Federal Digitization Moving to Common Guidelines

The U.S. Federal Agencies Digitization Initiative

http://www.digitizationguidelines.gov/

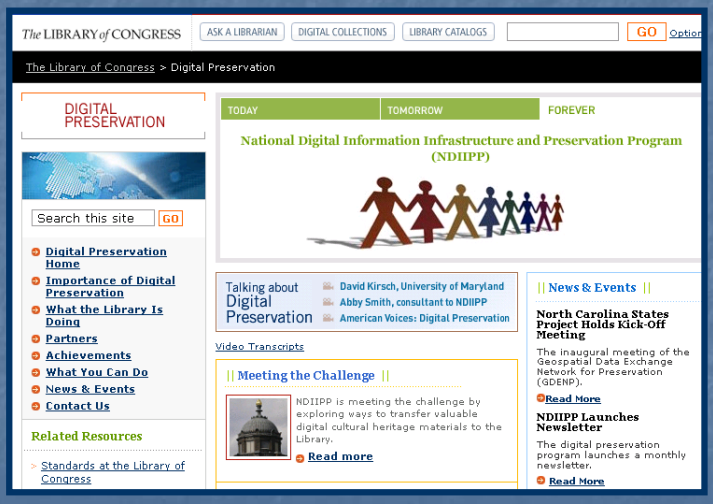
IS&T Archiving 2009 Conference

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

The Library of Congress

National Digital Information Infrastructure and Preservation Program (NDIIPP)

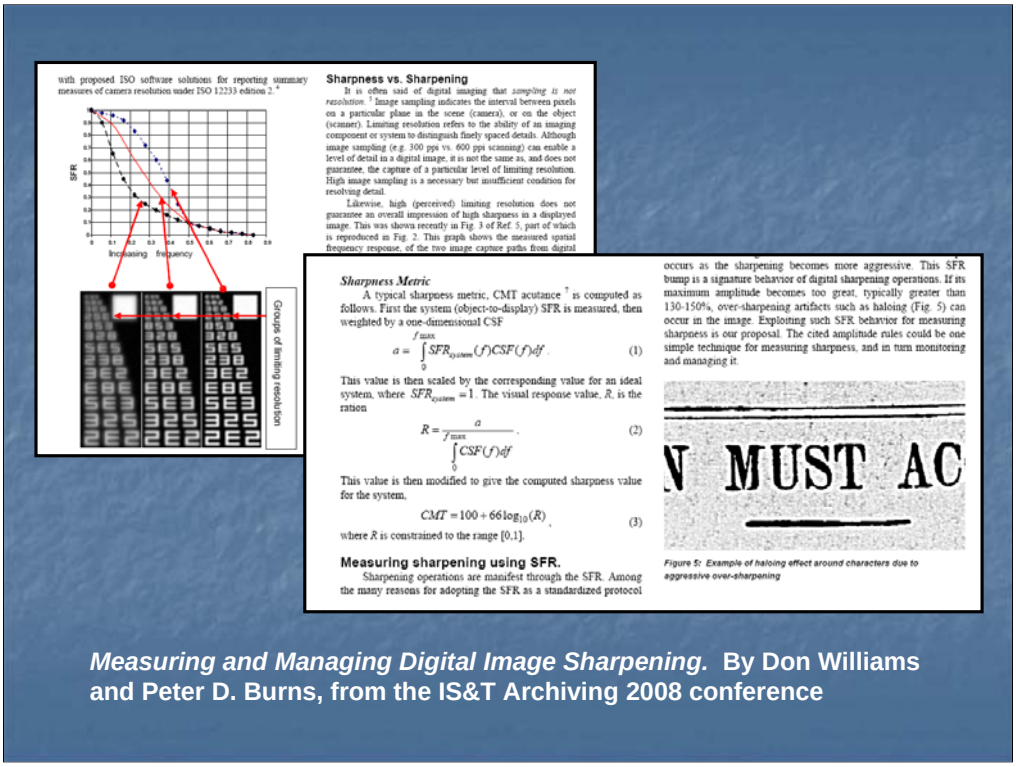
www.digitalpreservation.gov



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.

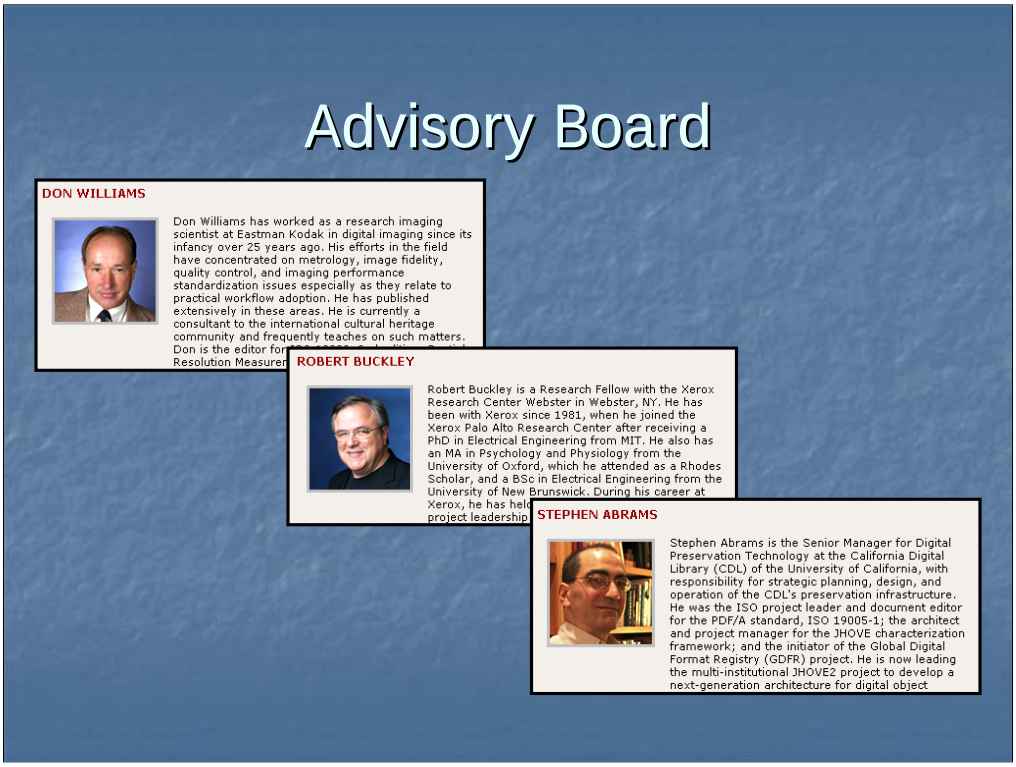
But it is very much a collaborative effort with participation from a number of federal agencies. I'll just name a few: the National Archives, the Government Printing Office, the National Gallery of Art, the Voice of America, the National Library of Medicine, the Smithsonian Institution, the Department of Defense Imagery Management Operations Center.

Our main emphasis is *digitization*--also known as digital reformatting--the conversion of analog originals into digital form. Michael Stelmach leads the still image group and they are looking at things like books, photos, and maps.



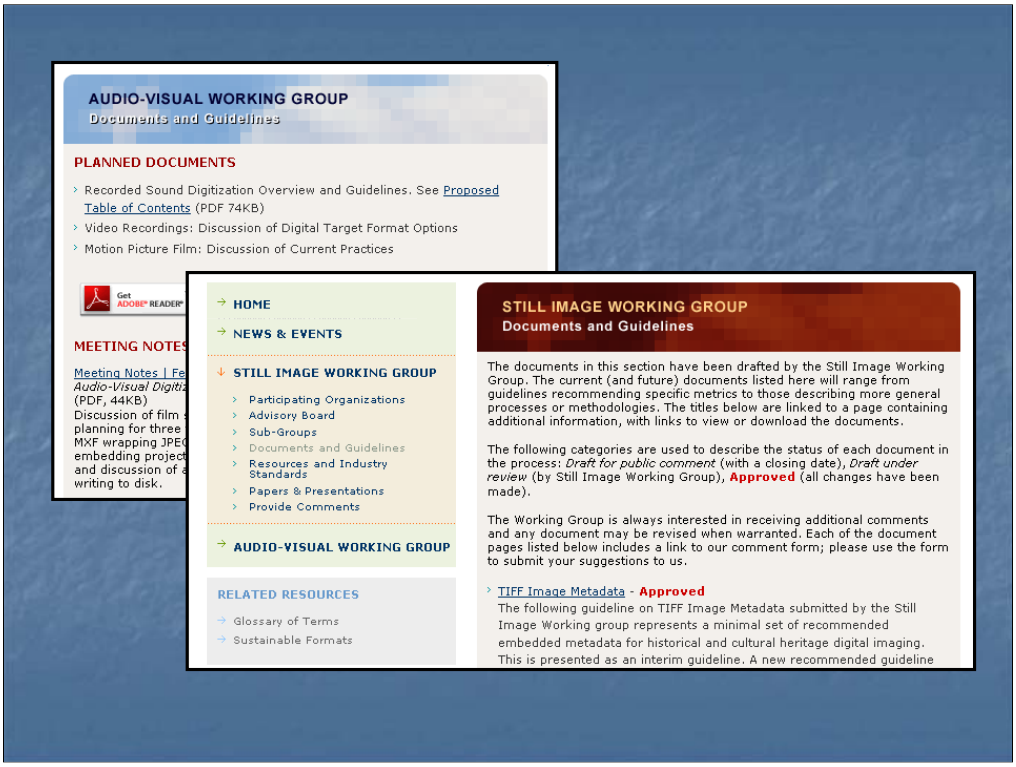
Measuring and Managing Digital Image Sharpening. **By Don Williams and Peter D. Burns, from the IS&T Archiving 2008 conference**

Digital still image reformatting is relatively mature and in consequence this group's work--aided by expert consultant Don Williams--moves in ever-more-sophisticated directions.

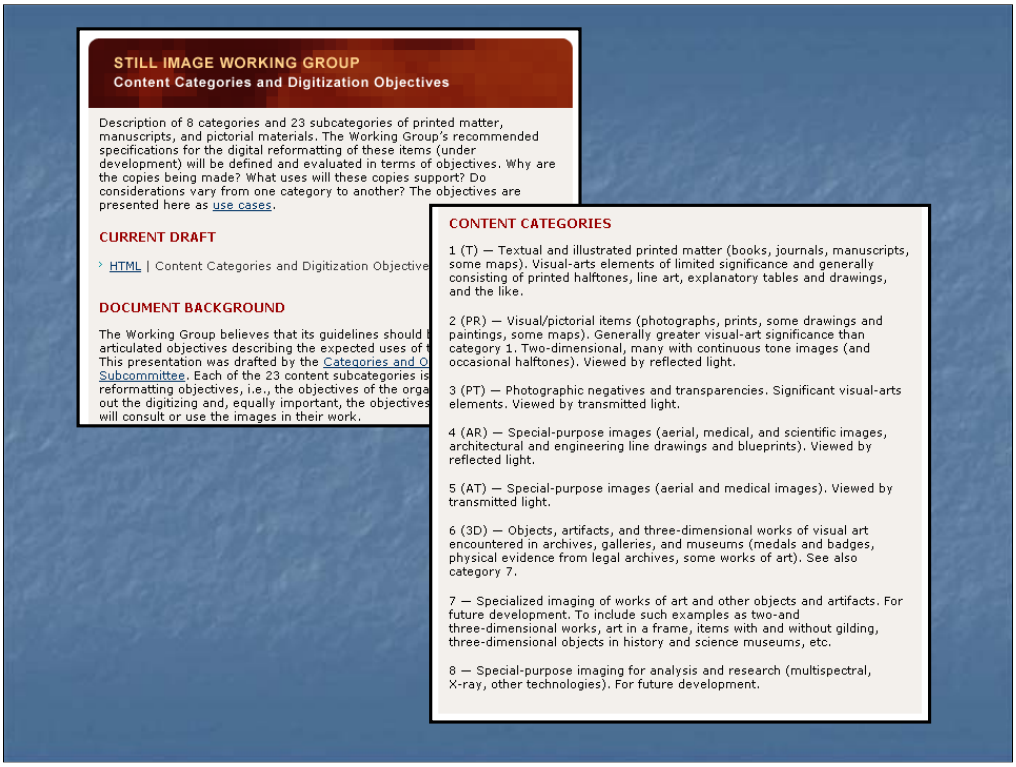


Don is one member of the group's advisory panel. The others are Rob Buckley from Xerox and Steve Abrams from the California Digital Library.

Meanwhile, I coordinate the group looking at audio and video. In terms of maturity--or “literacy” to use Don Williams’s elegant term--the reformatting of sound recording is catching up with still image digitization, while work with file-based video reformatting and motion picture scanning is just getting rolling in our cultural heritage institutions.



Both working groups will draft guidelines to support the use of comparable approaches from agency to agency, which will be especially helpful in the relationships with vendors who provide equipment and services.



We will define our recommended specifications in terms of objectives. It is not enough to just say "for preservation," as though that answered the question. The objectives will vary by category of content. We have 8 major categories . . .

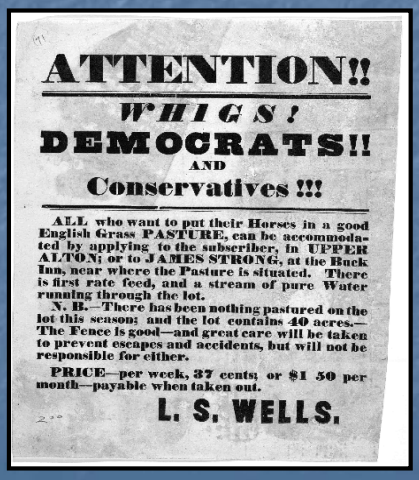
One of the subcategories

T.3. Documents with poor legibility or diffuse characters, e.g., carbon copies, Thermofax/Verifax, etc.; manuscripts or printed/typed pages with handwritten annotations or other markings; items with low inherent contrast, staining, fading, printed halftone illustrations, or included photographs.

. . . that splinter into 23 working subcategories. (We will try to collapse some of these together as we proceed.) Here's one from the still imaging side. We have benefited tremendously from the 2004 imaging guidelines from the National Archives and you may recognize this wording. "Documents with poor legibility or diffuse characters . . . items with low inherent contrast, staining, fading . . . ."

One of the subcategories

Valuation: determined by curator or end users to have informational and artifactual value, but not requiring color reproduction.



ATTENTION!!

WHIGS!

DEMOCRATS!!

AND

Conservatives !!!

ALL who want to put their Horses in a good English Grass PASTURE, can be accommodated by applying to the subscriber, in UPPER ALTON; or to JAMES STRONG, at the Buck Inn, near where the Pasture is situated. There is first rate feed, and a stream of pure Water running through the lot.

N. B.—There has been nothing pastured on the lot this season; and the lot contains 40 acres.—The Fence is good—and great care will be taken to prevent escapes and accidents, but will not be responsible for either.

PRICE—per week, 37 cents; or $1 50 per month—payable when taken out.

L. S. WELLS.

We added a new distinction: “determined by curator or end users to have informational **and** artifactual value, but **not** requiring color reproduction.” This means that we are not only interested in the words on the page but also in the “page-ness” itself. There are other subcategories for similar content with color, or when an items is of informational interest only.

Selected use case objectives for master images

- Digitizing organization (or successor/receiving agency with an archiving mission) sustains the master (or migrated copies) for the long-term without loss of essential features.

The objectives are stated as use cases and we differentiate between master images and derivative or service images. Here are a couple of for-instances for the master images. One is about longevity--the preservation angle: either this image, or one that we make in a future migration, must carry the underlying content forward over time.

Selected use case objectives for master images

- Digitizing organization uses master to produce derivative images for use cases like these:
  - (1) end-user-access interface
  - (2) other patron uses as listed
  - (3) OCR or other text-creation process
  - (4) document the condition of the original item

A second objective is that the master image is capable of supporting the production of a variety of derivative images. This conception is consistent with Don's semantics yesterday, when he said, "it's an information file, not an image-as-visual." The derivatives (not the master) are where you start thinking about output-referred and see images-as-visuals.

Selected use case objectives for derivative (service) images

- Patron sees inline image or image set in interface. Some view the complete work, a *virtual replica*.
- Patron prints images. Some require print-on-demand copy of complete work, a *physical replica*.
- Patron is confident that the content received is an authentic reproduction, also receives information on restrictions.
- Patron downloads a derivative image and, later, uses embedded metadata to identify content and determined technical provenance.

It's easier to sink your teeth into the use cases for the derivative images. Here are a few for this category, notice that we weave in the need for metadata:

-- patron examines images in the online interface

-- patron makes a hard copy

-- patron is confident that the content is authentic and receives information on rights and restrictions

-- patron consults embedded metadata

Selected use case objectives for derivative (service) images

- Publisher uses image to illustrate a book.
- Publisher uses image to illustrate a large poster.
- Exhibit designer uses image for display "mural."
- Broadcaster uses image in high-definition television program, zooming in for Ken Burns effect.

Then there are some objectives that drive quality factors:

-- publisher uses image for a high quality book illustration or for a large poster

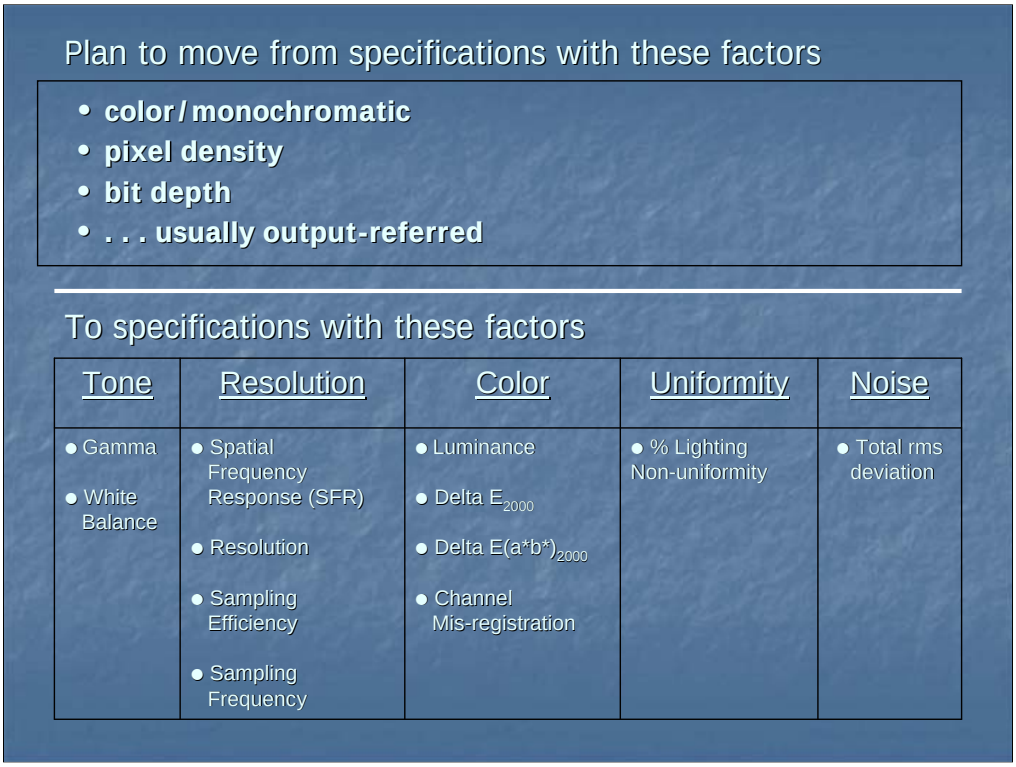
-- exhibit designer uses image for display "mural."

-- broadcaster uses image in high-definition television program

Image recommendation in 2004 guidelines from NARA

- 8-bit grayscale mode - adjust scan resolution to produce a QI of 8 for smallest significant character
  - or
- 8-bit grayscale mode - 400 ppi for documents with smallest significant character of 1.0 mm or larger
  - NOTE: Regardless of approach used, adjust scan resolution to produce a minimum pixel measurement across the long dimension of 4,000 lines for 8-bit files

The Working Group has not yet defined its specifications. The 2004 specs from the National Archives get us started with what is needed for derivative images. But a different spin will be in play when we specify the masters . . . the parents of all those derivative and migrated children.



This has brought us to consider what are the relevant quality factors. We need to move away from things like color vs. monochrome, pixel density, and bit depth, all of which are usually referenced (without saying so) in terms of outputs. Instead we want to be attentive to more appropriate ways to specify tonality, spatial resolution, color, uniformity, and noise.

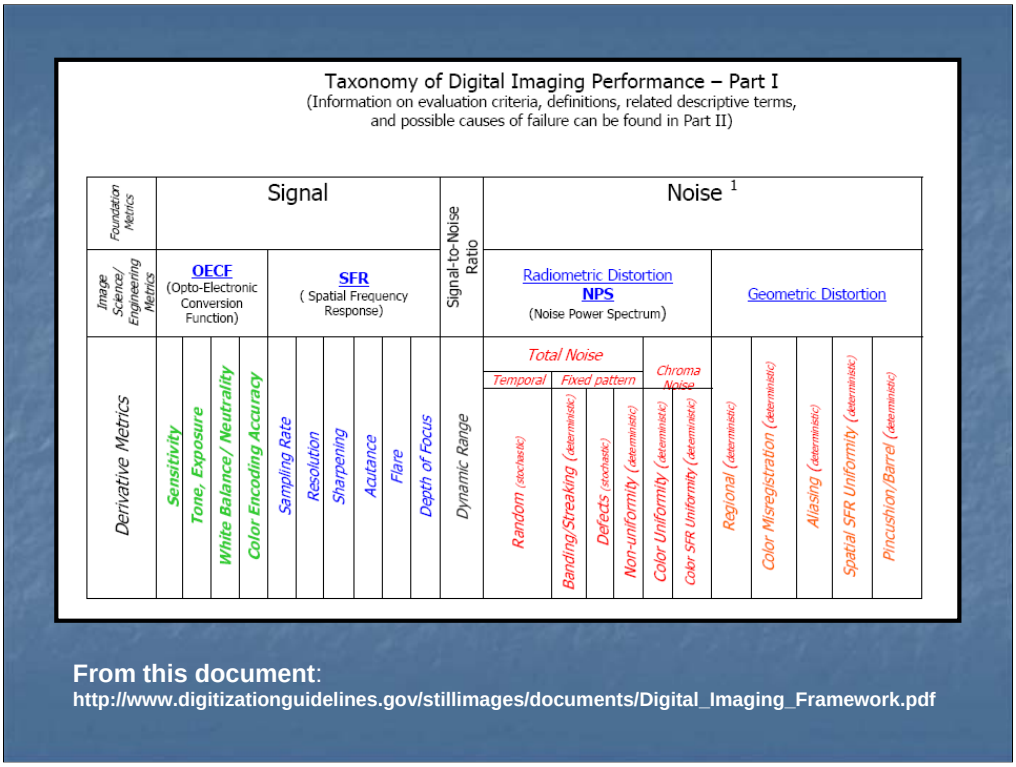
Bijlage 2. Schematisch overzicht toleranties en criteria Preservation Imaging Metamorfoze.

	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.

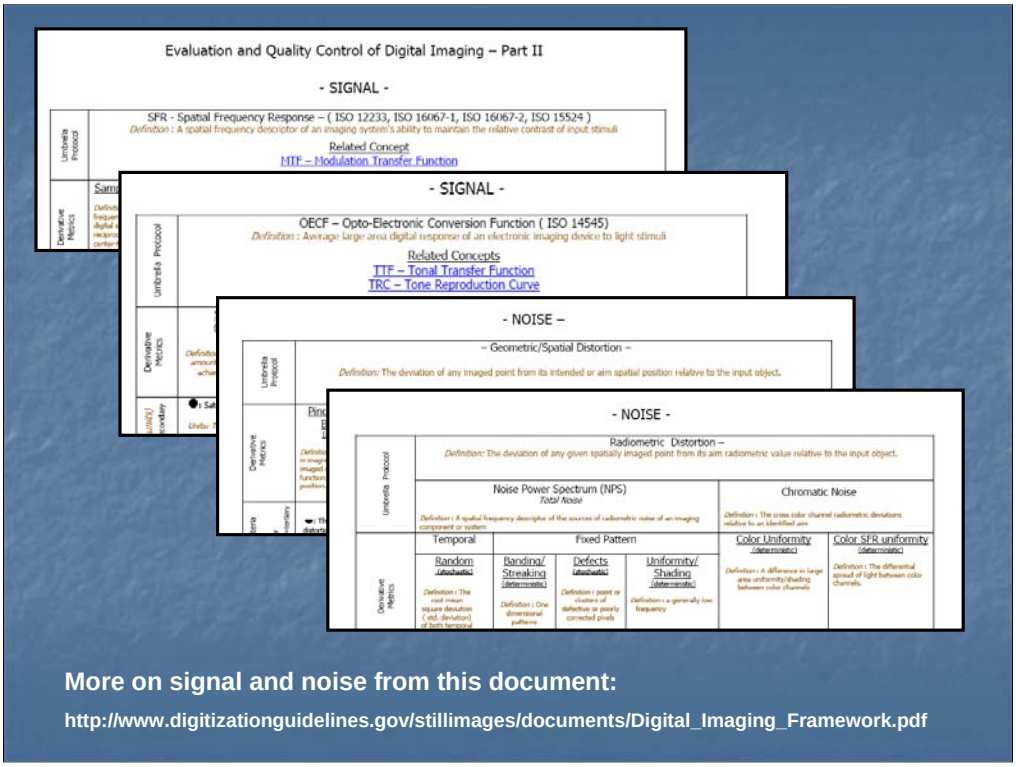
Three columns, three categories. Specifications in the various rows.

Where will the Working Group's specifications for masters end up? Too soon to tell, but here is a page from a very instructive document from the national library of the Netherlands: the columns are categories, the rows are the specs. Hans van Dormolen from the Metamorfoze project discussed this yesterday, very helpful.

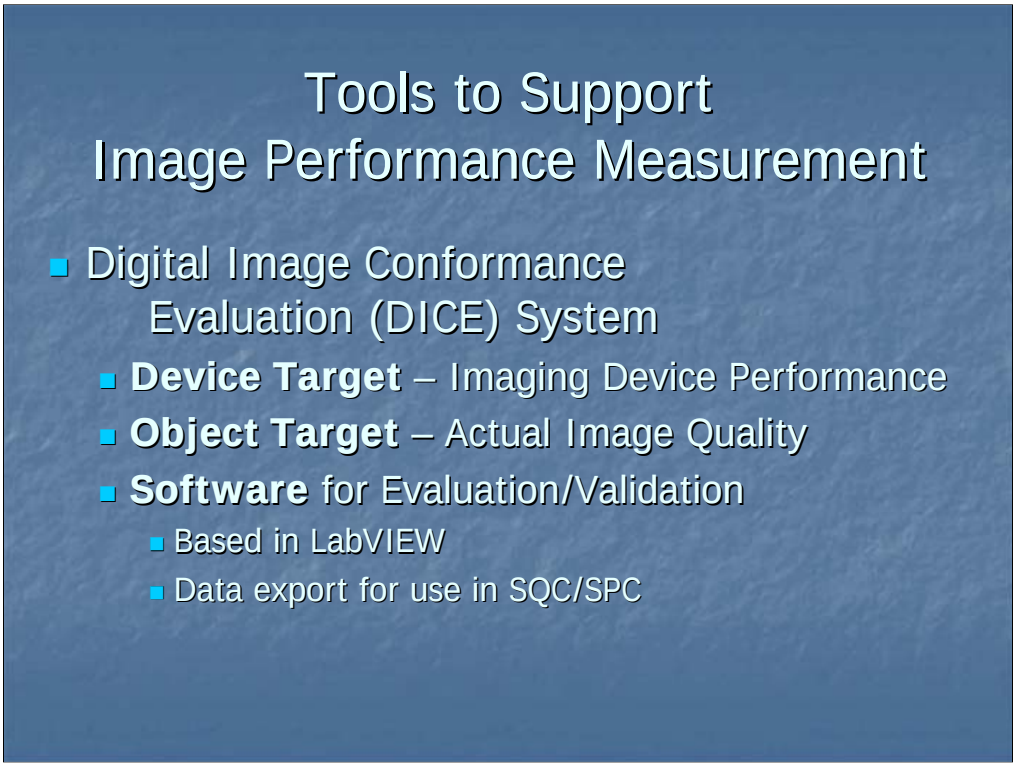


From this document:
http://www.digitizationguidelines.gov/stillimages/documents/Digital_Imaging_Framework.pdf

Don Williams helped us develop a framework to think through what is needed -- you will recognize it from his talk yesterday. Broadly speaking, it's all about *signal* and *noise*. This document has been posted online and we invite you to send us comments.

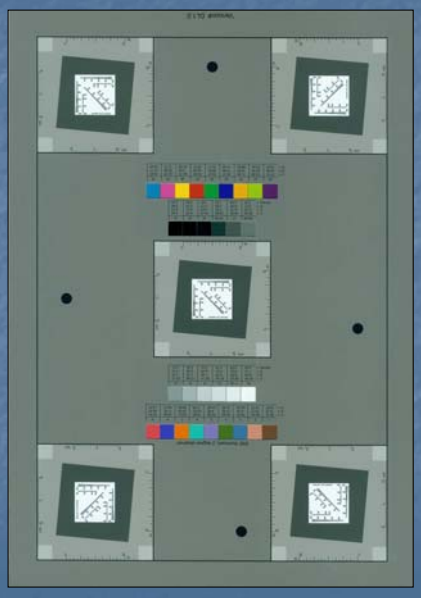


The framework describes four important measurement factors and lists relevant standards documents. For "signal," we have *spatial frequency response* or SFR and *opto-electronic conversion function* or OECF. For noise, there are several elements under the heading *geometric and spatial distortion* and another set of elements under *radiometric distortion*.



How will you know if your equipment and the work it produces conforms to the recommendations? Don has also been helping us with tools, including a pair of targets and supporting software. (It has been interesting to see the keen interest in targets at this conference. There is increased literacy in the broader community for sure.)

Device and Object Targets

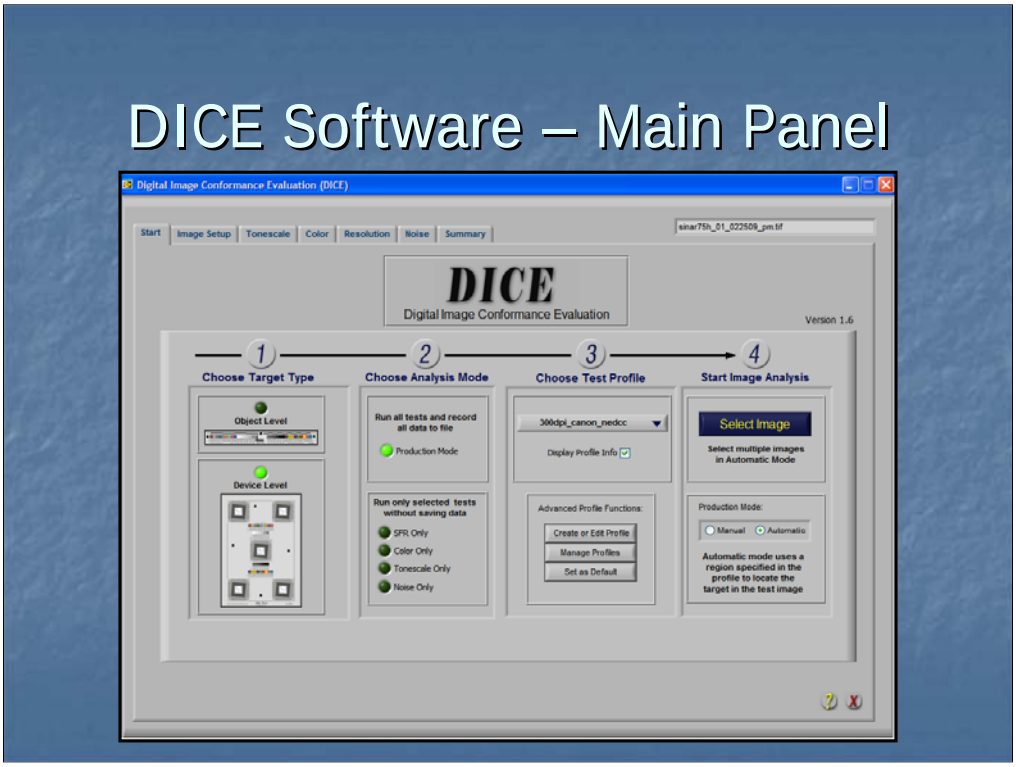


Object target as positioned for use

Thanks to OCLC for help with this part of the effort.

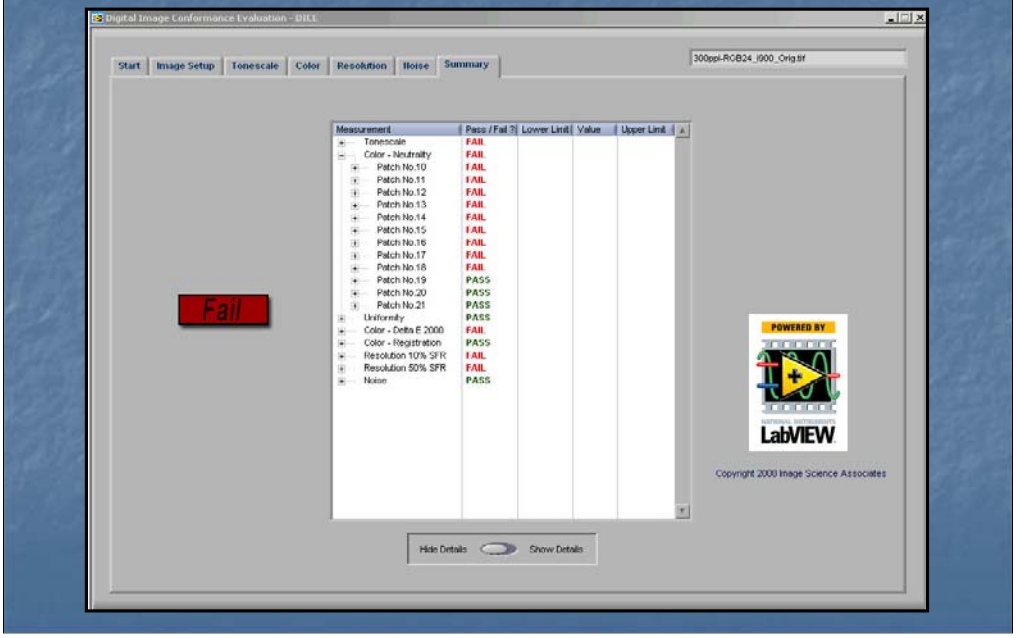
There's a big target for the device and small skinny one, to be imaged alongside the item being reformatted. They are both about ten inches long.

DICE Software – Main Panel

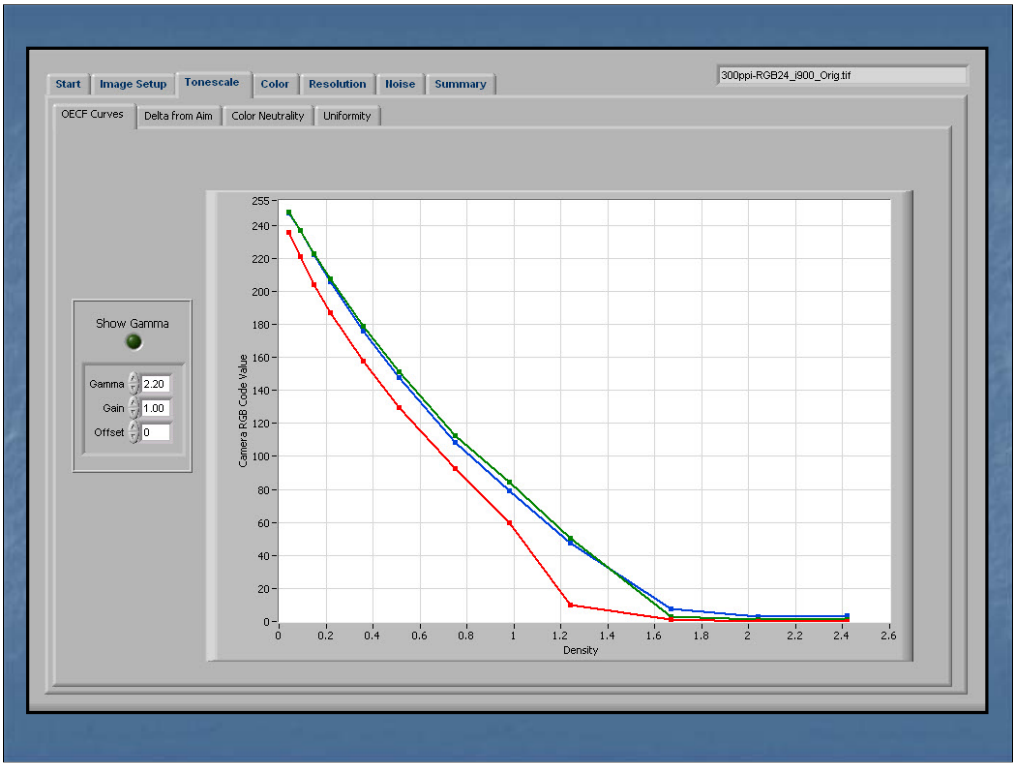


The images of the targets are analyzed by software built on top of LabVIEW, from National Instruments. Our custom package is still in an early beta stage but, when finished, will offer good functionality with an easy-to-use interface. For example, panel number 3 lets you build a job profile, setting the parameters for your current project, what we call “aim points.”

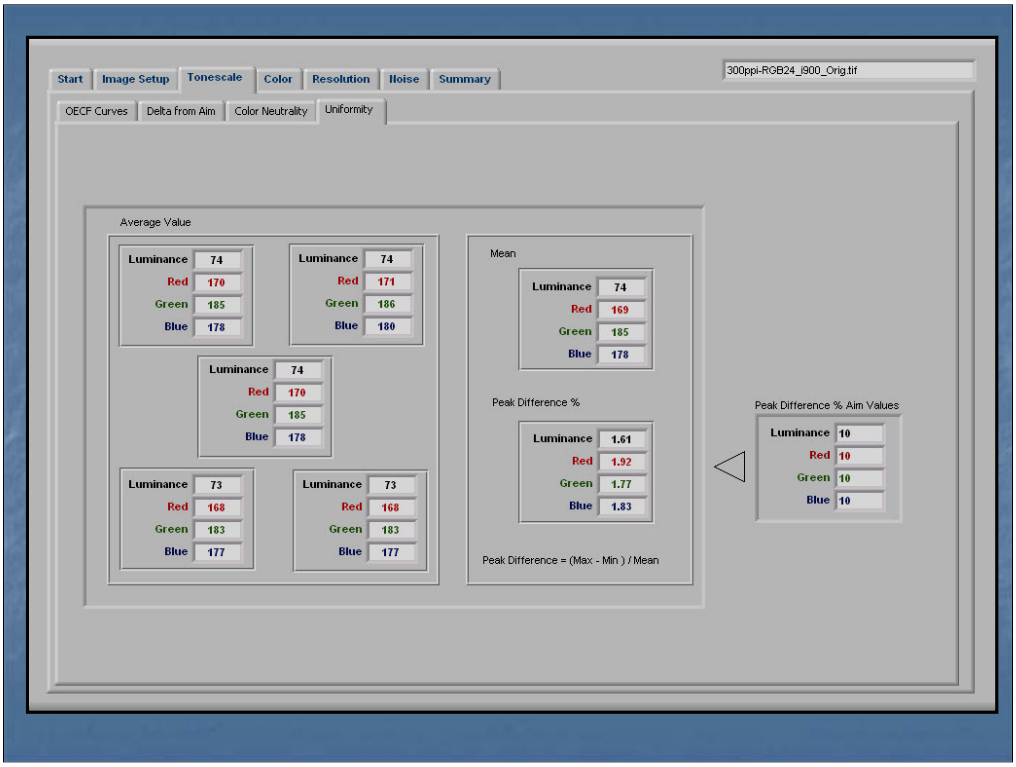
DICE – QC Summary Panel



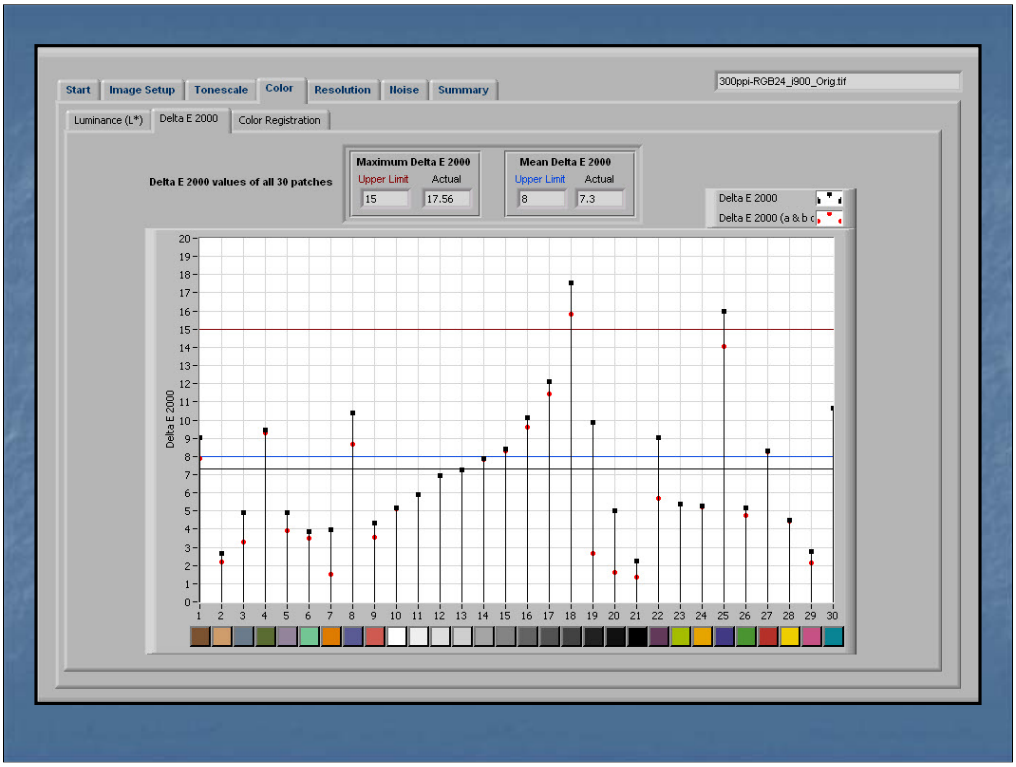
For many workers, the only display of results that they need is the one that tells them that their system passed or--in this case--failed.



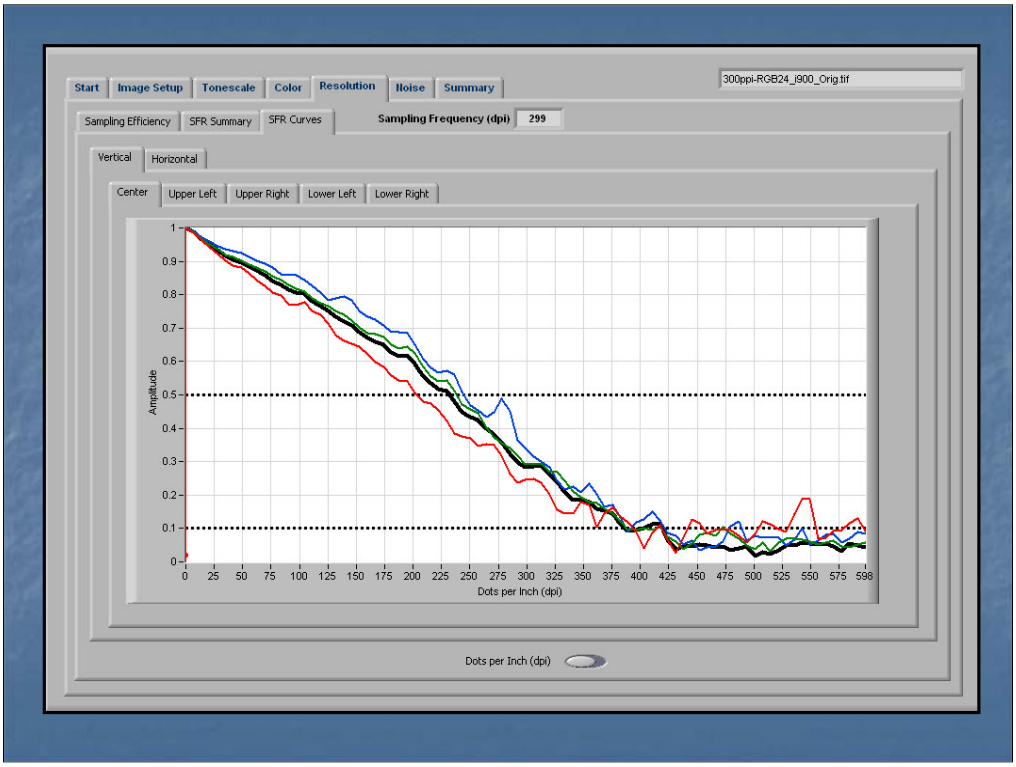
For workers who wish to know more, the software reports details. This display shows the OECF curves, part of what we call “tonescale” measurements.



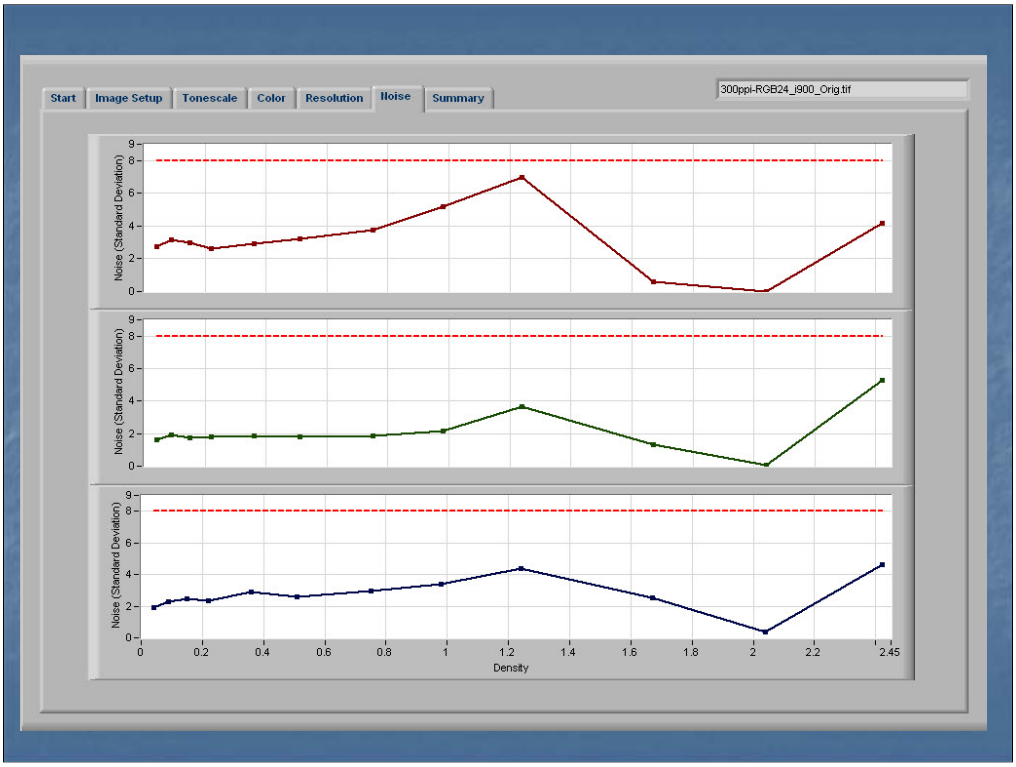
The tabs at the top tell you where you are -- in this case, still in tonescale, this time looking at uniformity.



This display, under color, provides Delta E or color difference information for the calibrated color patches on the target. The horizontal lines show the “aim points” from your profile. In this case, many values lie outside the profile zone; this is one reason that this sample failed.



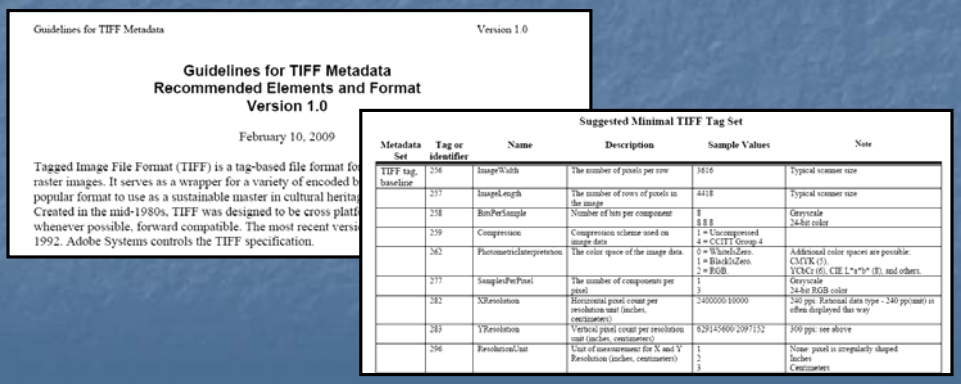
This one is for resolution, and shows the SFR curve. The upper horizontal line marks the point where the modulation transfer function drops to 50 percent, which we feel is significant for our types of images. Other image-dependent scientific fields use the 10 percent point (lower line) to determine the resolution of their imaging systems.



This display provides information about noise for red, green, and blue channels. This time the values from the target image fall below the level set for this profile--it gets a passing score on this measurement.

Beyond performance measurement

- Embedding metadata
  - TIFF header specification online now
  - Future: exploration of XMP



Guidelines for TIFF Metadata Version 1.0

Guidelines for TIFF Metadata
Recommended Elements and Format
Version 1.0

February 10, 2009

Tagged Image File Format (TIFF) is a tag-based file format fo
raster images. It serves as a wrapper for a variety of encoded b
popular format to use as a sustainable master in cultural herita
Created in the mid-1980s, TIFF was designed to be cross platf
whenever possible, forward compatible. The most recent versi
1992. Adobe Systems controls the TIFF specification.

Suggested Minimal TIFF Tag Set

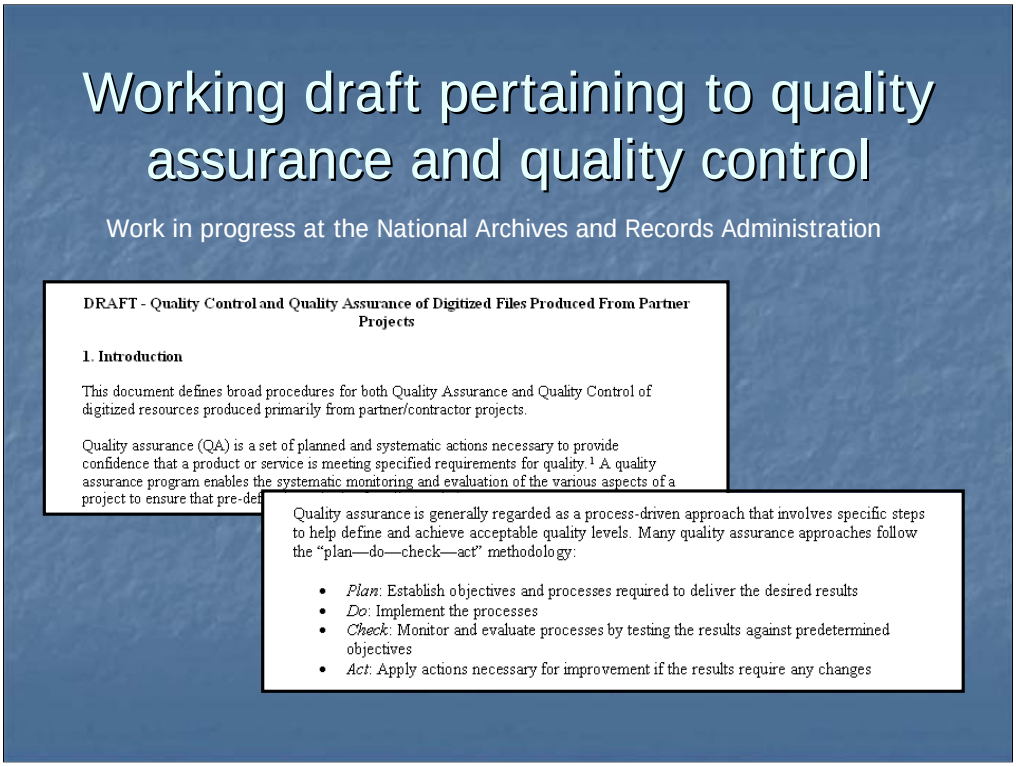
Metadata Set	Tag or identifier	Name	Description	Sample Values	Note
TIFF tag, baseline	256	ImageWidth	The number of pixels per row	3616	Typical scanner size
	257	ImageLength	The number of rows of pixels in the image	4418	Typical scanner size
	258	BitsPerSample	Number of bits per component	8 8 8 8	Grayscale 24-bit color
	259	Compression	Compression scheme used on image data	1 = Uncompressed 4 = CCITT Group 4	
	262	PhotometricInterpretation	The color space of the image data.	0 = WhiteIsZero. 1 = BlackIsZero. 2 = RGB.	Additional color spaces are possible: CMYK (5), YCbCr (6), CIE L*a*b* (8), and others.
	277	SamplesPerPixel	The number of components per pixel	1 3	Grayscale 24-bit RGB color
	282	XResolution	Horizontal pixel count per resolution unit (inches, centimeters)	2400000/10000	240 ppi. Rational data type - 240 ppi(unit) is often displayed this way
	283	YResolution	Vertical pixel count per resolution unit (inches, centimeters)	629145600/2097152	300 ppi. see above
	296	ResolutionUnit	Unit of measurement for X and Y Resolution (inches, centimeters)	1 2 3	None: pixel is irregularly shaped Inches Centimeters

Imaging performance is not the only game for the still image Working Group. Another activity concerns the metadata that ought to be embedded in image files. We have published a recommendation for TIFF headers--how to make the most of this rather thin opportunity--even as we begin an effort to explore the ways in which we might make use of XMP in our application.

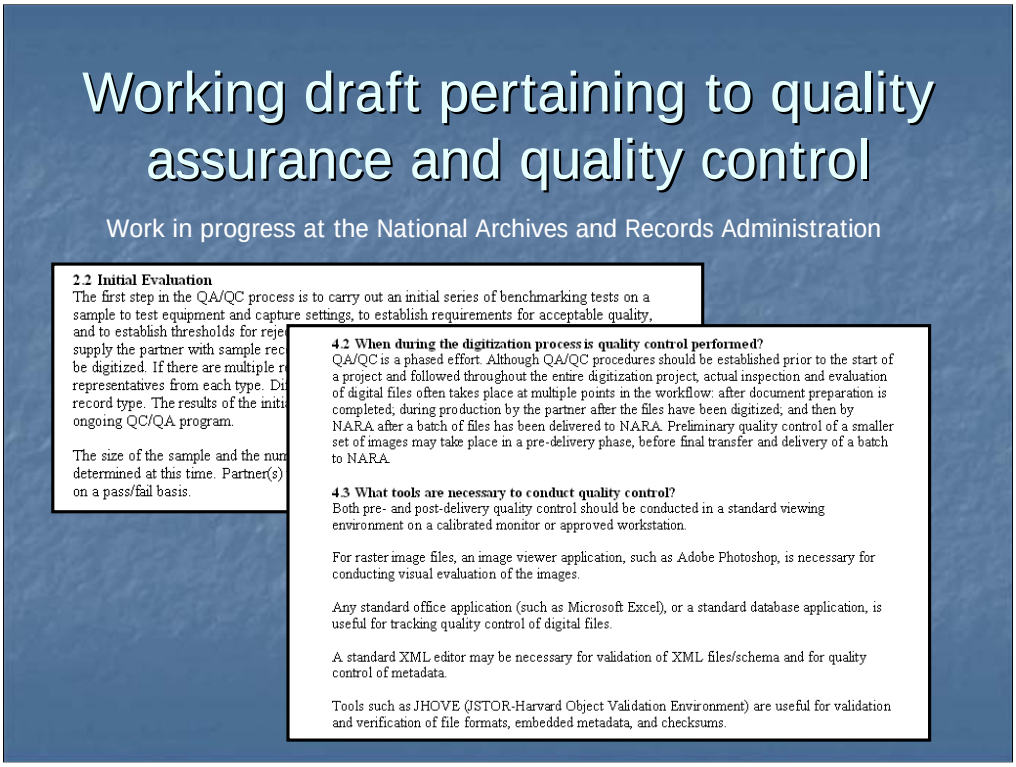
Beyond performance measurement

- Other “gaps” in prior guidelines to be investigated
  - Image Sharpening
  - Quality Management
  - Image Specification Metric Aims and Limits
  - Foldouts and Inserts in Bound Materials
  - Color Encoding Accuracy
  - Color Space Encoding
  - Selection Criteria for Master Image File Format

There were other “gaps” that we spotted when we reviewed existing guidelines. Here are a few of them, and we will pursue these as possible during the next year or two.



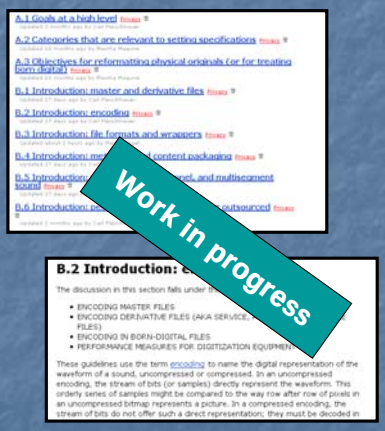
One area for additional development concerns quality assurance and quality review. Here the National Archives has taken the lead. The agency has entered into a number of partnerships with private sector partners. In order to be sure that the digital copies NARA receives from their partners meet appropriate standards, the agency has drafted a quality review specification.



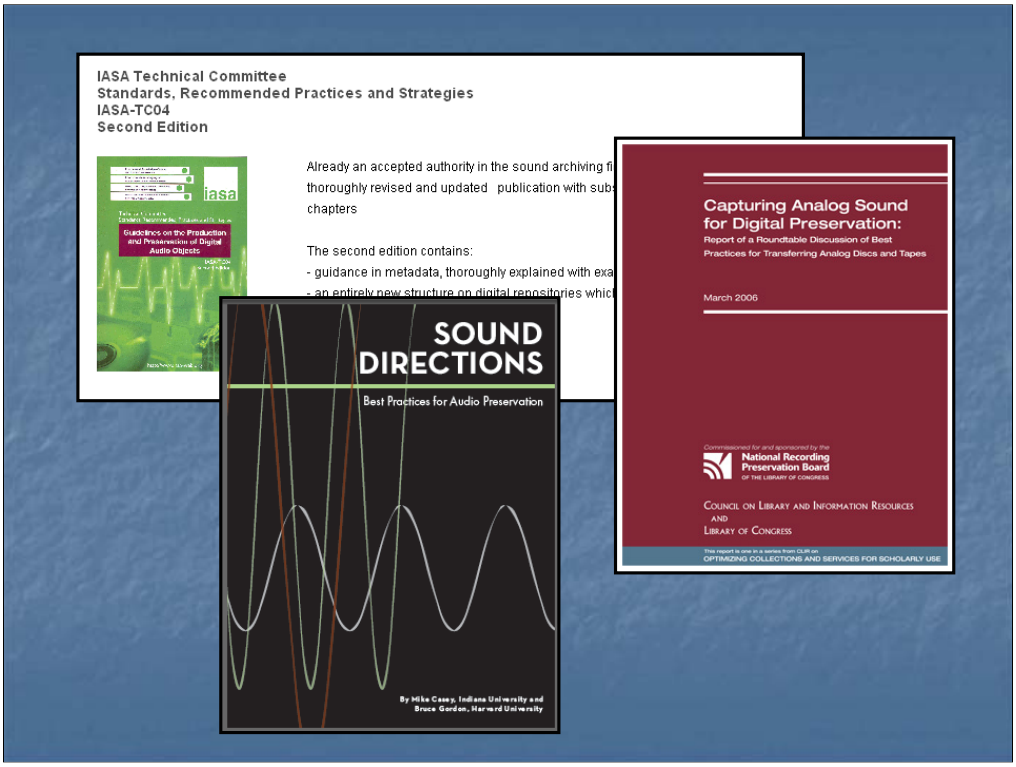
Steve Puglia is leading the NARA team on this topic, and they have shared their draft with the Working Group. This document will provide a starting point for our collective deliberations.

Quick note on audio-visual effort

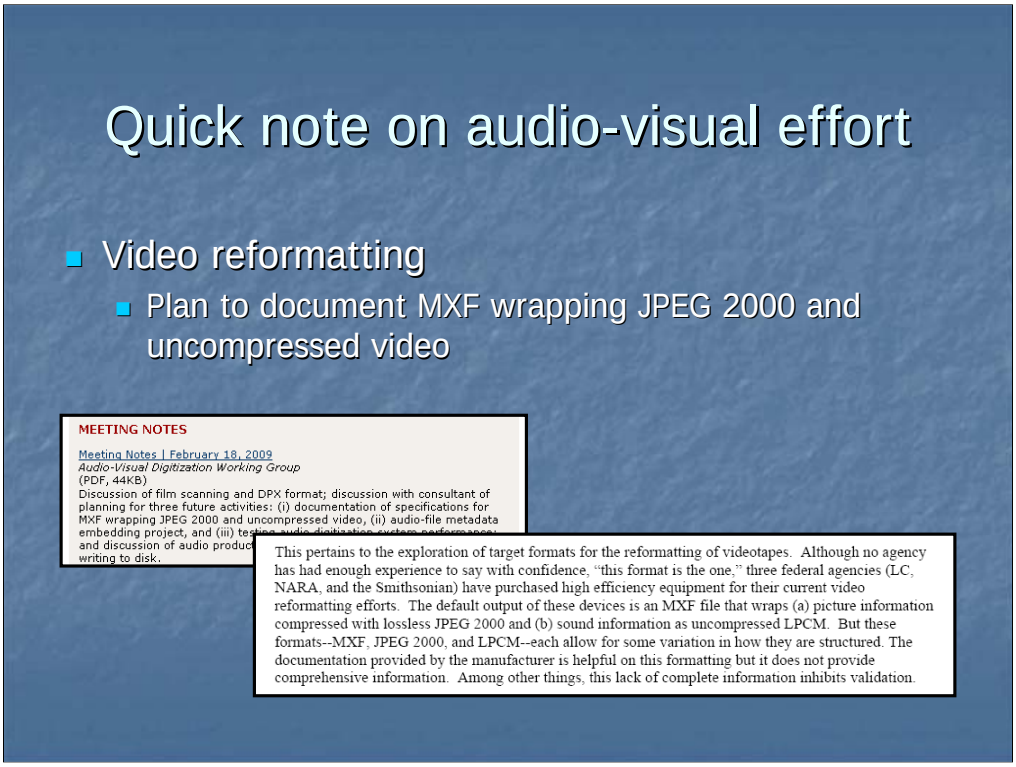
- Compile guidelines for recorded sound
  - Based on work by IASA and Sound Directions project
  - Parallels to still image effort
    - **Embedded metadata** (WAVE/BWF, the future)
    - **System performance testing**, considering use of IASA TC04 pass-fail specifications



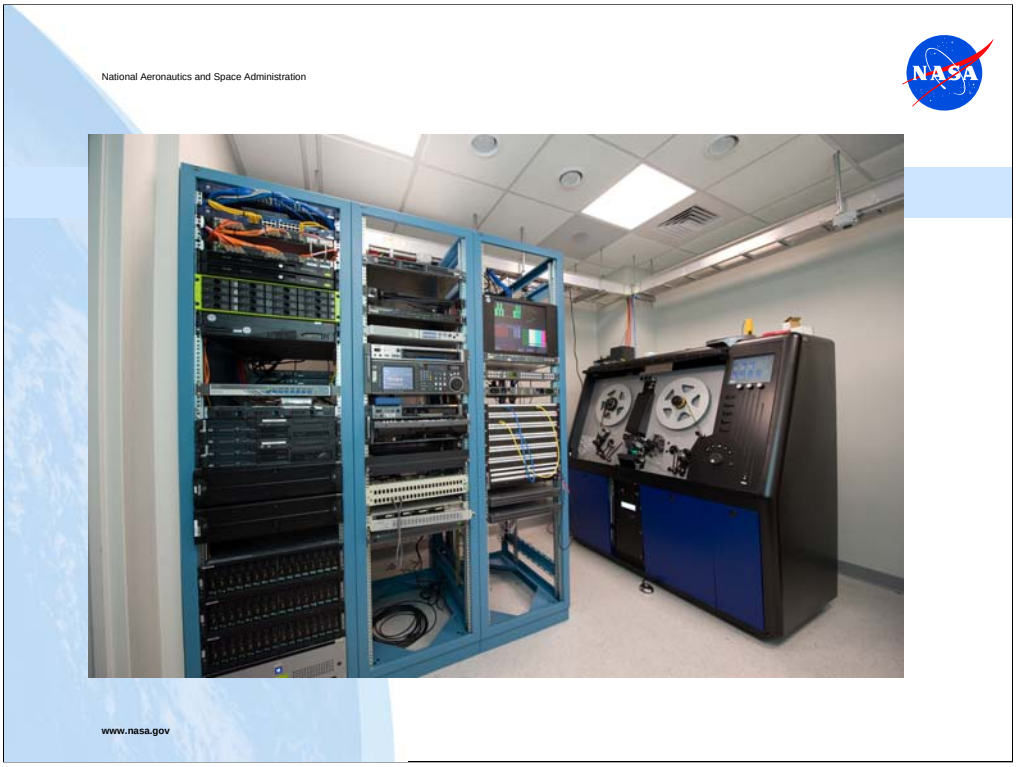
Meanwhile, on the audio-visual side, we have started to compile a guideline--as comprehensive as we can manage-- pertaining to sound recordings.



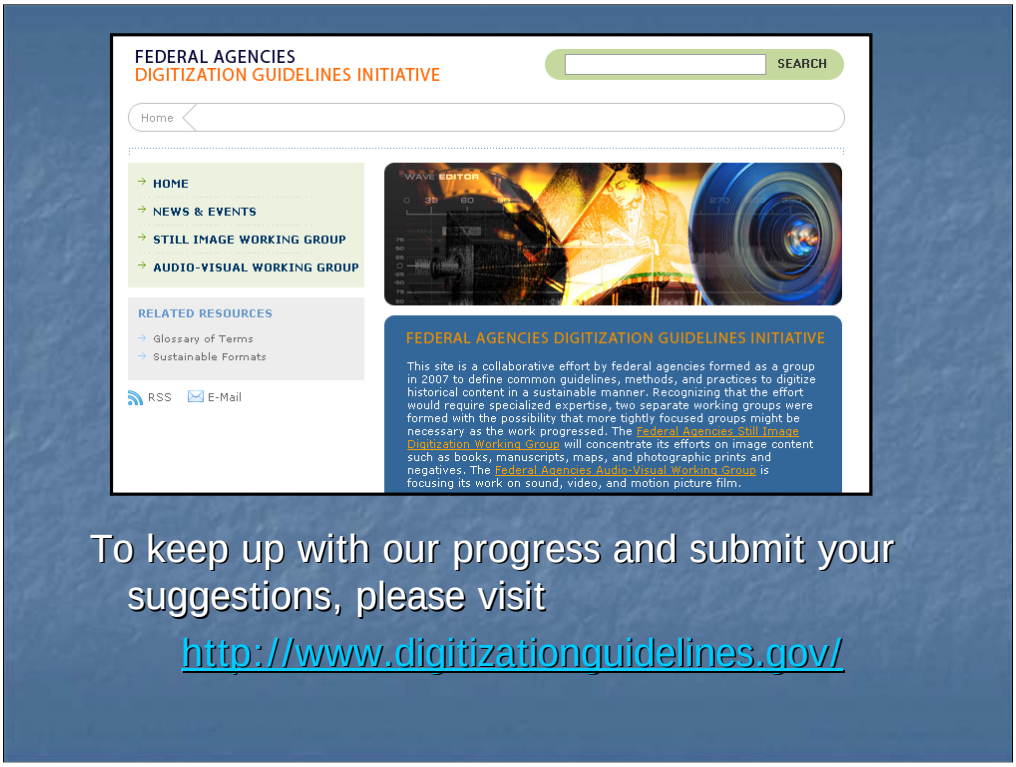
It will build on the great work produced by IASA--the International Association of Sound and Television Archives--the Sound Directions project from Indiana and Harvard Universities, and the Engineer's Roundtable organized by the National Recording Preservation Board.



It is too soon to be comprehensive on the moving image front but--for video--we want to define a profile or application specification for some of the reformatting going on already in three federal agencies. These are efforts that create MXF files, with the picture information in the form of frame images that may be losslessy compressed with JPEG 2000 or left uncompressed.



Regarding motion picture film scanning, we heard from the NASA Johnson Space Center in Houston. As far as we can tell, JSC is doing more of this than any other federal agency. Everyone who scans film seems to use the DPX format, and our Working Group discussion of the matter highlighted some issues. This area could benefit from the development of some best practices.



Altogether, there is a lot to do. Michael Stelmach and I hope to hear from you. Thank you for your interest.