The educational value of EDM in information science

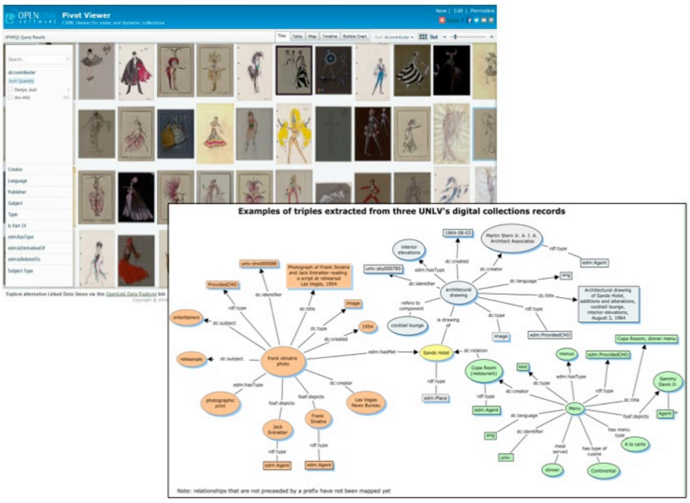
EDM is a data model designed for use in the cultural heritage sector and it supports the creation of linked open data. This case study looks at two initiatives that have found EDM a useful resource in teaching professionals and students about metadata and linked open data. One describes a project at the University of Nevada, Las Vegas (UNLV), to teach library staff about linked data in order to apply it locally; the second is a Massive Open Online Course (MOOC) offered by the University of North Carolina (UNC) about organizing and discovering data. Both have used EDM as an example data model.

Leading to Linking at UNLV

This project is described in the article "Leading to Linking: Introducing Linked Data to Academic Library digital Collections"[1]. The initial assumption is that the transition from a web of documents to a web of data has given users experience of, and raised expectations for, the kind of contextual information that comes from linking. It starts with the recognition that the creation of digital library collections and the possibility of linking data has entailed a huge shift away from traditional approaches to describing and organising information. This shift has been so profound that collection managers can be intimidated and deterred from taking action.

It is against this background that a group of academic library staff from the University of Nevada, Las Vegas, established a project to learn about linked data and evaluate its application in the area of their digital collections. The main objectives were to:

- "Study the feasibility of developing a common process that would allow the conversion of our collection records into linked data, *preserving their original expressivity and richness*
- Publish data from our collections as linked data to improve discoverability and connections with other related data sets on the Web."



[1] Cory K Lampert, Silvia B Southwick (2013) Leading to Linking: Introducing Linked Data to Academic Library digital Collections, Journal of Library Metadata, 13:2-3,230-253.

A study group of library staff used a series of workshops to learn the underlying principles and concepts of linked data. They then moved on to applying these to their data sets to create a set of RDF triples for publication. The decision was made to use the metadata of the digital collections rather than try to use the catalogue metadata (MARC), the latter seeming to require a lot more effort to convert.

To carry out an effective transformation of the data some initial decisions were required:

- choice of a data model and development of a mapping
- selection of technologies to transform and publish the data
- a process for URI assignment

Data Model and mapping

A survey of literature resulted in the choice of EDM as the data model. EDM is designed for use with digital objects in the cultural heritage sector so the content to be described was similar.

EDM offers three core classes: ProvidedCHO, for the original object being described, WebResource for the digital representations of the objects, and Aggregation, to link the whole together and provide data about the source of the description. There are classes for 'contextual' resources: Agent, Place, TimeSpan, Concept, Event and PhysicalThing. The classes and properties seemed a good fit with those in UNLV digital collection metadata.

The other valuable feature is that EDM re-uses well known and well defined properties from other namespaces such as Dublin Core and Friend of a Friend alongside the more Europeana-specific ones. Most of those originating in the Europeana namespace were created for Europeana-specific uses but are defined in general terms to enable their re-use by others.

UNLV concluded that for its initial approach it would use a partial version of EDM (not using ore:Proxym ore:Aggregation or edm:PhysicalThing) and with only minor changes otherwise. The version used is not fully documented yet and work is continuing to incorporate additional classes.

Once the customised version of the data model was defined mappings were created between the digital collections metadata and the classes and properties defined in the data model.

Technologies

Three technological components are required for creating, storing and publishing linked data. From amongst many contenders UNLV selected OpenRefine[2] for the data cleaning, manipulation and transformation; OpenLink Virtuoso[3] for storing, visualising and publishing the resulting RDF triples (the

[2] http://openrefine.org

[3] http://virtuoso.openlinksw.com

triplestore); and, for querying the data, the SPARQL[4] endpoint that is also offered by Virtuoso.

Data extraction and clean-up

A tool to manage the data extraction, manipulation and clean-up of the data is critical to the whole undertaking. The preparation of the data is the part of the process that takes most effort. UNLV selected OpenRefine with its RDF extension[5] as it can accept messy data, support the cleaning and manipulation of it and transform it from one format to another. An example of such cleaning or manipulation is the separation of the component parts of a MARC authority record for an author. This has the name in inverted form, followed by birth and death dates: it needs adjusting to produce three triples – the name, the birth date and the death date.

URI Assignment and Transformation

Transforming the data into RDF requires that each subject of a triple has a unique identifier – a URI. These could be assigned in the preparation phase or during transformation. Doing it during transformation saves human effort as it can be automated (or semi-automated) but the transformation process would be more efficient if URI assignment were carried out manually during the initial creation of the records. These options are still being explored as follow-up to the project.

UNLV are also still finalizing the rules for the creation of the URIs and have two situations: things that are unique to their own collection which therefore need a URI to be created, and things that may already have a URI assigned elsewhere which should be re-used.

For things unique to their collection a possible rule could be "*all URIs assigned to UNLV unique items will have the same root (e.g. "http://digital.library.unlv.edu) followed by the class of the item and a unique local identification of the item.* A costume design from the Showgirl digital collection could have a URI such as: http://digital.library.unlv.edu/ProvidedCHO/sho000005, where ProvidedCHO is the class of the item and sho000005 is its local unique identifier." These URIs serve only as identifiers at the moment but ultimately they will be dereferenceable.

Implementing the mapping

Using OpenRefine, the mapping guides the creation of the triples to be generated for each metadata element. A simple example:

Title: Costume design drawing, showgirl in Judy Garland costume, Las Vegas, June 5, 1980
Creator: Menefree, Pete
Genre: costume design drawings.

This results in three EDM classes:

[4] http://www.w3.org/TR/rdf-sparql-query/

[5] http://refine.deri.ie

- ProvidedCHO – the item being described
- Agent: the creator
- Concept: a term used to identify the item genre

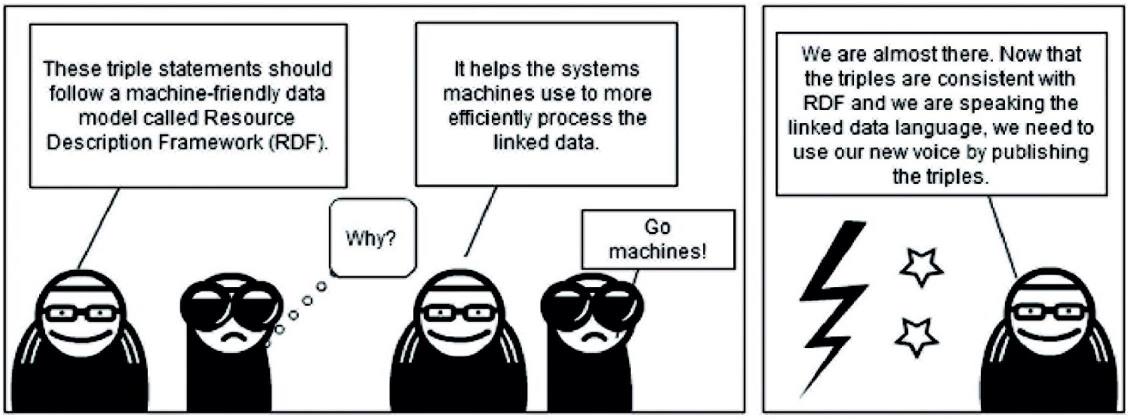
Following the rules for URI assignment a URI is created for the ProvidedCHO. Similarly, since no URI could be found for the Creator, another is created for him. For the concept however a URI is provided by the Library of Congress Thesaurus for Graphical Materials so can be re-used here. Giving (in human readable form) these triples:

http://digital.library.unlv.edu/ProvidedCHO/sho00242 dc:title "Costume design drawing, showgirl in Judy Garland costume, Las Vegas, June 5, 1980"

http://digital.library.unlv.edu/ProvidedCHO/sho00242 dc:creator http://digital.library.unlv.edu/Agent/Menefree_Pete

http://digital.library.unlv.edu/ProvidedCHO/sho00242 edm:hasType http://id.loc.gov/vocabulary/graphicMaterials/tgm002607

Note that for clarity the property labels have been used instead of their URIs.


These triple statements should follow a machine-friendly data model called Resource Description Framework (RDF).
Why?
It helps the systems machines use to more efficiently process the linked data.
Go machines!
We are almost there. Now that the triples are consistent with RDF and we are speaking the linked data language, we need to use our new voice by publishing the triples.

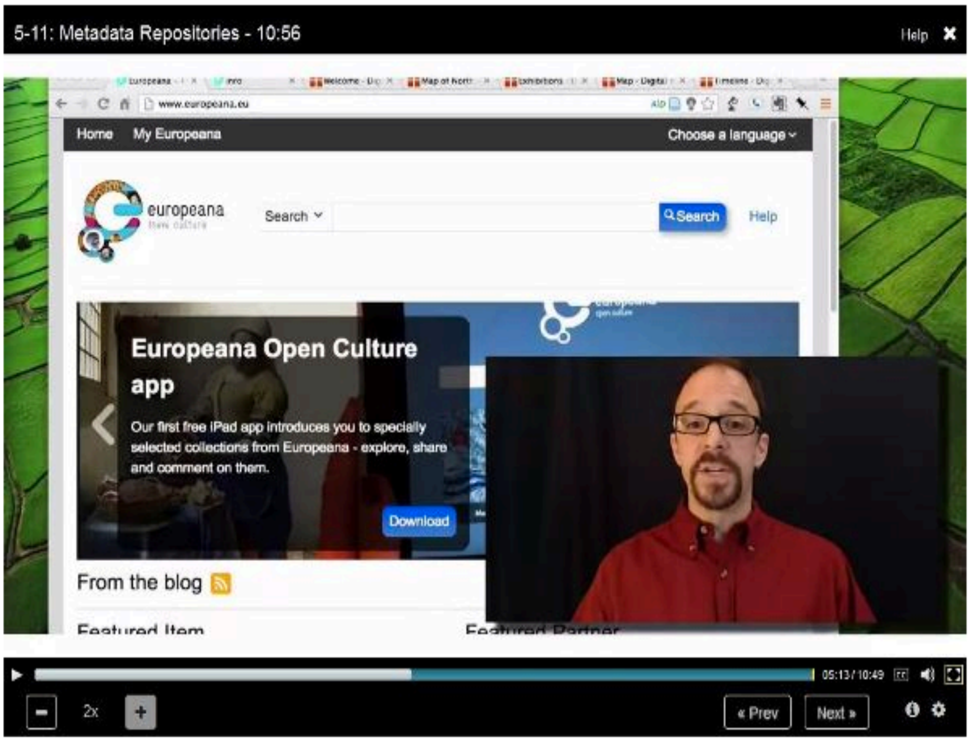
Excerpt from the UNLV comic strip giving the story of RDF

Findings

Discovering that most effort will be centered on the preparation of the data is a valuable finding. In addition is the importance of carefully defining rules for URI creation where necessary or connecting to existing identifiers where possible. The main lesson learned from the project however, is that although linked data is a complex area, librarians can create linked open data using a few simple, free technology tools.

Massive Online Open Course (MOOC) at the University of North Carolina (UNC)

The second educational use of EDM and Europeana records is not a typical EDM case study as it does not demonstrate the large-scale creation of EDM metadata. It is nonetheless interesting as it shows well-modeled metadata in a wider context. EDM was used to demonstrate aspects of metadata records as part of the MOOC run by the School of Information and Library Science from UNC at Chapel Hill. The course was called "Metadata: Organizing and Discovering Information" and was taught by Dr Jeffrey Pomerantz, Associate Professor and the Director of Undergraduate Studies.



Our first free iPad app introduces you to specially selected collections from Europeana - explore, share and comment on them.

Jeff introduces students to the Europeana portal

The course is offered via the Coursera platform using resources provided by UNC.[6] Twenty-six thousand students enrolled for the eight week course in 2013 and a further 15,000 have just finished the 2014 course. It was described in the course overview like this: "Metadata is the unsung hero of the modern world, the plumbing that makes the information age possible."[6] The introduction explains how we use and interact with metadata all the time in our daily lives, generally without realising it, when we interact with digital technology. Citing the use of ATMs, iTunes and Spotify the course places metadata for cultural heritage in the context of other more commercial uses. "We use and even create metadata constantly, but we rarely realize it. Metadata -- or data about data -- describes real and digital objects, so that those objects may be organized now and found later. Metadata is a tool that enables the information age functions performed by humans as well as those performed by computers."[6]

Dr Pomerantz discovered Europeana through colleagues in Europe who were involved in it. Having worked earlier in the field of scholarly publishing he became interested in the idea of enabling access to cultural heritage on a large scale. The Europeana connection was reinforced when he followed the

[6] https://www.coursera.org/course/metadata

development of the Digital Public Library of America[7]. That initiative chose to model some of their operations and metadata on Europeana.

Interviewed for the Europeana Blog[8] he explained why he had chosen Europeana as an example. "I used examples from both Europeana and the DPLA, first, because it allowed me to show two metadata records for the same resource: one from Europeana/DPLA and one from the owning institution. The Europeana records tend to be richer - that is, they have more elements with richer values provided. So metadata records from Europeana were very useful as examples, they allowed me to talk about things like metadata schemas, controlled vocabularies, and manually-created vs automatically-created metadata."

These are aspects of metadata creation and transformation that will be familiar to providers of data to Europeana. Various examples of how they were tackled by different providers can be read in the UNC case above and in the other Case Studies on this site.

[7] http://dp.la

[8] http://pro.europeana.eu/pro-blog/-/blogs/1944597/maximized?p_p_auth=o5BiAw1a