LIGHTWEIGHT INFORMATION DESCRIBING OBJECTS (LIDO): THE INTERNATIONAL HARVESTING STANDARD FOR MUSEUMS

LIGHTWEIGHT INFORMATION DESCRIBING OBJECTS (LIDO): THE INTERNATIONAL HARVESTING STANDARD FOR MUSEUMS

edited by
ATHENA WP3
Working Group
"Identifying standards
and developing
recommendations"

texts by
Gordon McKenna,
Collections Trust (UK)
Stefan Rohde-Enslin, Institut
für Museumsforschung,
SMB-PK (Germany)
Regine Stein, Bildarchiv Foto
Marburg (Germany)

The text of this booklet
is included in deliverable
3.3 of the ATHENA project.

ATHENA
general co-ordinator
Rossella Caffo

design
mt milani,
geo graphic sdf

ATHENA logo
Susan Hazan

LIDO name and logo
courtesy Rob Lancefield

web version
http://www.athenaeurope.
org/index.php?en/110/
promotional-material

table of contents

Introduction

It is the aim of the ATHENA project to support especially museums in providing object data for publication in Europeana.

ATHENA carried out a survey into which standards are in use with the partners in the ATHENA project especially and in museums in general. In the museum landscape there is a great variety of standards used. The survey showed that the harvesting formats CDWAlite and museumdat were in use.

ATHENA learnt that there existed a transnational CDWAlite/museumdat Working Group with the aim to further develop CDWAlite and museumdat to a joint standard "LIDO".

ATHENA also contributed to the development of LIDO by aligning it to the SPECTRUM schema.

LIDO is used as the format to aggregate museum data and transform and deliver them to EUROPEANA.

We hope that by applying the LIDO standard it will be easier for museums to provide their data to Europeana and other cultural heritage repositories.

Monika Hagedorn-Saupe

WP2 - Awareness and dissemination:
Enlarging the network and promoting the service

1. Why is LIDO needed?

> *"A museum object is more like an illustration, or witness of the past, than information in its own right. Cultural historical research means understanding 'possible pasts', the facts, events, material, social and psychological influences and motivations. It lives from understanding contexts, by pulling together bits and pieces of related facts from disparate resources, which can typically not be classified under subjects in an obvious way. It lives from taking into account all known facts."* [1]

In this statement the author, Martin Doerr, gives the case for the integration of **rich cultural information**.
The Mellon funded *ResearchSpace*[2] project within which this paper was written itself is "*aimed at supporting collaborative internet research, information sharing and publication for the cultural heritage scholarly community*".

To solve these challenges let us examine the three **metadata use environments** in which integration is taking place:

1 Martin Doerr,. *Technological Choices of the ResearchSpace Project*, ResearchSpace website, August 2010, see: https://sites.google.com/site/rspaceproject/researchspace-concepts/technological-choices-of-the-researchspace-project.

2 See: https://sites.google.com/site/rspaceproject/.

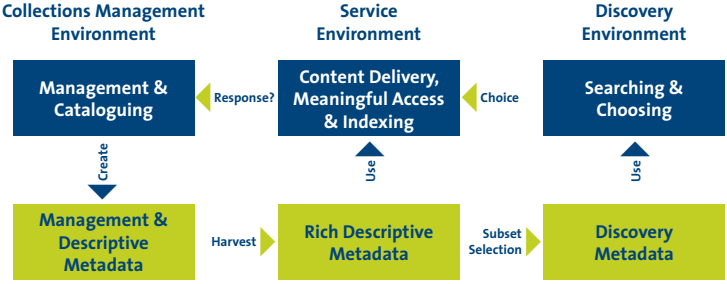


Figure 1. ***Metadata use environments***

1.1 Collections management

This is where metadata is **created**. The information recorded comes from a number of sources:

- Collections management activities of the organisation (for example: acquisition; loans; conservation, rights management and use)
- Descriptions of the object itself (for example: type; title; material; dimensions; subject of intellectual and visual content)
- Connections to events during its existence (for example: creation; field collection; use and association)
- Connections to persons, organisations, and places during its existence (these are often intimately connected to the events mentioned above).

Usually takes place at the collection holding organisation, within their own systems, and with a lot of human effort. Key concepts for metadata in this environment are:

- Maximum detail (all the relevant data)
- Preservation (of data)
- Domain specific schemes (museums, libraries and archives use different metadata schemes)
- Country specific schemes
- Organisation specific schemes (these might be in-house or adaptations of standards).

1.2 Service

This is where users are given **meaningful access** to a **single** piece of metadata describing an object or other piece of cultural material. Delivery usually includes a digital surrogate for the material.

Key concepts for metadata in this environment are:

- Cross-domain (probably contains material from more than one)
- Usable quality (for service being offered – often audience specific)
- Reasonable speed of delivery
- Rights protection (copyright statement or technical means).

Metadata here is a subset of the metadata in the collections management environment and should ideally be **harvested** from there.

This environment should also provide a means for collecting a user's response to the object which could feed back information into the collections management environment.
For example additional information about the content of a photograph might be provided by the user of the service which was unknown to its owning organisation.

1.3 Discovery

This is where users are given access to a set of metadata from many objects. Delivery is usually part of the result set of a search together with a thumbnail of some kind. Users **choose** a content they want to look at in the service environment.

Key concepts for metadata in this environment are:

- Cross-domain
- Maximum relevance of results
- Speed of choosing relevant resource (limited set of metadata elements).

Metadata here is a subset of the metadata in the service or collections management environments.

The appearance of the service and discovery environments in an organisation's website, portals, aggregators and Europeana is the same as that for the digital content.

From the above it can be seen that t he potentially rich metadata that is harvested from the collection management environment has the key role in providing a good service for users. The question that needs to be answered as a result of this analysis is: *Which metadata scheme should be used?*

Dublin Core (DC), in some version or other, is the commonly used metadata schema in both the service and discovery environment.

However the work of the ATHENA project questioned its use for museum content especially in the service environment.

There is a common view within the museum community that DC derived metadata schemas do not deliver a rich enough view of museum content. The importance of a museum object, especially outside the area of fine art, is often not covered adequately. DC-based systems 'flatten out' museum metadata, with most of the data going into 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 a Dublin Core based 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.

Different rich data will end up in the same element in a simple DC-based.
Also there is a loss of the relationships between the different 'who', 'what', 'when'

and ‘where’ classes of data and the events they relate to. So it becomes difficult to query the data in complex ways. Finally the ATHENA survey revealed the lack of standard DC-based metadata scheme, built into its design which allows for extensions. This is a barrier to interoperability.

It was to overcome this situation that LIDO was developed.

2. LIDO's background

LIDO – Lightweight Information Describing Objects Version 1.0 was delivered to the community during the ICOM/CIDOC conference in November 2010 in Shanghai/China.

Its definition in an XML schema,[3] together with the specification document,[4] can be found at http://www.lido-schema.org.

LIDO is the result of a collaborative effort of international stakeholders in the museum sector, starting in 2008, to create a common solution for contributing cultural heritage content to portals and other repositories of aggregated resources.
Being an application of the CIDOC Conceptual Reference Model (CRM) it provides an explicit format to deliver museum's object information in a standardized way.

Work that eventually led to LIDO started with the J. Paul Getty Trust and ARTstor developing CDWA Lite, an XML schema

3 See: http://www.lido-schema.org/schema/v1.0/lido-v1.0.xsd

4 See: http://www.lido-schema.org/schema/v1.0/lido-v1.0-specification.pdf

for describing cultural materials and their visual surrogates, to provide an easier and more sustainable model for contributing to union resources. It advanced with the Working Group Data Exchange of the German Museum Association's development of museumdat. Its main achievement was to generalize the scope to include the full range of object data, such as cultural history, natural history, or history of technology, and to include multilingual needs.

Led by the CDWA Lite Advisory Committee and the Documentation Committee of the German Museums Association, it was agreed to create a single schema that met the requirements articulated by CDWA Lite, museumdat, and feedback received from the greater community of information and technology professionals. As part of this effort, compliance with CIDOC-CRM was a major requirement. A working group was established for the development of LIDO.

Resulting from the report on existing standards applied in European museums, it was concluded, within the ATHENA project, that a metadata format for ATHENA would have to meet the needs of both museumdat and SPECTRUM. Consequently ATHENA decided to join the LIDO initiative and support further development that would subsequently integrate SPECTRUM requirements into the LIDO schema.

Here is an overview of standards and existing formats LIDO is based upon:

CDWA Lite

is an XML schema for encoding core records for works of art and material culture based on the data elements and guidelines in Categories for the Description of Works of Art (CDWA) and following the data content standard Cataloging Cultural Objects (CCO), provided by the J. Paul Getty Trust and ARTstor. It is intended as a low-barrier way to enable institutions to contribute their collections information to union catalogs using the Open Archives Initiatives Protocol for Metadata Harvesting (OAI/PMH). For more information see: http://getty.art.museum/research/conducting_research/standards/cdwa/cdwalite.html

museumdat

is an XML schema provided by the Documentation Committee of the German Museums Association, which builds largely upon CDWA Lite, but overcomes its specific focus on art mainly by a reconfiguration of the CDWA Lite elements that takes into account the event-oriented multi-disciplinary approach of the CIDOC Conceptual Reference Model.
For more information see: http://www.museumdat.org/index.php?ln=en

SPECTRUM XML schema

is based on the UK and international standard for collections management

with the same name from the Collections Trust. It provides a format for exchanging object records between different collections management systems and aggregating data. For more information see: http://www.collectionstrust.org.uk/spectrum

CIDOC CRM (ISO 21127)
provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation. It is intended to be a common language for domain experts and implementers to formulate requirements for information systems and to serve as a guide for good practice of conceptual modeling. Work is carried out through the CIDOC-CRM Special Interest Group. For more information see: http://www.cidoc-crm.org/

Therefore LIDO is the result of the joint effort of the CDWA Lite, museumdat, SPECTRUM and CIDOC CRM communities.
Being CIDOC CRM compliant, the schema combines the CDWA Lite and museumdat schemas, is informed by SPECTRUM.

3. LIDO's home

The information and activities that are related to LIDO are centralized within the Data Harvesting and Interchange Working Group of CIDOC[5], the International Committee for Documentation of ICOM[6].

The Working Group was installed at CIDOC's Annual General Meeting 2009 and serves as a place to bring together documentation and material for LIDO. The Working Group provides a history of developments and serves as the source for the most up-to-date information about advancements with LIDO, and as a place to evaluate its progress, implementations and understanding within the greater community.

Thus the Working Group brings together all the communities involved – CDWA Lite, museumdat, SPECTRUM and CIDOC CRM – in a committee that can sustain LIDO's understanding, advancement, exposure, and applicability. The group welcomes the participation from the wider heritage community and acts as the forum

5 http://cidoc.icom.museum

6 http://icom.museum

for LIDO's implementation. Working Group members are involved in several projects, expert groups, and local communities implementing LIDO and vice versa.

As central access point for information about LIDO a website is established at http://www.lido-schema.org.

More generally the Data Harvesting and Interchange Working Group provides information about, and supports the development and application of formats and techniques for harvesting and interchanging information of relevance to the museum community.

4. When to use LIDO

Organizations need to provide information on their objects to many portals or aggregators including those that are thematic, cross domain, regional, national and international. The difficulty, for the portal or aggregator owner, is that the object information is in the providers' own collections management systems and cataloguing databases. Each of these has potentially a different metadata format.
This means that it is both time consuming and costly to integrate information from all those organizations wanting to participate.

Using LIDO to provide a harvestable set of an item's descriptive metadata enables the organization to participate in such initiatives in a standard way.
LIDO provides a cost effective solution for the:

- Providing organisation, which can supply the same data to many portals and aggregators
- Portals and aggregators, which can be supplied with data from many organisations.

In both cases the benefit is gained by not having to carry out multiple metadata mappings.

LIDO is a 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.
It is not intended to be used as a basis for a collection management system or to support loan and acquisition activities.

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, architecture, cultural history, history of technology, and natural history).

Finally LIDO supports multilingual portal environments. It does this by having a language attribute that can be associated with each element, or more generally, with the group of descriptive elements for fully multilingual resources.

5. LIDO - the basic design principles

The following section is not intended for the sole use of an IT specialist. Neither is it intended here to present all details of the LIDO schema. The aim here is rather to give an overview of the construction principles of the schema and its general sections – and to do this in a way that all cultural heritage practioners will be able to understand why LIDO is the way it is.

To achieve the necessary flexibility, to meet the needs of various portals, LIDO organizes information in 'sets'. The "name of an object" for example is such a set of information. Let us assume that an organization has a dried hemp plant natural science specimen which has a 'preferred name' of "Hemp Plant". Conceptually in LIDO the information is stored in a set:

Object-Name-Set: Object-Name = Hemp Plant

Such sets in LIDO can easily be duplicated. Therefore the organization can have a second Object-Name-Set to record the specimen's 'scientific name':

Object-Name-Set: Object-Name = Cannabis Sativa

There might be even more names for any object. So a 'common name' might be:

Object-Name-Set: Object-Name = Ganja

One might expect there is confusion in allowing many names for the same object. In LIDO this is easily avoided. The different names are given a 'qualifier' (a type attribute). A more complete picture looks like this:

Object-Name-Set -> type: general name: Object-Name -> pref: preferred = Hemp-Plant
Object-Name-Set -> type: scientific name: Object-Name -> pref: alternate = Cannabis Sativa
Object-Name-Set -> type: colloquial name: Object-Name -> pref: alternate = Ganja

To further avoid confusion the Object-Name-Sets are grouped together, or 'wrapped', so we have:

Object-Name-Wrap-Begin
Object-Name-Set -> type: general name: Object-Name -> pref:preferred = Hemp-Plant
Object-Name-Set -> type: scientific name: Object-Name -> type: alternate = Cannabis Sativa
Object-Name-Set -> type: colloquial name: Object-Name -> type: alternate = Ganja
Object-Name-Wrap-End

This is the basic construction principle of LIDO:

1. A fine grain piece of information is recorded in a 'set'
2. The set is qualified and if necessary duplicated
3. Sets of the same kind are 'wrapped'.

Does this principle meet the aims of LIDO of taking all information from all objects and delivering it in a form that all portals might use it? While one organization's database might have one field for the names of objects another might have two fields, one called "common name", the other called "scientific name" ... in both cases LIDO can take the information and pass it to any portal. The receiving portal than can decide if it uses one or all of the names.

To be able to record multiple names for the same object really is important. With biological objects one may argue that the scientific name is the 'real name'. However what about paintings and other things?

There is a famous painting by Vermeer in the Rijksmuseum in Amsterdam. In the museum this painting is called "*Het melkmeisje*". There is also an English name for it "The kitchen maid" or "The milkmaid" and a German name "Die Milchmagd", a Spanish name "La lechera", and so on. These many names, in many languages, have possibly been established for many years.

With LIDO this information can be passed on to the portals. Simply repeat the name in different languages and qualify it with a language attribute (using the standard two-letter codes). The Rijksmuseum might give the information on Vermeer's painting as:

```
Object-Name-Wrap-Begin
    Object-Name-Set-Begin
        Object-Name -> lang: dutch = Het
            melkmeisje
        Object-Name -> lang: english = The
            Milkmaid
        Object-Name -> lang: german = Die
            Milchmagd
    Object-Name-Set-End
Object-Name-Wrap-End
```

More than one qualifier for a name set is possible. This way the many English names of Vermeer's painting can be passed on using LIDO. This situation would look like this:

```
Object-Name-Wrap-Begin
    Object-Name-Set-Begin
        Object-Name -> lang: dutch –> pref:
            preferred = Het melmeisje
        Object-Name -> lang: english -> pref:
            preferred = The Milkmaid
        Object-Name -> lang: german -> pref:
            preferred = Die Milchmagd
    Object-Name-Set-End
    Object-Name-Set-Begin
        Object-Name -> lang: english -> pref:
            alternate = The Kitchen Maid
    Object-Name-Set-End
Object-Name-Wrap-End
```

The principle is always the same. Theoretically LIDO can pass, to a portal, an infinite number of names for any one object. In real life it is rare to have many names for one object in the organization's own database.

The advantage with LIDO is that on the one hand it can take the information the organization's database holds and organize it in a structured way, but on the other hand it does not demand more than one single name for each object.

With the increase in the use of database driven multilingual websites the number of museum databases with more than one object name will grow. Also the numbers of databases which have object descriptions in more than one language will grow. Some already have it.

Again the basic LIDO design principle applies. Put a description translated into different languages into one set and assign each its language qualifier. Take the next description and qualify it accordingly. For instance, there might be one of the descriptions of a 'scientific' type, then assign this type attribute, and take the other description and qualify it as 'educational' type, and so on.

Support for multilingualism is a powerful feature of LIDO. The language qualifier can be put to each single element as shown above, or it can also be put at a higher level. So instead of qualifying each single element

it is possible to declare something like: "All of the following information is in language ABC except otherwise indicated". This makes life much easier!

Another powerful feature of LIDO is its ability to distinguish between 'display elements' and 'index elements'. The easiest way to explain the usefulness of this distinction is to look into how time information may be recorded. A point in time or a dates time span can be named very differently e.g. "The 30 years war" or "1618-1648" or "first half of 17th century" or "after 1617 and before 1649" or even "early Baroque". Because it is the aim of LIDO to be useful for all kinds of organizations it has to be able to accept and transport all these alternatives, no matter that they come from a history museum ("30 Years War"), a literature-museum ("first half of 17th Century"), or an art museum ("early Baroque"). In the database of an organization one of these ways to name a given time span is used and probably the organization wants its own naming to be displayed together with its object in a portal.

However portals can only meet the needs of their user if they make data contained in them searchable. So how can a successful search be made when data is, potentially, in many different forms? If a user is looking for objects between 1618 and 1648 how can he or she find an object whose creation date was given as "early Baroque" by the providing organization? This can be achieved by allowing

'indexing elements'. These indexing elements are used to define a point in time or a time span by simply defining an 'earliest date' and a 'latest date'. So for a time specified by an organization as "30 Years War" you would get:

Display.date: 30 Years War
Indexing.date.earliest: 1618
Indexing.date.latest: 1648

In case of a creation date of "1618-1648" the LIDO data would look like:

Display.date: 1618-1648
Indexing.date.earliest: 1618
Indexing.date.latest: 1648

Of course LIDO can only represent data in this way if it is inside an organization database in the same way. LIDO can then be used to pass the data to portals, and each portal can use the display element for display and the indexing elements for its search engine. Where an organization only has indexing elements or only display elements then LIDO can still be used.

Some organizations hold rich data or 'enriched data'. For example, an organization might not only have a location specified as "Paris", but it might also have information about which "Paris" is meant. "Paris" is ideally enhanced by referencing it in a standard terminology. The Getty's *Thesaurus of Geographic Names* (TGN) is often used, and the "Paris" which is the capital of France has a unique identity number there. LIDO can store

this information from the organization's database and pass it to a portal. Doing this again LIDO uses the basic principle described above:

Place-Name-Set-Begin
standard terminology ID -> source: TGN = 7008038
standard terminology term = Paris
Place-Name-Set-End

If another standard terminology is used in addition to TGN, e.g. the GeoNames Service, then that can be represented:

Place-Name-Set-Begin
standard terminology ID -> source: TGN = 7008038
standard terminology ID -> source: geonames = 2988507
standard terminology term = Paris
Place-Name-Set-End

Also the language of the data can be represented as above:

Place-Name-Set-Begin
standard terminology ID -> source: geonames = 2988507
standard terminology term -> lang: english = Paris
Place-Name-Set-End

Using a standard terminology, especially for persons, institutions and places is very valuable for portals. Giving the source, and the ID inside the source, defines a place more exactly than does a place name alone,

and it offers some more functionality. With the reference given as, e.g. “GeoNames” with an ID, then the portal’s database can be enriched by using GeoNames, which holds information about the name of “Paris” (the capital of France) in about 50 languages, including non-Latin scripts. Organizations can do the same and enrich their own databases automatically using these standard terminology sources. Best practice is for the organization to record the standard terminology used and the IDs in that source for all the database entries for places and persons. If this is followed LIDO is able to transport such data to a portal.

The examples above were about place and time. The same principles apply to persons or institutions. But there is something special in LIDO with regard to place, time or person-specification.

Many older collections management systems (CMS) only offer the possibility to store one date, or date range, per object. However the reality is often more complex. For example let us look at an art nouveau vase made from glass. The object’s name might be “Art Nouveau Vase” and the associated time span called “art nouveau”. However this Vase was designed by Daum Frères in Nancy, France, in the year 1904 but it was made by some unknown company in 1907. With a more modern CMS it is possible to store this additional information too. With archaeological objects, e.g. a flint axe, multiple dates become a more obvious need. Such an axe was made at a certain time,

e.g. 5000 BCE, and was excavated at another time. It was made at a place that is not necessarily the same place where it was found; and it was made by someone who for sure is not the one who excavated!

It was to take into account the multiplicity of associated dates, places, and actors that one of the main features of LIDO, its representation of 'events', came into being. The designing of the art nouveau vase would be such an event and its production would be a second event. Similarly the making of the flint axe would be one event, and the excavation of the same axe would be a second. Other types of events are possible, e.g. the use of an object. Such events are basically aggregations of information about an event: when the event took place, the place where the event happened, and the actors involved in the event.

Thinking in events has the advantage, conceptually, of allowing many dates or date ranges, places and actors to be associated with an object, whilst always keeping unrelated things apart. Therefore, for example, the exact relationship between an object and a person is always made clear. In the above example, the art nouveau vase, the LIDO would look like this:

Event-Set-Begin
Event-Type -> Designing
Event-Time -> 1904
Event-Place -> Nancy
Event-Actor -> Daum frères
Event-Set-End

Event-Set-Begin
Event-Type -> Production
Event-Time -> 1907
Event-Set-End

It is not necessary to have all information about it for representing an event in LIDO. It is sufficient to give the event 'type' and one additional piece of information, e.g. when the event took place. Only an event type would be meaningless, but correct, data.
It would be like a statement that said "This art nouveau vase was designed".

Events can be applied to all kinds of objects in all kinds of organizations. A production event for the flint axe could look like:

Event-Set-Begin
Event-Type -> Production
Event-Time -> 5000 BCE
Event-Set-End

Compare this to the form used to describe the production event of the art-nouveau vase. The structure of the data is identical. This is what LIDO is made for – being able to take data from all types of databases, in all types of organizations, and with all types of collections (art, archaeological, technical and natural science). With LIDO every organization can simply export their data into a standard structure. Each portal can incorporate this data easily because it can rely on LIDO's standardization.
An organization has to write only one export routine; from CMS to LIDO

(some commercially available CMS have a LIDO export already available). A portal has to write only one import routine instead of writing one routine for each organization it gets data from.

LIDO can transport all information relating to an object which might be worth for a portal to incorporate. The emphasis here is on 'can'. If an organization does not have a certain piece of information or does not want it to be handed on to a portal, it still can use LIDO. There are only very few mandatory elements in LIDO:

- The "object work type" which is the kind of the object, e.g. painting, drawing, printing machine
- The "object name" (or "object title") for example, "The kitchen maid", "At the banks of river; river Thames (Drawing by Alfred McCoy)", or "Printing Machine: Heidelberg KOR 40x57 cm"
- The "record identifier" together with the "record type" and the "record source" which is the name of the institution where the information is coming from.

Any LIDO record for an object must have these three pieces of information. All other information is optional. Of course an organization will want to give much more information to a portal than the minimum; however the exact types and amounts are at the discretion of the organization and the requirements of the portal.

All information inside LIDO is stored in five 'blocks'. These are:

- Object Classification – Containing: Object/Work Type, and Classification
- Object Identification – Containing: Title/Name, Inscriptions, Repository/Location, State/Edition, Object Description, and Measurements
- Events – Containing the above described events in a detailed structure
- Relations – Containing: Subject / Associations (these may be concepts, but also actors, places, events, or other objects), Related Works
- Administrative Metadata – Containing: Rights Information, Record Information and Resource Information.

These blocks are the superstructure for LIDO, and each sub-block has its own structure. It is not the purpose here to describe LIDO in detail. For explanations of the single elements please see LIDO website.

6. Practices with LIDO: the ATHENA experience[7]

6.1 ATHENA's choice of LIDO

The data model 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, a Dublin Core based model is not considered as appropriate within the museum community: museum metadata is 'flattened out', with most of the data going into a limited subset of elements. Consequently, the ATHENA project, following its own 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.
As a result ATHENA joined with the LIDO initiative to support further LIDO development. This developed LIDO was chosen as the metadata format for the delivery of museum content through ATHENA to Europeana.

7 This section is largely based upon the paper "Sharing Museum Information: Theory or Practice – A European Experience", given by Regine Stein at the CIDOC 2010 conference in Shanghai, November 10th 2010. For the full paper refer to http://cidoc.meta.se/2010/full_papers/stein.pdf.

6.2 The ATHENA mapping and ingestion process

The question that arises is: How to manage the mapping and ingestion process for content providers? To facilitate this process, particularly for content providers who may have only recently started sharing their data in a wider service environment, a software tool was developed by the technical partner of the ATHENA project, the National Technical University of Athens (NTUA).

Figure 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 their source data elements through 'drag and drop' to the target fields, including the mapping of structural elements holding no data, and conditions for the mapping and concatenation of data values and constants. A helpdesk mailing list allowed users to ask questions about the format, 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. The mapping results within the ATHENA project, showed that users appear to have grasped both the LIDO schema and how to map to it very well.

Yet the task of mapping data to LIDO, with the objective of including as much information as possible and avoiding any loss of granularity can be challenging. It requires the analysis not only of the full data structure, but also of how these data elements have been filled. Even with a documentation system based on a standard, everyday indexing practice tends to establish collection-specific, implicit rules and preconditions, which have to be reflected in the mapping.

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

So to get to a full and meaningful mapping that best reflects the source information in the target schema, several 'feedback loops' may be necessary between the local expert, who knows the source schema and content very well, and a LIDO expert who knows LIDO's structure in depth. This loop is considerably shortened by the ATHENA mapping tool, which reflects the target schema very clearly. The process is also considerably easier if the source schema is based on a documentation standard such as SPECTRUM or national standard. Moreover, features supporting data analysis and data value statistics, such as provided in the mapping tool, help immensely in this process.

6.3 Results

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 Dublin Core.

Not entirely surprisingly, there is a close connection between the level of control used 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 any information recorded, from the outset, as being used outside of its home context. The ease of connecting research information with other sources increases immensely when data structure and terminology standards are used.

LIDO serves in ATHENA as an intermediate layer between source formats and the Dublin Core-based ESE format. It thereby provides a more standardized representation of museum collections in the Europeana prototype even though the ESE format does not support the fine granularity of museum information. The practical implementation of the new Europeana Data Model EDM, that truly allows the LIDO format to be retrieved, will significantly improve resource discovery, providing more precise search results that carry meaningful links to associated resources.

Annex 1. LIDO overview

Descriptive and administrative groups of information of a LIDO record

– Object Classifications –

Object / Work Type (*mandatory*)

Classification

– Object Identifications –

Title / Name (*mandatory*)

Inscriptions

Repository / Location

State / Edition

Object Description

Measurements

– Events –

Event Set

– Relations –

Subject Set

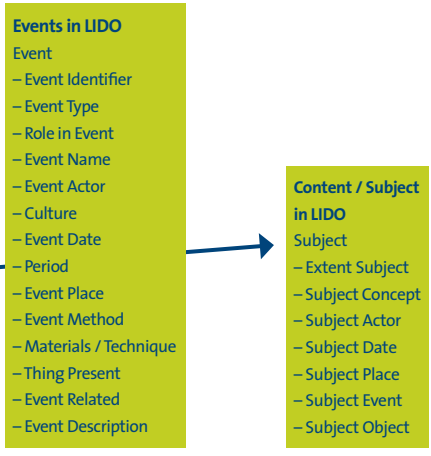
Related Works

– Administrative Metadata –

Rights

Record (*mandatory*)

Resource



Annex 2. The importance of persistent identifiers

The 'stars' of a museum's collections are its physical objects, with the obvious exception of 'born digital' material. The purpose of digital objects, together with descriptive metadata, is to act as surrogates which provide access to the physical. Therefore there is a need to provide a linking mechanism, preferably persistent, between the physical and digital.
For a user to have a meaningful experience it is often necessary to have many surrogates. These can include:

- Different image views of the object, including details, in different 'states' (e.g. open or closed), or X-rays.
- 3D models, reconstructions and replicas (these last two can be physical too).
- Moving images and audios of the object in operation.

To this can be added different metadata for a range of audiences: researchers, members of the public, children, and those speaking a language different from the organisation's native one(s).

Also there is the likelihood that there will be multiple copies of digital objects. In this situation there is a major management task that an organisation needs to address.

All these entities: physical and digital, and the links between them need to be managed. Persistent Identifiers (PIDs) have an important role in doing this. They provide the 'glue' that links the entities together and provide the access to the user over the Internet. Therefore it is vital that they are managed and supported by an organisation.

Printed in the month
of April 2011
by Repro Stampa Ind. Grafica,
Villa Adriana - Tivoli (Roma)