Identifiers: Types and Characteristics

By the Federal Agencies Audio-Visual Working Group
http://www.digitizationguidelines.gov/audio-visual/
November 21, 2011, links updated April 19, 2012

What is this document?

This is one of four documents pertaining to the embedding of metadata in digital audio files prepared by the Federal Agencies Audio-Visual Working Group. The three companion documents are:

- *Guideline for Federal Agency Use of Broadcast WAVE Files* (Version 2.0, April 2012)[1]
- *Introductory Discussion for the Proposed Federal Agencies Guideline* (updated April 2012)[2]
- *Consultant's report on embedding options in digital audio files* (June 2009)[3]

The content of this document was previously disseminated as an appendix to the 2009 version of *Embedding Metadata in Digital Audio Files: Introductory Discussion for the Federal Agencies Guideline*.[4]

Introduction

Different identifiers coexist and operate at varying levels of granularity and actionability in the digital content landscape. Identifiers from every part of this landscape are employed by federal agencies and other archives with little consistency, thus inhibiting the interoperability of our digital content. Some agencies use identifiers that link digital files to a metadata record in the system an agency uses to provide public access, while others use identifiers that connect solely to a local database, not accessible to the public. (Some content within an agency will be associated with both types of identifiers.) Some identifiers are attributes of the metadata while others are attributes of the digital content itself.

Identifier as attribute of a metadata record: LC example

Library of Congress Control Numbers (LCCNs) are the identifiers for the Library's bibliographic records (attribute of the metadata). They are not only "in" the record but the bibliographic records can be retrieved by the public using the PermaLink, a method for including the LCCN in an actionable URL. When the following URL is keyed into a browser, for example, the Permalink *http://lccn.loc.gov/mp76000002* returns a presentation of the bibliographic record for the motion picture Theodore Roosevelt's Return to New York, 1910. By adding *../marcxml*, *../mods*, or *../dc* to the URL, the system's action returns the same data in the form of a MARCXML record, a MODS record, or in the Dublin Core XML format. Thus, if an LCCN is embedded in digital content, it permits a user to pull up the bibliographic record.

[1] http://www.digitizationguidelines.gov/audio-visual/documents/Embed_Guideline_201204123.pdf
[2] http://www.digitizationguidelines.gov/audio-visual/documents/Embed_Intro_20120423.pdf
[3] http://www.digitizationguidelines.gov/audio-visual/documents/AVPS_Audio_Metadata_Overview_090612.pdf
[4] http://www.digitizationguidelines.gov/audio-visual/documents/Embed_Intro_090915.pdf

Identifier as attribute of digital content itself: LC example

The bibliographic record for Theodore Roosevelt's Return to New York, 1910 also contains an identifier for the digital content (attribute of the content). Scroll down in the display of the record to the MARC $856 field, and you will come to an additional actionable link in the form of a Handle identifier: *http://hdl.loc.gov/loc.mbrsmi/trmp.4170*. (Strictly speaking, this is a URL that incorporates the Handle.) Evoking this Handle brings back another presentation of the bibliographic record, this one in American Memory. That bib-record presentation includes further links based on the logical path- and file-name for the three files (at different resolutions and in different formats) that contain the movie, e.g., *http://memory.loc.gov/mbrs/trmp/4170s1.mpg*.

Examples in other agencies

Similar identifiers are in play in other agencies. At the National Archives and Records Administration, for instance, some recent practice identifies audio, motion picture, and video content by means of a string based on the archival Record Group, Series, and Item names and numbers. (At this writing, planning at NARA has begun for a new approach for digitized media items making use of non-mnemonic unique identifiers. All identifiers would be tracked in a related database and selected identifiers would also be embedded in the files.) Meanwhile, at least one Federal Agencies Working Group participant assigns the widely used international ISRC identifier to some of its digital audio items.

Digital packages, digital package parts, and digital files

Identifiers are associated with—identify—entities that exist at different levels of abstraction. The Working Group had adopted terms that refer to these levels, with definitions provided in the initiative’s online glossary:

- Digital package
  - http://www.digitizationguidelines.gov/term.php?term=digitalpackage
- Digital package part
  - http://www.digitizationguidelines.gov/term.php?term=digitalpackagepart
- Digital file
  - http://www.digitizationguidelines.gov/term.php?term=digitalfile

How do these named entities relate to identifiers? The following paragraphs describe some archetypes; in actual practice, there are often nuanced shadings from one type to another.

Digital-package-level identifier. In a library setting, packages generally correlate to what are called manifestations in the parlance of the Functional Requirements for Bibliographic Records (FRBR), and thus to identifiers like the LCCN (Library of Congress Control Number), ISBN, ISSN, ISRC, ISTC, ISWC, Superintendent of Documents number, etc.. In an archive, digital packages generally correlate to an item in, say, an EAD (Encoded Archival Description) finding aid, and to identifiers that reference the name for the collection, series, and item. Very often, this identifier will have been established prior to digitizing and will therefore be available for embedding by the digitization team.

Digital-package-part identifier. This identifier has to do with structural metadata; it is intended to answer the "what sub-part am I in relationship to other parts of the overall package?" Here's a hypothetical example: a Library of Congress book-scanning project uses the string 00220008 to identify scan-exposure 22, representing the book page with the number 8 printed on it. This identifier is not typically part of a bibliographic record or finding aid but (in this example) is more or less self-explanatory within the object set. The identifier is still an abstraction; it is not "at the file level." Multiple image files (e.g., archival master, derivative viewing file, and thumbnail) may all represent the same object sub-part, e.g., page 8. Here's a hypothetical recorded sound example: *gfc1972.cas14.sd2*. This identifier documents the fact that this sub-part is side *2* of original audiocassette *14*, from the 1972 portion of the Galax Fiddlers Convention Collection. Another subpart in the set would be identified as *gfc1972.cas14.sd1*, reproducing cassette side *1*. Generally speaking, the object-sub-part identifier is determined during the digitization process (including object-preparation stages) and the information is often available for embedding at production time.

File identifier. This is an identifier assigned to the file itself, i.e., a kind of license tag that may or may not refer directly to descriptive information. For some digitizing projects, this is assigned by the Digital Asset Management system at the time of digitization. If the file identifier does not directly link to descriptive information, the systems in use generally include a database or other look-up tool (some may be "manual") that can be used to connect back to descriptive information.

Identifiers that operate at the file level are found in METS documents. The METS Primer and Reference Manual states, "A <file> element may contain one or more <FLocat> elements which provide pointers to a content file and/or a <FContent> element which wraps an encoded version of the file." (p. 29) The FLocat elements has a LOCTYPE, for which the following types have been defined, some of which are persistent identifiers:

ARK: Archival Resource Key
URN: Uniform Resource Name
URL: Uniform Resource Locator
PURL: Persistent URL
HANDLE: a CNRI Handle
DOI: A Digital Object Identifier

The following example is provided at the METS Web site, with a FLocat of the LOCTYPE URL.

```
<mets:file ID="FID1" MIMETYPE="image/tiff" SEQ="1" CREATED="1999-06-
17T00:00:00" ADMID="ADM1A" GROUPID="GID1">
<mets:FLocat xlink:href="http://sunsite.berkeley.edu/masters/bkm00002773a.tif"
LOCTYPE="URL"/>
</mets:file>
```

File identifiers may be pre-assigned or assigned at production time. They may be assigned by a production-management, digital-asset-management, or repository-ingestion system, or by manual means.

Multiple identifiers and the desirability of typing

As suggested by the preceding, there may be more than one identifier to be embedded in a file or object-packaging metadata. A comparative example exists for digital photography, in the specification for International Press Telecommunications Council (IPTC) photo metadata.[5] The specification provides element definitions for the following types:

- Digital Image GUID (p. 32). Globally unique identifier for this digital image. It is created and applied by the creator of the digital image at the time of its creation . This value shall not be changed after that time.
- Image Registry Entry (p. 35). Typically an id from a registry is negotiated and applied after the creation of the digital image. Both a Registry Item Id and a Registry Organization Id [are required] to record any registration of this digital image with a registry.
- Related term: Item Id {registry entry detail} (p. 48). A unique identifier created by a registry and applied by the creator of the digital image. This value shall not be changed after being applied. This identifier is linked to a corresponding Registry Organization Identifier.
- Related term: Organization Id {registry entry detail} (p. 49). An identifier for the registry which issued the corresponding Registry Image Id.
- Related term: Registry Entry Details {data type} (p. 49). A structured datatype for an entry in a registry, includes the id for the image issued by the registry and the registry's id.
- Source Inventory Number {Artwork or Object detail} (p. 44). The inventory number issued by the organization or body holding and registering the artwork or object in the image.

Since there is a good chance that a given object or file will be associated with multiple identifiers, there will be considerable value in having a metadata encoding that allows for repeating elements and/or attributes, e.g., *identifier_type*, *identifier_value*. This approach would resemble that described for METS FLocat and LOCTYPE, above.

Meanwhile, another perspective on multiple identifiers is provided by the DLF wiki Best Practices for OAI Data Provider Implementations and Shareable Metadata.[6] The following recommendation pertains to circumstances in which multiple versions of a digital object may exist, looking from external Dublin Core metadata toward the digital content:

> [I]n the case of digital objects, if the identifiers resolve to multiple versions of the resource, it is important to identify a single primary identifier that a service provider can label or use as the primary link to the resource. For example, only one <dc:identifier> element should be included with an actionable identifier (i.e. a URL). Additional <dc:identifier> elements might be included with a local identifier if not actionable (i.e. an end-user cannot click on the identifier to arrive at the resource).

Identifier characteristics

The many categories and types of identifiers possess varying and often overlapping characteristics.

[5] http://iptc.cms.apa.at/cms/site/index.html?channel=CH0099 Consulted April 19, 2012.

[6] http://webservices.itcs.umich.edu/mediawiki/oaibp/index.php/IdentifyingTheResource. Consulted April 19, 2012; page marked as "last modified" in 2007.

Actionable. Roughly speaking, this has to do with the provision of an automated response at retrieval time, sometimes called resolution: can you plug the identifier into a query box or URL slot, and have a machine retrieve and send you back information or some other response? Reponses may include descriptive information intended for a human user, a machine-readable version of that description, or to a file of some kind. At the Library of Congress, Permalinks and Handles are actionable identifiers; see the examples provided earlier in this document.

Although actionable identifiers are almost always preferable, resolution is potentially expensive when it goes beyond a local system and especially if resolution involves more than redirection. Handles at the Library of Congress work by means of simple redirection and thus they are cheap. The version of the Handle adopted by publishers--the Digital Object Identifier (DOI)[7]--requires a combination of redirection and centrally stored key metadata, and thus is more expensive to support. There may be value for a digitizing unit or organization in having a locally managed resolvable identifier and participating, as appropriate, in broader systems of identifiers that may not be actionable.

Findable. Although actionable identifiers are desirable, non-actionable, "findable" identifiers can be very helpful in their stead. Non-actionable identifiers are sometimes called labels. Often you can use a search engine to look for information associated with an identifier. When a researcher writes about a specific work and uses something like the ISBN as a label, he or she has unambiguously identified the item under discussion even if neither the writer nor the reader try to resolve the ISBN through its naming authority.

Local, global. Roughly speaking, this has to do with how widely an identifier is known and understood. Some identifiers may only be understood by the staff of the digitizing unit or organization, or may only function in a local system for local users, e.g., when content is managed by a local-service Digital Asset Management System. Other identifiers may be global, like the ISBN, DOI, or patent registration numbers. Some of these may be actionable, some not.

Mnemonic, non-mnemonic. This refers to the "human-readability" of an identifier, which may vary by degree (some are easier to decipher than others) and the dependency on insider knowledge. For example, the digital ID for the American Memory sound recording of the fiddle tune Chicken Reel by Red Harmon and Willard Brewer is *AFCTS 4107b2*. The string *AFCTS* stands for American Folklife Center Todd-Sonkin collection, and *4107b2* stands for original disc-recording number 4107, side B, cut 2. In contrast, copyright registration numbers are not mnemonic, save for the prefix that defines the class. For example, *VAu 598-675* (aka *vau000598675*) is from the Visual Materials class. But to know what the number 598-675 references, you must look it up. This example is post-1977 and can be the subject of a search (http://cocatalog.loc.gov/); earlier examples require a visit to the 3x5-inch-card catalog in the Copyright Office.

Comment on production workflow

Common library and archive practices today feature two basic options: (1) identifiers for reproduction files are based on an identification scheme used for or derived from the originals

[7] http://www.doi.org/, consulted April 19, 2012.

(presumably making it easy to hook up the reproductions later with systems used to describe or manage the originals); and (2) new (and possibly non-mnemonic) identifiers are created for each file, and thereafter a database (and/or sidecar or embedded metadata that will end up in a database) is used to connect key identification information to this new identifier. Libraries and archives typically want to use option 1, but asset management systems usually assume 2. An organization must decide which it will use for a particular workflow (or whether it will carry both through the process) and stick with it.