EXPANDING YOUR TOOLKIT: USEFUL TOOLS FOR MANAGING BORN-DIGITAL CONTENT

Carol Kussmann

Upper Midwest Digital Collections Conference; Davenport, IA
August 10, 2016
What brought you to this workshop?
TODAY’S AGENDA

- Background information
- Lifecycle models
- Work through general workflow
- Learn about tools and use them throughout
- Wrap up
Why Manage Born-Digital Content?

Every Minute of the Day

- Pinterest users pin 3,472 images
- Vine users share 8,333 videos
- Skype users connect for 23,300 hours
- Yelp users post 26,380 reviews
- Apple users download 48,000 apps
- Pandora users listen to 61,141 hours of music
- Amazon makes $83,000 in online sales
- Instagram users post 216,000 new photos
- Twitter users tweet 277,000 times
- Tinder users swipe 416,667 times
- Facebook users share 2,460,000 pieces of content
- Google receives over 4,000,000 search queries
- YouTube users upload 72 hours of new video
- Whatsapp users share 347,222 photos
What is digital preservation?

- The process of actively maintaining a digital object for as long as required, in a form which is authentic, and accessible to users.


Access drives preservation – Brewster Kahle
**How Manage Born-Digital Content?**

- Understand digital lifecycle
- Follow workflows
- Use tools/system to help you

- Be consistent! But not afraid to change!
DIGITAL CURATION LIFECYCLE MODEL

http://www.dcc.ac.uk/resources/curation-lifecycle-model
**DPOE Model**

**Identify** - what digital content do you have?

**Select** - what portion of that content will be preserved?

**Store** - what issues are there for long-term storage?

**Protect** - what steps are needed to protect your digital content?

**Manage** - what provisions are needed for long-term management?

**Provide** - what considerations are there for long-term access?
INVENTORIES AND PRESERVATION PLANS

- Collect information about what you have
- Collect information about what you want to do
- Collect information about what you have done

- Understand your purpose and then create an inventory / preservation plan.
- Can be done at any time
# Inventory Examples

This spreadsheet contains multiple examples of inventories used in various situations with the general goal of learning more about digital holdings. Each sheet (tab) is a separate unrelated inventory. Details about the purpose and goals of each inventory are provided on their individual pages.

<table>
<thead>
<tr>
<th>Examples</th>
<th>Purpose</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archives and Special Collections (University of Minnesota)</td>
<td>Document incoming digital materials that will need to be addressed.</td>
<td>2014</td>
</tr>
<tr>
<td>Minnesota State Archives Preservation Plan</td>
<td>Documenting information about digital materials being held in the State Archives and Manuscript collections.</td>
<td>2013</td>
</tr>
<tr>
<td>National Digital Information Infrastructure Preservation Program (Minnesota State Archives)</td>
<td>Documenting information about record series being preserved by both the Minnesota State Archives and Minnesota Legislative Reference Library</td>
<td>2011</td>
</tr>
<tr>
<td>North Dakota State Library</td>
<td>Documenting information about digitized content being held in the North Dakota State Library Digital Collections.</td>
<td>2013</td>
</tr>
<tr>
<td>Wisconsin Historical Society</td>
<td>Documenting the digital collections at the Wisconsin Historical Society which serves as the WI State Archives. Unique to this inventory is the ranking system (criticality of the data) that assists with better understanding the risk of not being able to preserve the content as well as help plan storage requirements and incident recovery responses</td>
<td>2015</td>
</tr>
</tbody>
</table>

**Have you conducted an "inventory"? If you have an example to share and would like to include in this spreadsheet, please contact crkussmann@gmail.com.**

- [http://z.umn.edu/16yf](http://z.umn.edu/16yf)
INVENTORY TOOLS

Tools to record the information:
- Google Sheets
- Access Databases
- Collection Management Systems

Tools to collect information for the inventory:
- Windows Explorer / Get Info (Mac)
- NARA File Analyzer
- WinDirStat
- DROID
- HashMyFiles
Basic Inventory Information Exercise

- Determine basic information about the entire set of files
  - Number of files and folders
  - Amount of space the collection takes up

- Exercise...
  - Use Windows Explorer/Get Info on sample files folder
  - Collect information on spreadsheet (make a copy for yourself or create a new tab with your name on it)
  - http://z.umn.edu/17d5
WINDOWS EXPLORER / GET INFO
WINDIRSTAT

WinDirStat: http://windirstat.info/

Screenshot courtesy of University of Minnesota Libraries
WINDIRSTAT

- Explore content
- Understand file formats

Screenshot courtesy of University of Minnesota Libraries
WINDIRSTAT

- Explore specific files
**MN INVENTORY EXAMPLE**

- **Goal:** Understand information about the number of files, the amount of space they were occupying, and which file formats were in the collections.

- **Tools:**
  - Windows Properties / Get Info
  - WinDirStat
  - DROIDE
  - Database queries
  - Google Sheets
VISUAL REPRESENTATION

- 49,506 unique files
- 320 GB of content
**Visual Representation**

- 32,039 unique files
- 1769 GB of content
**Visual Representation**

Format Distribution by File Count

- **.jpg** 12,683 (3%)
- **.tif** 13,299 (1%)
- **.jp2** 24,474 (10%)
- **.txt** 35,823 (22%)
- **.dng** 178,915 (29%)
- **.pdf** 1,153,178 (26%)
- **.txt** 1,308,418 (19%)
- **.tif** 1,554,777 (26%)
- **.jpg** 1,771,829 (29%)

Screenshot courtesy of University of Minnesota Libraries
INGEST PROCESSES

“Get the Stuff”

Gather Info from Donor

Physical Transfer of Files
TOOLS TO HELP TRANSFER FILES

- **TeraCopy**
  - Copies files from one location to another
  - Can verify authentic transfer

- **Data Accessioner**
  - Copies files from one location to another
  - Verifies authentic transfer
  - Captures structural and technical metadata

  - Packages files into a zip file to transfer with metadata
  - Creates checksums to verify at a later date
EXERCISE: TRANSFER FILES

- **Data Accessioner**
  - Use Data Accessioner to transfer files in the sub-folder titled `_DataAccessionerTransferFiles`

- **TeraCopy**
  - Use TeraCopy to transfer the rest of the files

*Note: If you try and transfer more with DA, it will take a really long time.*
**Need to Know More: Metadata**

- You have the files, now what?
- You need to know more.
- Information = Metadata

**Preservation / Administrative**
Fixity (unchanged), Provenance (life story), Rights Information

**Technical**
(envt, format)

**Structural**
(understand, use)

**Descriptive**
(find, use)
GATHER BASIC INFORMATION

Virus / Malware Checks

Sensitive / Private Information

File Manifest / Metadata
COLLECT ADDITIONAL METADATA

- DROID
  - Checksums
  - Technical metadata
  - Structural metadata

- HashMyFiles
  - Checksums
  - Technical metadata
  - Structural metadata
  - Identical files (color coded)
DROID
## DROID

![DROID](image)

<table>
<thead>
<tr>
<th>Resource</th>
<th>Extension</th>
<th>Size</th>
<th>Last modified</th>
<th>Ids</th>
<th>Format</th>
<th>Version</th>
<th>Mime type</th>
<th>PUID</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>V16-n02_200705.pdf</td>
<td>pdf</td>
<td>1.2 MB</td>
<td>12/16/15 ...</td>
<td><img src="image" alt="Icon" /></td>
<td>Acrobat PDF</td>
<td>1.5</td>
<td>application...</td>
<td>fnt/19</td>
<td>Signature</td>
</tr>
<tr>
<td>V01-n12_199303.pdf</td>
<td>pdf</td>
<td>110.5 KB</td>
<td>12/16/15 ...</td>
<td><img src="image" alt="Icon" /></td>
<td>Acrobat PDF</td>
<td>1.1</td>
<td>application...</td>
<td>fnt/15</td>
<td>Signature</td>
</tr>
<tr>
<td>V08-n07_199910.pdf</td>
<td>pdf</td>
<td>472.3 KB</td>
<td>12/16/15 ...</td>
<td><img src="image" alt="Icon" /></td>
<td>Acrobat PDF</td>
<td>1.2</td>
<td>application...</td>
<td>fnt/16</td>
<td>Signature</td>
</tr>
<tr>
<td>V06-n10_199801.pdf</td>
<td>pdf</td>
<td>188 KB</td>
<td>12/16/15 ...</td>
<td><img src="image" alt="Icon" /></td>
<td>Acrobat PDF</td>
<td>1.1</td>
<td>application...</td>
<td>fnt/15</td>
<td>Signature</td>
</tr>
<tr>
<td>V22-n06-07_201309.pdf</td>
<td>pdf</td>
<td>4.9 MB</td>
<td>12/16/15 ...</td>
<td><img src="image" alt="Icon" /></td>
<td>Acrobat PDF</td>
<td>1.3</td>
<td>application...</td>
<td>fnt/17</td>
<td>Signature</td>
</tr>
<tr>
<td>V04-n05_199508.pdf</td>
<td>pdf</td>
<td>218.5 KB</td>
<td>12/16/15 ...</td>
<td><img src="image" alt="Icon" /></td>
<td>Acrobat PDF</td>
<td>1.1</td>
<td>application...</td>
<td>fnt/15</td>
<td>Signature</td>
</tr>
<tr>
<td>V14-n06-07_200509.pdf</td>
<td>pdf</td>
<td>1.7 MB</td>
<td>12/16/15 ...</td>
<td><img src="image" alt="Icon" /></td>
<td>Acrobat PDF</td>
<td>1.5</td>
<td>application...</td>
<td>fnt/19</td>
<td>Signature</td>
</tr>
<tr>
<td>V19-n07_201010.pdf</td>
<td>pdf</td>
<td>3.3 MB</td>
<td>12/16/15 ...</td>
<td><img src="image" alt="Icon" /></td>
<td>Acrobat PDF</td>
<td>1.5</td>
<td>application...</td>
<td>fnt/19</td>
<td>Signature</td>
</tr>
<tr>
<td>V04-n02_199505.pdf</td>
<td>pdf</td>
<td>159.5 KB</td>
<td>12/16/15 ...</td>
<td><img src="image" alt="Icon" /></td>
<td>Acrobat PDF</td>
<td>1.1</td>
<td>application...</td>
<td>fnt/15</td>
<td>Signature</td>
</tr>
<tr>
<td>V16-n10_200801.pdf</td>
<td>pdf</td>
<td>1.4 MB</td>
<td>12/16/15 ...</td>
<td><img src="image" alt="Icon" /></td>
<td>Acrobat PDF</td>
<td>1.5</td>
<td>application...</td>
<td>fnt/19</td>
<td>Signature</td>
</tr>
<tr>
<td>V18-n07_200910.pdf</td>
<td>pdf</td>
<td>2.4 MB</td>
<td>12/16/15 ...</td>
<td><img src="image" alt="Icon" /></td>
<td>Acrobat PDF</td>
<td>1.5</td>
<td>application...</td>
<td>fnt/19</td>
<td>Signature</td>
</tr>
<tr>
<td>V14-n02-03_200505.pdf</td>
<td>pdf</td>
<td>1.7 MB</td>
<td>12/16/15 ...</td>
<td><img src="image" alt="Icon" /></td>
<td>Acrobat PDF</td>
<td>1.5</td>
<td>application...</td>
<td>fnt/19</td>
<td>Signature</td>
</tr>
<tr>
<td>V10-n02-200105.pdf</td>
<td>pdf</td>
<td>467.4 KB</td>
<td>12/16/15 ...</td>
<td><img src="image" alt="Icon" /></td>
<td>Acrobat PDF</td>
<td>1.2</td>
<td>application...</td>
<td>fnt/16</td>
<td>Signature</td>
</tr>
<tr>
<td>V05-n05_199608.pdf</td>
<td>pdf</td>
<td>250.9 KB</td>
<td>12/16/15 ...</td>
<td><img src="image" alt="Icon" /></td>
<td>Acrobat PDF</td>
<td>1.1</td>
<td>application...</td>
<td>fnt/15</td>
<td>Signature</td>
</tr>
</tbody>
</table>

Screenshot courtesy of University of Minnesota Libraries
## DROID

### File count and sizes

**Profile Summary**

<table>
<thead>
<tr>
<th>Name</th>
<th>Signature version</th>
<th>Container version</th>
<th>Started</th>
<th>Finished</th>
<th>Filters</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCU-IMAGES</td>
<td>82</td>
<td>20150307</td>
<td>14 Aug 2015</td>
<td>15 Aug 2015</td>
<td></td>
</tr>
</tbody>
</table>

### File count and sizes

<table>
<thead>
<tr>
<th>Report field</th>
<th>Grouping fields</th>
</tr>
</thead>
<tbody>
<tr>
<td>FILE_SIZE</td>
<td></td>
</tr>
</tbody>
</table>

**Filter fields:**

<table>
<thead>
<tr>
<th>Field</th>
<th>Operator</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>RESOURCE_TYPE</td>
<td>NONE_OF</td>
<td>&quot;Folder&quot;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Profile</th>
<th>Count</th>
<th>Sum</th>
<th>Min</th>
<th>Max</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCU-IMAGES</td>
<td>2867306</td>
<td>76267393859435</td>
<td>0</td>
<td>254240458061</td>
<td>26598972</td>
</tr>
<tr>
<td>Profile totals</td>
<td>2867306</td>
<td>76267393859435</td>
<td>0</td>
<td>254240458061</td>
<td>26598972</td>
</tr>
</tbody>
</table>
### DROID

<table>
<thead>
<tr>
<th>Profile</th>
<th>Count</th>
<th>Sum</th>
<th>Min</th>
<th>Max</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aerial</td>
<td>9945</td>
<td>192763922072</td>
<td>4395</td>
<td>218831972</td>
<td>19382998</td>
</tr>
<tr>
<td>Profile totals</td>
<td>9945</td>
<td>192763922072</td>
<td>4395</td>
<td>218831972</td>
<td>19382998</td>
</tr>
</tbody>
</table>

### pft

<table>
<thead>
<tr>
<th>Profile</th>
<th>Count</th>
<th>Sum</th>
<th>Min</th>
<th>Max</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aerial</td>
<td>110</td>
<td>25020172</td>
<td>313</td>
<td>2102111</td>
<td>227456</td>
</tr>
<tr>
<td>Profile totals</td>
<td>110</td>
<td>25020172</td>
<td>313</td>
<td>2102111</td>
<td>227456</td>
</tr>
</tbody>
</table>

### pfx

<table>
<thead>
<tr>
<th>Profile</th>
<th>Count</th>
<th>Sum</th>
<th>Min</th>
<th>Max</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aerial</td>
<td>62</td>
<td>4647504</td>
<td>24</td>
<td>599536</td>
<td>74959</td>
</tr>
<tr>
<td>Profile totals</td>
<td>62</td>
<td>4647504</td>
<td>24</td>
<td>599536</td>
<td>74959</td>
</tr>
</tbody>
</table>
## Hash My Files

### File List

<table>
<thead>
<tr>
<th>Filename</th>
<th>Full Path</th>
<th>Iden</th>
<th>Extension</th>
<th>Created Time</th>
<th>Modified Time</th>
<th>MD5</th>
<th>SHA-256</th>
<th>File Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>XING_T24.TIF</td>
<td>CA\Users\cor0...</td>
<td>32</td>
<td>TIF</td>
<td>7/22/2016 12:53:49</td>
<td>1/20/2015 9:23:49</td>
<td>417d41c525ea3d4b444caac556b6fcf4</td>
<td>9a7a63d3b7f4a88a3879e8630f0d4f0</td>
<td>118,248</td>
</tr>
<tr>
<td>XING_B24.BMP</td>
<td>CA\Users\cor0...</td>
<td>32</td>
<td>BMP</td>
<td>7/22/2016 12:53:49</td>
<td>1/20/2015 9:23:59</td>
<td>a579b903eb8f52a2e69f3e590a05</td>
<td>572d990a1d55f4d7e547507585d4cd7</td>
<td>113,134</td>
</tr>
<tr>
<td>XING.JPG</td>
<td>CA\Users\cor0...</td>
<td>32</td>
<td>JPG</td>
<td>7/22/2016 12:53:49</td>
<td>1/20/2015 9:24:02</td>
<td>b32a4a4c8e1a2e6e576a2e630b2ef</td>
<td>7c5d5c9e3a17855c4f6b560847b5f157</td>
<td>45,778</td>
</tr>
<tr>
<td>WHAT.BMP</td>
<td>CA\Users\cor0...</td>
<td>32</td>
<td>BMP</td>
<td>7/22/2016 12:53:49</td>
<td>1/20/2015 9:26:21</td>
<td>3116b521d8222ff03033ce8376b7</td>
<td>685c4e686b4f9e5e7cb927986e17c7</td>
<td>86,454</td>
</tr>
<tr>
<td>W1_RTFO.DOC</td>
<td>CA\Users\cor0...</td>
<td>32</td>
<td>DOC</td>
<td>7/22/2016 12:53:49</td>
<td>1/20/2015 9:29:32</td>
<td>1f15e8e1e5c39f359e5d500653839f6</td>
<td>51eac30b17856e9c656a2d7e</td>
<td>2,932</td>
</tr>
<tr>
<td>tone24bit.flac</td>
<td>CA\Users\cor0...</td>
<td>32</td>
<td>flac</td>
<td>7/22/2016 12:53:49</td>
<td>1/20/2015 9:32:20</td>
<td>0073395c74a22e6f5d041637d866</td>
<td>728f62b6a6d6b0cd7c7b6a4e781d766</td>
<td>450,922</td>
</tr>
<tr>
<td>Daylight.jpg</td>
<td>CA\Users\cor0...</td>
<td>32</td>
<td>jpg</td>
<td>7/22/2016 12:53:49</td>
<td>1/20/2015 9:35:07</td>
<td>5ab317fca94de60268b1e1310398c51</td>
<td>54ec9531e74ef0133445f66f37e6a66</td>
<td>3,616,493</td>
</tr>
</tbody>
</table>

### File List (continued)

<table>
<thead>
<tr>
<th>Filename</th>
<th>Full Path</th>
<th>Iden</th>
<th>Extension</th>
<th>Created Time</th>
<th>Modified Time</th>
<th>MD5</th>
<th>SHA-256</th>
<th>File Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-1-2.jpg</td>
<td>CA\Users\cor0...</td>
<td>11</td>
<td>jpg</td>
<td>7/25/2016 11:41:00</td>
<td>1/20/2015 9:17:17</td>
<td>2e248b2537aa0c01c7e4d737d9a0d6</td>
<td>24a3f4782b6dc6c5b71500dd886ecb72e2</td>
<td>3,967</td>
</tr>
<tr>
<td>2-2-1.jpg</td>
<td>CA\Users\cor0...</td>
<td>12</td>
<td>jpg</td>
<td>7/22/2016 12:50:12</td>
<td>1/20/2015 9:17:68</td>
<td>db7a52e79313ec4e30355952943c8a0</td>
<td>4651c68f893c2bf1cd6e7a8835d9e1</td>
<td>1,855</td>
</tr>
<tr>
<td>2-2-1.jpg</td>
<td>CA\Users\cor0...</td>
<td>12</td>
<td>jpg</td>
<td>7/22/2016 12:50:12</td>
<td>1/20/2015 9:17:68</td>
<td>db7a52e79313ec4e30355952943c8a0</td>
<td>4651c68f893c2bf1cd6e7a8835d9e1</td>
<td>1,855</td>
</tr>
<tr>
<td>2-2-1.jpg</td>
<td>CA\Users\cor0...</td>
<td>13</td>
<td>jpg</td>
<td>7/22/2016 12:50:12</td>
<td>1/20/2015 9:16:59</td>
<td>eb7e5ed938960c7786e41979e4df7dace</td>
<td>2dc4d1473820988e7ba3b86c38707e2e0</td>
<td>492</td>
</tr>
<tr>
<td>2-2-1.jpg</td>
<td>CA\Users\cor0...</td>
<td>13</td>
<td>jpg</td>
<td>7/22/2016 12:50:12</td>
<td>1/20/2015 9:16:59</td>
<td>eb7e5ed938960c7786e41979e4df7dace</td>
<td>2dc4d1473820988e7ba3b86c38707e2e0</td>
<td>492</td>
</tr>
<tr>
<td>0000023.bmp</td>
<td>CA\Users\cor0...</td>
<td>14</td>
<td>bmp</td>
<td>7/22/2016 12:47:59</td>
<td>1/20/2015 9:20:58</td>
<td>70ae5f3903bbf4511619c7ce601b6</td>
<td>4859432b1099e030ccdb83b3e535f41007</td>
<td>2,842,962</td>
</tr>
<tr>
<td>23.bmp</td>
<td>CA\Users\cor0...</td>
<td>14</td>
<td>jpg</td>
<td>7/22/2016 12:47:59</td>
<td>1/20/2015 9:15:39</td>
<td>70ae5f3903bbf4511619c7ce601b6</td>
<td>4859432b1099e030ccdb83b3e535f41007</td>
<td>2,842,962</td>
</tr>
<tr>
<td>wordDoc_1.docx</td>
<td>CA\Users\cor0...</td>
<td>15</td>
<td>docx</td>
<td>7/22/2016 12:04:46</td>
<td>1/20/2015 9:05:34</td>
<td>d12a73a3e87c6209a827519b2e86</td>
<td>65da6c59619f4ed42e1e83e1380572b1d90</td>
<td>12,714</td>
</tr>
<tr>
<td>234.doc</td>
<td>CA\Users\cor0...</td>
<td>15</td>
<td>docx</td>
<td>7/22/2016 12:05:34</td>
<td>1/20/2015 9:05:34</td>
<td>d12a73a3e87c6209a827519b2e86</td>
<td>65da6c59619f4ed42e1e83e1380572b1d90</td>
<td>12,714</td>
</tr>
<tr>
<td>0382.docx</td>
<td>CA\Users\cor0...</td>
<td>15</td>
<td>docx</td>
<td>7/22/2016 12:05:34</td>
<td>1/20/2015 9:05:34</td>
<td>d12a73a3e87c6209a827519b2e86</td>
<td>65da6c59619f4ed42e1e83e1380572b1d90</td>
<td>12,714</td>
</tr>
<tr>
<td>BMP W24b - P...</td>
<td>CA\Users\cor0...</td>
<td>16</td>
<td>bmp</td>
<td>7/22/2016 12:48:01</td>
<td>1/20/2015 8:15:22</td>
<td>df747777c591370d1b2151347b6d4176</td>
<td>459741486f4e5e06c3228d7c387072b0</td>
<td>2,359,352</td>
</tr>
<tr>
<td>BMP - ILLC544...</td>
<td>CA\Users\cor0...</td>
<td>16</td>
<td>bmp</td>
<td>7/22/2016 12:48:01</td>
<td>1/20/2015 8:15:54</td>
<td>df747777c591370d1b2151347b6d4176</td>
<td>459741486f4e5e06c3228d7c387072b0</td>
<td>2,359,352</td>
</tr>
<tr>
<td>BMP_OS 24b - P...</td>
<td>CA\Users\cor0...</td>
<td>17</td>
<td>bmp</td>
<td>7/22/2016 12:48:01</td>
<td>1/20/2015 8:15:24</td>
<td>df747777c591370d1b2151347b6d4176</td>
<td>459741486f4e5e06c3228d7c387072b0</td>
<td>2,359,352</td>
</tr>
<tr>
<td>BMP_OS 2 - PS...</td>
<td>CA\Users\cor0...</td>
<td>17</td>
<td>bmp</td>
<td>7/22/2016 12:48:01</td>
<td>1/20/2015 8:15:36</td>
<td>df747777c591370d1b2151347b6d4176</td>
<td>459741486f4e5e06c3228d7c387072b0</td>
<td>2,359,352</td>
</tr>
</tbody>
</table>
EXERCISE: COLLECT METADATA

- Run DROID and/or HashMyFiles on the entire collection
- Review how to save the information
- Determine if there is a format you prefer
- Which tool works better for you?
Evaluate
PROCESSING / FILE CLEAN UP

Content and Organization

- Personally Identifiable Information
- Duplicate Files
- Organizational Needs

Images: OpenClipArt.org
**Find and Remove Duplicates**

- **Duplicate File Finder**

![Screenshot of Duplicate File Finder](http://www.ashisoft.com/)

Screenshot courtesy of University of Minnesota Libraries
EXERCISE: FIND AND REMOVE DUPLICATES

- Using Duplicate File Finder, search the collection for duplicates.
- Make choices on what to remove and remove them.

Screenshot courtesy of University of Minnesota Libraries
RENAMe FILES

- Bulk Rename Utility

Screenshot courtesy of University of Minnesota Libraries

http://www.bulkrenameutility.co.uk/
EXERCISE: RENAME FILES

- Use Bulk Rename Utility to...
  - Add a prefix
  - Remove spaces
  - Add a date
CREATE NEW DOCUMENTATION

- If warranted, might need to re-run documentation...
- Do you want to have reports that document new file names, new organization, removal of files (lack of existence)...
- What might you use to do this?
PROTECT CONTENT

- Moved to long-term ‘storage’

Protecting from:
- Physical Disasters
- Natural Disasters
- Security Risks
  - Internal and External
- Technological Risks
  - Obsolescence / Degradation

CC license: http://creativecommons.org/licenses/by-nc-sa/2.0
PROTECT CONTENT

- Have multiple copies
  - Lots of Copies Keeps Stuff Safe (LOCKSS)
- Know where your content is located
  - Onsite/offsite; online/offline; preservation/access
- Know who has access to it
  - DP staff, IT staff, others?
- Understand obsolescence concerns
  - Formats; hardware; software
How do we know if a file changed over time?

Calculate and monitor its ‘fixity’

**Fixity**: the property of a digital object being fixed or unchanged

Use checksums.

**Checksum**: A mathematical algorithm run on/against a file and its resulting value.

Multiple types: MD5, SHA-1, SHA-256…
**Fixity**

- **Document**
  - HashMyFiles
  - DROID
  - Data Accessioner
  - Many more...

- **Document and Monitor**
  - MD5 Summer
  - Exact File
  - Systems themselves...

Image: CC Attribution: www.digitalbevaring.dk
MD5 SUMMER

http://www.md5summer.org/

Screenshot courtesy of Wisconsin Historical Society
MD5 SUMMER

- Select location to calculate checksum on
- Select checksum algorithm from dropdown
MD5 SUMMER
EXACT FILE

ExactFile 1.0.0.15 BETA

- Single File Checksums
- Create Digest
- Test Digest
- Create TestFile Applet
- Benchmark
- About ExactFile

Checksum Digest for Files in a Folder

- C:\Users\corr013\Desktop\Chrome Downloads\...
- Include subfolders
- Include full paths in output
- Output to File: C:\Users\corr013\Desktop\Chrome Downloads\checksum

Method and Format: MD5 (MD5 checksums in md5sum format)

Creating file checksums...
- Fotolia_41749254_X1.jpg
- ASC Presentation_Final.pdf
- 20130109_CavePhotos (1).zip
- DPOE Sample Files.zip
- DCP-Acronyms-1.doc
- DPOE Sample Files (1).zip
- Enterprise Data Resource, Maturity Assessment (draft 20140222).pdf
- 20130109_CavePhotos.zip
- 108 of 436 files completed...

Screenshot courtesy of University of Minnesota Libraries
**Exercise: Checksums**

- Exact File or MD5 Summer
- Generate checksums for a folder of the sample files
- Change a file name
- Recheck the checksums
- Open a Word document in the folder
- Add or delete something and resave
- Recheck the checksums
ACCESS

- Preservation = Access
- The reason we are trying to manage these collections is to provide access
MANAGEMENT: WHO WILL DO THIS WORK??

- Management is ongoing work. Who does this?
- Be flexible

Ingest Workflow Example
TOOL SELECTION SUGGESTIONS

- Understand your goals / purpose
- Create a workflow
- Test tools (don’t select tools first)
- Select best tools to help
- Be flexible
Questions? Other Tool Suggestions?

Contact Information

Carol Kussmann
University of Minnesota Libraries
kussmann@umn.edu

Image: CC Attribution: www.digitalbevaring.dk
WORKSHOP RESOURCES

- List of Exercises: http://z.umn.edu/17hs
- Exercise Worksheet: http://z.umn.edu/17d5
Upper Midwest Digital Collections Conference 2016
Expanding Your Toolkit: Useful Tools for Managing Born-Digital Content

Workshop Exercises
The following lists the exercises presented in the Expanding Your Toolkit Workshop at the Upper Midwest Digital Collections Conference 2016 in Davenport, Iowa.

Exercise worksheet: http://z.umn.edu/17d5

Exercise 1: Inventory Basics

It is important to understand on a general level what you are collecting or bringing into an archives. The most basic information we need is the number of files and the total size that they take up.

Use Windows Explorer Properties menu to figure out the number of files and total file size of the collection.

Exercise 2: Transfer Files

Most electronic records are moved around via external hard drive or flash drive which then need to be transferred to an archives or another location. When we move files, we want to limit the amount of changes we make to a file and its metadata.

Use Data Accessioner to transfer the sample files from the folder they are now in into a new folder you create. NOTE: run this program only on the sub folder titled _DataAccessionerTransferFiles. If you do not do this, it will take a really long time!

Exercise 3: Collect Additional Metadata

After files are in our possession we need to know more about them.

Use one or more programs to see how much information you can gather and what forms/formats you can save it in.

DROID: File name (resource), extension, size, last modified date, IDs, format, version, MIME type, PUID, method, hash

HashMyFiles: File name, extension, size, modified time, version, hash value(s), full path, created time, identical

WinDirStat: File name, extension

CRKussmann (kussmann@umn.edu)
August 5, 2016
**Exercise 4: Finding and Removing Duplicate Files**

If you ran HashMyFiles you will have an understanding of how many duplicates might be in your collection. With these files being from a shared hard drive, lets assume that there is some duplication.

Use Duplicate File Finder to find the duplicate files and remove them.

**Exercise 5: Renaming Files**

There are many reasons why renaming files might be necessary. You may want to add a prefix or suffix to the original name that helps identify the collection. You may want to add a date. You may want to remove special characters.

Use **Bulk Rename Utility** to rename some files in folder x…
- Add prefix  [suggest add “Transferred_”]
- Remove spaces
- Add date  [suggest for some or all files]

**Exercise 6: Checksums**

Checksums should be run on content at the point of ingest or soon after. In this example, we are going to run checksums on the content we are deciding to keep for our archives so we have a complete list.

Use a checksum tool of your choice, try more than one for various outputs, and generate checksums and think about how you would use the outputs to verify that the content has not changed over time.

**Tools Discussed:**
- MD5 Summer
- Exact File
- DROID
- HashMyFiles
- Fixity (not provided)