Surprise!
Look What We Have!?

Carol Kussmann
Digital Preservation Analyst
University of Minnesota Libraries

March 16, 2016
Library Technology Conference; Macalester College, St. Paul, MN
Background

- Building our digital preservation program
  - Foundational work
  - Policy development
  - Requirements for hardware and software

- Everything is based on “the stuff”
  - Need to know where the stuff is
  - Need to know how much stuff
  - Need to know more about the types of files
The “Eventually” Plan...

- Digital Preservation Outreach and Education Modules
  - Identify
  - Select
  - Store
  - Protect
  - Manage
  - Provide

Image: Digital Preservation Outreach & Education curriculum materials
The Inventory Process

- Determine whose files are being cared for long-term
- Determine where those files are being stored
- Determine best way to calculate total size, file count, and file format list
- Run tools/reports to capture required information
- Compile results and review
- Share results
“Collections” with Long-Term Value

- Institutional repository
- Data repository
- UMedia repository
- AgEcon materials
- Minnesota Digital Library
- Archives and Special Collections materials
- Digitization projects
- Specific projects with digital components

Location of Materials

- Library servers
- OIT servers
- Network drives
- Standalone computers

Questions to ask
- Are these duplicated anywhere?
- What types of files are included (access/preservation)?
- Is there a common tool to use to capture information?
- Who do I need to work with to get the information?
Tools Used

- Windows Properties / Mac Get Info
- WinDirStat
- DROID*
- Database query*
- Google Sheets*
Windows Properties/Get Info

- Used for size estimates
- Does not give info on file formats
WinDirStat

<table>
<thead>
<tr>
<th>Name</th>
<th>Subtree Percent</th>
<th>Percenta...</th>
<th>&gt; Size</th>
<th>Items</th>
<th>Files</th>
<th>Subdirs</th>
<th>Last Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q:\ASC</td>
<td>63.3%</td>
<td>[1:12 s]</td>
<td>1.6 TB</td>
<td>20,449</td>
<td>16,415</td>
<td>4,034</td>
<td>2/1/2016 9:19:16 PM</td>
</tr>
<tr>
<td>PAA</td>
<td>27.8%</td>
<td>1017.5 GB</td>
<td>3,529</td>
<td>3,231</td>
<td>153</td>
<td>135</td>
<td>12/2/2015 7:20:48 PM</td>
</tr>
<tr>
<td>IHRCA</td>
<td>5.6%</td>
<td>446.6 GB</td>
<td>527</td>
<td>374</td>
<td>466</td>
<td>153</td>
<td>10/14/2015 9:15:22 PM</td>
</tr>
<tr>
<td>Givens</td>
<td>2.2%</td>
<td>90.3 GB</td>
<td>3,694</td>
<td>3,228</td>
<td>526</td>
<td>12/16/2015 5:43:07 PM</td>
<td></td>
</tr>
<tr>
<td>UMJA</td>
<td>0.6%</td>
<td>36.0 GB</td>
<td>4,366</td>
<td>3,840</td>
<td>386</td>
<td>12/2/2015 7:15:07 PM</td>
<td></td>
</tr>
<tr>
<td>Trettler</td>
<td>0.4%</td>
<td>9.1 GB</td>
<td>2,659</td>
<td>2,273</td>
<td>244</td>
<td>12/4/2015 4:31:43 PM</td>
<td></td>
</tr>
<tr>
<td>YMCA</td>
<td>0.0%</td>
<td>6.5 GB</td>
<td>2,681</td>
<td>2,437</td>
<td>202</td>
<td>12/4/2015 4:31:43 PM</td>
<td></td>
</tr>
<tr>
<td>CBI</td>
<td>0.0%</td>
<td>430.9 MB</td>
<td>258</td>
<td>252</td>
<td>6</td>
<td>2/1/2016 9:19:16 PM</td>
<td></td>
</tr>
<tr>
<td>CLRC</td>
<td>0.0%</td>
<td>183.8 MB</td>
<td>371</td>
<td>352</td>
<td>19</td>
<td>10/22/2015 4:19:01 PM</td>
<td></td>
</tr>
<tr>
<td>UA</td>
<td>0.0%</td>
<td>166.7 MB</td>
<td>2,315</td>
<td>401</td>
<td>1,914</td>
<td>1/7/2016 6:08:10 PM</td>
<td></td>
</tr>
<tr>
<td>MSS</td>
<td>0.0%</td>
<td>59.5 MB</td>
<td>33</td>
<td>25</td>
<td>8</td>
<td>5/13/2015 9:05:11 PM</td>
<td></td>
</tr>
<tr>
<td>&lt;Files&gt;</td>
<td>0.0%</td>
<td>4.7 MB</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>12/17/2015 8:56:31 PM</td>
<td></td>
</tr>
<tr>
<td>Kuss test</td>
<td>0.0%</td>
<td>0.0 MB</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>9/3/2015 3:18:55 PM</td>
<td></td>
</tr>
<tr>
<td>NAA</td>
<td>0.0%</td>
<td>0.0 MB</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10/31/2014 3:18:36 PM</td>
<td></td>
</tr>
<tr>
<td>SCRIB</td>
<td>0.0%</td>
<td>0.0 MB</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10/31/2014 3:18:51 PM</td>
<td></td>
</tr>
<tr>
<td>SWHA</td>
<td>0.0%</td>
<td>0.0 MB</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10/31/2014 3:19:02 PM</td>
<td></td>
</tr>
</tbody>
</table>

Image: Screenshots by CRKussmann
WinDirStat
WinDirStat

<table>
<thead>
<tr>
<th>Name</th>
<th>Subtree Percent</th>
<th>Percenta</th>
<th>&gt; Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kinetic cromlech 1983.mov</td>
<td></td>
<td>4.0%</td>
<td>10.7 GB</td>
</tr>
<tr>
<td>Urban Sky Harvest 1991 .mov</td>
<td></td>
<td>3.8%</td>
<td>10.1 GB</td>
</tr>
<tr>
<td>Bobcat Dance- edited 1992.mov</td>
<td></td>
<td>2.6%</td>
<td>6.9 GB</td>
</tr>
<tr>
<td>Duluth .mov</td>
<td></td>
<td>2.5%</td>
<td>6.6 GB</td>
</tr>
<tr>
<td>Mother's Day Dance 2001 edited.mov</td>
<td></td>
<td>2.4%</td>
<td>6.4 GB</td>
</tr>
<tr>
<td>Dance Dance 1988 b.mov</td>
<td></td>
<td>2.3%</td>
<td>6.3 GB</td>
</tr>
<tr>
<td>Trumpets and Tectonic Terpsichore- Hubert H...</td>
<td></td>
<td>2.1%</td>
<td>5.6 GB</td>
</tr>
<tr>
<td>Dance for Peace mov</td>
<td></td>
<td>1.9%</td>
<td>5.2 GB</td>
</tr>
<tr>
<td>Work Samples 1983-1990.mov</td>
<td></td>
<td>1.9%</td>
<td>5.1 GB</td>
</tr>
<tr>
<td>Dance for Peace Sarajevo 1996.mov</td>
<td></td>
<td>1.9%</td>
<td>5.0 GB</td>
</tr>
<tr>
<td>Opus on a May Morning 1987.mov</td>
<td></td>
<td>1.9%</td>
<td>5.0 GB</td>
</tr>
<tr>
<td>Kinetic cromlech b .mov</td>
<td></td>
<td>1.5%</td>
<td>4.1 GB</td>
</tr>
</tbody>
</table>

Image: Screenshots by CRKussmann
WinDirStat

![WinDirStat screenshot](image.jpg)
WinDirStat

- Capture information about collection size, number of files

- Capture information about file types
DROID

- Worked for networked drives, library servers, and standalone computers
- Reporting capabilities

Image: Screenshots by CRKussmann
# DROID Reports

## 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

### Report field

<table>
<thead>
<tr>
<th>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>

### Profile

<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>

Image: Screenshots by CRKussmann
# DROID Reports

### pdf

<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>
Database Query

- Sample query reports
  - Total size (GB)
  - File size (MB)
  - File count

```plaintext
49506 bitstreams found.
totalSize = 319.5875 Gb
application/matlab: 130 files, 556.69037 Mb
image/png: 10 files, 6.048507 Mb
application/octet-stream: 383 files, 45906.457 Mb
application/pdf: 47572 files, 197918.3 Mb
application/msword: 228 files, 41.648846 Mb
image/jpeg: 39 files, 90.5604 Mb
text/epub: 4 files, 23.549309 Mb
text/xml: 8 files, 0.200381 Mb
text/mobi: 2 files, 13.790194 Mb
application/vnd.ms-powerpoint: 17 files, 82.0181 Mb
application/postscript: 2 files, 0.366856 Mb
application/x-tex: 2 files, 0.371093 Mb
text/html: 12 files, 0.883703 Mb
video/mpeg: 6 files, 119.456795 Mb
image/tiff: 96 files, 22394.445 Mb
text/csv: 116 files, 313.24612 Mb
audio/x-wav: 14 files, 1239.0375 Mb
application/zip: 354 files, 34339.824 Mb
application/vnd.ms-excel: 56 files, 36.22191 Mb
video/quicktime: 28 files, 15452.748 Mb
text/plain: 409 files, 1047.4673 Mb
text/richtext: 18 files, 4.138414 Mb
```

88257 records found.
88344 bitstreams found.
totalSize = 98.188774 Gb
application/vnd.ms-excel: 2 files, 1.10327 Mb
text/richtext: 1 file, 0.034616 Mb
application/msword: 21 files, 20.69224 Mb
application/octet-stream: 7 files, 7.830204 Mb
text/html: 2 files, 0.203356 Mb
application/vnd.ms-powerpoint: 3 files, 22.706623 Mb
application/pdf: 88308 files, 98136.195 Mb
Process for Combining Information

- Pulled file format, number of files, and size from DROID
- Used database query reports to get file format, number of files, and size
- Put file formats/extension in an alphabetical list in google sheets
- Each collection had its own tab
- Combined on another tab
- Used to create reports
Putting it All Together

- Results for one collection in a Google Sheet
- Totals, file format, number of files from report
- Converted to GB (Displayed 2x for use later)
# Google Sheets Totals

![Image: Screenshots by CRKussmann](image.png)

## Table

<table>
<thead>
<tr>
<th>File Type</th>
<th>Total File Size (bytes)</th>
<th>Number of Files</th>
<th>MDL</th>
<th>Total File Size (bytes)</th>
<th>Number of Files</th>
<th>Aerial</th>
<th>Total File Size (bytes)</th>
<th>Number of Files</th>
<th>DCImages</th>
<th>Total File Size (bytes)</th>
<th>Number of Files</th>
</tr>
</thead>
<tbody>
<tr>
<td>adf</td>
<td>369,000</td>
<td>223</td>
<td></td>
<td>289,000</td>
<td></td>
<td></td>
<td>289,000</td>
<td></td>
<td></td>
<td>289,000</td>
<td></td>
</tr>
<tr>
<td>aux</td>
<td>164,000</td>
<td>266</td>
<td></td>
<td>238,000</td>
<td></td>
<td></td>
<td>238,000</td>
<td></td>
<td></td>
<td>238,000</td>
<td></td>
</tr>
<tr>
<td>avi</td>
<td>256,000</td>
<td>286</td>
<td></td>
<td>289,000</td>
<td></td>
<td></td>
<td>289,000</td>
<td></td>
<td></td>
<td>289,000</td>
<td></td>
</tr>
<tr>
<td>or2</td>
<td>90,000</td>
<td>234</td>
<td></td>
<td>238,000</td>
<td></td>
<td></td>
<td>238,000</td>
<td></td>
<td></td>
<td>238,000</td>
<td></td>
</tr>
<tr>
<td>mov</td>
<td>120,000</td>
<td>202</td>
<td></td>
<td>204,000</td>
<td></td>
<td></td>
<td>204,000</td>
<td></td>
<td></td>
<td>204,000</td>
<td></td>
</tr>
<tr>
<td>pdf</td>
<td>256,000</td>
<td>266</td>
<td></td>
<td>289,000</td>
<td></td>
<td></td>
<td>289,000</td>
<td></td>
<td></td>
<td>289,000</td>
<td></td>
</tr>
<tr>
<td>png</td>
<td>164,000</td>
<td>266</td>
<td></td>
<td>238,000</td>
<td></td>
<td></td>
<td>238,000</td>
<td></td>
<td></td>
<td>238,000</td>
<td></td>
</tr>
</tbody>
</table>

**Image:** Screenshots by CRKussmann
Final “Report”

- Explanatory text
  - Why and what’s included
- Represents most common formats
- Totals
- Individual repositories
- Charts

The University of Minnesota Libraries is responsible for the long-term access and preservation of a variety of materials including those in digital format. This report represents the current holdings of the Libraries’ digital assets that Digital Preservation and Repositories Technology department is responsible for. (This does not include content that other repositories / services are responsible for including HathiTrust and ArchiveIt.) This number will continue to increase as more and more work is done natively in the digital realm and as expectations for availability of digital files continues to increase.

Areas represented include:
- University Digital Conservancy / Data Repository of Minnesota
- UMedia Archive
- AgEcon Repository
- Library initiated digitization projects of Library materials
- Archives and Special Collections Materials
- Minnesota Digital Library master scans
- Master files of special projects initiated/hosted by the Libraries (e.g. Aerial photographs)

A variety of file formats are found within the Libraries collections - the formats displayed in the charts below represent only the top formats by number of files and by total amount of space they use. The majority of the files in our holdings represents images and textual based documents; while video and image files consume 90% of the total space. It is important to note that by file count, audio and video files did not make the top ten list, however 45% of our total storage space it taken up by video files alone.

Individual inventories were also created and shared with the associated repository or collecting area. Moving forward these baseline inventories can be used to see growth and pattern changes as we continue to collect and produce digital content that requires long-term management.
Grand Totals

- 6,095,978 unique files
- ~130 Tb of content
Sample Report 1

- 49,506 unique files
- 320 GB of content

Format Distribution by File Count

Format Distribution by File Size (GB)
Sample Report 2

- 32,039 unique files
- 1769 GB of content
Sample Report 3

- 2,867,306 unique files
- 69 TB of content

Image: Screenshots by CRKussmann
Summary

- Collaborative effort
- Understand the ‘problem/question’ then find a tool to solve/answer

- Inventory as a first step

- Given a better idea of what we have and where
- Provides us with more information on which to base next steps
  - Foundational work
  - Policy development
  - Requirements for hardware and software
Digital Preservation and Outreach Education Training

- Managing Digital Content Over Time: An Introduction to Digital Preservation
- Three part webinar series
- April/May
Resource Links

- **DROID**: [http://z.umn.edu/147h](http://z.umn.edu/147h)
- **WinDirStat**: [http://windirstat.info/](http://windirstat.info/)
- **Tools Sheet**: [http://z.umn.edu/147i](http://z.umn.edu/147i)
Questions?

Contact Me
Carol Kussmann
kussmann@umn.edu

Image: https://openclipart.org/detail/194097/grey-bird-with-question-marks