A Demonstration of Linked Data Source Discovery and Integration*

Jason Slepicka, Chengye Yin, Pedro Szekely, and Craig A. Knoblock

University of Southern California
Information Sciences Institute and Department of Computer Science, USA
{knoblock,pszekely,slepicka}@isi.edu
{chengyey}@usc.edu

Abstract. The Linked Data cloud is an enormous repository of data, but it is difficult for users to find relevant data and integrate this data into their own datasets. Datasets in the Linked Data cloud are represented using ontologies, but lack accurate descriptions of the data they contain. We present an approach that leverages R2RML mappings to describe the contents of datasets in order to find relevant data and integrate it with a user's own datasets. Our demonstration shows how users can easily create R2RML mappings for their datasets and then use these mappings to augment their datasets with data from the Linked Data cloud.

1 Introduction

The Linked Data cloud contains an enormous amount of data about many topics. Consider museums, which often have detailed data about their artworks but may only have sparse data about the artists who created them. Museums typically have tombstone data about artists (name, birth/death years, and places) but may lack biographies, influences, etc. Museums could use additional information about their artists in the Linked Data cloud and integrate it with their own to produce a richer, more complete dataset.

Our approach to this use case, built into our KARMA data integration system [8], leverages R2RML mappings [7] to describe user's datasets and datasets in the Linked Data cloud. We envision a world where such datasets include richer descriptions than what is available today. Today, datasets include, at best, a VoID description [1] with basic metadata, such as access method and vocabularies used. We propose adding R2RML mappings, which contain significantly richer descriptions about each subject and its properties. Even though R2RML was defined to specify mappings from relational databases to RDF, recent work [2] has proposed extensions to handle other types of data, including CSV, JSON, XML and Web APIs. Consequently, it is reasonable to expect that more datasets in the Linked Data cloud could be published with R2RML-style descriptions.

In this demonstration we show how museum users can use KARMA to quickly define an R2RML mapping of a dataset (our previous work), how they can use

* A video demonstration is available at http://youtu.be/sr-XDBKeNCY

R2RML mappings to find more information about the artists in their dataset, and how they can augment their dataset with additional information.

2 Datasets

For our demonstration we will integrate a CSV file containing 197 artists with Linked Data published by the Smithsonian American Art Museum (SAAM). In previous work [8], we mapped the SAAM dataset, including over 40,000 artworks and 8,000 artists to the CIDOC CRM ontology [3] using R2RML and made it accessible by a SPARQL endpoint, along with a repository for the R2RML mappings. The SAAM LOD here is a proxy for the Linked Data cloud to illustrate the vision of a Linked Data cloud populated with R2RML models.

3 Demonstration

We will show how a user can interactively model an artist dataset, discover the Smithsonian's data for those artists, and then integrate the Smithsonian's data.

Step 1: Modeling a New Source. The user begins by using KARMA's existing capability to model the artists in the CSV file as crm:E21_Person in an R2RML mapping shown in Figure 1. KARMA can use this mapping to generate RDF, and can also compare it to retrieve other mappings, discovering new related sources that can be integrated with the artist dataset.

Step 2: Discovering Data Sources. The user then clicks on E21_Person1 in the R2RML mapping and selects Augment Data to discover new data to integrate into artist records. KARMA retrieves R2RML mappings in its repository that describe crm:E21_Person, and uses these mappings to generate a candidate set of linked data sources to integrate, identifies meaningful object and data properties, and presents them to the user as illustrated in Figure 2. To help the users select properties to integrate, Karma uses Bloom filters to estimate the number of artists that have each of the properties listed in Figure 2.

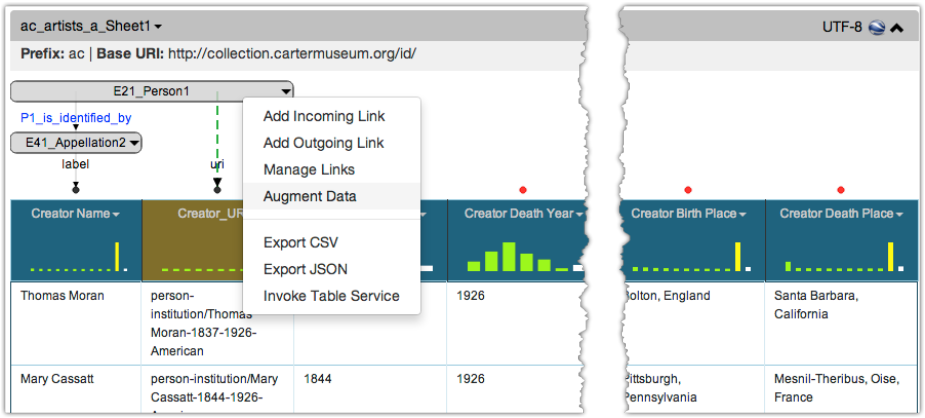

Fig. 1. A KARMA user creates an R2RML mapping for a CSV file of a museum's artists' biographical records and clicks 'Augment Data' to discover new data sources

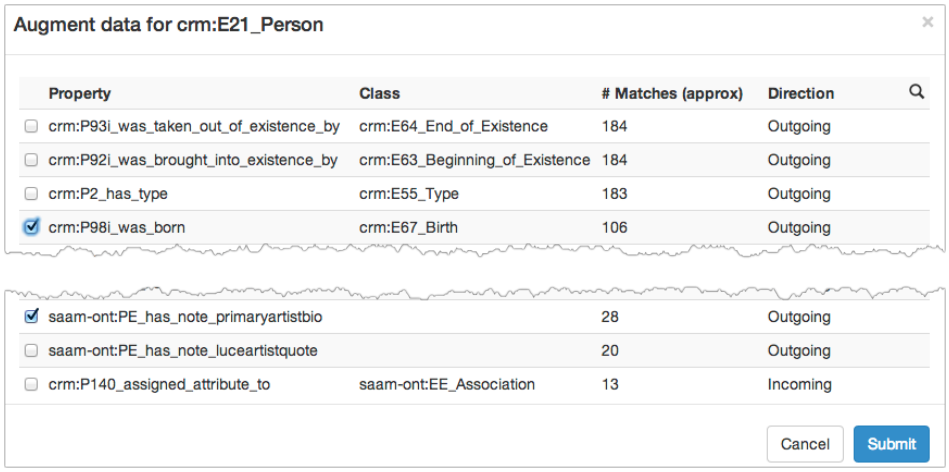

Fig. 2. A Karma user selects CIDOC CRM object and data properties discovered from other sources to augment crm:E21_Person

Step 3: Integrating Data Sources. The user selects the artist's biography (for completeness) and birth (for validation). KARMA automatically constructs SPARQL queries to retrieve the data, integrates it into the worksheet, and augments the R2RML mapping accordingly (Figure 3). To support the integrated SPARQL queries, we generated owl:sameAs links between the artists in the CSV file and the Smithsonian dataset using LIMES [5] (we plan to integrate LIMES with KARMA to enable users to perform all integration steps within KARMA).

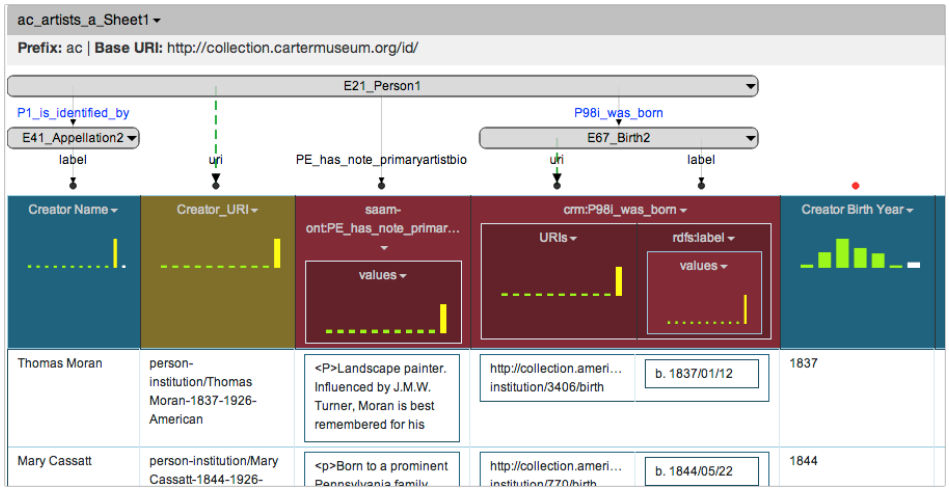

Fig. 3. A Karma user has integrated biographical data from the Smithsonian as new columns in their dataset. The columns contain artists' biographies and birth dates.

4 Related Work and Conclusions

We see similarities in our approach with those used in relational database integration and semantic service composition. ORCHESTRA[4] starts, like R2RML, by aligning database tables to a schema graph. For integration, heuristics are used to translate keyword searches over the graph into join paths using its Q query system. However, these joins are not guaranteed to be semantically meaningful, unlike the integration paths KARMA finds using R2RML.

Platforms such as iServe[6] capture Linked Services and make them discoverable and queryable by annotating them with their Minimal Service Model. However, the past work on service discovery and composition only uses a semantic model of the inputs and outputs of the services. In contrast, KARMA service descriptions [9] also capture the relationship between the attributes, which allows us to automatically discover semantically meaningful joins.

By building on KARMA's ability to quickly model many source types, we demonstrate how a user can discover other linked data sources, select the desired attributes from those sources, and then integrate the data from those sources into their own dataset. Through this source discovery and integration, a user can transparently compose and join other sources and services in a semantically meaningful, interactive way that was not previously possible.

References

1. ALEXANDER, K., CYGANIAK, R., HAUSENBLAS, M., AND ZHAO, J. Describing linked datasets with the VoID vocabulary. W3C note, W3C, Mar. 2011. http://www.w3.org/TR/2011/NOTE-void-20110303/.
2. DIMOU, A., SANDE, M. V., COLPAERT, P., MANNENS, E., AND DE WALLE, R. V. Extending r2rml to a source-independent mapping language for rdf. In *International Semantic Web Conference (Posters and Demos)* (2013), vol. 1035 of *CEUR Workshop Proceedings*, CEUR-WS.org, pp. 237–240.
3. DOERR, M. The cidoc conceptual reference module: An ontological approach to semantic interoperability of metadata. *AI Mag. 24*, 3 (Sept. 2003), 75–92.
4. IVES, Z. G., GREEN, T. J., KARVOUNARAKIS, G., TAYLOR, N. E., TANNEN, V., TALUKDAR, P. P., JACOB, M., AND PEREIRA, F. The orchestra collaborative data sharing system. *ACM SIGMOD Record 37*, 3 (2008), 26–32.
5. NGOMO, A.-C. N., AND AUER, S. Limes: a time-efficient approach for large-scale link discovery on the web of data. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Three* (2011), AAAI Press, pp. 2312–2317.
6. PEDRINACI, C., LIU, D., MALESHKOVA, M., LAMBERT, D., KOPECKY, J., AND DOMINGUE, J. iserve: a linked services publishing platform. In *CEUR workshop proceedings* (2010), vol. 596.
7. SUNDARA, S., CYGANIAK, R., AND DAS, S. R2RML: RDB to RDF mapping language. W3C recommendation, W3C, Sept. 2012. http://www.w3.org/TR/2012/REC-r2rml-20120927/.
8. SZEKELY, P., KNOBLOCK, C. A., YANG, F., ZHU, X., FINK, E., ALLEN, R., AND GOODLANDER, G. Connecting the Smithsonian American Art Museum to the Linked Data Cloud. In *Proceedings of the 10th Extended Semantic Web Conference* (Montpellier, May 2013).

9. Taheriyan, M., Knoblock, C. A., Szekely, P., and Ambite, J. L. Semi-automatically modeling web apis to create linked apis. In *Proceedings of the ESWC 2012 Workshop on Linked APIs* (2012).