TECHNOLOGY AND SCALING

PUBLISHING THE DATA OF THE SMITHSONIAN AMERICAN ART MUSEUM TO THE LINKED DATA CLOUD

PEDRO SZEKELY, CRAIG A. KNOBLOCK, FENGYU YANG, ELEANOR E. FINK, SHUBHAM GUPTA, RACHEL ALLEN AND GEORGINA GOODLANDER

Abstract *Museums around the world have built databases with metadata about millions of objects, their history, 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. This paper describes the process of publishing the data of the Smithsonian American Art Museum (SAAM). We describe the database-to-RDF mapping process, discuss our experience linking the SAAM dataset to hub datasets such as DBpedia and the Getty Vocabularies, and present our experience in allowing SAAM personnel to review the information to verify that it meets the high standards of the Smithsonian. Using our tools, we helped SAAM publish high-quality linked data of their complete holdings: 41,000 objects and 8,000 artists.*

Keywords: Semantic Web, Linked Data, Resource Description Framework (RDF), Web Ontology Language (OWL), Entity resolution, Record Linking, Schema Mapping, Data Curation, Extraction, transformation and loading (ETL); Cultural Heritage; Museum; Collection Management Software

International Journal of Humanities and Arts Computing 8 (2014) Supplement: 152–166
DOI: 10.3366/ijhac.2014.0104

www.euppublishing.com/ijhac

INTRODUCTION

Recently, several efforts seek to publish metadata about the objects in museums as Linked Open Data (LOD). LOD provides an approach to publishing data in a standard format (called RDF) using a shared terminology (called a domain ontology) and linked to other data sources. The linking is particularly important because it relates information across sources, breaks down data silos and enables applications that provide rich context.

Some notable LOD efforts include the Euopeana project[1], which published data on 1,500 of Europe's museums, libraries, and archives, the Amsterdam Museum[2], which published data on 73,000 objects, and the LODAC Museum[3], which published data from 114 museums in Japan. Despite the many recent efforts, significant challenges remain. Mapping the data of a museum to linked data involves three steps:

1. Map the Data to RDF: The first step is to map the metadata about works of art into RDF. This involves selecting or writing a domain ontology with standard terminology for works of art and converting the data to RDF according to this ontology. De Boer et al.[2] note that the process is complicated because many museums have richly-structured data including attributes that are unique to a particular museum, and the data is often inconsistent and noisy because many individuals have maintained the data over a long period of time. In past work, the mapping is typically defined using manually written rules or programs.
2. Link to External Sources: Once the data is in RDF, the next step is to find the links from the metadata to other repositories, such as DBpedia or GeoNames. In previous work, developers define a set of rules for performing the mapping. Because the problem is difficult, the number of links in past work is actually quite small as a percentage of the total set of objects that have been published.
3. Curate the Linked Data: The third step is to curate the data to ensure that both the published information and its links to other sources within the LOD are accurate. Because curation is so labor intensive, this step has been largely ignored in previous work and as a result links are often inaccurate.

Our goal is to develop technology to allow museums to map their own data to LOD. The contribution of this paper is an end-to-end approach that maps museum source data into high quality linked data. In particular, we describe the process of mapping the metadata that describes the 41,000 objects of the Smithsonian American Art Museum (SAAM). This work builds on our previous work on a system called Karma for mapping structured sources to RDF. In terms of linking, we found that mapping the entities, such as artist names, to DBpedia could not be easily or accurately performed using existing tools, so we developed a specialized mapping approach to achieve high accuracy. Finally, to

ensure that the Smithsonian publishes high quality linked data, we developed a curation tool that allows museum staff to easily review and correct any errors in the automatically generated links to other sources.

In the remainder of this paper, we describe our approach and present the approach to mapping, linking, and curating museum data. For each of these topics, we describe our approach and evaluate its effectiveness. We then compare our work to previous work and conclude with a discussion of the contributions and future work.

MAPPING THE DATA TO RDF

In this section we describe our approach to mapping the data of the Smithsonian American Art Museum to Linked Open Data. This includes the selection of a domain ontology and then relating this data to the domain ontology to build the RDF.

Building a Museum Domain Ontology

To create an ontology for the SAAM data, we start with the Europeana Data Model (EDM[4]), the metamodel used in the Europeana project[5] to represent data from Europe's cultural heritage institutions. EDM is a comprehensive OWL ontology that reuses terminology from several widely-used ontologies: SKOS[6] for the classification of artworks, artist and place names; Dublin Core[7] for the tombstone data; FOAF[8] and RDA Group 2 Elements[9] to represent biographical information; ORE[10] from the Open Archives Initiative, used by EDM to aggregate data about objects.

The SAAM ontology[11] (Figure 1) extends EDM with subclasses and subproperties to represent attributes unique to SAAM (e.g., identifiers of objects) and incorporates classes and properties from schema.org[12] to represent geographical data (city, state, country). We chose to extend EDM because this maximizes compatibility with a large number of existing museum LOD datasets.

One of the most challenging tasks in the project was selecting and extending the ontologies. We considered EDM and CIDOC CRM[13]; both are large and complex ontologies, but neither fully covers the data that we need to publish. We needed vocabularies to represent biographical and geographical information, and there are many to choose from. Following the lead of the Amsterdam Museum[2], we used RDA Group 2 Elements for the biographical information. We didn't find guidance for representing the geographical information in the cultural heritage community so we selected schema.org as it is a widely used vocabulary. Our extensions (shown in boldface/shaded in Figure 1) are subclasses or subproperties of entities in the ontologies we reuse.

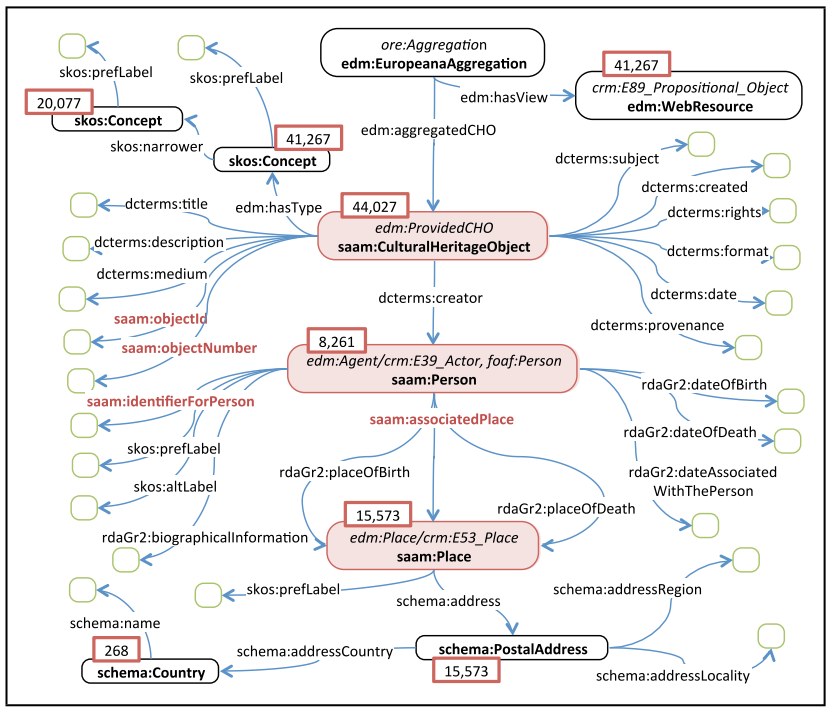

Figure 1. The SAAM ontology. Named ovals represent classes, un-named green ovals represent literals, arcs represent properties, boxes contain the number of instances generated in the SAAM dataset, italicized text shows superclasses, all properties in the SAAM namespace are subproperties of properties in standard vocabularies.

USING KARMA TO MAP THE SAAM DATA TO RDF

In previous work[14], we developed Karma, a tool to map structured data to RDF according to an ontology of the user's choice. The goal is to enable data-savvy users (e.g., spreadsheet users) to do the mapping, shielding them from the complexities of the underlying technologies (SQL, SPARQL, graph patterns, XSLT, XPath, etc). Karma addresses this goal by automating significant parts of the process, by providing a visual interface (Figures 2 & 3) where users see the Karma-proposed mappings and can adjust them if necessary, and by enabling users to work with example data rather than just schemas and ontologies. The Karma approach to map data to ontologies involves two interleaved steps: one, assignment of semantic types to data columns and two, specification of the relationships between the semantic types.

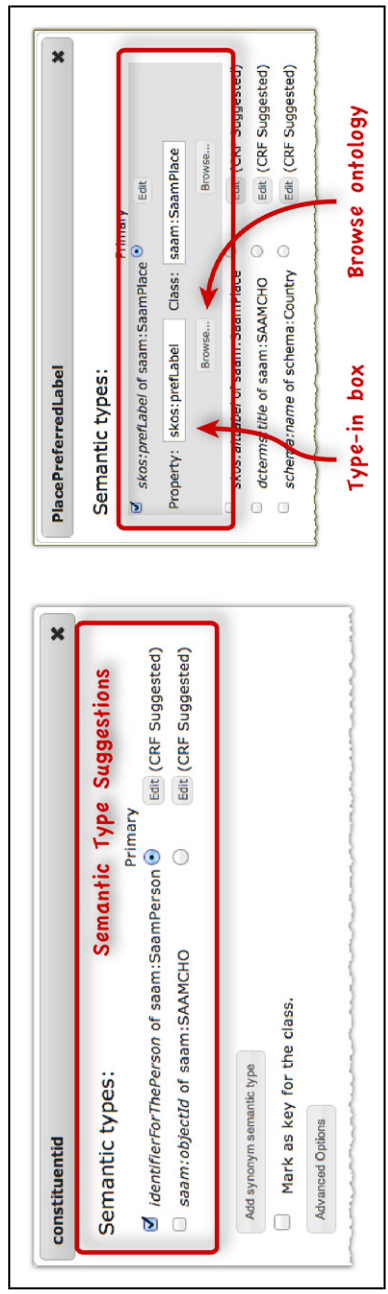

Figure 2. Semantic types map data columns to classes and properties in an ontology. Left: Karma suggestions to model the constituentid column in a SAAM table (the first choice is correct). Right: user interface for editing incorrect suggestions.

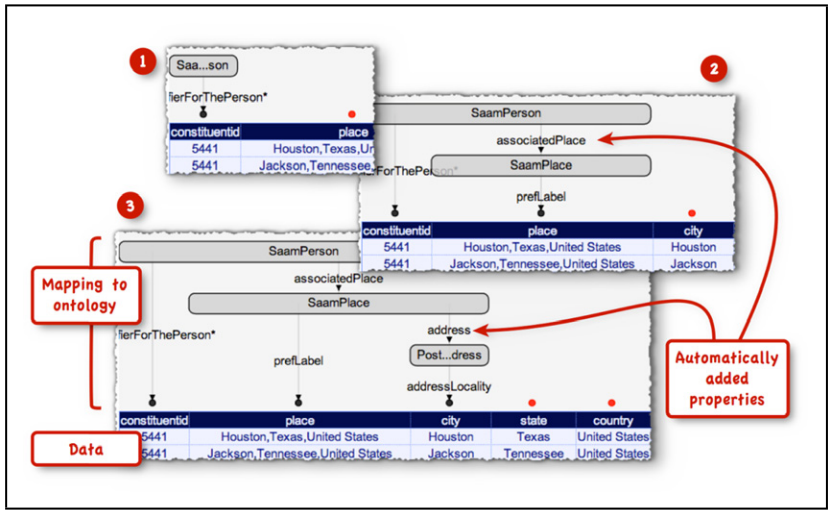

Figure 3. Each time the user adds new semantic types to the model, Karma connects them to the classes already in the model.

A semantic type can be either an OWL class or the range of a data property (which we represent by the pair consisting of a data property and its domain). Karma uses a conditional random field[15] (CRF) model to learn the assignment of semantic types to columns of data from user-provided assignments[16]. Karma uses the CRF model to automatically suggest semantic types for unassigned data columns (Figure 2). When the desired semantic type is not among the suggested types, users can browse the ontology to find the appropriate type. Karma automatically re-trains the CRF model after these manual assignments.

The relationships between semantic types are specified using paths of object properties. Given the ontologies and the assigned semantic types, Karma creates a graph that defines the space of all possible mappings between the data source and the ontologies[14]. The nodes in this graph represent classes in the ontology, and the edges represent properties. Karma then computes the minimal tree that connects all the semantic types, as this tree corresponds to the most concise model that relates all the columns in a data source, and it is a good starting point for refining the model (Figure 3). Sometimes, multiple minimal trees exist, or the correct interpretation of the data is defined by a non-minimal tree. For these cases, Karma provides an easy-to-use GUI to let users select a desired relationship (an edge in the graph). Karma then computes a new minimal tree that incorporates the user-specified relationships.

Mapping Columns to Classes

Mapping columns to the ontology is challenging because in the complete SAAM ontology there are 407 classes and 105 data properties to choose from. Karma addresses this problem by learning the assignment of semantic types to columns. Figure 2 shows how users define the semantic types for the constituentid (people or organizations) and place columns in one of the SAAM tables. The figure shows a situation where Karma had learned many semantic types. The left part shows the suggestions for constituentid. The SAAM database uses sequential numbers to identify both constituents and objects. This makes them indistinguishable, so Karma offers both as suggestions, and does not offer other irrelevant and incorrect suggestions. The second example illustrates the suggestions for the place column and shows how users can edit the suggestions when they are incorrect.

Connecting the Classes

Connecting the classes is also challenging because there are 229 object properties in the ontology to choose from. Figure 3 illustrates how Karma automatically connects the semantic types for columns as users define them. In the first screen the user assigns a semantic type for consitutentid. In the second screen, the user assigns a semantic type for place, and Karma automatically adds to the model the associatedPlace object property to connect the newly added SaamPlace to the pre-existing SaamPerson. Similarly, when the user specifies the semantic type for column city, Karma automatically adds the address object property. Each time users model the semantic type of a column, Karma connects it to the rest of the model[14].

Evaluation

We evaluated the effectiveness of Karma by mapping 8 tables (29 columns) to the SAAM ontology (Table 1). We performed the mapping twice: in Run 1, we started with no learned semantic types, and in Run 2 we ran Karma using the semantic types learned in the first run. The author of the paper that designed the ontology performed the evaluation. Even though he knows which properties and classes to use, when Karma didn't suggest them he used the browse capability to find them in the ontology instead of typing them in. It took him 18 minutes to map all the tables to RDF, even in the first run, when Karma's semantic type suggestions contained the correct semantic type 24% of the time. The second run shows that the time goes down sharply when users don't need to browse the ontology to find the appropriate properties and classes. The evaluation also

shows that Karma's algorithm for assigning relationships among classes is very effective (85% and 91% correct in Run 1 and Run 2).

Linking to External Resources

The RDF data will benefit the Smithsonian museum and the community if it is linked to useful datasets. We focused on linking SAAM artists to DBpedia[17] as it provides a gateway to other linked data resources and it is a focus for innovative applications. We also linked the SAAM artists to the Getty Union List of Artist Names (ULAN®) and to the artists in the Rijksmuseum dataset.

Museums pride themselves in publishing authoritative data, so SAAM personnel manually verified all proposed links before they became part of the dataset. To make the verification process manageable, we sought high-precision algorithms. We matched people using their names, including variants, and their birth dates and death dates. The task is challenging because people's names are recorded in many different ways, multiple people can have the same name, and birth dates and death dates are often missing or incorrect.

Our approach involves estimating the ratio of people in DBpedia having each possible value for the properties we use for matching (e.g., ratio of people born in 1879). We compare names using the Jaro-Winkler string metric[18], and for them compute the ratios as follows: we divide the interval [0, 1] in bins of size ε, and for each bin we estimate the number of pairs of people whose names differ by a Jaro-Winkler score less than ε. Empirically, we determined that $\varepsilon = 0.01$ and 10 million samples yield good results in our ground truth dataset.

The matching algorithm is simple. Given a SAAM and a DBpedia person, their matching score is $s = 1-d*n$ where d is the date score and n is the name score. If the dates match exactly, d is the fraction of people in DBpedia with those dates. Otherwise, d is the sum of the fractions for all the intervening years. n is the fraction of people in DBpedia whose Jaro-Winkler score is within ε from the score between the given pair of people.

Evaluation

To evaluate our algorithm we constructed ground truth for a dataset of 535 people in the SAAM database (those whose name starts with A). We manually searched in Wikipedia using all variant names and verified the matches using the text of the article and all fields in the SAAM record, including the biography. We found 176 matches in DBpedia.

Figure 4 shows the evaluation results on the ground truth (note that the matching score s decreases from left to right). The highest F-score .96 achieves a precision of .99 and a recall of .94 (166 correct results, 1 incorrect result). As the matching score decreases, precision suffers (more incorrect results), but

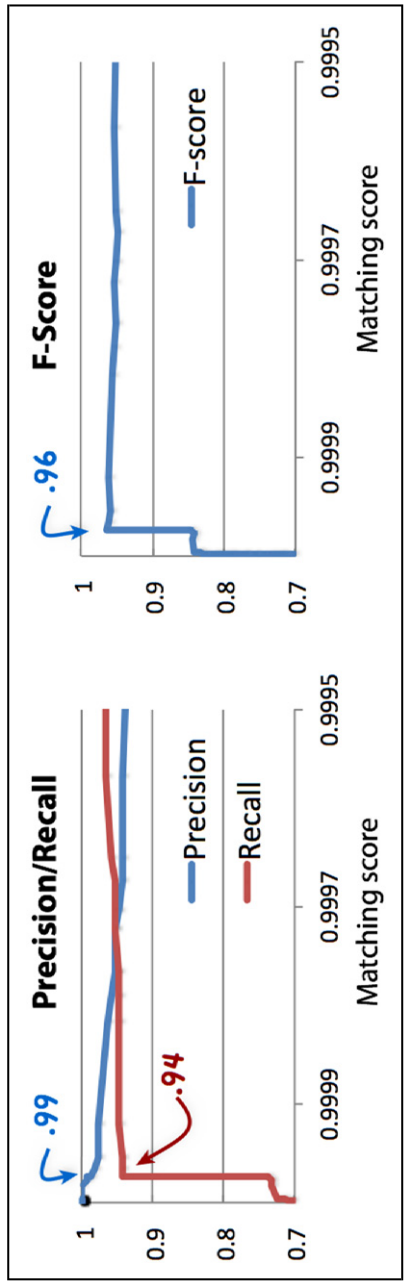

Figure 4. Precision/Recall and F-score as a function of our algorithm's matching score.

recall improves (more links identified). We linked the complete datasets using a matching score of 0.9995 because the loss of precision is relatively small and in the curation step users can easily identify the comparatively small number of incorrect matches that get introduced. This process identified 2,807 links to DBpedia, 1,759 links to Getty ULAN® and 321 links to the Rijksmuseum.

Curating the Linked Data

Museums need the ability to ensure that the linked data they publish are of high quality. The first aspect of the curation process is to ensure that the RDF is correct. Museum personnel can easily browse individual RDF records on the Web, but without understanding the relationship between an RDF record and the underlying database records, it is hard to assess whether the RDF is correct. Karma helps museum personnel understand these relationships at the schema level by graphically showing how database columns map to classes and properties in the ontology (e.g., Figures 2 & 3). Karma also lets users click on individual worksheet cells to inspect the RDF generated for it, helping them understand the relationships at the data level. These graphical views also enable SAAM personnel and the Semantic Web researchers to communicate effectively while refining the ontology and the mappings. Our goal by the end of the project is that SAAM personnel will use Karma to refine the mappings on their own.

The second aspect of the curation process is to ensure that links to external sources are correct. Our approach is to 1) record the full provenance of each link so that users (and machines) can record links and inspect them when the data sources or the algorithm change, and 2) make it easy for users to review the results of the linking algorithm. We use the PROV ontology[19] to represent provenance data for every link including revisions, matching scores, creation times, author (human or system/version), and data used to produce a link. Users review the links using the Web interface depicted in Figure 5. The interface is a visualization and editor of the underlying PROV RDF records. Each row represents a link. The first cell shows the records being linked: the top part shows links to information about the SAAM record and the bottom part shows links to information for a record in an external source. The next columns show the data values that were used to create the link and information about its revision history. The last column shows buttons to enable users to revise links and provide comments. SAAM personnel used this interface to verify all 2,807 links to DBpedia.

Related Work

There has been much recent interest in publishing museum data as Linked Open Data. Europeana[20], one of the most ambitious efforts, published the metadata

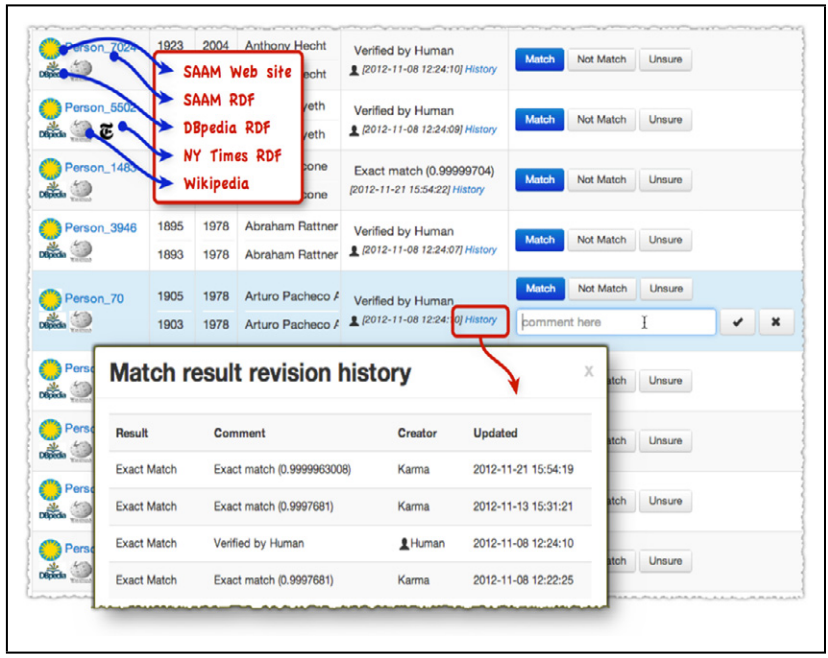

Figure 5. The Karma interface enables users to review the results of linking.

on 17 million items from 1,500 cultural institutions. This project developed a comprehensive ontology, called the Europeana Data Model (EDM) and used it to standardize the data that each organization contributes. This standard ontology enables Europeana to aggregate data from such a large number of cultural institutions. The focus of that effort was on developing a comprehensive data model and mapping all of the data to that model. Several smaller efforts focused on mapping rich metadata into RDF while preserving the full content of the original data. This includes the MuseumFinland, which published the metadata on 4,000 cultural artifacts[20] and the Amsterdam Museum[2], which published the metadata on 73,000 objects. In both of these efforts the data is first mapped directly from the raw source into RDF and then complex mapping rules transform this RDF into an RDF expressed in terms of their chosen ontology. The actual mapping process requires using Prolog rules for some of the more complicated cases. Finally, the LODAC Museum published metadata from 114 museums and research institutes in Japan. They defined a relatively simple ontology that consists of objects, artists, and institutions to simplify the mapping process.

In our work on mapping the 41,000 objects from SAAM, we went beyond the previous work in several important ways. First, we developed an approach that supports the mapping of complex sources (both relational and hierarchical) into rich domain ontologies[14]. This approach is in contrast to previous work, which first maps the data directly into RDF[21] and then aligns the RDF with the domain ontology[22]. As described earlier, we build on the EDM ontology, a rich and easily extensible domain ontology. Our approach makes it possible to preserve the richness of the original metadata sources, but unlike the MuseumFinland and the Amsterdam Museum projects, a user does not need to learn a complex rule language.

Second, we performed significantly more data linking than these previous efforts. There is significant prior work on linking data across sources and the most closely related is the work on Silk[23] and the work on entity coreference in RDF graphs[24]. Silk provides a nice framework that allows a user to define a set of matching rules and weights that determine whether two entities should be matched. We tried to use Silk on this project, but we found it extremely difficult to write a set of matching rules that produced high quality matches. The difficulty was due to a combination of missing data and the variation in the discriminability of different data values. The approach that we used in the end was inspired by the work on entity coreference by Song and Heflin[25], which deals well with missing values and takes into account the discriminability of the attribute values in making a determination of the likelihood of a match.

Third, because of the importance to the Smithsonian of producing high-quality linked data, we developed a curation tool that allows an expert from the museum to review and approve or reject the links produced automatically by our system. Previous work has largely ignored the issue of link quality (Halpin et al.[25] reported that in one evaluation roughly 51% of the same-as links were found to be correct). The exception to this is the effort by the NY Times to map all of their metadata to linked data through a process of manual curation. In order to support a careful evaluation of the links produced by our system, we developed the linking approach that allows a link reviewer to see the data that is the basis for the link and to be able to drill down into the individual sources to evaluate a link.

CONCLUSIONS AND FUTURE WORK

In this paper we described our work on mapping the data of the Smithsonian American Art Museum to Linked Open Data. We presented the end-to-end process of mapping this data, which includes the selection of the domain ontologies, the mapping of the database tables into RDF, the linking of the data to other related sources, and the curation of the resulting data to ensure

high-quality data. This initial work provided us with a much deeper understanding of the real-world challenges in creating high-quality link data.

For the Smithsonian, the linked data provides access to information that was not previously available. The Museum currently has 1,123 artist biographies that it makes available on its website; through the linked data, we identified 2,807 links to people records in DBpedia, which SAAM personnel verified. The Smithsonian can now link to the corresponding Wikipedia biographies, increasing the biographies they offer by 60%. Via the links to DBpedia, they now have links to the New York Times, which includes obituaries, exhibition and publication reviews, auction results, and more. They can embed this additional rich information into their records, including 1,759 Getty ULAN® identifiers, to benefit their scholarly and public constituents.

The larger goal of this project is not just to map the SAAM data to Linked Open Data, but rather to develop the tools that will enable any museum or other organization to map their data to linked data themselves. We have already developed the Karma integration tool, which greatly simplifies the problem of mapping structured data into RDF, a high-accuracy approach to linking datasets, and a new curation tool that allows an expert to review the links across data sources. Beyond these techniques and tools, there is much more work to be done. First, we plan to continue to refine and extend the ontologies to support a wide range of museum-related data. Second, we plan to continue to develop and refine the capabilities for data preparation and source modeling in Karma to support the rapid conversion of raw source data into RDF. Third, we plan to generalize our initial work on linking data and integrate a general linking capability into Karma that allows a user to create high-accuracy linking rules and to do so by example rather than having to write the rules by hand.

We also plan to explore new ways to use the linked data to create compelling applications for museums. A tool for finding relationships, like EverythingIsConnected.be[26], has great potential. We can imagine a relationship finder application that allows a museum to develop curated experiences, linking artworks and other concepts to present a guided story. The Museum could offer pre-built curated experiences or the application could be used by students, teachers, and others to create their own self-curated experiences.

ACKNOWLEDGEMENTS

This research was funded by the Smithsonian American Art Museum. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Smithsonian Institution.

END NOTES

[1] Haslhofer, B., Isaac, A.: data.europeana.eu - The Europeana Linked Open Data Pilot. In: *Multiple values selected*. The Hague, The Netherlands (Jul 2011)

[2] 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. In: Simperl, E., Cimiano, P., Polleres, A., Corcho, O., Presutti, V. (eds.) *Lecture Notes in Computer Science*, pp. 733–747. Springer Berlin Heidelberg (2012)

[3] Matsumura, F., Kobayashi, I., Kato, F., Kamura, T., Ohmukai, I., Takeda, H.: Producing and Consuming Linked Open Data on Art with a Local Community. In: *Proceedings of the Third International Workshop on Consuming Linked Data (COLD2012)*. CEUR Workshop Proceedings (2012)

[4] http://www.europeana.eu/schemas/edm/

[5] http://europeana.eu

[6] http://www.w3.org/2004/02/skos/

[7] http://purl.org/dc/elements/1.1/ and http://purl.org/dc/terms/

[8] http://xmlns.com/foaf/0.1/

[9] http://rdvocab.info/ElementsGr2

[10] http://www.openarchives.org/ore/terms/

[11] http://americanart.si/linkeddata/schema/

[12] http://schema.org/

[13] http://www.cidoc-crm.org

[14] Knoblock, C.A., 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. In: *Proceedings of the 9th international conference on The Semantic Web: research and applications*. pp. 375–390. Springer-Verlag, Berlin, Heidelberg (2012)

[15] Lafferty, J., McCallum, A., Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: *Proceedings of the International Conference on Machine Learning* (2001)

[16] Goel, A., Knoblock, C.A., Lerman, K.: Exploiting Structure within Data for Accurate Labeling Using Conditional Random Fields. In: *Proceedings of the 14th International Conference on Artificial Intelligence (ICAI)* (2012)

[17] http://dbpedia.org

[18] Cohen, W.W., Ravikumar, P., Fienberg, S.E., Others: A comparison of string distance metrics for name-matching tasks. In: *Proceedings of the IJCAI-2003 Work- shop on Information Integration on the Web (IIWeb-03)*. pp. 73–78 (2003)

[19] http://www.w3.org/TR/prov-o/

[20] Haslhofer, B., Isaac, A.: data.europeana.eu - The Europeana Linked Open Data Pilot. In: *Multiple values selected*. The Hague, The Netherlands (Jul 2011)

[21] Bizer, C., Cyganiak, R.: D2R Server–publishing relational databases on the semantic web. In: Poster at the 5th International Semantic Web Conference (2006)

[22] Bizer, C., Schultz, A.: The R2R Framework: Publishing and Discovering Mappings on the Web. 1st International Workshop on Consuming Linked Data (2010)

[23] Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Silk–a link discovery framework for the web of data. In: *Proceedings of the 2nd Linked Data on the Web Workshop*. pp. 559–572 (2009)

[24] Song, D., Heflin, J.: Domain-independent entity coreference for linking ontology instances. *ACM Journal of Data and Information Quality* (ACM JDIQ) (2012)

[25] Halpin, H., Hayes, P., McCusker, J., Mcguinness, D., Thompson, H.: When owl: same as isn't the same: An analysis of identity in linked data. *Proceedings of the 9th International Semantic Web Conference* pp. 305–320 (2010)

[26] Sande, M.V., Verborgh, R., Coppens, S., Nies, T.D., Debevere, P., Vocht, L.D., Potter, P.D., Deursen, D.V., Mannens, E., and Walle, R.: Everything is Connected. In: *Proceedings of the 11th International Semantic Web Conference (ISWC)* (2012)