Using open source tools to expose cross collection data in the LIDO schema
CIDOC Helsinki, June 12, 2012

Part I: LIDO at the Yale Center for British Art: modeling collection information for linked open data
Emmanuelle Delmas-Glass, Collection Information Manager
Emmanuelle.delmas-glass@yale.edu

Effective December 2011, the Yale Center for British Art started exposing and vending collection data in the LIDO schema. We believe that we are the first American museum to do so. To date, the data has been harvested by the Google Art Project, ARTstor and the Yale University single search portal, which harvests data from Yale’s 20 libraries, 1 natural history museum, 2 art museums as well as public affairs units, and then builds a central database for cross collection discovery (CCD): http://discover.odai.yale.edu/ydc/.

This paper will be two fold, with the first half focusing on the role of LIDO as the metadata schema supporting research and access to scholarly resources across the YCBA collections as well as paving the way toward linked data. The second half will go over the preparation, tools and techniques used to expose collection information with readily available open source tools. The goal is to offer a roadmap to institutions that are using LIDO on how to share their collections information.

The Yale Center for British Art (YCBA) is both a public museum and a university research institute. It holds the largest and most comprehensive collection of British art outside the United Kingdom, presenting the development of British art and culture from the Elizabethan period to the present day. Its collections consist of about 2,000 paintings, 200 sculptures, 40,000 prints and drawings, 2,000 frames, 30,000 rare books and manuscripts and 30,000 reference library materials. Together the collections provide an exceptional resource for understanding the story of British art.

Architecture:

The YCBA manages and catalogues its art collections in the Gallery Systems TMS 2010 collections management system and the YCBA library materials are catalogued in the Yale University Library central ORBIS system. In terms of architecture/data flows, the YCBA partnered with the Yale Office of Digital Assets and Infrastructure (ODAI). ODAI manages 2 Yale wide systems which feed our online collections catalog: the cross collections repository, which effectively works as a data pump for us, and the Media Manager 7 system. The YCBA has set up a low barrier data exchange protocol using OAI-PMH provider with LIDO metadata for the art collections, and the Yale central Library system set up an OAI-PMH provider with MARC metadata for ORBIS records, including the YCBA’s Orbis records. The data is harvested nightly by ODAI. ODAI also indexes the data and provides Apache Solr interface for access to the data.

YCBA online collections catalogue:

The YCBA online collections catalogue is sourced from the CCD database, although our collections catalogue displays a richer data set than the one exposed through CCD. Only a subset of the data we

provide to CCD is displayed in CCD: for example, our online catalogue displays inscriptions when CCD does not. The YCBA’s online collections catalogue is a single search box where, for the first time in the history of the museum, one can search across all the museum’s collections, i.e. works of art but also rare books, manuscripts and reference library materials.

Transition from CDWA Lite to LIDO:

When the YCBA launched its own cross collections online catalogue in May 2011 the metadata schema behind it for the art collection was CDWA Lite. In December 2011 we switched from CDWA Lite to LIDO. I will explain the rationale between the various steps LIDO enabled us to take, going from the scholarly cataloguing of the art collection to cross collection discovery to linked data. For each of these phases I will discuss the implications for the data management. Clearly each new phase came with increasingly challenging goals and I will go over the metadata choices that allowed us to deliver results.

LIDO as a metadata delivery schema:

Being an XML schema, LIDO is perfectly poised to serve our data exchange goals since we needed to contribute our collection information to Yale’s central metadata repository the way we had with CDWA Lite until now.

LIDO is designed specifically for the purpose of exporting collection metadata and harvesting it to union catalogs. Because we are the only institution on campus to contribute LIDO data to the Yale wide cross collection discovery service (CCD), ODAI, the office maintaining that service had to rewrite our driver. We compromised with ODAI and in exchange we did not request any changes to the CCD interface. We did request, however, that ODAI index more than the data it displays through CCD, such as locations, dimensions, geographic information system data, i.e. new LIDO data that we want to display in our online collections catalogue, either in data fields, facets or both.

LIDO as a support for richer data set:

When we started evaluating LIDO about 2 years ago the goal was primarily to use it as a schema supporting a richer data set than CDWA Lite. LIDO would offer us the possibility to publish more research oriented data about our works of art. At this point we had been conducting our basic cataloguing campaign for 3 years and we were ready to expand the scope of the project from the basic description of our art objects to include events in the lives of these objects.

The same way CDWA Lite had become our de facto metadata structure for basic or minimal cataloguing, conceptually LIDO became our metadata structure for scholarly cataloguing, which allows us to be in step with the research oriented mission of the YCBA.

Events data modeling:
To proceed with a manageable set of data, we decided to target a few of the events suggested by the LIDO event type terminology, namely provenance, publication and exhibition history. The TMS data structure more or less follows CDWA. LIDO’s structure, however, is different from CDWA because of LIDO’s emphasis on events, which are absent from CDWA.

For events in general, we did not have much readily available data in our collections management system as it had been used primarily to record basic descriptive facts about objects and their locations. For example, for provenance, we had to set in place new cataloguing guidelines as well as a research project to extract the information from the curatorial objects files. For exhibition history, we decided to reuse only a subset of the exhibition information entered by the Registrar’s Office: name and date of the exhibition. For publication, our Reference Library has started a few years ago to record in our CMS bibliographic citations whenever our paintings on display in the galleries were either discussed or illustrated in books, articles or videos. Other events, such as relationships to other objects outside of the YCBA collection, will have to be managed differently and I envision the use of the full fledged TMS Events Module, rather than just a field in the Object Module. Clearly, focusing on events means extending the scope of what had, so far, been under the "cataloguing project" umbrella.

Mapping:

Once our collection information had been modeled for LIDO, the next step was to map CDWA Lite and LIDO. Because we had made sure to be standard compliant in our CDWA Lite phase, i.e. we used the data content standard Cataloguing Cultural Objects and data value standards (controlled vocabularies) to produce sharable metadata, and because LIDO is partly built upon CDWA Lite, mapping from CDWA Lite to LIDO was a natural fit.

LIDO as a cross collection discovery tool:

LIDO supports the YCBA’s cross collection discovery goal, which is essential to the YCBA. The museum’s founder, Paul Mellon, did not discriminate between books and works of art when collecting. Indeed he indicated several times that he was quite keen on the cross-pollination of the YCBA collections. Furthermore, prints, for example, are held in our Prints & Drawings department as well as in our Rare Books & Manuscripts department. Consequently, at the YCBA, prints can be catalogued with different standards, which are largely dictated by the cataloguing systems used in the curatorial departments rather than by the type of the object at hand. Nonetheless, the online users should be able to seamlessly find all prints, no matter how they are catalogued. Patrons should not know or have to worry about the data management issues that our data silos have brought on over the years.

Art collections data expressed in LIDO still had to map to our Rare Books & Manuscripts data to at least the extent of what CDWA Lite permitted. In the first phase of our online collections catalogue, CDWA Lite was mapped to MARC whenever possible providing a few crosswalk/common fields as a way to search across all of our collections at once. LIDO was not an impediment to cross collection discovery, to the contrary. Because LIDO is designed to work with other collection types than art collections, it offers more opportunities for cross collection discovery such as provenance and credit line... if only the Yale library policy would permit. Indeed, LIDO might trigger a university policy change so that the Yale Library System can release its holdings records, where books copy ID numbers, work types, credit line and provenance are recorded.
In our online collections catalogue, the All Collections search page is the expression of our LIDO-MARC mapping.

LIDO’s strength in terms of being able to handle data from different collection types proved to be even more efficient in the Yale wide CCD context, where a wider variety of collections is represented. Indeed LIDO uncovered many more cross collection access points than previously known between Yale collections that traditionally have little in common. Our library for examples does not include geographic information system (GIS) data in its MARC records but GIS data is essential to the Yale Peabody Museum of Natural History to track places of excavation for example. The YCBA art collections also indexes places, especially places as subjects. Mapping more data to LIDO allowed the YCBA art collections to share previously unsuspected common access points with anthropological collections, at the GIS level but also through various events.

LIDO as a crosswalk schema:

The other advantage of using LIDO as the Yale wide cross collection discovery service (CCD) schema is that it could potentially provide ODAI with only one schema to maintain to which the CCD partners would map their native schemas. These native schemas include DublinCore, MARC, MODS, CDWA Lite, iTunes, DarwinCore, and YouTube. Because we have tested that, for the scope of data contributed to CCD by the Yale partners, LIDO does work for all the native schemas originally mapped to the CCD schema, LIDO has been in February 2012, on principle, accepted by all CCD participants as the new default CCD crosswalk schema. This is a much better situation for ODAI than having to maintain a complex crosswalk between more than 10 metadata schemas.

LIDO and Linked Data:

Finally we are looking at LIDO to pave the way toward linked open data and offer solutions as to how to harmonize vocabularies and enable a platform for the scholarship of British art through federated query.

The issue of vocabulary harmonization has always been a hot topic at the YCBA because the art collection cataloguing uses both published controlled vocabularies as well as local terms.

The published controlled vocabularies the YCBA uses include Getty Art and Architecture Thesaurus, Getty Thesaurus of Geographic Names, ICONCLASS, Library of Congress Subject Headings, Oxford Dictionary of National Biography, and Getty Union List of Artists Names. The necessity to stray outside of the standards and utilize local terms reflects the YCBA’s intention of describing in a useful way the works of art it cares for as visual objects. Others have resorted to the same policy. The Yale Lewis Walpole Library, for example, which has an extensive print collection that overlaps with ours, uses both Library of Congress Subject Headings and local terms in its MARC records. When the cataloguing project was launched 5 years ago, the YCBA made the conscious decision to utilize controlled vocabularies whenever useful but also to look elsewhere whenever these were not helpful or lacking. Consequently our subject analysis protocol includes an element of tagging, which the Public Cataloguing Foundation defines as “choosing your own words to describe what can be seen in a selection of paintings and classifying paintings in a variety of ways.” (http://tagger.thepcf.org.uk//faq/1175)
In some cases, we have had to develop our own local controlled vocabularies because there were no electronically published ones that fulfilled our needs (the AAT emotion vocabulary for example is rather lacking). This is true also for the highly specialized frame vocabulary that we are developing in partnership with our frame expert, that we will make publicly available.

We manage these local terms the same way we manage published controlled vocabularies: in our collections management system’s thesaurus manager. Our collections management system came loaded with the Getty Art and Architecture Thesaurus and offered a readily available hierarchical organization of terms into which we nest out local terms as well as other published controlled vocabularies.

How to offer these locally managed vocabularies as linked open data resources?

The ideal situation would have been to be able to reference vocabularies already published as linked open data but since none of the vocabularies that we use are, except for the Library of Congress Subject Headings, we had to resort to another option.

Modeling our data for linked open data meant that we carefully respect the Conceptual Reference Model’s expectations (CRM). Our next step was to respect one of the tenets of the CRM that is model our controlled and local vocabularies the way the CRM expects concepts to be modeled: each with a concept source and a unique identifier or conceptID.

After some prodding we discovered that our CMS thesaurus manager provides the conceptIDs of the AAT and TGN terms that come loaded with the system. Again, we could reuse this structure to record the conceptIDs of our local terms and assign them local YCBA conceptIDs.

This presented some interesting system challenges, however, because only a few of our CMS fields are truly linked to our thesaurus manager tables. Classification, Culture, Period, Object Name fields for example are not.

Another issue is that in our CMS constituents data originates from a different field/table whether that person is a creator or a subject. In real life though, an artist can be a painter as well as the subject of a painting. No matter what the role of the artist is, creator or sitter, our goal of course is that this artist gets only one unique identifier/conceptID. We solved this problem by recording the unique conceptIDs of the constituents who are creators in the thesaurus, the way the CMS already does for other concepts, such as AAT keywords. In the case of constituents, because the Library of Congress publishes its name authority file as linked data, we add Library of Congress’ conceptIDs for our constituents. This ensures that searching on J.M.W. Turner for example returns all results, whether if he is the creator or subject. Because our CMS’ thesaurus manager only came with the Getty AAT and TGN loaded in, we had to find a way to add the conceptIDs of new terms. In order to compensate for our CMS deficiencies and provide conceptIDs for all concepts, whether controlled or local, we created an Access tool that allows cataloguers to supplement terms’ entries with their unique identifiers/conceptIDs and vocabulary name/source. This Access tool writes into the thesaurus manager tables, which in turn provide them to COBOAT.

Generating unique identifiers for all concepts is the foundation that allows us to craft URIs for the concepts, URIs which are the essence of linked open data. URIs will allow us to disambiguate one constituent from another, as well as harmonize vocabularies.
Modeling our data for linked open data also meant that we carefully respect the Conceptual Reference Model’s expectation about “false negatives”. Since the CRM is about modeling facts about an object, and anything which is true about it, we have had to model in LIDO only what is compliant with CRM.

Our CMS currently offers only one simple free text field to record inscriptions transcriptions as well as inscriptions descriptions. Consequently, our TMS Inscriptions field carries values such as ‘Inscribed lower right: “Turner”’ as well as ‘no inscription’, which is not to say that the words “no inscription” are actually written on the work of art of course but that we have looked for inscriptions and did not find any. Arguably, this might of value to researchers. In a third yet common scenario, our CMS Inscriptions field is empty because there are still many thousands of prints and drawings which might have inscriptions but which have not been looked at in detail yet. In order to be CRM compliant we have decided for now to not model the false negatives, i.e. the empty inscription field nor the “no inscription”. While this case forces us to think of a way to supplement our CMS, in the meantime, we might be withholding valuable information from scholars.

CONCLUSION:

One of the benefits of LIDO is that it forced us to think of our data as data not just as display. We are now much more agile with our data and not generating long concatenations that cannot be queried, which we did with CDWA Lite. In order to generate data that can be used in a linked data environment we have embraced the CRM expectation of conceptIDs, which is well supported by LIDO. Of course being more thorough with data translates into data management projects, which adds to the metadata creation workflow. It also points out some of our CMS deficiencies. Finally LIDO made us understand that, as the authority on the YCBA collections, the YCBA needs to publish its own URIs for its objects and exhibitions. With time, others will step in the breech and publish more controlled vocabularies as linked data that we will then refer to.

LIDO allows us to meet many goals. It provides an extended metadata schema for scholarly cataloguing, especially through the concept of event, it supports cross collection discovery, and it paves the way for linked open data, which is crucial for harmonizing our data set with other cultural heritage data sets.