Convert it to preserve it:
Digitization and file conversion

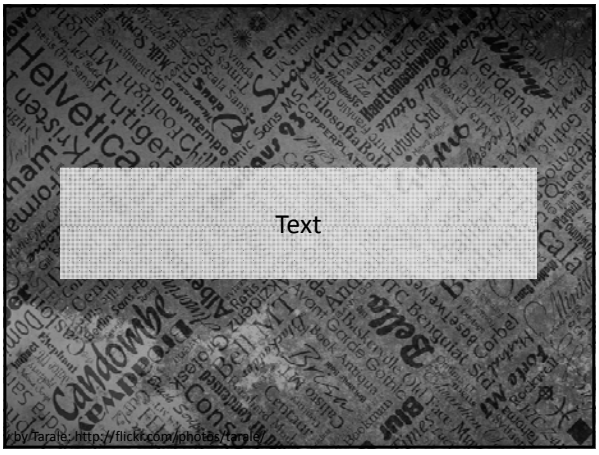
Session Outline

- Review major formats
  - Text
  - Images
  - Audio
- Brief discussion of Video, Data, and Interactive Systems
  - These lack a preservation consensus. There are many technical questions and fewer guarantees.
- For each, we'll examine
  - How the format is designed
  - What risks and advantages it entails for preservation
  - Key specifications for creating (or working with vendors to create) files in these formats

Photo: John Keogh
http://www.flickr.com/people/jvk/

Medium Definition of Digital Preservation

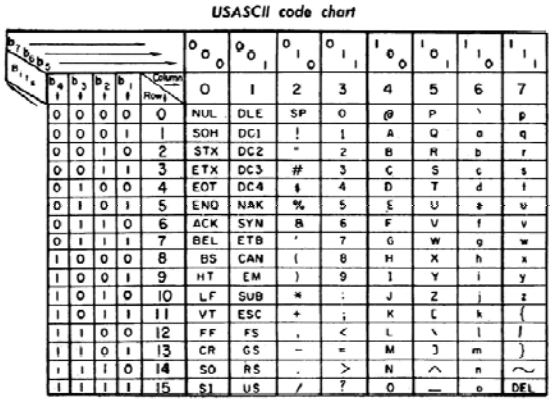
Digital preservation combines policies, strategies and actions to ensure access to **reformatted** and **born digital** content regardless of the **challenges of media failure** and **technological change**. The goal of digital preservation is the **accurate rendering** of **authenticated content** over time.



by Tarale: http://flickr.com/photos/tarale/

Text

- UTF-8, a way of representing Unicode, is standard
- Digital text is purely character data
  - No font or layout information is stored in a pure text file
  - Critical for searching and manipulation
- XML is a UTF-8 text format



USASCII code chart

Bits b7 b6 b5	0 0 0	0 0 1	0 1 0	0 1 1	1 0 0	1 0 1	1 1 0	1 1 1
b4 b3 b2 b1 / Row / Column	0	1	2	3	4	5	6	7
0 0 0 0	0 NUL	DLE	SP	0	@	P	`	p
0 0 0 1	1 SOH	DC1	!	1	A	Q	a	q
0 0 1 0	2 STX	DC2	"	2	B	R	b	r
0 0 1 1	3 ETX	DC3	#	3	C	S	c	s
0 1 0 0	4 EOT	DC4	$	4	D	T	d	t
0 1 0 1	5 ENQ	NAK	%	5	E	U	e	u
0 1 1 0	6 ACK	SYN	&	6	F	V	f	v
0 1 1 1	7 BEL	ETB	'	7	G	W	g	w
1 0 0 0	8 BS	CAN	(	8	H	X	h	x
1 0 0 1	9 HT	EM	)	9	I	Y	i	y
1 0 1 0	10 LF	SUB	*	:	J	Z	j	z
1 0 1 1	11 VT	ESC	+	;	K	[	k	{
1 1 0 0	12 FF	FS	,	<	L	\	l	|
1 1 0 1	13 CR	GS	-	=	M	]	m	}
1 1 1 0	14 SO	RS	.	>	N	^	n	~
1 1 1 1	15 SI	US	/	?	O	_	o	DEL

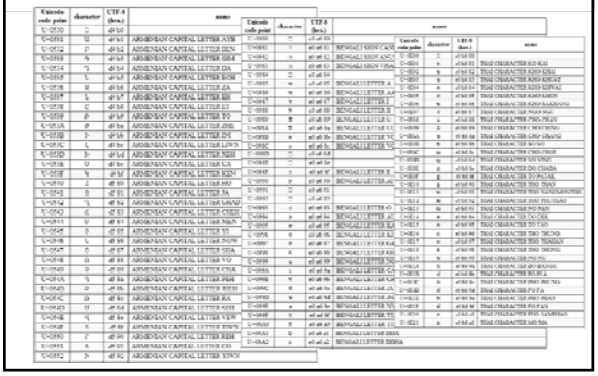
Text: UTF-8

- 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

(some of) The UTF-8 Character Set



Images and Text

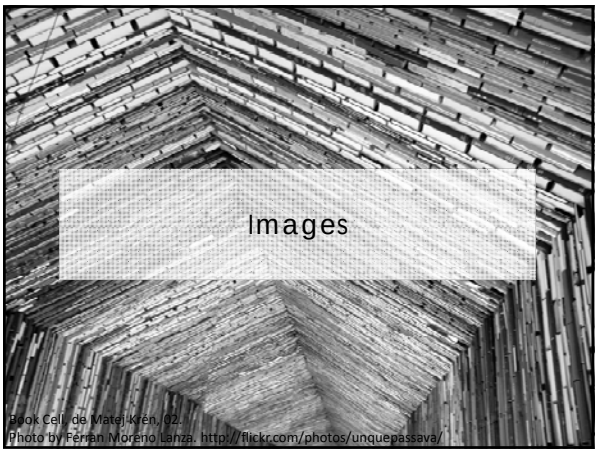
- The portions of the unicode character set that we just saw were, of course, an image.
- Computers don't read; they encode and decode
- So, digitized books are page images plus text transcriptions plus the metadata that holds all of that together.
- To get text from images, you have to re-key it or use Optical Character Recognition (OCR)
  - OCR accuracy is reported as **character-level** accuracy from ideal sources
  - Actual outcomes for accurately transcribed words from less-than-perfect sources is usually lower!

Two sides of text

- Format for building digital library systems (XML, HTML/CSS, UTF-8, PHP, etc.)
- Documents in a digital library
  - Microsoft Word: a "de facto" standard, especially with the move to Office XML in recent versions
  - PDF: a format with an open license, that can contain text, images, audio, video, forms, etc.
    - PDF/A is based on PDF 1.4, and contains a limited set of PDF features that are considered preservable

Key Specifications

- UTF-8 encoded Unicode text
- XML-based formats for markup
- Clarity about how text capture was performed (OCR or re-keying)
  - Metadata that carries the details of this!
- For documents, know your versions.
  - PDF/A, when possible
  - .docx, when possible



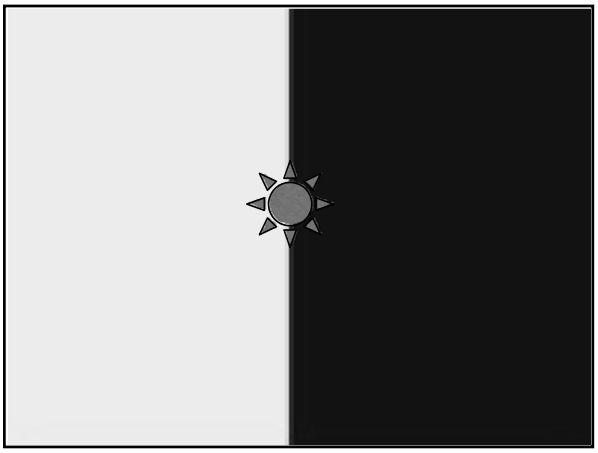
Book Cell, de Matej Kren, 02.
Photo by Ferran Moreno Lanza, http://flickr.com/photos/unquepassava/

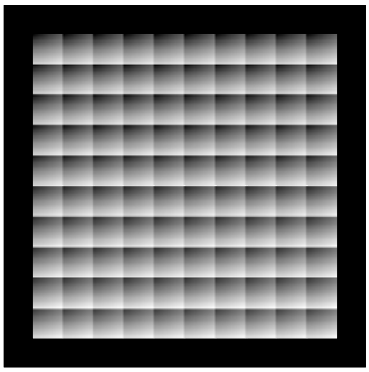
Images

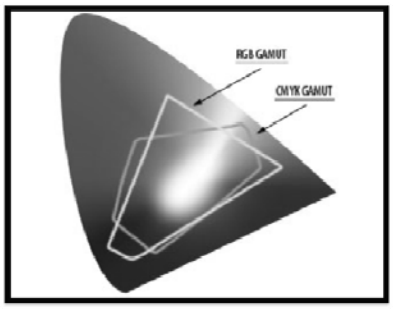
- Two types of images: Raster (which we'll talk about today) and Vector (which we won't)
- TIFF is the standard preservation format
- JPEG2000 emerging as a new alternative
- File should:
- Contain uncompressed image data (TIFF and JPEG2000 can both store compressed data)
- Be at least 300 pixels per inch (ppi/dpi), 24-bit color
  - Higher pixel count effectively allows more "zoom-in" without pixilation
- Color calibrated and profiled with an ICC color profile.

Capturing Good Images

- Set up and profile equipment, then leave it alone!
- Master should be an unaltered capture – color profiled but not "color corrected"
- Editing, retouching, and color correction should be done on a secondary copy for a particular use-case, web or print, for example.

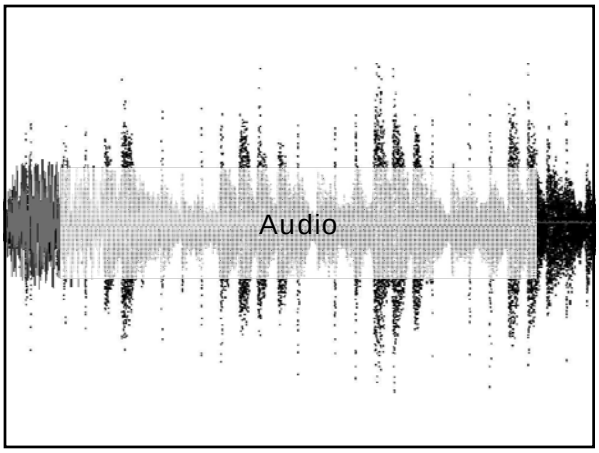






Key Specs

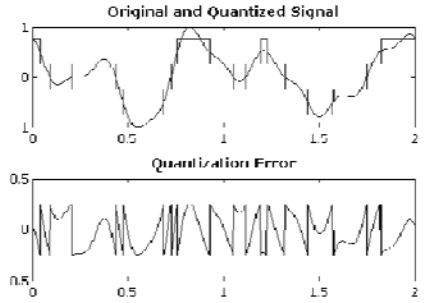
- Uncompressed, 24-bit RGB
- Color – managed (icc profile)
- Usually TIFF format
- 300+ DPI

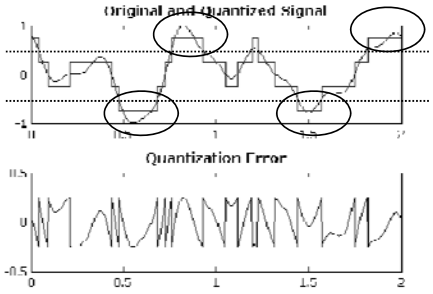


Audio

- Uncompressed Pulse Code Modulation
- Broadcast WAV (BWAV) – Wave file with a metadata header
- Resolution of at least 44.1 kHz (CD quality), preferably 96 kHz
- Bit Depth of at least 16-bit (CD quality), pref. 24-bit

Why Frequency and Depth Matter





Resolution

- Waveforms, one per channel.
  - Mono = 1, Stereo = 2, 5.1 = 6 channels
- Perhaps some metadata, in BWF especially
- CD audio is 44.1 kHz (44,100 samples per second)
- Most digital preservation engineers favor 96 kHz
  - Extra sampling capacity helps avoid errors, provides finer reproduction of sound

Bit Depth

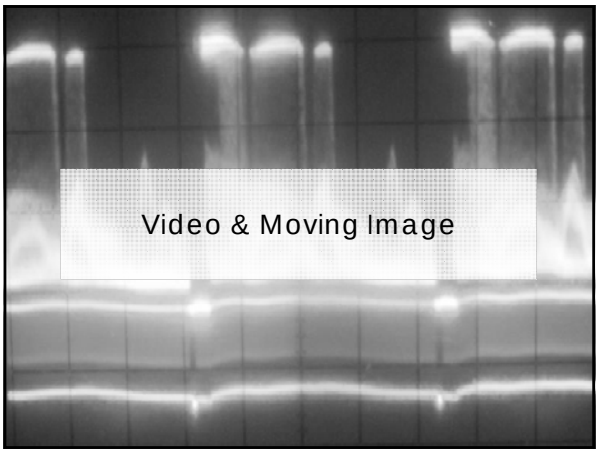
- CD audio is 16-bit, which allows up to 65,536 levels of amplitude, between 0-96dB
- 24-bit audio has a theoretical maximum of 16.7 million levels from 0-144 dB
  - Current digital audio converters are limited to ~120 dB because of practical limits on integrated circuit design
- 96khz/24-bit surpasses limits of *human* hearing
  - Some signals encode data not meant for humans

Capturing Audio

- Source: Tape, LP, Microphone, etc.
  - Compare: Photo, document, etc.
- **Digital Audio Converter (DAC): This will determine the bit depth and resolution, and basic quality of your capture**
  - Compare: Scanner, Digital Camera
- Audio Mastering/Editing Software
  - Compare: Image editing software

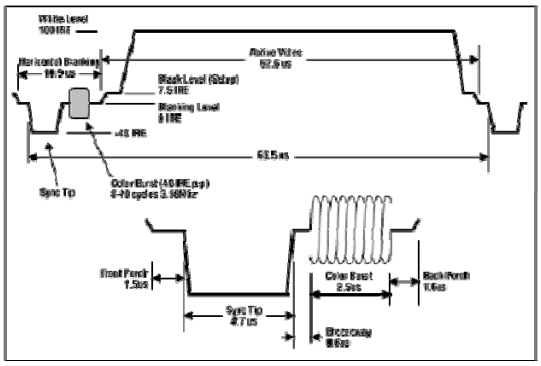
Key Specs

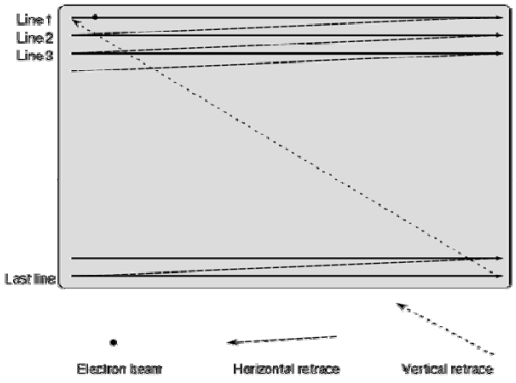
- Broadcast WAV (BWAV) – Wave file with a metadata header
  - WAV audio is Pulse Code Modulation (PCM), the universal format for uncompressed audio
- Resolution of at least 44.1 kHz (CD quality), preferably 96 kHz
- Bit Depth of at least 16-bit (CD quality), pref. 24-bit



Two different sources

- Motion picture is a series of optical image frames with a sound track (usually optical)
- Video is a series of magnetically recorded signals as waveforms for image and sounds
- Video has a specific resolution derived from a fixed number of scan lines
- 720x480i60 from 486 scan lines is SECAM standard
  - 720x480 are picture dimensions
  - i60 indicates interlacing
  - 6 lines for (graphical, not textual) metadata



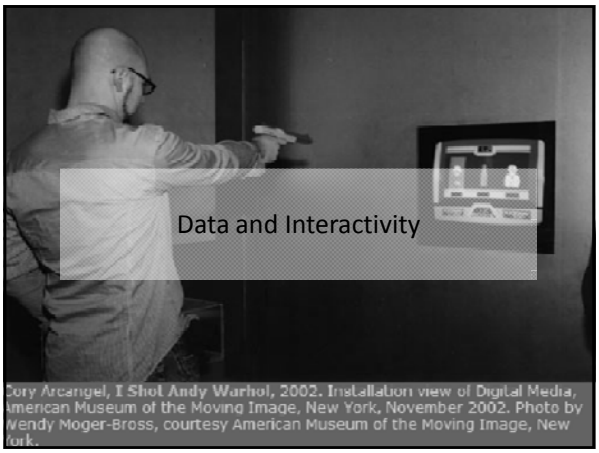


Video & Moving Image

- Standards and practices developing
  - Uncompressed desirable, but high storage costs
  - Compression is normal in video, but may cause preservation problems
- Uncompressed .AVI is the current safe bet
  - Motion JP2K & MPEG21 may be options
  - H.264 becoming the standard for service copies
- Pick one, but plan on a migration

Digital Video (Delivery)

- H.264 is standard
  - Often delivered via Flash Video (FLV)
- Several more-or-less proprietary options (Quicktime, Real, Windows Media)
- HTML 5 is emerging video delivery platform
- Pick one, but plan on a migration



Data and Interactivity

- Need to decide if fixed points in time are required: Are you storing an instance of data?
- Need to decide if active system is required: Are you maintaining and experience or immersive environment?
  - Or, are you doing both?
- Examples of how this affects you: Email and Social Media.
  - Email is a known but loosely defined set of standards, the use of which is tightly coupled to client application
  - Social media is "closed but free", with no cost to use, but no provision to move data to other systems, either.
- Where to learn more:
  - ICPSR: www.icpsr.umich.edu/icpsrweb
  - CDL: www.cdlib.org/services/uc3/datamanagement
  - Variable Media Network: variablemedia.net

Online Resources

- Sources of standards and specifications
  - PARS Minimum Capture Guidelines (this year): http://www.ala.org/alcts/mgrps/pars
  - FAGDI (right now): http://www.digitizationguidelines.gov/
  - Google Digital Curation group: http://groups.google.com/group/digital-curation
  - DigiPres listserv (via ALA): http://lists.ala.org/wws/info/digipres
  - CDL Digitization Guidelines: http://www.cdlib.org/inside/diglib/guidelines/
  - Connecting to Collections: http://www.connectingtocollections.org/all-topics/care-for-digital-materials/
- Professional Groups:
  - AMIA: http://www.amianet.org/
  - ARSC: http://www.arsc-audio.org/
  - SIST: http://www.imaging.org/

Jacob Nadal
http://jacobnadal.com/342