Digitization for Conservators

Preservation of Books - Photos - Archival Material (SCIE0047)
Donna MacLeod, MLIS

June 24, 2016

What will we look at?

- What will we digitize and how will we prioritize it?
- How do I do it? What do I need?
- What metadata do I have, do I need?
- What safety precautions do I need to take (material, people, files)?
- Who should do the work? In-house vs. hiring a company
- What happens once it's digitized?

Why do we digitize?

- Preservation
- Access
- Funding (If the reason for the project is “because we got funding”, clarify the actual goals of the project and ensure that long-term storage is taken into account.)

- Knowing WHY you are doing something helps you set the priorities and objectives for your project. It also helps you to define what “success” looks like.

What do we digitize?

- The first consideration should be: “Can we digitize it safely?”.
- You might also consider: “Has it already been digitized?” and “Is it in copyright?”.
- Types of materials you may be asked to digitize
  (WARNING: Library Speak ahead)
  - Books and manuscript material
  - Pictures (photographs, paintings of different media, drawings, etc.)
  - Maps
  - Ephemera (printed matter not meant to be preserved)
  - Realia (three-dimensional objects from real life that do not easily fit into the orderly categories of printed material). Think Ming Vase.
  - Audio-visual materials like cassette and VHS tapes

Before you start

- Take an inventory of all your materials
- Prioritize your collections
  - Significant & unique items
  - Popular or high-demand items
  - Fragile items or obsolete items
  - Community initiatives
- Figure out copyright (you may need to get permission to reproduce or publish material)
- Make a plan based on your priorities and include a budget
  - Equipment: purchase, lease, share with other community members
  - You may need to do a conservation treatment to stabilize the items before digitizing them.
- Decide what you want the end-product to look like and do, and how you will deliver it online.
- Document early, document often, document everything. (And then, do it again.)

When: Project deliverables and timelines

- Expectation-management is very important, communicate early and often
- Use actual measurements with your own system and processes whenever possible. Numbers are your friend, timing helps you project more realistically.
- Build time into the estimate for the inevitable delays and complications
- Things you need to consider
  - Timelines you are required to meet (helps define scope of project: must-have vs. nice-to-have)
  - Funding (the conditions of funding often help define the must-haves, timelines, staffing, etc.)
  - Equipment (Do you have enough of it? Is it shared? What if it breaks? Can you get service or replacement parts? Do you have a backup plan?)
  - Staff availability and suitability

Who: In-house vs. hiring externally

- Hiring an external provider is sometimes cheaper, faster and produces better quality
  - Do your due-diligence, make sure
    - They have the equipment to do the job properly
    - They conform to standards
    - They have a proven track record and agree to follow care & handling guidelines to the letter
  - Ask to speak to some of their clients, to see samples of work they have done for others
  - All your decisions must be made ahead of time and thoroughly documented, including formats, quality control, delivery and timelines
- Having in-house staff means you don't have to send your material out. You can control the project and directly supervise the work.
- Either way, provide detailed care & handling training for everyone involved

Who: Staffing a digitization project

- Do they have to share equipment?
- Staff must pay close attention and be consistent for long periods of time. This is very hard work.
  - Do they have the skills (image, metadata, organization, project management)?
  - Do they have the patience and endurance required?
  - Are they detail-oriented and consistent, or easily bored and motivated by novelty?
  - Will they handle the material properly?
  - Can they resolve problems independently and reliably?
  - It's best if you have one person or team work on a project, take responsibility and assume ownership of it. A large number of temps is NOT the way to go. For example, using volunteers or students might mean that your product is inconsistent.

…. and now ….

- HOW DO YOU DO IT?

How: Finding it again

- So, your goal is a bunch of digital stuff.
- You need to find it again, maybe publish it so other people can find it too.
- That means
  - Every item must be identifiable (distinguishable from all the others, that is, unique)
  - Every item must be described in such a way that retrieval is possible
  - Every item must have a location
  - Every item must belong to a larger structure
- You'll need a web server, a content management system, access tools and interfaces for: browsing, searching, navigation, maybe streaming, access controls for restricted information, persistent URLs.

How: Naming it

- The most basic form of identification is the filename
  - You can't save the file without it
  - Most file systems only permit unique names. This also avoids ambiguity and reduces the risk you will over-write a file. Don't make the folder structure part of what makes the name unique, because you will eventually move the file and then problems occur.
    - Don't use: m:/BigProject1/First-file.tif and then use m:/BigProject2/First-file.tif
    - Instead use: m:/BigProject1/BP1-f1.tif and m:/BigProject2/BP2-f2.tif or some other convention to uniquely identify the items.
  - Changing filenames afterwards is difficult (and confusing).

Filenames

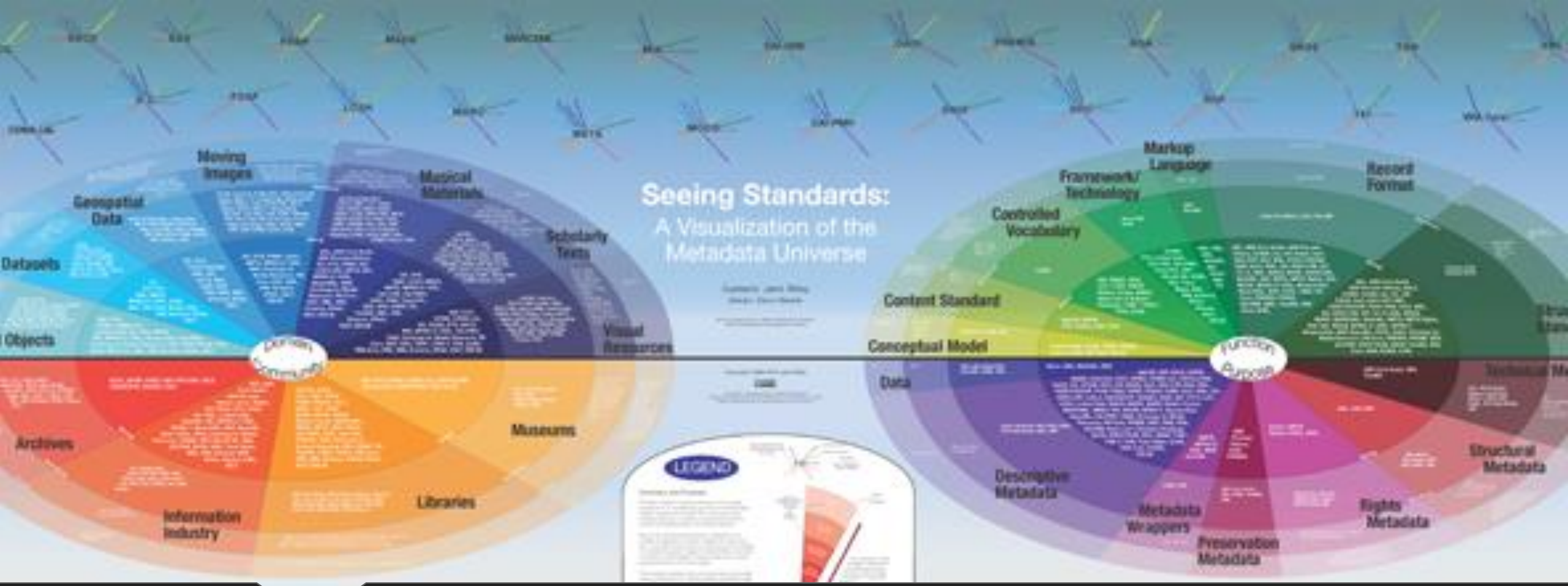
- Two ways to name a file
  - Human-readable
    - Easy to read and share
    - Must be created by hand (easy to make a mistake)
    - Tempting to change later on
  - Machine-readable
    - Can be assigned automatically or generated quickly
    - Have to copy it (not type it)
    - Meaningless
    - Can stay the same regardless of later situations, less tempting to change it

Metadata

- Metadata is "data about data"
- Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource
- Different kinds (how many different kinds depends on who you ask)
  - Descriptive: describes the item for discovery (e.g. title, creator, date, format, abstract, keywords)
  - Structural: describes how items fit together (e.g. files in a box in an archival collection)
  - Administrative
    - When and how the file was created, equipment used, etc.
    - Rights management
    - Preservation information (needed to preserve the digital file, including migration history, encoding, etc.)
    - Usage metadata (for stats)

Metadata is all about standards

- Find a standard, understand it, and apply it rigorously
- Standards help people use metadata in the same way
  - Field definitions are maintained in one place (machine-readable)
  - Data that use the fields in the same way can be combined into larger collections
- Different communities of practice use different standards
  - Dublin Core is big in libraries (http://dublincore.org/)
  - EAD is widely used in archives (encoded archival description http://www.loc.gov/ead/)



The sheer number of metadata standards in the cultural heritage sector is overwhelming, and their inter-relationships further complicate the situation.

Each of the 105 standards listed here is evaluated on its strength of application to defined categories in each of four axes: community, domain, function, and purpose.

Source: Jenn Riley & Devin Becker, Indiana University, 2009-10. Released under a CC 3.0 BY/NC/SA license. http://jennriley.com/metadatamap/

How: Image format

- Image format
  - TIFF (Tagged Image File Format) used for the uncompressed master archival file format for digital reproductions from paper and photographic media such as negatives. These are not modified. Include a ruler for scale and a standardized colour chart.
  - JPEG (File Interchange Format) or JPEG200 files, derived from the master TIFF are used to deliver images online. These can be modified (for example, in Photoshop) as required. They do not include ruler or colour chart and are usually cropped. This is format used for distribution, it is NOT good for preservation.
  - JPEG2000 provides the ability to zoom without loss of image quality
  - PDF or PDF-A (Portable Document Format – Archival) are used for primarily textual information or for multi-page items. Use optical character recognition (OCR) to make them full-text searchable.
- Standard formats for audio, video and other types of file have also been identified. There is less consensus around what the best preservation format for these items is.

How: Image compression and resolution

- Image compression causes data loss but makes the files smaller
  - Master files are not compressed to preserve the most information possible (scan as high as possible and don't mess with it)
  - Distribution files are compressed to make them easier to transmit (changes depending on purpose)
- Image quality is determined by resolution (the number of pixels per square inch)
  - For display on a screen, you need a resolution of 72 dpi (dots per inch, also called ppi)
  - For printing at a high quality, you need from 300 to 600 dpi

Image resolution

- Changing the resolution of an image affects the dimensions of the image
  - An 8MP image at 72 dpi might be 48" x 32"
  - The same image at 300 dpi might be 11.5" x 7.7"
  - Think of it like making the pixels smaller so that more of them fit into the same 1-inch space
- Changing the resolution of an image affects the file size
  - An image at 72 dpi is a lot smaller than the same image at 300 dpi

How: Equipment

- Two main types of digital reprography equipment
  - Scanner
    - Captures the image in a line that moves across the page
    - Slow, variable resolution, mainly for flat objects document scanner images
    - Item must be placed flat on the glass (fixed focal plane)
  - Digital camera
    - Captures the image all at the same time
    - Faster, constant resolution, more adjustable
    - Item does not need to be flat (you can use book cradles, etc.)
- Book scanners are generally camera-based systems

How: Storing it

- Long-term storage is a requirement for a digitization program
  - When planning, be sure to calculate adequate space on a stable, long-term storage media
  - Unless the project is finite, plan to add additional storage as you go (do not use more than 90-95% of available space)
  - Digitizing for quality of image (fine art) or digitizing for quantity (books, textual material)
- Additional space for regular backups is also required
  - Online or offline storage
  - Back up offsite or at least on a network drive
  - Make at least two copies (ideally, stored in different location) — LOCKSS
  - Automate backups so you don't forget
- You can archive on your machine or you can archive in the cloud (somebody else's computer) but no industry-wide cloud archiving standards exist, so read the fine print

How: Storing it — File management

- Best-practice is to use a content management system (CMS)
- Decide ahead of time on your file-management structure and stick to it consistently (don't let the digitizers do their own thing, everyone must do it the same way)
- Digital items have many parts (image, metadata, preservation files). Keep them together.
- Establish a regular back-up schedule (LOCKSS)

Where: The room

- Do you have enough physical space? (safety of people, objects)
- Is the lighting consistent, indirect? (not daylight, task lighting that can be turned on for work but off for the safety of the items)
- Are there enough tables or work area for preparation?
- Is there an adequate power supply (enough outlets)?
- Do you have the proper equipment? (book cradles, supports, magnets)
- Are there secure overnight storage and acceptable (monitored) environmental conditions?

Digitization

- Consider each type of item
  - Are you digitizing for quantity (text) or quality (images)?
  - Equipment used varies with type of material
    - Flat, one sheet, possibly folded
    - Book material (problems: large, tight gutters, foldouts, brittle, hard to open)
    - Large items like maps and some images may need to be done on an overhead camera or scanner
- Consider individual items, evaluate needs and consider safety of each item
- Is your process designed to scan and process (post-production, metadata, publication, storage) one item, or to scan multiple items and then process them?

Post-production: Adjusting the image

- Often, when digitising historical material, it is important to maintain the look and feel of the original, including defects and imperfections, for ethical and image-integrity reasons
- Usually, only these adjustments are made
  - Brightness and contrast
  - Cropping that does not interfere with the integrity of the object (remove white borders or mounting boards)
- In certain circumstances, distribution images — but never master images — can be retouched (label removed, digital tear repairs, etc.) depending on their purpose. This can be time-consuming, however.

Post-production: Quality control

- How you do quality control depends on the goals of the project
- Develop a consistent approach (make a checklist)
  - Missing pages
  - Image quality acceptable (resolution, colour and tone, overall appearance)
  - Compare master to derivatives

Publication & storage

- Publication
  - Make sure the digital item meets your established standards and contains no errors
  - Make sure the metadata correctly describes the item
  - Double-check that you have the right to publish the material
- Storage and backups
  - Store the documents you might need access to in an accessible location
  - Archiving should be done regularly and consistently (automated, if possible)
  - Keep at least two copies, in different locations, if you can
  - Know how long it takes to have your backups restored

Contact

- Donna MacLeod
- E-mail: donna@dmacleod.ca
- Phone (day): 416-393-7105

- Slides and a list of suggested readings are available at http://www.dmacleod.ca/dma-wp/digitization-for-conservators-2016/