A Template for Handling Statistical Data in RDF

Yu Asano[1], Makoto Iwayama[1], Hideaki Takeda[2, 3],
Seiji Koide[2, 3], Fumihiro Kato[2, 3], and Iwao Kobayashi[3]

[1] Hitachi, Ltd., Central Research Laboratory, Tokyo, Japan
{yu.asano.ko, makoto.iwayama.nw}@hitachi.com
[2] National Institute of Informatics, Tokyo, Japan
{takeda, koide, fumi}@nii.ac.jp
[3] Linked Open Data Initiative, Inc., Tokyo, Japan
iwao@scholex.com

Abstract. A template is proposed in this paper for handling the statistical data in the Resource Description Framework (RDF). The governments of many nations publish vast amounts of statistical data in tables, and the statistical data in RDF enables the software to easily process it. However, it is quite difficult for users to browse and edit this data. We have created a template with an original structure that can include the necessary data for RDF. After the necessary data is transferred into the template, the statistical data in RDF can be automatically exported and can be intuitively corrected by modifying the data in the template.

Keywords: Statistics, RDF, RDF Data Cube Vocabulary, Linked Data

1 Introduction

The governments of many nations publish vast amounts of statistical data [1, 2] that can be used for policy prediction, planning, and adjustment by investigating the relations between the relevant data. However, searching for and connecting the relevant data is costly, because most statistical data are published in tables with different structures using different vocabularies. Therefore, there is a strong interest in publishing statistical data in a machine-understandable format using a unified vocabulary. As a standard mean, the World Wide Web Consortium (W3C) proposed the RDF Data Cube Vocabulary [3], which breaks a standard 2D table into a collection of subject-predicate-object triples. The statistical data in RDF using the Data Cube Vocabulary enables for the software to easily process the data.

However, it is not easy for users to browse and edit the statistical data in RDF. Without software, it is hard for us to see the distribution of the data or to collect the necessary data for modification. It is much easier for people to deal with the statistical data in tables than in RDF. Therefore, a user-friendly interface that can handle RDF data is required. In this paper, we propose a template that stays as close to the original table format as possible in order to create this type of interface. Additionally, the necessary data for RDF, some of it actually missing in the original table, can be embedded in this template, and then automatically converted into RDF. For add-

ing/modifying/deleting RDF data, the users can intuitively perform the corresponding operation on the template that is almost exactly the same as that for the original table.

In the next section, we present our template for handling the statistical data in RDF. Then, we discuss the features of our proposed template, the experiments conducted while using our template on real data, and the related works.

2 Proposal

We propose creating a new template from an original table for inputting the necessary information for RDF in order to convert the statistical data in a table into RDF triples. This template features additional rows and columns for the dimensions and measures that are necessary for RDF but not included in the original table. After users add the rows and columns in the template, our conversion software uses the table to automatically produce triples. Since the template maintains the observation cells with the same structure as the original table, the users can browse and edit the data in the same way as they would on the original table. This is convenient because the users are more familiar with the original table than the converted RDF triples.

Fig. 1 is an example of a template produced from [1]Table 1. The basic structure of the original table (Table 1) is maintained in the template (Fig. 1). In this figure, each cell in the rows/columns inserted in step 1 takes on the URI of the upper/left cell that refers to a dimension or measure value. Furthermore, our template has cells for inputting the information that is necessary for RDF but not included in the original table. This is one of main features of our template. Each cell in the row/column inserted in step 2 takes on a URI that refers to a dimension when the corresponding column/row heading takes on the dimension values. For example, "eg:refYear" in the inserted column is the URI of a dimension whose values are years ("2010" and "2005"). This dimension does not appear in the original table. Another example is for measures. In the second row of Fig. 1, we input the URIs of the corresponding measures ("Population" and "Area"). However, when the original table has a single measure, the name of the measure may not be written in the table, but only written in the caption of the table. In such cases, the URI of the measure can be input into the inserted cell, which is marked as "#" in Fig. 1.

Table 1. Populations and areas (measures) by prefecture and year (dimensions).

	Population (people)		Area (km^2)
	2010	2005	2010
Saitama	7,194,556	7,054,382	3,767.92
Chiba	6,216,289	6,056,462	5,081.91
Tokyo	13,159,388	12,576,611	2,102.95
Kanagawa	9,048,331	8,791,587	2,415.86

[1] In Table 1, the populations are extracted from a legislative investigation conducted by the Japanese Ministry of International Affairs and Communications and the areas are from an area investigation conducted by the Ministry of Land, Infrastructure, Transport and Tourism.

			Population (people)		Area (km^2)	
			eg:population		eg:areas	← Step 1
			2010	2005	2010	
		eg:refYear	eg:year-2010	eg:year-2005	eg:year-2010	← Step 1
	eg:refPrefecture	#				← Step 2
Saitama	eg:prefecture-11		7,194,556	7,054,382	3,767.92	
Chiba	eg:prefecture-12		6,216,289	6,056,462	5,081.91	
Tokyo	eg:prefecture-13		13,159,388	12,576,611	2,102.95	
Kanagawa	eg:prefecture-14		9,048,331	8,791,587	2,415.86	
	↑ Step 1	↑ Step 2				

Fig. 1. Creating new template from original Table 1 content.

RDF triples can be automatically produced from the template. For example, the RDF triples for the cell refering to "Tokyo's population in 2010 is 13,159,388" are follows.

```
eg:dataset-02-003 a qb:Observation;
  qb:dataset eg:dataset-02;
  eg:refPrefecture eg:prefecture-13;
  eg:refYear eg:year-2010;
  qb:measureType eg:population;
  eg:population 13159388.
```

3 Discussion

Our template has the following characteristics.

- The necessary data for RDF are embedded into a table. The users or software can easily manage the data without referring to any additional data.
- URIs from the same group are listed in a row/column and their labels are also listed in the adjacent row/column. These characterstics for overview and correspondence enable the users to easily and consistently input the necessary URIs.
- Various kinds of tables can be handled in a uniform way. Particularly for a table with a single measure, our template has a space for the measure that is not contained in the original table.
- The data structure of the original table is maintained. Users can edit the statistical data just as they would on the original table, and the editing results are immediately reflected on RDF.

As an experiment, we converted six statistical tables published by the Japanese Ministry of Economy, Trade and Industry into RDF triples using our template. We could successfully convert them into about 3 million triples. However, three problems arose during the conversion processes. First, there was a case where a physical table (a spreadsheet) had multiple logical tables for easiness of browsing by users. In these cases, we made a template for each logical table. Second, one table included comments in the rows/columns that were not necessary in RDF. This problem could be

solved by enabling the users to specify whether or not each row/column is necessary. Finally, two tables had different dimensions for the same row/column. For example, a broader classification and a narrower classification, such as for prefectures and municipalities, were in the same column. In this case, these classifications needed to be handled as different dimensions. In order to do so, we extended our template to insert multiple rows/columns to distinguish between these dimensions.

Several tools have been developed to clean up data in tables or to convert them into RDF. The tools [4-6] are powerful but not designed for handling RDF data in Data Cube Vocabulary. The tool [7] is for converting the structural data in RDF by using the RDF Data Cube Vocabulary. Users import the original statistical table into the tool and, on the tool, add the necessary data for conversion into RDF. On the other hand, our template is made by inserting rows/columns to the original table. On the template, the users can input the necessary data for RDF and handle the statistical data in the same way as they would on the original table. Furthermore, the users can use any spreadsheet software they want for handling our template because the template is a simple table format and does not need additional data other than the table.

4 Conclusion and Future Work

We proposed a template for handling statistical data in RDF in this paper. Our template is suitable for users to browse and edit data because it maintains as much of the original table structure as possible. Additionally, it can contain the data necessary for RDF that does not appear in the original table. After the users input the data for RDF, the statistical data in RDF can automatically be exported. When the users need to modify the data in RDF, they can intuitively edit it in the proposed template. In the future, we intend to expand the application range of our proposal by introducing a function that automatically provides the necessary data.

References

1. Takeda, H., Kato, F., Koide, S., Matsumura, F., Ohmukai. I., Kobayashi, I., Iwayama, M., Asano, Y., Hamasaki, M.: Presentation of Statistical Data and their Relationship as LOD. 27th JSAI, N4-OS-10b-6 (2013)
2. Hoefler, P., Granitzer, M., Veas, E., Seifert, C.: Linked Data Query Wizard: A Novel Interface for Accessing SPARQL Endpoints. In: LDOW 2014, http://ceur-ws.org/Vol-1184/ldow2014_paper_06.pdf, CEUR-WS.org (2014)
3. Cyganiak, R., Reynolds, D. (eds.): The RDF Data Cube Vocabulary. W3C Recommendation 16 January 2014, http://www.w3.org/TR/vocab-data-cube/, World Wide Web Consortium (2014)
4. Open Refine. http://openrefine.org/, (accessed 2014-06-30)
5. RDF Refine. http://refine.deri.ie/, (accessed 2014-06-30)
6. Han, L., Finin, T., Parr, C., Sachs, J., Joshi, A.: RDF123: From Spreadsheets to RDF. In: ISWC 2008, LNCS, vol. 5318, pp 451-466. Springer, Heidelberg (2008)
7. Salas, P.E.R, Mota, F.M.D, Martin, M., Auer, S., Breitman, K., Casanova, M. A.: Publishing Statistical Data on the Web. In: IEEE Sixth International Conference on Semantic Computing, pp 285-292. IEEE Press, New York (2012)