

Federal Agencies Technical Metadata Subgroup

Participation from both the AV and Still Image Working Groups

Notes from the meeting held at the National Archives (Archives 1, main building downtown), September 1, 2009.

Abstract: Nine agencies represented, discussion of scope for the subgroup; glimpses of relevant work at various agencies; about metadata "leveling" and the need for specifications for vendors.

Opening comments

The meeting was chaired by Kate Murray from the National Archives and Records Administration (NARA). She reported on recent digital-content developments in the agency that reflected a desire to standardize what is done with technical metadata. NARA's interest not only concerns the reformatting of historical collections but also accommodating the types of born digital content that are being added to the Electronic Records Archive (ERA). Why is this important? Murray reported that one NARA document states that technical metadata is "necessary to ensure the continued usability of an object, or to reconstruct the object if it is damaged." Meanwhile, she also noted that concepts and practices regarding technical metadata vary between agencies and this may prevent the development of a single detailed guideline due to variety of agency missions.

The discussion turned to the definitions currently posted in the glossary on the Federal Agencies Web site (<http://www.digitizationguidelines.gov/glossary.php>; readers should note that the definitions are edited from time to time and may have changed subsequent to September 2009). Among other terms, the glossary defines *technical*, *source*, and *process* metadata. Although semantics and usage vary from one agency to another--not everyone uses these terms in the same way--all parties agree that two critical types of metadata document the circumstances and processes used to create the file (in a reformatting activity) and the various "facts" about the file's technical characteristics. One attendee noted that this conceptualization mapped to three of the metadata subcategories in the Metadata Encoding and Transmission Standard (METS) standard: (1) *techMD* that documents the attributes of the final digital item, (2) *digiProvMD* that documents the attributes of the conversion process, and (3) *sourceMD* that documents the attributes of the source item. In the audio-visual world, these distinctions are echoed in two emerging Audio Engineering Society (AES) specifications: X09B covers (1) and (3) and X098C covers (2).

Glimpses of work at various agencies

Representatives from the Library of Congress and other agencies provided snapshots of their efforts to identify the sets of technical metadata that ought to be captured and maintained. Regarding the scope, one person said, "We take a holistic view, all metadata is pertinent."

One Library of Congress representative highlighted the importance of working with the Society of Motion Picture and Television Engineers (SMPTE): "We should engage their process; we ought to engage the commercial sector that produces the content that we receive (or manufactures the devices that produce the content we create or receive)." This person went on to

point out that the SMPTE MXF format is important in his agency's work and that we ought to look at what may be placed in that format's "header" slots. Another attendee called attention to the extensive SMPTE metadata dictionary. "There are classes where an individual or organization can register their own elements, with class 13 for public data and class 14 for private. This might be a good way to add elements to the dictionary if we wish to extend it."

Another speaker referred to the continued development of practices pertaining to the still images produced when scanning documents, books, or photographs. This had to do with working with the commercial standards that are already there, e.g., EXIF. "Shall we try to get more scanner manufacturers to adopt the EXIF," this person asked, "thus embedding this valuable data in our from-scanner images?" One attendee spoke of his group's creation of new digital photographs (e.g., at agency events) and how they used Adobe software to maintain the native EXIF data from the camera and then added some additional metadata "on top."

An attendee from the Library of Congress preservation unit described her multi-spectral camera, used for scientific/analysis photography, to support the conservation of physical collections. Their goal is to create a reference collection of scientific images and data that can be referred and used by others over time, and they wish to capture "core data" rather than just "more data." Their camera generates EXIF and XMP sidecar data, she said, but to some degree the images and metadata it produces is proprietary format. However, the manufacturer is keen to see some standardization.

EXIF and XMP metadata came up in other remarks. One NARA staff member noted that her agency uses the ANSI Z39.87 data dictionary that defines the technical data for images (this is the basis for the XML schema called *MIX*; see <http://www.loc.gov/standards/mix/>). "The data about the source object," she said, "goes into Telescope, a data management system that some in my agency use. But we may not describe or list all of the facts: for images, we just say this image is compliant with this or that specification. In addition, we use XMP in a limited way and this data goes into Telescope.."

A representative from NASA's Kennedy Space Center noted that many of the formats they encounter or create are new or emerging. They find that EXIF and XMP are helpful for still images but pointed out that NASA's moving imagery is shot in several different ways. "We have unique types of high speed cameras," the speaker said, "and these mean there are special requirements. We gather what we can but most is kept external to the files."

Regarding the extent of data to be collected, an attendee from the Smithsonian Institution said, "We often ask what is the minimum technical metadata needed? We don't want to try for too much." A NARA representative who uses the SAMMA device to reformat video noted that this system generates technical metadata "not only at the file level but also on the frame-by-frame level. We ask: Do we keep all that metadata? There are 30-odd fields of data per second." This was followed by some added comments on NARA's use of the DPX format when scanning film and the agency's desire to find a best practice for technical metadata in that activity.

The meeting attendees included a representative from the California Digital Library (CDL; a state government entity) who is active in the development of the JHOVE tool

(<https://confluence.ucop.edu/display/JHOVE2Info/Home>). The tool examines and extracts information about/from a file. In the work of the CDL, he said, "We rely on JHOVE to examine and extract what is submitted. We are managing the resulting metadata in an external database but we feel that this is not ultimately sustainable. We ask: what is the metadata subset we should maintain externally? And: what will be needed by repository mgrs to support repos management? Any data beyond that set can be re-extracted in the future."

Metadata *levels* and relationships with vendors

The group talked about two contextual factors. One was referred to as *levels*: some metadata is true for a whole collection or scanning project, other metadata is special to each item or even each file. Depending on how metadata is handled, however, there is a risk that metadata "at the project level" may not be visible when someone looks at a specific item. The second contextual factor--an echo of the Smithsonian comment earlier--concerned automation of metadata capture. One agency representative said, "We hope to build a lot of tools." Another said, "We have to automate 99 percent of the work; we plan to develop a broad umbrella, the same in every workflow."

One person noted that the conversation thus far concerned in-house systems. "But some of us go to outsourcing," he said, "and we need vendor specs. Vendors are often not familiar with metadata." A similar need for a specification was voiced by a NARA staff member, noting that their agency exchanges or receives files between or from agencies. Another attendee noted that the need for specifications was one important reason for launching the Federal Agencies initiative: "Among other things, many of us have to deal with vendors and present a unified front. Service bureaus can deal with complex metadata, but they rather not have 15 or 20 different requirements from 15 or 20 customers." Some at the meeting noted that it is not possible to persuade vendors to adopt or support specification, especially in the case of system manufacturers, where large-scale commercial customers carry more weight.

The Working Group's consultant is a former employee of a service bureau, who now consults with a number of universities and other customers of such service providers. "Our clients are saying the same things," he reported, "they are concerned with technical metadata as compared to source metadata as compared to process metadata."

What next for this group?

Kate Murray said that there was probably value in having sub-subgroups but that it was not yet clear how best to form these. We could consider "by equipment," "by format," "concerning terminology," "for tools," or in consideration of factors like metadata "leveling." This will be taken up at the next meeting. Meanwhile two voices from the floor noted that the discussion thus far had not highlighted the DNG still image format and the PBCore video metadata format. These two entities may be of interest to this subgroup.