Sharing Museum Information: Theory or Practice?

A European Experience

Regine Stein
Deutsches Dokumentationszentrum für Kunstgeschichte – Bildarchiv Foto Marburg
Philipps-Universität Marburg
Germany

Abstract:
Sharing museum information in a broad context – thematic, regional or cross-sectoral – without losing its rich semantics has been much discussed recently. In particular the XML harvesting schema LIDO has been developed to ease the delivery of such information to portals. What is the practical experience of sharing content: Which problems do we meet during implementation, and do we succeed in integrating the data?

The discussion of these questions will be based upon the biggest use case currently under implementation for the museum community in Europe. The ATHENA project is one in a series of projects to build on Europeana as a common access point to European cultural heritage and provides a mechanism for harvesting museum holdings for Europeana. The metadata format used in the ATHENA ingestion process is LIDO. Museum data ingested through ATHENA has to be converted to LIDO.

The Bildarchiv Foto Marburg contributes its own collection of about 800.000 digital photographs on European art and architecture through ATHENA, and provides support for the implementation of LIDO in ATHENA. So the institution is a well placed to reflect on the complete process, starting from practical data conversion of its own information system, the comparison of data coming in from different institutions to final presentation in Europeana.

The paper will try to deduce from this experience practical conclusions on how to prepare museum data for sharing it in a broad context.

CIDOC 2010
ICOM General Conference
Shanghai, China
2010-11-08 – 11-10

1. Introduction

The idea of presenting museum information within a broader context than one's own institution is far from being new – over the last 10-15 years an indefinite number of online services have provided access to cultural heritage information in a thematic, regional and / or cross-sectoral context, not least the "Bildindex of Art and Architecture" run by the author's institution. However with recent efforts to build on Europeana as a common access point to European cultural heritage, digitization and online publication of museum objects has gained a much wider base, and the Europeana prototype provides the biggest use case for analysis of how the sharing museum information can be put into practice in a comprehensive way.

2. The Europeana project framework and ATHENA

ATHENA, Access to cultural heritage networks across Europe, is one of a number of projects run by different cultural heritage institutions within the Europeana framework. ATHENA provides content to Europeana by establishing a mechanism for harvesting museum holdings. It involves partners from over 23 different countries, using 20 different languages, with the objective of supporting and encouraging museums' participation. A set of tools, recommendations and guidelines is produced, and it is hoped that these will be used by museums to support internal digitization projects and to facilitate the integration of their digital content. One major goal is to develop an infrastructure that will enable semantic interoperability with Europeana while preserving museum object specifics.

The data model currently used in the Europeana prototype, ESE, is based on the Dublin Core metadata format. Although initially created strictly for the description of web resources, Dublin Core has become the most common format in cultural heritage service environments. However, the ESE model is not considered as appropriate within the museum community: museum metadata is 'flatten out', with most of the data going into a limited subset of elements. For example, a number of different persons and institutions are usually associated with a museum object: the creator or finder of an object, important persons who have used it, the museum currently holding it, previous owners, and so on. All this qualified information is lost in the ESE format. Moreover, the lack of structure that allows elements to be grouped according to their semantic content leads to substantial information loss. A particular problem is the fact that Dublin Core does not allow information about the object itself and its digital surrogate to be clearly differentiated – the creator of the object appears in the same field than the photographer of its image.

Consequently, the ATHENA workpackage on metadata formats, following a best practice report on metadata formats used by the partners, came to the conclusion that a more appropriate data model for museum information should be used. Since the LIDO development already underway was primarily an effort to harmonize CDWA Lite and museumdat into one single schema, ATHENA decided to join the LIDO initiative and support further development that would subsequently integrate SPECTRUM requirements into the schema. Thus LIDO was chosen as the metadata format for the delivery of museum content through ATHENA to Europeana

3. The LIDO format

LIDO is an XML schema intended for delivering metadata, for use in a variety of online services, from an organization's online collections database to portals of aggregated resources, as well as exposing, sharing and connecting data on the web. The strength of LIDO lies in its ability to support the full range of descriptive information about museum objects; it can be used for all kinds of object, e.g. art, cultural, technology and natural science. Moreover, it supports multilingual portal environments.

LIDO defines 14 groups of information of which just three are mandatory. This allows for the widest and most comprehensive range of information possible. Organizations can decide on how rich – or how light – they want their contributed metadata records to be.

The schema consists of a nested set of 'wrapper' and 'set' elements, many of them repeatable, which organizes information about an object into a tree-like structure. This allows any degree of detail to be recorded in a logically correct, semantically coherent way. An important part of its design is the concept of events, taken from the CIDOC CRM. Information about actors, dates and places related to a museum object is mediated through an event: the creation, collection, and use of an object are seen as events occuring during the object's lifecycle. An exception is events that are depicted or referred to directly, considered as subject matter.

Another important construction principle is the distinction between indexing information that is optimized for searching and retrieval, and display information that is optimized for online presentation. Each information unit contains distinct sub-elements for indexing and display.

The structural elements of LIDO contain 'data elements' which hold actual data *values*. LIDO also allows the recording of information about data *sources* (e.g. in a book) and references to controlled terminology (e.g. the identification code for a term in a thesaurus). Conceptually the information in a LIDO record is organized in 7 areas, of which 4 have descriptive and 3 an administrative character:

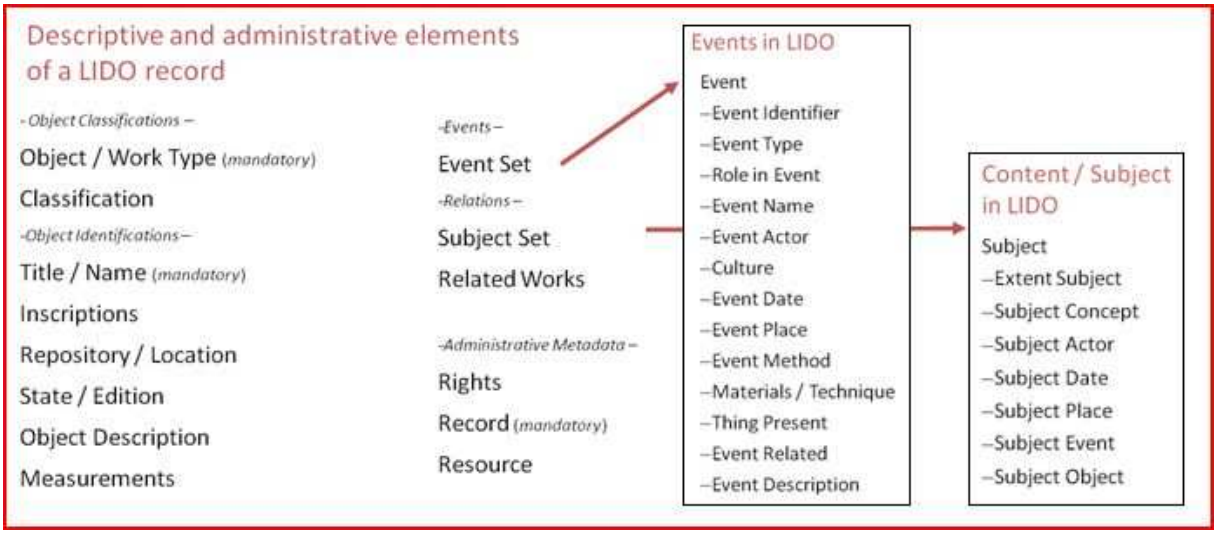

Fig. 1: LIDO overview

The descriptive information section holds:

- Object classification information such as object type and other classifications,
- Object identification information such as titles, inscriptions, repository information, descriptions, and measurements.
- Event information about events where the object was present or in which it participated, such as creation, modification, acquisition, finding, or use. This section holds a number of sub-elements including event type and name, participating actors, cultures involved, date and place information as well as materials and techniques used (typically in the creation/production event).
- Relation information links to related objects, but also to the subject – that is the content of a work: what is depicted in or by a work or what the work is about.

The administrative information section holds:

- Rights associated with the object
- Record information about the source providing the metadata
- Resource information, in particular about digital resources being supplied to the service environment for representing an object online.

The result of a joint effort of several international key institutions and groups dealing with museum documentation standards, e.g. the CDWA, museumdat, SPECTRUM and CIDOC CRM communities, the release of LIDO v1.0 during this year's CIDOC conference can be seen as a clear reward to the community. It provides a single, common schema for contributing content to cultural heritage repositories. This enables museums and other content providers, using different data structures and software systems, to express and deliver a wide variety of information in a standardized and machine-readable format. Furthermore, this information can easily be accessed, harvested and recontextualized by semantic-aware services. Apart from the exciting promise of new applications, LIDO promises time- and cost-savings for museums interchanging object information in different daily work contexts.

4. Contributing content I: Bildarchiv Foto Marburg

As part of the Philipps University in Marburg, the German Documentation Center for Art History "Deutsches Dokumentationszentrum für Kunstgeschichte - Bildarchiv Foto Marburg" is a national and international research and service institute. Its mission is to collect, index and make available photographs related to European art and architecture, as well as to conduct research on the history, practice and theory of how visual cultural assets are transmitted. Holding roughly 1.7 million photographs, Foto Marburg is one of the largest image archives on European art and architecture. Through the cooperative structures it has established, Foto Marburg supports the documentary work of museums, offices for the protection of historic monuments, libraries and research institutes and serves the community by publishing the pictorial material and the indexing data of more than 80 partner institutions.

Within the ATHENA project Foto Marburg has contributed to Europeana object descriptions and related digital images from its own photographic collection; 326.608 LIDO records describing distinct objects, accompanied by 796488 digital images

providing different views and details of these objects. This is particularly relevant for architectural objects, but also for complex art objects such as triptychs consisting of multiple paintings on several panels. Although collecting only photographs, the Foto Marburg documentation system focuses on detailed description and indexing of the work of art or architecture itself: its creation and modification history, its provenance and its visual contents. Multiple photographs, offering distinct details and perspectives, as well as historical views over time, are attached to the object record. Each image is associated with specific resource information such as photographer and date taken.

The task of mapping our own data to LIDO, with the objective of including as much information as possible and avoiding any loss of granularity, has been a challenging piece of work. It requires analysis not only of the full data structure, but also of how these data fields have been filled. Even with a documentation system based on a standard, such as in our case the (German) MIDAS standard, everyday indexing practice tends to establish collection-specific, implicit rules and preconditions, which have to be repected in the mapping.

The fundamental task is to identify which data elements or groups of elements in the source structure correspond directly tot LIDO elements information groups, and which source elements have a qualifying character: their data values having a direct influence on the choice of LIDO target. Consequently a *conditional* mapping is needed. This is particularly important for the grouping of events, e.g. That nature of an event can often be deduced from the role of an associated actor. A commonly used data structure, also found in Foto Marburg's data, is to use a specific data field for the name of an object's "Creator", while placing date and place information related to the act of creation in general date and place fields along with qualifying sub-elements values such as "Creation", "Find", "Use". These sub elements can be used to regroup the information into the event-based LIDO structure.

The mapping and data conversion tasks have now been successfully accomplished: the relatively complex original information structure has been converted into the LIDO structure and the data has been submitted to ATHENA via an OAI-PMH harvesting service.

5. Contributing content II: The ATHENA mapping and ingestion process

Turning now from the perspective of a single institution with considerable experience in merging heterogenous data, the question arises as to how manageable the mapping and ingestion process is for content providers who have only recently started sharing their data in a wider service environment. To facilitate this process a mapping tool has been developed by the technical partner of the ATHENA project, the National Technical University of Athens.

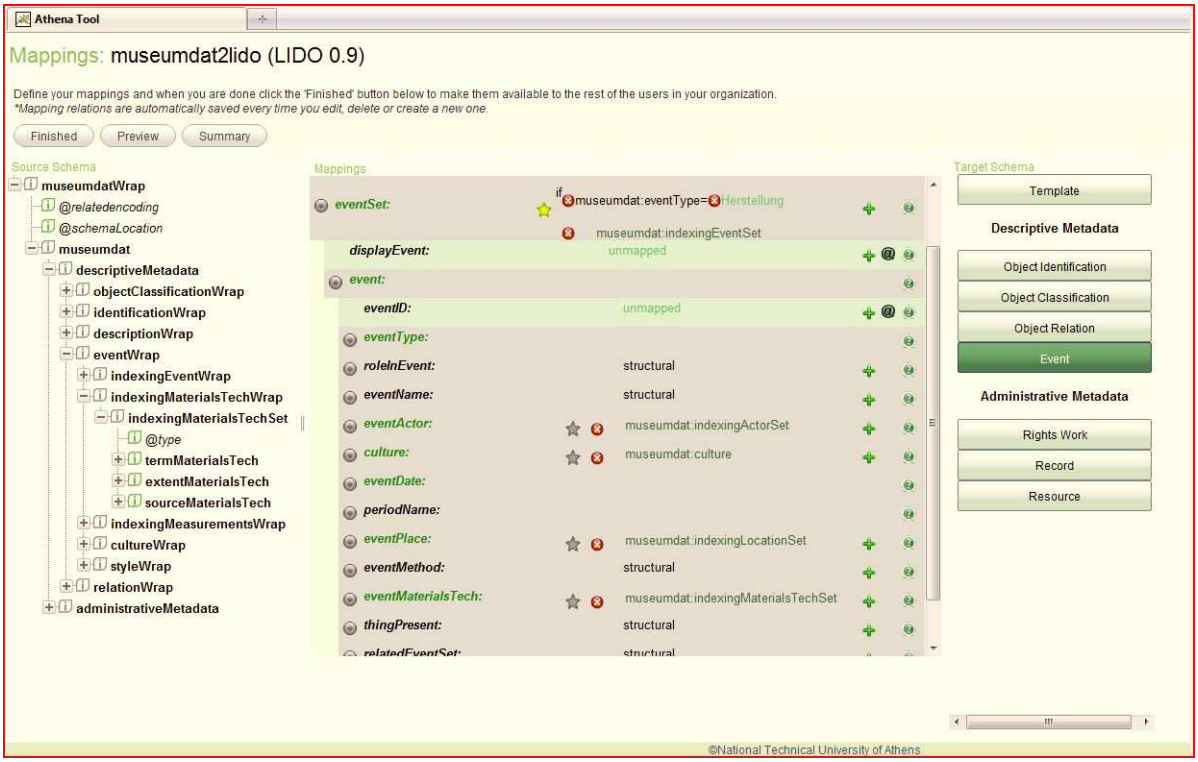

Fig. 2: ATHENA mapping tool

Any kind of data provided in an XML format can be loaded into the system. The tool then visualizes, on the left, the incoming source data structure and, on the right, the LIDO target schema. The content provider can then map its source data fields through drag and drop to the target fields, including mapping of structural elements holding no data, and conditions for the mapping and concatenation of data values and constants. A helpdesk mailing list allows users to ask questions about the format and the tool, and to help each other.

Combining a comprehensive metadata format with a customized technical solution for practical mapping is an exciting effort. It enables semantic interoperability of content from many different collections and from different management systems with different data structures. It is difficult to evaluate how the process will evolve over the next few months of the ATHENA project's activities and beyond, but some preliminary statements may be given here for discussion, both, positive and instructive. The overall mapping results are good and the questions on the helpdesk list comprehensive, so users appear to have grasped, from the material and the tool provided, both the LIDO schema and how to map to it.

Yet to get to a meaningful mapping that best reflects the source information in the target schema, several feedback loops are often needed between the local expert, who knows the source schema and content very well, and a LIDO expert who knows the LIDO structure in depth. This loop is considerably shortened by the ATHENA mapping tool, the result of a close cooperation between LIDO schema developers and technical implementers, which reflects the target schema very clearly. The process is considerably easier if the source schema is based on a documentation standard such as SPECTRUM, CDWA, or national standard. Moreover, features supporting data analysis and data value statistics, provided in the mapping tool, help immensely in this process. This kind of quality management is crucial and may be further developed.

Overall it seems that it is both appropriate and simpler for content providers to map their data to a well structured metadata format, instead of randomly choosing some corresponding field in a flat structure such as ESE.

Presently, LIDO serves in ATHENA as an intermediate layer between source formats and the DublinCore-based ESE format. It thereby provides a more standardized representation of museum collections in Europeana. Since the ESE format does not support the fine granularity of museum information and fails to make a clear distinction between the museum object itself and its digital surrogate in an online service, standardized presentation helps to improve search and display quality considerably. Beyond this LIDO effectively prepares the ground for new, data quality focused approaches.

6. The European experience: Conclusions

ot entirely surprisingly, here is a close connection between the level of control initially practised in a source format, e.g. in data structure and data values, and its comprehensive mapping to a standardized harvesting format. So try to think of your data, from the outset, as being used outside of your own home context. The ease of connecting your research information with other sources increases immensely when data structure and terminology standards are used.

LIDO, considered as a format for delivering machine-readable data, is an important piece in the whole framework, but standing alone it does not guarantee interoperability. It does not resolve the issues of multilingualism in data provided across 20 different European countries using 20 different languages. In the first instance, this is a question of data value control.

To evaluate the use of LIDO within the Europeana service environment it will be crucial to see the practical implementation of the new Europeana Data Model, EDM. EDM will supplement and enhance the currently used ESE model with a meta-structure that truly allows the LIDO format to be retrieved. It is a clear expectation that the implementation of this data model will significantly improve resource discovery, providing more precise search results that carry meaningful links to associated resources.

Developers of standards such as LIDO will have to focus particularly on providing documentation and training for the standard, and supporting museum practitioners as well as technical expert users such as software developers. Used in conjunction with increasing opportunities to participate in linked data environments, this will enable museums to recontextualize their collections in a meaningful way and hence improve understanding of these collections within the greater cultural heritage context.

Visit the websites of discussed projects here:
http://www.lido-schema.org/
http://www.bildindex.de/
http://www.athenaeurope.org/
http://www.europeana.eu/