

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

Defining Digital Preservation

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.

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

Webinar 2: Convert it to preserve it:

Digitization and file conversion

USASCII code chart

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

Unicode code point	decimal	UTF-8	name	Unicode code point	decimal	UTF-8	name
U+0000	0	00 00	UNICODE CHARACTER U+0000	U+0001	1	00 01	UNICODE CHARACTER U+0001
U+0002	2	00 02	UNICODE CHARACTER U+0002	U+0003	3	00 03	UNICODE CHARACTER U+0003
U+0004	4	00 04	UNICODE CHARACTER U+0004	U+0005	5	00 05	UNICODE CHARACTER U+0005
U+0006	6	00 06	UNICODE CHARACTER U+0006	U+0007	7	00 07	UNICODE CHARACTER U+0007
U+0008	8	00 08	UNICODE CHARACTER U+0008	U+0009	9	00 09	UNICODE CHARACTER U+0009
U+000A	10	00 0A	UNICODE CHARACTER U+000A	U+000B	11	00 0B	UNICODE CHARACTER U+000B
U+000C	12	00 0C	UNICODE CHARACTER U+000C	U+000D	13	00 0D	UNICODE CHARACTER U+000D
U+000E	14	00 0E	UNICODE CHARACTER U+000E	U+000F	15	00 0F	UNICODE CHARACTER U+000F
U+0010	16	00 10	UNICODE CHARACTER U+0010	U+0011	17	00 11	UNICODE CHARACTER U+0011
U+0012	18	00 12	UNICODE CHARACTER U+0012	U+0013	19	00 13	UNICODE CHARACTER U+0013
U+0014	20	00 14	UNICODE CHARACTER U+0014	U+0015	21	00 15	UNICODE CHARACTER U+0015
U+0016	22	00 16	UNICODE CHARACTER U+0016	U+0017	23	00 17	UNICODE CHARACTER U+0017
U+0018	24	00 18	UNICODE CHARACTER U+0018	U+0019	25	00 19	UNICODE CHARACTER U+0019
U+001A	26	00 1A	UNICODE CHARACTER U+001A	U+001B	27	00 1B	UNICODE CHARACTER U+001B
U+001C	28	00 1C	UNICODE CHARACTER U+001C	U+001D	29	00 1D	UNICODE CHARACTER U+001D
U+001E	30	00 1E	UNICODE CHARACTER U+001E	U+001F	31	00 1F	UNICODE CHARACTER U+001F
U+0020	32	00 20	UNICODE CHARACTER U+0020	U+0021	33	00 21	UNICODE CHARACTER U+0021
U+0022	34	00 22	UNICODE CHARACTER U+0022	U+0023	35	00 23	UNICODE CHARACTER U+0023
U+0024	36	00 24	UNICODE CHARACTER U+0024	U+0025	37	00 25	UNICODE CHARACTER U+0025
U+0026	38	00 26	UNICODE CHARACTER U+0026	U+0027	39	00 27	UNICODE CHARACTER U+0027
U+0028	40	00 28	UNICODE CHARACTER U+0028	U+0029	41	00 29	UNICODE CHARACTER U+0029
U+002A	42	00 2A	UNICODE CHARACTER U+002A	U+002B	43	00 2B	UNICODE CHARACTER U+002B
U+002C	44	00 2C	UNICODE CHARACTER U+002C	U+002D	45	00 2D	UNICODE CHARACTER U+002D
U+002E	46	00 2E	UNICODE CHARACTER U+002E	U+002F	47	00 2F	UNICODE CHARACTER U+002F
U+0030	48	00 30	UNICODE CHARACTER U+0030	U+0031	49	00 31	UNICODE CHARACTER U+0031
U+0032	50	00 32	UNICODE CHARACTER U+0032	U+0033	51	00 33	UNICODE CHARACTER U+0033
U+0034	52	00 34	UNICODE CHARACTER U+0034	U+0035	53	00 35	UNICODE CHARACTER U+0035
U+0036	54	00 36	UNICODE CHARACTER U+0036	U+0037	55	00 37	UNICODE CHARACTER U+0037
U+0038	56	00 38	UNICODE CHARACTER U+0038	U+0039	57	00 39	UNICODE CHARACTER U+0039
U+003A	58	00 3A	UNICODE CHARACTER U+003A	U+003B	59	00 3B	UNICODE CHARACTER U+003B
U+003C	60	00 3C	UNICODE CHARACTER U+003C	U+003D	61	00 3D	UNICODE CHARACTER U+003D
U+003E	62	00 3E	UNICODE CHARACTER U+003E	U+003F	63	00 3F	UNICODE CHARACTER U+003F
U+0040	64	00 40	UNICODE CHARACTER U+0040	U+0041	65	00 41	UNICODE CHARACTER U+0041
U+0042	66	00 42	UNICODE CHARACTER U+0042	U+0043	67	00 43	UNICODE CHARACTER U+0043
U+0044	68	00 44	UNICODE CHARACTER U+0044	U+0045	69	00 45	UNICODE CHARACTER U+0045
U+0046	70	00 46	UNICODE CHARACTER U+0046	U+0047	71	00 47	UNICODE CHARACTER U+0047
U+0048	72	00 48	UNICODE CHARACTER U+0048	U+0049	73	00 49	UNICODE CHARACTER U+0049
U+004A	74	00 4A	UNICODE CHARACTER U+004A	U+004B	75	00 4B	UNICODE CHARACTER U+004B
U+004C	76	00 4C	UNICODE CHARACTER U+004C	U+004D	77	00 4D	UNICODE CHARACTER U+004D
U+004E	78	00 4E	UNICODE CHARACTER U+004E	U+004F	79	00 4F	UNICODE CHARACTER U+004F
U+0050	80	00 50	UNICODE CHARACTER U+0050	U+0051	81	00 51	UNICODE CHARACTER U+0051
U+0052	82	00 52	UNICODE CHARACTER U+0052	U+0053	83	00 53	UNICODE CHARACTER U+0053
U+0054	84	00 54	UNICODE CHARACTER U+0054	U+0055	85	00 55	UNICODE CHARACTER U+0055
U+0056	86	00 56	UNICODE CHARACTER U+0056	U+0057	87	00 57	UNICODE CHARACTER U+0057
U+0058	88	00 58	UNICODE CHARACTER U+0058	U+0059	89	00 59	UNICODE CHARACTER U+0059
U+005A	90	00 5A	UNICODE CHARACTER U+005A	U+005B	91	00 5B	UNICODE CHARACTER U+005B
U+005C	92	00 5C	UNICODE CHARACTER U+005C	U+005D	93	00 5D	UNICODE CHARACTER U+005D
U+005E	94	00 5E	UNICODE CHARACTER U+005E	U+005F	95	00 5F	UNICODE CHARACTER U+005F
U+0060	96	00 60	UNICODE CHARACTER U+0060	U+0061	97	00 61	UNICODE CHARACTER U+0061
U+0062	98	00 62	UNICODE CHARACTER U+0062	U+0063	99	00 63	UNICODE CHARACTER U+0063
U+0064	100	00 64	UNICODE CHARACTER U+0064	U+0065	101	00 65	UNICODE CHARACTER U+0065
U+0066	102	00 66	UNICODE CHARACTER U+0066	U+0067	103	00 67	UNICODE CHARACTER U+0067
U+0068	104	00 68	UNICODE CHARACTER U+0068	U+0069	105	00 69	UNICODE CHARACTER U+0069
U+006A	106	00 6A	UNICODE CHARACTER U+006A	U+006B	107	00 6B	UNICODE CHARACTER U+006B
U+006C	108	00 6C	UNICODE CHARACTER U+006C	U+006D	109	00 6D	UNICODE CHARACTER U+006D
U+006E	110	00 6E	UNICODE CHARACTER U+006E	U+006F	111	00 6F	UNICODE CHARACTER U+006F
U+0070	112	00 70	UNICODE CHARACTER U+0070	U+0071	113	00 71	UNICODE CHARACTER U+0071
U+0072	114	00 72	UNICODE CHARACTER U+0072	U+0073	115	00 73	UNICODE CHARACTER U+0073
U+0074	116	00 74	UNICODE CHARACTER U+0074	U+0075	117	00 75	UNICODE CHARACTER U+0075
U+0076	118	00 76	UNICODE CHARACTER U+0076	U+0077	119	00 77	UNICODE CHARACTER U+0077
U+0078	120	00 78	UNICODE CHARACTER U+0078	U+0079	121	00 79	UNICODE CHARACTER U+0079
U+007A	122	00 7A	UNICODE CHARACTER U+007A	U+007B	123	00 7B	UNICODE CHARACTER U+007B
U+007C	124	00 7C	UNICODE CHARACTER U+007C	U+007D	125	00 7D	UNICODE CHARACTER U+007D
U+007E	126	00 7E	UNICODE CHARACTER U+007E	U+007F	127	00 7F	UNICODE CHARACTER U+007F
U+0080	128	00 80	UNICODE CHARACTER U+0080	U+0081	129	00 81	UNICODE CHARACTER U+0081
U+0082	130	00 82	UNICODE CHARACTER U+0082	U+0083	131	00 83	UNICODE CHARACTER U+0083
U+0084	132	00 84	UNICODE CHARACTER U+0084	U+0085	133	00 85	UNICODE CHARACTER U+0085
U+0086	134	00 86	UNICODE CHARACTER U+0086	U+0087	135	00 87	UNICODE CHARACTER U+0087
U+0088	136	00 88	UNICODE CHARACTER U+0088	U+0089	137	00 89	UNICODE CHARACTER U+0089
U+008A	138	00 8A	UNICODE CHARACTER U+008A	U+008B	139	00 8B	UNICODE CHARACTER U+008B
U+008C	140	00 8C	UNICODE CHARACTER U+008C	U+008D	141	00 8D	UNICODE CHARACTER U+008D
U+008E	142	00 8E	UNICODE CHARACTER U+008E	U+008F	143	00 8F	UNICODE CHARACTER U+008F
U+0090	144	00 90	UNICODE CHARACTER U+0090	U+0091	145	00 91	UNICODE CHARACTER U+0091
U+0092	146	00 92	UNICODE CHARACTER U+0092	U+0093	147	00 93	UNICODE CHARACTER U+0093
U+0094	148	00 94	UNICODE CHARACTER U+0094	U+0095	149	00 95	UNICODE CHARACTER U+0095
U+0096	150	00 96	UNICODE CHARACTER U+0096	U+0097	151	00 97	UNICODE CHARACTER U+0097
U+0098	152	00 98	UNICODE CHARACTER U+0098	U+0099	153	00 99	UNICODE CHARACTER U+0099
U+009A	154	00 9A	UNICODE CHARACTER U+009A	U+009B	155	00 9B	UNICODE CHARACTER U+009B
U+009C	156	00 9C	UNICODE CHARACTER U+009C	U+009D	157	00 9D	UNICODE CHARACTER U+009D
U+009E	158	00 9E	UNICODE CHARACTER U+009E	U+009F	159	00 9F	UNICODE CHARACTER U+009F

Images and Text

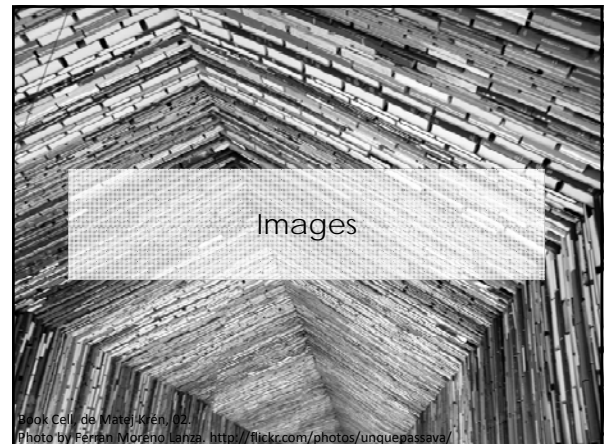
- The portions of the unicode character set that we just saw were, of course, an image.
- Computers don't read; they encode and decode
- So, digitized books are page images plus text transcriptions plus the metadata that holds all of that together.
- To get text from images, you have to re-key it or 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

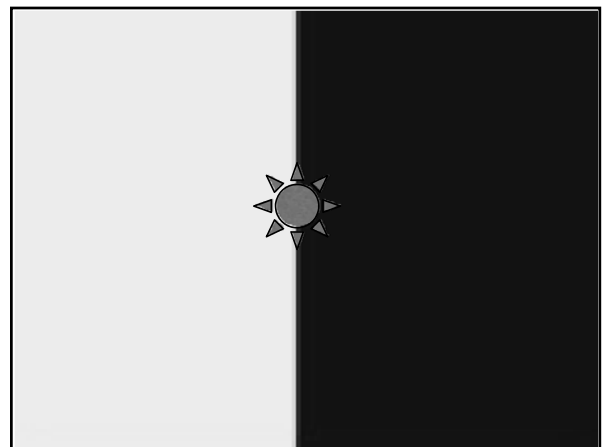


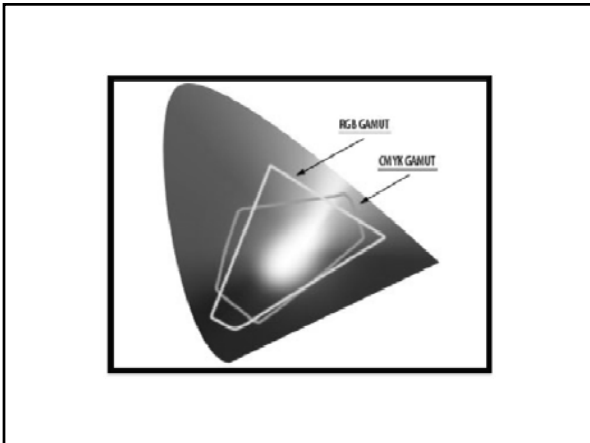
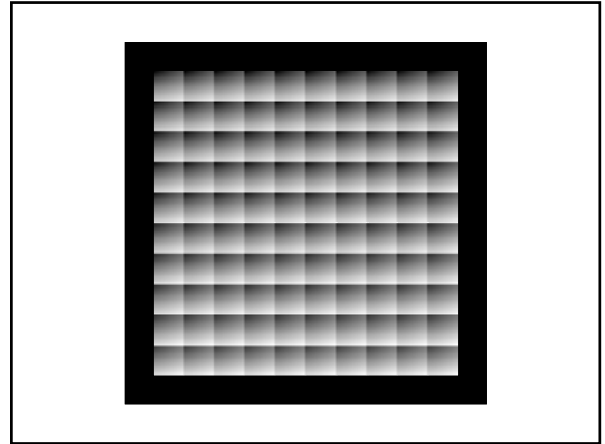
Images

- 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

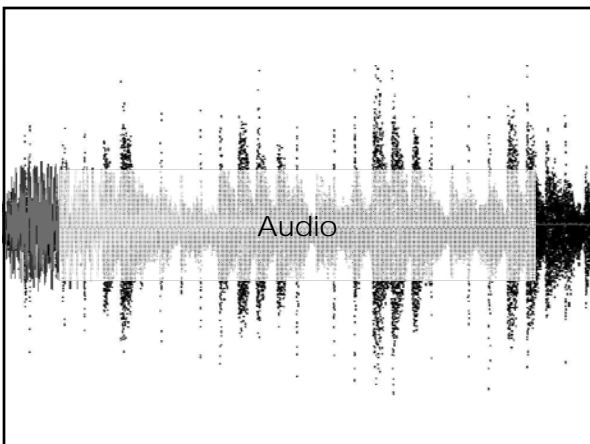
- Set up and profile equipment, then leave it alone!
- Master should be an unaltered capture – color profiled but not "color corrected"
- Editing, retouching, and color correction should be done on a secondary copy for a particular use-case, web or print, for example.





Key Specs

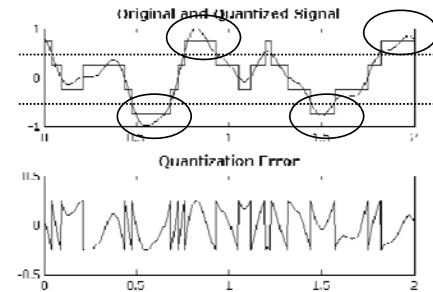
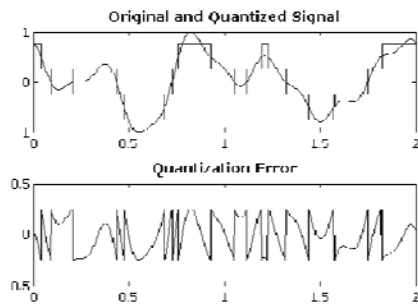
- Uncompressed, 24-bit RGB
- Color – managed (icc profile)
- Usually TIFF format
- 300+ DPI



Audio

- Uncompressed Pulse Code Modulation
- Broadcast WAV (BWA) – 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

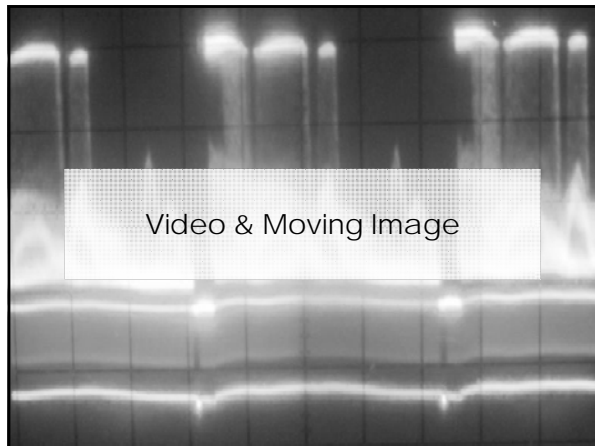
- CD audio is 16-bit, which allows up to 65,536 levels of amplitude, between 0-96dB
- 24-bit audio has a theoretical maximum of 16.7 million levels from 0-144 dB
 - Current digital audio converters are limited to ~120 dB because of practical limits on integrated circuit design
- 96kHz/24-bit surpasses 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

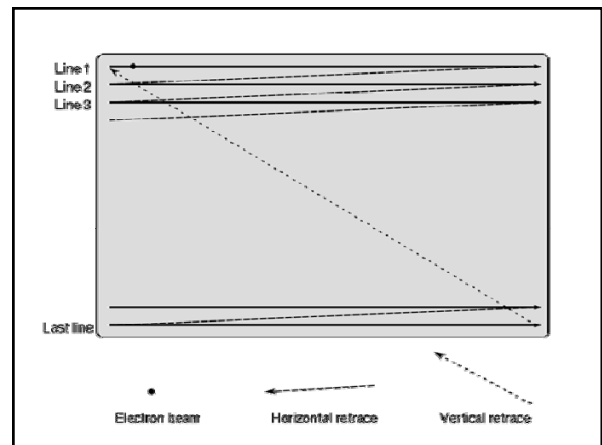
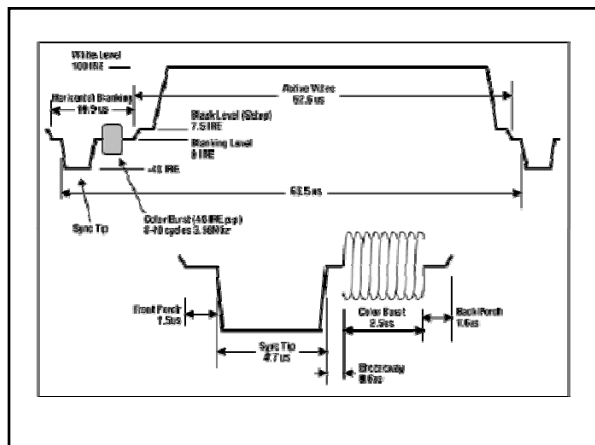
Key Specs

- Broadcast WAV (BWF) – 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

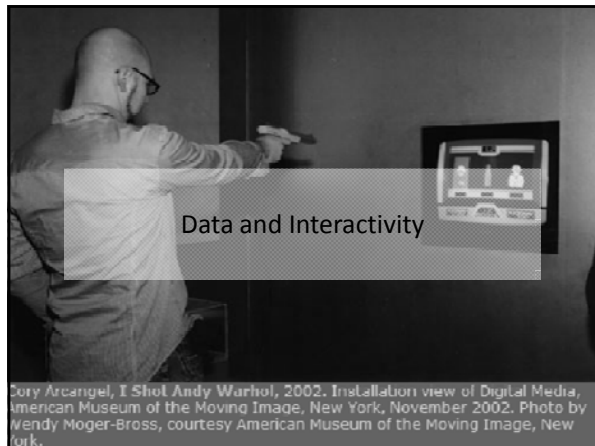


Video & Moving Image

- 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/www/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>