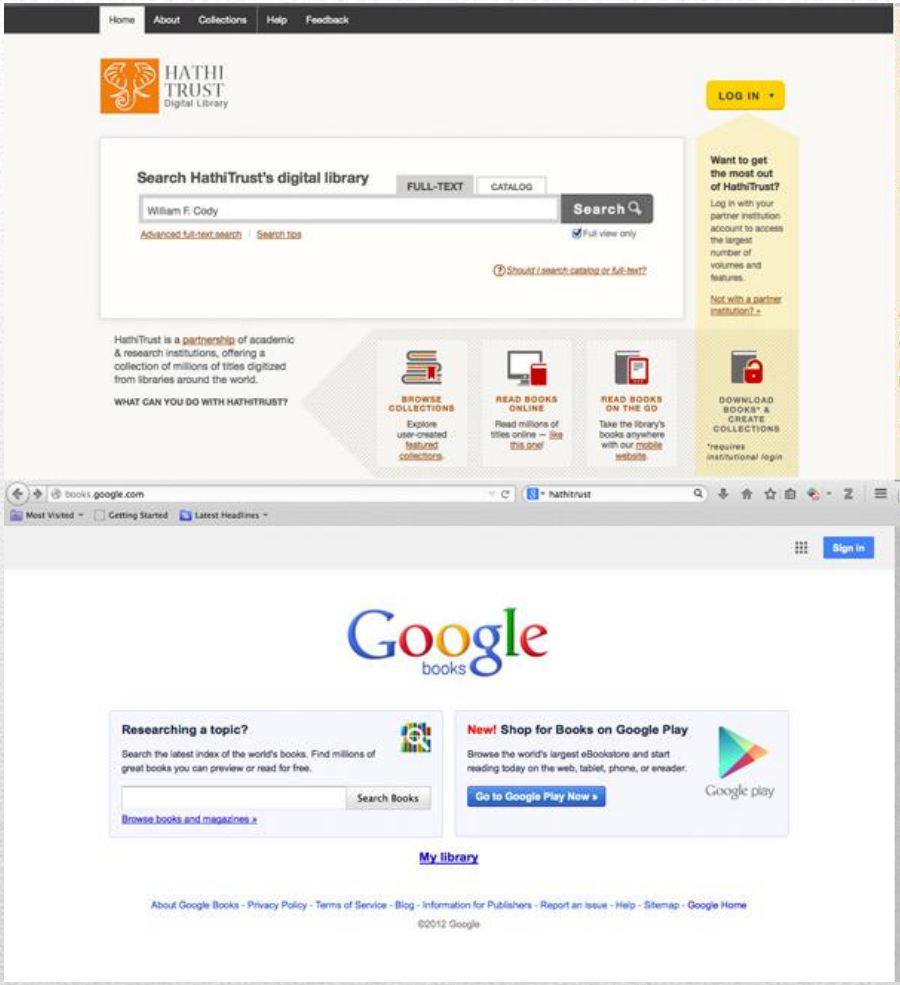
Fall 9-17-2014

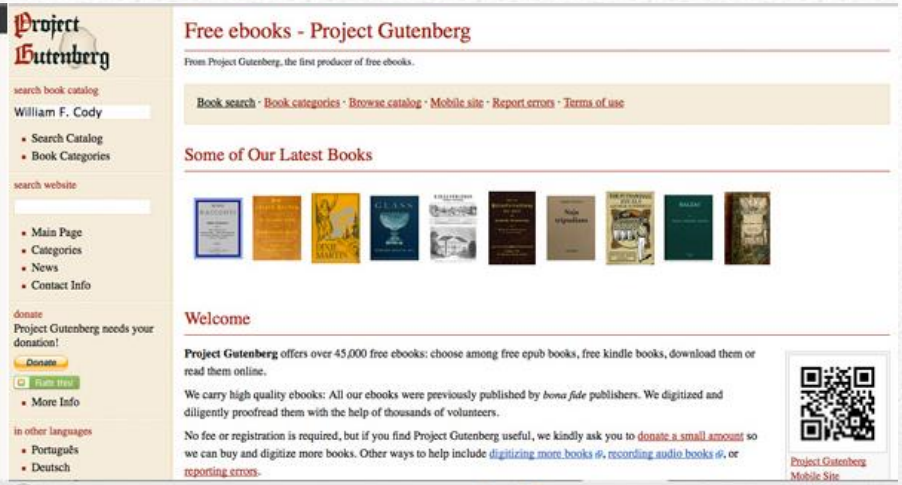
Digitization Fundamentals: Text September 17, 2014

Laura Weakly
University of Nebraska-Lincoln

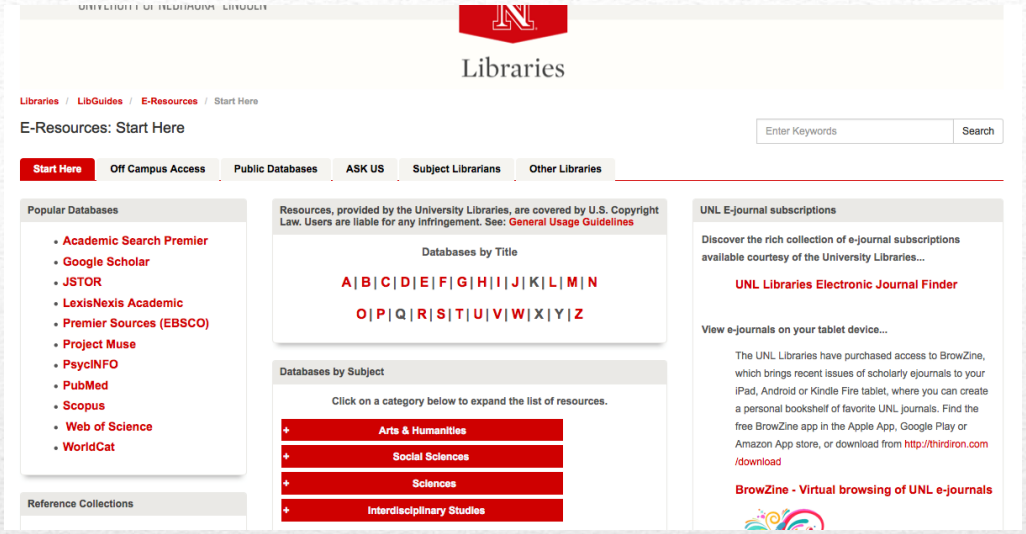
Digitization Fundamentals: Text

Laura Weakly





Do you need to digitize?



Has it already been digitized? Can you use it?

What is the source? What do you want to do with it?

Resources on campus

- Media Services (221 Love Library South) flatbed, microform, large format, 3D
- Geology Library (10 Bessie Hall)
- New Media Center (116 Architecture Hall)
- Pixel Lab (123 Henzlik Hall)

Archival Standards

- National Archives and Records Administration http://www.archives.gov/preservation/technical/guidelines.pdf
- Library of Congress http://memory.loc.gov/ammem/about/techStandards.pdf

Table 1: Summary of LoC Image Quality Standards by Document Type and Expected Outcome

Document Type	Expected Outcome	Image Parameters Standards				Notes
		Resolution	Bit Depth	Grayscale Factors	Color Accuracy	
Printed text: books w/illustrations, pamphlets, typed pages, newspapers,	Image of text	300 ppi minimum	8-bit grayscale	minimum 18 steps minimum 5.5 f-stops Y channel noise <=5%		
	OCR'ed text	400 ppi	8-bit grayscale	minimum 18 steps minimum 5.5 f-stops Y channel noise <=5%		
Music: sheet music, annotated scores, music manuscripts	Access to content	300 ppi minimum	*8-bit grayscale	minimum 18 steps minimum 5.5 f-stops Y channel noise <=5%		*24-bit color should be used where color is an important attribute of the document.
	Recognition of artifactual features	400 ppi	8-bit grayscale	minimum 18 steps minimum 5.5 f-stops Y channel noise <=5%		
Manuscripts: handwritten, typewritten copies	Access to content	300 ppi minimum	*8-bit grayscale	minimum 18 steps minimum 5.5 f-stops Y channel noise <=5%	If 24-bit color Delta-E < 8	*24-bit color should be used where color is an important attribute of the document.
	Recognition of artifactual features	400 ppi	8-bit grayscale	minimum 18 steps minimum 5.5 f-stops Y channel noise <=5%	If 24-bit color Delta-E < 8	
Maps: printed tones printed color up to D-size 22" x 34" *oversized	Content Research	250 ppi minimum	24-bit color		Delta-E < 8	*ppi is dependant on map size – particularly when map sections must be stitched together and map filesize increases to 500 MBs and more
	Map reproduction	400 ppi	24-bit color		Delta-E < 6 ICC Profile	

LOC Chart

What does this mean?

- Pixels, Resolution, Bit Depth, DPI
- File Formats, Compression

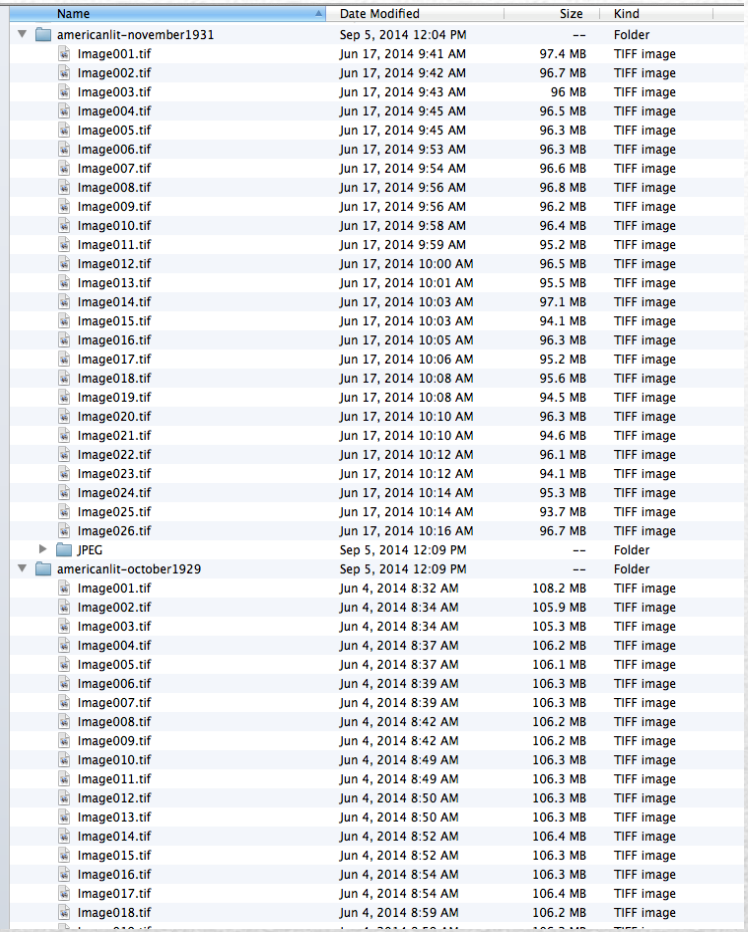
Canadian Heritage Information Network
http://www.pro.rcip-chin.gc.ca/cours-courses/fondamentales_numerisation-digitization_fundamentals/index-eng.jsp

File naming & organizing

- Think about your naming convention BEFORE digitizing
- Be descriptive, but not too descriptive
- If using dates, standardize for computer sorting YYYYMMDD
- Use leading 0s
- Rename, if necessary (Automator on Mac)

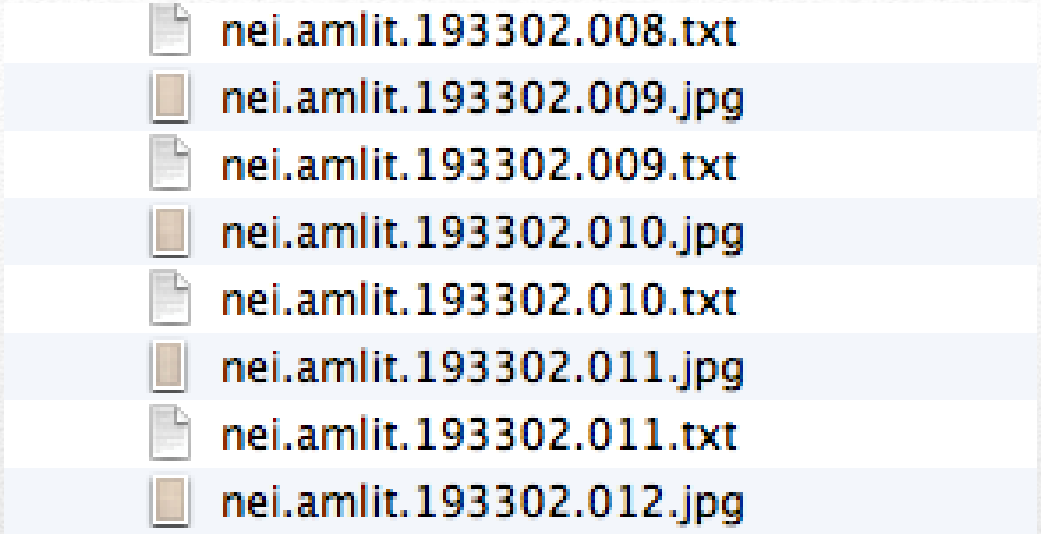
12 1918, Theodore Roosevelt Presidential Papers Microfilm, Series 3A, Reel 410, vol
:h 11 1905, Theodore Roosevelt Presidential Papers Microfilm, Series 2, Reel 337, v5
McGinty May 23 1912, Roosevelt Presidential Papers Microfilm, Series 3A, Reel 376,
Ginty Feb 25 1913, Theodore Roosevelt Presidential Papers Microfilm, Series 2, Reel 3
Ginty June 18 1912, Theodore Roosevelt Presidential Papers Microfilm, Series 3A, Ree
Ginty May 1, 1906, Theodore Roosevelt Presidential Papers Microfilm, Series 2, Reel
.L Scott Feb 26 1916, Woodrow Wilson Presidential Papers Microfilm, Series 4, Reel 3
'illiam Howard Taft Nov 4 1908, William Howard Taft Presidential Papers Microfilm, Se
t to William F Cody Nov 10 1908, William H...residential Papers Microfilm, Series 8, Re
:o TR June 14 1912, Theodore Roosevelt Presidential Papers, Series 1, Reel 146.tif
:o TR March 14 1912, Theodore Roosevelt Presidential Papers Microfilm, Series 1, Ree
:o TR May 16 1912, Theodore Roosevelt Presidential Papers Microfilm, Series 1, Reel
:o TR May 30 1906, Theodore Roosevelt Presidential Papers Microfilm, Series 1, Reel

Too much (above);
Too little (right)



Name	Date Modified	Size	Kind
americanlit-november1931	Sep 5, 2014 12:04 PM	--	Folder
Image001.tif	Jun 17, 2014 9:41 AM	97.4 MB	TIFF image
Image002.tif	Jun 17, 2014 9:42 AM	96.7 MB	TIFF image
Image003.tif	Jun 17, 2014 9:43 AM	96 MB	TIFF image
Image004.tif	Jun 17, 2014 9:45 AM	96.5 MB	TIFF image
Image005.tif	Jun 17, 2014 9:45 AM	96.3 MB	TIFF image
Image006.tif	Jun 17, 2014 9:53 AM	96.3 MB	TIFF image
Image007.tif	Jun 17, 2014 9:54 AM	96.6 MB	TIFF image
Image008.tif	Jun 17, 2014 9:56 AM	96.8 MB	TIFF image
Image009.tif	Jun 17, 2014 9:56 AM	96.2 MB	TIFF image
Image010.tif	Jun 17, 2014 9:58 AM	96.4 MB	TIFF image
Image011.tif	Jun 17, 2014 9:59 AM	95.2 MB	TIFF image
Image012.tif	Jun 17, 2014 10:00 AM	96.5 MB	TIFF image
Image013.tif	Jun 17, 2014 10:01 AM	95.5 MB	TIFF image
Image014.tif	Jun 17, 2014 10:03 AM	97.1 MB	TIFF image
Image015.tif	Jun 17, 2014 10:03 AM	94.1 MB	TIFF image
Image016.tif	Jun 17, 2014 10:05 AM	96.3 MB	TIFF image
Image017.tif	Jun 17, 2014 10:06 AM	95.2 MB	TIFF image
Image018.tif	Jun 17, 2014 10:08 AM	95.6 MB	TIFF image
Image019.tif	Jun 17, 2014 10:08 AM	94.5 MB	TIFF image
Image020.tif	Jun 17, 2014 10:10 AM	96.3 MB	TIFF image
Image021.tif	Jun 17, 2014 10:10 AM	94.6 MB	TIFF image
Image022.tif	Jun 17, 2014 10:12 AM	96.1 MB	TIFF image
Image023.tif	Jun 17, 2014 10:12 AM	94.1 MB	TIFF image
Image024.tif	Jun 17, 2014 10:14 AM	95.3 MB	TIFF image
Image025.tif	Jun 17, 2014 10:14 AM	93.7 MB	TIFF image
Image026.tif	Jun 17, 2014 10:16 AM	96.7 MB	TIFF image
JPEG	Sep 5, 2014 12:09 PM	--	Folder
americanlit-october1929	Sep 5, 2014 12:09 PM	--	Folder
Image001.tif	Jun 4, 2014 8:32 AM	108.2 MB	TIFF image
Image002.tif	Jun 4, 2014 8:34 AM	105.9 MB	TIFF image
Image003.tif	Jun 4, 2014 8:34 AM	105.3 MB	TIFF image
Image004.tif	Jun 4, 2014 8:37 AM	106.2 MB	TIFF image
Image005.tif	Jun 4, 2014 8:37 AM	106.1 MB	TIFF image
Image006.tif	Jun 4, 2014 8:39 AM	106.3 MB	TIFF image
Image007.tif	Jun 4, 2014 8:39 AM	106.3 MB	TIFF image
Image008.tif	Jun 4, 2014 8:42 AM	106.2 MB	TIFF image
Image009.tif	Jun 4, 2014 8:42 AM	106.2 MB	TIFF image
Image010.tif	Jun 4, 2014 8:49 AM	106.3 MB	TIFF image
Image011.tif	Jun 4, 2014 8:49 AM	106.3 MB	TIFF image
Image012.tif	Jun 4, 2014 8:50 AM	106.3 MB	TIFF image
Image013.tif	Jun 4, 2014 8:50 AM	106.3 MB	TIFF image
Image014.tif	Jun 4, 2014 8:52 AM	106.4 MB	TIFF image
Image015.tif	Jun 4, 2014 8:52 AM	106.3 MB	TIFF image
Image016.tif	Jun 4, 2014 8:54 AM	106.3 MB	TIFF image
Image017.tif	Jun 4, 2014 8:54 AM	106.4 MB	TIFF image
Image018.tif	Jun 4, 2014 8:59 AM	106.2 MB	TIFF image

File naming examples



File naming examples

Metadata

- Record in spreadsheet (Excel, GoogleDoc)
- Title, Subject, Description, Creator, Source, Publisher, Date, Contributor, Rights, Relation, Format, Language, Type, Identifier, Coverage
- ScanFileName, Title, Subject, Description, Creator, Publisher, PubPlace, Contributors, OriginalDate, Type, Format, OriginalSize, Source Identifier, CollectionTitle, CollectionCreator, Copyright Ownership, ScanResolution, ScanDate, Publisher, Gray/RGB, ScanningNotes, ManipulationNotes, Operator, Scanner/ColorBar, MasterFileLocation

Metadata

- Dublin Core
  http://omeka.org/codex/Working_with_Dublin_Core
- Text Encoding Initiative
  http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf
- Encoded Archival Description
  http://www.loc.gov/ead/

Optical Character Recognition (OCR)

- Transcription
- ABBYY Finereader http://www.abbyy.com
- OmniPage http://www.nuance.com/for-business/by-product/omnipage/standard/index.htm
- Tesseract https://code.google.com/p/tesseract-ocr/

Thank you!