Core Issues in Digital Preservation:
Text and Images

Jacob Nadal, Preservation Officer
UCLA Library

Text

- Digital text encodings have their roots in telegraph codes (really)
- ASCII (American Standard Code for Information Interchange) dates from 1968
  - 7-bit code
  - 32 control characters
  - 94 printable characters

This material has been created for the Infopeople Project [infopeople.org], and has been supported in part by the U.S. Institute of Museum and Library Services under the provisions of the Library Services and Technology Act, administered in California by the State Librarian. This material is licensed under a Creative Commons 3.0 Share & Share-Alike license. Use of this material should credit the author and funding source.
Text: UTF-8

• Unicode is an unlimited way of encoding characters
• The Unicode Transmission Format - 8 bit (UTF-8) is the most common way to encounter Unicode
  – UTF-8 transmits using 1 to 4 “octets,” 8-bit bytes
  – First 128 of these are US-ASCII, and then there are lots of other things

Text: UTF-8

• Easy to identify
  – Given an unknown text string, a simple search pattern identifies UTF-8 over 99.5% of the time
• Default, native encoding for XML
• Multi-language support

This material has been created for the Infopeople Project [infopeople.org], and has been supported in part by the U.S. Institute of Museum and Library Services under the provisions of the Library Services and Technology Act, administered in California by the State Librarian. This material is licensed under a Creative Commons 3.0 Share & Share-Alike license. Use of this material should credit the author and funding source.
Images and Text

• That unicode character set that just scrolled by was, of course, an image.
• Computers don’t read; they encode and decode
• So, digitized books are page images plus text transcriptions plus the metadata that holds all of that together.
Images

**TIFF**

- Developed by Aldus in 1986, and passed to Adobe.
- Version 6.0 published in 1992 and has no IP restrictions
- TIFF may include compressed parts; **be diligent about using uncompressed TIFF.**
  - LZW (lossless) compression debatable.

**JPEG 2000**

- Developed in 2000, released as ISO standard with a no-cost license for its core components
- Wavelet-based, so can hold several levels of compression within one file
- Shortage of authoring tools
Digital Negative

- Developed by Adobe to provide a non-proprietary format for RAW camera data
- May be valuable as a digital preservation format for the specific use-case of born-digital photography

The Other Image Formats and...

- JPEG (not JPEG2000)
- RAW (Camera sensor data)
- PNG (Portable network graphics)
- PSD (Photoshop document)

... Their Problems

- Compression or size limits (JPEG, PNG)
- Intellectual property / manufacturers proprietary standards (PSD, RAW)

This material has been created for the Infopeople Project [infopeople.org], and has been supported in part by the U.S. Institute of Museum and Library Services under the provisions of the Library Services and Technology Act, administered in California by the State Librarian. This material is licensed under a Creative Commons 3.0 Share & Share-Alike license. Use of this material should credit the author and funding source.
And then there’s PDF

- Lots of PDF types, with varying levels of preservability. Currently in version 1.7.
- PDF is (simplistically) a metadata wrapper for text and graphic content.
  - PDF can contain almost any media – raster and vector graphics, forms, audio, video, and more
- PDF 1.4 has an off-shoot called PDF/A that is used for archiving

What to put into an image

- Resolution
  - 300 dpi bare minimum, 600 dpi standard, 1200+ for special circumstances
- Bit-Depth (color)
  - 8-bit (256 grays) or 24-bit (256 Reds, 256 Greens, and 256 Blues for 16 million combinations)

Resolution

- Scanners
  - Limited by the number of sensors in the scanner’s array (top to bottom) and the motion of its motor (left to right)
- Cameras
  - Limited by physical size (H” x W”) and sensor density (pixels per inch) of the imaging chip

This material has been created for the Infopeople Project [infopeople.org], and has been supported in part by the U.S. Institute of Museum and Library Services under the provisions of the Library Services and Technology Act, administered in California by the State Librarian. This material is licensed under a Creative Commons 3.0 Share & Share-Alike license. Use of this material should credit the author and funding source.
Color

- Color needs to be calibrated
- The eye, the image sensor, and the image rendering device all have different color sensitivity
- None of these are a perfect match for the source spectra
  - And those vary depending on the type of illumination
- Best practice is to calibrate all devices and **not** edit color on the initial capture
- Create derivatives for each use-case: web delivery (RGB), high-res. display (RGB), print (CMYK), etc.

**CHROMATIC ADAPTATION**

Don't trust your eyes
Seeing and Recording and Transmission

- The eye processes light in two ways
  - Hue and saturation (color shade and depth; cones)
  - Luminance (brightness, like “black & white”; rods).
- Computers and digital imaging devices process light as three color channels: red, green, and blue
  - A fixed amount of data is assigned to each color
  - “24-bit” color has 8 bits worth of R, G, and B (256 levels each; 16.7 million combinations)
- Colors are returned as RGB (digital) or CMYK (print)
Multi-spectral imaging

- Light is radiation. Our visible spectrum ranges from 390 to 750 nanometers.
  - Immediately below (longer freq.) is infrared, which we encounter as heat, above is ultraviolet
- Under different types of radiation, media reflect, refract, fluoresce in different ways.
  - Infrared, Ultraviolet, X-radiation, Polarization, and more can produce different imaging effects
  - More image capture in more spectra means more complete digital representation
- But mostly, we just need the visible spectrum.
Starting Background Color

Note how much your eye adjusted, and how quickly.

Ending Background Color

http://www.jacobnadal.com/247

IMAGE Q&A

This material has been created for the Infopeople Project [infopeople.org], and has been supported in part by the U.S. Institute of Museum and Library Services under the provisions of the Library Services and Technology Act, administered in California by the State Librarian. This material is licensed under a Creative Commons 3.0 Share & Share-Alike license. Use of this material should credit the author and funding source.