Follow up Information

Caring for Digital Materials:
Metadata, finding aids, and digital asset management

Danielle Cunniff Plumer, instructor

Caveat: In the Q&A below, I generally discuss tools for Windows computers (and sometimes for Linux), as I am not as familiar with Macintoshes. If you have specific questions about Mac software, feel free to email me at danielle@dcplumer.com and I will do my best to research the options and get back to you.

Metadata

Q. *Wikipedia says: The term metadata refers to "data about data". http://en.wikipedia.org/wiki/Metadata . Can you explain if this definition is same as archival?*

As I noted in the presentation, in the cultural heritage community we like to add the qualifiers that our metadata is structured, designed for a particular community of users, and based on formal standards. These standards are designed to help us with the long-term management of our physical and electronic resources.

Metadata Standards

Q. *We've talked about putting a MARC record at the library across the street which has collections directly related with ours. Would this 'work'? Can it be done from an EAD or is it something entirely new in terms of managing the metadata?*

It can be done and has been done by many institutions. There is a free tool that many library catalogers use, called MarcEdit (http://people.oregonstate.edu/~reeset/marcedit/html/index.php), that can create MARC records from EAD [Encoded Archival Description] finding aids. You can see a video tutorial on this at http://www.youtube.com/watch?v=9LP4TTHx7e0.

Q. *How does the adoption of RDA relate to EAD finding aid creation?*

At this time, the Society of American Archivists has not published the revision of their content standard (DACS), to include updates related to RDA, the library content standard that updates the Anglo-American Cataloging Rules (AACR2); the new revision of DACS should be available this summer and I believe will be offered as a free download from the SAA website. Most archives use DACS when creating finding aids. However, there was a webinar last year which addressed ways to use RDA in DACS; you can find the link to the recording at http://www.ala.org/alcts/confevents/upcoming/webinar/cat/053012.

Note: In the presentation I referenced a site that gives you instructions on creating EAD from Excel spreadsheets; it is http://orbiscascade.org/index/northwest-digital-archives-tools.

Q. *Can you use two standards? I work in a history museum and would use DACS and CCO.*

You certainly can. Some institutions develop "application profiles" or local cataloging guides that include elements from different content standards and metadata schema. You may have to implement this in your collections management database using custom fields, however, so it is essential that you develop and maintain complete documentation on how the fields should be used.

Nomenclature and other controlled vocabularies

Q. *I use Past Perfect and know that Nomenclature 3.0 is part of it - I was wondering if something similar to it could be created specifically for digital material as nomenclature does for artifacts.*

PastPerfect and other content management systems that allow you to create custom controlled vocabularies and metadata can be adapted to use vocabularies like the ones recommended for use with the Dublin Core elements format and type or metadata elements from the NISO Data Dictionary and PREMIS. However, I'm not aware of any collections or digital asset management systems that come with these types of features fully implemented because the implementations tend to be very institution-specific (your institution does it one way, mine does it another, etc.)

Crosswalks

Q. *I loved your example in your slide showing the Dublin Core versions of museum fields, are there other examples that show where common museum fields could fit into the Dublin Core fields?*

I found a guide from the Sewall-Belmont House and Museum that documents their cataloging practice and references both the PastPerfect field and the Dublin Core element, when applicable (http://www.sewallbelmont.org/wp-content/uploads/2011/02/SBHM-Metadata-Plan.pdf). It also provides examples of how the field should be used. This is an example of what I call an "application profile," and it is a great example of the sort of documentation you should try to create for your projects.

The standard publication describing metadata crosswalks is "Crosswalks, Metadata Harvesting, Federated Searching, Metasearching: Using Metadata to Connect Users and Information" by Mary S. Woodley in *Introduction to Metadata* (Version 3.0) from the Getty Institute, available at http://www.getty.edu/research/publications/electronic_publications/intrometadata/path.html.

It looks at the Categories for the Description of Works of Art (CDWA), MARC, EAD, and Dublin Core.

The Library of Congress and other institutions have developed a number of standard crosswalks. Users of the MarcEdit program described above can access many of these.

File naming conventions

Q. *We are scanning obituaries and saving them as pdfs. We name them with the surname, first name, then maiden name, and date of death and put each in the alpha file according to first letter of the surname. What problems will we have in the future using this method?*

As one of the other participants noted, it is best to make sure that the filenames you are using are unique. It is not unheard of for newspapers to run an obituary more than once, for example. However, if you are confident that your system gives you unique filenames, then my recommendation is to make sure that you don't use spaces or special characters other than hyphen (-) and underscore (_) in the filenames and that you put the dates in ISO 8601 format (YYYY-MM-DD). Sample: Plumer_Eliane_Dessaux_2000-09-22.pdf . Be consistent and document your practice.

Q. *What other resources are there for file naming assistance?*

As I mentioned in the presentation, file naming is one of the most difficult institutional practices to standardize, so spend some time researching it and making your decisions. One participant in the session recommended www.controlledvocabulary.com/imagedatabases/filenaming.html, which is a very good source. The site is maintained by David Riecks, who is very knowledgeable about metadata.

Another good review of file naming standards comes from the JISC Digital Media site: http://www.jiscdigitalmedia.ac.uk/guide/choosing-a-file-name. JISC is a United Kingdom organization with the mission of supporting higher education and research. Many of their guides to digitization (which they spell "digitisation") are excellent.

Preservation Metadata

Q. *Do you consider metadata for monitoring physical media (e.g., when to verify media is still viable; when to migrate, etc.)?*

This is a good example of metadata used specifically for preservation. The PREMIS data dictionary includes entities such as eventType which are used to record actions taken on a digital object. For a given object, one might create metadata for an event such as:

- Object
  - ObjectIdentifier = 014010058365

- StorageMedium = MagneticTape
- Event
  - EventIdentifier = 002764
  - EventType = mediaRefreshment
  - EventDateTime = 2008-04-16

This would tell the repository manager that the object, stored on magnetic tape, was last "refreshed" or copied onto fresh storage media in 2008. An alert could be set to remind the manager to refresh the object again prior to the expected end of life for that medium (magnetic tape is expected to have a lifespan of approximately 10-20 years, so ideally the object would be copied to new media in 2018 or so. Multiple copies are, of course, still needed).

Embedding metadata

Q. *If we are accepting born digital materials would you recommend saving the file before monkeying around with the metadata in headers?*

I personally do think that it is important to keep a copy of the original file without any modification and to keep separate copies that have changes, including revised filenames and metadata. In the OAIS model for digital preservation, you can think of these as "Submission Information Packages" and "Archival Information Packages." However, not all institutions can afford the additional file storage for this. The best practice is to document how you treat electronic files and then to be consistent in your practice.

Note that while changing the name of a file will not change the checksum for the file, altering the embedded metadata (file properties) **will** change the checksum!

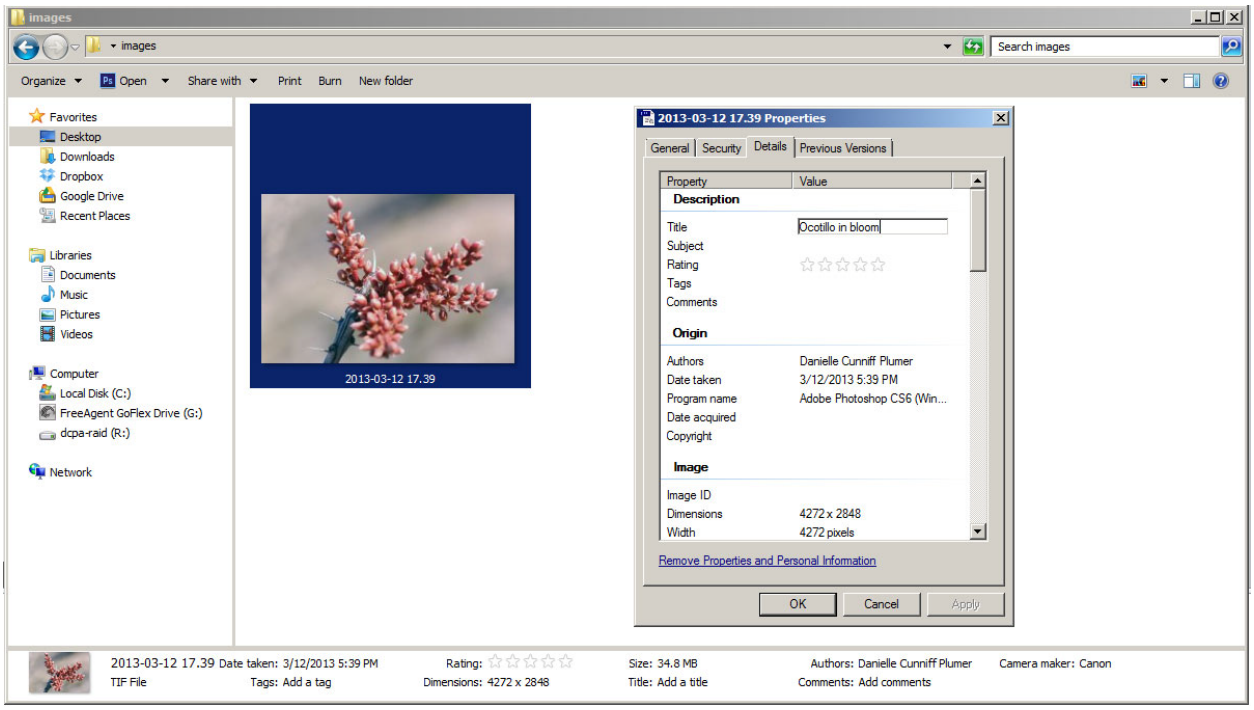
Q. *I used Windows to add data to photos but then couldn't find that information when in Adobe Bridge.*

I should have been clearer about this in the presentation; I apologize! Starting with Windows 7, the metadata is written in IPTC format in the header of the file, and should be available to other programs, such as Adobe Bridge. In earlier versions of Windows, the metadata was in a proprietary format that other programs couldn't read. In general, it is better to use a full-featured editor like Adobe Bridge, Lightroom, or Aperture, because they offer more standards-compliance.

Tip: If you work in a small non-profit institution, be sure to check out TechSoup (http://www.techsoup.org/), which offers free or greatly discounted versions of software offered by major software companies, including Adobe. Libraries should go to TechSoup for Libraries (http://techsoupforlibraries.org/).

Q. *I'm a little unclear on entering metadata into the header of digital files. Does this have to be done file by file? Can it be auto-generated?*

If you are using the Windows Explorer to generate metadata as in the example below, the metadata is usually entered one file at a time.



However, you can select multiple files, right-click on them, and enter metadata for all the files you have selected. The metadata values will be the same for all the files, however, so this may not be useful to you. Similar "batch" operations can be done in Adobe Bridge and other programs.

Q. *Is there a way to create some kind of global inventory of the metadata that is embedded within Windows Explorer files? Is there a way to export it all into an Excel file?*

There are many ways to extract embedded metadata. Most of the tools used by digital archivists write the metadata to an XML file, which can then be stored in the same folder as the image. Two of my favorite tools for extracting technical metadata are the New Zealand Metadata Extractor (http://meta-extractor.sourceforge.net) and JHOVE2. Both of these are fairly well-documented, though the configuration options may be a bit complicated for new users. Both programs run in a Java Runtime Environment, so you will need to install Java to use them (http://java.com/en/download/index.jsp). They are really intended to let you harvest technical and preservation metadata, though, so they may not extract all the descriptive metadata you are interested in, and you can't easily import that metadata into Excel.

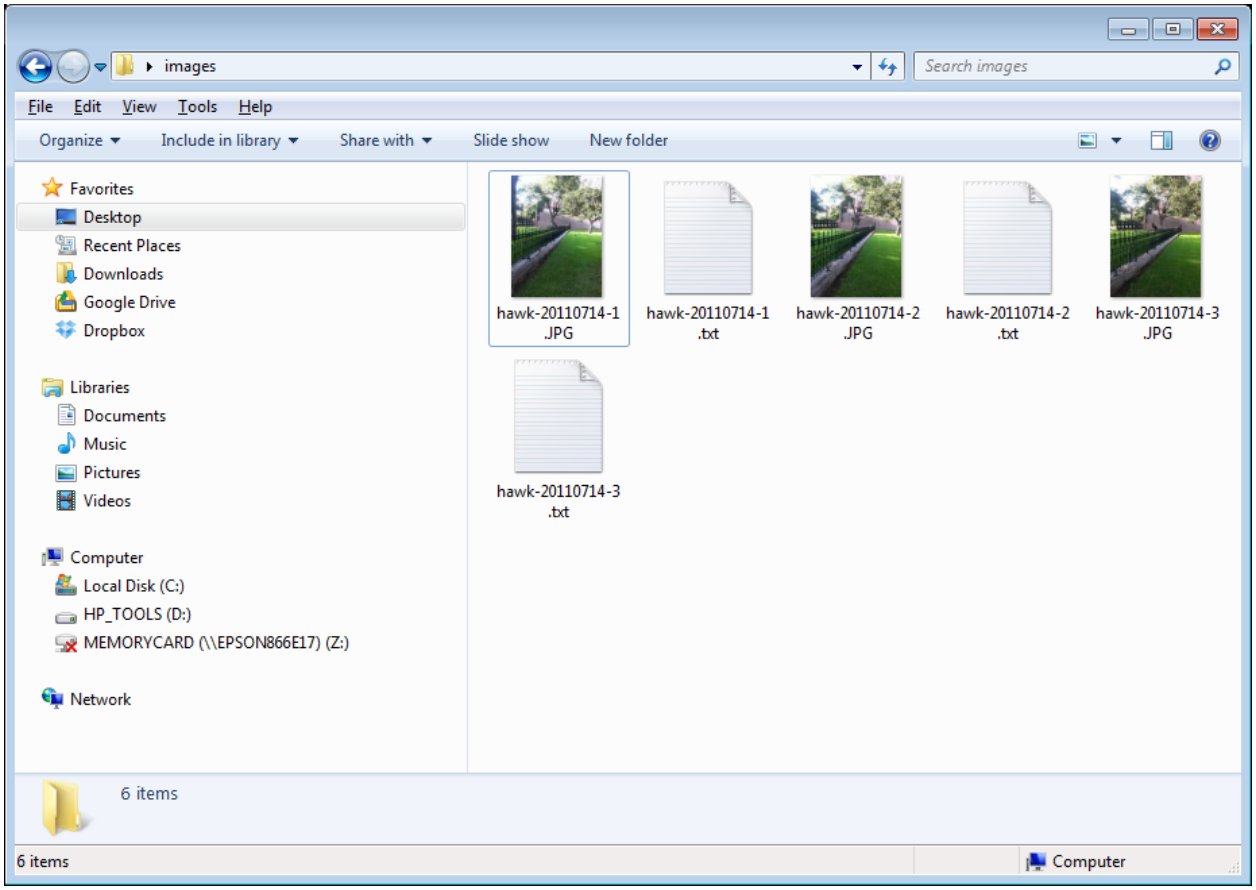
The free ExifTool (http://www.sno.phy.queensu.ca/~phil/exiftool/index.html) software by Phil Harvey is a command-line utility (available for Windows, Macintosh, and Linux) that

lets you both export metadata from and import metadata to a variety of file types. You can export metadata to a CSV format that can then be opened by Excel. This is not the easiest program to use if you are not reasonable proficient at command-line operations, but it is an extremely powerful tool for metadata work.

Other participants recommended some tools to help users create lists of files, and in some cases metadata, from Windows systems. I tested the free versions of two of the recommended tools, TreeSize (http://www.jam-software.com/treesize/) and DirectoryListPrint (http://www.infonautics.ch/directorylistprint/), and they do allow you to get a list of the files on your system, much as the command "ls –al | file" would do on a Linux system, but they don't output metadata.

Q. *What does non-embedded metadata look like - a text file in the same folder as digital image files - one text file with metadata for all files in that folder?*

There are many different ways to create non-embedded metadata. The simplest way is to create a text or XML file corresponding to every digital object, either stored in the same directory as in the image below or with one digital object and metadata record per directory. You might also create a CSV file with all the metadata for all the items in a directory, using Exiftool or some other tool.



Advanced: One current best practice for storing digital objects plus their metadata comes from the BagIt specification developed by the Library of Congress, Stanford University and the California Digital Library (see http://www.digitalpreservation.gov/series/challenge/data-transfer-tools.html for details). In this specification, a folder (or directory, for Linux users) is created for every digital object. A digital object may consist of one or more digital files (images, audio, video, etc.). A brief metadata record may also be created in that same folder, as will a manifest including the checksum for the collection of objects included in the "bag." The Library of Congress has created a video introduction to BagIt, available at http://www.digitalpreservation.gov/multimedia/videos/bagit0609.html.

Q. *If photo metadata is embedded as part of the photo, should we attempt to separate it?*

This is something you need to decide on in your local practice. Most institutions manage their metadata in a dedicated collections management database or digital asset management system, for convenience. You may want to copy the technical and preservation metadata from the digital objects into this system when you first add them. After that, you may choose to update the embedded metadata periodically or store metadata files with the digital objects. As always, document your decisions and be consistent.

Embedded metadata for electronic documents

- Microsoft Office. Metadata for Microsoft Word and other Microsoft Office documents can be accessed through the document "properties." Some metadata is automatically generated – for example, the "author" is automatically entered with the name of the person who first created the document. If Word does not know the correct name (such as might happen when an IT department installs the software onto a new computer), the metadata will be wrong. For instructions on accessing the metadata, see http://office.microsoft.com/en-us/word-help/view-or-change-the-properties-for-an-office-document-HA010047524.aspx. In newer versions of the Office suite, there is an option to delete all the properties to prepare a file for sharing; use this option carefully.

- Adobe PDF. Metadata for Adobe PDF documents is similarly accessed through document "properties." You may need a full version of the PDF software to be able to add or edit metadata this way. For instructions, see http://help.adobe.com/en_US/acrobat/X/standard/using/WS58a04a822e3e50102bd615109794195ff-7c66.w.html

Embedded metadata for image files

Q. *Do you embed technical metadata in derivative JPEG files as well?*

Technical metadata is most typically automatically generated by the programs you use to create the derivative files, and you may not be able to change the values. You certainly can and probably should embed additional descriptive metadata if you will be making the files

available over the Internet, to help ensure that copyright and other information remains with the image (note, however, that it is difficult to prevent others from modifying this metadata). However, you will probably need to do this as a separate step in your processing workflow, as most programs don't automatically transfer metadata from the master image to the derivatives.

Metadata for audio and video files

Q. *Several Oral History institutions that I have referenced suggest stripping metadata from the audio file to preserved fixity. What are your thoughts?*

I'm not quite sure what the recommended practice is, here. Inclusion or exclusion of metadata does not affect the fixity of the digital object, though the object's checksum will change if metadata in the header of the object changes. It may be that the institutions are recommending that the original digital file be preserved intact, without adding metadata, and that additional metadata either be handled externally or added to the header of a **copy** of the digital object.

Q. *What are some of the best metadata editors out there for audiovisual materials? I know, for example, that Audacity, has a great metadata editor for audio files. What common audiovisual applications (e.g. - Adobe Premiere, etc.) have good metadata editors for audiovisual (moving image) files.*

Audacity's Metadata Editor will let you modify descriptive metadata for audio files (see http://manual.audacityteam.org/man/Metadata_Editor for details), but it does not really support the detailed preservation metadata we need for long-term preservation. For audio files, I particularly recommend using BWF MetaEdit (http://sourceforge.net/projects/bwfmetaedit/), a free, open source tool that supports embedding, validating, and exporting of metadata in Broadcast WAVE Format (BWF) files. It was developed by the Federal Agencies Digitization Guidelines Initiative, supported by AudioVisual Preservation Solutions (http://www.avpreserve.com/). For more about embedding metadata in Broadcast WAVE Format files, see http://digitizationguidelines.gov/guidelines/digitize-embedding.html.

For other tools, including some that work with video files, see http://www.avpreserve.com/avpsresources/tools/.

Q. *I'm specifically interested in metadata for digital video files - learning about MXF but not sure we'll end up using software that is capable of this, besides for a separate database, do we have other options?*

You may want to look at the University of Texas at Austin's Metadata Guidelines for Video, developed as part of their work on the Human Rights Documentation Initiative project (http://www.lib.utexas.edu/schema/Video_Metadata_Guidelines_v1.pdf). They store the metadata in separate METS files, not embedded into the video itself. You could also store the metadata elements in a database.

Many institutions are using the PBCore metadata schema for digital video. Even though PBCore was specifically developed for use by public broadcast video, it is a flexible standard that can be adapted to many uses. For information about the standard, see http://www.pbcore.org/. Unfortunately, work on PBCore has currently stopped due to funding and governance issues.

Q. *Are there any good examples of how PBCore is being used?*

There are many case studies listed on the PBCore website; see http://www.pbcore.org/category/case-studies/. There's also a very nice write-up of one project that used PBCore as part of the American Archive pilot digitization project at http://www.pbcoreresources.org/article/use_of_pbcore_in_the_american_archive_pilot_project/ .

Accessibility

Q. *How are others writing descriptions to make them accessible to deaf or blind patrons?*

Metadata is how we provide access to a lot of our digital materials, and full descriptions are a really good way to provide that access. Some institutions are legally obligated to provide accessible alternatives and for others it's still a good idea (among other things, it helps your objects show up better in internet search engine results). There have been a number of projects that have tested different types of metadata, transcription of handwritten materials, and rich description for digital objects.

Copyright

Q. *Will circulating materials donated from different sources in digital format violate the copyright?*

Copyright depends on many factors. If materials are not in the public domain, your deeds of gift or purchase agreements should describe the copyright status and permissions available for the original items and should also explain what you can do with digital copies of the items. The Society of American Archivists has a useful guide to deeds of gift available online at http://www.archivists.org/publications/deed_of_gift.asp. Many institutions beginning digitization projects have found that they need to update their deeds of gift to permit putting copies of the items online.

Other Questions

Q. *Off the topic, how do you get speech recognition of the speaker (software, etc.)?*

Doug Oard, a researcher in search and information retrieval at the College of Information Studies and the Institute for Advanced Computer Studies at the University of Maryland, College Park, wrote a very nice article discussing the state-of-the-art in automated speech recognition (ASR) software in 2012 for the *Oral History in the Digital Age* project. It is available at http://ohda.matrix.msu.edu/2012/06/automatic-speech-recognition/. Currently, there are no consumer systems that can do better than a poor job at recognizing speech in uncontrolled environments. However, this is a very active area of research, and better systems may be available in a few years.

Q. *Is there a technical metadata capture tool for utf-8?*

> You may be thinking of textMD, which is a schema for technical metadata for textual objects (http://www.loc.gov/standards/textMD/). It includes elements for character set used in the textual object, such as UTF-8, along with other information.

Other questions? Email info@heritagepreservation.org and include "Caring for Digital Materials" in the subject. Or join the community at http://www.connectingtocollections.org/ and ask your questions there!