Generating LOD from Web: A Case Study on Building Integrated Museum Collection Data

Fuyuko Matsumura[1], Fumihiro Kato[2], Tetsuro Kamura[3,4], Ikki Ohmukai[1,3], and Hideaki Takeda[1,3]

[1] National Institute of Informatics, {fuyuko, i2k, takeda}@nii.ac.jp
[2] Research Organization of Information and Systems, fumi@nii.ac.jp
[3] The Graduate University for Advanced Studies
[4] Tokyo University of the Arts, kamura.tetsuro@noc.geidai.ac.jp

Abstract. In this paper, a workflow was developed to enable efficient data extraction from web and integration them with the cooperation of web developers and data professionals who specialized in a certain field. This paper introduces how we applied the workflow to build Linked Data for "LODAC Museum", a dataset on museum collection data in Japan.

1 Introduction

The Linked Open Data for ACademia (LODAC)[1] project has started to integrate and publish academic information of Japan as Linked Open Data (LOD) datasets to enhance interdisciplinary sharing and reuse of various datasets. Linked Data Integration Framework (LDIF)[2] aims to automatically integrate and map data to common vocabularies using RDF data included in websites; however, most of websites do not have RDF data and how to generate LOD from websites written in HTML is important to increase useful LOD. Therefore, a workflow to generate LOD from web is developed in this paper. It is characterized by separating of web programming and metadata mapping. This paper introduces the workflow to generate LOD consisting of key-value pairs taking "LODAC Museum"[1], a dataset on museum collection data in Japan as an example.

2 The Workflow for Generation of Linked Data on Museum Collection

Museum collection data are transformed into LODAC Museum by the following steps. The original collection data are obtained by web scraping for museum sites and translated into Resource Description Framework (RDF).

1. Extracting data from web pages: Collect key-value pairs data from web pages of artworks in multiple sources using Apache Nutch and Solr.
2. Mapping vocabularies: Map keys in extracted data to a common schema by museum professionals.
3. Integrating unique items: Identify the same items (artwork, creator, museum) across museum collections and associate them to single identifiers.

[1] http://lod.ac

4. Publishing: Publish data as Linked Data with permalinks that work as identifiers for items, accessible through the SPARQL endpoint.
5. Versioning: Store snapshots of crawled original HTML files, extracted key-value pairs, and transformed RDF files into the git repository to enable to trace the process of data transformation and retry data generation from original files.

The main feature of our system is that the key-value pairs extraction from museum websites and metadata mapping are completely divided and can be independently processed, because it is assumed that two different types of professionals are needed. Prior to the data extraction from HTML, it is needed to find parts which include key-value pairs from HTML and write XPath for the parts into the setting files of Nutch by web developers. On the other hand, when extracted key-value pairs are mapped to RDF, it is desired that the professionals in the target domain handle metadata mapping because vocabularies and sentences of websites are possibly quite hard to understand for people who are not museum professionals.

When the construction of LODAC Museum was started, this workflow had not implemented yet and scraped data were manually transformed into RDF in programs. After the workflow was implemented, it additionally extracted 578,500 key-value pairs from websites of 38 institutions. Since some of them were mapped to multiple properties, those data were converted into 890,588 RDF triples according to the mapping rule, and appended to the RDF store.

3 Discussion

Some methods have been proposed to generate RDF from websites and some of them are trying to automatically extract desired information from websites using machine learning techniques. However, expressions of collection data are different between each website of museums in Japan. For instance, when collection data is expressed as a table, some sites display the creator's name with the title of a work in the same cell while others display the birth year with the title. Moreover, each museum uses different terms to describe each attribute. Therefore, it seems suitable that museum professionals manually implement mapping rules and those generated rules can also be used as training data for automatic extraction of museum collection data from websites in the future.

Furthermore, we plan to develop a user interface for metadata mapping and semi-automatic data integration based on text matching; moreover, further verification of our data conversion system are needed using other datasets.

References

1. Kamura, T., Takeda, H., Ohmukai, I., Kato, F., Takahashi, T., and Ueda, H.: Study Support and Integration of Cultural Information Resources with Linked Data, Proc. of the 2nd International Conference on Culture and Computing, pp.177–178 (2011)
2. Schultz, A., Matteini, A., Isele, R., Bizer, C. and Becker, C.: LDIF - Linked Data Integration Framework, Proc. of the 2nd International Workshop on Consuming Linked Data (COLD2011) (2011)