Questions from the April 2, 2013 Digital Preservation Webinar "Preservation Planning and an Introduction to PREMIS"

Answers provided by Lisa Gregory

On Metadata

I am a self-taught Archivist (except for a Western Archives Institute course) and have not had formal training in Archives. That is probably why it is so confusing to me. I sort of fell into this position and am taking several "basic" webinars to learn all I can digest. My question today was about Metadata -- it is confusing to me -- I know how to find it/work with it on photos (e.g. tiff or jpeg) but am unsure about text docs. AND I am sure there is more I could/should learn about image files as well! Is there some website that has really BASIC info that I could read/download? The last portion of the webinar hour today helped some - but I would still like to get more. . .

It might be helpful to start thinking of metadata in terms of the standard fields requested in metadata schemas, and not so much based on the format of the object you're describing. If you know how to describe images, many of the same fields will apply to documents, or audio/video.

I'm not sure where you are in your exploration of metadata, but I've listed a few suggestions for introductory information:

http://blogs.scientificamerican.com/information-culture/2012/12/17/what-is-metadata-a-christmas-themed-exploration/

http://wiki.dublincore.org/index.php/User_Guide

What's a data dictionary? An ontology?

A data dictionary is a place (sometimes a document) where the metadata elements used in a system are explained and defined.

An ontology is "a formal specification of a shared conceptualization" (http://www-ksl.stanford.edu/kst/what-is-an-ontology.html). That's pretty high level, so I like to think of ontologies when it comes to metadata as a structured set of metadata elements that includes expressions of relationships between those elements.

When we scan an image, the camera captures the metadata, such as exposure. I know that for digital preservation, there is some additional metadata required. Where does this come from? Is it typically entered by the person doing the scanning, would it be done in Indexing, or is there another recommended protocol?

A good deal of preservation metadata can be automatically extracted as you described. However yes, the preservation metadata that can't be automatically extracted is often entered by the person doing the capturing.

Can you recommend examples of metadata dictionaries?
I'm not sure if you mean local dictionaries created by institutions for their own internal use or widely adopted standards. As to the latter, there are a ton out there – many are subject-specific. Other than the ones mentioned in the webinar (PREMIS, Dublin Core), here's a list of some additional metadata standards – all are googleable:

EAD, VRA Core, Darwin Core, TEI, NISO MIX, METS, DDI

On PREMIS

In PREMIS, what's the difference between a file and a bitstream? DSpace seems to make no practical distinction, so I'm not used to these two things being talked about as different.
A file is made up of a bitstream + data headers that stipulate the file's structure and provide additional metadata. PREMIS describes a bitstream this way: "A bitstream cannot be transformed into a standalone file without the addition of file structure (headers, etc.) and/or reformatting the bitstream to comply with some particular file format." P. 7

What's the relationship between PREMIS and MIX and METS?
These three are all metadata schemas, with some overlap between. As described in the webinar, PREMIS provides semantic units which can hold values related to digital preservation metadata. MIX is used for description of digital images – it's as robust as PREMIS but include many metadata bits automatically captured by digital cameras and/or unique to image files (things like tile width, granular color profile information, reference points, etc.).

METS is a way of encoding metadata, whether it's technical, administrative or descriptive. What that means is it provides a structural way, using XML, to organize metadata so that it's machine readable. It can express relationships within a single object's metadata and/or between multiple objects. Jody (webinar co-host) has a great page on METS: http://jodyderidder.com/metadata/mets.html

PREMIS can be used for administrative metadata in METS. So you can choose to record your preservation metadata using PREMIS semantic units, and structure that information in METS (as administrative metadata). http://www.loc.gov/standards/premis/premis-mets.html

The same is true of MIX. If you're using MIX to describe digital images, you can structure that metadata in METS' administrative, descriptive, and/or technical metadata.

You may also want to read through this article, which gives a use case mentioning all three of these schemas: http://www.dlib.org/dlib/march08/pearce/03pearce.html

Software/Tool Recommendations

What's your recommendation for OCR software?

The industry heavyweight is ABBYY FineReader. We use it frequently here, and it works well enough. Adobe Acrobat will OCR .pdf files as well.

I've never used Tesseract, but it's an open source alternative if you have the infrastructure to install and run it. http://code.google.com/p/tesseract-ocr/

What's your recommendation for conversion to PDF/A-1a (apart from Adobe Acrobat or ABBY Fine Reader)
We've only ever used Adobe Acrobat. However I'm also aware of advocates for PDFtk, which I understand is a powerful open source command line tool.
http://www.pdflabs.com/tools/pdftk-the-pdf-toolkit/

Is there a list of tools divided up by each step of the process?
Unfortunately, I'm not sure about the part of the presentation to which this question applies. I'll point to the general digital preservation tool showcase at the Library of Congress – I hope that helps. http://www.digitalpreservation.gov/tools/

Other digital preservation tools we use on a regular basis include
TeraCopy: http://codesector.com/teracopy
Hashmyfiles: http://www.nirsoft.net/utils/hash_my_files.html
DROID: http://digital-preservation.github.io/droid/
XENA: http://xena.sourceforge.net/
Jhove: http://jhove.sourceforge.net/
ReNamer: http://www.den4b.com/?x=products&product=renamer
Bulk Rename Utility: http://www.bulkrenameutility.co.uk/Main_Intro.php

Digital Preservation
What are the first steps we should take? Can you give examples for small and one-person libraries?
First and foremost, audit where your content is and then get it redundantly stored in disparate locations. I'd suggest checking out the 2012 ASERL Digital Preservation Webinars http://www.aserl.org/archive/, which outline first steps.

How can we access the OAIS model?
The OAIS model is found in all of its glory within CCSDS 650.0-M-2. However if you're new to the model I recommend going to Brian Lavoie's classic Introductory Guide.
http://www.dpconline.org/component/docman/doc_download/91-introduction-to-oais

Can a repository take the first steps toward preserving digital files (converted to open formats) even though it hasn't yet set up any kind of a digital asset management system? This might be a situation in which what you have to work with in terms of infrastructure is digital storage space set up for 2 copies not collocated, with security in terms of differential levels of user access, as well as some system for recording /monitoring (?) bit-level fixity. If so, how do you associate metadata with the digital files in the absence of an infrastructure

for doing this? Including a spreadsheet in the same directory as the files is all I could come up with.
Yes! The situation you've described is a perfectly acceptable workflow. Structured data stored in a spreadsheet, saved in an open format (.csv, for example) and stored with the files can be enough to later act upon. Many content management systems or other tools that needs machine-actionable data to start building relationships between files and metadata will accept open spreadsheet formats.

I'd always recommend a structured text file, spreadsheet, or database over an unstructured text file. Will save you a lot of trouble later.

When someone hands you a jump drive with files on it, what are your first steps?
This is a hard question, because it depends on how much I know about it already and also what I'm intending to do with it. For example's sake, let's say I've been told that I must preserve the contents of the drive, that I don't know what's on it, and that I need to inventory it but don't have to make the contents usable/accessible at this point.

I'll start off with the assumption that you have a computer that has a USB compatible port. If you're a purist – it would be a clean scrubbed computer which is *only* used for digital forensics software and activities. Few folks work under that condition, though.

In general, my first steps include virus scanning, creating an image of the drive, and triaging the files included.

In creating a disk image, you're using software to essentially lift the entire drive, as is, into a single discreet unit off of the media. If you google "create a disk image" there are a number of software utilities that accomplish this for you. You may also want to check out BagIt. Here's a detailed user guide:
http://www.records.ncdcr.gov/erecords/Using_BagIt_ver2%202_generic_final_20110414.pdf

Once you've created a disk image (and assuming you don't have a digital forensics lab at your disposal) I personally would recommend creating an alternative copy of the disk to work with. Remember, you're saving the first image untouched as your preservation copy. Think of this second copy as the one you're going to autopsy.

You can use file validation tools like DROID or JHOVE (see above) on the files on the jump drive if you need to extract metadata and also to find out if the files are what they say they are. You can start assigning more granular metadata at this point, as you would if you were scanning a bunch of photographs, if that's what the project calls for.

Those are the first steps, in general, you can take. What happens afterward depends on preservation workflows at your institution, and what you find on the disk.

Do you have suggested sources / strategies for funding digital preservation activities?

This is the $64,000 question. Endemic to digital preservation is longevity – which means ideally funding should be built into your program, and not project-based. For many, that just isn't a reality at the beginning. If you haven't already looked at "Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information" I'd suggest starting there. http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf

I see funding acquisition as hand in hand with being able to articulate your needs in a way administrators will understand. In my experience, the "someday it'll all be gone if we don't act" works about on par with insurance sales - it's hard to make that argument compelling with people who don't work with preservation or cultural heritage on a regular basis.

What I've found instead is to start by trying to establish your own credibility as an information professional. This means getting yourself at the table when these sorts of decisions are being discussed, whether it's with IT folks or your institution's administration.

Make a formal business case backed up by the work that's been accomplished in the digital preservation field. Hopefully some of that plan comes out of the suggestions I made in the webinar. You're establishing that you're not making this stuff up, you've thought it out, and you have a direction you're confident in.

You're also going to have to find a way to put things into cost-benefit terms. As much as we've tried to avoid it, that's what it ends up boiling down to in many cases, at least on some level.

As far as specific revenue streams, we've been fortunate enough to receive funding from IMLS (http://www.imls.gov/applicants/available_grants.aspx) for creating a digital preservation tool through a Sparks! Ignition grant.

Also, we have received funding through the LSTA program (http://www.imls.gov/programs/, funded by IMLS) for some digital preservation research and workflow implementation. Again, this isn't for ongoing infrastructure costs but for efforts to move our own digital preservation agenda forward.