Federal Digitization
Moving to Common Guidelines

The U.S. Federal Agencies Digitization Initiative

http://www.digitizationguidelines.gov/

IASA 40th Annual Conference
With HeNAA, Athens, Greece
September 25, 2009

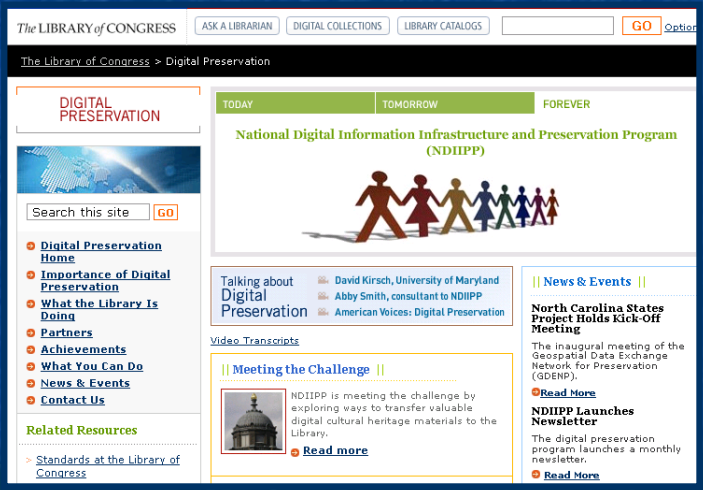
Carl Fleischhauer
cfle@loc.gov
Library of Congress
Washington, DC

TABLE OF CONTENTS:

The Library of Congress

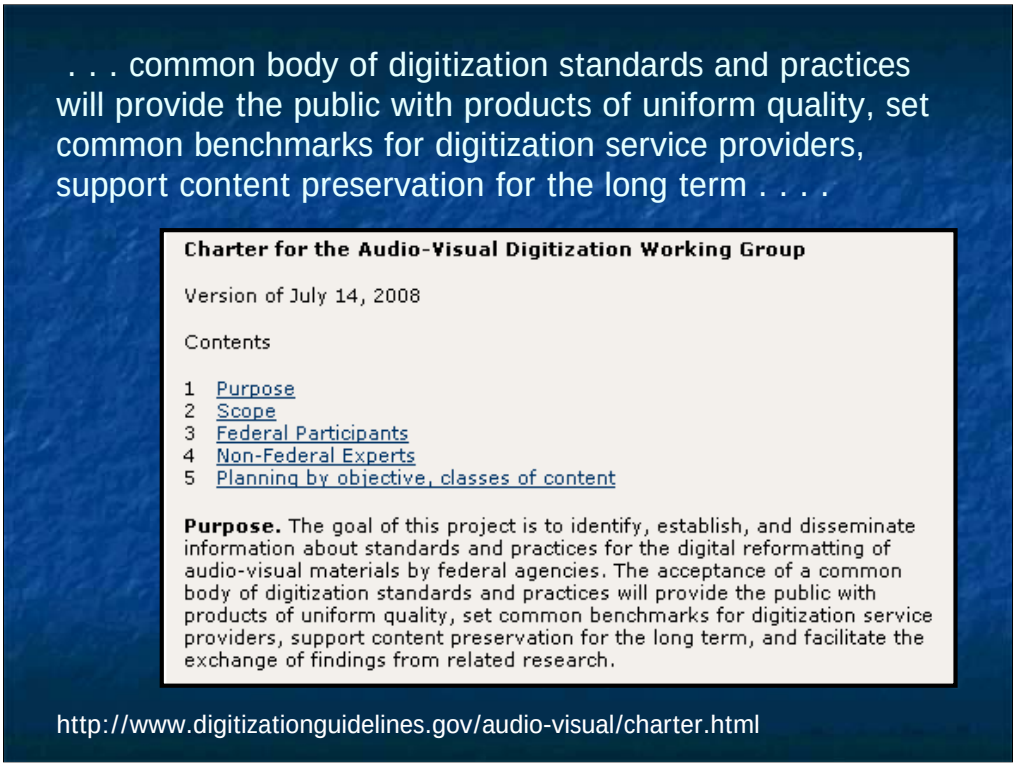
National Digital Information Infrastructure and Preservation Program (NDIIPP)

www.digitalpreservation.gov



The Federal Agencies Digitization Guidelines Initiative was launched in 2007 under the auspices of the National Digital Information Infrastructure and Preservation Program (NDIIPP) at the Library of Congress.

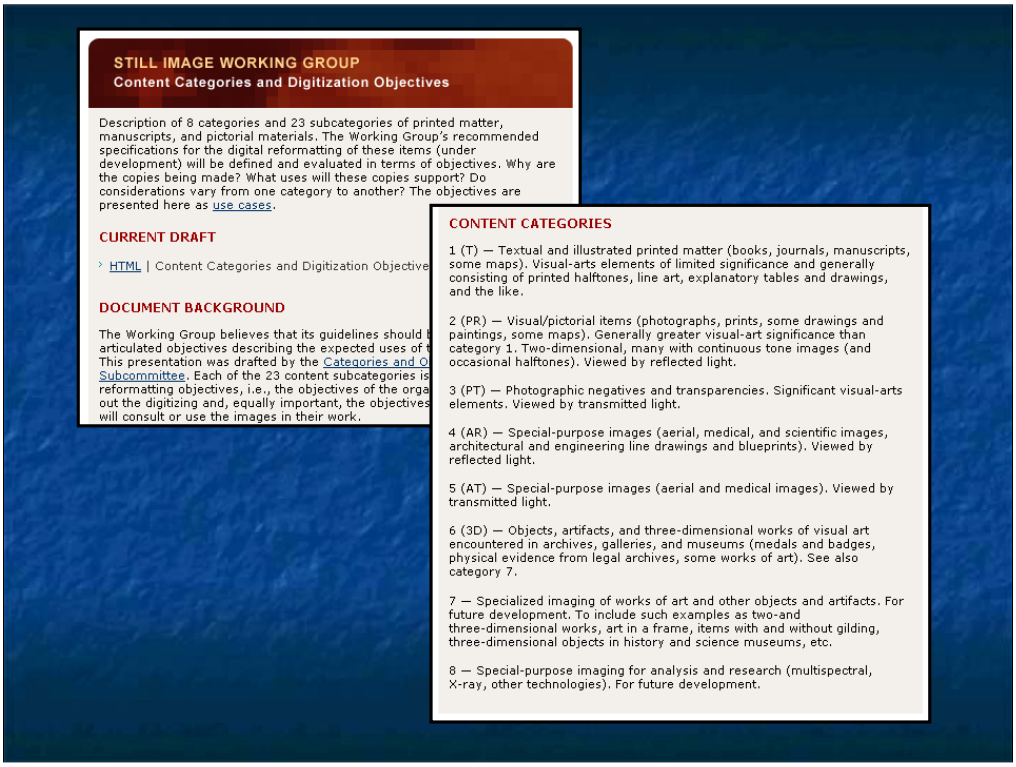
It is a collaborative effort with participation from a number of federal agencies, including the U.S. National Archives, the National Gallery of Art, the Voice of America, the National Library of Medicine, the Smithsonian Institution, and several others.



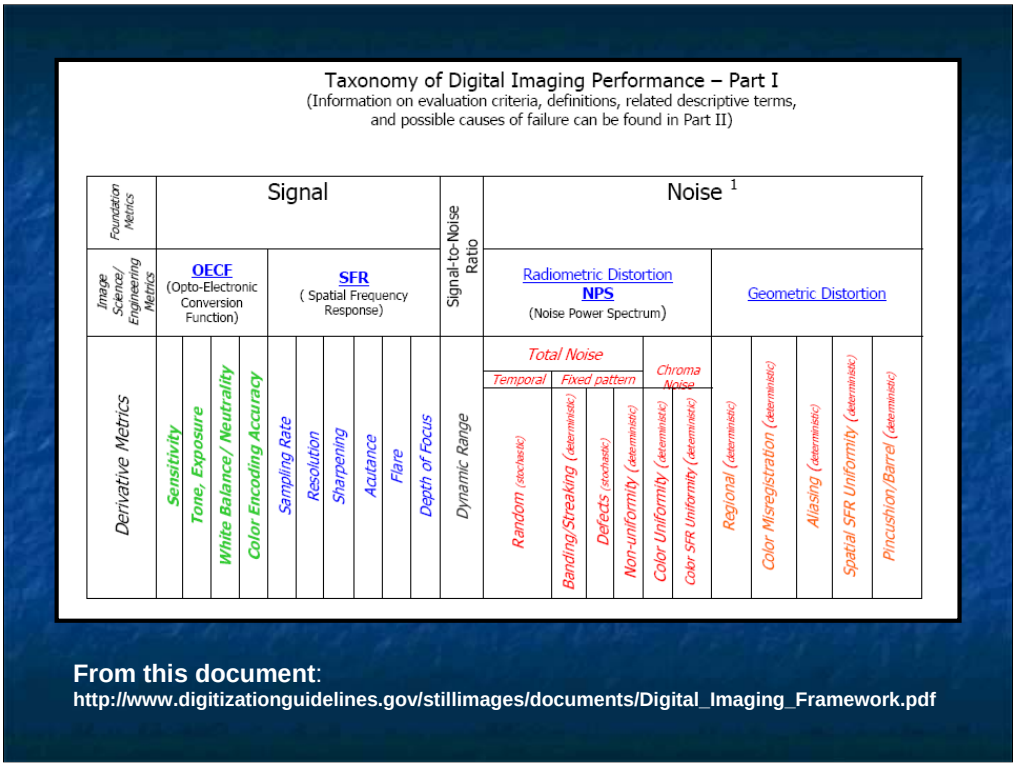
We want to develop guidelines that are comparable from agency to agency, for the sake of uniformity and to make it easier for the vendors who provide equipment and services.

Our main emphasis is *digitization*--the conversion of analog originals into digital form. There are two working groups. Michael Stelmach at the Library leads the still image working group; they look at things like scanning books, photos, and maps.

I lead the Audio-Visual Working Group, focused on sound and video recordings and motion picture film.



Both groups will define their recommended specifications in terms of objectives. For still images, the objectives (and thus the specifications) will vary by category of content. There are 8 major categories for still images.



From this document:
http://www.digitizationguidelines.gov/stillimages/documents/Digital_Imaging_Framework.pdf

What elements will be part of the still image specifications? They will move away from a reliance on "output" measures like pixel density and bit depth. Instead, they want to be attentive to more appropriate ways to specify tonality, spatial resolution, color, uniformity, and noise.

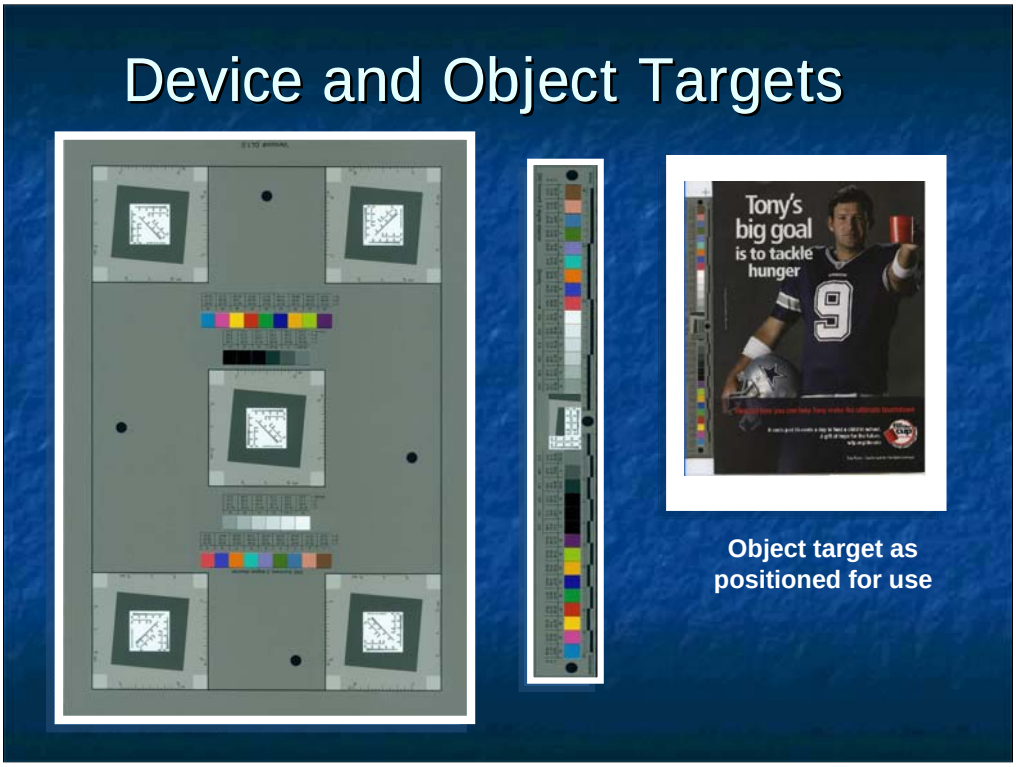
Bijlage 2. Schematisch overzicht toleranties en criteria Preservation Imaging Metamorfoze.

	Artwork,.	Unique librarian and archival material.	Not Unique librarian and archival material. uniek bibliotheek en archief materiaal	
				For access only*
Color space	eciRGBv2	eciRGBv2	eciRGBv2	Not specified
Bitdepth	16/8 Bit	8 bit	8 bit	Not specified
Tonalcapture				255
1. pixel value vak A (reflectiewaarde = 0,89)	249 - 238	249 - 238	249 - 238	
2. Highlight gamma	0,8 – 1,08	0,8 – 1,08	0,8 – 1,08	Not specified
3. pixelvalue vak 1 (reflectiewaarde = 0,70)	228 - 218	228 - 218	228 - 218	255
4. D-max	1,95/2.15	1,50	1,50	Not specified
noise	STD < 4	STD < 4	STD < 4	Not specified
Sn/r	> 1	> 1	> 1	Not specified
Uniform illumination	7	7	7	Not specified
• **< A-2**				
• **< A-1**	11	11	11	Not specified
• **≥A-1**	13	13	13	Not specified
Colorcast	+3 en -3	+3 en -3	+3 en -3	Not specified
Color accuracy (formula cie 1976)	Gemiddelde ΔE < 4 Maximale ΔE < 12	Gemiddelde ΔE < 6 Maximale ΔE < 15	Niet gespecificeerd	Not specified
lppmm	Minimaal 5 lppmm	Minimaal 5	Minimaal 5	Not specified

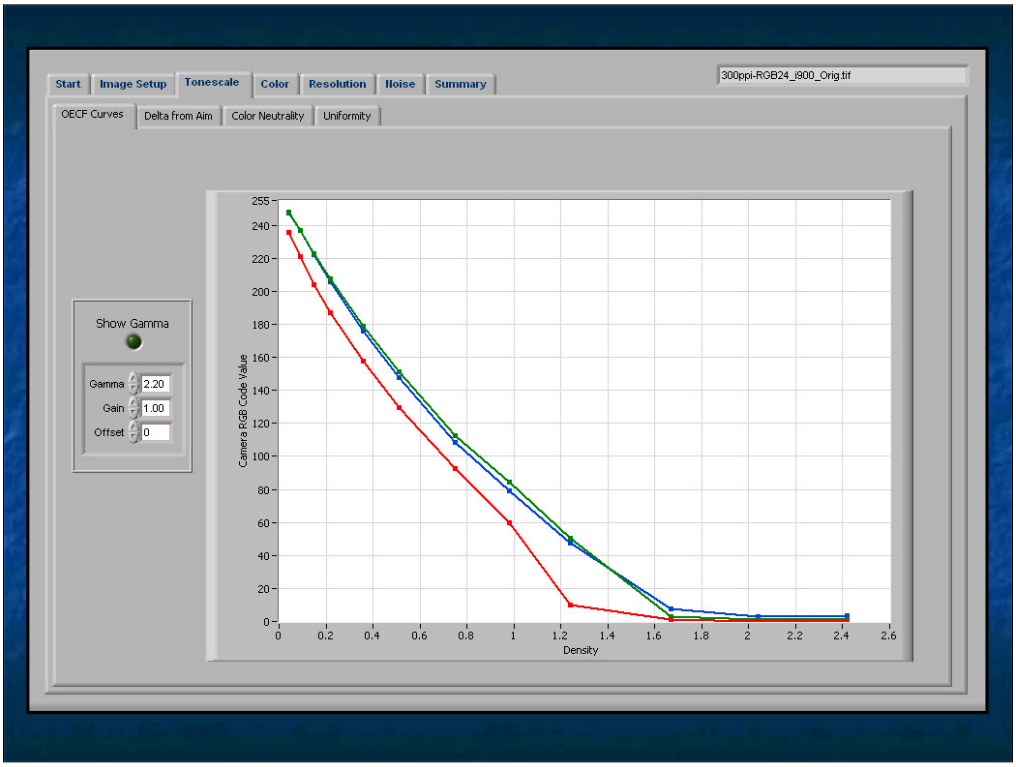
Working document from the National Library of the Netherlands.

Three columns, three categories. Specifications in the various rows.

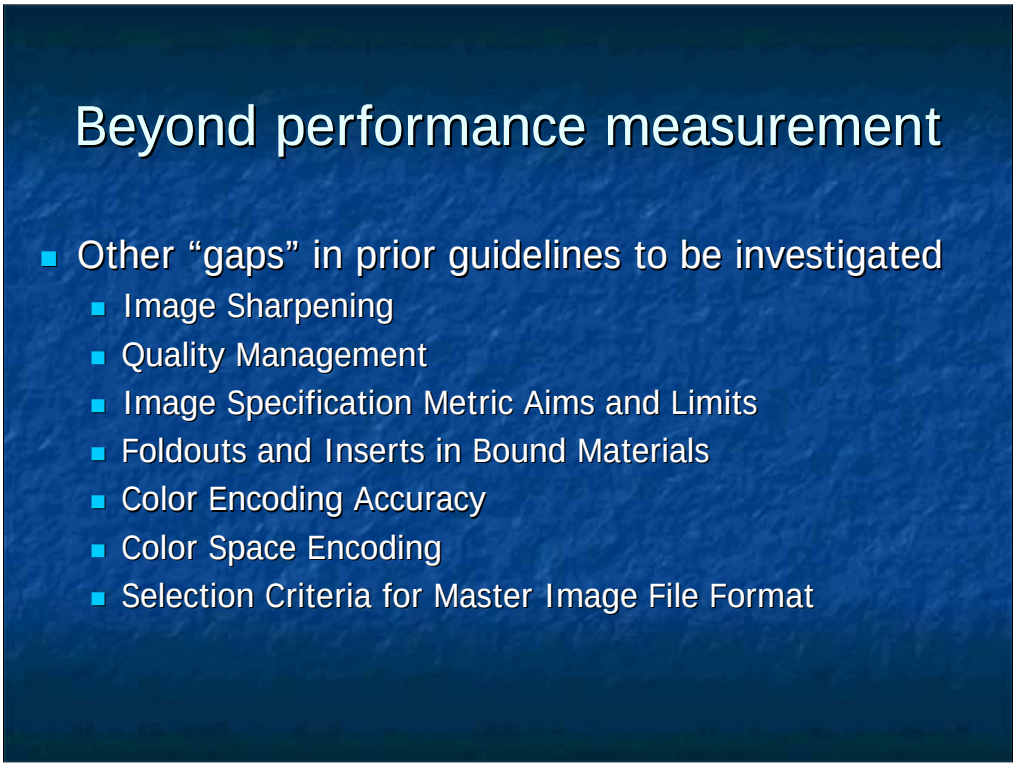
The Federal Agencies still image specifications are not yet ready, but we see that others are thinking in the same way. Here is a page from the Metamorfoze project at the national library of the Netherlands: the columns are categories, the rows are the specifications, which vary by category.



Meanwhile, how will you know if your equipment and the work it produces conform to the recommendations? The still images group is developing tools, including a pair of targets--both about ten inches long--and supporting software.



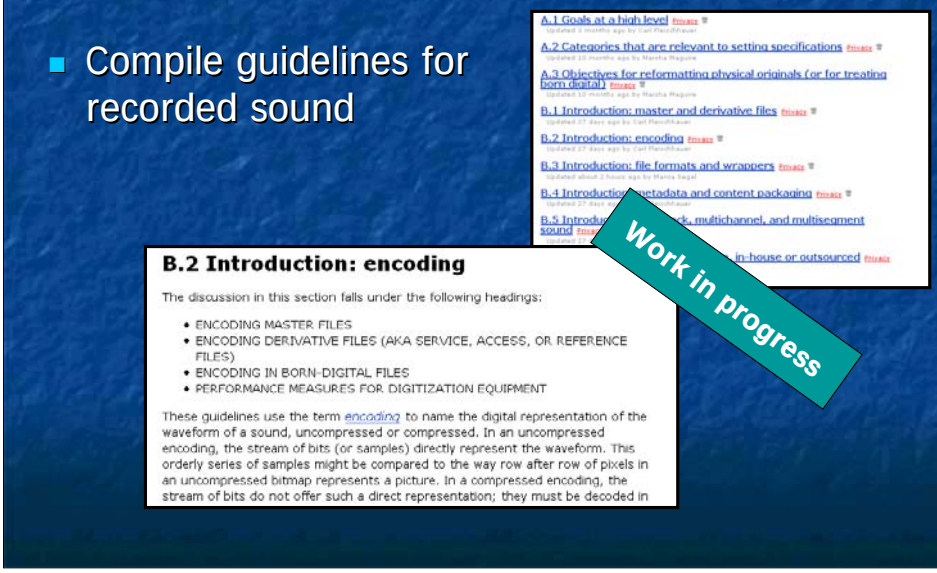
The images of the targets are analyzed by a custom application built on top of LabVIEW, from National Instruments. It reports if your scanning device passes or fails and provides more detail if you wish.



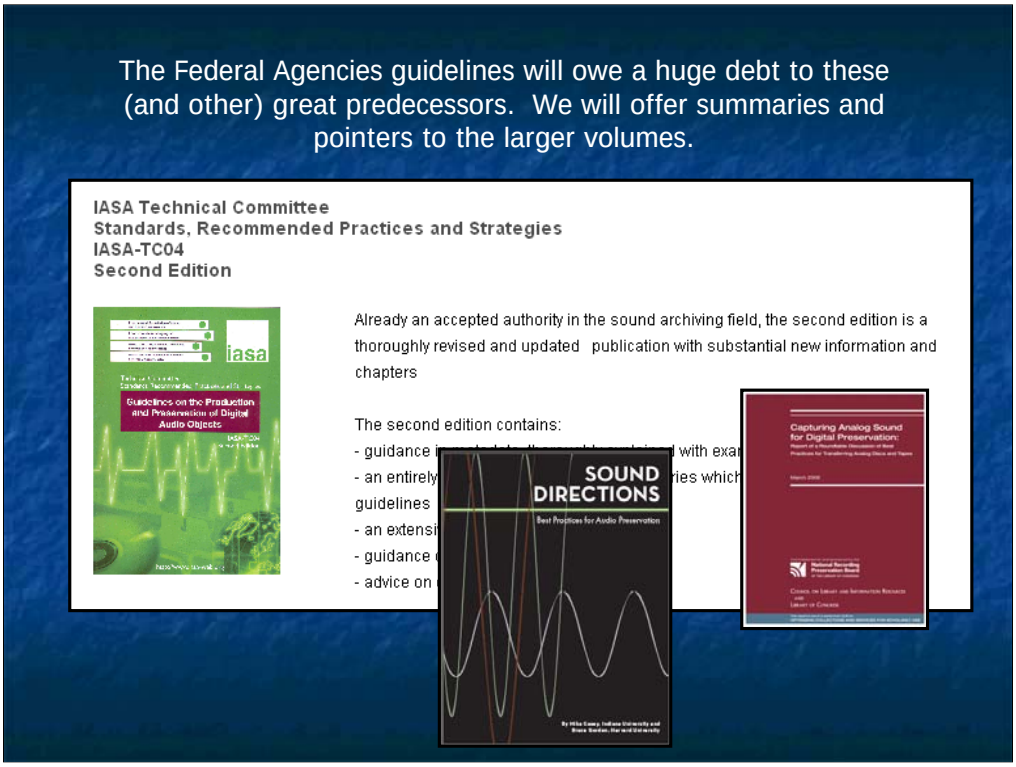
Imaging performance is not the only game for the Still Images Working Group. They have also published a recommendation for metadata to be embedded in image file headers, and there is a list of “gaps” in existing guidelines, to be filled during the next year or two.

Audio-visual effort: recorded sound

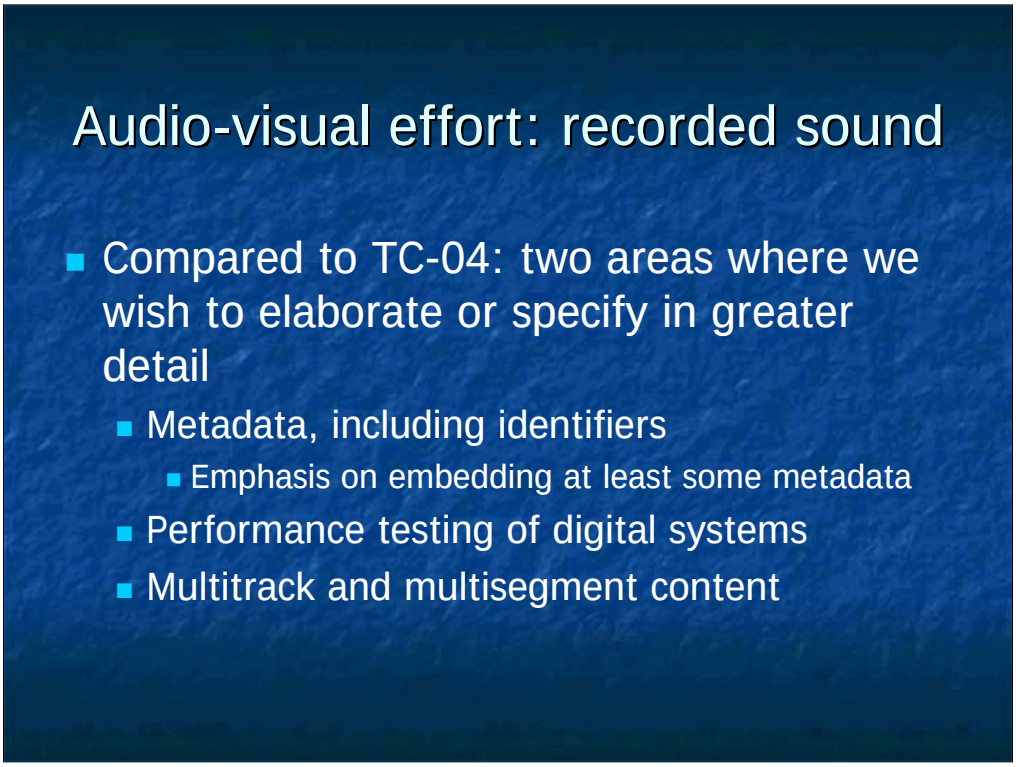
- Compile guidelines for recorded sound



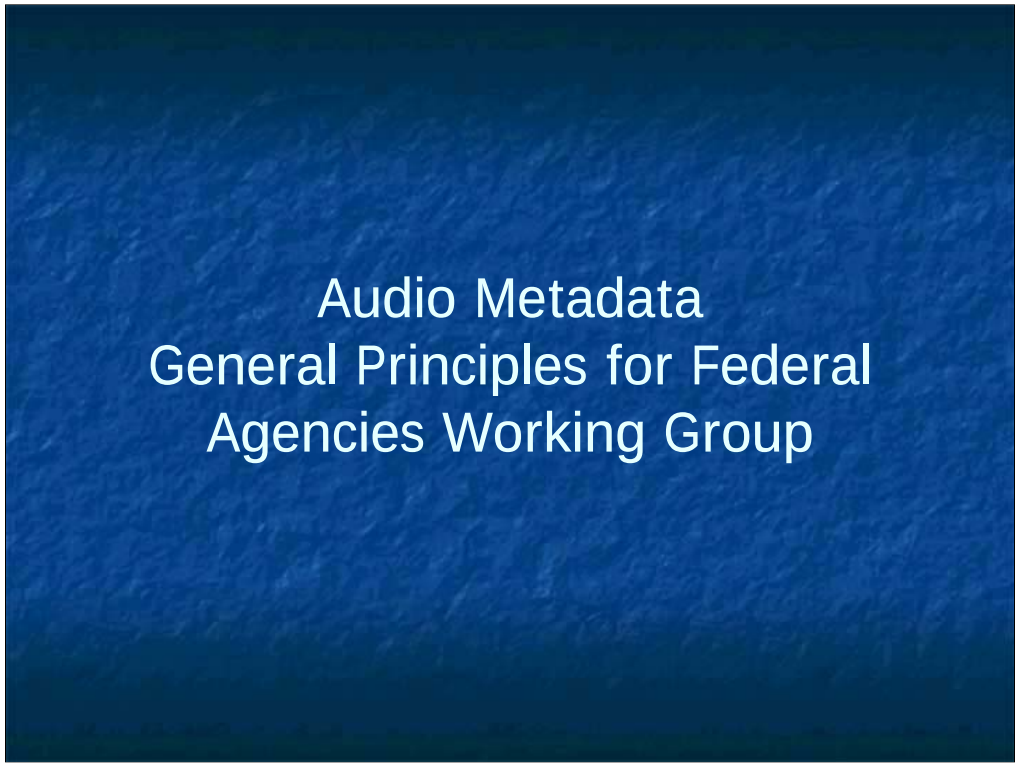
Meanwhile, in the Audio-Visual Working Group, we have started to compile a guideline--as comprehensive as we can manage-- pertaining to sound recordings.



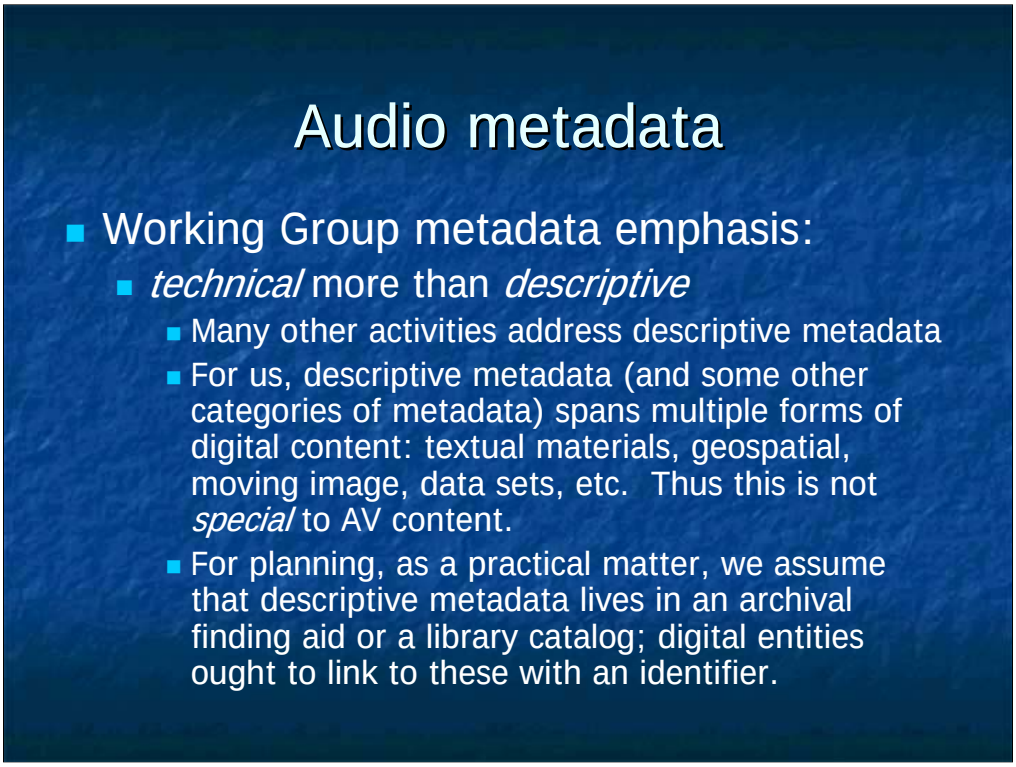
It will build on the great work produced by IASA, the Sound Directions project from Indiana and Harvard Universities, the Engineer's Roundtable organized by the National Recording Preservation Board, and other documents.



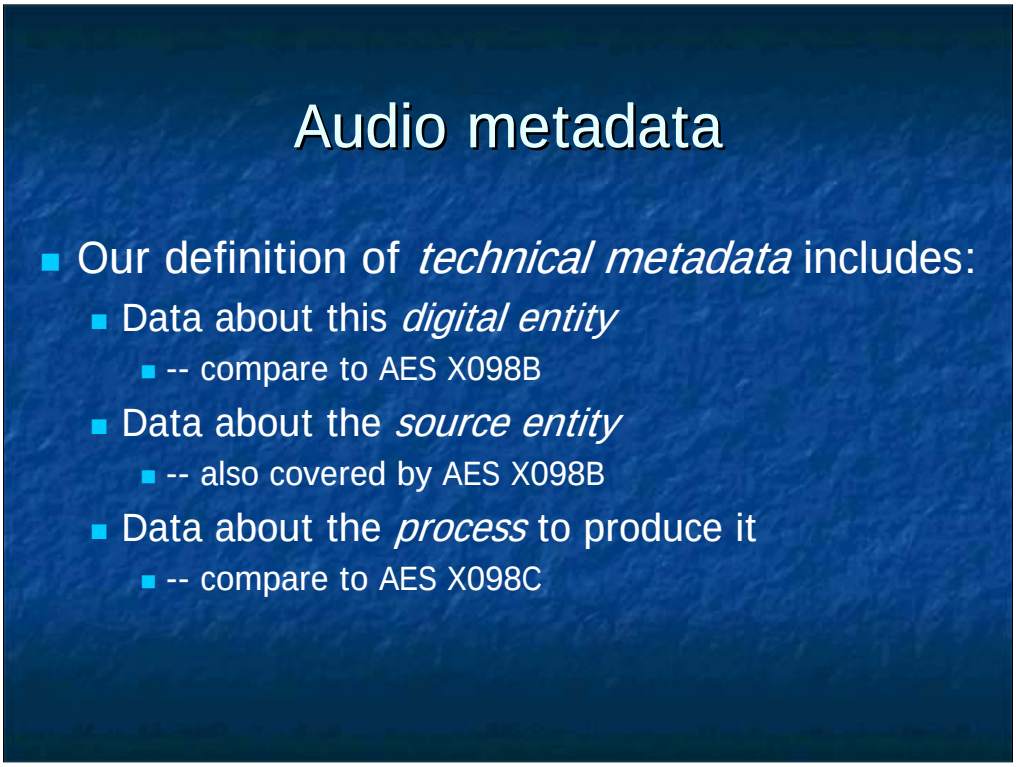
As we drafted our document, we saw a few areas where we want to elaborate or add detail to what is offered in TC-04. These include metadata and digital system performance testing, which I will discuss in this talk. We also see a need for more guidance on multitrack and multisegment recordings but we have nothing to offer at this time.



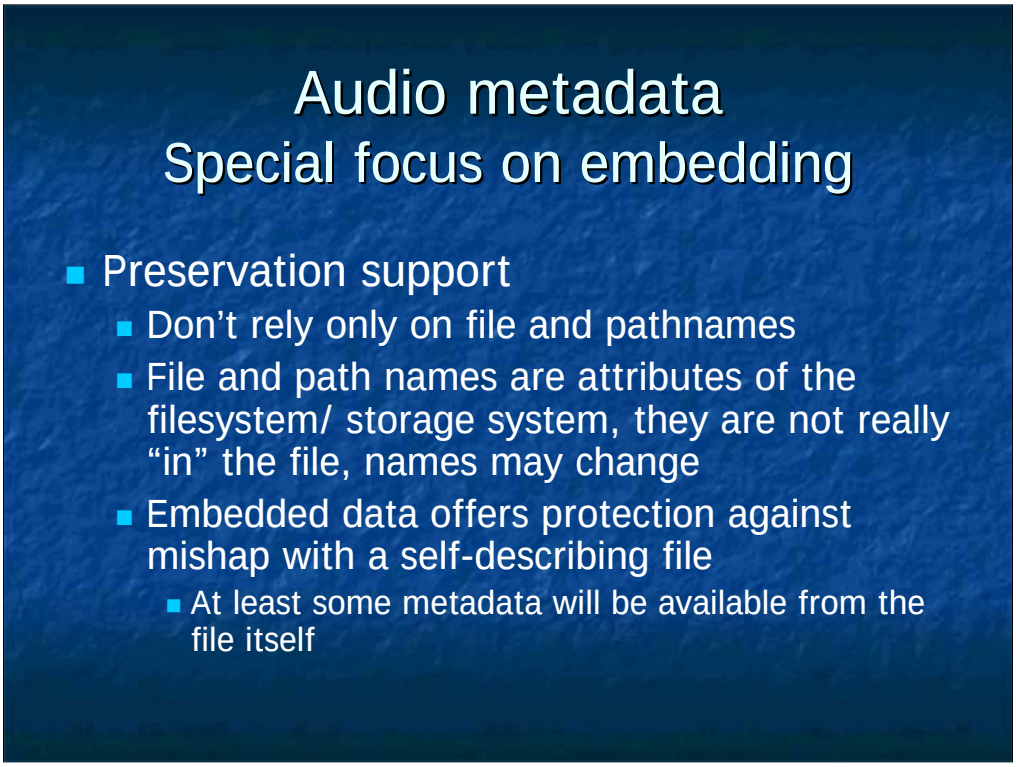
In TC-04, Chris Clark's chapter on metadata covers a broad swath, reminding us all how big this topic is.



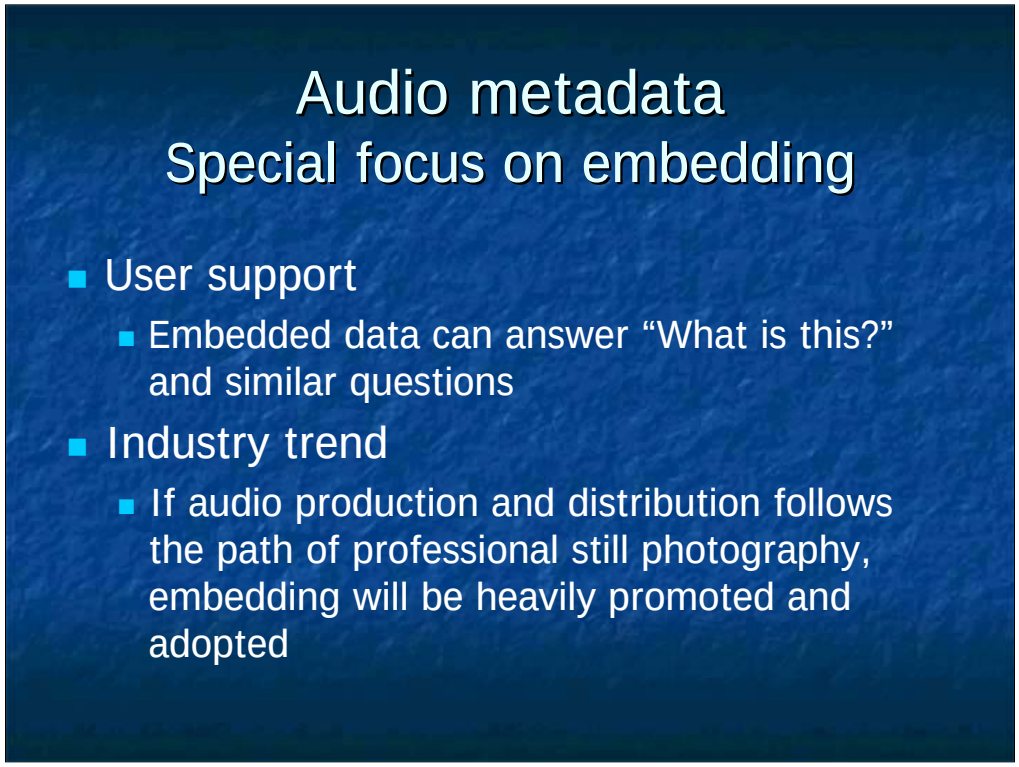
As a practical matter, our Working Group decided to take on a smaller piece of the whole. We think our expertise can be applied to technical metadata -- we will leave descriptive (mostly) to others. In any case, in institutions like ours, descriptive metadata practices tend to sweep across *all* types of digital content, and there are a lot of other players in that game.



Our definitions for technical metadata are in accord with the new specifications from the Audio Engineering Society. The metadata categories pertain to the *digital entity* you have produced, the *source entity* you started with, and the *process* used to digitize.



Like the Still Images Working Group, we have also begun to specify what ought to be embedded in digital entities. This is partly to support preservation. Identifiers in file and path names may be helpful . . . but they change or are subject to change by (for example) a digital asset management system. We think their is added safety in having identifying metadata *in* the file itself.

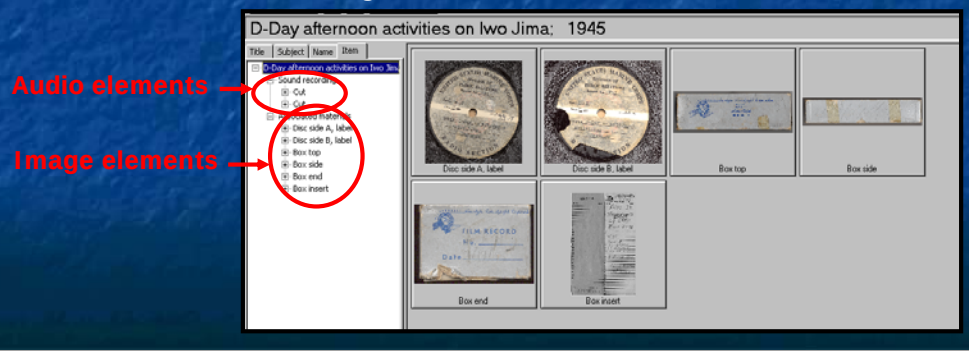


Embedded metadata can also help endusers who may have downloaded a file and, if professional still photography is any kind of indicator, we will see more and more embedding in the sound and video industries.

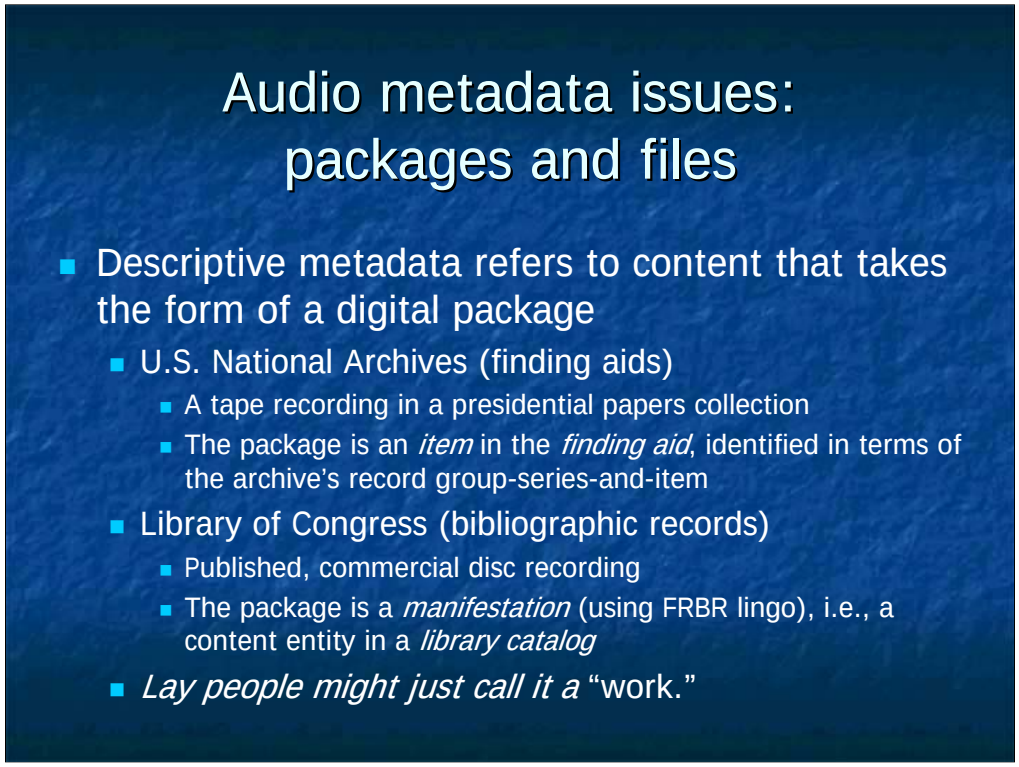
As we proceeded, we considered some issues.

Metadata planning issues: packages and files

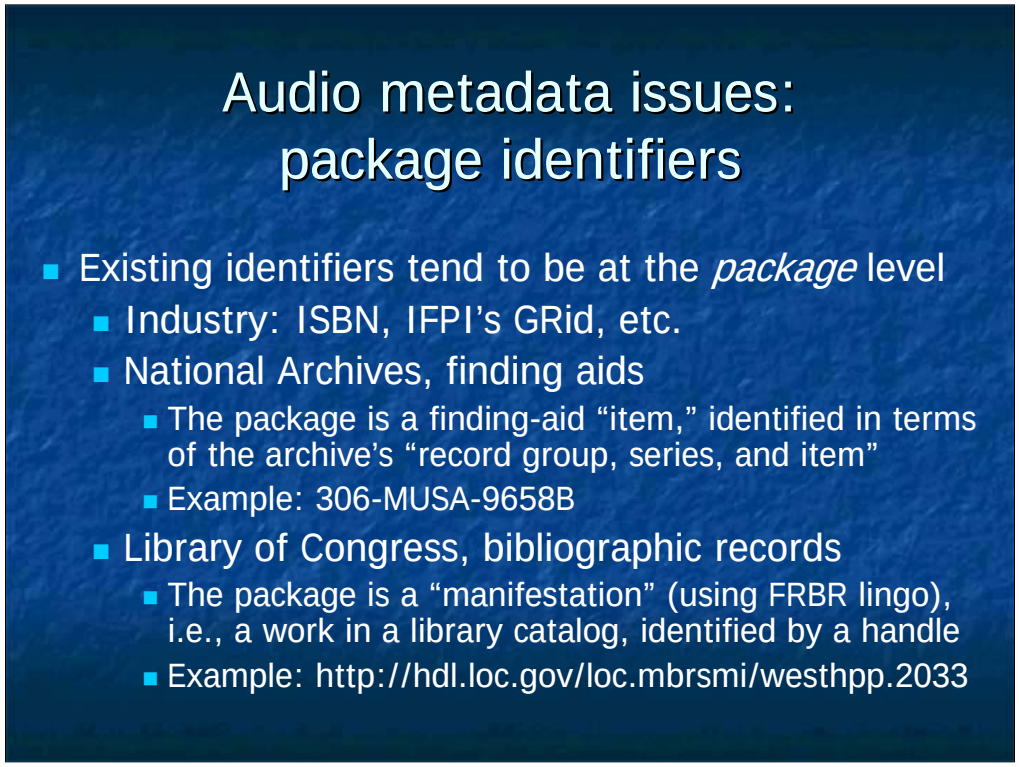
- *packages* as compared to *files*
  - An package is an entity (often) represented by multiple files
  - A side, B side, images of labels



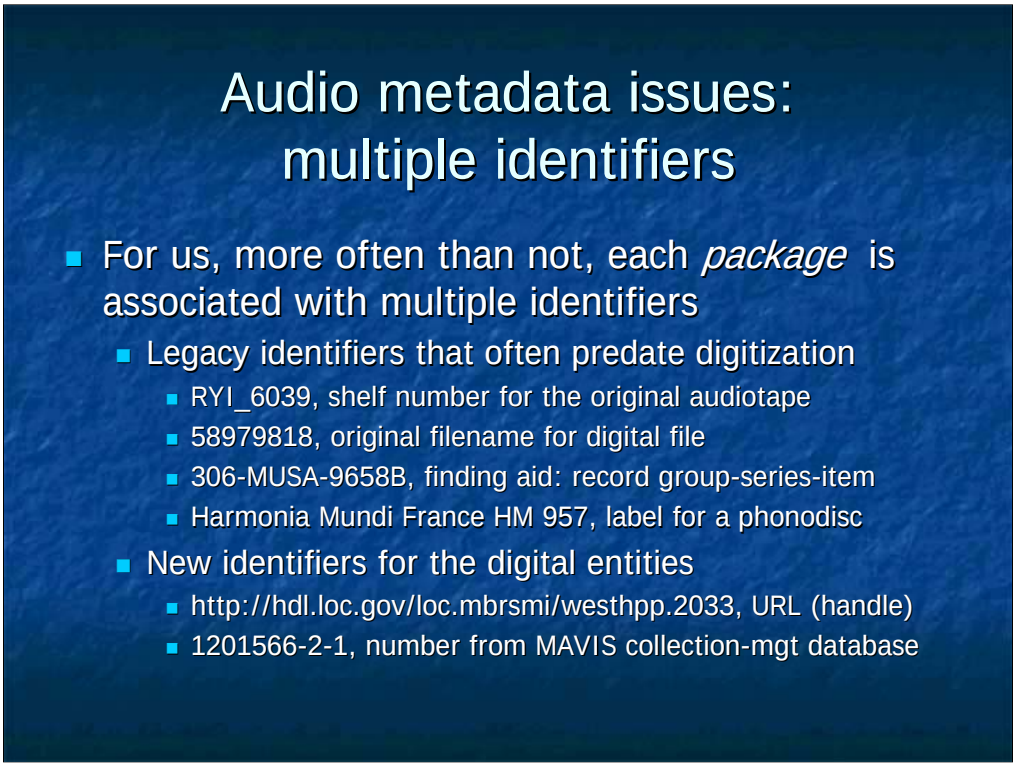
We see digital *entities* as taking (at least) two forms: *packages* and *files*. Our simple definition of package is "a digital entity made up of multiple files."



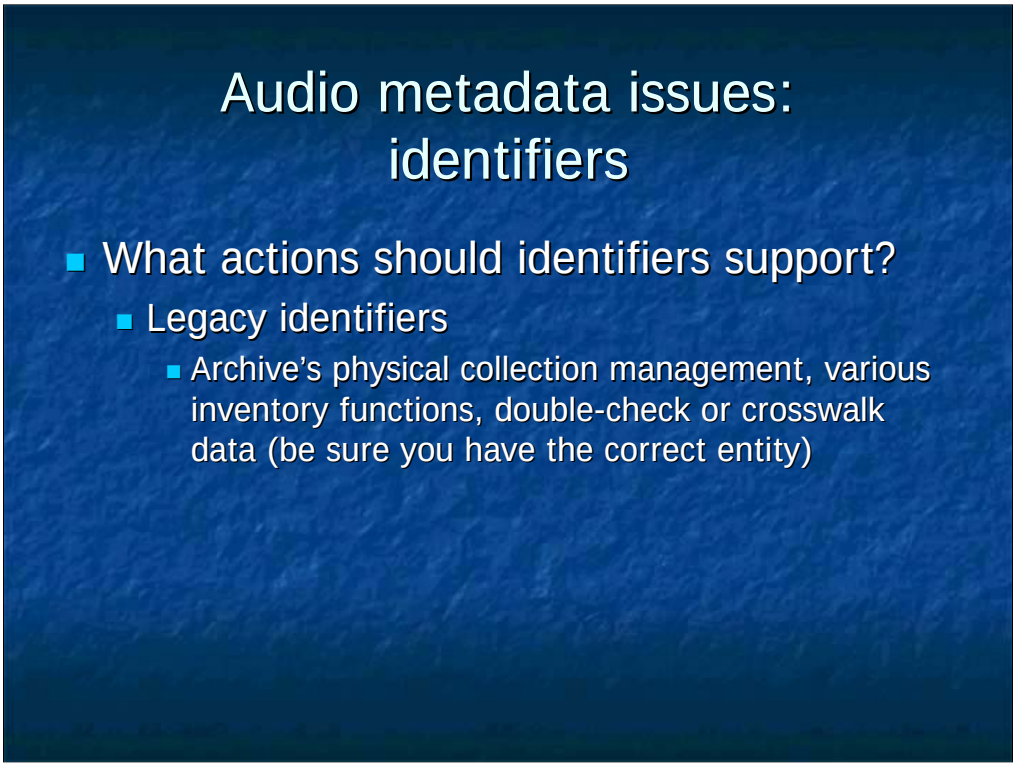
More often than not, descriptive metadata refers to an entity that--once it is in digital form--takes the form of a *package*. Archivists who use a finding aid often call this content entity an *item.* Librarians who use bibliographic records in a catalog (and who use FRBR terminology) may call it a *manifestation*. The rest of us just call the package entity a "work."



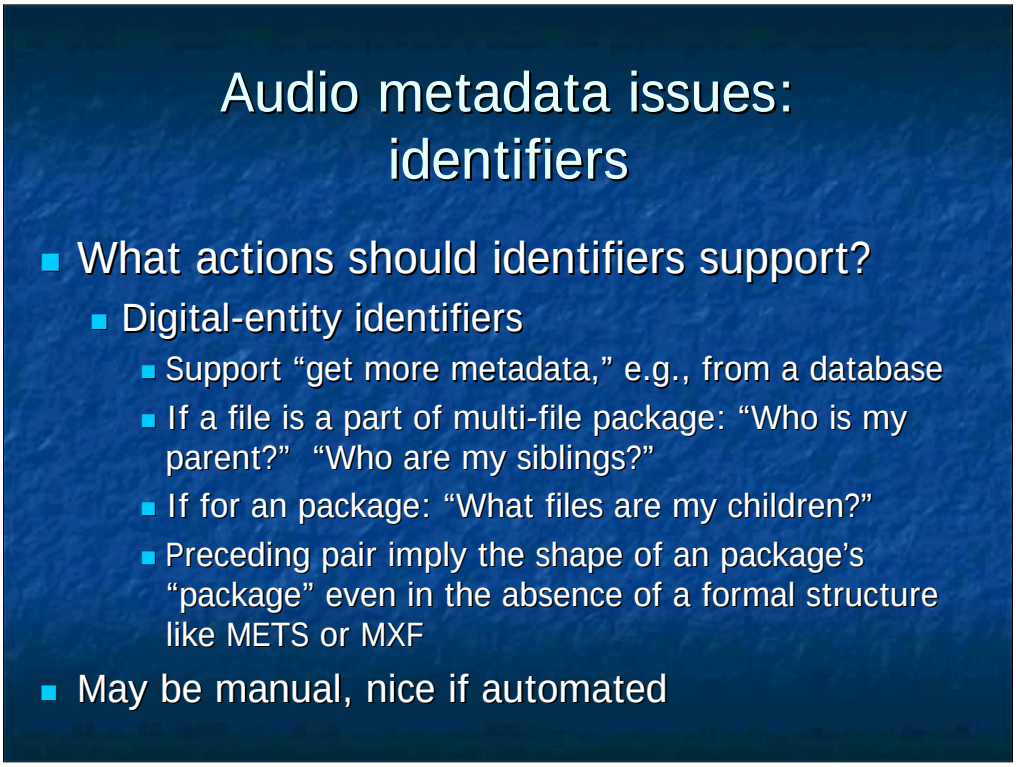
Many pre-existing identifiers are associated with content at the same level as the descriptive metadata—if the content is digital: *packages*. Examples from industry are the ISBN or the IFPI's GRid (Global *Release Identifier*). And from memory institutions we have the U.S. National Archives' representation of *Record Group, Series, and Item*; and from the Library of Congress, *a handle*.



In addition, we found that our organizations usually have multiple identifiers in play, ranging from the shelf number for the tape to the filename this one had "last time."



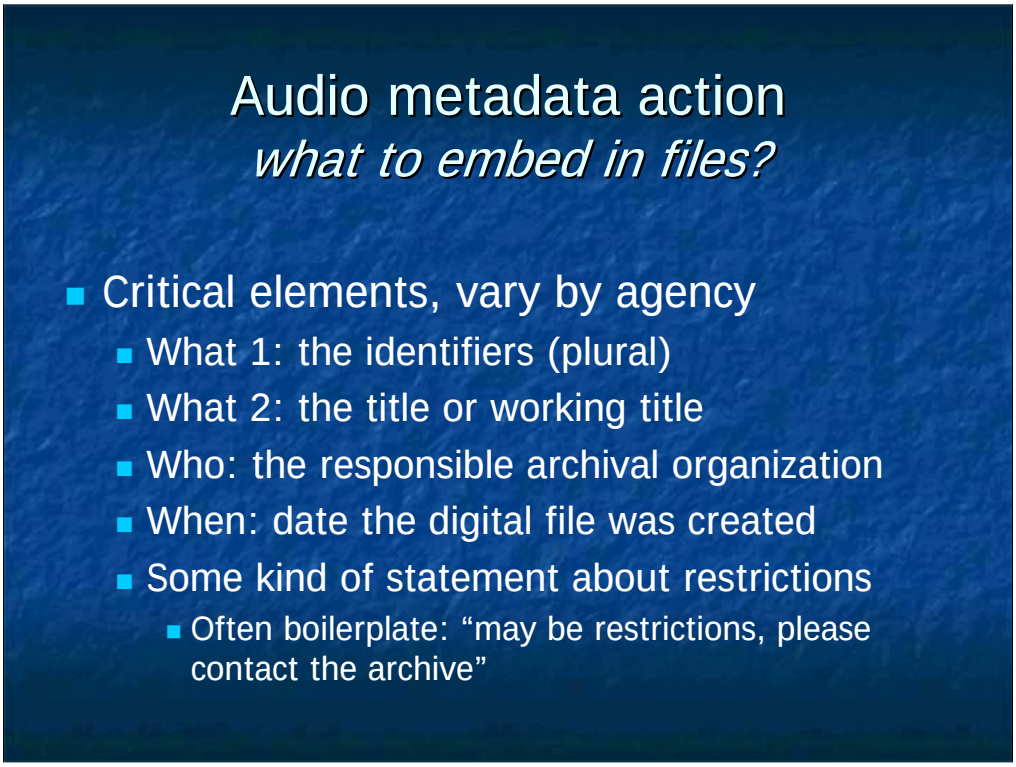
. . . . as a little side-topic: let's remind ourselves what identifiers help people do. Legacy identifiers like the shelf number give archivists another way to cross-check the inventory.



For digital-entity identifiers, the most important job is "go and get more metadata," for example, from a database. With content in a package, identifiers might help a "parent" find the "children files" or might let one of the "children" be connected to the "parent package."

So what is our Working Group actually doing?

We found that package-level practices were in their infancy; they were too local to build upon at this time. None of our organizations have really implemented packaging schemes like METS or MXF. So we decided to start with metadata at the file level.



When we talked about what our agencies want to embed, there was some variation. Here's the set of core elements we ended up with--some will be optional:

What is this file? Answered (above all) by the identifiers and, for some, by a title or working title.

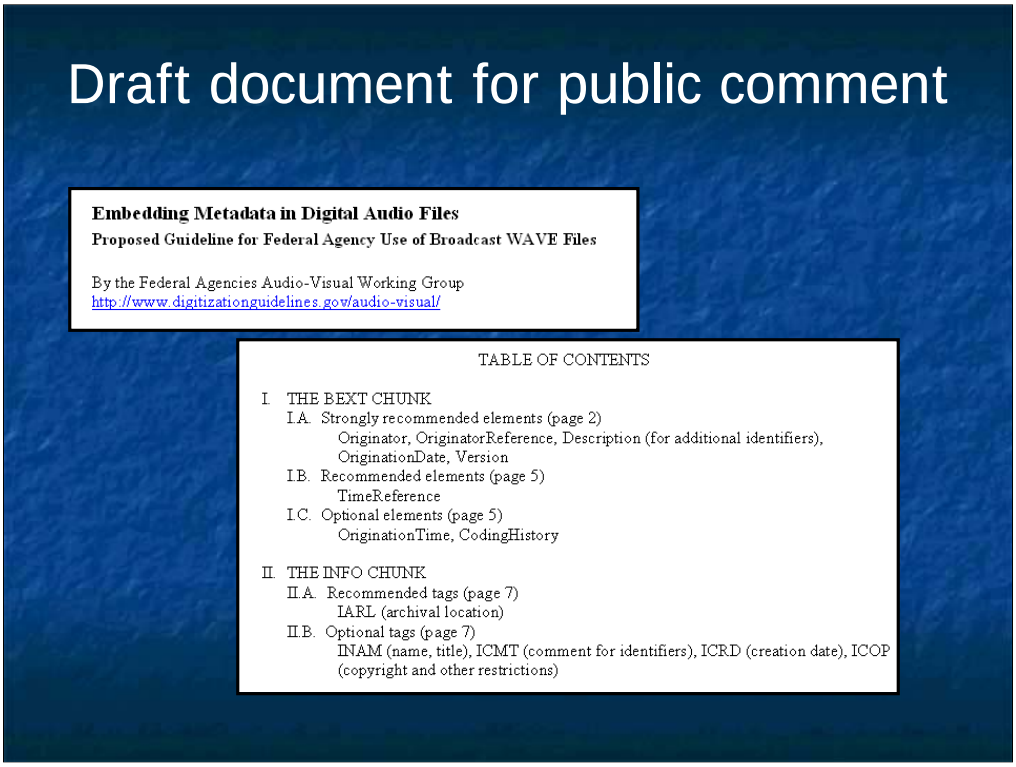
Who is responsible? The name of the archiving organization.

When was the digital entity created?

And some kind of a statement about restrictions on access or use. For a lot of our materials, this will be boilerplate: "May be restricted, check with the archive."

We have a draft document for public comment.

Draft document for public comment



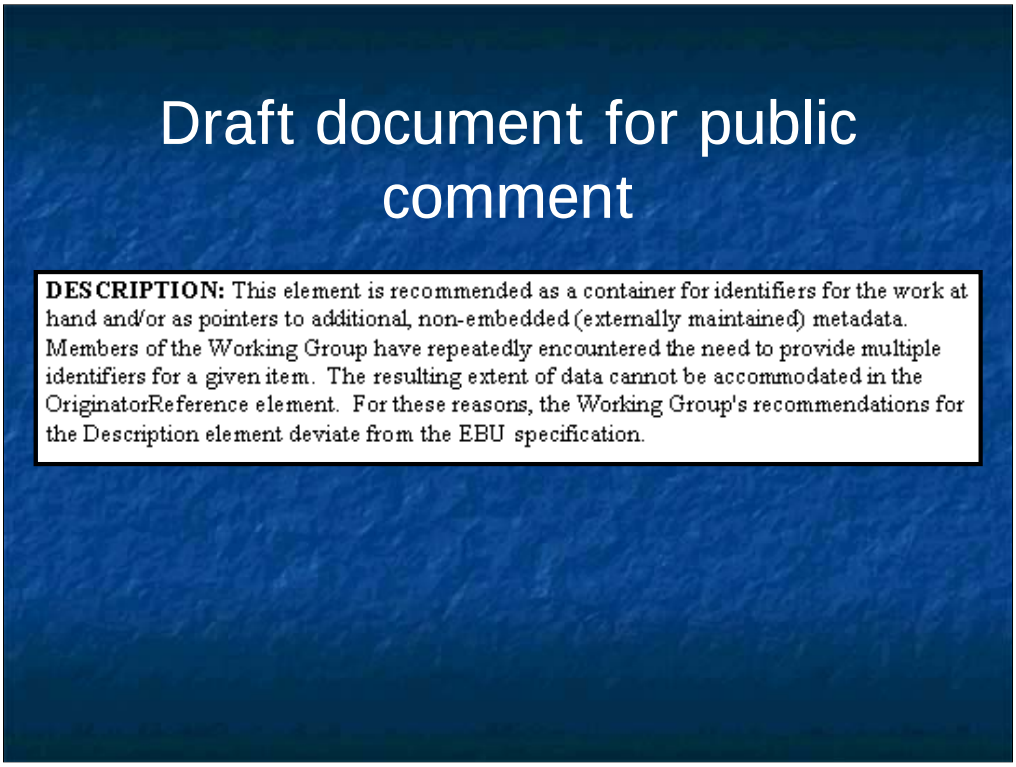
Embedding Metadata in Digital Audio Files

Proposed Guideline for Federal Agency Use of Broadcast WAVE Files

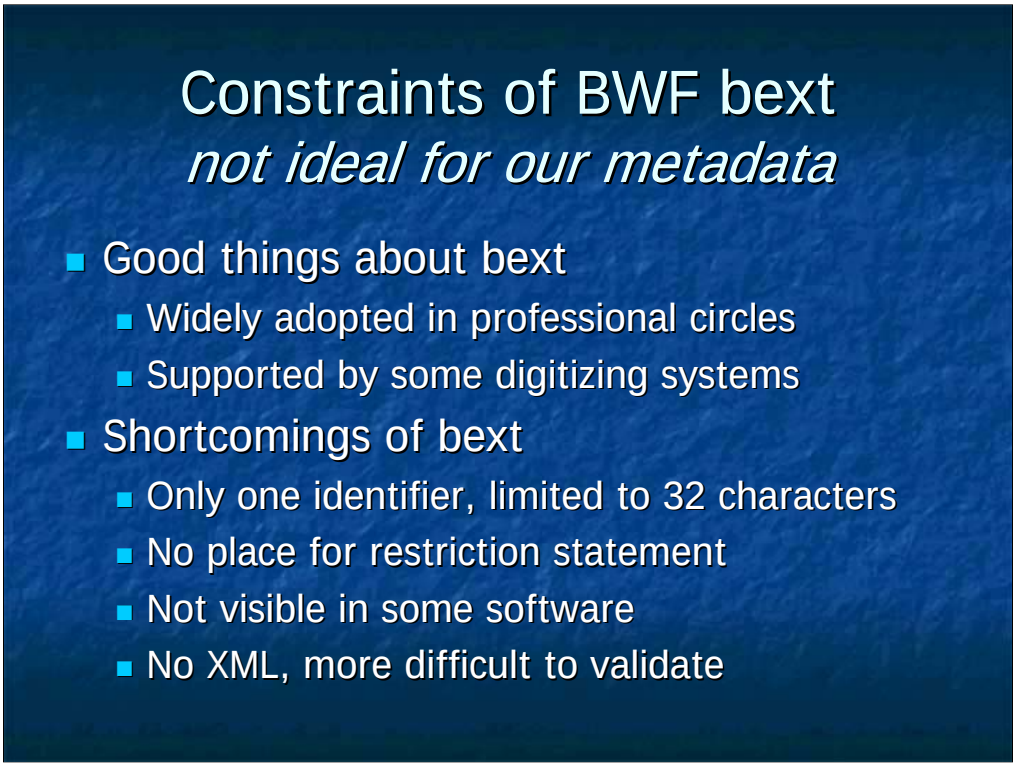
By the Federal Agencies Audio-Visual Working Group
http://www.digitizationguidelines.gov/audio-visual/

TABLE OF CONTENTS

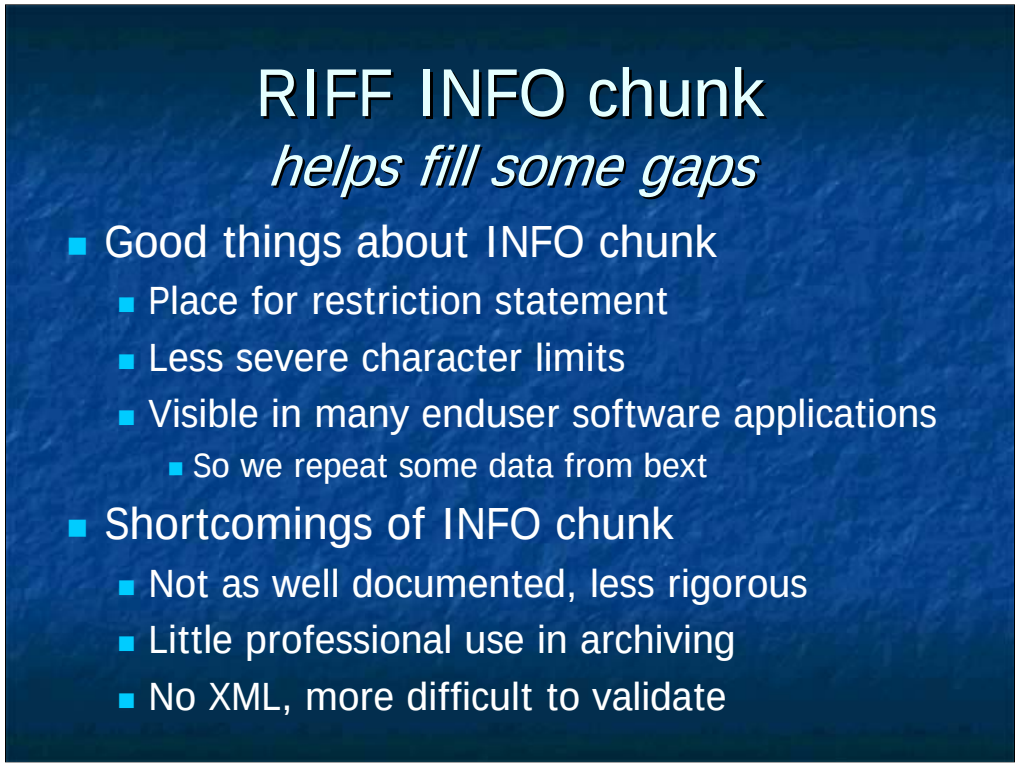
It recommends the placement of certain types of information in elements in both the Broadcast WAVE bext chunk and also the RIFF INFO chunk.



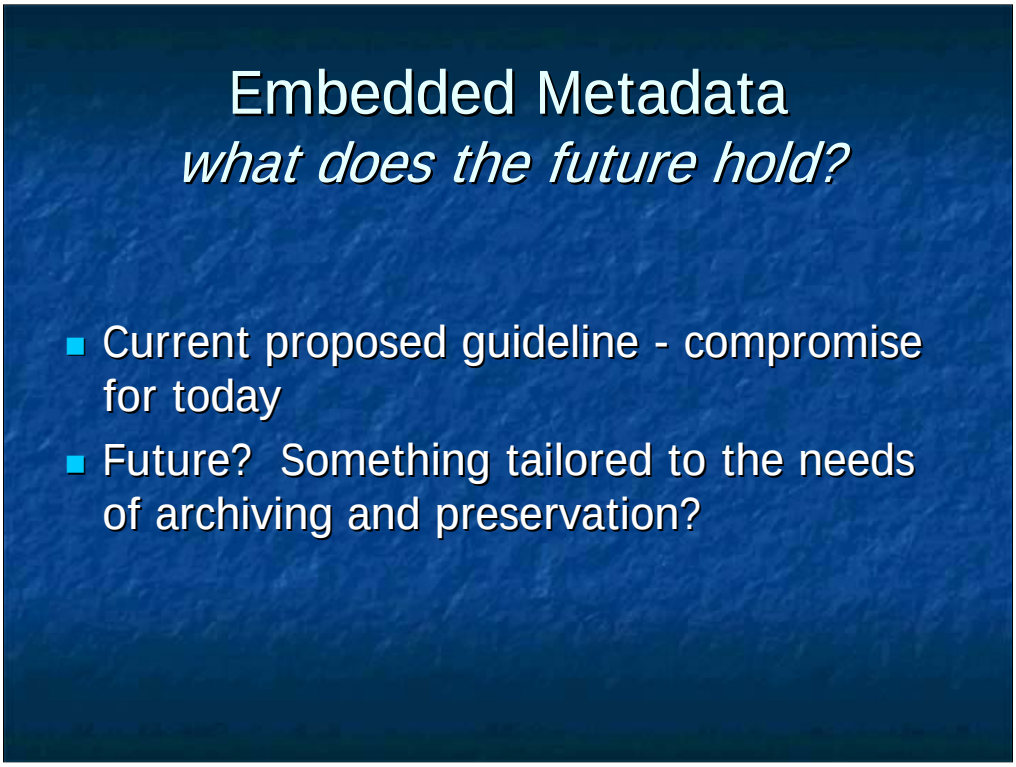
Our greatest departure from the EBU specification is for the Description element. We needed a place to put multiple identifiers, and we decided to use the Description element to meet that need.



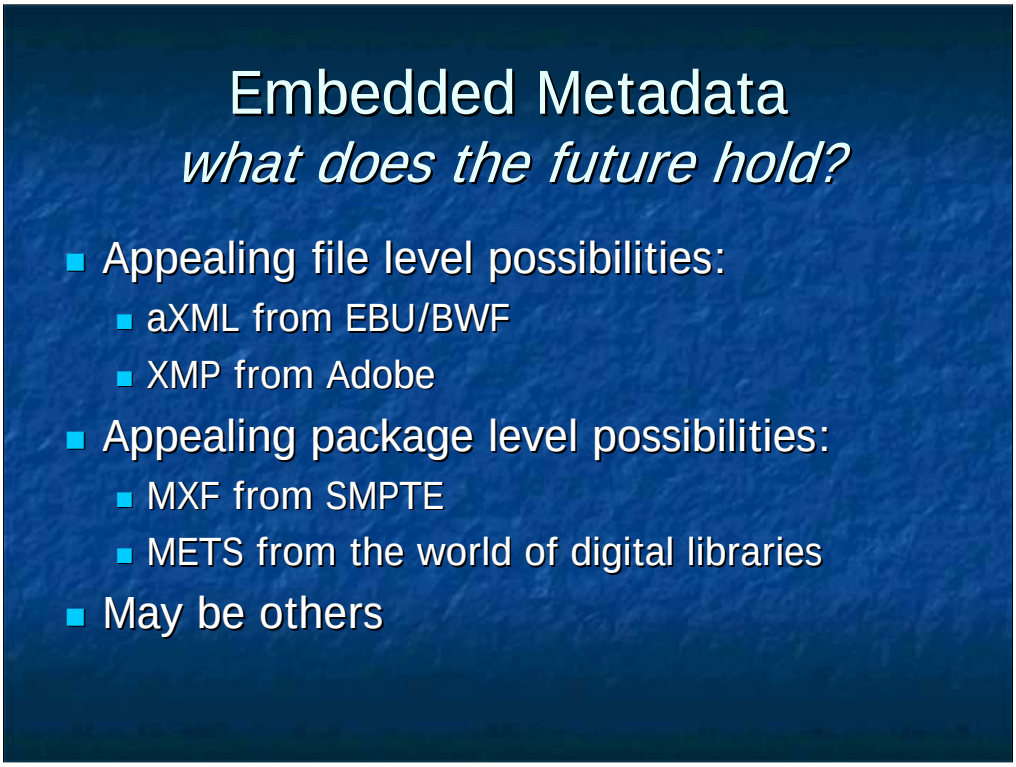
As the preceding suggests, we found that the bext chunk fell short of our requirements. Other shortfalls include the lack of a place for a restriction statement, and the fact that bext metadata is not visible in all application software.



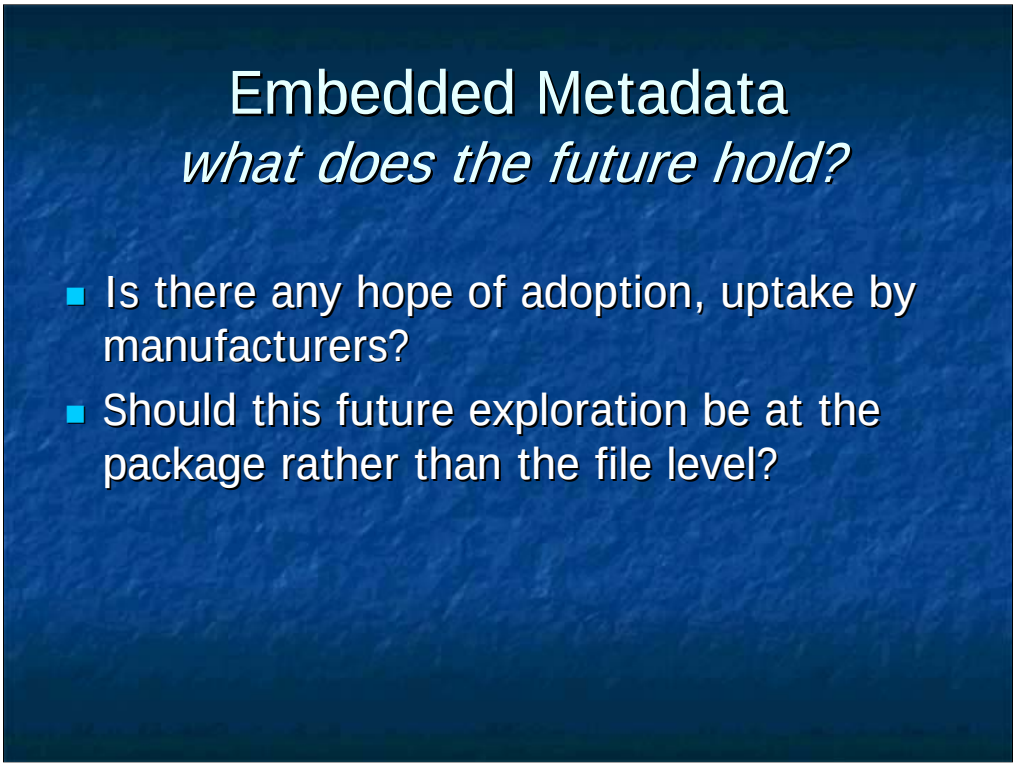
This is why we allow for optional use of the RIFF INFO chunk that is part of all WAVE files. INFO offers a few additional places to park metadata, and this metadata is visible in common end-user software. But neither bext nor INFO are expressed as XML and cannot be validated.



Thus we see our guideline as a compromise for now. We fervently wish for something tailored to the needs of preservation archiving.



There are some other metadata options to consider, some for embedding, some not.

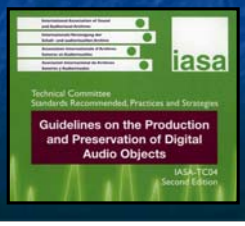


But we worry about adoption. The implementation of any of those options depends upon support from manufacturers, notably those who make digital audio workstations or playback software.

Next . . . we have started to look at digitization system performance testing.

Performance testing of audio digitization systems

- System performance testing
  - Considering IASA TC04 pass-fail specifications
  - Digital systems (A-to-D convertor, DAW, and other elements in path)



2.4.1.1 **Total Harmonic Distortion + Noise (THD+N)**

With signal 997 Hz at -1 dB FS, the A/D converter THD+N will be less than -105 dB unweighted, -107 dB A-weighted, 20 Hz to 20 kHz bandwidth limited.

With signal 997 Hz at -20 dB FS, the A/D converter THD+N will be less than -95 dB unweighted, -97 dB A-weighted, 20 Hz to 20 kHz bandwidth limited.

2.4.1.2. **Dynamic Range (Signal to Noise)**

The A/D converter will have a dynamic range of not less than 115 dB unweighted, 117 dB A-weighted. (Measured as THD+N relative to 0 dB FS, bandwidth limited 20 Hz to 20 kHz, stimulus signal 997 Hz at -60 dB FS).

2.4.1.3. **Frequency Response**

For an A/D sampling frequency of 48 kHz, the measured frequency response will be better than ± 0.1 dB for the range 20 Hz to 20 kHz.

For an A/D sampling frequency of 96 kHz, the measured frequency response will be better than ± 0.1 dB for the range 20Hz to 20 kHz, and ± 0.3 dB for the range 20 kHz to 40 kHz.

For an A/D sampling frequency of 192 kHz, the frequency response will be better than ± 0.1 dB for the range 20Hz to 20 kHz, and ± 0.3 dB from 20 kHz to 50 kHz (reference audio signal = 997 Hz, amplitude -20 dB FS).

2.4.1.4 **Intermodulation Distortion IMD (SMPTE/DIN/AES17)**

The A/D converter IMD will not exceed -90 dB. (AES17/SMPTE/DIN twin-tone test sequences, combined tones equivalent to a single sine wave at full scale amplitude):

2.4.1.5 **Amplitude Linearity**

The A/D converter will exhibit amplitude gain linearity of ± 0.5 dB within the range -120 dB FS to 0 dB FS. (997 Hz sinusoidal stimuli).

2.4.1.6 **Spurious Aharmonic Signals**

Better than -130 dB FS with stimulus signal 997 Hz at -1 dBFS

2.4.1.7 **Internal Sample Clock Accuracy**

For an A/D converter synchronised to its internal sample clock, frequency accuracy of the clock measured at the digital stream output will be better than ±25 ppm.

2.4.1.8 **Jitter**

Interface jitter measured at A/D output <5ns.

2.4.1.9 **External Synchronisation**

Where the A/D converter sample clock will be synchronised to an external reference signal, the A/D converter must react transparently to incoming sample rate variations ± 0.2% of the nominal sample rate. The external synchronisation circuit must reject incoming jitter so that the synchronised sample rate clock is free from artefacts and disturbances.

TC-04 provides our starting point with its list of pass-fail numbers for analog-to-digital convertors. It states the minimum levels required for total harmonic distortion plus noise, intermodulation distortion, jitter, and so on.

Performance testing of audio digitization systems

- Desire: imitate still image working group approach: affordable target, easy-to-use software
- For lay persons, not engineers
- Requires affordable tone generation at IASA pass-fail levels
- Requires easy software to read and report THD+N, IMD, jitter, etc.
- Alas, this is challenging . . . to be continued

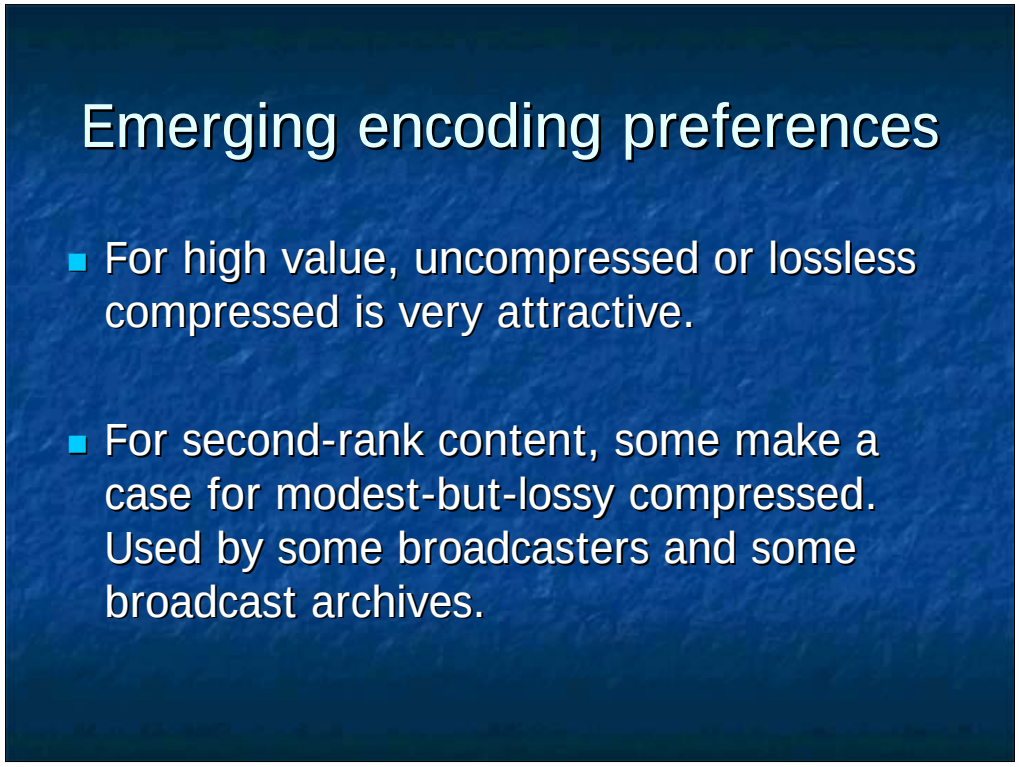
Here we wanted to imitate the Still Image Working Group with their target and software. We wanted something affordable, for non-engineers to use, so the staff at an archive could perform testing. But a solution has not come easily in view . . . we will continue to pursue this topic.

Meanwhile . . . with video content . . .

Audio-visual effort: video

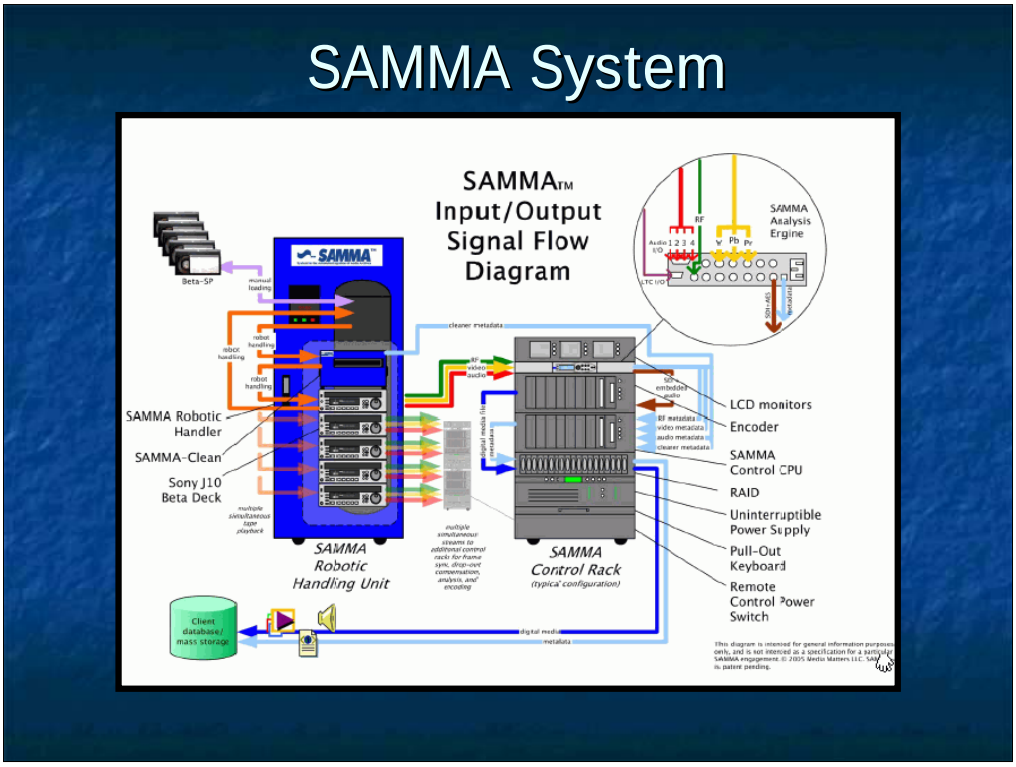
- Exploration of “target formats”
- We watch and wait while agencies to gain experience . . .

. . . our general approach is to wait for our members to gain some experience. It would be premature to make recommendations. But we do want to explore target formats.

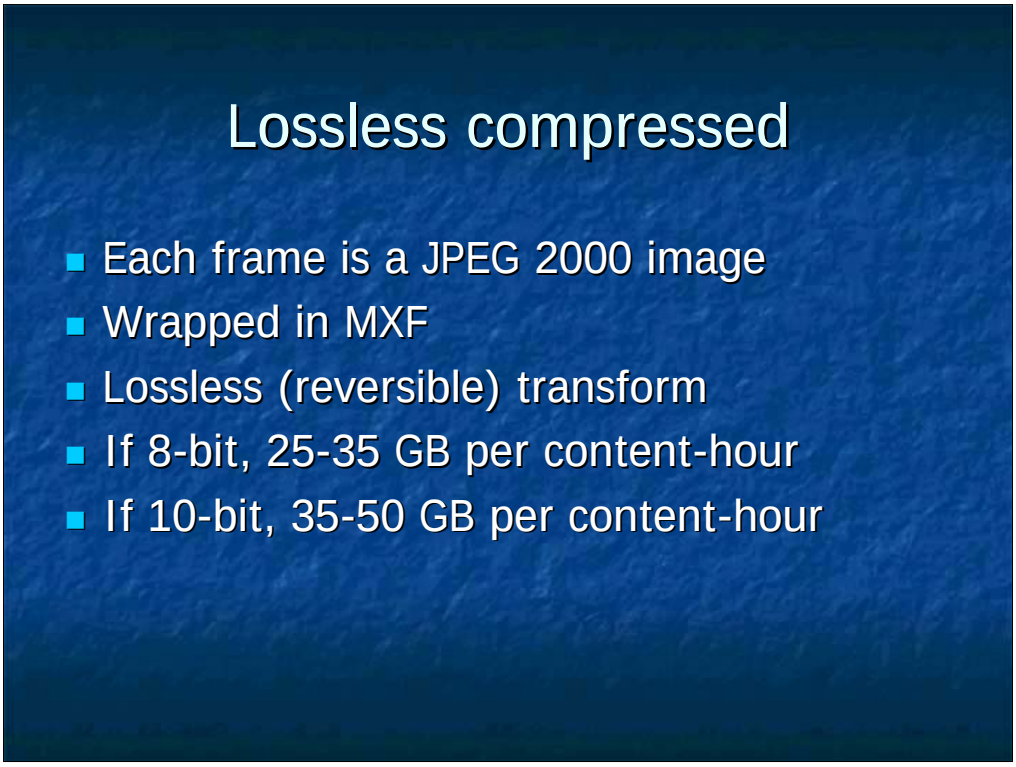


Our preferences--like those for still images and audio--are for essences that are uncompressed or compressed in a lossless manner.

Three federal agencies are engaged in some initial work: the Library of Congress, National Archives, and Smithsonian Institution.



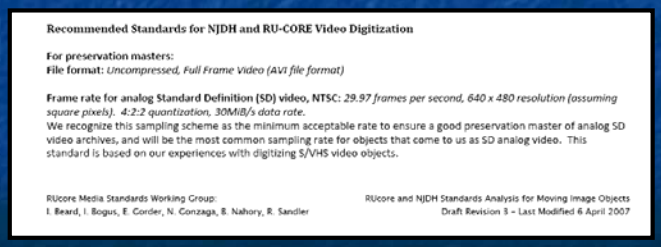
All three have purchased SAMMA systems and are starting to use them.



SAMMA produces a stream of video-frame images, each of which is encoded in the lossless JPEG 2000 format, wrapped in MXF. Early indications are that the file size ranges from 25 to 50 gigabytes per hour, depending on variables like bit depth.

Also of interest: uncompressed video

- U.S. activities: Stanford Univ., Rutgers Univ.
- 4:2:2 or 4:4:4, 10-bit SDI stream
- About 100 GB per content-hour
  - Another source reported 70 GB for 8-bit video



Recommended Standards for NJDH and RU-CORE Video Digitization

For preservation masters:
File format: *Uncompressed, Full Frame Video (AVI file format)*

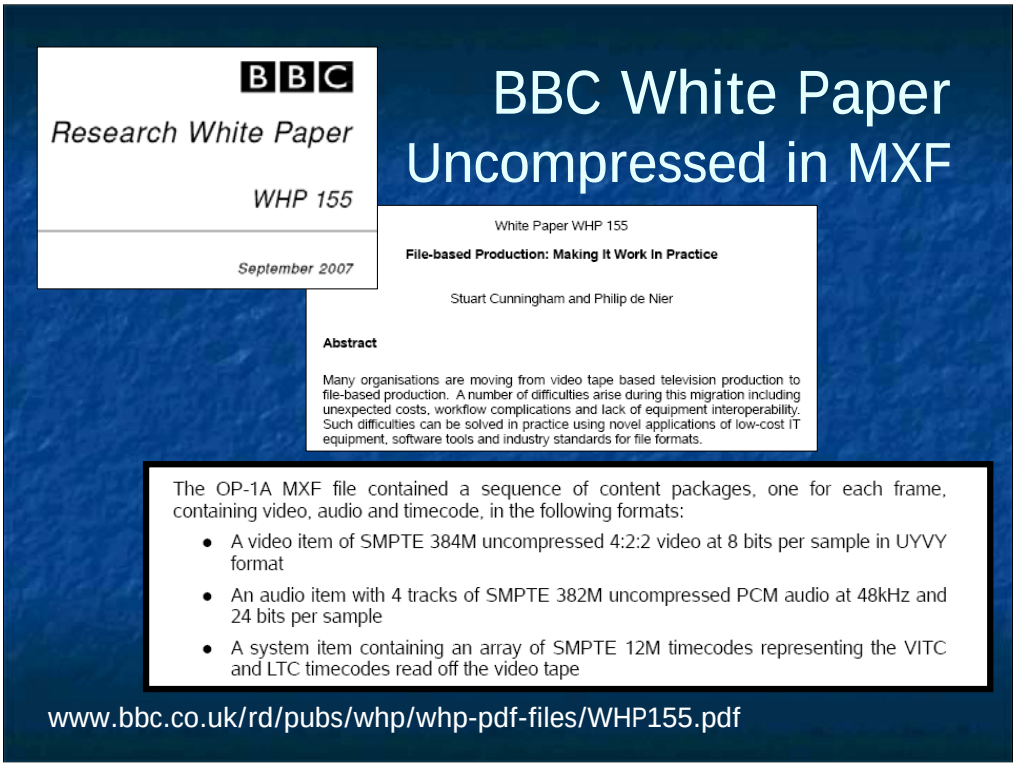
Frame rate for analog Standard Definition (SD) video, NTSC: *29.97 frames per second, 640 x 480 resolution (assuming square pixels). 4:2:2 quantization, 30MiB/s data rate.*
We recognize this sampling scheme as the minimum acceptable rate to ensure a good preservation master of analog SD video archives, and will be the most common sampling rate for objects that come to us as SD analog video. This standard is based on our experiences with digitizing S/VHS video objects.

RUcore Media Standards Working Group:
I. Beard, I. Bogus, E. Corder, N. Gonzaga, B. Nahory, R. Sandler

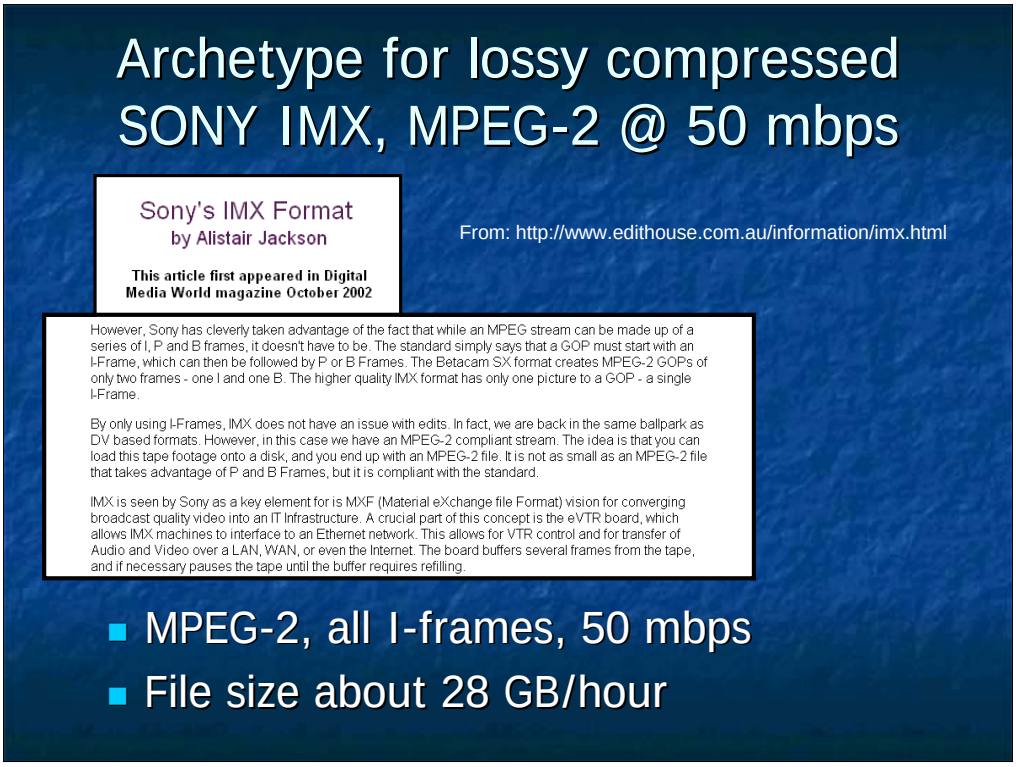
RUcore and NJDH Standards Analysis for Moving Image Objects
Draft Revision 3 – Last Modified 6 April 2007

Rutgers spec: http://rucore.libraries.rutgers.edu/collab/ref/dos_avwg_video_obj_standard.pdf

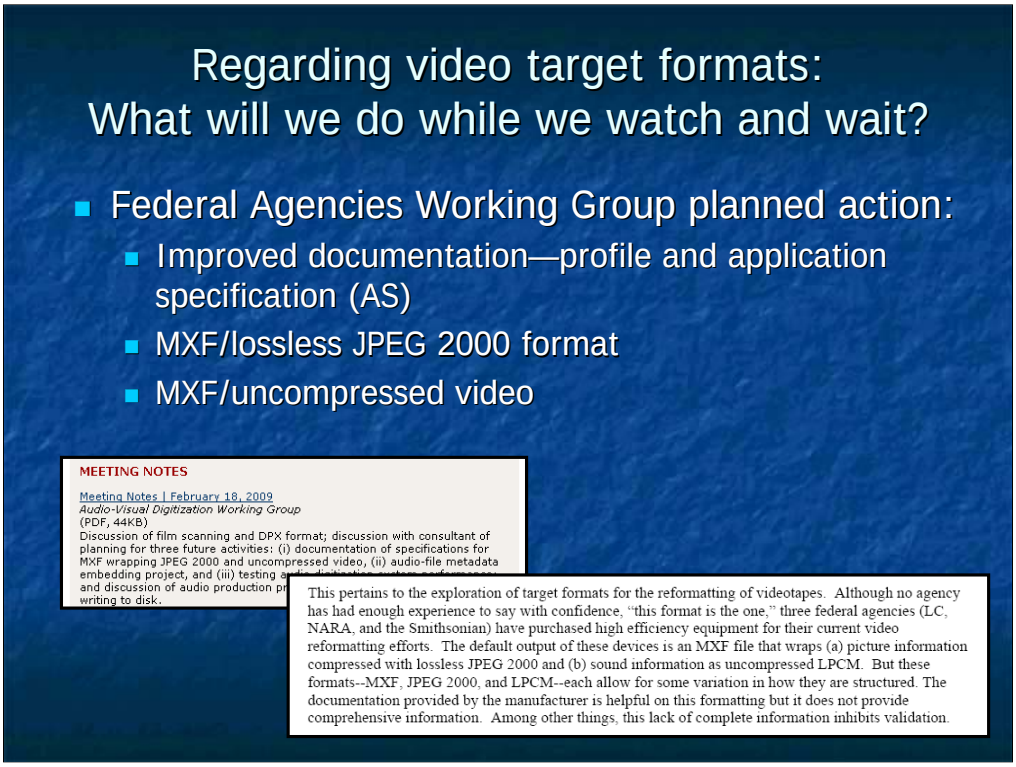
Members of our group are also interested in uncompressed video files. Some work is being done at two American universities, with file sizes reported in the range of 70 to 100 gigabytes per hour.



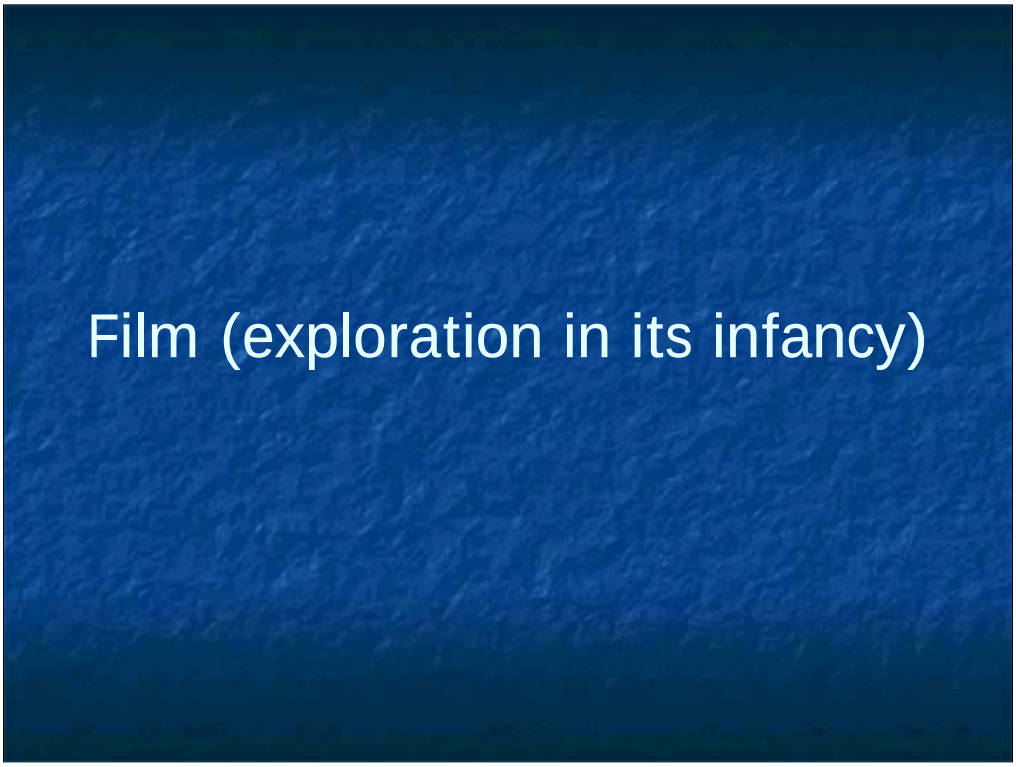
The most thorough discussion of this approach that I have seen comes from the BBC.



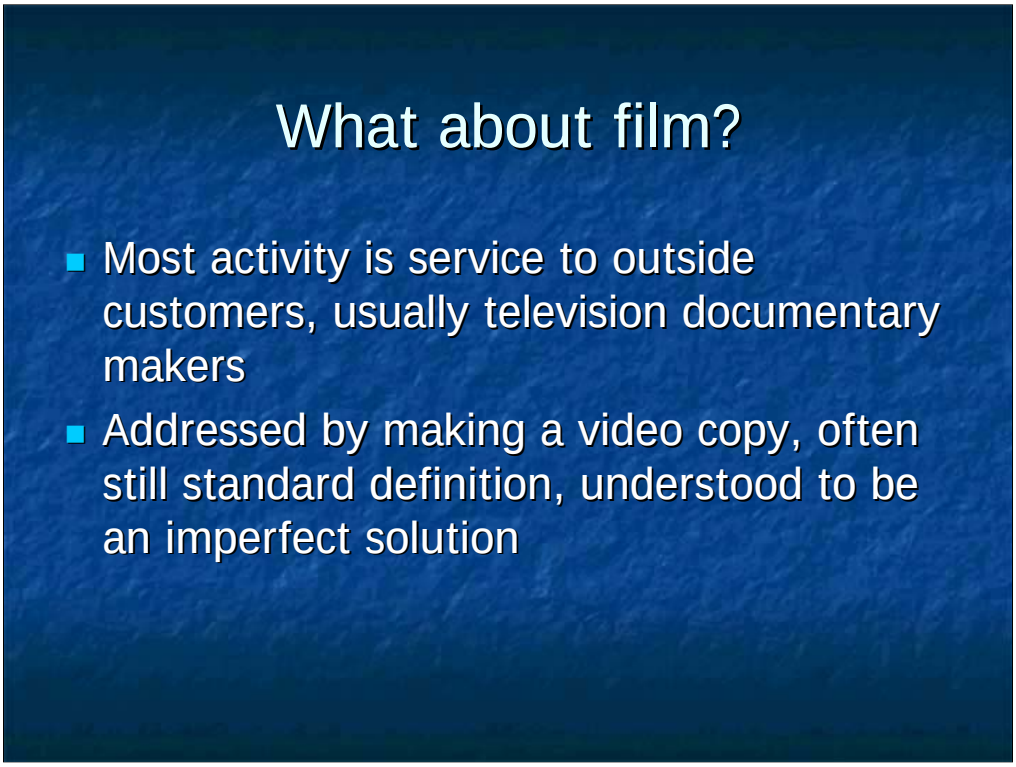
Meanwhile, we hear about high-resolution-but-lossy compression, often in broadcast archives. This usually employs an MPEG-2, all-I-frame approach, at 50 megabits per second, a format that owes a great debt to SONY's IMX systems. File sizes here are said to run about 28 gigabytes per hour.



As we watch these developments, we have been starting an effort to document the MXF/JPEG 2000 approach (and perhaps a similar uncompressed approach). We believe that a JPEG 2000 profile and an MXF applications specification would support making tools to validate of the files that SAMMA produces, and might encourage other vendors to build similar video conversion devices.



[Finally] . . . . Motion picture film scanning is in its infancy for our members.



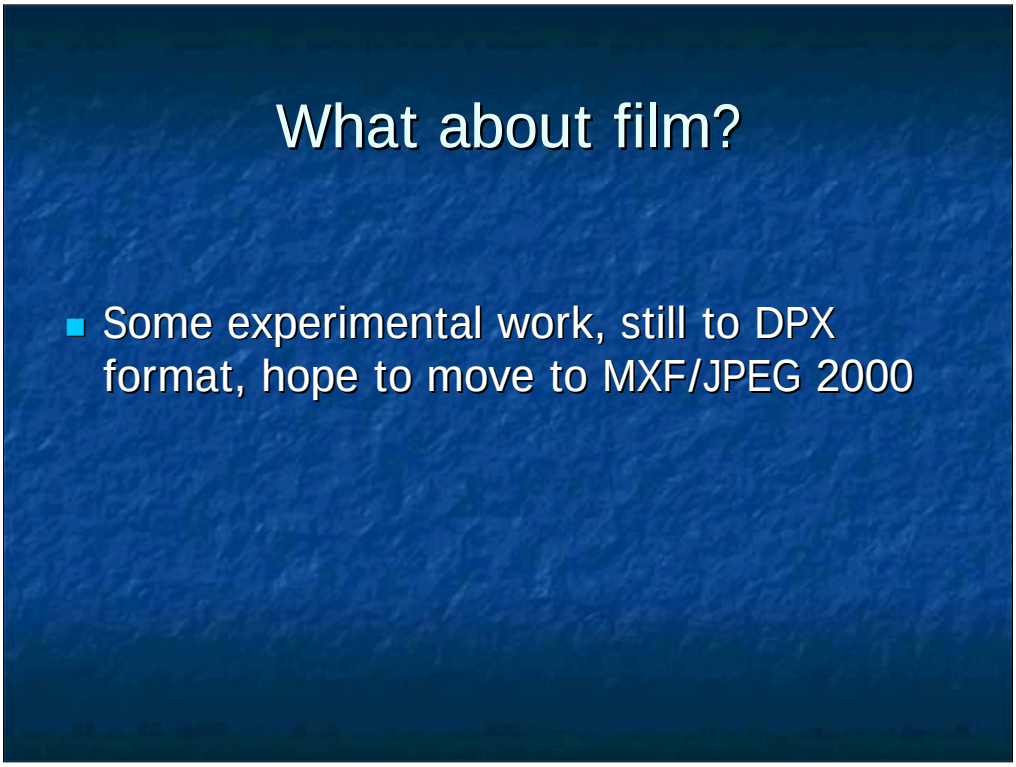
In federal agencies today, most digital reformatting of film is done in response to requests from the makers of video documentaries seeking historical footage. This need is addressed by making video copies but no one sees that as a perfect solution.

National Aeronautics and Space Administration

Most active high-resolution film scanning program: NASA Johnson Space Center

www.nasa.gov

We heard from the National Aeronautics and Space Administration (NASA) Johnson Space Center in Houston. As far as we can tell, JSC is doing more high-resolution film scanning than any other federal agency.



Everyone who scans film seems to use DPX as the target format, and our Working Group discussion of the matter highlighted some problems. We'd love to move in the direction of MXF and perhaps JPEG 2000 as we proceed.

Altogether, there is a lot to do. We would love to hear from you. Thank you for your interest.