Publishing Data from the Smithsonian American Art Museum as Linked Open Data*

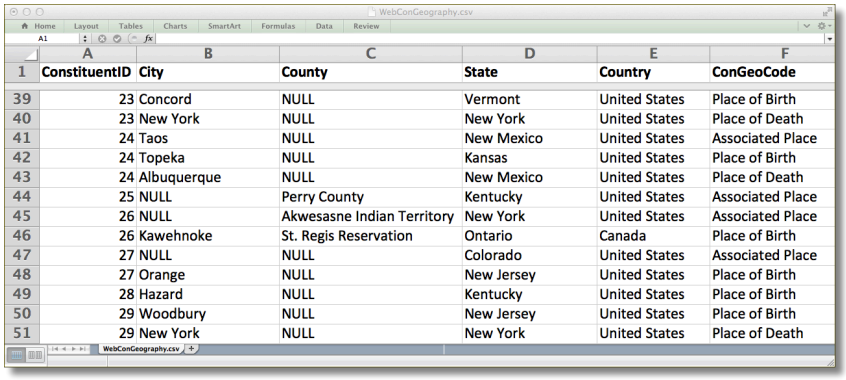
Craig A. Knoblock[1], Pedro Szekely[1], Shubham Gupta[1], Animesh Manglik[1], Ruben Verborgh[2], Fengyu Yang[3], and Rik Van de Walle[2]

[1] University of Southern California
Information Sciences Institute and Department of Computer Science, USA
[2] Multimedia Lab – Ghent University – iMinds, Belgium
[3] Nanchang Hangkong University, Nanchang, China
{knoblock,pszekely,shubhamg}@isi.edu, manglik@usc.edu,
{ruben.verborgh,rik.vandewalle}@ugent.be, frueyang@gmail.com

Abstract. Museums around the world have built databases with metadata about millions of objects, the people who created them, and the entities they represent. This data is stored in proprietary databases and is not readily available for use. Recently, museums embraced the Semantic Web as a means to make this data available to the world, but the experience so far shows that publishing museum data to the Linked Data cloud is difficult: the databases are large and complex, the information is richly structured and varies from museum to museum, and it is difficult to link the data to other datasets. We have been collaborating with the Smithsonian American Art Museum to create a set of tools that allow museums and other cultural heritage institutions to publish their data as Linked Open Data. In this demonstration we show the end-to-end process of starting with the original source data, modeling the data with respect to a ontology of cultural heritage data, linking the data to DBpedia, and then publishing the information as Linked Data.

Recently there have been a number of efforts to publish metadata about the objects in museums as Linked Open Data (LOD). Some notable efforts include the Euopeana project [2], which published data on 1,500 of Europe's museums, libraries, and archives, the Amsterdam Museum [1], which published data on 73,000 objects, and the LODAC Museum [4], which published data from 114 museums in Japan. Despite the many recent efforts, there are still significant challenges in publishing data about artwork to the Linked Data cloud. In particular, the past work requires a user to manually convert their data into RDF and there has been almost no work on actually linking the cultural heritage data across sources. The contributions of this work are that we developed a very quick and efficient method for mapping museum data to a cultural heritage ontology and we created tools for linking and validating the links to other sources [5].

* This research was funded by the Smithsonian American Art Museum. A demonstration video is available at: http://youtu.be/1Vaytr09H1w



	A	B	C	D	E	F
1	ConstituentID	City	County	State	Country	ConGeoCode
39	23	Concord	NULL	Vermont	United States	Place of Birth
40	23	New York	NULL	New York	United States	Place of Death
41	24	Taos	NULL	New Mexico	United States	Associated Place
42	24	Topeka	NULL	Kansas	United States	Place of Birth
43	24	Albuquerque	NULL	New Mexico	United States	Place of Death
44	25	NULL	Perry County	Kentucky	United States	Associated Place
45	26	NULL	Akwesasne Indian Territory	New York	United States	Associated Place
46	26	Kawehnoke	St. Regis Reservation	Ontario	Canada	Place of Birth
47	27	NULL	NULL	Colorado	United States	Associated Place
48	27	Orange	NULL	New Jersey	United States	Place of Birth
49	28	Hazard	NULL	Kentucky	United States	Place of Birth
50	29	Woodbury	NULL	New Jersey	United States	Place of Birth
51	29	New York	NULL	New York	United States	Place of Death

Fig. 1. Source data from the Smithsonian

Datasets. In this demonstration we will use datasets extracted from Smithsonian American Art Museum collection management database. Figure 1 shows a sample of raw data about artists. One of the challenges in mapping this particular table to RDF is that the meaning of each row is encoded in the last column of the table, which shows whether the data specifies the place of birth, place of death, or an associated place. Each row must be mapped to a different property in the ontology.

Mapping the Data to RDF. The first step in the process is to model the raw data using an ontology of cultural heritage data. We developed a tool called Karma [3], which semi-automatically builds a semantic description of a data source using machine learning techniques. Karma makes it possible to quickly model the dataset shown in Figure 1. The resulting model, shown in Figure 2, illustrates several capabilities needed for real datasets, such as customizing the URL generation and generating different properties for different rows in a table. After modeling the data, users can ask Karma to publish the model as R2RML and publish the data as RDF.

Linking to External Datasets. Once the RDF is published, the next step is to link the information about entities, including artists and locations, to the corresponding entities in other sources. In particular we link artists to people in DBpedia and locations to the corresponding locations in Geonames. The name fields of the artists, possibly augmented with other identifying information such as birth year, is sent to a reconciliation service that has a fuzzy index of DBpedia information. Figure 3 shows a user invoking the reconciliation service on the Person class, and it shows the resulting links and scores for each match to DBpedia. The interface also allows a user to verify each of the links and to drill down into the individual data sources to determine whether the links are correct.

Publishing the Linked Data Once the links have been verified, the information is then published and made available as Linked Open Data. We created a user-friendly version of Pubby[4] that describes the content and shows the images

[4] `http://wifo5-03.informatik.uni-mannheim.de/pubby/`

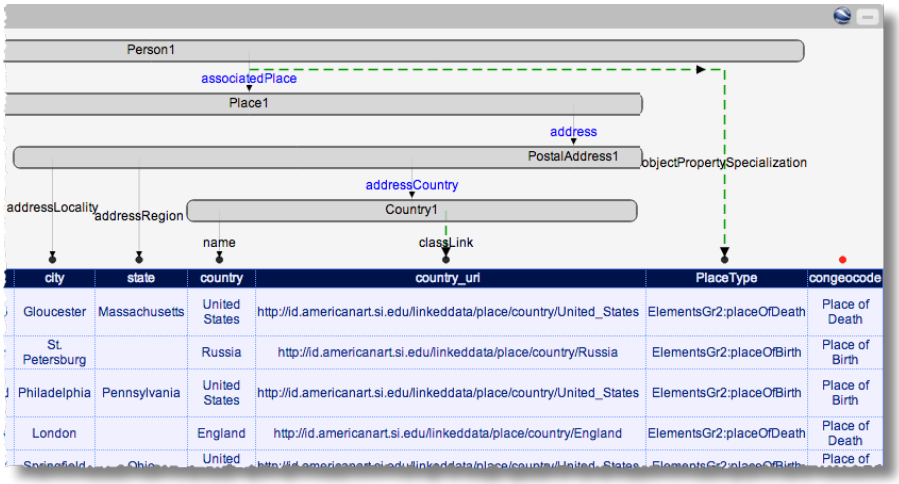


Fig. 2. Source model interactively generated using Karma

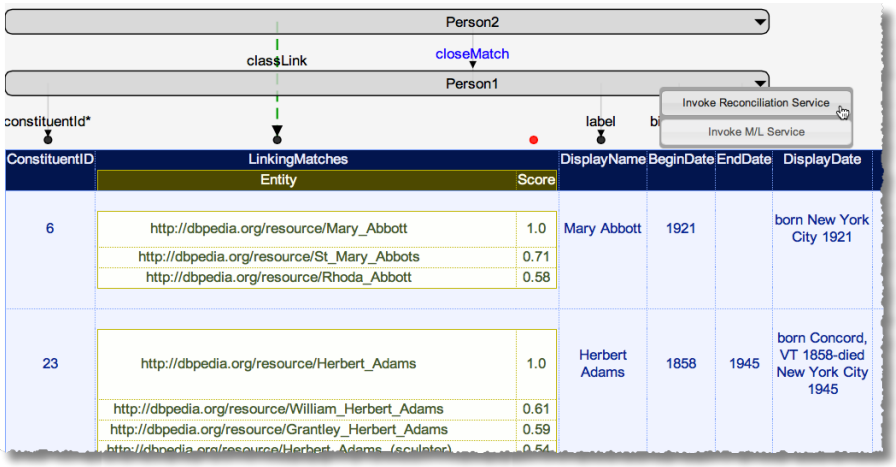


Fig. 3. Screen shot of the interface for linking the museum data to DBPedia

and labels directly on the page to make the resulting linked data more readable. Figure 4 shows a screen shot of the published data from the same data source shown in Figure 1.

Using the Linked Data Once the linking is complete, the Linked Data can also be used to augment other sources of data. In our project with the Smithsonian, they found that one use of the Linked Data is to augment their current Web pages with the additional information available from Wikipedia and the New York Times, which follows directly from the linking of their data about artists

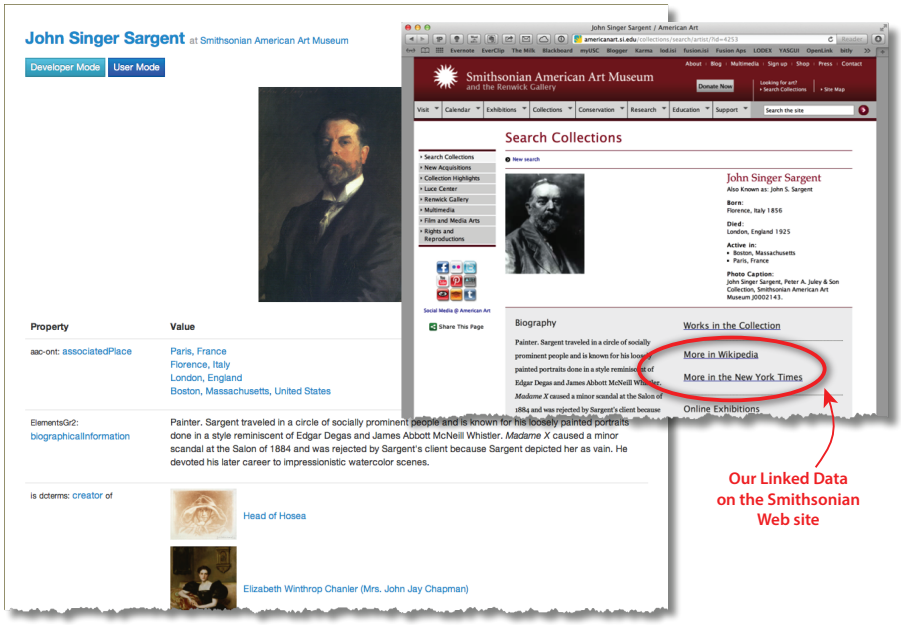


Fig. 4. Linked data viewed in a user-friendly version of Pubby and page of the Smithsonian Web site that includes the Linked Data

to DBpedia. Figure 4 shows screen shots of our user friendly version of Pubby and the Smithsonian American Art Museum's Web site with the additional links generated directly from the Linked Data that we produced for them. The Linked Data can also be used to curate the Smithsonian's own database, create virtual online museums, or create new applications that build on the Linked Data.

References

1. Boer, V., Wielemaker, J., Gent, J., Hildebrand, M., Isaac, A., Ossenbruggen, J., Schreiber, G.: Supporting Linked Data Production for Cultural Heritage Institutes: The Amsterdam Museum Case Study. *Lecture Notes in Computer Science*, pp. 733–747. Springer Berlin Heidelberg (2012).
2. Haslhofer, B., Isaac, A.: The Europeana Linked Open Data Pilot. *Proceedings of the International Conference on Dublin Core and Metadata Applications*, (2011)
3. Knoblock, C., Szekely, P., Ambite, J.L., Goel, A., Gupta, S., Lerman, K., Muslea, M., Taheriyan, M., Mallick, P.: Semi-Automatically Mapping Structured Sources into the Semantic Web. *Proceedings of the 9th Extended Semantic Web Conference* (2012)
4. Matsumura, F., Kobayashi, I., Kato, F., Kamura, T., Ohmukai, I., Takeda, H.: Producing and Consuming Linked Open Data on Art with a Local Community. *Proceedings of the Third International Workshop on Consuming Linked Data* (2012)
5. Szekely, P., Knoblock, C.A., Yang, F., Zhu, X., Fink, E., Allen, R., Goodlander, G.: Connecting the Smithsonian American Art Museum to the Linked Data Cloud. *Proceedings of the 10th Extended Semantic Web Conference* (2013)