Department of
Information Engineering
and Computer Science
DISI

UNIVERSITY
OF TRENTO - Italy

DISI - Via Sommarive, 5 - 38123 POVO, Trento - Italy
http://disi.unitn.it

PROVENANCE IN OPEN DATA ENTITY-CENTRIC AGGREGATION

Fausto Giunghiglia and Moaz Reyad

May 2014

Technical Report # DISI-14-008

1 Motivation and goals

Recently an increasing number of open data catalogs appear on the Web [1]. These catalogs contain data that represents real world entities and their attributes. Data can be imported from several catalogs to build web services; hence there is a need to trace the source of each entity and attribute value in a way that handles also the possible conflicts between attribute values coming from overlapping sources [2]. For open data, source tracing requires capturing both the provenance [3] of the attribute values and the identity links [4] between entities. Moreover, resolving the conflicts manually becomes harder with the increasing size of data.

We propose a source tracing module that extends any existing import process by making it tracing-aware. The source tracing module contains three tools: authority, provenance and evidence. Authority provides rules for overriding attribute values, provenance specifies the source of an attribute value and evidence provides identity links between entities.

2 Problem

The problem of tracing sources is studied with respect to an import process that takes an open data catalog and extracts entities and their attribute values from its contents. The extracted entities and attribute values are imported into a database called *entity base*.

A common category of the open data repositories is the DCAT catalog. DCAT[1] (Data Catalog Vocabulary) is an RDF vocabulary for describing datasets in a data catalog. A DCAT catalog can have one or more datasets, a dataset can have one or more distributions. DCAT catalogs exist within a Web-based system called CKAN. CKAN[2] (Comprehensive Knowledge Archive Network) is a dataset distribution system. Datasets are distributed as packages. Each package has one or more resource groups, and each resource group has one or more resources.

Open data catalogs contain data that represents objects from the real world. We refer to real world objects that are of enough importance to be given a name as *entities*. An example for entities is Italy. There are different *entity types*, such as Locations. Italy is an entity of type Location. The

[1] http://www.w3.org/TR/vocab-dcat/

[2] http://ckan.org

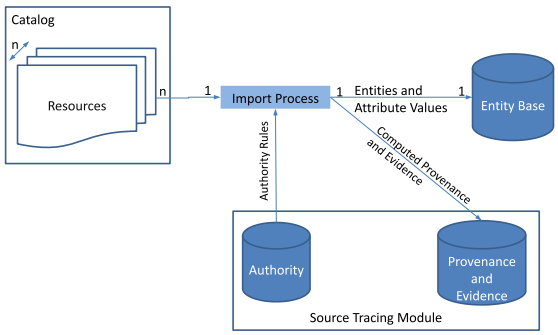

Figure 1: Extending an import process with the source tracing module

type of entity gives the list of *attribute definitions* that can be assigned to an entity of this type. Location entities may have the attribute Area which holds the value of the total area of the location. The values of the attribute definitions for a specific entity are called *attribute values*.

The entity base is populated with entities through an import process which can be, for instance, a generic work flow for importing any dataset or a custom procedure for importing a specific dataset. We consider any import process that has the following three aspects:

1. Partiality: The import process may take a partial input.
2. Overlap: Imported data may be disjoint or overlapped with existing entities and attribute values in the entity base.
3. Multiple Imports: The import process may run multiple times on the same catalog.

3 Our approach

We propose a source tracing module that extends any existing import process by making it tracing-aware (see Figure 1). The source tracing module contains three tools: authority, provenance and evidence.

3.1 Authority

Authority is a meta-attribute of an element (entity type, an attribute definition, an entity or an attribute value) that provides a connection between

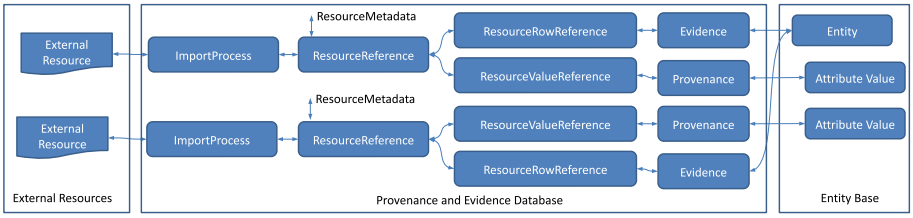

Figure 2: Extending an import process with the source tracing module

the element and the resource which has the authority to create or update it. Authority is specified through a set of authority rules. An authority rule is a relation between a resource and one or more elements which is called the scope, with a ranking value that is called the priority.

The scope specifies the set of elements that are affected by an authority rule. We support four ordered levels of authority scope: (1) entity type, (2) a set of entities, (3) attribute definition and (4) attribute value. The three aspects of the import process (partiality, overlap and multiple imports) can happen at any scope. The priority is a ranking value that is assigned to order if multiple sources are given authority for the same scope. This ranking is a total order. Authority should be defined for each element. Its purpose is to help in finding a winning resource if there is a conflict between two resources in an attribute value.

3.2 Provenance and Evidence

An import process runs on an external resource and extracts entities and their attribute values from it. Before creating or updating the entities and their attribute values in the entity base, a tracing-aware import process creates a graph of elements between the external source and the entity base. This graph is shown in Figure 2. The ultimate goal of this graph is to trace the sources of each element in the entity base. The graph is connected to the entity base through provenance and evidence. Provenance is a meta-attribute that specifies the source of an attribute value; while evidence is an attribute that links an entity with another external entity which represents the same real world object.

4 Acknowledgments

This work is partially sponsored by the CUBRIK (Human-enhanced time-aware multi-media search), an IP funded project within the European Community's Seventh Framework Programme FP7/2007-2013 under grant agreement n287704.

References

[1] Braunschweig, K., Eberius, J., Thiele, M., Lehner, W.: The state of open data limits of current open data platforms (2012)

[2] Bleiholder, J., Naumann, F.: Data fusion. ACM Comput. Surv. **41**(1) (January 2009) 1:1–1:41

[3] Moreau, L., Groth, P.T.: Provenance: An Introduction to PROV. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan and Claypool Publishers (2013)

[4] Halpin, H., Place, B.: When owl:sameas isn't the same: An analysis of identity links on the semantic web. In: In Linked Data on the Web (LDOW). (2010)