclosed text. Certain HTML tags are also allowed such as the line break `<br>`.

For historical and conventional reasons, some infobox templates of different names are in fact identical and share the same set of fields. In this case, infobox template of name *A* can be redirected to another infobox template of name *B* where *A* is treated as an alias to *B*. An example is the infobox template “Prime Minister” which is redirected to another infobox template “officeholder”. In our approach, a predicate is formed by concatenating the infobox name and the infobox field name, where the infobox name is used to qualify the infobox field name. Our tool will follow these infobox template redirections in order to replace the infobox

```
{{Infobox Bridge
|bridge_name= Tsing Ma Bridge
|image=Tsing Ma Bridge 2008.jpg
|caption=Tsing Ma Bridge at night
|official_name= Tsing Ma Bridge
|also_known_as=
|carries= 6 lanes of roadway (upper)
<br>2 [[MTR]] rail tracks,
2 lanes of roadway (lower)
|crosses= [[Ma Wan Channel]]
|locale= [[Ma Wan|Ma Wan Island]]
and [[Tsing Yi Island]]
|design= Double-decked
[[suspension bridge]]
|mainspan= {{convert|1377|m|ft|0}}
|width= {{convert|41|m|ft|0}}
|clearance= {{convert|62|m|ft|0}}
|open= [[April 27]], [[1997]]
|toll= HK$30 (cars)
|coordinates= {{coord|22|21|05|N|
114|04|27|E|region:HK_type:landmark}}
}}
```

```
<mediawiki>
...
<page>
<title>Tsing Ma Bridge</title>
<id>91180</id>
<infobox>
<name>Bridge</name>
<entry>
<property>bridge_name</property>
<value>Tsing Ma Bridge</value>
</entry>
...
<entry>
<property>also_known_as</property>
<value></value>
</entry>
<entry>
<property>locale</property>
<value>[[Ma Wan|Ma Wan Island]] and
[[Tsing Yi Island]]</value>
</entry>
...
</infobox>
</page>
...
</mediawiki>
```

**Figure 3: A fragment in the infobox data XML file**

alias names by their canonical name.

Our data cleansing process consists of two steps. The first step is to remove all markups for text formatting. It also removes other noise, such as comments written by authors. The second step is to resolve infobox redirections. If an infobox template is redirected to another infobox template using the directive `#REDIRECT target_infobox_template`. This step builds a redirection map and replaces all infobox alias names by their canonical names. Since `<nowiki>` and `<pre>` tags can be used to escape the WikiMedia markups, no cleansing will be done for the content enclosed by these two tags. The following gives some examples of infobox data before and after cleansing.

*HTML comments are removed:*

```
<value>1679 <!-- Ballistics data source --></value>
```

is cleansed to

```
<value>1679</value>
```

*Removal of all HTML tags recognized by Wikipedia; i.e., each `<b>`, `<i>`, `<font>`, etc. is simply discarded while each `<br>`, `<h1>`, `<li>`, `<p>`, etc. is replaced by a space:*

```
<value>Grace Fletcher Webster<br/>
Caroline LeRoy Webster</value>
<value>2.27 km<sup>2</sup></value>
```

are cleansed to:

```
<value>Grace Fletcher Webster Caroline LeRoy Webster</value>
<value>2.27 km2</value>
```

*Removal of WikiMedia markups:*

```
<value>'''Don River''' watershed</value>
<value>__NOGALLERY__ RaymondPremru.jpg</value>
```

are cleansed to:

```
<value>Don River watershed</value>
<value>RaymondPremru.jpg</value>
```

### 2.3 Conversion to RDF Triplets

The cleansed infobox data XML file is then processed. For each <page> fragment, a number of RDF triplets can be generated.

The subject and predicate of any RDF triplet are Uniform Resource Identifiers (URIs) while the object can be either a URI or a string literal. We use <...> to denote a URI and "<...>" to denote a literal. For storage efficiency, we only store the names of the subjects and predicates instead of their full URIs. The RDF triplets are generated from an infobox field (tagged <entry> in the infobox data XML file) using the following rules.

*When the value of the infobox field is empty*, this field is ignored. For example, in Fig. 3, the `also_known_as` field generates no triplet.

*When the value of the infobox field contains text only and no interwiki links*, one triplet is generated. The subject is the wiki page title. The predicate is the concatenation of the infobox name, the symbol “#” and the infobox field name. The object is a literal representing the field value. For example, in Fig. 3, the `bridge_name` field generates the following triplet:

```
<Tsing_Ma_Bridge> <Bridge#bridge_name> "Tsing Ma Bridge" .
```

*When the value of the infobox field contains some interlinks*, one triplet is generated from each distinct interwiki link, and one additional triplet is generated from the infobox value as a whole. For each interwiki link, the subject is the title of the current wiki page; the predicate is the concatenation of the infobox name, “#” and the infobox field name; the object is the URI of the wiki page title in the link. For the additional link, the subject and the predicate are generated as above. The object is a literal transformed from the field value where each link is replaced by the link label or the wiki page title if the label is absent. For example, the `locale` field in Fig. 3 generates the following 3 triplets:

```
<Tsing_Ma_Bridge> <Bridge#locale> <Ma Wan> .
<Tsing_Ma_Bridge> <Bridge#locale> <Tsing Yi Island> .
<Tsing_Ma_Bridge> <Bridge#locale>
"Ma Wan Island and Tsing Yi Island" .
```

Each predicate is formed from an infobox field name qualified with the infobox name, to distinguish two infobox fields with the same name but defined in different infobox templates. For example, the infobox fields with the same name “length” are defined in both “Song” and “UK Bus” infobox templates. The semantics of these two infobox fields are different (i.e., the time duration of a song vs. the physical length of a bus) and are thus represented by two different predicates, i.e., <Song#length> and <UK\_Bus#length>.

### 3. RELATED WORK

DBpedia[1] is another work on RDF data extraction from Wikipedia data dump for query purpose. However, in DBpedia the following issues were observed:

*Predicates of different semantics may use the same URI.* This is because only the infobox field name is used to form a predicate URI, e.g.,

```
<http://dbpedia.org/resource/Amusement_Parks_USA>
<http://dbpedia.org/property/length> "2:29"@en .
```

```
<http://dbpedia.org/resource/Dennis_Trident_3>
<http://dbpedia.org/property/length>
"10.3m, 10.6m, 11.3m or 12m"@en .
```

In the above, although both predicates use the same URI (<http://dbpedia.org/property/length>), the first predicate comes from the infobox “Song” and refers to the time duration of a song, while the second predicate comes from the infobox “UK Bus” and refers to the physical length of a bus. Our approach creates each predicate from a distinct infobox field name qualified with the infobox name.

*Noise is induced from non-infobox data.* DBpedia processes not only infobox data but also other template data, which has generated ambiguous RDF triplets. For example,

```
<http://dbpedia.org/resource/Dell>
<http://dbpedia.org/property/name> ‘‘Dell, Inc.’’@en .
```

```
<http://dbpedia.org/resource/Dell>
<http://dbpedia.org/property/name> ‘‘Texas’’@en .
```

```
<http://dbpedia.org/resource/Dell>
<http://dbpedia.org/property/name> ‘‘Companies’’@en .
```

In the above, the triplets are produced from the wiki page about Dell. The first triplet comes from the infobox. The second and third triplets come from a *portal box* (template) (rather than the infobox) and do not make sense. There are many triplets like this example, which have generated inaccurate semantics in the DBpedia dataset.

### 4. RESULTS

We used the Wikipedia data dump archived on 2009-06-18 for RDF triplets extraction. The triplets extracted were stored in a SQLite database table with columns for subject, predicate, and object respectively. We also imported the infobox RDF dataset provided by DBpedia (version 3.3, produced from the Wikipedia data dump dated 2009-05-20) into a SQLite database with identical table schema for comparison. Table 1 shows some statistics collected from the RDF dataset extracted by our tool and from the DBpedia dataset.

	Our approach	DBpedia
Database file size	2.9GB	11GB
RDF triplet count	20,321,291	39,855,823
Distinct subject count	1,014,251	3,798,547
Distinct predicate count	93,198	49,122
Mean no. of triplets per subject	20.04	10.49
Mean no. of triplets per predicate	218.04	811.36
Mean no. of distinct predicates per subject	14.86	6.99
Mean no. of distinct subjects per predicate	161.67	540.91

**Table 1: Statistics for the extracted infobox dataset and DBpedia dataset**

The DBpedia dataset nearly doubles the number of triplets since it extracts triplets from non-infobox templates. Its number of subjects is also more than tripled of that of our approach because DBpedia creates subjects for nested templates in infobox data. In our approach, one subject is created for each wiki page. We qualify each infobox field name qualified with its infobox name to create one predicate while DBpedia treats each distinct infobox field name (regardless of its infobox name) as one predicate. Therefore, the number of distinct predicates of our approach is almost

twice that of DBpedia. The last two figures are the average number of distinct predicates one subject has, and the average number of subjects sharing a particular predicate. In our approach, every subject represents a page and every predicate represents a unique infobox field, so these two figures can be interpreted as that on average a page has about 15 distinct infobox fields (qualified with its infobox name), and a distinct infobox field appears in about 162 pages.

## 5. CONCLUSION

Since Wikipedia requires little domain knowledge and its size is of the Internet scale, it is a potential data source for semantic web experiments. Our approach produces huge datasets with more predicates than DBpedia. All the subject and object URIs represent Wikipedia pages properly, and each predicate URI uniquely identifies a particular infobox field. Therefore one can easily map a triplet back to an infobox field in a page from which it originates. Also because of the availability of the dump files from regularly updated Wikipedia snapshots, our datasets are more useful for experiments than the others.

The datasets generated from our tool can be applied in a variety of semantic web studies such as semantic web searching and RDF data storage. For example, our datasets can be used in [6] to identify whether a column store is scalable for RDF data with a huge number of predicates. Another example of using our datasets is the semantic search experiments conducted in [9] to obtain the index size and search time for Internet scale data. Currently, we are using the datasets to conduct research on database schema mining for RDF data. We will open source our tool so that researchers can freely use it to produce Wikipedia infobox RDF datasets from Wikipedia dump files for experimental work.

## 6. REFERENCES

- [1] DBpedia. <http://dbpedia.org/>.
- [2] RDF/XML Syntax Specification (Revised). W3C Recommendation. <http://www.w3.org/TR/rdf-syntax-grammar/>.
- [3] Uniprot Data. <http://www.uniprot.org/downloads/>.
- [4] Wikipedia Data Dump. <http://download.wikimedia.org/backup-index.html>.
- [5] Wikipedia Markup Specification. [http://www.mediawiki.org/wiki/Markup\\_spec](http://www.mediawiki.org/wiki/Markup_spec).
- [6] D. J. Abadi, A. Marcus, S. R. Madden, and K. Hollenbach. Scalable Semantic Web Data Management Using Vertical Partitioning. *VLDB*, 2007.
- [7] D. J. Abadi, A. Marcus, S. R. Madden, and K. Hollenbach. Using The Barton Libraries Dataset As An RDF benchmark. MIT-CSAIL-TR-2007-036. Technical report, MIT, 2007.
- [8] C. Bizer, T. Health, D. Ayers, and Y. Raimond. Interlinking Open Data on the Web. *4th European Semantic Web Conference*, 2007.
- [9] H. Wang, K. Zhang, Q. Liu, T. Tran, and Y. Yu. Q2Semantic: A Lightweight Keyword Interface to Semantic Search. *ESWC*, 2008.