TOPIC: LINKED DATA FOR LIBRARIES, ARCHIVES, AND MUSEUMS

- LINKED DATA VOCABULARY MANAGEMENT
- EUROPEANA MOVING TO LINKED OPEN DATA
- LINKING LIVES END-USER INTERFACE
- JOINING THE LINKED DATA CLOUD
- OCLC’S LINKED DATA INITIATIVE
- LC’S BIBLIOGRAPHIC FRAMEWORK INITIATIVE
Linked Data hold the promise to derive additional value from existing data throughout different sectors, but practitioners currently lack a straightforward methodology and the tools to experiment with Linked Data. This article gives a pragmatic overview of how general purpose Interactive Data Transformation tools (IDTs) can be used to perform the two essential steps to bring data into the Linked Data cloud: data cleaning and reconciliation. These steps are explained with the help of freely available data (Cooper-Hewitt National Design Museum, New York) and tools (Google Refine), making the process repeatable and understandable for practitioners.

**Linked Data comes at a cost**

Many institutions are now aware of the importance of having their data available as Linked Data. The five-star scheme proposed by Tim Berners-Lee seems a valuable tool to assess the reusability of data for current and future applications. However, some goals are more difficult to reach than others: for example, linking to other data is currently a rare practice, yet this is crucial to provide context for both human and machine data consumers. The evaluation scheme unfortunately makes abstraction of the quality of the data: while most institutions have data available in a structured format, consistency issues within individual data fields present tremendous hurdles to create links in between datasets in an automated manner.

Streamlining and cleaning data to enhance the linking process in between heterogeneous data sources used to be a task that had to be performed by people with both high domain and technological skills. Often, many thousands of records require similar operations. This is either a tedious manual task or something that needs to be automated on a per-case basis. Luckily, the advent of Interactive Data Transformation tools (IDTs) allows for rapid and inexpensive operations on large amounts of data, even by domain experts who do not have in-depth technical skills. But exactly how much can be achieved with IDTs and how
reliable are the results? These are the questions we have been investigating in the scope of the Free Your Metadata initiative. We were able to verify that IDTs can assist with cleaning and linking of large datasets, leading to a high success percentage at a minimal cost. (For more on this, see our forthcoming JASIST article: Evaluating the success of vocabulary reconciliation for cultural heritage collections. Pre-print at: freeyourmetadata.org/publications/)

The Interactive Data Transformation revolution

Interactive Data Transformation tools resemble the desktop spreadsheet software we are all familiar with. While spreadsheets are designed to work on individual rows and cells, IDTs operate on large datasets at once. These tools offer a homogeneous and non-expert interface through which domain experts can perform both the cleaning and reconciliation operations. Several general-purpose tools for interactive data transformation have been developed over the last years, such as Potter’s Wheel ABC and Wrangler. Here we want to focus specifically on Google Refine (formerly Freebase Gridworks), as it has recently gained a lot of popularity and is rapidly becoming the tool of choice to efficiently process and clean large amounts of data in a browser based interface. Google Refine further allows the reconciliation of data with existing knowledge bases, creating the connection with the Linked Data vision. The DERI research group has developed an RDF extension for Google Refine, which can be downloaded for free. The RDF extension allows users to add SPARQL endpoints to the reconciliation process. DBpedia is added, for example, so that the content of a keyword type of field can be matched to terms described as SKOS concepts in DBpedia. More specialized sources such as the LC Subject Headings (LCSH) and the Art & Architecture Thesaurus (AAT)® can also be used.

CASE STUDY  Cooper-Hewitt National Design Museum, New York

The use of Google Refine for metadata cleaning and reconciliation will be demonstrated with the help of the metadata of the Cooper-Hewitt National Design Museum, which released its collection metadata as a downloadable file through GitHub in February 2012, using the Creative Commons Attribution Share Alike (CCASA) license. The file contains basic metadata (29 fields) for 137,571 objects.

The cleaning operations are performed on the entire dataset but for reasons of simplicity and performance the reconciliation process specifically focuses on the metadata record for Design for a candelabrum by Michelangelo, as it is one of the iconic objects from the collection. The drawing is available through the Google Art project.

Cooper-Hewitt National Design Museum
Design for a candelabrum by Michelangelo
Data cleaning

Once the data are imported into Google Refine, a diverse set of filters and facets can be applied on the individual fields. All manipulations are performed through a clear and straightforward interface, allowing domain experts without any technical background to experiment with data normalization and cleaning.

The following operations illustrate some of the most recursive issues with data and how Google Refine can be used to both identify, and where possible, solve them in an automated manner.

Deduplication

After loading the data into the application, the first operation we need to perform is to detect and remove duplicates. This can easily be done by performing the Duplicates facet on fields such as objectid and invno (inventory number). For example, 6,215 records were identified through this facet that have a duplicate inventory number.

Atomization

A quick glance at the medium field, which typically has content such as “Quill-work, silver, glass, and black-painted pine,” or the content of the geography field (e.g., “London England”), illustrates one of the biggest hurdles for automated data analysis and reconciliation: field overloading. These values need to be split out into individual cells through the function Split multi-valued cells on the basis of separation characters, which are a comma and a whitespace in the case of the medium field and a whitespace in the case of the geography field.

Applying facets and clustering

Once the content of a field has been properly atomized, filters, facets, and clusters can be applied to give a quick and straightforward overview of classic formal data issues. By applying the custom facet facet by blank, one can in a matter of seconds measure the completeness of the fields; for example, 72% of the description fields and 93% of the movement fields from the Cooper-Hewitt collection are left blank.

The text facet is one of the most powerful features of Google Refine, as it instantly identifies both the most recurrent values and the outliers of a field. When applied on the names field for our collection, we see a total number of 3,785 different values composed of a small number of terms that are heavily used (e.g., “drawing” is used to describe 27% of the objects) and a long tail of object names which are only used once. After the application of a facet, Google Refine proposes to cluster facet choices together based on various similarity methods, such as nearest neighbor or key-collision. The two or more related values are presented and a merge is proposed, which can either be approved or manually overridden. Figure 1 illustrates the clustering and how it allows resolution of case inconsistencies, incoherent use of either the singular or plural form, and simple spelling mistakes. However, a manual check of the proposed clusters is necessary as attention needs to be given to near-duplicates such as toaster – coaster. The application of the nearest neighbor clustering method, which is considered as the least aggressive, typically reduces the number of variant values by 10 to 15%.
Reconciling data with the Linked Data cloud

Once the data has been cleaned, the moment has come to give meaning to the field values. We as humans understand what “Italian” and “Renaissance” mean, but to machines both terms are just strings of characters. With Linked Data, meaning is created by providing context in the form of links. For example, for “Renaissance,” we mean the cultural movement in Europe during the 14th to 17th centuries, as defined by the link http://en.wikipedia.org/wiki/Renaissance.

The process of matching text strings to concepts is called reconciliation in Google Refine. Reconciliation can be performed automatically on all records on a per-field basis. Among interesting columns to reconcile in the Cooper-Hewitt collection are name, culture, and period. For each column, you can specify the reconciliation source, and the type of entity contained in the column (see Figure 2).

We will illustrate reconciliation on the object “Design for a Candelabrum.” If we reconcile the name field with the LCSH vocabulary, the value “Drawing” becomes linked to the LCSH concept Drawing. The latter is more than simply a string; it is a concept in a hierarchy, with relations to other terms. The word “Italian” can be reconciled automatically to the Freebase entry of Italy. If we try to reconcile “Late Renaissance” with the LCSH, Refine offers us two alternatives between which it cannot choose automatically: Art, Late Renaissance and Painting, Late Renaissance. While we need to select our choice manually, Refine does limit the number of choices we have to make.

The links that result from the reconciliation process not only help machines, they also eventually help people consume information faster and smarter. For example, if the maker field is reconciled to the Michelangelo article in Wikipedia, people have access to relevant information directly. If many items from different collections are linked this way, people can browse related works automatically. Reconciliation thereby connects each collection to the Linked Data cloud.

CONTINUED »
Start picking the low hanging fruit!

With the help of freely available data and tools, we demonstrated in a straightforward manner how non-technical people can bring their own data into the Linked Data cloud. The arrival of IDTs and Google Refine, in particular, has made data cleaning and reconciliation available for the masses. Concrete examples showed how recurrent data quality issues can be handled by Google Refine and how to transform strings of text into links pointing to external data sources that are already a part of the Linked Data cloud.

The quickly evolving landscape of standards and technologies certainly continues to present challenges to non-technical domain experts wishing to derive additional value out of their data on the Web. We do not wish to make light of the inherent complexities involved in the interlinking of data, but we do want to point out the low hanging fruit that is currently right in front of our noses. This small case study demonstrates that significant results can be obtained at a minimal cost through freely available tools.

Acknowledgements

The authors would like to thank the Cooper-Hewitt National Design Museum for making their metadata freely available and therefore allowing us to perform the case study on the basis of metadata which can be used under the CCASA license.

The research activities described in this paper were partly funded by Ghent University, the Interdisciplinary Institute for Broadband Technology (IBBT), the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), the Fund for Scientific Research Flanders (FWO Flanders), and the European Union. doi:10.3789/isqv24n2-3.2012.04

SETH VAN HOOLAND (svhoolan@ulb.ac.be) is Digital Information Chair at Université Libre de Bruxelles. RUBEN VERBORGH (ruben.verborgh@ugent.be) is a PhD Student at Ghent University, Interdisciplinary Institute for Broadband Technology (IBBT) Multimedia Lab. RIK VAN DE WALLE (rik.vandewalle@ugent.be) is Senior Full Professor at Ghent University and Research Director of the Interdisciplinary Institute for Broadband Technology (IBBT) Multimedia Lab.