

Quick Guide to Image Metadata

Contents

Why metadata?.....	2
Why so many standards?	3
Key list: metadata/vocabulary standards for image libraries.....	4
Developing a metadata schema: Top Tips	5
Simple Guide to the published schemas	6
Applying metadata: Top Tips	8
Approaches to multilingual provision.....	9
Developing and applying controlled vocabularies: Top Tips	10

Why metadata?

How to make the case for investing in good metadata practices:

- helps you manage workflow economically and effectively
- speeds up the import of data from your suppliers, and reduces errors
- helps you to export your data in a format that complies with your customers' needs
- enables you to reutilise the same data/images for multiple purposes
- helps your data to reach potential users all over the world
- improves the quality and consistency of your data
- helps you provide multilingual access to your collections
- enables sophisticated search techniques, with subtle distinctions where needed
- promotes interoperability over the Internet and other networks
- helps you comply with copyright laws
- helps with management and preservation of records

To achieve all the above, you must plan carefully and exploit standards when possible.

Why so many standards?

We have always found it helpful to standardize practices and conventions in our cataloguing of museum pieces, photographs, etc., so that people can make sense of the display and keep our records in order. But in a networked world, computers rather than humans handle most of the transactions, and they need standards much more than we do.

Data for computer handling need to be presented exactly the same way every time, and different standards may be added at each stage in the process. MILE has not dealt with underlying codes and protocols such as Unicode, HTTP, HTML and XML, that underpin networked communications in all sectors. The focus is on those needed specifically for the metadata of images, works of art and other items of cultural heritage held in galleries and museums. Even for these, a huge number of different standards apply. The communication chain does require several, that is to say, combinations of standards that work together harmoniously. But try the wrong combination and you'll find conflicts which impede communication.

So why not select just one standard for each component of the communication chain, and ensure that each of these is compatible with the next? The simple explanation is that needs are not the same across the sector. More specifically:

- Some image libraries depict fine art, needing metadata of interest to art historians and curators, while others show botanical specimens (or medical cases, or oceanographic pictures, etc) with metadata needed by scientists, and others have simply stock photography for the media and general public.
- As well as needing different metadata schemas, the controlled vocabularies to be applied with the schemas need to vary from one sector to another. For example, medical images need terms from MeSH (Medical Subject Headings); art images need artist names from ULAN (Union List of Artist Names); and many libraries need a vocabulary specifically designed to cover the breadth and depth of their own collection. Because the terms in everyday language have so many ambiguities, and there are so many different ways of expressing the same concept, the controlled languages are typically incompatible with each other.
- The sector cannot operate in isolation. Images are often delivered embedded in text and accompanied by other materials. Across the Internet we connect with the multimedia world. Image libraries must be flexible in providing data in formats acceptable to users and systems from other sectors.
- However logical and convenient a new standard may seem, the reality is that existing collections have mostly been catalogued using older standards, some of them homegrown. Retrieving items from the legacy collections cannot be neglected as new practices come in.
- Internally, every organisation has the idiosyncrasies of its own history, customer base and suppliers to accommodate.

Given all this variety, it is unrealistic to hope that everyone will conform to exactly the same schemas and/or vocabularies. The best hope of achieving interoperability across the sector relies on a twin strategy:

- encourage all image libraries, museums, galleries etc to be consistent in applying the standards they choose;
- develop cross-walks that enable mapping between schemas and vocabularies.

Key list: metadata/vocabulary standards for image libraries

Guides to practice

SPECTRUM: The UK Museum Documentation Standard
PAS 197: Code of practice for cultural collections management
CCO (Cataloging Cultural Objects)
AACR2 (Anglo-American Cataloguing Rules)
RDA (Resource Description and Access)

Metadata schemas

Dublin Core
IPTC
VRA Core
CDWA/CDWA Lite
IEEE-LOM
ESE (Europeana Semantic Elements)
museumdat
METS (Metadata Encoding and Transmission Standard)

Metasearch /harvesting protocols

OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting)
SRU (Search and Retrieval via URL)
SPARQL (Simple Protocol and RDF Query Language)
SKOS-API (Simple Knowledge Organization Systems - application programming interface)
ADL thesaurus protocol (Alexandria Digital Library)

Standards for controlled vocabularies

ISO 2788:1986 (for monolingual thesauri)
ISO 5964:1985 (for multilingual thesauri)
BS 8723 (in 5 parts, published 2005/2007, monolingual + multilingual)
ISO 25964 (in 2 parts, expected 2010+)
ANSI/NISO Z39.19 (2005, monolingual only)
IFLA Guidelines for multilingual thesauri (2009)

Vocabulary exchange formats

SKOS (Simple Knowledge Organization Systems)
DD 8723-5
Zthes

Controlled vocabularies designed for art/image collections

Iconclass
AAT (Art & Architecture Thesaurus)
ULAN (Union List of Artist Names)
TGN (Thesaurus of Geographic Names)
TGM (Thesaurus of Graphic Materials)

Reference models

CIDOC-CRM (Conceptual Reference Model)
FRBR (Functional Requirements for Bibliographic Records)

Developing a metadata schema: Top Tips

- Plan to use one master metadata schema throughout your internal processes, with mappings to and from other (external) schemas, which will be used at the output and input ends of your workflow, respectively.
- Do not expect to find a ready-made schema that meets your internal needs in full. But do look for one that comes close.
- Do not reinvent the wheel; base your internal schema on the existing published schema closest to your needs. You can adapt this by adding some elements and omitting others.
- Try to use exactly the same name and definition of each element as in the published schema or you will end up confusing everybody.
- If possible, set your schema up as an “application profile” of an existing published schema. The general approach and methodology for application profiles is well explained in the Dublin Core literature and you can copy this approach, even if you do not choose DC as your base schema.
- Developing the application profile may be easier if you draw on just one published schema, but optionally you can draw on two or more.
- Enlist the support of a champion at top management level.
- Involve representatives of all key interests: users; sales team; cataloguers; IT, etc.
- Take into account the requirements of the customers and collaborators to whom you supply data. This should include an understanding of the search preferences of users.
- Study the foibles of the data to be imported from your suppliers.
- Know your own in-house needs.
- Go for granularity, i.e. subdivide each element into the smallest sub-elements you may need. (You can always dumb down by merging elements later; but you cannot so easily “dumb-up”!)
- Don’t expect too much of your cataloguers. If they can’t tell the difference between a geranium and a pelargonium, don’t expect them to make that distinction reliably in the metadata!
- Provide for more than one level of description within a single record. Very often the metadata associated with a photograph (e.g. date, IPR, name of photographer, etc.) differs from the metadata applicable to the object(s) shown in the photograph. This is a common problem with pictures of museum pieces or other works of art, and with composite resources such as a newspaper article or an educational package. The schema may need to “nest” the metadata for several objects within that of the image as a whole.
- Provide for user-generated metadata (such as bookmarks or social tagging) to be added at a later stage. Usually separate fields are needed for such data.

Simple Guide to the published schemas

This page picks out key features of some commonly available schemas, that you should bear in mind when choosing your base schema, or when mapping outputs for your customers. The schemas described here are only a selection especially relevant to images and cultural heritage. You should also consider schemas specific to your country (e.g. the e-GMS if you supply to the UK government) or sector (e.g. AGMES from FAO if you are into agriculture).

CDWA and CDWA Lite (Categories for the Description of Works of Art)

http://www.getty.edu/research/conducting_research/standards/cdwa/index.html

- Well adapted to works of art;
- Supported by good cataloguing guidance and other resources from the Getty Research Institute;
- CDWA Lite provides an XML schema for outputting the core elements in the full CDWA.

Overall: Could form the basis of your internal schema if you are in the business of fine art, and your customers may well request data formatted according to CDWA Lite.

VRA Core 4.0 (Visual Resources Association)

<http://www.vraweb.org/projects/vracore4/index.html>

- Well suited to images of cultural objects;
- Provides well for nested metadata ;
- Follows the same cataloguing rules as CDWA.

Overall: Could form the basis of your internal schema if you are in the business of photography, and your customers may well request data using this format.

IPTC Photo Metadata Standard (International Press Telecommunications Council)

[http://www.iptc.org/std/photometadata/specification/IPTC-PhotoMetadata\(200907\)_1.pdf](http://www.iptc.org/std/photometadata/specification/IPTC-PhotoMetadata(200907)_1.pdf)

- Updated 2009 to support metadata of artworks within the metadata of a photograph;
- Popular with news agencies worldwide.

Overall: Could form the basis of your internal schema if you are in the business of supplying images, and your customers may well request data using this format.

METS (Metadata Encoding and Transmission Standard)

<http://www.loc.gov/standards/mets/mets-home.html>

- Its strength is in managing and exchanging data about composite digital objects;
- Caters for embedding metadata drawn from another schema, e.g. VRA or DC;

Overall: Choose METS if your users need to find their way around complex digital “documents” made up of several chapters, images, tables, audio files, etc.

Dublin Core (DC)

<http://dublincore.org/documents/dces/>

- Popular with academic institutions and other organisations worldwide;
- Works best for text resources;
- Lacks many elements needed for images and for works of art.

Overall: Even if the schema is inadequate for your in-house needs, you should be able to supply outputs in DC format for customers who require it.

ESE (Europeana Semantic Elements)

<http://dev.europeana.eu/>

- Based on the Dublin Core, extended to meet needs of Europeana Project;
- Applicable to all kinds of cultural heritage, including literary works.

Overall: Not designed for your internal use, but for outputs if you wish to join in Europeana.

IEEE-LOM (Learning Object Metadata, Institution of Electrical and Electronics Engineers)

<http://ltsc.ieee.org/wg12/>

- Enables learners or instructors to search, evaluate, acquire, and utilize Learning Objects;
- Allows metadata for a simple object, e.g. an image, to be nested within a package of educational materials.

Overall: Unlikely to meet your internal needs, but useful for outputs if you supply to educational institutions or aggregators of materials.

museumdat

<http://www.museumdat.org/>

- Adapted from data elements in CDWA Lite and VRA Core;
- Used by museum portals.

Overall: Not designed for your internal use, but for outputs if you supply to certain museum portals.

Applying metadata: Top Tips

- To accompany your schema you need a set of cataloguing rules. For each element that you have drawn from a published schema, your definition and cataloguing rule should be compatible with those of the source schema. (For example, if your schema includes elements drawn from CDWA or VRA Core, you should follow the corresponding rules from CCO.) You may need to extend the rules to cover particular issues that arise within your own image collection.
- Set up editorial and quality control procedures to ensure that all the cataloguing records you produce conform to the rules – consistently!
- Test the schema and the rules thoroughly before it is too late to change them. Use some of your quirkiest images for the testing. When the images are catalogued following the rules, does the metadata display correctly on your website and are your customers happy with the records you supply?
- Organize your workflow to maximize efficiency. For example:
 - Use automated processes wherever possible without jeopardizing quality
 - Avoid duplication of effort
 - Integrate controlled vocabularies wherever possible to speed data entry and avoid errors
- Beware of unnecessary perfectionism, which adds to costs without a proportionate increase in value. Subjective decisions are required to determine which rules/procedures are important to your business.
- Get the right balance between complexity and simplicity. Complexity is fine if it works effectively and is hidden in the backroom, but make sure that simplicity is delivered to the user. Likewise, high development (one-off) costs may be justified in building robust schemas, vocabularies and systems, provided they result in low costs at the operating (much-repeated) stage.
- Exploit the detail of your metadata in a specialised interface for users who want to distinguish between, for example, pictures of Vienna and works housed in Vienna, or between a recent painting in Renaissance style and one that was actually painted during the Renaissance period. But do not impose a complicated interface on the first-time user!
- Remember the potential of harvesting and metasearch for reaching new audiences. Your data need to be in the right format to be reach portals and other networks across the globe.
- Remember the dangers of harvesting and metasearch. When your images are captured and displayed in a completely different environment, your context will be lost and extra metadata may be needed. For example, a UK collection of Victoriana might include pictures of “The Queen at home”; for an international audience the metadata needs more explanation of *which* queen.
- When you receive batches of images for input, use mapping techniques to feed the supplier’s data into the correct metadata field (with validation and other checks as necessary). This is easier if the supplier uses a standard schema such as IPTC. You may find it useful to develop an “input crosswalk” from each of the commonly used schemas to your own internal schema. But if your supplier uses a text-oriented schema such as Dublin Core, which does not make the distinctions commonly needed for works of art or for images, the result of the mapping may need hand-editing to fit it for your purposes.
- Mapping is also useful when you are supplying completed records in the formats required by your customers. Mappings often work in one direction only, and so the “output crosswalk” **from** your data **to** any of the commonly used schemas usually differs from the input crosswalk.
- The best quality can be obtained by preparing mappings from your internal schema directly to each required output schema. However, you can save development time by consulting a published crosswalk (e.g. the “Metadata standards crosswalk” from CDWA to other

standards, on the Getty Research Institute website). Avoid chained mapping if you can. (In other words, map directly from A to C rather than indirectly via B.)

- Be prepared for change as time passes, and design systems accordingly. For example you may need to add more metadata elements, or retrieve and correct all items referring to a certain copyright holder, or add another language to your search vocabularies.

Approaches to multilingual provision

Several different approaches are possible, with different implications for metadata:

- a. Ignore the needs of users whose first language is not the same as the main language used in your records. In this way you avoid doing anything special or incurring extra costs with your metadata. The main disadvantage is loss of market share.
- b. Provide for users who search in different languages, by enabling them to get their queries translated automatically while they search. Unless you enable automatic translation of the outputs, the records they retrieve will be presented in your original language, which may or may not be acceptable to the users. With this approach, there may be doubts about the quality of the translation and there will be a significant development cost and running cost of the automatic translation, but you avoid the (initially larger) cost and delay of translating the metadata.
- c. Translate all the metadata into the languages wanted by your customers, and provide a user interface in the same languages. This way you have significant up-front development costs, and a small overhead associated with translation of vocabulary updates, but the extra running costs are minimal. In the long run, it is the cheapest way of enabling multilingual access to your collection.

Assuming you choose option c, the simplest way to translate the metadata is to translate the controlled vocabularies. Thus the same translations can be used over and over again, each time the same names or keywords are used in cataloguing. Only the fields which do not use multilingual controlled vocabularies will be left without translations.

Developing and applying controlled vocabularies: Top Tips

- Typically you will need 3-4 different types of vocabulary:
 - simple lists (at most 20 items in each, as a flat list)
 - name authority lists (e.g. of artist names)
 - thesauri (to support search functions)
 - taxonomies (to support browse functions)
- All but the simple lists are expensive to build and maintain. Take advantage of existing lists wherever possible.
- Maintenance/updating should be built into your procedure and your budget, in perpetuity.
- The overhead cost of a controlled vocabulary is justified over time by:
 - a. reduction in operating costs, as the vocabulary supplies synonyms, broader terms, narrower and related terms automatically
 - b. similarly, a multilingual vocabulary supplies term translations automatically
 - c. improved consistency and error detection
- For geographic names, as well as normal synonyms and translations, your vocabulary should supply geographical coordinates, latitude and longitude, postcodes, and any other popular coding system.
- Thesaurus construction needs specialized software. A modest purchase saves you a lot of time and errors.
- Thesaurus construction needs tight editorial control. Appoint someone meticulous, good at spelling, who enjoys playing with words and has relevant experience.
- When developing your own vocabularies, follow the national and international standards. (See list)
- Include common misspellings (so long as they are unambiguous) among the other lexical variants of preferred terms that you admit.
- A multilingual vocabulary should give equal status to all the languages. To avoid the dominance of one language, take all the required languages into account from the start.
- Some of your customers may want to view and select terms from your vocabularies, but mostly you should try to integrate the vocabularies into your search and cataloguing interfaces so that they convert terms automatically behind the scenes.
- While mapping from one metadata schema to another is usually feasible, mapping from one controlled vocabulary is usually a much bigger job. The development expense of mapping between vocabularies is only justified if you routinely have to provide outputs or search capabilities in more than one.