# 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

LIBRARIES
UNIVERSITY OF MINNESOTA

# 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

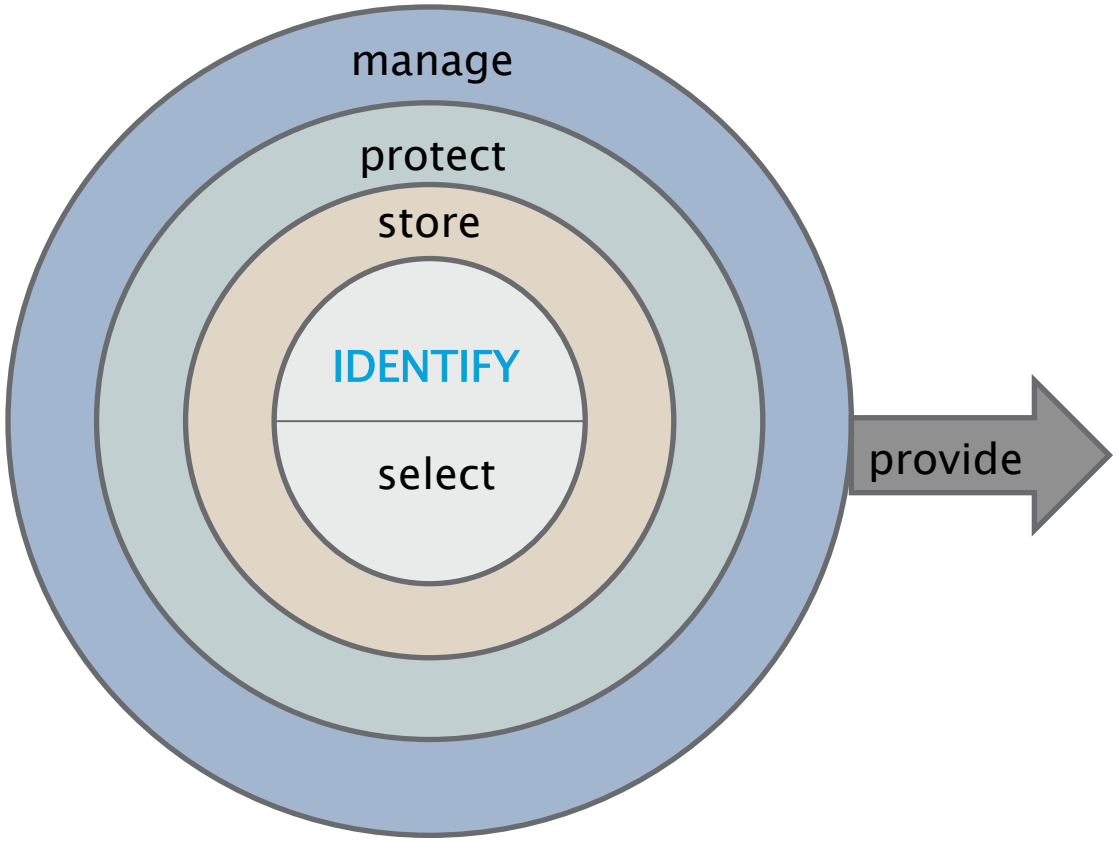



Image: Digital Preservation Outreach & Education curriculum materials

# 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

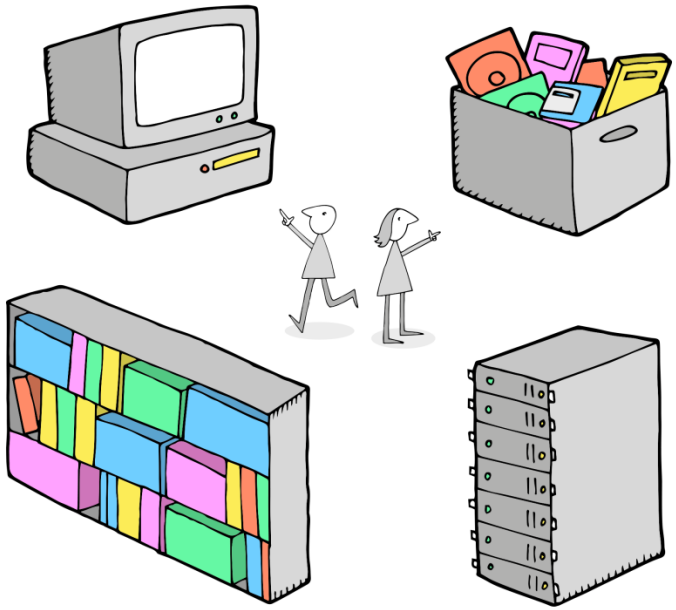

Image: (CC BY 2.5 Denmark) www.digitalbevaring.dk. Illustration by Jorgen Stamp.

# “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*

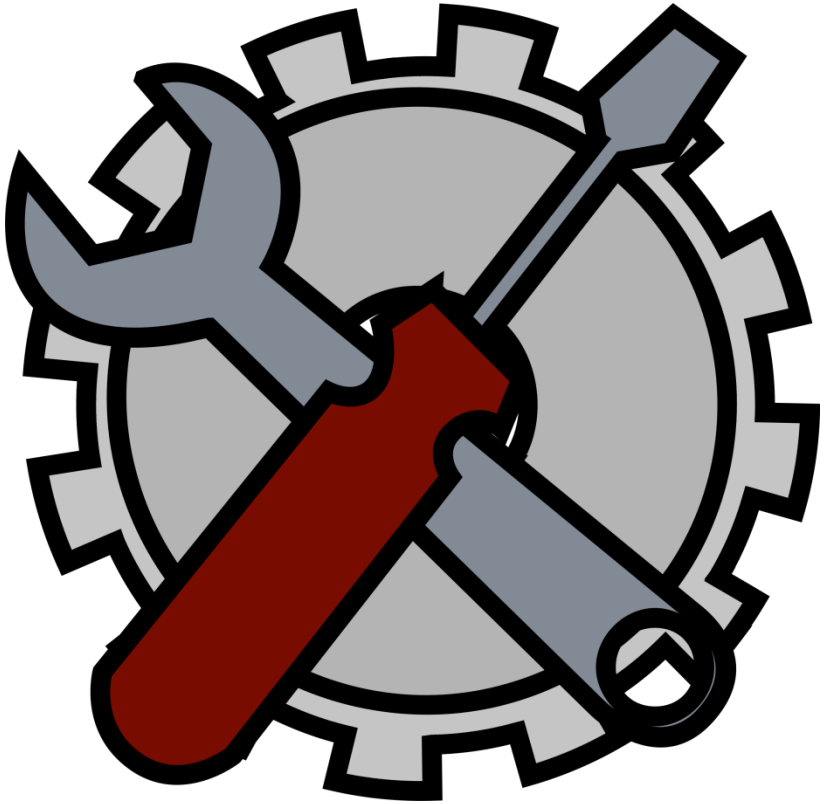
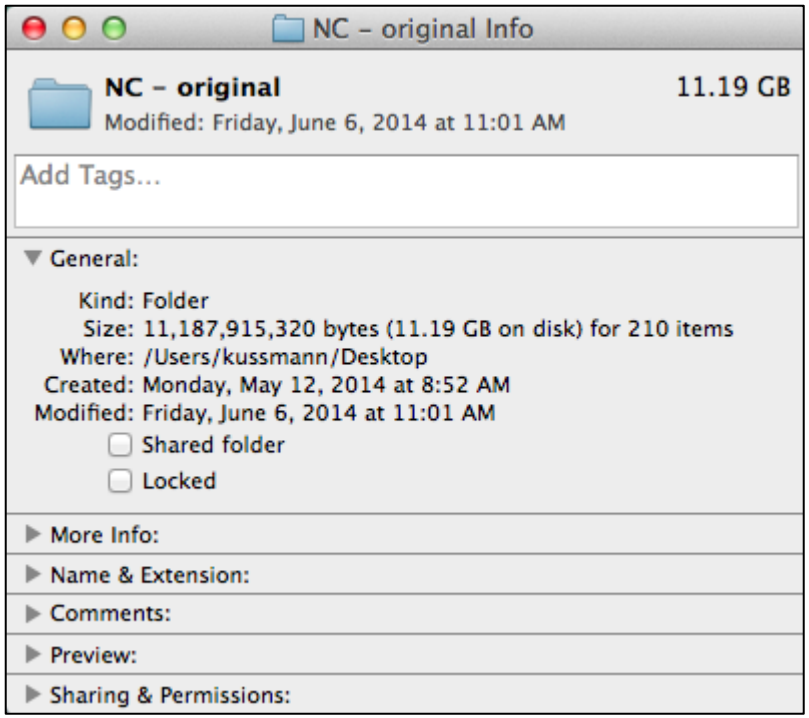


Image: https://openclipart.org/detail/143623/admin-tools-icon

# Windows Properties/Get Info

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



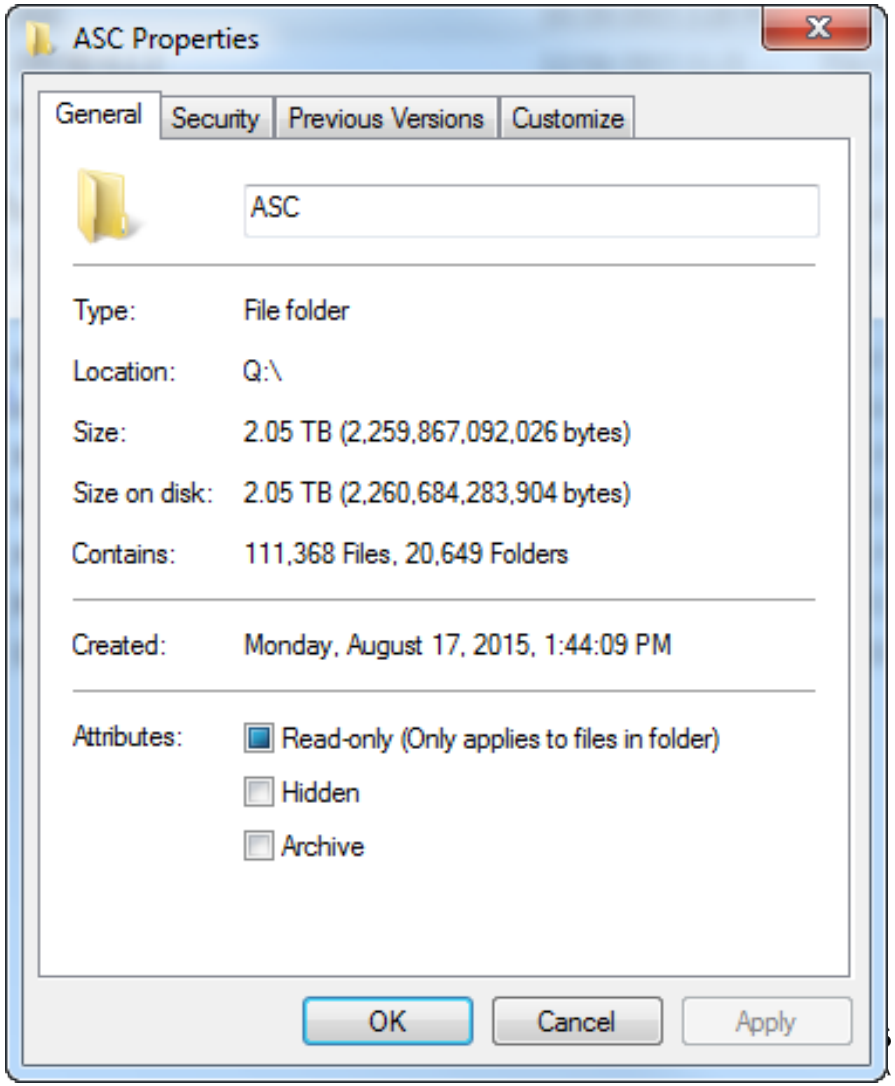




Image: Screenshots by CRKussmann

# WinDirStat

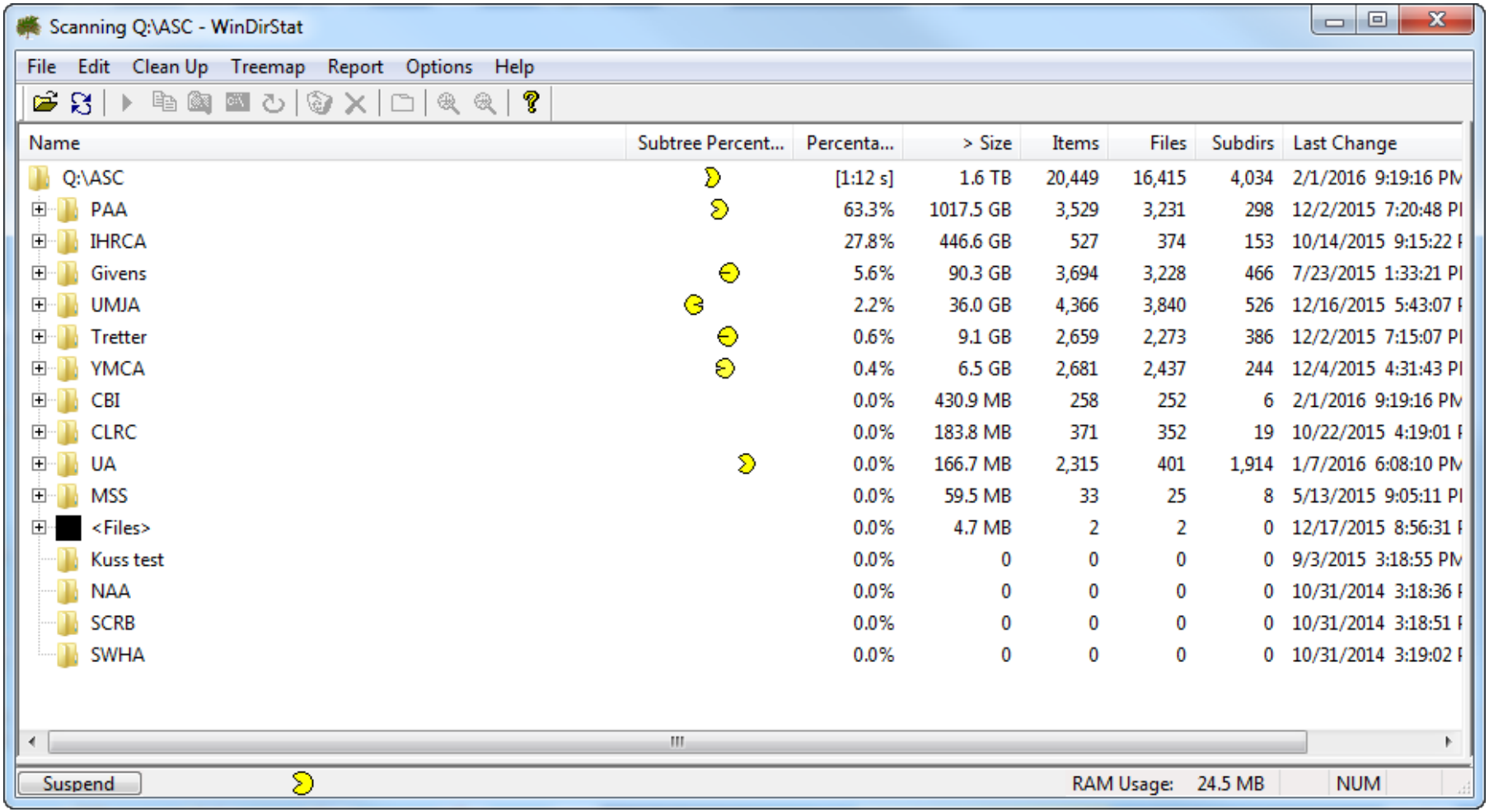



Image: Screenshots by CRKussmann

# WinDirStat

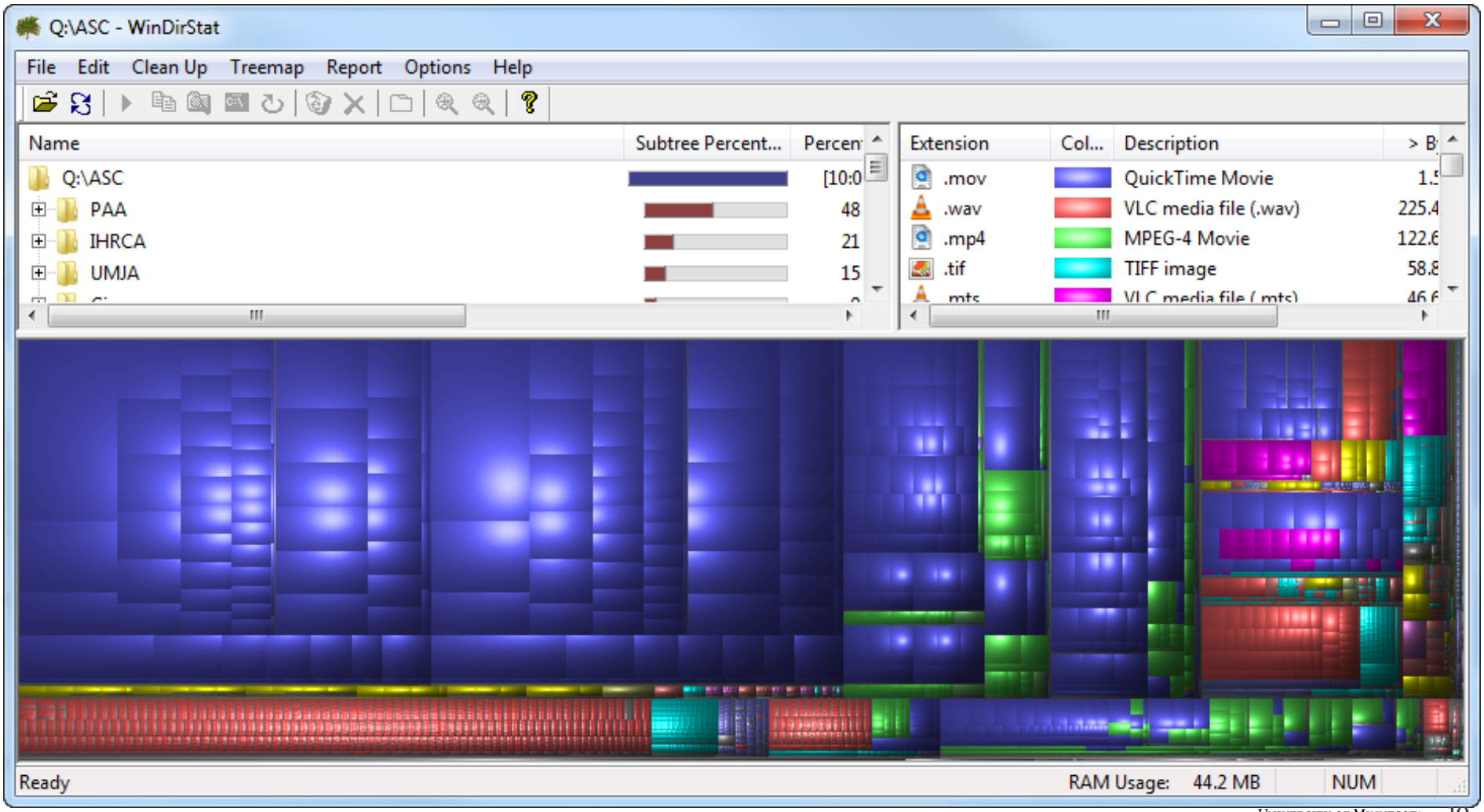



Image: Screenshots by CRKussmann

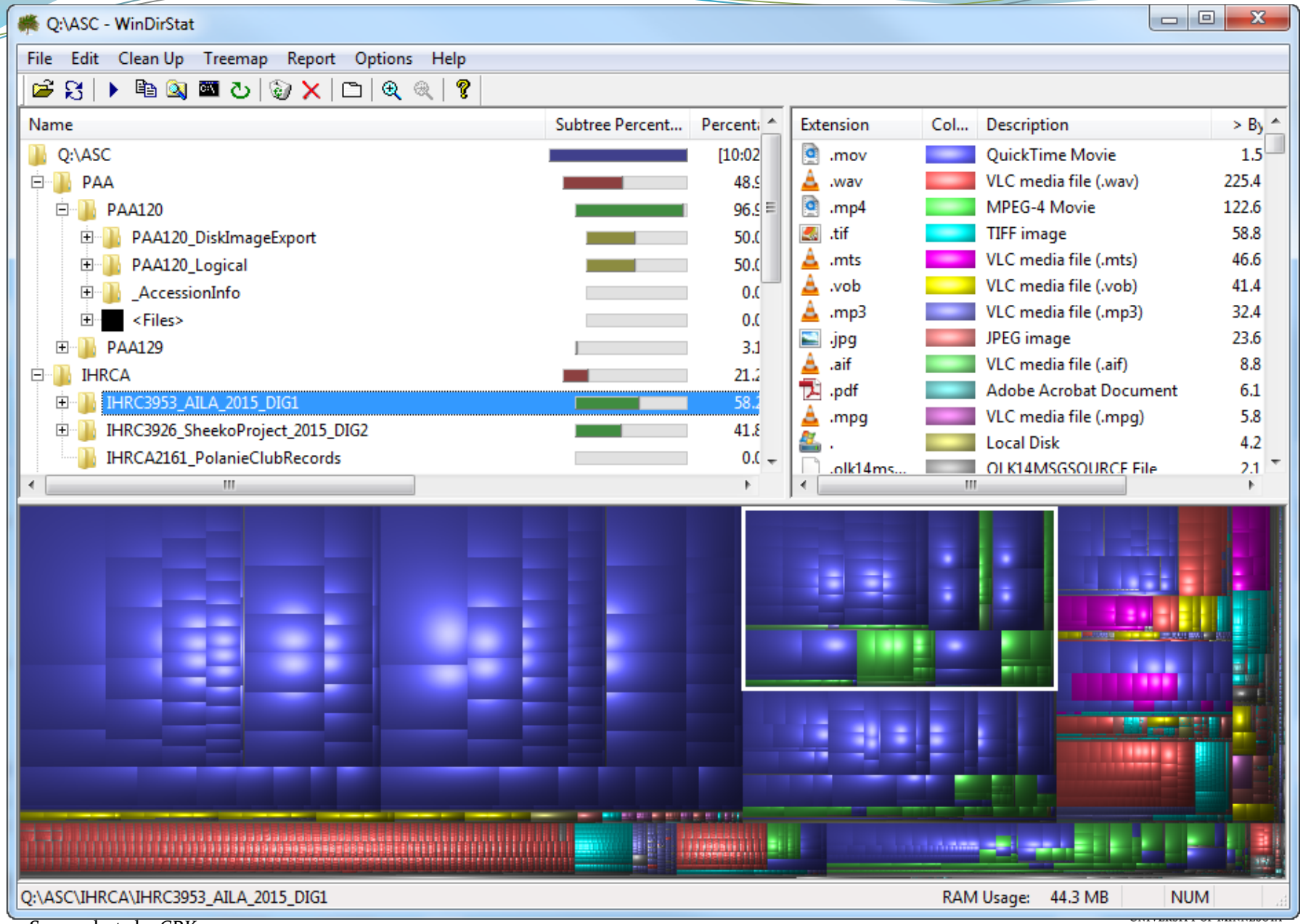



Image: Screenshots by CRKussmann

# WinDirStat

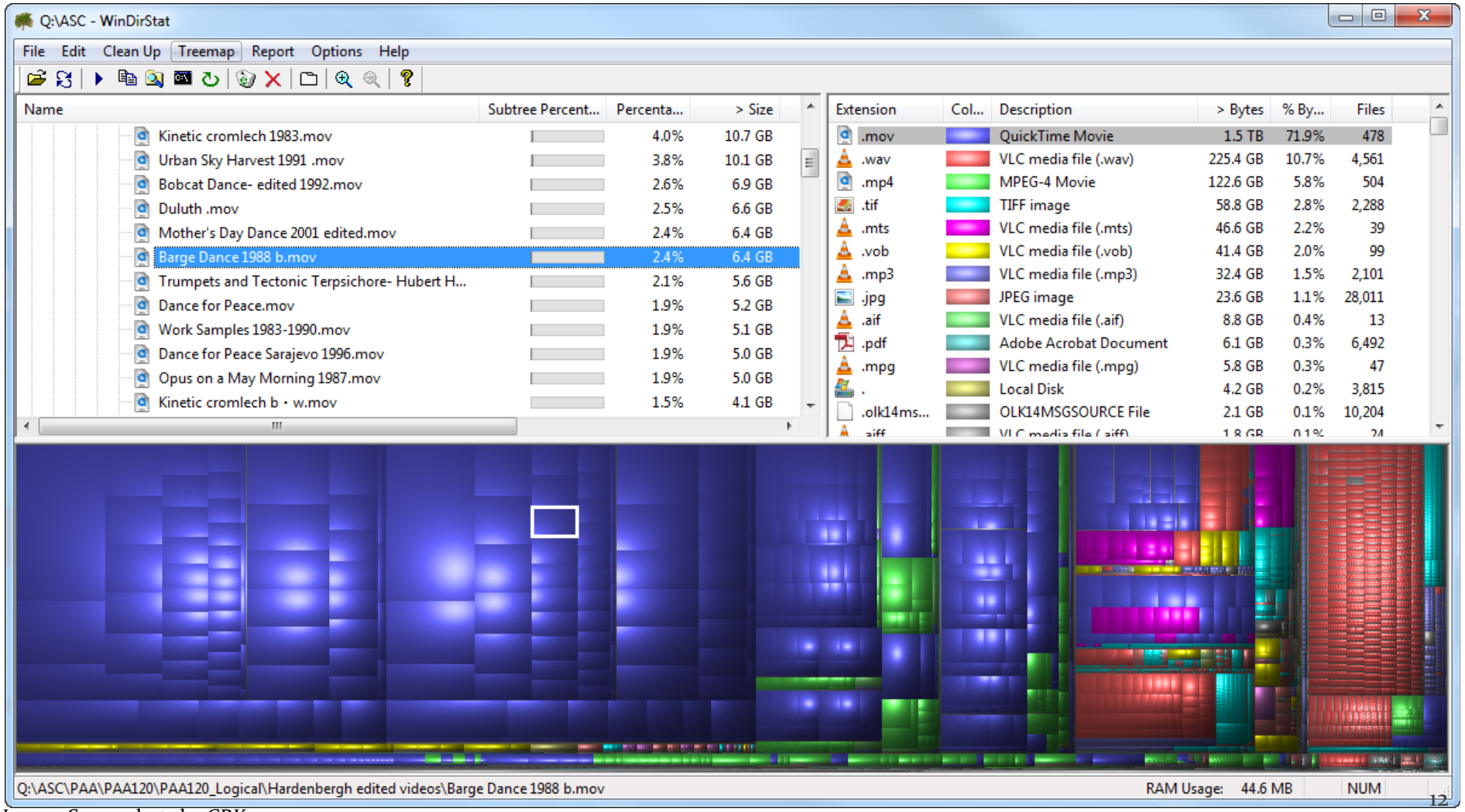



Image: Screenshots by CRKussmann

# WinDirStat

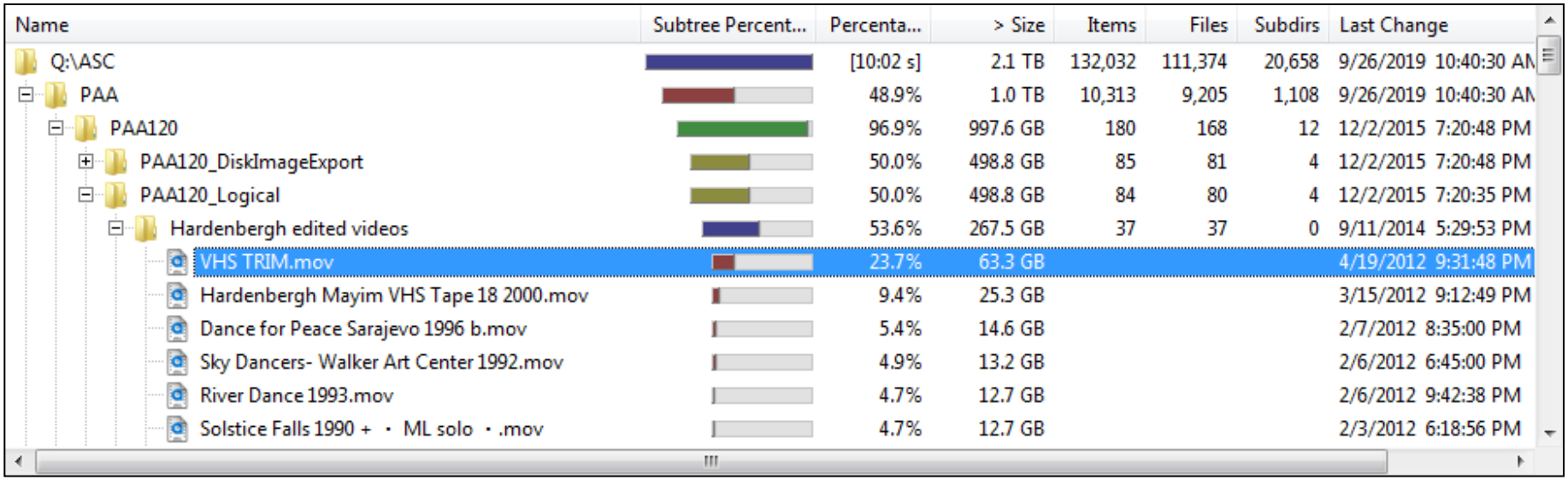
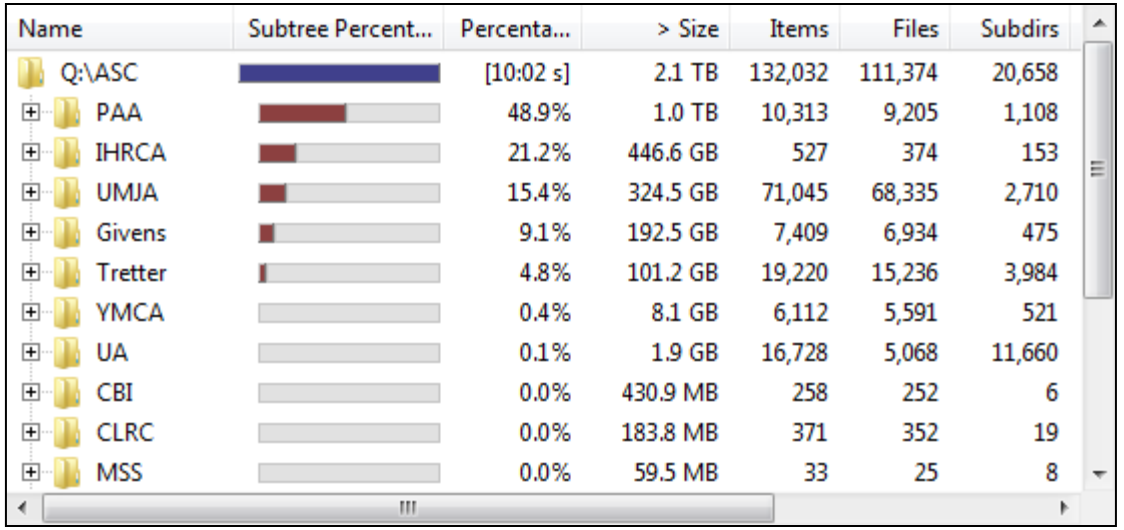



Image: Screenshots by CRKussmann

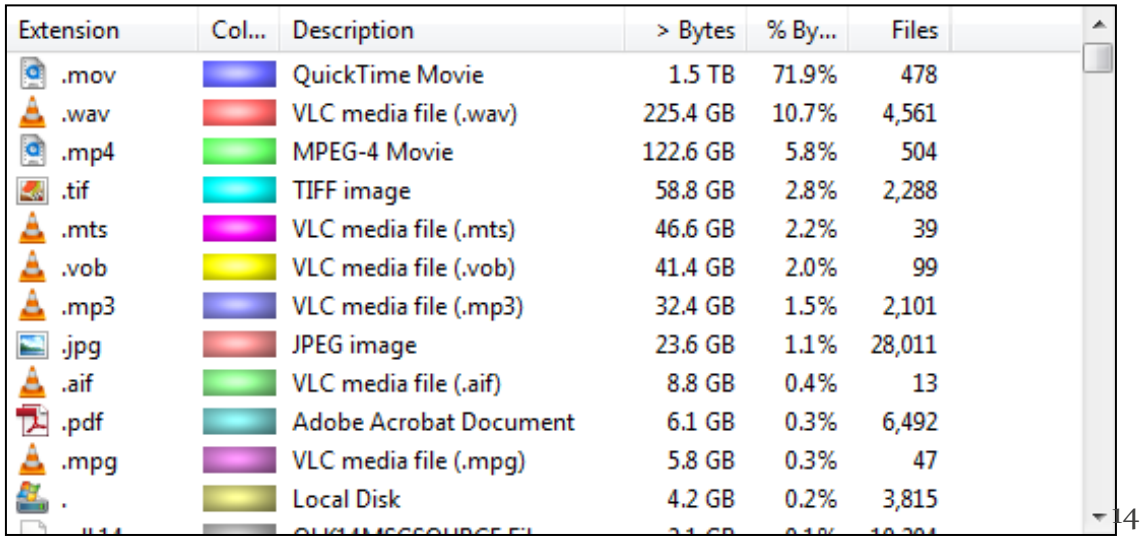
# WinDirStat

- Capture information about collection size, number of files



| Name | Subtree Percent... | Percenta... | > Size | Items | Files | Subdirs |
|---|---|---|---|---|---|---|
| Q:\ASC | | [10:02 s] | 2.1 TB | 132,032 | 111,374 | 20,658 |
| PAA | | 48.9% | 1.0 TB | 10,313 | 9,205 | 1,108 |
| IHRCA | | 21.2% | 446.6 GB | 527 | 374 | 153 |
| UMJA | | 15.4% | 324.5 GB | 71,045 | 68,335 | 2,710 |
| Givens | | 9.1% | 192.5 GB | 7,409 | 6,934 | 475 |
| Tretter | | 4.8% | 101.2 GB | 19,220 | 15,236 | 3,984 |
| YMCA | | 0.4% | 8.1 GB | 6,112 | 5,591 | 521 |
| UA | | 0.1% | 1.9 GB | 16,728 | 5,068 | 11,660 |
| CBI | | 0.0% | 430.9 MB | 258 | 252 | 6 |
| CLRC | | 0.0% | 183.8 MB | 371 | 352 | 19 |
| MSS | | 0.0% | 59.5 MB | 33 | 25 | 8 |

- Capture information about file types



| Extension | Col... | Description | > Bytes | % By... | Files |
|---|---|---|---|---|---|
| .mov | | QuickTime Movie | 1.5 TB | 71.9% | 478 |
| .wav | | VLC media file (.wav) | 225.4 GB | 10.7% | 4,561 |
| .mp4 | | MPEG-4 Movie | 122.6 GB | 5.8% | 504 |
| .tif | | TIFF image | 58.8 GB | 2.8% | 2,288 |
| .mts | | VLC media file (.mts) | 46.6 GB | 2.2% | 39 |
| .vob | | VLC media file (.vob) | 41.4 GB | 2.0% | 99 |
| .mp3 | | VLC media file (.mp3) | 32.4 GB | 1.5% | 2,101 |
| .jpg | | JPEG image | 23.6 GB | 1.1% | 28,011 |
| .aif | | VLC media file (.aif) | 8.8 GB | 0.4% | 13 |
| .pdf | | Adobe Acrobat Document | 6.1 GB | 0.3% | 6,492 |
| .mpg | | VLC media file (.mpg) | 5.8 GB | 0.3% | 47 |
| . | | Local Disk | 4.2 GB | 0.2% | 3,815 |

Image: Screenshots by CRKussmann

# DROID

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

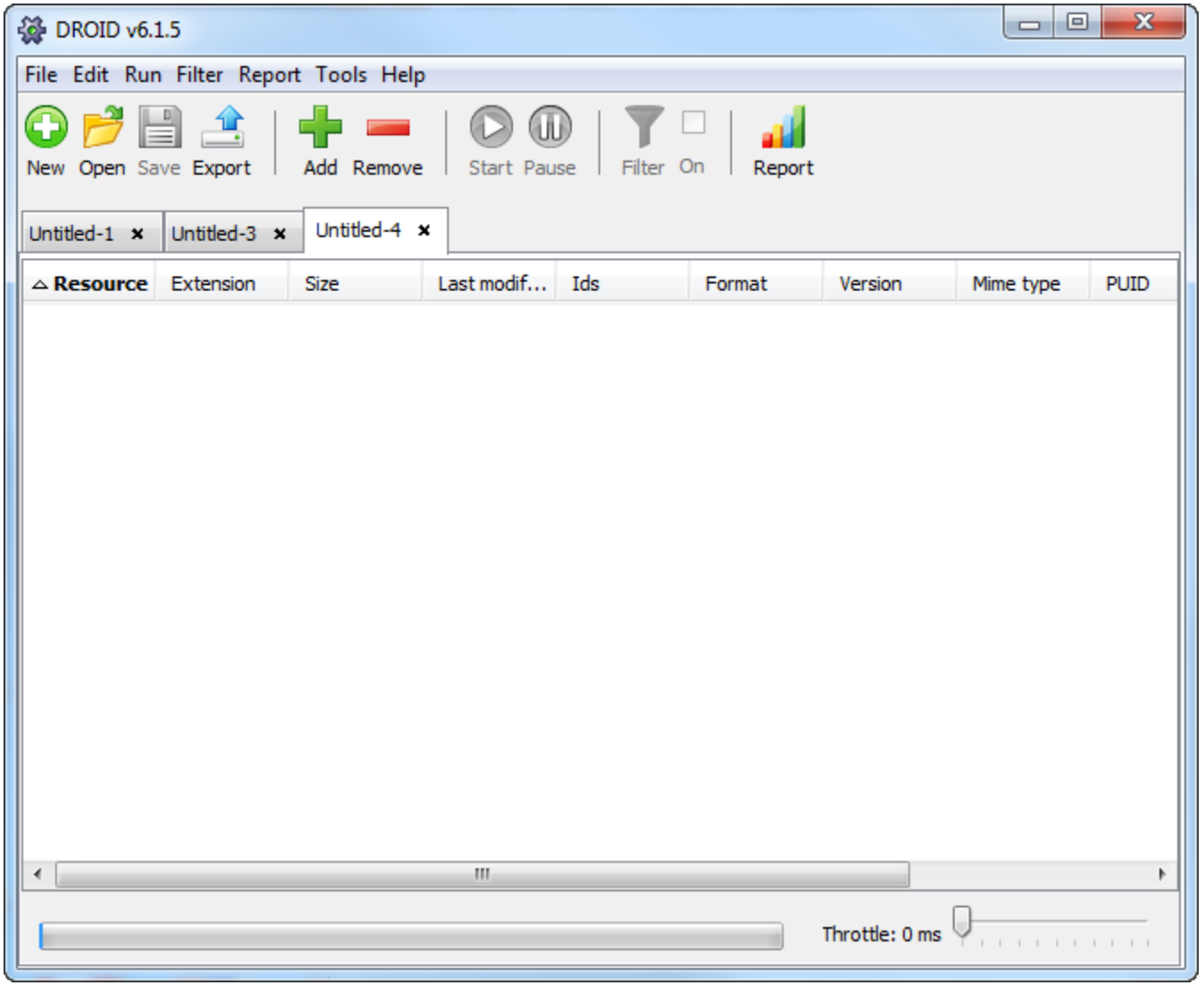



Image: Screenshots by CRKussmann

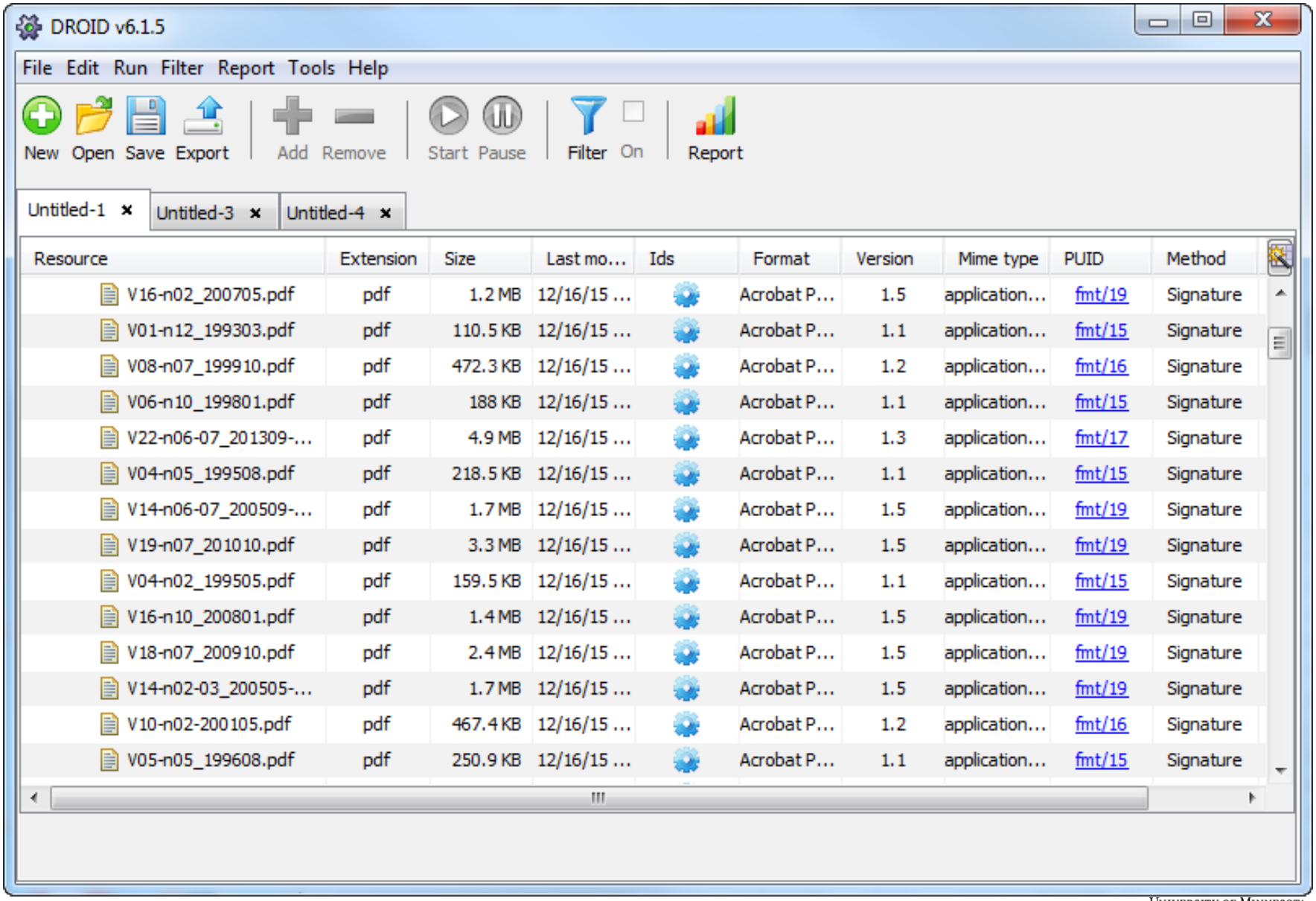
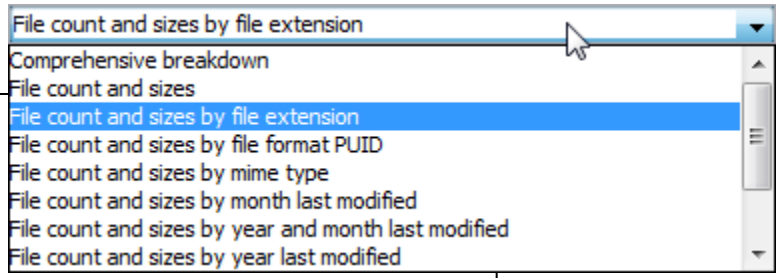



Image: Screenshots by CRKussmann

# DROID Reports




## File count and sizes

### Profile Summary

| Name | Signature version | Container version | Started | Finished | Filters |
|---|---|---|---|---|---|
| DCU-IMAGES | 82 | 20150307 | 14 Aug 2015 | 15 Aug 2015 | |

### File count and sizes

| Report field | Grouping fields |
|---|---|
| FILE_SIZE | |

Filter fields:

| Field | Operator | Values |
|---|---|---|
| RESOURCE_TYPE | NONE_OF | "Folder" |

| Profile | Count | Sum | Min | Max | Average |
|---|---|---|---|---|---|
| DCU-IMAGES | 2867306 | 76267393859435 | 0 | 254240458061 | 26598972 |
| **Profile totals** | **2867306** | **76267393859435** | **0** | **254240458061** | **26598972** |

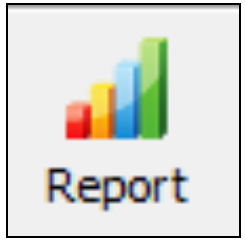



Image: Screenshots by CRKussmann

# DROID Reports

| pdf | | | | | |
|---|---|---|---|---|---|
| Profile | Count | Sum | Min | Max | Average |
| Aerial | 9945 | 192763922072 | 4395 | 218831972 | 19382998 |
| **Profile totals** | **9945** | **192763922072** | **4395** | **218831972** | **19382998** |

| pft | | | | | |
|---|---|---|---|---|---|
| Profile | Count | Sum | Min | Max | Average |
| Aerial | 110 | 25020172 | 313 | 2102111 | 227456 |
| **Profile totals** | **110** | **25020172** | **313** | **2102111** | **227456** |

| pfx | | | | | |
|---|---|---|---|---|---|
| Profile | Count | Sum | Min | Max | Average |
| Aerial | 62 | 4647504 | 24 | 599536 | 74959 |
| **Profile totals** | **62** | **4647504** | **24** | **599536** | **74959** |



Image: Screenshots by CRKussmann

# Database Query

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

```
88257 records found.
88344 bitstreams found.
totalSize = 98.188774 Gb
    application/vnd.ms-excel: 2 files, 1.10327 Mb
    text/richtext: 1 files, 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
```

```
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
```

Image: Screenshots by CRKussmann

# 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)

| Total File Size per file format(bytes) | File Format | Number of Files | Total File Size (GB) | File Format | Total File Size (GB) |
|---|---|---|---|---|---|
| | adf | | 0 | adf | 0 |
| | aft | | 0 | aft | 0 |
| | aiff | | 0 | aiff | 0 |
| | aux | | 0 | aux | 0 |
| 1503238554 | avi | 48 | 1.4 | avi | 1.4 |
| | cr2 | | 0 | cr2 | 0 |
| 32820428.8 | csv | 52 | 0.03056640625 | csv | 0.03056640625 |
| | dng | | 0 | dng | 0 |
| 134637158.4 | doc | 1096 | 0.125390625 | doc | 0.125390625 |
| 391118848 | docx | 4932 | 0.3642578125 | docx | 0.3642578125 |
| | dv | | 0 | dv | 0 |
| | ecw | | 0 | ecw | 0 |
| | epub | | 0 | epub | 0 |
| 118803660.8 | html | 5720 | 0.1106445313 | html | 0.1106445313 |
| | lrprev | | 0 | lrprev | 0 |
| 1258291.2 | jp2 | 1 | 0.001171875 | jp2 | 0.001171875 |
| 15518924.8 | jpeg | 78 | 0.014453125 | jpeg | 0.014453125 |
| 6979321856 | jpg | 6352 | 6.5 | jpg | 6.5 |
| | m4a | | | m4a | |
| 585944268.8 | m4v | 5 | 0.545703125 | m4v | 0.545703125 |

*DSpace - UDC/DRUM | *MDL | *Aerial | *DCImages | *St

Image: Screenshots by CRKussmann

# Google Sheets Totals

| DSpace | Total File Size per file format(bytes) | Number of Files |
|---|---|---|
| adf | | |
| aft | | |
| aiff | | |
| aux | | |
| avi | | |
| or2 | | |
| osv | 328462363.5 | 116 |
| dng | | |
| doc | 43671960.34 | 228 |
| docx | | |
| dv | | |
| eow | | |
| epub | 24699340.23 | 4 |
| html | 9266297.569 | 12 |
| Irprev | | |
| jp2 | | |
| jpeg | 94869461.99 | 39 |
| jpg | | |
| m4a | | |
| m4v | | |
| matlab | 58373216.4 | 130 |
| mobi | 14460066.46 | 2 |
| mov (quicktime) | 16203380687 | 28 |
| mp3 (mpeg) | 1252596283 | 6 |
| mp4 | | |
| mpg | | |
| octet-stream | 48136409065 | 383 |
| pdf | 2075323793.41 | 47572 |
| png | 6342319.276 | 10 |
| ppt (ms-powerpoi | 86002211.23 | 17 |

| MDL | Total File Size per file format(bytes) | Number of Files |
|---|---|---|
| adf | | |
| aft | | |
| aiff | | |
| aux | | |
| avi | 1842502914.42 | 296 |
| or2 | | |
| osv | 52932940 | 664 |
| dng | | |
| doc | 2277549.68 | 8863 |
| docx | 3700746.52 | 19248 |
| dv | 63366511145.45 | 508 |
| eow | | |
| epub | | |
| html | 2018479 | 24 |
| Irprev | | |
| jp2 | 9240259948 | 613 |
| jpeg | 9908 | 2 |
| jpg | 5918418774 | 40771 |
| m4a | | |
| m4v | | |
| matlab | | |
| mobi | | |
| mov (quicktime) | 7194282048.8 | 374 |
| mp3 (mpeg) | 6983724.4176 | 2183 |
| mp4 | 5437240476.3 | 1060 |
| mpg | | |
| octet-stream | | |
| pdf | 516028408.3 | 5228 |
| png | 2706968 | 46 |
| ppt (ms-powerpoint) | | |

| Aerial | Total File Size per file format(bytes) | Number of Files |
|---|---|---|
| adf | 8369676704 | 228 |
| aft | 49504224 | 144 |
| aiff | | |
| aux | 25594650 | 1631 |
| avi | | |
| or2 | | |
| osv | 9624710 | 47 |
| dng | | |
| doc | 36918238 | 643 |
| docx | | |
| dv | | |
| eow | 7504463956 | 20 |
| epub | | |
| html | | |
| Irprev | | |
| jp2 | 4007007358951 | 1625446 |
| m4a | | |
| m4v | | |
| matlab | | |
| mobi | | |
| mov (quicktime) | | |
| mp3 (mpeg) | | |
| mp4 | | |
| mpg | | |
| octet-stream | | |
| pdf | 1927639220.72 | 9945 |
| png | 4505399 | 548 |
| ppt (ms-powerpoint) | | |

| DCImages | Total File Size per file format(bytes) | Number of Files |
|---|---|---|
| adf | | |
| aft | | |
| aiff | 14651031.36 | 15 |
| aux | | |
| avi | 3491096766.66 | 63 |
| or2 | 1445486337.29 | 8331 |
| osv | | |
| dng | 6452280769.66 | 36623 |
| doc | 366219757 | 2255 |
| docx | 25176103.2 | 294 |
| dv | 4182975715.280 | 2107 |
| eow | | |
| epub | | |
| html | | |
| Irprev | 31689768.89 | 12683 |
| jp2 | 6497694588.314 | 675281 |
| jpeg | 43417.632 | 24 |
| jpg | 3805883888.948 | 94036 |
| m4a | | |
| m4v | 122717841.390 | 204 |
| matlab | | |
| mobi | | |
| mov (quicktime) | 4079134950.480 | 1672 |
| mp3 (mpeg) | 550843.2154 | 1247 |
| mp4 | 1844842.596.185 | 1347 |
| mpg | | |
| octet-stream | | |
| pdf | 182.478425.944 | 25381 |
| png | 1.364.657.365 | 366 |
| ppt (ms-powerpoi | 100.31.12.960 | 10 |

| A3C Data Drive | Total File Size per file format(bytes) | Number of Files |
|---|---|---|
| adf | | |
| aft | | |
| aiff | | |
| aux | | |
| avi | 150.32.38.65.4 | 48 |
| or2 | | |
| osv | 32.62.04.28.8 | 52 |
| dng | | |
| doc | 1.34.63.71.58.4 | 1.086 |
| docx | 39.11.18.648 | 4.932 |
| dv | | |
| eow | | |
| epub | | |
| html | 1.18.80.36.60.8 | 5.720 |
| Irprev | | |
| jp2 | 1.26.82.91.2 | 1 |
| jpeg | 155.18.92.4.8 | 78 |
| jpg | 6.97.93.21.856 | 6.352 |
| m4a | | |
| m4v | 5.85.94.42.68.8 | 5 |
| matlab | | |
| mobi | | |
| mov (quicktime) | 1.53.93.16.27.88.86 | 425 |
| mp3 (mpeg) | 22.22.64.65.757 | 2.277 |
| mp4 | 1.32.71.44.89.446 | 503 |
| mpg | | |
| octet-stream | | |
| pdf | 9.46.23.49.62.4 | 1.966 |
| png | 93.11.35.48.8 | 608 |
| ppt (ms-powerpoi | 1.61.48.07.0.4 | 8 |

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)

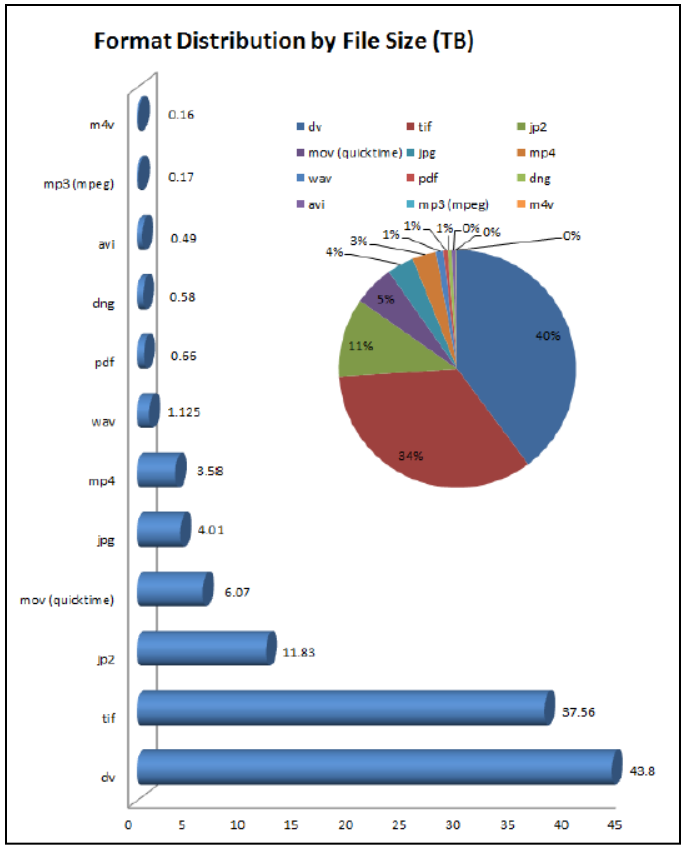
A variety of file formats are found within the Libraries collections - the formats displayed in the charts below represent only the top formats by number of files and by total amount of space they use. The majority of the files in our holdings represents images and textual based documents; while video 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.

Image: Screenshots by CRKussmann

# Grand Totals

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



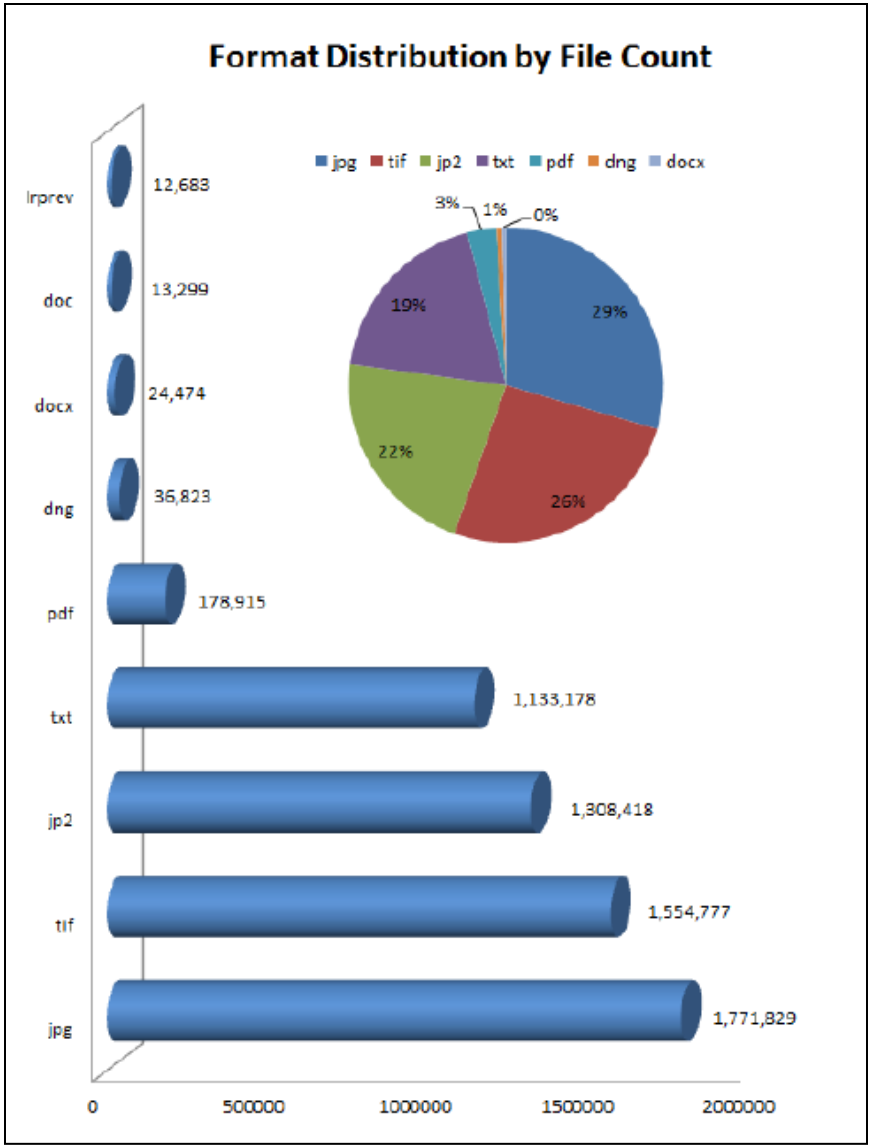
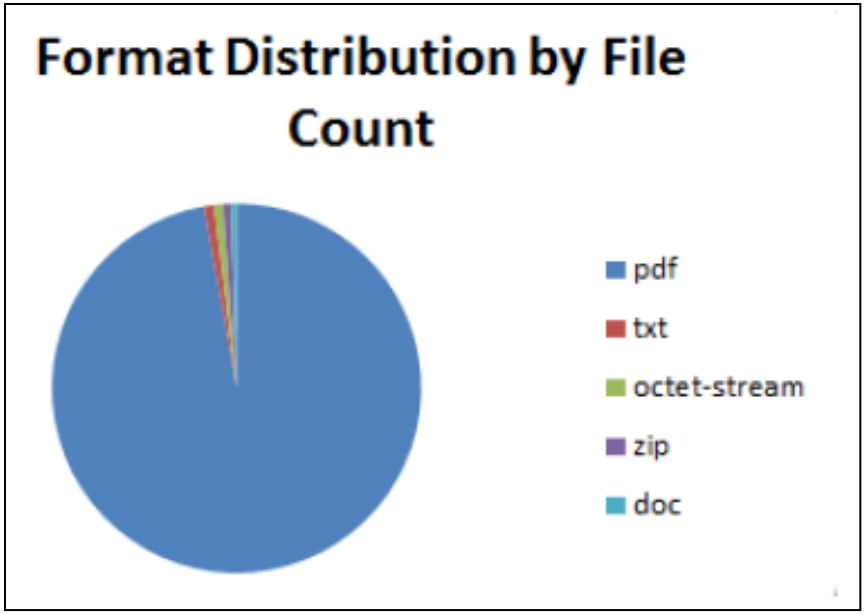




Image: Screenshots by CRKussmann

# Sample Report 1

- 49, 506 unique files
- 320 GB of content



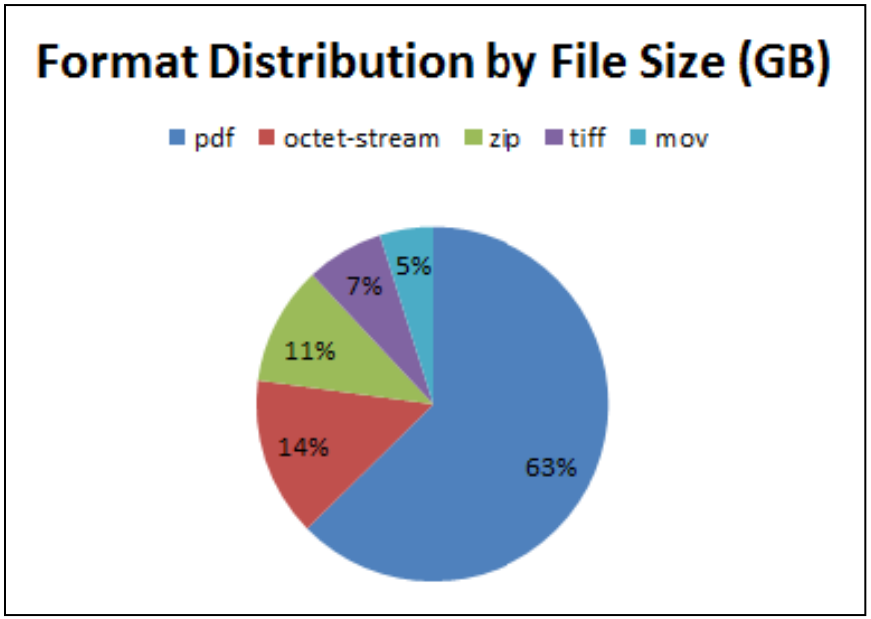
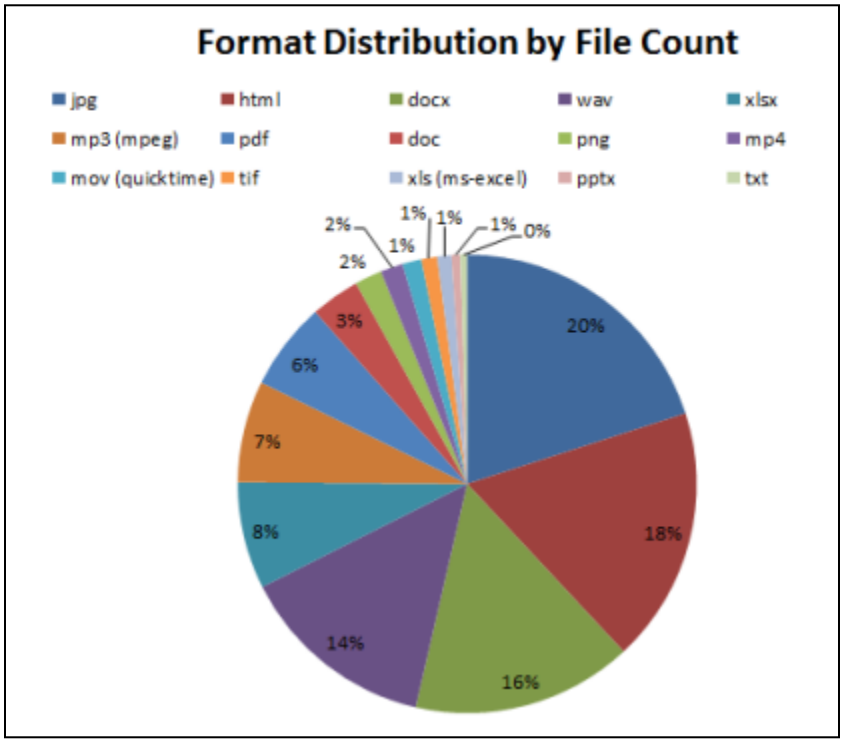




Image: Screenshots by CRKussmann

# Sample Report 2

- 32, 039 unique files
- 1769 GB of content



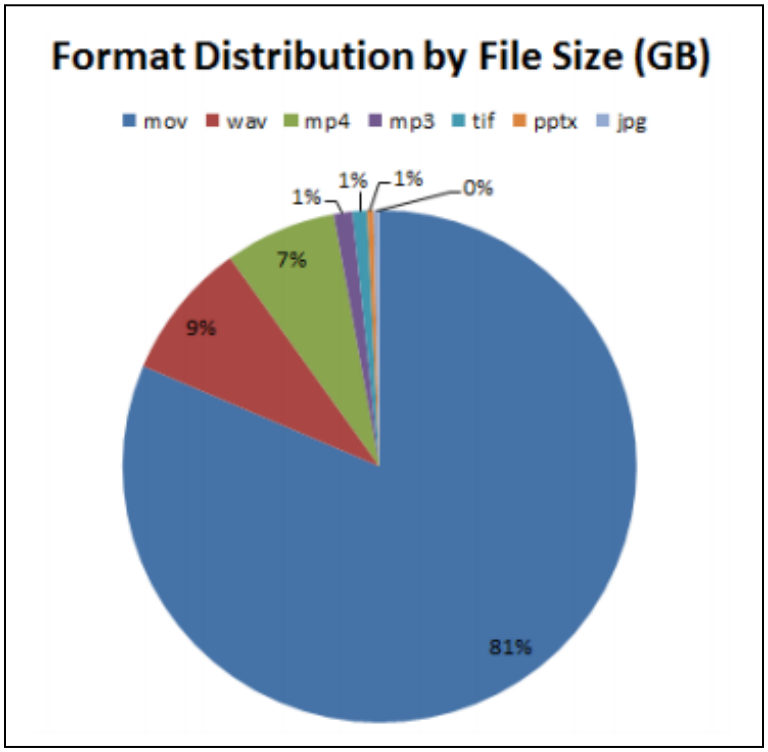
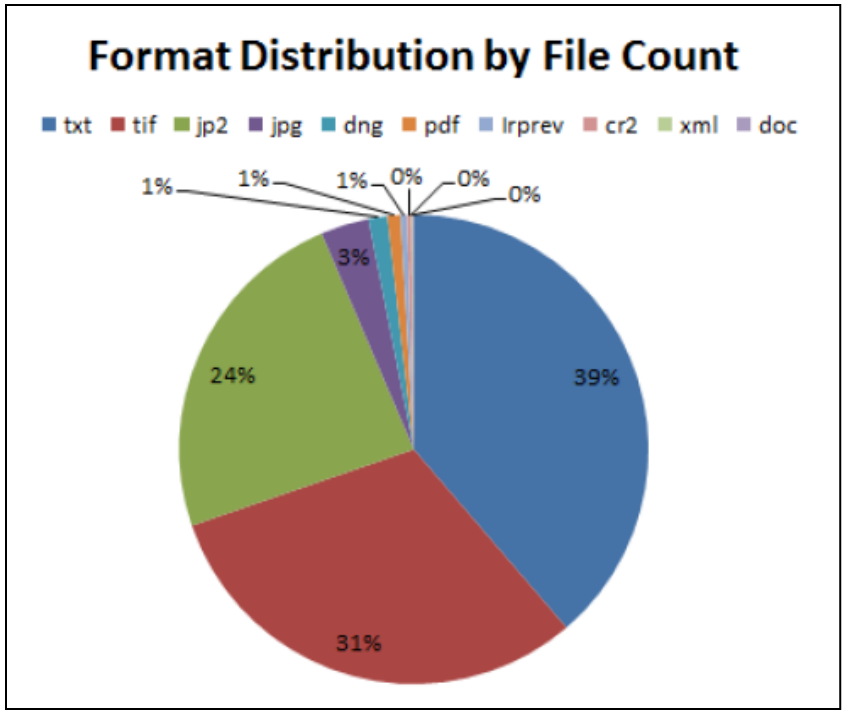




Image: Screenshots by CRKussmann

# Sample Report 3

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



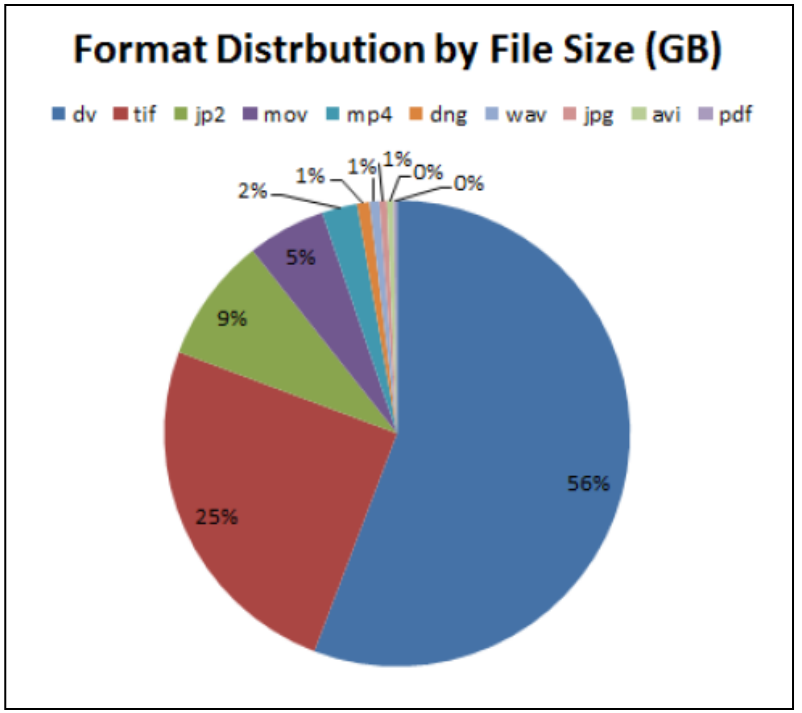




Image: Screenshots by CRKussmann

# 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

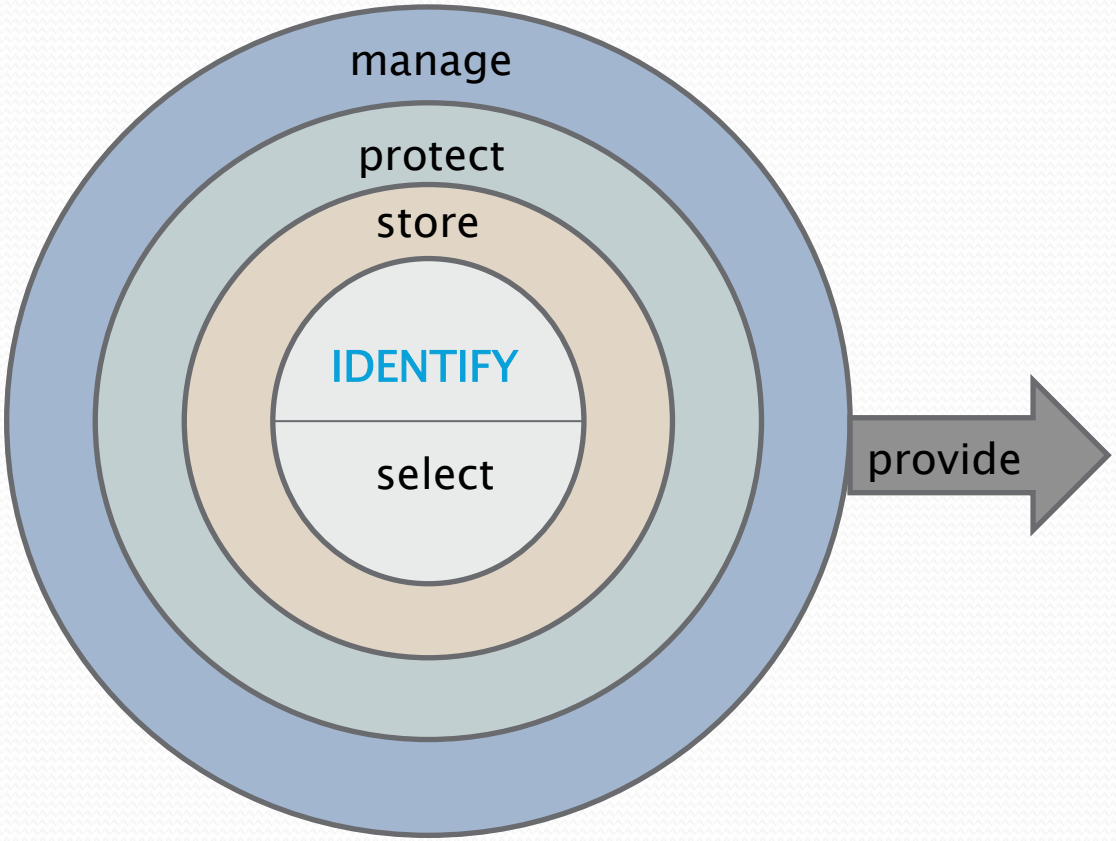
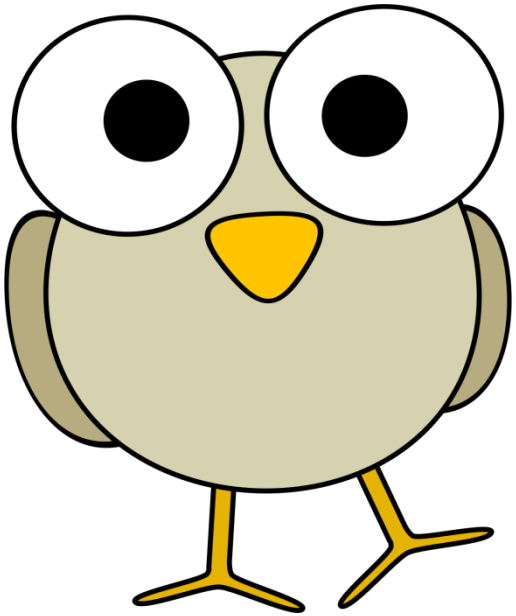



Image: Digital Preservation Outreach & Education curriculum materials

# Resource Links

- DROID: http://z.umn.edu/147h
- WinDirStat: http://windirstat.info/
- Tools Sheet: http://z.umn.edu/147i

# Questions?



??

**Contact Me**

Carol Kussmann
kussmann@umn.edu

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