



# Surprise! Look What We Have!?

Carol Kussmann  
Digital Preservation Analyst  
University of Minnesota Libraries

March 16, 2016

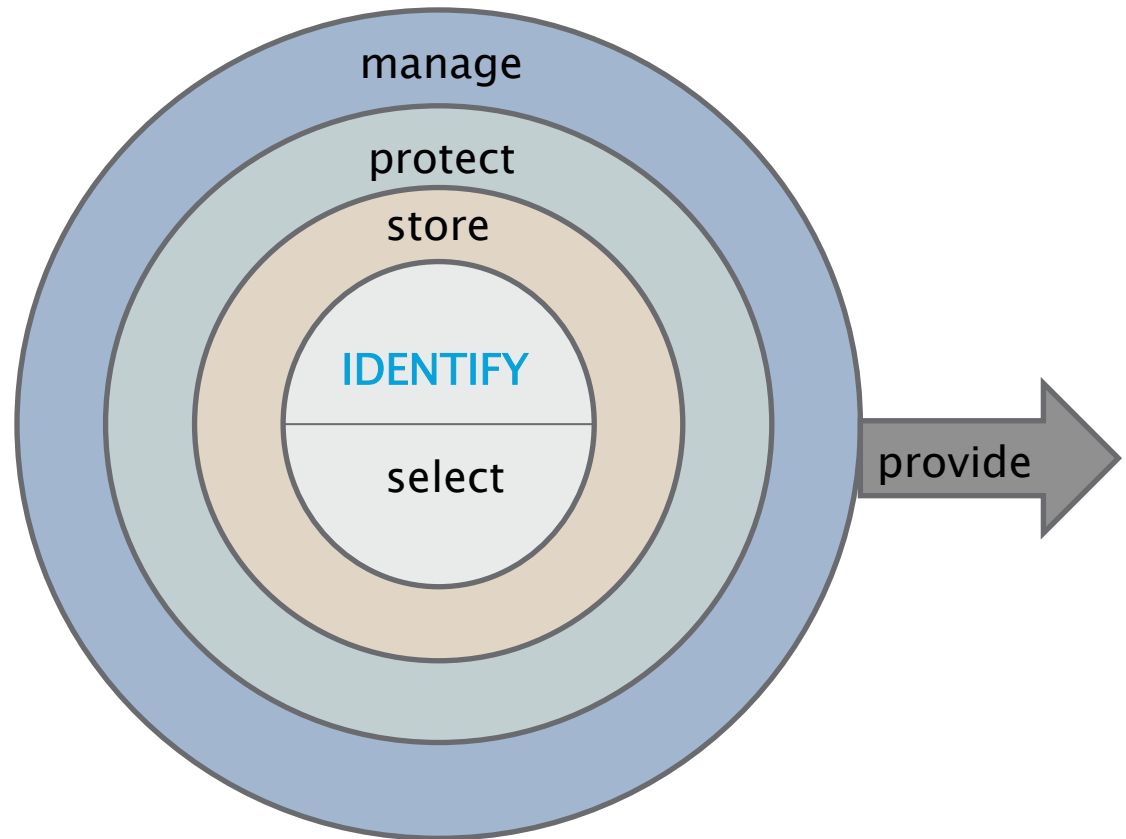
Library Technology Conference; Macalester College, St. Paul, MN

# 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



# 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

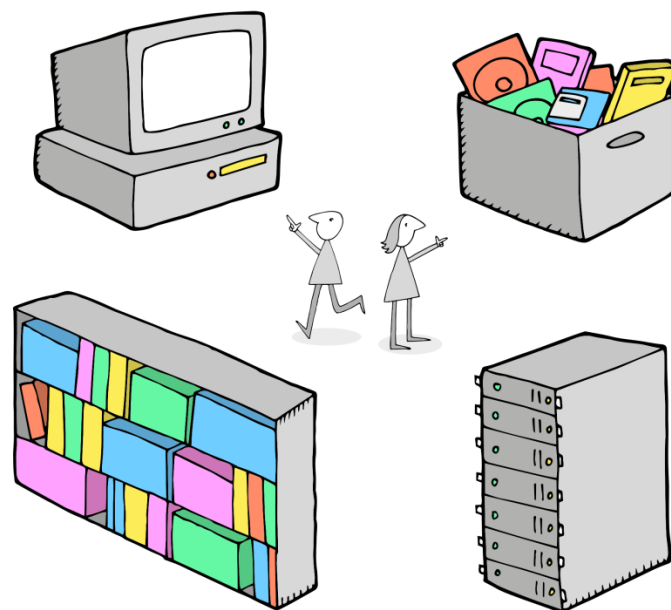


# Location of Materials

- 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?



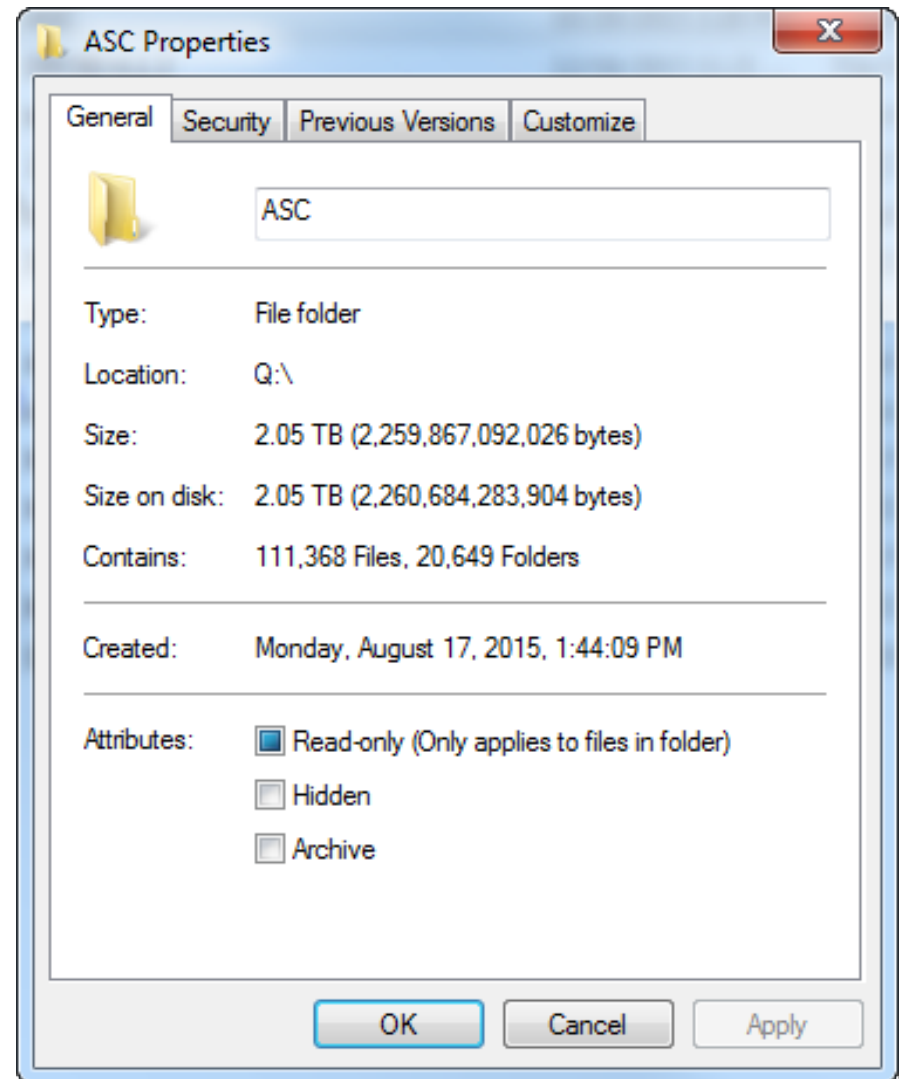
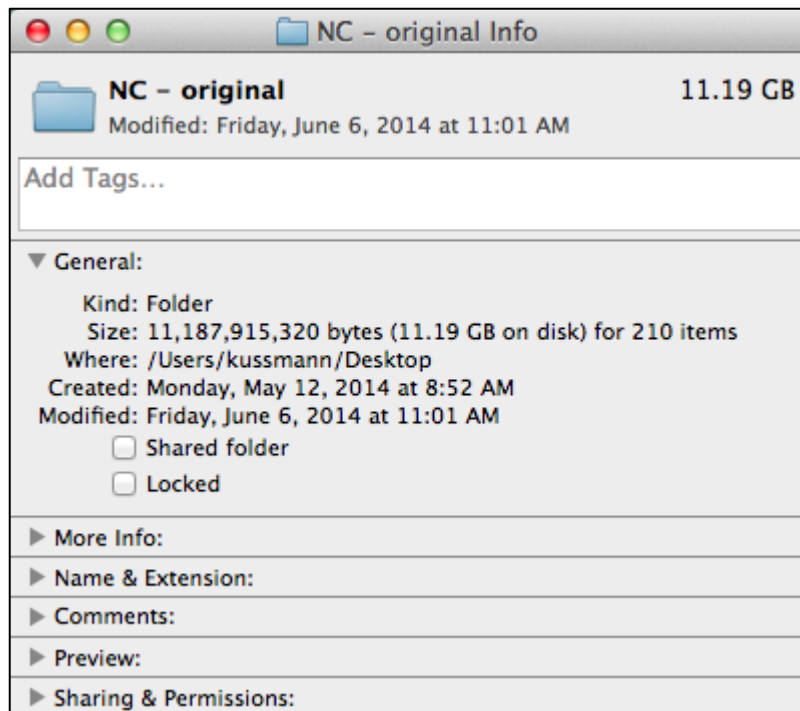
# Tools Used

- 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

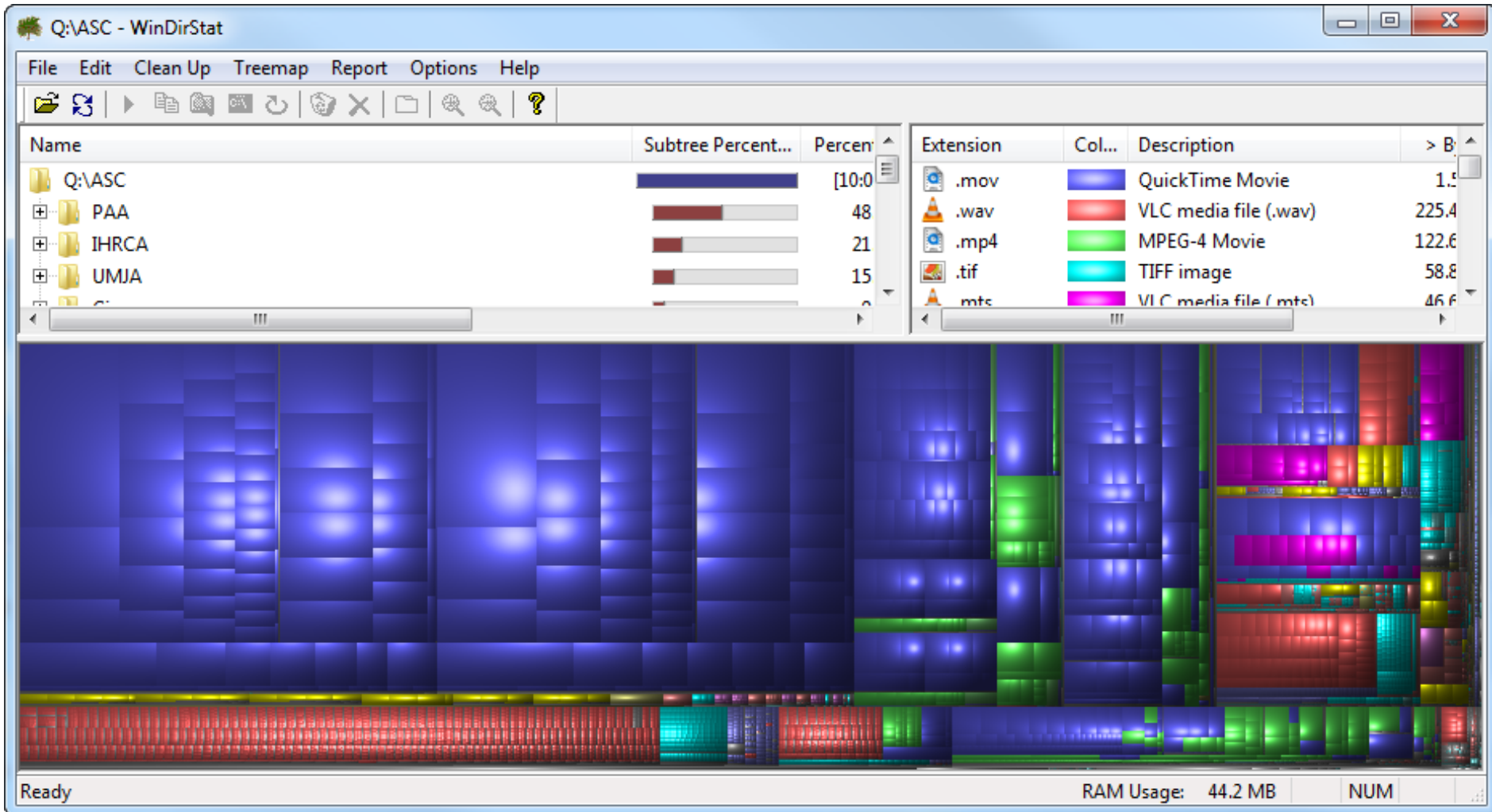




# WinDirStat

Name	Subtree Percent...	Percent...	> Size	Items	Files	Subdirs	Last Change
Q:\ASC		[1:12 s]	1.6 TB	20,449	16,415	4,034	2/1/2016 9:19:16 PM
PAA		63.3%	1017.5 GB	3,529	3,231	298	12/2/2015 7:20:48 PM
IHRCA		27.8%	446.6 GB	527	374	153	10/14/2015 9:15:22 PM
Givens		5.6%	90.3 GB	3,694	3,228	466	7/23/2015 1:33:21 PM
UMJA		2.2%	36.0 GB	4,366	3,840	526	12/16/2015 5:43:07 PM
Tretter		0.6%	9.1 GB	2,659	2,273	386	12/2/2015 7:15:07 PM
YMCA		0.4%	6.5 GB	2,681	2,437	244	12/4/2015 4:31:43 PM
CBI		0.0%	430.9 MB	258	252	6	2/1/2016 9:19:16 PM
CLRC		0.0%	183.8 MB	371	352	19	10/22/2015 4:19:01 PM
UA		0.0%	166.7 MB	2,315	401	1,914	1/7/2016 6:08:10 PM
MSS		0.0%	59.5 MB	33	25	8	5/13/2015 9:05:11 PM
<Files>		0.0%	4.7 MB	2	2	0	12/17/2015 8:56:31 PM
Kuss test		0.0%	0	0	0	0	9/3/2015 3:18:55 PM
NAA		0.0%	0	0	0	0	10/31/2014 3:18:36 PM
SCRB		0.0%	0	0	0	0	10/31/2014 3:18:51 PM
SWHA		0.0%	0	0	0	0	10/31/2014 3:19:02 PM

# WinDirStat



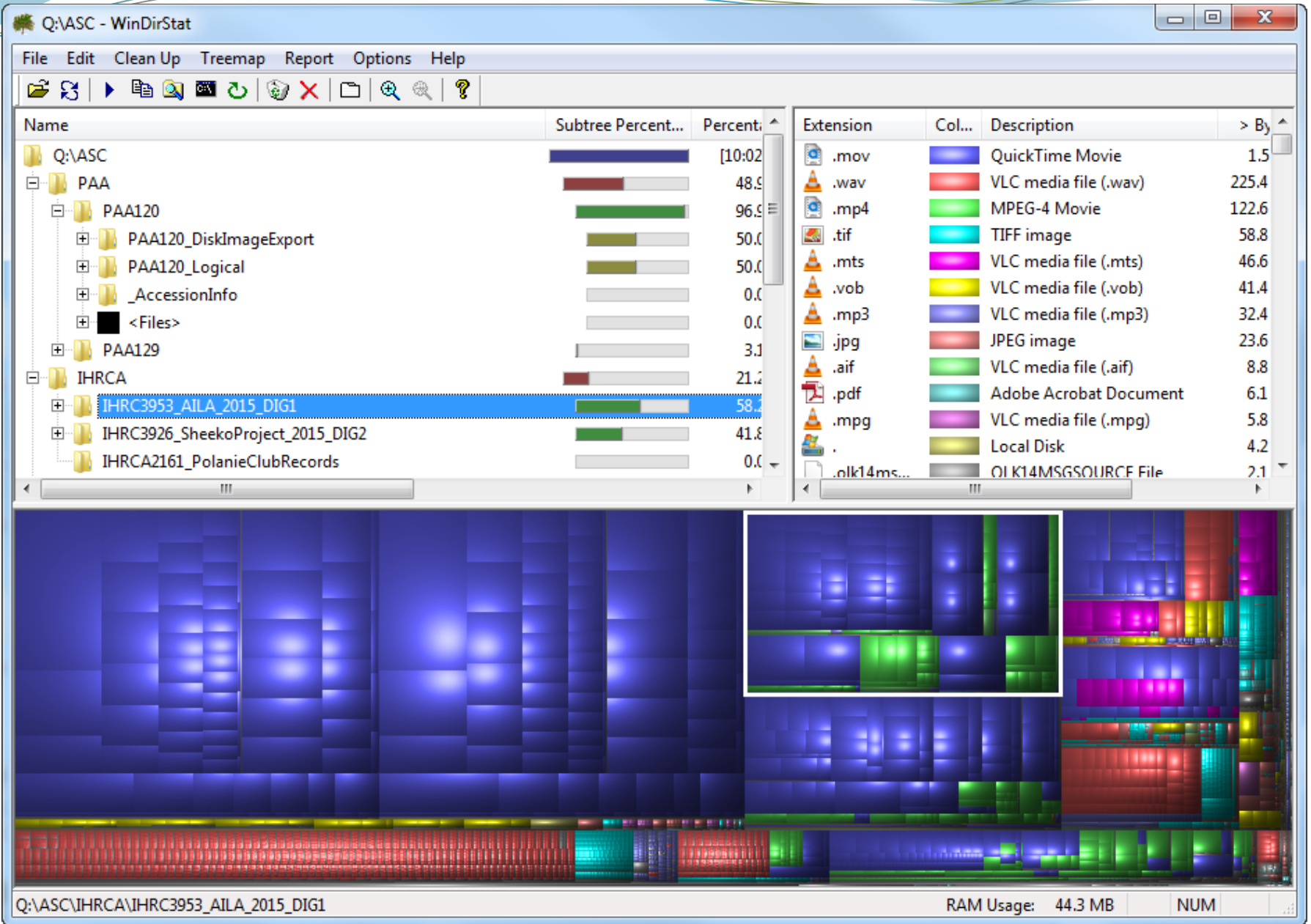


Image: Screenshots by CRKusmann

# WinDirStat

The screenshot displays the WinDirStat application window. The top menu bar includes File, Edit, Clean Up, Treemap, Report, Options, and Help. The main window is divided into two primary sections: a file list on the left and a treemap visualization on the right.

**File List:**

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

**Extension List:**

Extension	Col...	Description	> Bytes	% By...	Files
.mov	Blue	QuickTime Movie	1.5 TB	71.9%	478
.wav	Red	VLC media file (.wav)	225.4 GB	10.7%	4,561
.mp4	Green	MPEG-4 Movie	122.6 GB	5.8%	504
.tif	Cyan	TIFF image	58.8 GB	2.8%	2,288
.mts	Magenta	VLC media file (.mts)	46.6 GB	2.2%	39
.vob	Yellow	VLC media file (.vob)	41.4 GB	2.0%	99
.mp3	Dark Blue	VLC media file (.mp3)	32.4 GB	1.5%	2,101
.jpg	Red	JPEG image	23.6 GB	1.1%	28,011
.aif	Green	VLC media file (.aif)	8.8 GB	0.4%	13
.pdf	Cyan	Adobe Acrobat Document	6.1 GB	0.3%	6,492
.mpg	Purple	VLC media file (.mpg)	5.8 GB	0.3%	47
.	Yellow-Green	Local Disk	4.2 GB	0.2%	3,815
.olk14ms...	Grey	OLK14MSGSOURCE File	2.1 GB	0.1%	10,204
.aiff	Dark Grey	VLC media file (.aiff)	1.8 GB	0.1%	24

**Treemap Visualization:**

The treemap visualization at the bottom of the window shows a hierarchical view of the file system. The largest area is dominated by blue blocks, representing the .mov files. Other colors include red, green, cyan, magenta, yellow, and purple, corresponding to the file extensions listed in the table above. A white square highlights a specific block within the blue area.

**System Information:**

Q:\ASC\PA120\PA120\_Logical\Hardenbergh edited videos\Barge Dance 1988 b.mov  
RAM Usage: 44.6 MB  
NUM 12

# WinDirStat

Name	Subtree Percent...	Percenta...	> Size	Items	Files	Subdirs	Last Change
Q:\ASC		[10:02 s]	2.1 TB	132,032	111,374	20,658	9/26/2019 10:40:30 AM
PAA		48.9%	1.0 TB	10,313	9,205	1,108	9/26/2019 10:40:30 AM
PAA120		96.9%	997.6 GB	180	168	12	12/2/2015 7:20:48 PM
PAA120_DiskImageExport		50.0%	498.8 GB	85	81	4	12/2/2015 7:20:48 PM
PAA120_Logical		50.0%	498.8 GB	84	80	4	12/2/2015 7:20:35 PM
Hardenbergh edited videos		53.6%	267.5 GB	37	37	0	9/11/2014 5:29:53 PM
VHS TRIM.mov		23.7%	63.3 GB				4/19/2012 9:31:48 PM
Hardenbergh Mayim VHS Tape 18 2000.mov		9.4%	25.3 GB				3/15/2012 9:12:49 PM
Dance for Peace Sarajevo 1996 b.mov		5.4%	14.6 GB				2/7/2012 8:35:00 PM
Sky Dancers- Walker Art Center 1992.mov		4.9%	13.2 GB				2/6/2012 6:45:00 PM
River Dance 1993.mov		4.7%	12.7 GB				2/6/2012 9:42:38 PM
Solstice Falls 1990 + · ML solo · .mov		4.7%	12.7 GB				2/3/2012 6:18:56 PM

# WinDirStat

- Capture information about collection size, number of files

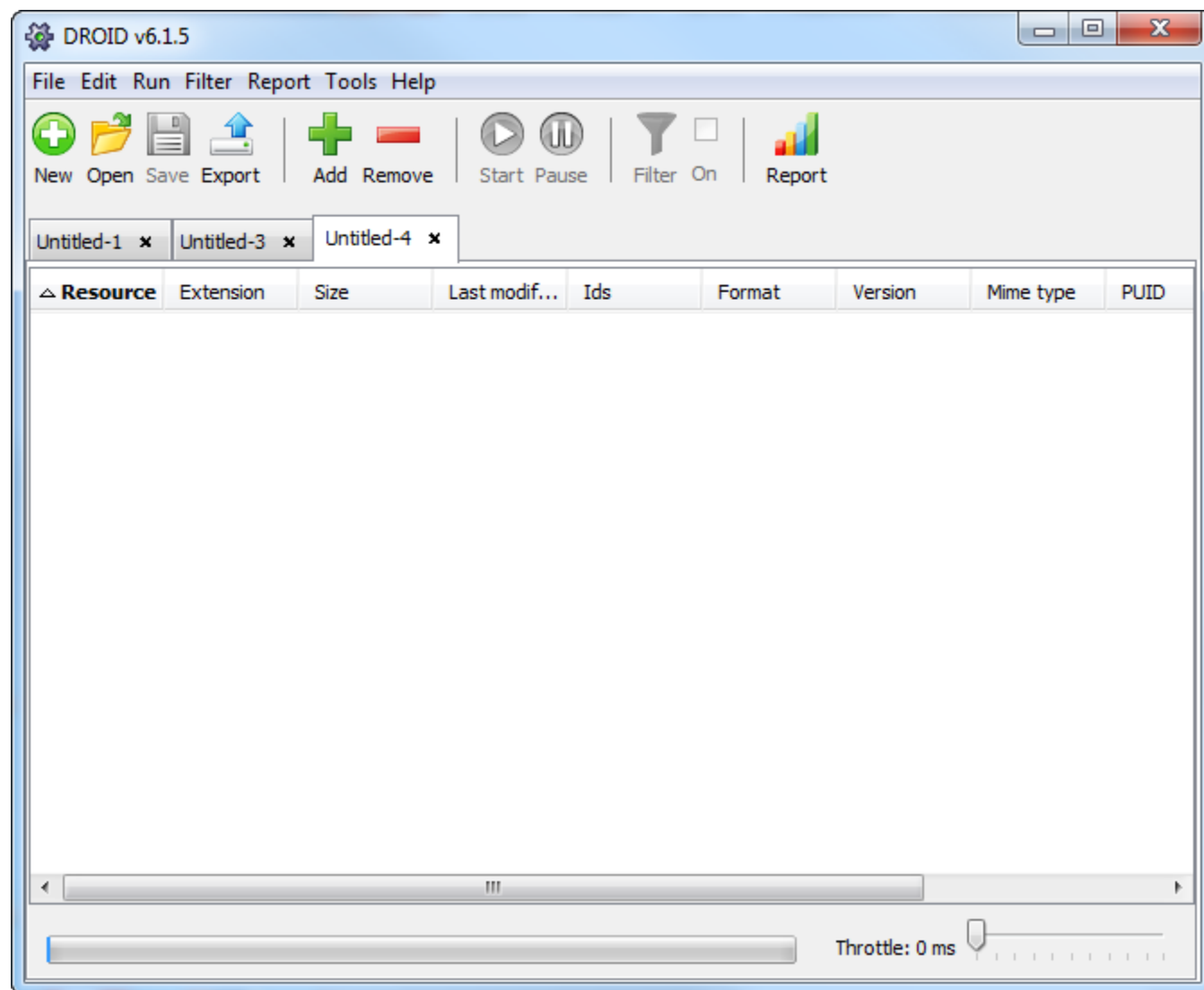
Name	Subtree Percent...	Percenta...	> Size	Items	Files	Subdirs
Q:\ASC		[10:02 s]	2.1 TB	132,032	111,374	20,658
PAA		48.9%	1.0 TB	10,313	9,205	1,108
IHRCA		21.2%	446.6 GB	527	374	153
UMJA		15.4%	324.5 GB	71,045	68,335	2,710
Givens		9.1%	192.5 GB	7,409	6,934	475
Tretter		4.8%	101.2 GB	19,220	15,236	3,984
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

# DROID

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



**DROID v6.1.5**

File Edit Run Filter Report Tools Help

New Open Save Export Add Remove Start Pause Filter On Report

Untitled-1 x Untitled-3 x Untitled-4 x

Resource	Extension	Size	Last mo...	Ids	Format	Version	Mime type	PUID	Method
V16-n02_200705.pdf	pdf	1.2 MB	12/16/15 ...		Acrobat P...	1.5	application...	<a href="#">fmt/19</a>	Signature
V01-n12_199303.pdf	pdf	110.5 KB	12/16/15 ...		Acrobat P...	1.1	application...	<a href="#">fmt/15</a>	Signature
V08-n07_199910.pdf	pdf	472.3 KB	12/16/15 ...		Acrobat P...	1.2	application...	<a href="#">fmt/16</a>	Signature
V06-n10_199801.pdf	pdf	188 KB	12/16/15 ...		Acrobat P...	1.1	application...	<a href="#">fmt/15</a>	Signature
V22-n06-07_201309-...	pdf	4.9 MB	12/16/15 ...		Acrobat P...	1.3	application...	<a href="#">fmt/17</a>	Signature
V04-n05_199508.pdf	pdf	218.5 KB	12/16/15 ...		Acrobat P...	1.1	application...	<a href="#">fmt/15</a>	Signature
V14-n06-07_200509-...	pdf	1.7 MB	12/16/15 ...		Acrobat P...	1.5	application...	<a href="#">fmt/19</a>	Signature
V19-n07_201010.pdf	pdf	3.3 MB	12/16/15 ...		Acrobat P...	1.5	application...	<a href="#">fmt/19</a>	Signature
V04-n02_199505.pdf	pdf	159.5 KB	12/16/15 ...		Acrobat P...	1.1	application...	<a href="#">fmt/15</a>	Signature
V16-n10_200801.pdf	pdf	1.4 MB	12/16/15 ...		Acrobat P...	1.5	application...	<a href="#">fmt/19</a>	Signature
V18-n07_200910.pdf	pdf	2.4 MB	12/16/15 ...		Acrobat P...	1.5	application...	<a href="#">fmt/19</a>	Signature
V14-n02-03_200505-...	pdf	1.7 MB	12/16/15 ...		Acrobat P...	1.5	application...	<a href="#">fmt/19</a>	Signature
V10-n02-200105.pdf	pdf	467.4 KB	12/16/15 ...		Acrobat P...	1.2	application...	<a href="#">fmt/16</a>	Signature
V05-n05_199608.pdf	pdf	250.9 KB	12/16/15 ...		Acrobat P...	1.1	application...	<a href="#">fmt/15</a>	Signature



# DROID Reports

- File count and sizes by file extension
- Comprehensive breakdown
- File count and sizes
- File count and sizes by file extension**
- File count and sizes by file format PUID
- File count and sizes by mime type
- File count and sizes by month last modified
- File count and sizes by year and month last modified
- File count and sizes by year last modified

## File count and sizes

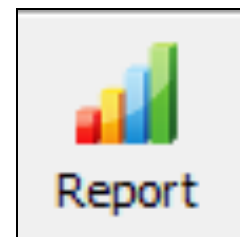
### Profile Summary

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
<b>Profile totals</b>	<b>2867306</b>	<b>76267393859435</b>	<b>0</b>	<b>254240458061</b>	<b>26598972</b>



# DROID Reports

pdf					
Profile	Count	Sum	Min	Max	Average
Aerial	9945	192763922072	4395	218831972	19382998
<b>Profile totals</b>	<b>9945</b>	<b>192763922072</b>	<b>4395</b>	<b>218831972</b>	<b>19382998</b>

pft					
Profile	Count	Sum	Min	Max	Average
Aerial	110	25020172	313	2102111	227456
<b>Profile totals</b>	<b>110</b>	<b>25020172</b>	<b>313</b>	<b>2102111</b>	<b>227456</b>

pfx					
Profile	Count	Sum	Min	Max	Average
Aerial	62	4647504	24	599536	74959
<b>Profile totals</b>	<b>62</b>	<b>4647504</b>	<b>24</b>	<b>599536</b>	<b>74959</b>

# 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
```

# 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



# 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:

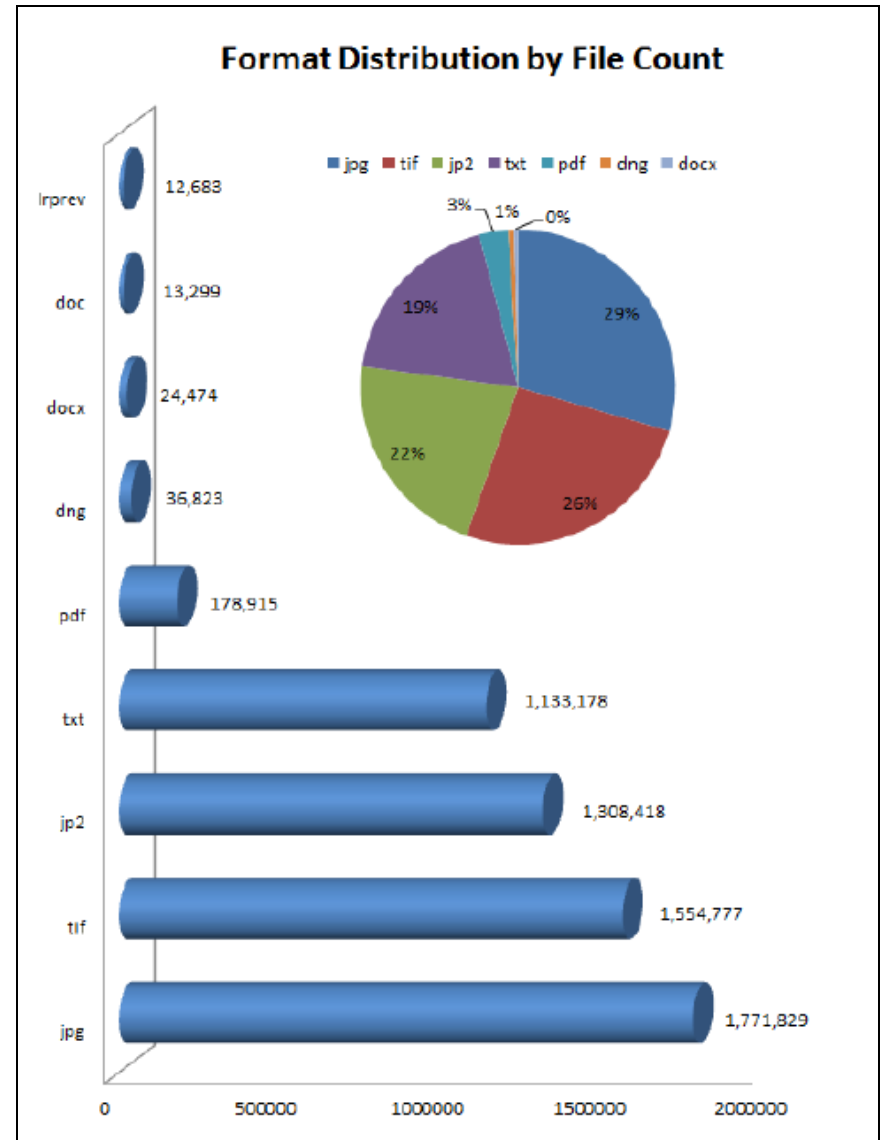
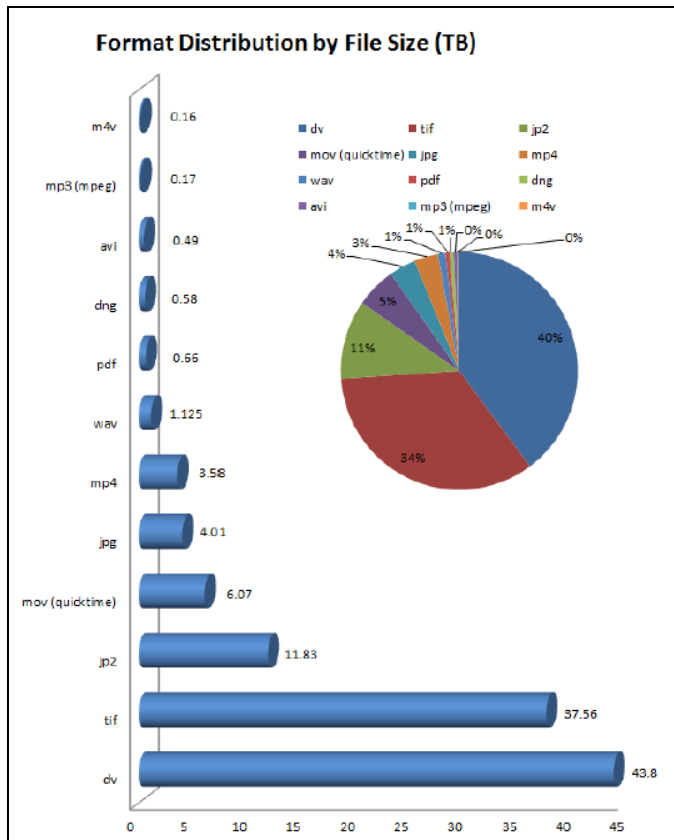
- University Digital Conservancy / Data Repository of Minnesota
- UMedia Archive
- AgEcon Repository
- Library initiated digitization projects of Library materials
- Archives and Special Collections Materials
- Minnesota Digital Library master scans
- Master files of special projects initiated/hosted 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

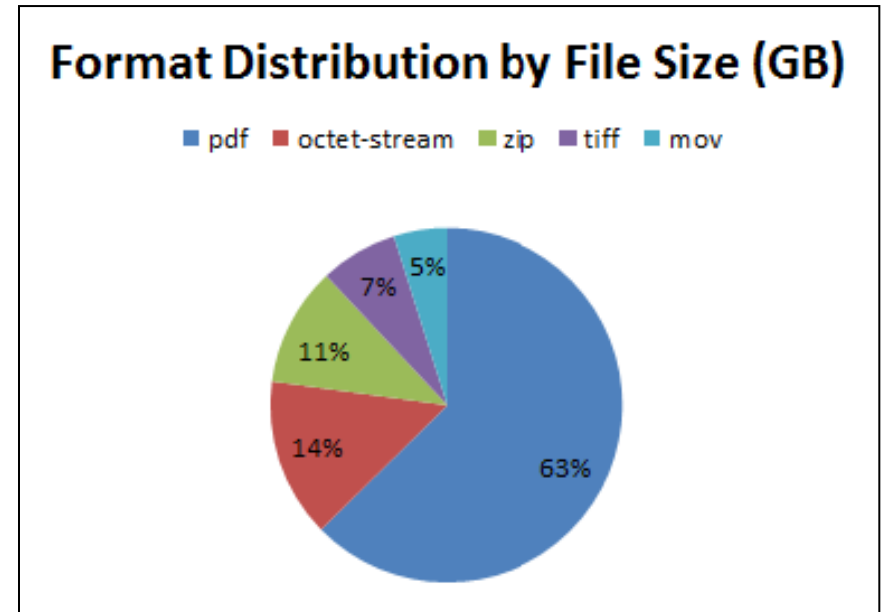
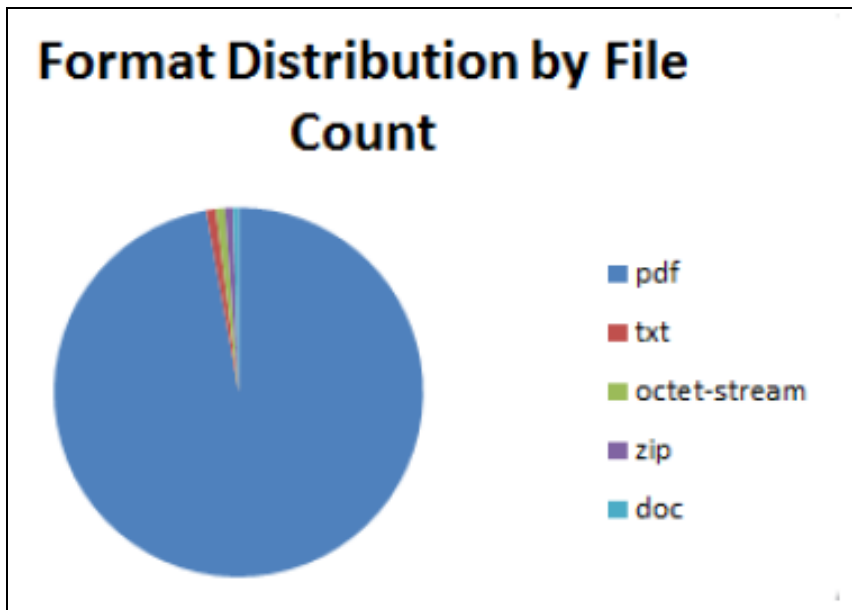
- 6, 095,978 unique files
- ~130 Tb of content





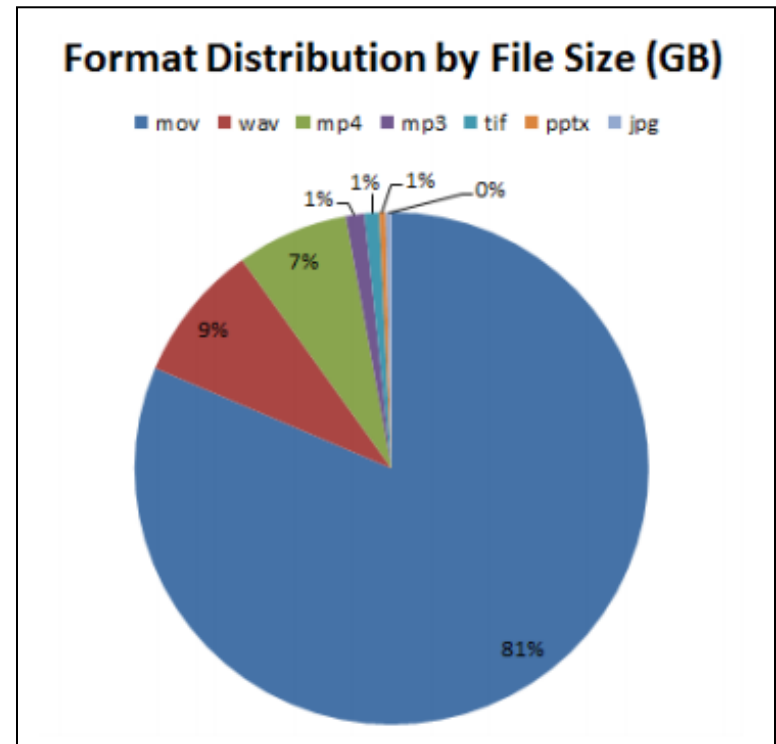
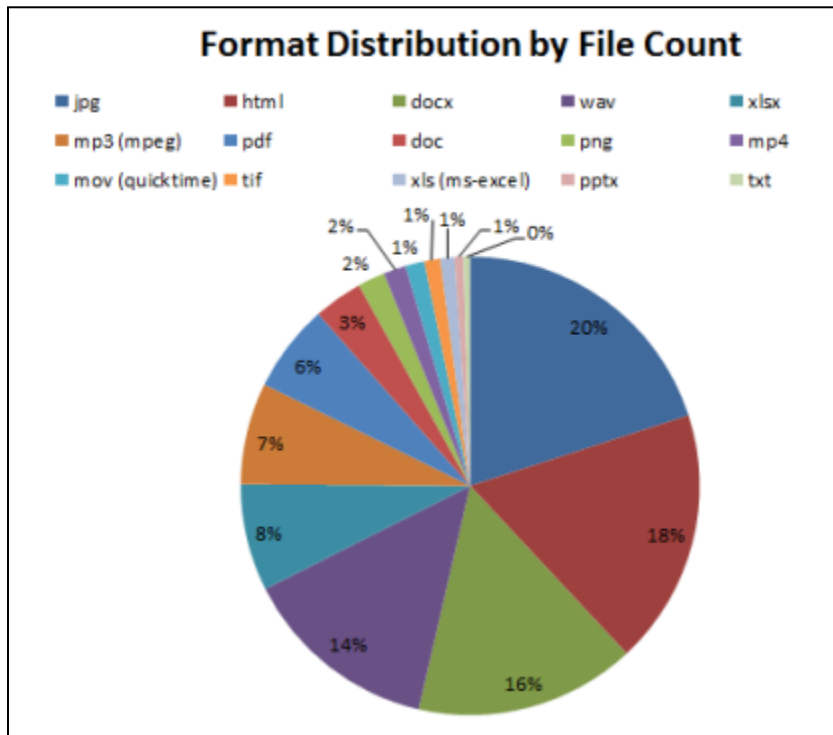
# Sample Report 1

- 49,506 unique files
- 320 GB of content



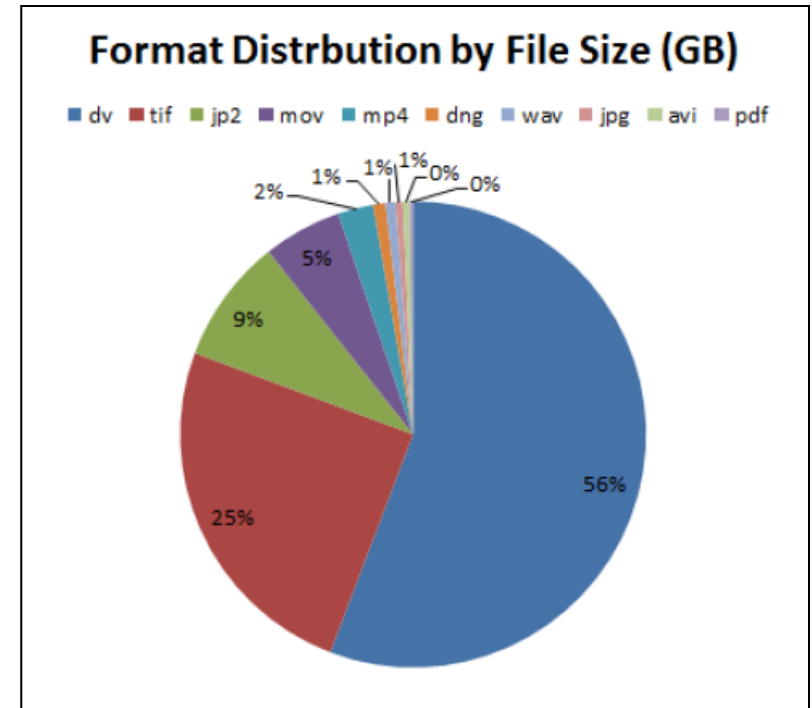
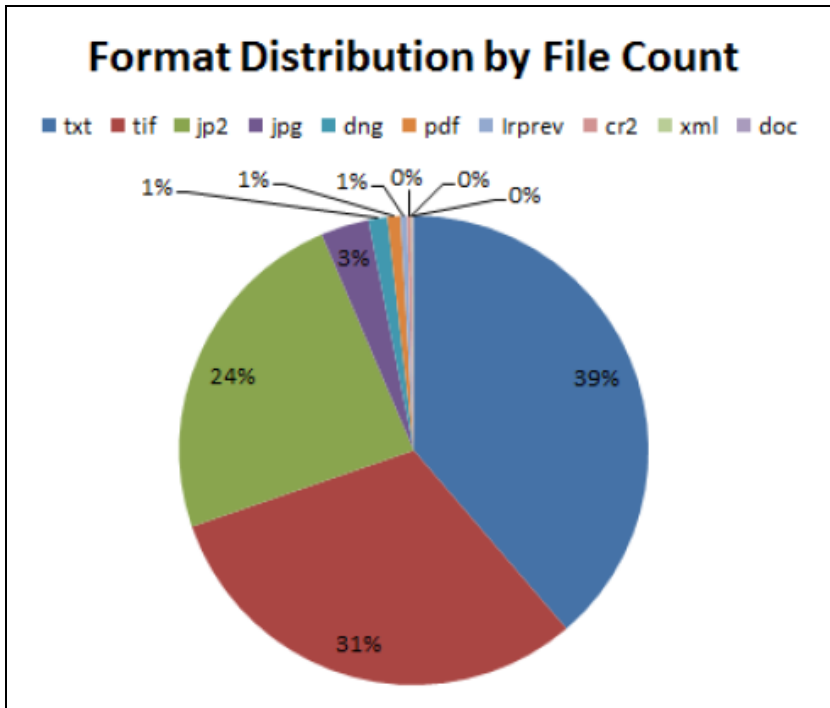
# Sample Report 2

- 32, 039 unique files
- 1769 GB of content



# Sample Report 3

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

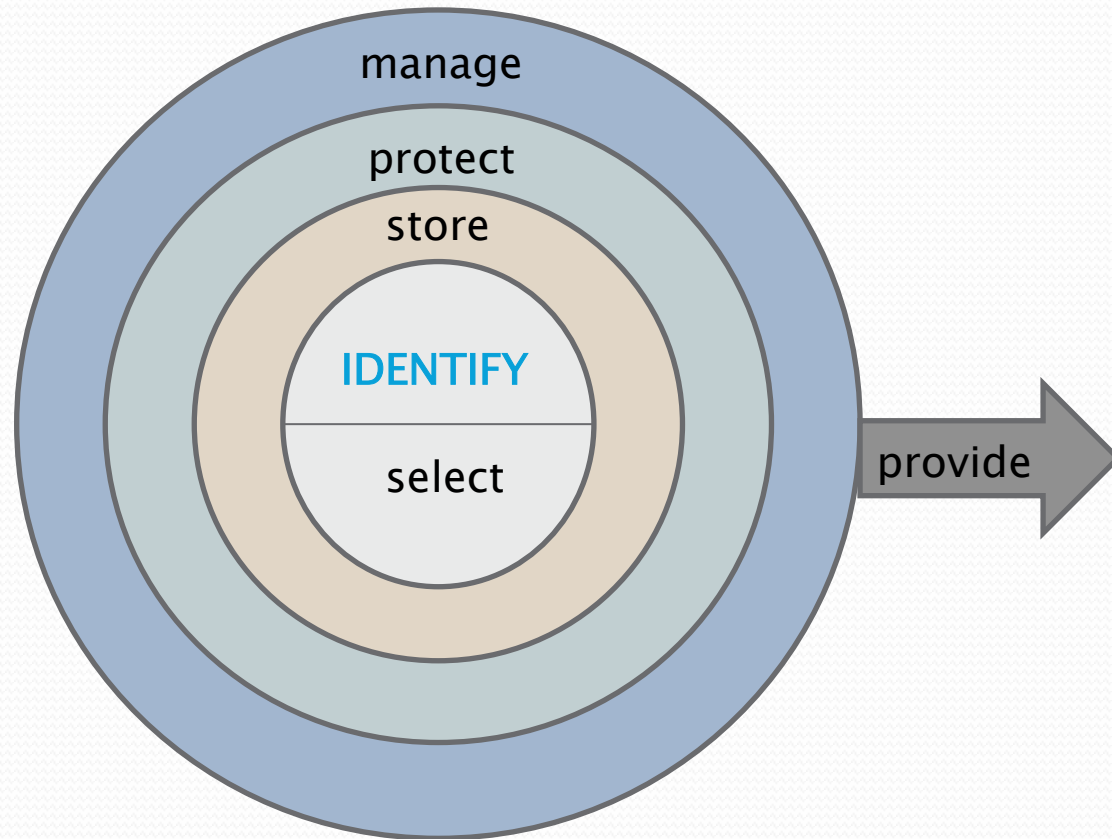


# Summary

- 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

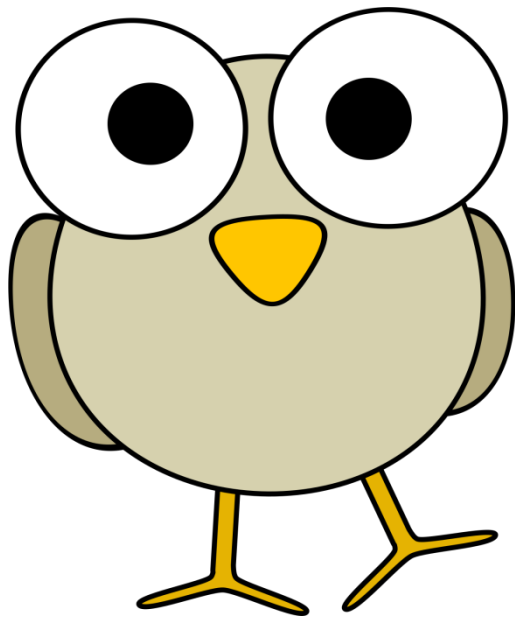
- Managing Digital Content Over Time: An Introduction to Digital Preservation
- Three part webinar series
- April/May



# Resource Links

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

# Questions?



**Contact Me**

Carol Kussmann  
[kussmann@umn.edu](mailto:kussmann@umn.edu)