

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), March 8, 2010.

Abstract: Eight agencies represented, discussion of the terminology for technical metadata categories; use cases (why have technical metadata, what are the purposes for capturing and keeping it?); metadata "leveling" (how different metadata elements may apply at the different levels; video technical metadata.

Terminology

The meeting was chaired by Kate Murray from the National Archives and Records Administration (NARA). The lead-off topic was terminology, referring to the definitions at the Federal Agencies Web site (<http://www.digitizationguidelines.gov/glossary.php>). There was considerable back and forth about the terms *technical*, *source*, and *process* metadata. In the end, a number of voices counseled that *technical metadata* be a generic term. It would be the umbrella over (at least) three other entities for which we need the perfect terms. This trio consists of what might be called *file characteristics metadata*, i.e., information about the digital entity in hand, *source metadata*, information about the source entity (could be analog or digital), and *process metadata*, i.e., information about the process used for production.

There was a call for volunteers to join a committee on semantics/terminology and to propose adjustment to the current glossary. Nine individuals volunteered.

Use cases

This discussion concerned the purposes for technical metadata. What might it be used for? What activities ought it support? Although not full-blown uses cases with actors and actions, attendees highlighted some key activity zones in which technical metadata would play a role:

- Quality control/quality assurance
 - One NARA representative reported that their agency is exploring the use of automated batch production tools such as Interra System's Baton, Tektronix's Cerify, and Dobbin from AudioCube/Quadriga. We need to compare information provided by that type of system; this means sorting out the metadata elements. We want to know, "does our new object pass muster?" That info will eventually be passed onto the NARA ERA digital repository.
- Management of data during its life cycle
 - This can include actions to transform and encode; an agency will wish be sure that essential technical characteristics have been maintained.
- Digital preservation assessment within a repository
- Source information, i.e., metadata about the entity you started with.
 - Someone noted that the TRAC auditing of trusted repositories specification (http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf) requires this, metadata about processes applied, sometimes called provenance metadata.
- Fitness for use, suitability for use

- We want to be able to say that this file succeeds in filling this request, e.g., for a still image, we want to be sure that we have the DPI needed to meet a given requirement?
- Identifiers

In the discussion, a variety of ideas and agency experiences were mentioned, including these:

- Mention of JHOVE and tools built "on top of" JHOVE, for quality management and statistical process control. These tools extract or generate technical metadata, which is needed over the long term to support sustainability.
- Regarding ongoing data management, one speaker said that an agency will want to document encoding and compression. In a repository, you might decide that you will no longer support this form of encoding. This kind of assessment of tracking metadata could be expanded to cover what is going in the repository.
- If you do transform data, another person said, we may want to know what a given item was originally for legal or forensic reasons; you will need provenance or process metadata to retrace your steps.
- One attendee said that this discussion highlighted the possibility that there is another class of technical metadata, what might be called *data management metadata*. But this might not be a category of its own. Like Premis, just pluck it out from other subcategories.
- Others talked about how we--collectively--might use technical metadata for obsolescence monitoring. Could we track and compare statistical data across organizations, i.e., monitor technical/formatting trends and see where problems arise. This would be the empirical monitoring of formats. "Remember," one person said, "the OAIS model calls for tracking information in the general digital-content ecosystem."
- Following up, another attendee said he had been thinking about the role for a "technical metadata aggregator." In the consumer domain, there are apps that report back on tech metadata. There ought to be simple ways to this. Look at <http://www.codecsdb.com/>. It is a simple database reporting on commonly used codecs by the general public. The reporting mechanism is through a tool called Video Inspector which will report on technical metadata of files.
- Someone asked, would you ever want to identify content produced by a particular make or model of a conversion device, in case it was later discovered that this device had a defect? Would you want to identify and remake those files?
- Another attendee expressed concern about recommendations to produce extensive, detailed metadata. This can be labor intensive; will there be tools to make it easy?

There was a call for volunteers to join a committee on use cases. Six individuals volunteered.

Metadata "leveling"

Steve Puglia from NARA led the discussion of this topic, referring to a draft document from his agency that outlines a metadata hierarchy. Different types of metadata apply at different levels. One aspect pertains to the tree-like structures with their respective levels, as seen in archival collections when digitized: collections or projects, representations or items, digital objects and digital copies, and data or file format. At the low "file format" level, for example, there is metadata about bitstream encoding: RGB vs. YCrCb. A second aspect pertains to the fact that

some metadata is for systems (machine readable, coded, machine actionable) while some is for people (human readable, expressed in a manner that is understandable by people). Meanwhile, some metadata is external to the file, some embedded in files. The metadata for people or data management, for example, may work better outside the file, for easier access.

In the discussion, a variety of ideas were expressed, including these:

- As we proceed, we might head for a recommendation about where to "put" certain types of metadata in order to meet the needs of certain use cases.
- Perhaps there is what might be called "class" metadata, i.e., a block of information that applies to a whole batch or class of objects.

There was a call for volunteers to join a committee on leveling. Nine individuals volunteered.

Video technical metadata

Kate Murray introduced this topic, noting that NARA has been puzzling over what might constitute a minimal technical metadata set for video. Their discussions focused on "file properties," including things not dependent on file format and encoding, e.g., aspect ratio, line count, and picture size. What are the technical elements that should be tracked across all video objects?

In the discussion, a variety of ideas were expressed, including these:

- For raster images, we have ANSI/NISO Z39.87 (Data Dictionary - Technical Metadata for Digital Still Images). There are also the emerging Audio Engineering Society specifications: X098B and X098C. We need something like these for video.
- The drop down menus in MAVIS (an audiovisual collection-management application) offers a nice set of elements. But we must remember that these are somewhat volatile. This application is used by the Library of Congress and all of the term lists tend to be changed on a regular basis.
- This topic is made a little more challenging since we do not have solid inter-agency consensus on preferred target file formats when reformatting video. This makes it especially difficult to say what to "put into a header."

At the meeting's end, the group found it difficult to be clear on next steps and, for now, action was deferred.