VRA Bulletin

Volume 38
Issue 1 *The inaugural electronic edition of the VRA Bulletin*

Article 4

March 2012

Cultural Objects Digitization Planning: Metadata Overview

Janice L. Eklund
University of California - Berkeley, janice.l.eklund@gmail.com

Cultural Objects Digitization Planning: Metadata Overview

Abstract
This document offers an overview of image metadata types, applications, and best practice considerations for planning cultural object digitization projects.

Keywords
metadata, digitization, image cataloging, digital curation

Author Bio & Acknowledgements

Jan Eklund served as Curator of Visual Resources for the History of Art department at the University of California, Berkeley, from 1993-2010. She holds a Masters Degree in History of Art from the same institution. She has been an active member of the Visual Resources Association (VRA) for more than 20 years, and served as chair of the VRA Data Standards Committee from 2005-2007 during which time she co-authored version 4.0 of the VRA Core Categories Metadata Standard and XML schema. In addition, she served on the editorial advisory board for the Cataloging Cultural Objects metadata content standard from 2002-2006, and was a member of the California Digital Library's UC Shared Images Metadata Working Group in 2008-09. Jan has presented papers and taught metadata workshops at several VRA annual conferences, and taught metadata workshops for the joint ARLIS/VRA Summer Educational Institute (SEI) in 2006-2007. She is now an independent consultant living and working in Berkeley, California.

This feature articles is available in VRA Bulletin: https://online.vraweb.org/vrab/vol38/iss1/4

Cultural Objects Digitization Planning: Metadata Overview

It has often been said that the most important part of any digitization project is the planning that precedes it. When the project involves digitizing unique cultural objects, this cannot be emphasized strongly enough. By far the largest cost associated with any archival digitization project is the cost associated with collecting, organizing, entering, and managing the data associated with the images that depict those cultural objects. The time invested up front in the planning process will more than pay for itself over time by resulting in data that is flexible and robust enough to be readable, exportable, and still discoverable across new systems as technology changes. All around the world, millions of images of works of art, architecture, literature, and material culture are captured digitally every day and added to the global electronic landscape. Without effective pre-planning for capturing good metadata to accompany these images, they will remain inaccessible. A picture may be worth a thousand words, but a few well-chosen words associated with an image are what allow that image to be isolated and retrieved from the blizzard of images that exist in cyberspace. There are many things to consider when planning a digitization project and metadata is one of them. This document provides an overview of the issues surrounding image metadata and poses some questions to consider and resolve before embarking on such a project.

First steps

One of the most important questions to ask at the beginning of any project is *why* it is being done and what it hopes to accomplish. Who is the audience? Who will benefit from having these images and their accompanying data in electronic form? How will the audience use these images and for what purpose? Where will this collection live, who will manage it, and how will it be sustained and kept viable in the future? Determining both immediate and long-term needs and uses at the outset will help keep the project focused and will help answer more specific questions about what kinds of resources will be needed to accomplish the project goals.

Pieces of the Puzzle

Quality digital collections share the following essential elements.

1. *Quality content.* A critical mass of high-quality images and data attracts end-users who return repeatedly to the same collection. They will return because they are able to find what they are looking for quickly and efficiently and what they find is worth finding because each quality image is accompanied by complete and accurate metadata.

2. *Quality container*. The data structure used to hold metadata about the images of cultural objects should not only be capable of recording enough data to fully describe the content, but should also be flexible enough to allow that data to be repurposed. Such structures are often described as “granular” because they generally store data in many small pieces that can be combined and re-combined

automatically in different ways at different times to meet different needs. A simple example might be storing a telephone number in three discrete pieces: area code, prefix and last four digits. This allows efficient sorting by area code or prefix, as well as printing and/or screen display with or without area code.

3. *Quality cataloging*. The information recorded about a cultural object is essential but *how* it is recorded is just as important. Consistent data yields consistent search results. Basic information such as artist's name, if recorded inconsistently or incompletely from record to record, will produce incomplete search results and frustrated end users. Precision and quantity of cataloging data (sometimes referred to as specificity and exhaustivity) are also important. The words individuals choose to search for information about cultural objects vary widely and are determined by the individual point of view of the end user and *why* they are seeking this information in the first place. Good catalog data strives to be both precise and as thorough as possible, and provides multiple access points to accommodate the widest possible number of end users.

4. *Quality curation*. Quality data is "curated" data. Curated data is data selected, vetted, and created with long-term viability in mind by dedicated and knowledgeable professional catalogers. Digital curation involves "maintaining, preserving and adding value to digital research data throughout its lifecycle."[1] It is an ongoing process that starts with the planning process and continues through the entire data lifecycle, including data creation, access, preservation, migration, storage, reappraisal, and transformation. The active management of metadata insures long-term research value and mitigates the risk of digital obsolescence.

5. *Quality terminology*. Terminology used to describe cultural objects is an important component of quality cataloging and curation. The use of controlled vocabularies to describe digital collections insures data consistency by making sure that like items are described using the same terms. The terms "notebook," "journal," "diary," and "blog" might all describe a personal daily written account so it is important when describing such objects that the same terms be used for like items to insure consistent search results later on.

6. *Quality rights data*. Clear and accurate copyright data is especially important for a quality digital collection. Digital images are easily copied, edited, and transmitted around the globe. Image consumers who wish to use these images for different purposes need to know who owns the rights to the image, and who to contact about how they may or may not use it. In addition to image rights, there may also be rights to the *content* of those images. Images of works of art not in the public domain may have several layers of copyright involved, such as the creative rights of the artist as creator of the work depicted, creative rights of

[1] "What is Digital Curation? | Digital Curation Centre," http://www.dcc.ac.uk/digital-curation/what-digital-curation.

the photographer who has uniquely captured this work of art in an image and/or creative rights of the photographer who creates unique or born-digital art, as well as distribution or publication rights granted by the artist or photographer to a publisher.

Questions to Ask

When planning a digitization project there are a number of questions that will help determine size and scope of the project. The answers to these questions will help determine the size and sophistication of the information system(s) that will be needed.

1. *What is being managed?*
   a. Is it a collection of **physical objects**? Physical objects, in addition to basic descriptive cataloging of the objects themselves, often have additional information management needs, such as managing conservation, valuation, loan, or exhibition data.

   b. Is it a collection of **digital assets**? A digital asset is any form of content and/or media that has been formatted into a binary source that includes the right to use it. A digital file without the right to use it is not an asset. Digital assets are categorized in three major groups which may be defined as textual content (digital assets), images (media assets), and multimedia (media assets).[2] Digital assets frequently have multiple layers of rights information that are important to record, such as the creative rights of the author of a manuscript, and the creative rights of the photographer who captures this manuscript in digital form. Distribution and/or usage rights may also be involved if this asset is offered for sale or use in publication.

   c. Do collection items have **intellectual property** considerations? Will management of permissions for use of source material or images of source material for publication or other commercial use be required? Collaborative and multi-media works often involve multiple rights that may require a more robust rights data structure.

   d. Will management of **people and processes** be involved? Tracking usage data, either online usage or physical loans, will require transaction records. Algorithmic routines that transform or repurpose the data automatically into different formats or digital objects may require close attention to versioning and regular updates to avoid digital obsolescence.

   e. Will the project require management of **records contained in different information systems**? This will require attention to cross-platform issues, such as how the data displays on different computers using different

[2] van Niekerk, A. J. (2006) The Strategic Management of Media Assets; A Methodological Approach. Allied Academies, New Orleans Congress, 2006

software, fonts, and applications. As with automatic processes, close attention to regular updates will be needed to keep pace with technology changes.

2. *How will the data be used*?
   a. **Who is the intended audience?** It is important to consider both the immediate audience, such as the faculty and students in an individual department, and any audience that might be served in the future, e. g., the entire university, K-12 schools, or the general public. Different audiences require different search strategies to accommodate multiple levels of knowledge and sophistication.

   b. **Will multiple levels of permission be required?** If certain groups need access to information restricted to others then it will need to be determined whether access should be restricted to certain classes of information or individual fields in specific tables. Not all database software packages allow permissions to be set at the field level so this requirement will limit the choice of information systems.

3. *What kind of information system(s) will be needed*? There are many different kinds of information systems, each designed for a different purpose. Answers to the questions posed above should help determine what kind of information system(s) will be required.

   a. **Collections management system (CMS).** A collection comprised of physical objects will need some sort of collections management system to record data about each of the objects in the collection. Most CMSs are applications built on or with generic database software (MS Access, Filemaker Pro, Oracle, Sybase, etc.) and hold basic object information, as well as rights, conservation, loan, and exhibition data if needed. Examples include The Museum System and EmbARK from Gallery Systems, Past Perfect, Willoughby Multi MIMSY, and EMu from KE Software.

   b. **Cataloging/production system.** Depending upon the size and complexity of the collection's administrative organization, a separate cataloging and/or production system that feeds into the collections management system may be required. These are most often applications that are built on top of the collections management system to facilitate data entry.

   c. **Digital asset management system (DAMS).** If the project will be managing digital assets alone or in addition to physical objects, it may need both a collections management system to manage and record data about the objects and a separate digital asset management system that can manage and record data about the digital assets. Common desktop examples include such products as Extensis Portfolio and Canto Cumulus. Enterprise systems include NetXposure, North Plains Telescope, EMC Documentum, Interwoven, MediaBin, Open Text Artesia, and ContentDM from OCLC.

d. **Discovery, access, and presentation system (DAPS).** If the digital assets will be made available to others beyond immediate staff, a separate discovery and access system that allows an end user to browse or search the collection and retrieve images by category, key word(s), or a combination of both may be required. Examples include Luna Insight, Madison Digital Image Database (MDID), and Almagest. All three of these packages also include tools that allow the end user to create their own PowerPoint-style presentations using selected images. Some integrated tools such as ContentDM and Extensis Portfolio advertise themselves as being both a Digital Collections Management System and a Discovery and Access System by providing a web publishing tool that pushes content from the DAMS to a searchable website that may be made available separately. These two products do not include any presentation tools.

e. **Online catalog.** Depending upon the size and complexity of the organization, a tool that brings together records from multiple databases, DAMS, and DAPS into one searchable online catalog may be required.

f. **Preservation repository.** If long-term archival preservation is desired, consideration should be given to establishing or contributing to a separate preservation repository. For security reasons, preservation repositories are generally separate from both public access and production data systems. Submission formats that conform to national or international standards are generally required; regular, redundant, and geographically dispersed backups are performed; and documented strategies for systematic data refreshment and data migration to prevent data loss and/or obsolescence are established. Institutional repository software such as dSpace or Fedora is commonly used for preservation repositories. Because their primary function is long-term data storage, these products can store vast amounts of data but generally have fewer tools that facilitate access for end users.

g. **All of the above.** Very large, complex organizations with large and complex needs may want all of the above. Smaller organizations will be able to function well with perhaps only one or two. A phased approach will allow gradual growth by adding systems over time. Thinking through both immediate and future needs will help determine which systems are most important to implement first.

4. *What kind of resources will be needed?*

a. **Hardware and Software**. Most small to medium-sized collections can manage with common off-the-shelf desktop computer hardware and software. If multiple staff members will be responsible for data entry, the addition of a file server will be essential. The exact configuration will depend upon the software selected and the functional needs of the organization. If the digitization will be done in-house, purchase of scanning equipment appropriate to the materials to be scanned (flatbed scanners, film scanners, digital cameras, copy stands, book cradles, etc.) will be required, and this equipment will need to be networked to communicate with the server and desktop computers.

b. **Staffing and Facilities**. While it is certainly possible (and sometimes necessary) that permanent staff already employed by an organization can be trained to perform all the tasks involved in a digitization project, it can also pay to outsource certain portions of it. Outsourced production scanning can often be more cost-effective than using permanent or even casual employees to do the work when the cost of purchasing scanning hardware and software, the space needed to house all of this equipment, staff training, the number of staff needed, and the hours needed to accomplish the job are all taken into consideration. Production scanning companies with access to high-end volume scanning equipment, facilities, and trained personnel can often turn the scanning portion of the project around more quickly and economically. Entering catalog data can be a labor-intensive and time-consuming activity that is best accomplished by staff who know the material well, but even that can be cost-effectively outsourced if there are pre-existing catalog records either on paper or in some electronic form. Even if only minimal catalog records are established for each item in a large project, this can save time and money in the long run. Outsourcing the basic production aspects frees local collection staff to enter the data that requires their specialized subject expertise, rather than bogging them down with repetitive production work.

Database Considerations

If the digitization project involves a physical collection that does not yet have any collection data in electronic form, a simple database to capture minimal metadata about each item to be digitized should be established. This will help keep track of collection items as they are digitized, and provide a starting point for more complete descriptive metadata to be added at a later date. Many off the shelf database products provide the basic tools to accomplish this. It is beyond the scope of this document to provide a thorough discussion of database design and current database products, but the following general discussion should provide a starting point for making an informed product selection that will accommodate both the management and budget needs for a collection of any size.

Data structure: flat vs. relational. Database products come in two basic forms, flat or relational. A flat database is basically one large spreadsheet or table, with each row representing one record and each column representing fields within that record. Each row or record holds information about one item in a collection and each column holds individual pieces of that information for each row or record. A good example of the flat structure is an Excel spreadsheet. This kind of simple product is suitable for collections of like items where the information about each item is relatively simple, with only one value per column needed. An example might be a simple stamp collection where each row represents an individual stamp, and each column contains data about the country of issue, date issued, condition, purchase price, purchase date, modern valuation, etc., for each stamp *(fig. 1).*

Name	Country	Year	Condition	Purchase Price	Purchase Date	Current Valuation
Stamp1	Australia	1974	Fair	$2.50	2003-18-09	$2.55
Stamp2	Great Britain	1990	Fair	$2.35	2010-02-04	$2.55
Stamp3	Great Britain	1876	Good	$7.00	2010-04-10	$52.50

Figure 1: Flat Data Model

Relational databases consist of multiple tables that are related to one another by sharing common key fields. Relational databases are better suited for collections with complex works or different kinds of works that require different kinds of fielded information. Examples include museum collections, archives, and library collections. Any collection that contains many different kinds of objects, complex objects, or documents that have different parts (pages of a notebook, panels of an altarpiece, or multiple specimens of a particular species) will need a relational model database to capture complete descriptive information about each object. Complex collection data cannot be captured efficiently in a flat data model because space must be allowed *in every record* to accommodate the most complex object in the collection. This adds up to a lot of wasted space, and wasted space means more money and hardware needed for storage, backup, preservation, etc. It is much more efficient to catalog in a relational environment, where data can be entered once and then linked to many other records. Data consistency is an added advantage of this model because data that repeats from record to record is linked, rather than recorded over and over again in each record. This reduces the chance for error and makes updating multiple records much easier. Even simple collections can benefit from a relational database by taking advantage of related tables to hold controlled lists of values that repeat for each record *(fig. 2).*

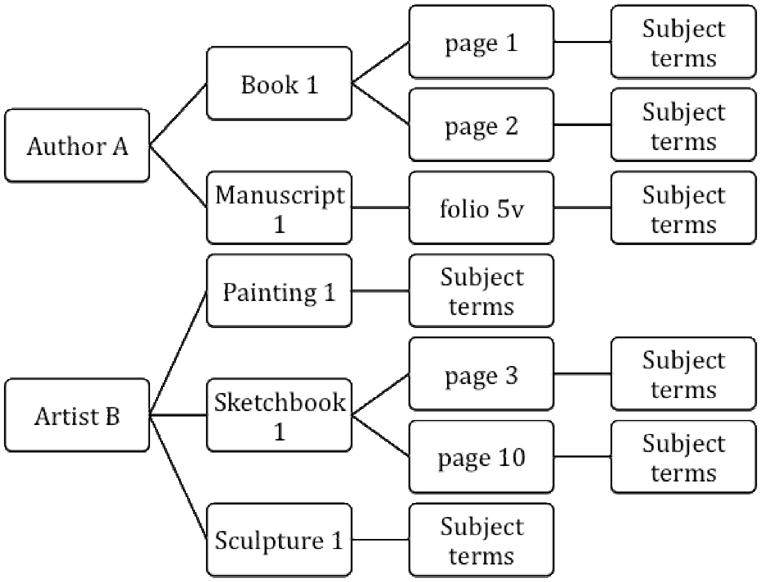


Figure 2: Relational Data Model

Portability and interoperability. Another important consideration when choosing a database product is data portability and interoperability with other systems. Will the product be able to "talk" and exchange data with other information systems such as a DAMS, DAPS, or online catalog? Does the database product support export formats that the other systems can accept and will it export the data in the desired format? Will the data be locked into a proprietary system or format that might become obsolete or too expensive to maintain in the future? New versions of software will eventually lose backward compatibility with earlier versions of the same product making regular upgrades a necessity. If a product switch is required at a later date to accommodate a new information system, the ability to migrate data from one product to another is an important consideration. Data stored for long periods of time for preservation purposes will need to be periodically refreshed to insure data integrity and part of that refreshment will ultimately involve exporting from an older version or system into a newer one.

While relational databases are able to capture and record rich and robust metadata about works, component works, related works, and all of their associated images, most image delivery systems are essentially flat. Their structure typically only supports one record per image and that record must contain all the information about that image *and* its content. This allows for fast and efficient search and retrieval of images. In order to include even minimal information about both work and image in one record, relational data is typically "flattened" by conflating or concatenating data from one or more fields and then exporting it in a tab- or comma-delimited format for ingesting into the DAMS. For example, work title, work creator (in natural language order), and image title might be concatenated and put into a single DAMS *Title* field, e. g.,

Garden at Sainte-Adresse, by Claude Monet. Detail view of signature at lower right.

It can be a challenge to include even minimal descriptive metadata about a work and its associated image from a relational structure into one image record without creating confusion. For example, if there is only one field called "Creator" in the DAMS, should that field contain the name of the creator of the work depicted in the image or the creator of the image? Does the Rights statement refer to the rights holder of the image or the rights holder of the work depicted in the image? Decisions about what and how much metadata to export and where it should go in the flat file structure must be made before exporting. An end user who enters "Botticelli" in the Creator field expecting to find images of Italian painting may find nothing if the Creator field contains the name of the photographer.

There are a number of strategies that may be employed to keep relational work and image metadata separate and distinct in a flat image delivery system. Most DAMS include a generic Description field, so one popular method is to simply concatenate all the work metadata into one long string and export it to the image Description field. This allows the remaining DAMS fields such as Creator, Date, Rights, etc., to be populated exclusively by image metadata, but still allows an end user to search the Description field for work metadata by key word.

Embedded metadata. An alternative to exporting work and image metadata to a DAMS is to embed metadata, or a link to metadata, in the image file itself. Embedding a link to the object metadata record in the image file insures that any recipient of the digital image file, regardless of where or how it is obtained, will have a pointer to the most current descriptive metadata record available from an authoritative source. More complete descriptive metadata may also be flattened and exported from the relational work record and stored in the image file itself. But the advantage of only embedding a pointer is that the pointer always remains the same and always refers to a single, current catalog record, whereas embedding complete descriptive metadata in the image file makes possible the creation of multiple, seemingly identical image files that may contain conflicting descriptive metadata, depending upon when the image was made and the work record last updated.

Support. Any database product selected will need to have some kind of technical support as well as regular database backups. Whether tech support is provided by one person, a contract service, or an in-house team of support professionals depends upon the size, complexity, and sophistication of both the budget and the organization. Tech support is often required during initial setup; during software or hardware upgrade and/or migration; or after hardware failure, theft, data corruption, or other disaster. Tech support should also help insure that regular backups are kept to minimize downtime after any kind of business interruption. Data is only as safe as the last backup, and maybe not safe enough if the last backup was made a month ago on an external hard drive next to the computer on which a staff member just spilled coffee.

Content Management

Just as important as the selection of hardware, software, information system, and database structure is the management of the metadata that is put into those products. There are different types of metadata that serve different functions.

1. *Technical metadata*. Technical metadata is data that describes the image file itself. Much of it is recorded at the time of capture and stored in the file itself. Examples of technical metadata include (but are not limited to): file size, bit depth, resolution, capture device, capture device settings, capture date, etc.

2. *Administrative metadata*. Administrative metadata is generally local data that is used to manage data records within a collection and is almost always unique to the collection. Accession numbers, classification categories, ownership and rights data are examples of administrative metadata. This kind of data is most commonly found in the collections management or digital asset management system and is either linked to images of collection items or embedded into the image file itself.

3. *Descriptive metadata*. Information that describes a work, be it a physical object or the content of an image, is called descriptive metadata. Descriptive metadata answers the "who, what, when, where, and what is it of or about" questions for the object or image content. Descriptive metadata is generally stored in a collections management system and linked to digital images. Descriptive metadata can be very simple or quite complex, depending upon the work in question. Cultural objects by their very nature are generally complex, having been created by different cultures at different time periods by named or unnamed persons for different purposes. Cultural objects frequently have multiple values associated with certain fields of information such as multiple types of names (creator names, donor names, publisher names, etc.), multiple types of dates (creation dates, restoration dates, publication dates, etc.), multiple location values (discovery locations, repository locations, exhibition locations, etc.), multiple subjects, materials, and components. The complexity of this kind of descriptive metadata makes it difficult to fully record in anything but a relational data model.

4. *Preservation metadata*. Preservation metadata is often wrapped around a technical, administrative, and descriptive metadata core and contains information about the date and time of ingest to the preservation repository, version (if multiple versions of the same digital object exist) and information related to the hardware and software requirements needed to open, view, and transform the data if necessary for refreshment or migration at a later date.

5. *Transport metadata*. Transport metadata is metadata that facilitates the transmission of data records across the network and across different hardware platforms. The Library of Congress METS (Metadata Encoding Transfer

Standard) is a good example.[3] METS employs the World Wide Web Consortium's XML (Extensible Markup Language)[4] schema to wrap around a technical, administrative and descriptive metadata core that describes a digital object.

6. *Usage metadata*. Usage metadata records statistics about how often a digital resource is visited; how, when, and how often a digital object is accessed, opened, or downloaded; etc. Usage metadata is used to rank objects by popularity (determined by how often they are accessed), allocate resources dynamically, and determine peak load times, among other things.

Importance of Metadata Standards

Metadata standards are the rules of the road for how the various types of metadata described above should be recorded and transmitted. Metadata standards are community-developed guidelines designed to insure data that is viable and can travel across disparate networks via a variety of hardware and software platforms without losing context or meaning in the process. Standards have been developed for nearly every type of metadata and they generally come in one of two flavors: rule-based or principle-based. Rule-based standards, as the name might indicate, provide specific rules for how data should be structured, formatted, or encoded. Principle-based standards provide guiding principles for these functions rather than hard and fast rules.

1. *Data Structure standards*. Data structure standards provide the architecture that holds the metadata. Data structure standards control how data is categorized or modeled into table and field structures, or metadata element sets. Some examples include:

   a. **Rule-based**: MARC (Machine Readable Cataloging),[5] MODS (Metadata Object Description Schema),[6] EAD (Encoded Archival Description)[7]
   b. **Principle-based**: CDWA (Categories for the Description of Works of Art), [8] VRA Core 4.0 (Visual Resources Association Core Categories) ,[9] CIDOC CRM (CIDOC Conceptual Reference Model)[10]

2. *Data Content Standards.* Data content standards guide the selection, organization, and formatting of data entered into data structures and typically include recommendations for both display and index values. These are often referred to as cataloging rules.

[3] See http://www.loc.gov/standards/mets/
[4] See http://www.w3.org/XML/
[5] See http://www.loc.gov/marc/
[6] See http://www.loc.gov/standards/mods/
[7] See http://www.loc.gov/ead/
[8] See http://www.getty.edu/research/conducting_research/standards/cdwa/
[9] See http://www.vraweb.org/projects/vracore4/index.html
[10] See http://www.cidoc-crm.org/

a. **Rule-based**: AACR2 (Anglo-American Cataloging Rules), [11] soon to be RDA (Resource Description and Access), [12] DACS (Describing Archives: A Content Standard)[13]
b. **Principle-based**: CDWA, CCO (Cataloging Cultural Objects)[14]

3. *Data Value Standards*. Data value standards are vocabularies, authorities, and taxonomies used to guide the selection of terms used in descriptive cataloging. While the formatting and organization of these terms is the province of data content standards, the terms themselves exist as either controlled lists of acceptable values, or full-blown thesauri, which contain terms in a hierarchical structure. Some examples include:

a. **AAT** (Art and Architecture Thesaurus) [15]
b. **ULAN** (Union List of Artist's Names)[16]
c. **TGN** (Thesaurus of Geographic Names)[17]
d. **LCSH** (Library of Congress Subject Headings)[18]
e. **LCNAF** (Library of Congress Name Authority File)[19]

4. *Authority Files.* In addition to national or international community-developed data value sources, local authority files are often constructed from a combination of terminology derived from several sources. For example, if a collection wanted to include both Library of Congress Subject Headings plus local subject terms not included in LCSH, they might both be included in a local subject authority from which catalogers may select subject terms. Similar local authorities are often constructed for Personal and Corporate Names, Places, and Concepts.

Cataloging Considerations

Workflow. Part of the planning that goes into any digitization project should include the cataloging process. Primary among these is determining who will be doing the cataloging and at what point in the workflow it will be accomplished. Will minimal data records be created at the time collection items are digitized, or will the capture process be independent of the cataloging process? Depending upon the size and sophistication of the collection and collection staff, entering catalog metadata can be one of the most time-consuming and costly aspects of a digitization project. Even if only minimal data

[11] See http://www.aacr2.org/
[12] See http://www.rdatoolkit.org/
[13] See http://www.archivists.org/governance/standards/dacs.asp
[14] See http://www.vrafoundation.org/ccoweb/index.htm
[15] See http://www.getty.edu/research/conducting_research/vocabularies/aat/
[16] See http://www.getty.edu/research/conducting_research/vocabularies/ulan/
[17] See http://www.getty.edu/research/conducting_research/vocabularies/tgn/
[18] See http://authorities.loc.gov/
[19] Ibid.

records are created for each item, how those minimal records are created can make a huge difference. This is where having the right kind of metadata structure in place can help.

If your collection has a large number of like items with only minor variations from item to item, then a relatively simple flat structure may suffice. But even a small collection, if comprised of many different unique items, items that contain multiple component parts, or multiple discrete collections of items, will need a much more robust data structure to be able to efficiently capture the full scope and nuance of everything it contains. Complete data about each item need not be entered all at once, but it is important that there be a structure in place that will accommodate all the data that will need to be recorded, even if it has to wait until a later date. A good catalog record should:

- Provide a level of description sufficient to identify an object or group of objects and its differences from other, similar objects
- Provide an historic archive relating to an object or cross-references to sources where information can be found
- Be held in a system that allows convenient access, e. g., using indexes or free-text retrieval.[20]

The first question to ask is what exactly is being cataloged? When dealing with cultural heritage materials this is a fundamental question. Is it a physical *object or work*, needing information such as what kind of object or work it is, when it was made, where it was made, and by whom? Or is it the *image* of an object or work, needing information such as what kind of image it is, who made the image, and where and when it was made? The fundamental distinction between work and image metadata is an important one to resolve before beginning in order to avoid confusion among cataloging staff and end users alike. It is certainly possible to capture both, but care must be taken to keep the distinction an explicit one. The data model can help maintain this distinction by storing work metadata in one set of relational tables, and image metadata in another.

Images that depict collection objects, linked to collection object records, can inherit information about the objects through their association of image and collection item records. This helps keep the cataloging efficient and the data consistent because one object record informs all the image records linked to it. If object data needs to be added or changed, the change is made in one object record rather than in hundreds of individual image records.

The same holds true for collection-level records that describe the collection as a whole, or unique objects with component parts. For example, an object record that contains

[20] *SPECTRUM. The UK Museum Documentation Standard.* Revised ed. 3.2, edited by Gordon McKenna and Efthymia Patsatzi. Cambridge: The Collections Trust. c. 2009, p. 99.

basic information about a medieval manuscript, such as who wrote the text, who illustrated it, when it was created and where, can be linked to multiple component part records each representing a page or folio. This way each page or folio record can contain its own unique data (such as page or folio number, illuminator, and subject terms specific to that page or folio) but will also inherit the basic information about the manuscript as a whole from the parent record automatically. Not having to store redundant data in each component record can represent a huge time and cost savings and insure data consistency at the same time.

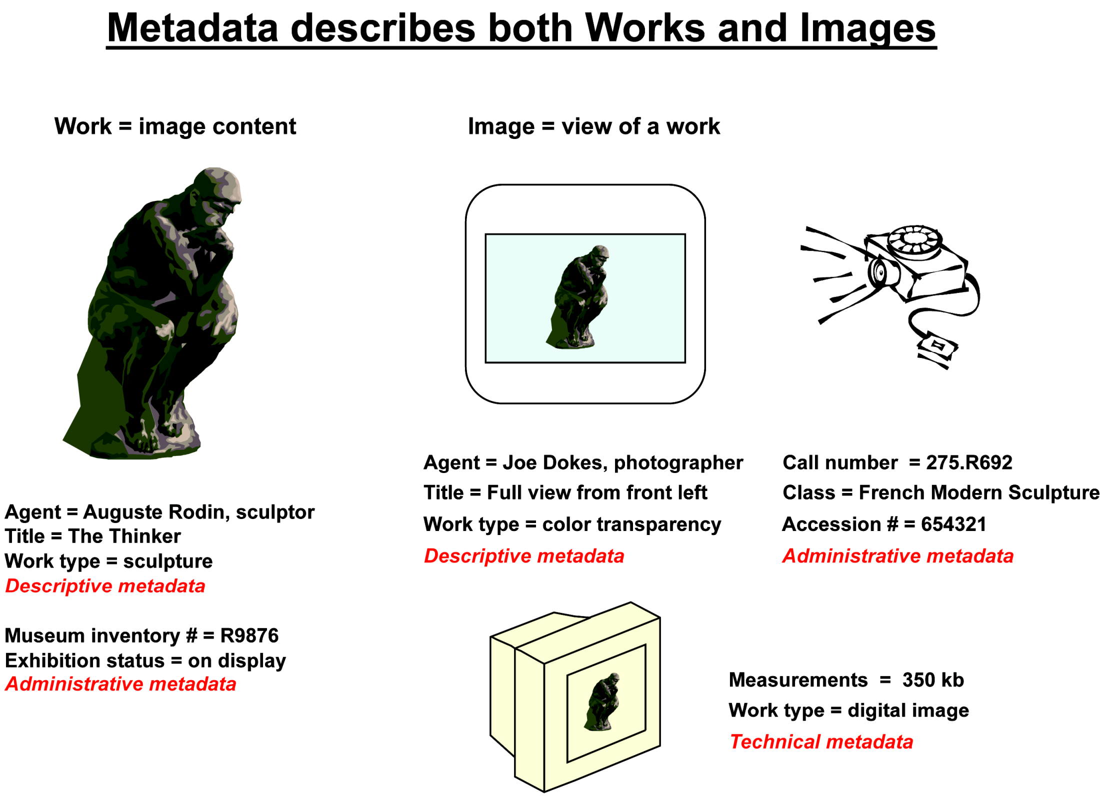
Minimal Catalog Records. Practical considerations often limit the ability of a cataloging staff to enter as much metadata about works and images in a collection as might be desirable. Deciding just how much information to record can often be a challenge. What constitutes a "minimal" record varies according to the type and composition of a collection and will largely be determined by the project timeline, budget, and retrieval needs. Establishing local guidelines about what constitutes a minimal record for the collection at the beginning of a project can ease the process. A good descriptive catalog record should contain enough information to be able to answer the following questions about a work or image:

1. What is it called?
2. What is it?
3. Who made it?
4. When was it made?
5. Where was it made/where was it found/where is it now?
6. What is it of or about?
7. Who owns it?

These questions are answered by descriptive metadata categories such as:

1. Title (what is it called?)
2. Work type (what is it?)
3. Creator (who made it?)
4. Date (when was it made?)
5. Creation location/Discovery location/Current location (where was it made/where was it found/where is it now?)
6. Subject (what is it of or about?)
7. Copyright (who owns the rights to it?)

It is important to note that the same questions apply to both works depicted in an image and the image file itself. The questions are the same, but the answers are different. If the data model stores these values in different tables there will be no confusion about which values pertain to the work and which to the image *(fig. 3).*



Metadata describes both Works and Images

Work = image content

Image = view of a work

Agent = Auguste Rodin, sculptor
Title = The Thinker
Work type = sculpture
Descriptive metadata

Museum inventory # = R9876
Exhibition status = on display
Administrative metadata

Agent = Joe Dokes, photographer
Title = Full view from front left
Work type = color transparency
Descriptive metadata

Call number = 275.R692
Class = French Modern Sculpture
Accession # = 654321
Administrative metadata

Measurements = 350 kb
Work type = digital image
Technical metadata

Figure 3: The Same Metadata Elements Describe both Works and Images

Minimal Metadata Records

Different kinds of works have different information needs. While a minimal descriptive catalog record should contain enough information to identify an object, or group of objects, and identify its differences from other similar objects, the data fields needed to accomplish this disambiguation will vary depending upon the type of item being cataloged. For example, an excavation site name might be the only piece of data that distinguishes one potshard from another, whereas the creation date might be the distinguishing factor between two landscape paintings by the same artist in the same museum collection or between two versions of the same poem. However, there are some basic categories of information that are considered “core” for certain work types, and they all serve to answer the same basic questions.

1. What is it called? - Title
2. What is it? – Work type
3. Who made it? - Creator
4. When was it made? - Date
5. Where was it made/where was it found/where is it now? - Location
6. What is it of or about? - Subject
7. Who owns it? - Rights

The following examples illustrate what minimal display value metadata records might look like for a few broad content categories. Illustrated examples of both full and minimal cataloging records may be found at http://www.vraweb.org/projects/vracore4/vracore_examples.html.

Works of Visual Art and Architecture

1. Title (preferred and alternate for both work and component works)
2. Work type (e. g., painting, sculpture, cathedral, cookie jar)
3. Creator (artists' names and roles)
4. Date (creation, discovery)
5. Location (creation, discovery, repository)
6. Subject
7. Rights (creative)

Additional categories that may be important for disambiguation

8. Repository number
9. Related textual reference (e. g., corpus or catalog raisonné number)
10. Related works (larger entities)
11. Measurements (framed and unframed, with or without base, etc.)
12. Materials and techniques (media type)
13. Inscriptions (e. g., hallmarks, monograms, signatures, captions)
14. State and/or edition (for prints)
15. Physical description (e. g., shape, condition, unique physical characteristics)

Published and Unpublished Print Materials

1. Title (preferred and alternate for both work and component works)
2. Object type (e. g., book, article, letter, diary, engraving)
3. Creator(s) (name and role)
4. Date (creation, publication, edition)
5. Location (creation, repository, publication, discovery)
6. Subject(s)
7. Rights (creative, publication, performance)

Additional categories that may be important for disambiguation

8. Repository number
9. Measurements (e. g., size, number of pages, plate marks)
10. Materials and techniques (e. g., engraving, linotype, vellum)
11. Inscriptions (e. g., watermarks, monograms, signatures)
12. State and/or edition
13. Physical description (e. g., shape, condition, unique physical characteristics)

Time Based media (moving images, audio recordings, performance art)

1. Title
2. Work type (e. g., motion picture, performance, sound recording)
3. Creator(s) (names, roles, and extent for each)
4. Date(s) (creation, recording, distribution, performance)
5. Location (performance location)
6. Rights (creative, distribution, performance)
7. Subject(s)

Additional categories that may be important for disambiguation

8. Physical Description (free-text description of plot, content, track listing)
9. Measurements (running time, duration)

Electronic Resources (websites, born-digital works)

1. Title
2. Work type (e. g., website, sound file, electronic art)
3. Creator(s) (name and role)
4. Date (creation, version)
5. Location (URL, URI, PID)
6. Subject(s)
7. Rights (creative, licensing)

Additional categories that may be important for disambiguation

8. State/Edition (version)
9. Measurements (bit depth, pixel dimensions)
10. Physical description

Final thoughts

There are many reasons for embarking on a digitization project. But regardless of the reason, thinking through all the metadata issues at the outset will help clarify and identify what information is important for you to record in order to achieve a satisfactory end result. Whether your project is large or small, a collection of rare museum objects or a shoebox full of old family photos, considering these issues in the beginning will help you identify appropriate tools and workflows that enable you to create clear, consistent metadata records that will meet your needs both now and well into the future.

Resources

American Library Association. *Anglo-American Cataloguing Rules,* 2nd ed., 2002 revision. Chicago: American Library Association.

Baca, Murtha, and Visual Resources Association. *Cataloging Cultural Objects: A guide to describing cultural works and their images*. Chicago: American Library Association, 2006.

Baca, Murtha, and Patricia Harpring, eds. Categories for the Description of Works of Art. [online]. Los Angeles: J. Paul Getty Trust and College Art Association. 2000. http://www.getty.edu/research/conducting_research/standards/cdwa/ (accessed August 3, 2010)

CDWA Lite: XML Schema Content for Contributing Records via the OAI Harvesting Protocol. Los Angeles: J. Paul Getty Trust. 2005. http://www.getty.edu/research/conducting_research/standards/cdwa/cdwalite.html (accessed August 3, 2010)

Digital Curation Centre. http://www.dcc.ac.uk/.

Getty Vocabulary Program. *Art and Architecture Thesaurus* (AAT). Los Angeles: J. Paul Getty Trust. 1988-. http://www.getty.edu/research/conducting_research/vocabularies/aat/ (accessed August 3, 2010)

__________*Editorial Guidelines.* Los Angeles: J. Paul Getty Trust, 2003. http://www.getty.edu/research/conducting_research/vocabularies/editorial_guidelines.html (accessed August 9, 2010)

__________*Getty Thesaurus of Geographic Names* (TGN). Los Angeles: J. Paul Getty Trust. 1988-. http://www.getty.edu/research/conducting_research/vocabularies/tgn/ (accessed August 3, 2010)

__________*Metadata Standards Crosswalk*. Los Angeles: J. Paul Getty Trust, 2009-. http://www.getty.edu/research/conducting_research/standards/intrometadata/crosswalks.html (accessed August 9, 2010)

__________*Union List of Artist Names* (ULAN). Los Angeles: J. Paul Getty Trust. 1988-. http://www.getty.edu/research/conducting_research/vocabularies/ulan/ (accessed August 3, 2010)

Library of Congress. *Encoded Archival Description. Version 2002 Official Site.* Washington, D.C.: Library of Congress, 2010. http://www.loc.gov/ead/ (accessed August 23, 2010)

Library of Congress Authorities. *Library of Congress Authorities.* Washington, DC: Library of Congress, 2005. http://authorities.loc.gov/ (accessed August 9, 2010).

Society of American Archivists. *Describing Archives: A Content Standard.* Chicago: Society of American Archivists, 2004.

SPECTRUM. The UK Museum Documentation Standard. Revised ed. 3.2, edited by Gordon McKenna and Efthymia Patsatzi. Cambridge: The Collections Trust, c. 2009. http://www.collectionstrust.org.uk/spectrum (accessed August 9, 2010).

Visual Resources Association. VRA Core 4.0. http://www.vraweb.org/projects/vracore4/index.html (accessed August 9, 2010)

Weber, Mary Beth. *Cataloging Nonprint and Internet Resources.* New York: Neal-Schuman Publishers, Inc., 2002.