Developing a Born-Digital Preservation Workflow

Bill Donovan & Jack Kearney

Digital Imaging & Curation Manager
Thomas P. O'Neill, Jr. Library

Audiovisual Archives Assistant
John J. Burns Library

Boston College Libraries

July 23, 2014

Goals

1. Develop a systematic approach to digital preservation (DP) of born-digital collections.
2. Gain experience with a variety of DP hardware and software and figure out practical protocols --- how does all of this stuff work?
3. Use a real-life example, the electronic records of the Mary O'Hara papers. (MOH)

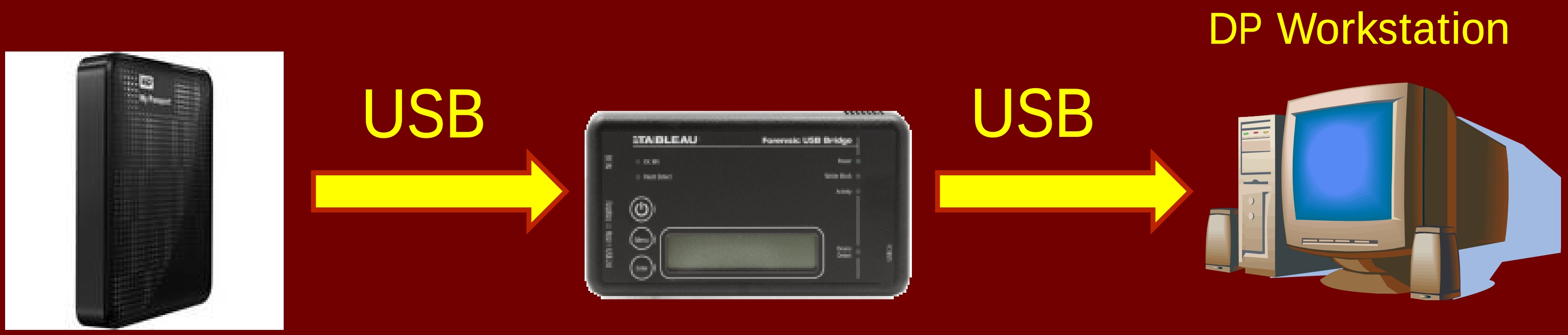
Mary O’Hara electronic records

An Irish soprano and harpist of international renown, Mary O'Hara has appeared on many of the world's major stages, including Royal Albert Hall, New York's Carnegie Hall, Sydney Opera House, and Toronto's Roy Thompson Hall.

Hard drive donated to the Burns Library Irish Music Center at Boston College.

Protecting the original

Used a forensic “write-blocker.”
Permits reading but not writing.
Prevents changes to the files on the hard drive.

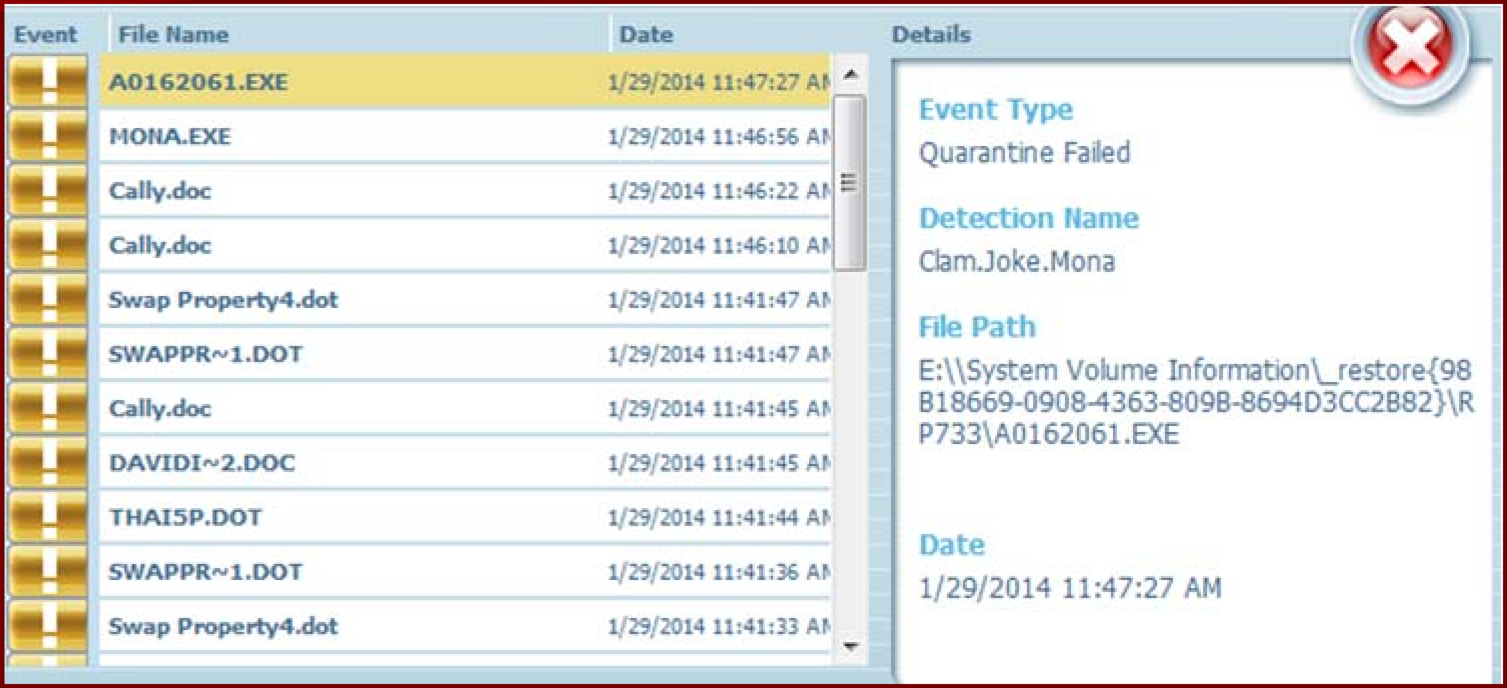


USB

USB

DP Workstation

But is the original virus-free?



Event	File Name	Date
!	A0162061.EXE	1/29/2014 11:47:27 AM
!	MONA.EXE	1/29/2014 11:46:56 AM
!	Cally.doc	1/29/2014 11:46:22 AM
!	Cally.doc	1/29/2014 11:46:10 AM
!	Swap Property4.dot	1/29/2014 11:41:47 AM
!	SWAPPR~1.DOT	1/29/2014 11:41:47 AM
!	Cally.doc	1/29/2014 11:41:45 AM
!	DAVIDI~2.DOC	1/29/2014 11:41:45 AM
!	THAI5P.DOT	1/29/2014 11:41:44 AM
!	SWAPPR~1.DOT	1/29/2014 11:41:36 AM
!	Swap Property4.dot	1/29/2014 11:41:33 AM

Details

Event Type
Quarantine Failed

Detection Name
Clam.Joke.Mona

File Path
E:\\System Volume Information_restore{98B18669-0908-4363-809B-8694D3CC2B82}\RP733\A0162061.EXE

Date
1/29/2014 11:47:27 AM

Decision: Delete 33 infected files from the hard drive (after first creating a comprehensive inventory of all files on the drive)

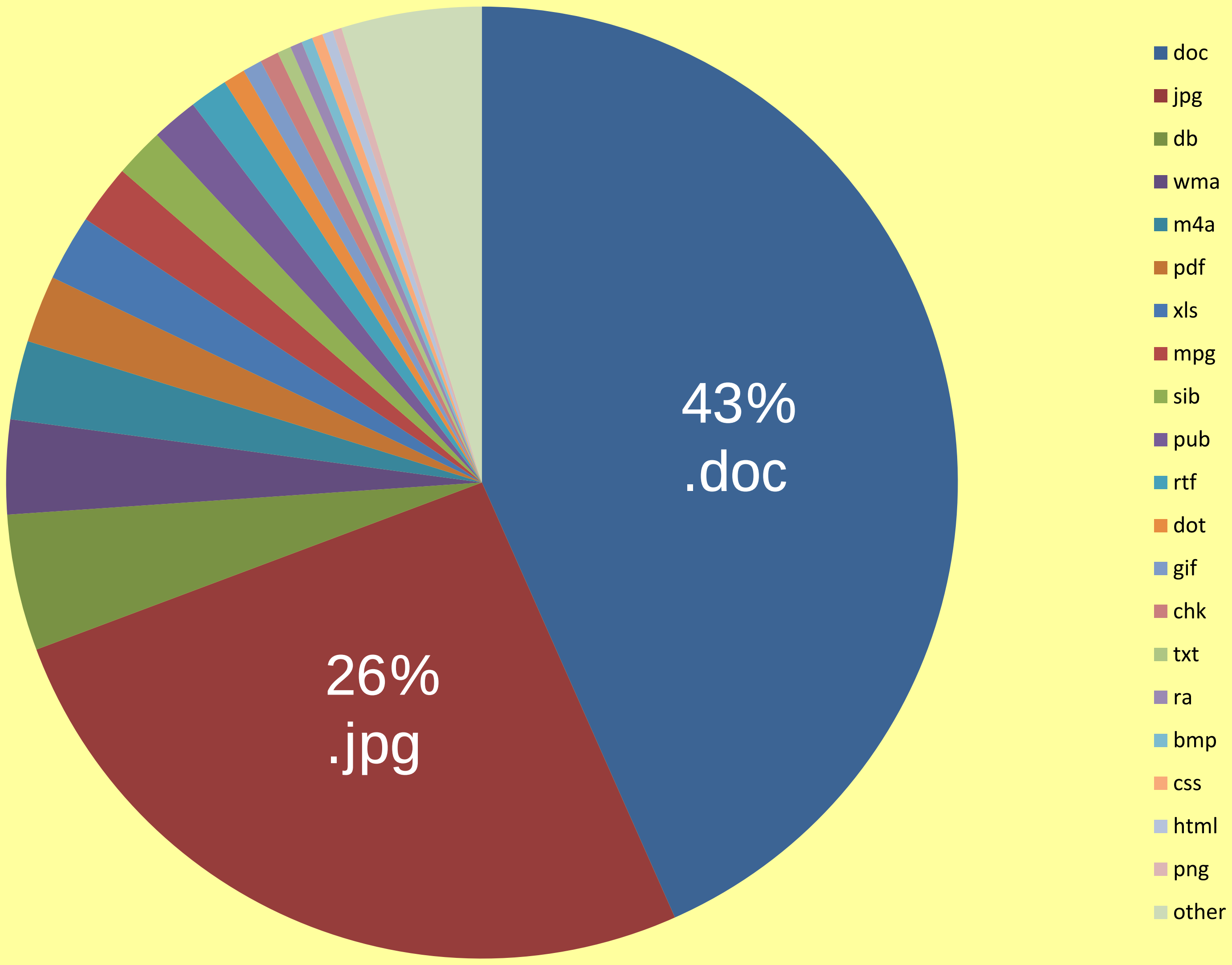
What’s on the MOH hard drive?

Inventory taken with this Unix command:

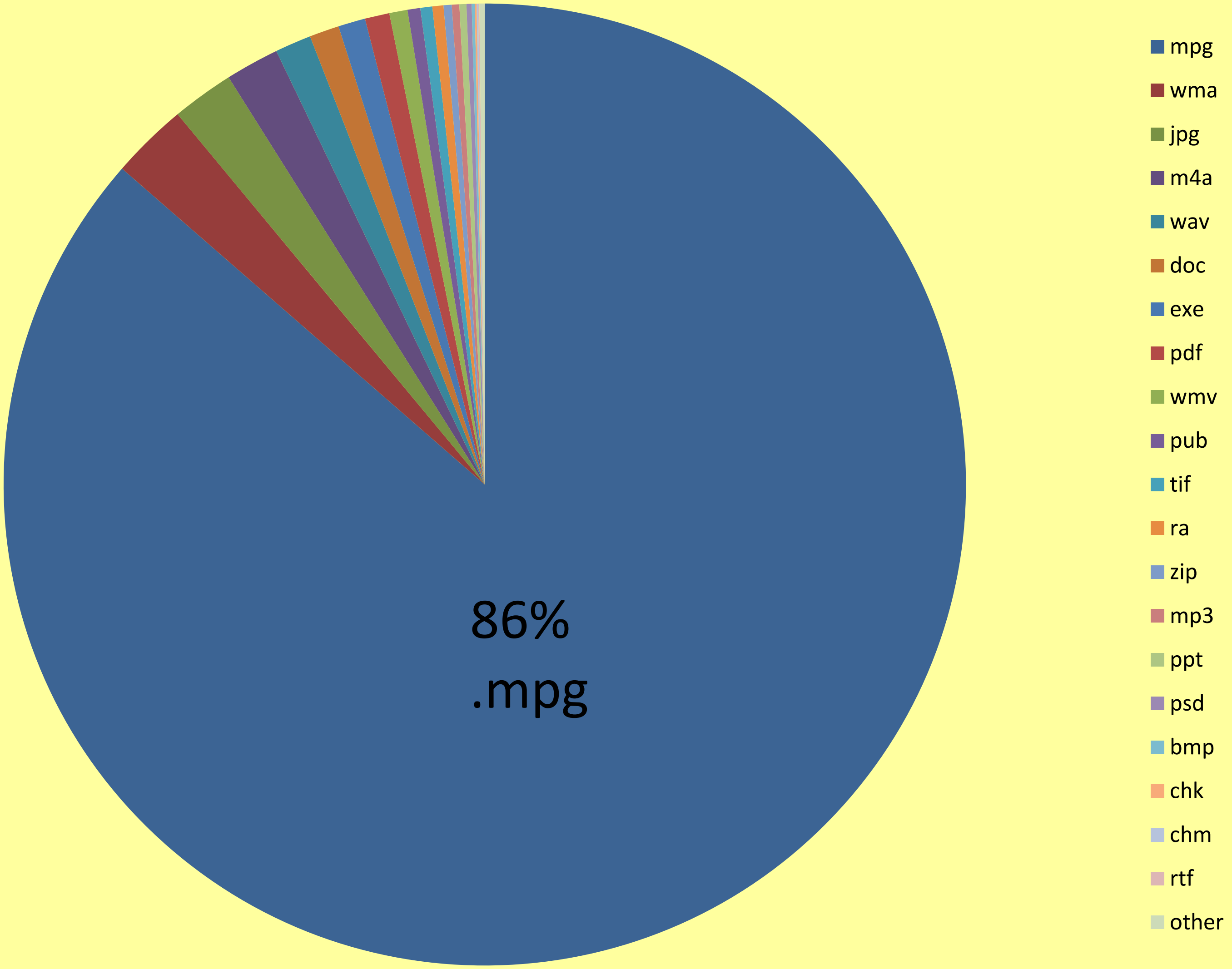
```
find directory-name  -type f -exec ls -l {} ; >c:\data\MOH\inventory.txt
```

21,988 files that are on the hard drive, for a total of 104.3 GB

20 most frequent file types, plus other



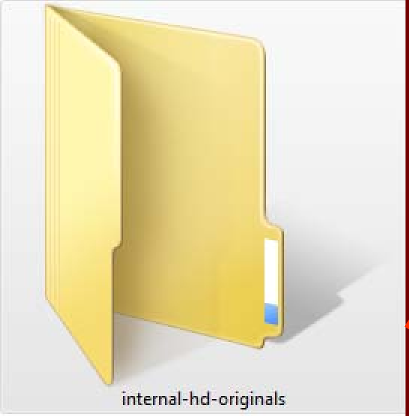
20 file types with most number of bytes, plus other



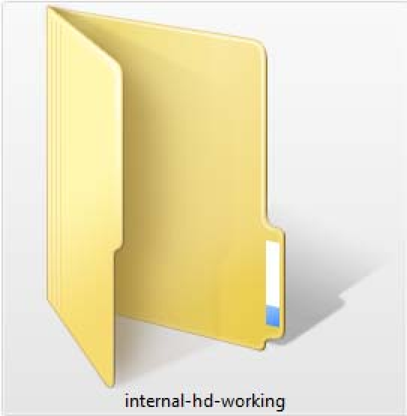
Back-ups of files

MOH hard drive
with original files

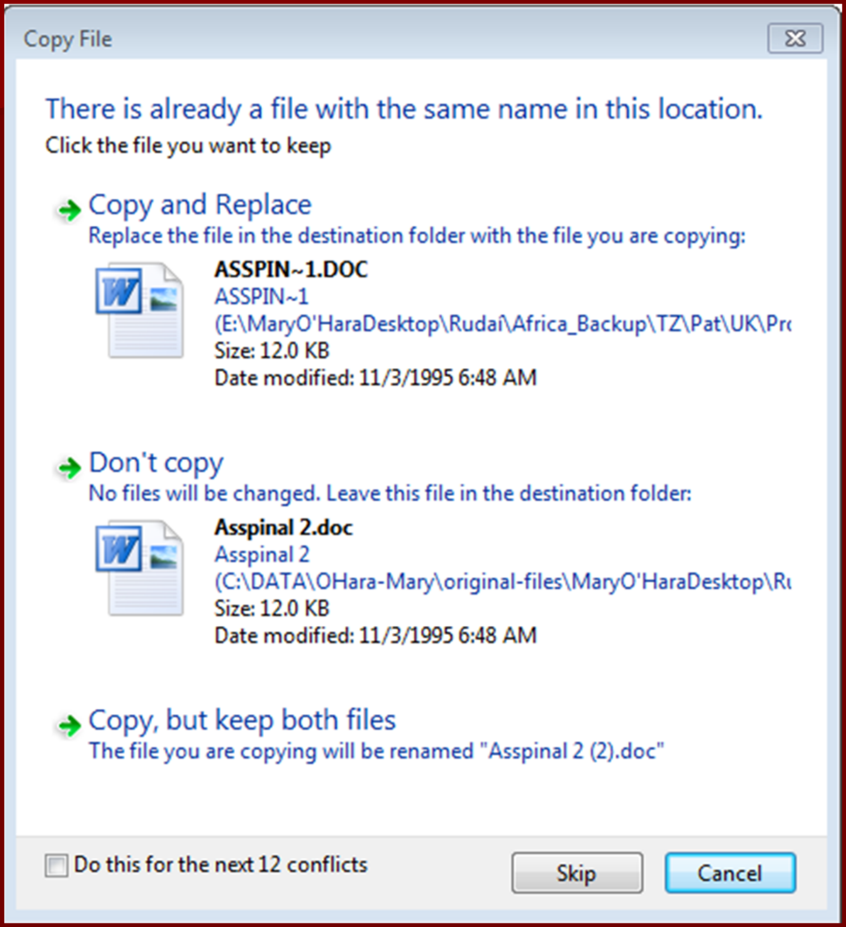
DP Workstation’s
internal hard drive:
“Original files”



DP Workstation’s
internal hard drive:
“Working files”



Duplicate files



There is already a file with the same name in this location.

Click the file you want to keep

Copy and Replace

Replace the file in the destination folder with the file you are copying:

ASSPIN~1.DOC
ASSPIN~1
(E:\MaryO'HaraDesktop\Rudai\Africa_Backup\TZ\Pat\UK\Pr
Size: 12.0 KB
Date modified: 11/3/1995 6:48 AM

Don't copy

No files will be changed. Leave this file in the destination folder:

Asspinal 2.doc
Asspinal 2
(C:\DATA\OHara-Mary\original-files\MaryO'HaraDesktop\Ru
Size: 12.0 KB
Date modified: 11/3/1995 6:48 AM

Copy, but keep both files

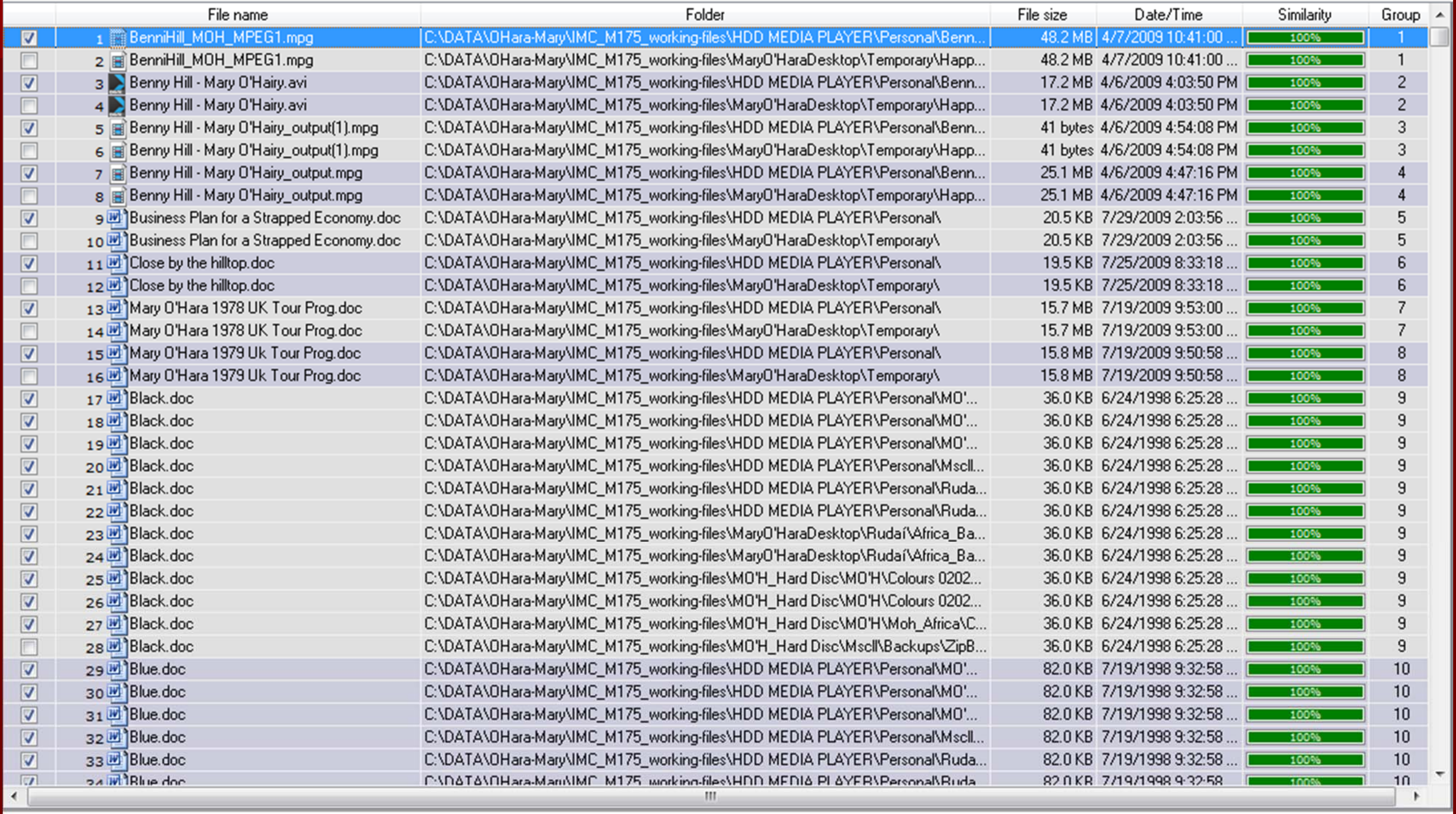
The file you are copying will be renamed "Asspinal 2 (2).doc"

□ Do this for the next 12 conflicts

"8.3" constraint

Decision: keep only the copy with the more complete filename

De-duping software



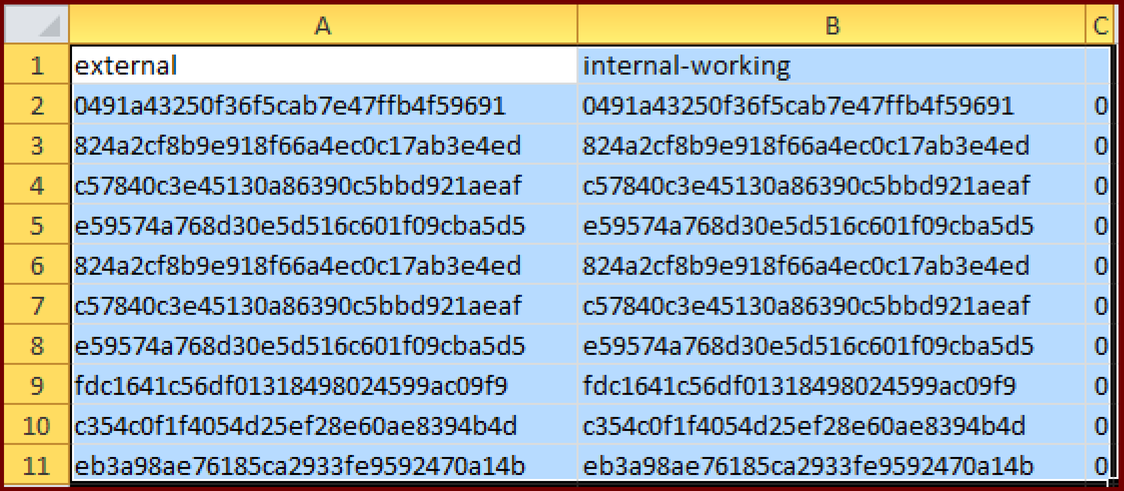
		File name	Folder	File size	Date/Time	Similarity	Group
☑	1	BenniHill_MOH_MPEG1.mpg	C:\DATA\OHara-Mary\IMC_M175_working-files\HDD MEDIA PLAYER\Personal\Benn...	48.2 MB	4/7/2009 10:41:00 ...	100%	1
□	2	BenniHill_MOH_MPEG1.mpg	C:\DATA\OHara-Mary\IMC_M175_working-files\MaryO'HaraDesktop\Temporary\Happ...	48.2 MB	4/7/2009 10:41:00 ...	100%	1
☑	3	Benny Hill - Mary O'Hairy.avi	C:\DATA\OHara-Mary\IMC_M175_working-files\HDD MEDIA PLAYER\Personal\Benn...	17.2 MB	4/6/2009 4:03:50 PM	100%	2
□	4	Benny Hill - Mary O'Hairy.avi	C:\DATA\OHara-Mary\IMC_M175_working-files\MaryO'HaraDesktop\Temporary\Happ...	17.2 MB	4/6/2009 4:03:50 PM	100%	2
☑	5	Benny Hill - Mary O'Hairy_output(1).mpg	C:\DATA\OHara-Mary\IMC_M175_working-files\HDD MEDIA PLAYER\Personal\Benn...	41 bytes	4/6/2009 4:54:08 PM	100%	3
□	6	Benny Hill - Mary O'Hairy_output(1).mpg	C:\DATA\OHara-Mary\IMC_M175_working-files\MaryO'HaraDesktop\Temporary\Happ...	41 bytes	4/6/2009 4:54:08 PM	100%	3
☑	7	Benny Hill - Mary O'Hairy_output.mpg	C:\DATA\OHara-Mary\IMC_M175_working-files\HDD MEDIA PLAYER\Personal\Benn...	25.1 MB	4/6/2009 4:47:16 PM	100%	4
□	8	Benny Hill - Mary O'Hairy_output.mpg	C:\DATA\OHara-Mary\IMC_M175_working-files\MaryO'HaraDesktop\Temporary\Happ...	25.1 MB	4/6/2009 4:47:16 PM	100%	4
☑	9	Business Plan for a Strapped Economy.doc	C:\DATA\OHara-Mary\IMC_M175_working-files\HDD MEDIA PLAYER\Personal\	20.5 KB	7/29/2009 2:03:56 ...	100%	5
□	10	Business Plan for a Strapped Economy.doc	C:\DATA\OHara-Mary\IMC_M175_working-files\MaryO'HaraDesktop\Temporary\	20.5 KB	7/29/2009 2:03:56 ...	100%	5
☑	11	Close by the hilltop.doc	C:\DATA\OHara-Mary\IMC_M175_working-files\HDD MEDIA PLAYER\Personal\	19.5 KB	7/25/2009 8:33:18 ...	100%	6
□	12	Close by the hilltop.doc	C:\DATA\OHara-Mary\IMC_M175_working-files\MaryO'HaraDesktop\Temporary\	19.5 KB	7/25/2009 8:33:18 ...	100%	6
☑	13	Mary O'Hara 1978 UK Tour Prog.doc	C:\DATA\OHara-Mary\IMC_M175_working-files\HDD MEDIA PLAYER\Personal\	15.7 MB	7/19/2009 9:53:00 ...	100%	7
□	14	Mary O'Hara 1978 UK Tour Prog.doc	C:\DATA\OHara-Mary\IMC_M175_working-files\MaryO'HaraDesktop\Temporary\	15.7 MB	7/19/2009 9:53:00 ...	100%	7
☑	15	Mary O'Hara 1979 Uk Tour Prog.doc	C:\DATA\OHara-Mary\IMC_M175_working-files\HDD MEDIA PLAYER\Personal\	15.8 MB	7/19/2009 9:50:58 ...	100%	8
□	16	Mary O'Hara 1979 Uk Tour Prog.doc	C:\DATA\OHara-Mary\IMC_M175_working-files\MaryO'HaraDesktop\Temporary\	15.8 MB	7/19/2009 9:50:58 ...	100%	8
☑	17	Black.doc	C:\DATA\OHara-Mary\IMC_M175_working-files\HDD MEDIA PLAYER\Personal\MO'...	36.0 KB	6/24/1998 6:25:28 ...	100%	9
☑	18	Black.doc	C:\DATA\OHara-Mary\IMC_M175_working-files\HDD MEDIA PLAYER\Personal\MO'...	36.0 KB	6/24/1998 6:25:28 ...	100%	9
☑	19	Black.doc	C:\DATA\OHara-Mary\IMC_M175_working-files\HDD MEDIA PLAYER\Personal\MO'...	36.0 KB	6/24/1998 6:25:28 ...	100%	9
☑	20	Black.doc	C:\DATA\OHara-Mary\IMC_M175_working-files\HDD MEDIA PLAYER\Personal\Mscll...	36.0 KB	6/24/1998 6:25:28 ...	100%	9
☑	21	Black.doc	C:\DATA\OHara-Mary\IMC_M175_working-files\HDD MEDIA PLAYER\Personal\Ruda...	36.0 KB	6/24/1998 6:25:28 ...	100%	9
☑	22	Black.doc	C:\DATA\OHara-Mary\IMC_M175_working-files\HDD MEDIA PLAYER\Personal\Ruda...	36.0 KB	6/24/1998 6:25:28 ...	100%	9
☑	23	Black.doc	C:\DATA\OHara-Mary\IMC_M175_working-files\MaryO'HaraDesktop\Rudaí\Africa_Ba...	36.0 KB	6/24/1998 6:25:28 ...	100%	9
☑	24	Black.doc	C:\DATA\OHara-Mary\IMC_M175_working-files\MaryO'HaraDesktop\Rudaí\Africa_Ba...	36.0 KB	6/24/1998 6:25:28 ...	100%	9
☑	25	Black.doc	C:\DATA\OHara-Mary\IMC_M175_working-files\MO'H_Hard Disc\MO'H\Colours 0202...	36.0 KB	6/24/1998 6:25:28 ...	100%	9
☑	26	Black.doc	C:\DATA\OHara-Mary\IMC_M175_working-files\MO'H_Hard Disc\MO'H\Colours 0202...	36.0 KB	6/24/1998 6:25:28 ...	100%	9
☑	27	Black.doc	C:\DATA\OHara-Mary\IMC_M175_working-files\MO'H_Hard Disc\MO'H\Moh_Africa\C...	36.0 KB	6/24/1998 6:25:28 ...	100%	9
□	28	Black.doc	C:\DATA\OHara-Mary\IMC_M175_working-files\MO'H_Hard Disc\Mscll\Backups\ZipB...	36.0 KB	6/24/1998 6:25:28 ...	100%	9
☑	29	Blue.doc	C:\DATA\OHara-Mary\IMC_M175_working-files\HDD MEDIA PLAYER\Personal\MO'...	82.0 KB	7/19/1998 9:32:58 ...	100%	10
☑	30	Blue.doc	C:\DATA\OHara-Mary\IMC_M175_working-files\HDD MEDIA PLAYER\Personal\MO'...	82.0 KB	7/19/1998 9:32:58 ...	100%	10
☑	31	Blue.doc	C:\DATA\OHara-Mary\IMC_M175_working-files\HDD MEDIA PLAYER\Personal\MO'...	82.0 KB	7/19/1998 9:32:58 ...	100%	10
☑	32	Blue.doc	C:\DATA\OHara-Mary\IMC_M175_working-files\HDD MEDIA PLAYER\Personal\Mscll...	82.0 KB	7/19/1998 9:32:58 ...	100%	10
☑	33	Blue.doc	C:\DATA\OHara-Mary\IMC_M175_working-files\HDD MEDIA PLAYER\Personal\Ruda...	82.0 KB	7/19/1998 9:32:58 ...	100%	10
☑	34	Blue.doc	C:\DATA\OHara-Mary\IMC_M175_working-files\HDD MEDIA PLAYER\Personal\Ruda	82.0 KB	7/19/1998 9:32:58	100%	10

Initial checksums

Decide which checksum algorithm to use

Compute initial checksums for each set of files

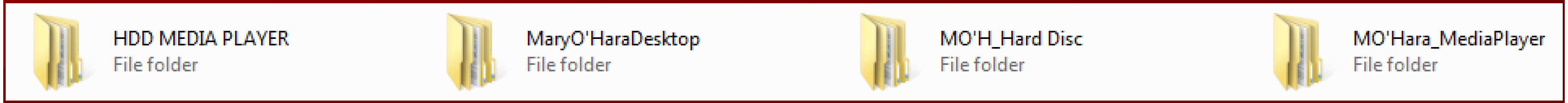
Compare checksums to verify original files and copies are identical



	A	B	C
1	external	internal-working	
2	0491a43250f36f5cab7e47ffb4f59691	0491a43250f36f5cab7e47ffb4f59691	0
3	824a2cf8b9e918f66a4ec0c17ab3e4ed	824a2cf8b9e918f66a4ec0c17ab3e4ed	0
4	c57840c3e45130a86390c5bbd921aeaf	c57840c3e45130a86390c5bbd921aeaf	0
5	e59574a768d30e5d516c601f09cba5d5	e59574a768d30e5d516c601f09cba5d5	0
6	824a2cf8b9e918f66a4ec0c17ab3e4ed	824a2cf8b9e918f66a4ec0c17ab3e4ed	0
7	c57840c3e45130a86390c5bbd921aeaf	c57840c3e45130a86390c5bbd921aeaf	0
8	e59574a768d30e5d516c601f09cba5d5	e59574a768d30e5d516c601f09cba5d5	0
9	fdc1641c56df01318498024599ac09f9	fdc1641c56df01318498024599ac09f9	0
10	c354c0f1f4054d25ef28e60ae8394b4d	c354c0f1f4054d25ef28e60ae8394b4d	0
11	eb3a98ae76185ca2933fe9592470a14b	eb3a98ae76185ca2933fe9592470a14b	0

For final-state collections, use automated tool for integrity-checking

Odd file/folder names

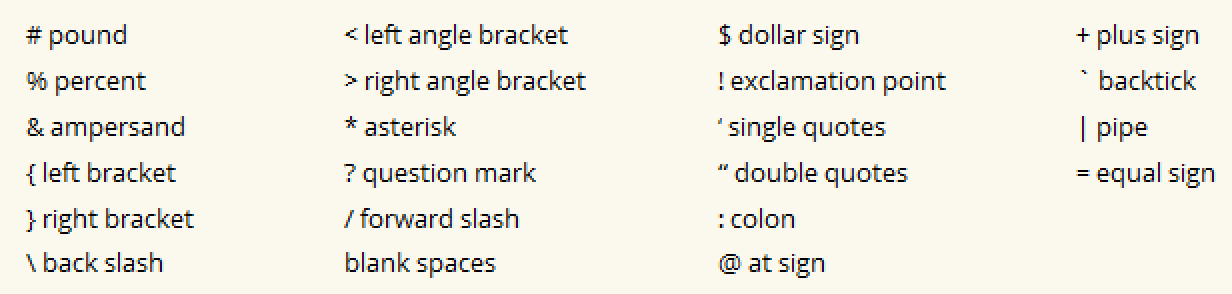


HDD MEDIA PLAYER
File folder

MaryO'HaraDesktop
File folder

MO'H_Hard Disc
File folder

MO'Hara_MediaPlayer
File folder



# pound	< left angle bracket	$ dollar sign	+ plus sign
% percent	> right angle bracket	! exclamation point	` backtick
& ampersand	* asterisk	' single quotes	| pipe
{ left bracket	? question mark	“ double quotes	= equal sign
} right bracket	/ forward slash	: colon	
\ back slash	blank spaces	@ at sign	

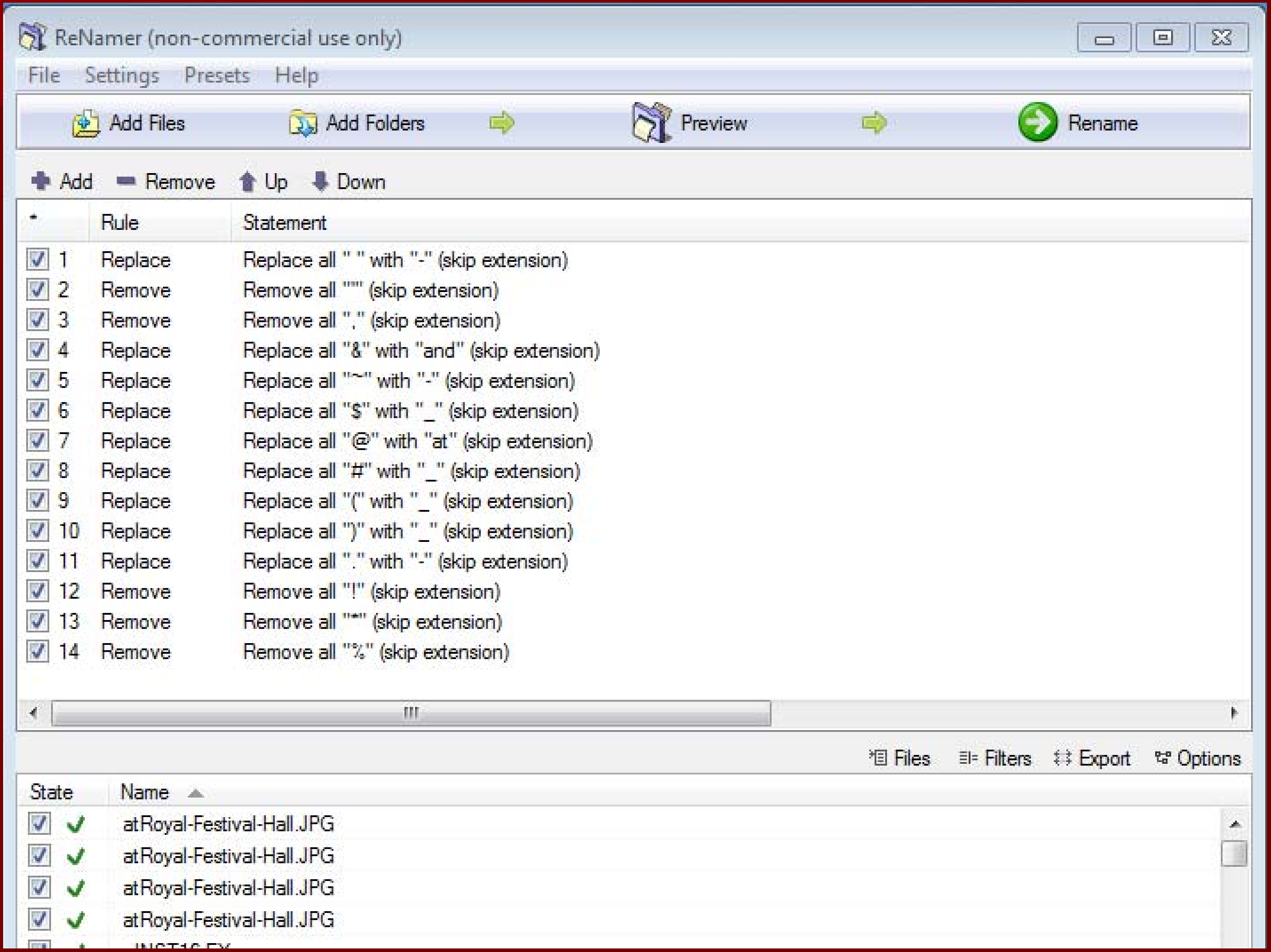
Some
“forbidden
characters”

http://www.mtu.edu/umc/services/web/cms/characters-avoid/

Had to “escape” these characters in folder name in our Unix commands.

Decision: remediate folder and file names, <u>but only for the working copies</u>.

Renamer: remediating folder/file names



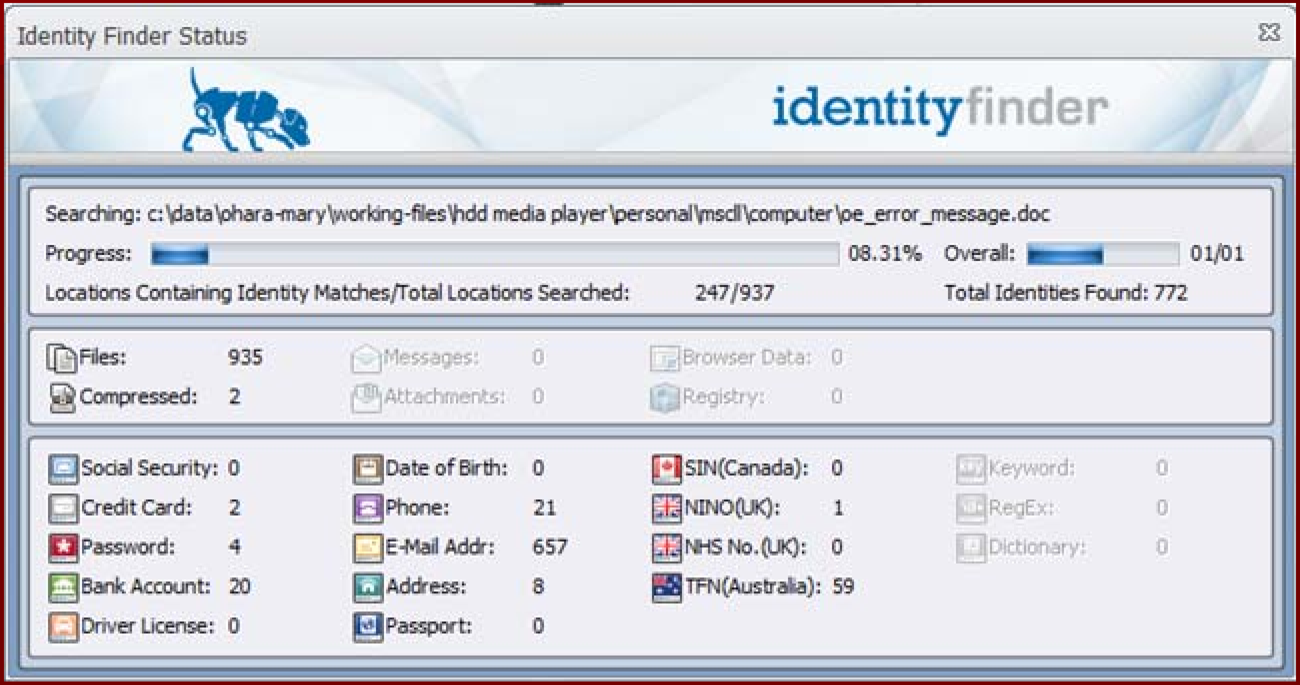
	#	Rule	Statement
☑	1	Replace	Replace all " " with "-" (skip extension)
☑	2	Remove	Remove all "'" (skip extension)
☑	3	Remove	Remove all "," (skip extension)
☑	4	Replace	Replace all "&" with "and" (skip extension)
☑	5	Replace	Replace all "~" with "-" (skip extension)
☑	6	Replace	Replace all "$" with "_" (skip extension)
☑	7	Replace	Replace all "@" with "at" (skip extension)
☑	8	Replace	Replace all "#" with "_" (skip extension)
☑	9	Replace	Replace all "(" with "_" (skip extension)
☑	10	Replace	Replace all ")" with "_" (skip extension)
☑	11	Replace	Replace all "." with "-" (skip extension)
☑	12	Remove	Remove all "!" (skip extension)
☑	13	Remove	Remove all "*" (skip extension)
☑	14	Remove	Remove all "%" (skip extension)

State		Name
☑	✓	atRoyal-Festival-Hall.JPG
☑	✓	atRoyal-Festival-Hall.JPG
☑	✓	atRoyal-Festival-Hall.JPG
☑	✓	atRoyal-Festival-Hall.JPG

Any files off-limits or expendable?

- Confidential information
  - ➢ Social security numbers
  - ➢ Financial information
- Files that have nothing to do with MOH per se
  - System Files
- Files that have no archival value
  - Thumbs.db

Personally Identifiable Information (PII)



Identity Finder Status

identityfinder

Searching: c:\data\phara-mary\working-files\hdd media player\personal\mscl\computer\oe_error_message.doc

Progress: 08.31% Overall: 01/01

Locations Containing Identity Matches/Total Locations Searched: 247/937 Total Identities Found: 772

Files:	935	Messages:	0	Browser Data:	0
Compressed:	2	Attachments:	0	Registry:	0

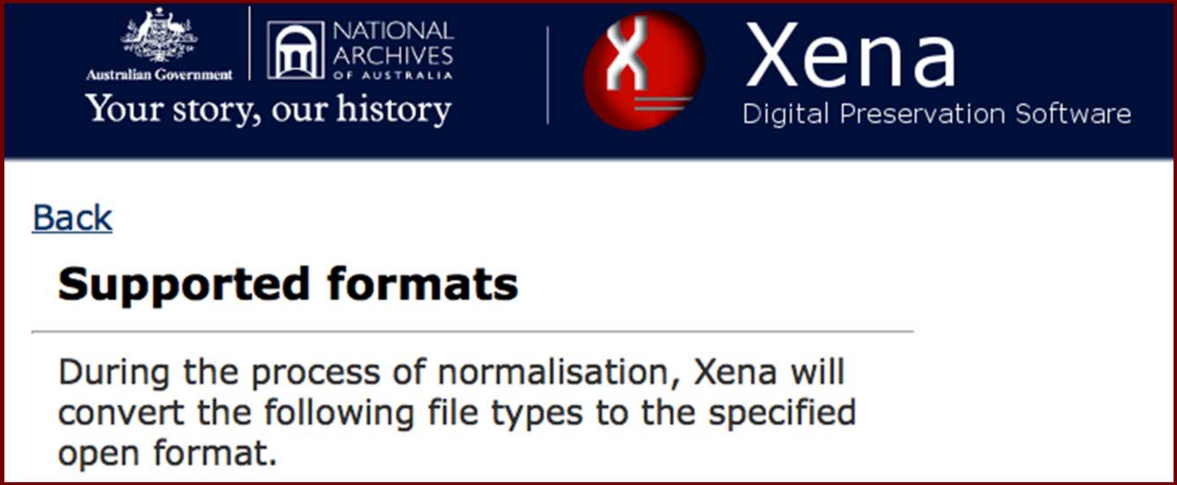
Social Security:	0	Date of Birth:	0	SIN(Canada):	0	Keyword:	0
Credit Card:	2	Phone:	21	NINO(UK):	1	RegEx:	0
Password:	4	E-Mail Addr:	657	NHS No.(UK):	0	Dictionary:	0
Bank Account:	20	Address:	8	TFN(Australia):	59		
Driver License:	0	Passport:	0				

Policy decisions to be made by archivists:

- Based upon the PII findings, which files will eventually made open to the public, or not?

Any proprietary file formats?

Normalize using Xena ? (see Gregory 2010)



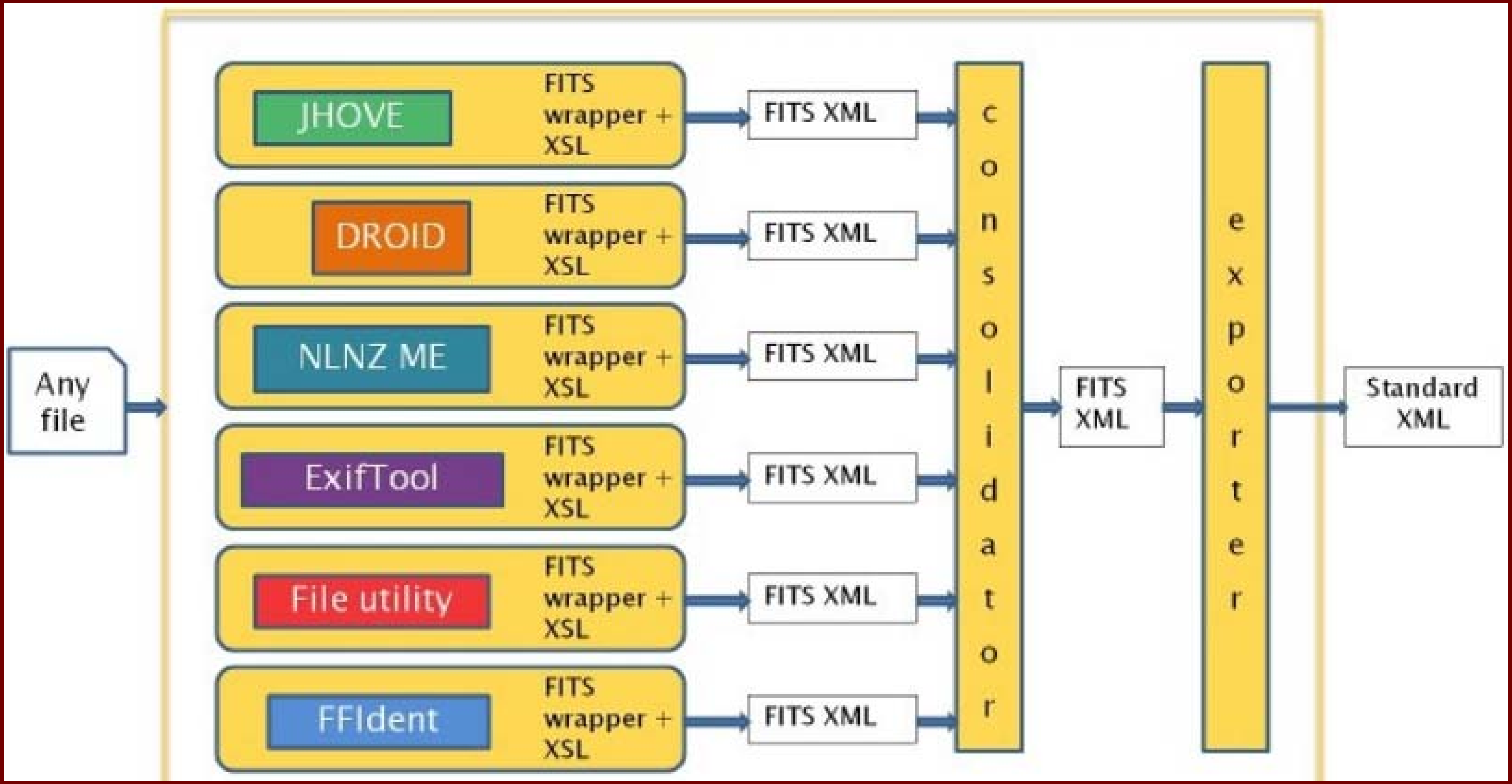
Supported formats

During the process of normalisation, Xena will convert the following file types to the specified open format.

For RealAudio files, normalize with dBpoweramp?

Policy decision: Which files to normalize, and which formats to preserve?

File formats --- identify, validate, and extract metadata --- using FITS



http://projects.iq.harvard.edu/fits/fits-processing

Output of FITS --- identity of the file and the current tool versions in FITS

```
...

  <identity format="Tagged Image File Format" mimetype="image/tiff" toolname="FITS"
toolversion="0.6.2">
   <tool toolname="Jhove" toolversion="1.5" />
   <tool toolname="file utility" toolversion="5.03" />
   <tool toolname="Exiftool" toolversion="9.06" />
   <tool toolname="Droid" toolversion="3.0" />
   <tool toolname="NLNZ Metadata Extractor" toolversion="3.4GA" />
   <tool toolname="ffident" toolversion="0.2" />

...
```

Output of FITS --- file information

```
...
    <size toolname="Jhove" toolversion="1.5">1795770</size>
    <creatingApplicationName toolname="Jhove" toolversion="1.5">Omniscan 11.12 SR2
Build13</creatingApplicationName>
    <lastmodified toolname="Exiftool" toolversion="9.06"
status="SINGLE_RESULT">2013:08:14 14:15:38-04:00</lastmodified>
    <filepath toolname="OIS File Information" toolversion="0.1"
status="SINGLE_RESULT">C:\DATA\FITS_test_folder\fits_test_imagefile.tif</filepath>
    <filename toolname="OIS File Information" toolversion="0.1"
status="SINGLE_RESULT">C:\DATA\FITS_test_folder\fits_test_imagefile.tif</filename>
    <md5checksum toolname="OIS File Information" toolversion="0.1"
status="SINGLE_RESULT">ccfca47fb4f2597c04e299c99f4043ce</md5checksum>
    <fslastmodified toolname="OIS File Information" toolversion="0.1"
status="SINGLE_RESULT">1376504138000</fslastmodified>
...
```

Output of FITS --- file status

```
<filestatus>
    <well-formed toolname="Jhove" toolversion="1.5"
status="SINGLE_RESULT">true</well-formed>

    <valid toolname="Jhove" toolversion="1.5" status="SINGLE_RESULT">true</valid>
  </filestatus>
```

From JHOVE website (http://jhove.sourceforge.net/):

A digital object is well-formed if it meets the purely syntactic requirements for its format.

An object is valid if it is well-formed and it meets additional semantic-level requirements.

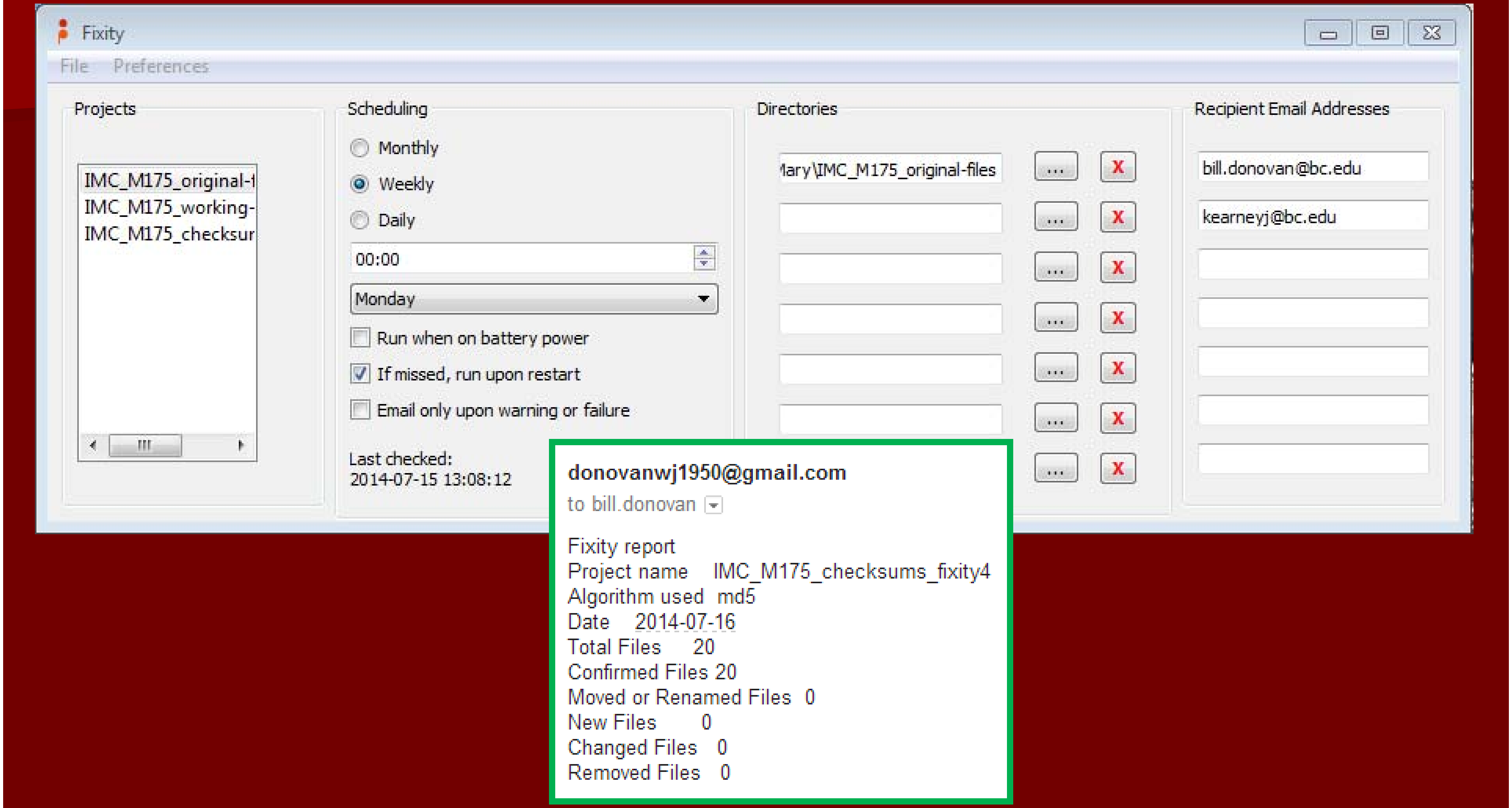
Output of FITS --- metadata

```
...
    <compressionScheme toolname="Jhove" toolversion="1.5">Uncompressed</compressionScheme>
    <imageWidth toolname="Jhove" toolversion="1.5">1598</imageWidth>
    <imageHeight toolname="Jhove" toolversion="1.5">373</imageHeight>
    <colorSpace toolname="Jhove" toolversion="1.5">RGB</colorSpace>
    <referenceBlackWhite toolname="Jhove" toolversion="1.5" status="SINGLE_RESULT">0.0 255.0 0.0 255.0 0.0
255.0</referenceBlackWhite>
    <iccProfileName toolname="Exiftool" toolversion="9.06" status="SINGLE_RESULT">sRGB IEC61966-2.1</iccProfileName>
    <orientation toolname="Jhove" toolversion="1.5">normal*</orientation>
    <samplingFrequencyUnit toolname="Jhove" toolversion="1.5" status="CONFLICT">in.</samplingFrequencyUnit>
    <samplingFrequencyUnit toolname="Exiftool" toolversion="9.06" status="CONFLICT">inches</samplingFrequencyUnit>
    <xSamplingFrequency toolname="Jhove" toolversion="1.5">300</xSamplingFrequency>
    <ySamplingFrequency toolname="Jhove" toolversion="1.5">300</ySamplingFrequency>
    <bitsPerSample toolname="Jhove" toolversion="1.5">8 8 8</bitsPerSample>
    <samplesPerPixel toolname="Jhove" toolversion="1.5">3</samplesPerPixel>
    <imageProducer toolname="Jhove" toolversion="1.5" status="SINGLE_RESULT">Zeutschel Omniscan 11</imageProducer>
    <scanningSoftwareName toolname="Jhove" toolversion="1.5">Omniscan 11.12 SR2 Build13</scanningSoftwareName>
    ...
```

Workflow: checklist

- ☑ Take inventory
- ☑ Check for viruses
- ☑ Make back-up copies
- ☑ Leave originals as-is
- ☑ Fingerprint each file (checksum)
- ☑ Protect privacy (personally identifiable information)
- ☑ Fix unorthodox folder/file names
- ☑ Avoid proprietary file formats
- ☑ Validate file formats
- ☑ Weed out extraneous/duplicate files
- ☑ Integrate into collection
- ☑ Provide access to users

Fixity: automated integrity-checking for final state collections



Projects

IMC_M175_original-f
IMC_M175_working-
IMC_M175_checksur

Scheduling

○ Monthly
● Weekly
○ Daily

00:00

Monday

□ Run when on battery power
☑ If missed, run upon restart
□ Email only upon warning or failure

Last checked:
2014-07-15 13:08:12

Directories

lary\IMC_M175_original-files

Recipient Email Addresses

bill.donovan@bc.edu
kearneyj@bc.edu

donovanwj1950@gmail.com
to bill.donovan

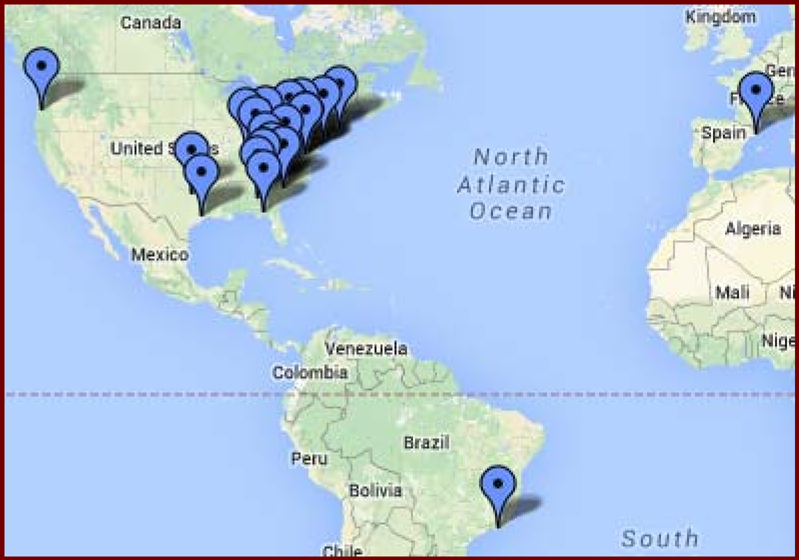
Fixity report
Project name IMC_M175_checksums_fixity4
Algorithm used md5
Date 2014-07-16
Total Files 20
Confirmed Files 20
Moved or Renamed Files 0
New Files 0
Changed Files 0
Removed Files 0

http://www.avpreserve.com/

Distributed Digital Preservation

LOCKSS-based MetaArchive Cooperative

- At least 6 copies, geographically dispersed
- Fixity-checking (not just back-ups)



Keep track of DP actions

- File migrations
  - Obsolete file formats
  - Proprietary file formats
- Metadata changes

Possibilities for documenting DP events:

- ArchiveSpace
- PREMIS Event Service

Future Plans

Additional electronic records from the Mary O'Hara Papers (e.g. data DVDs or CDs)

Replicate DP system but portable

DP in a box

Beyond preservation...

- Hand off to our archivists (policy questions)
- Provide access to the MOH electronic records
- Create links within the MOH finding Aid

Additional Resources

ASERL Webinars re: Digital Preservation (Spring 2013)
http://www.aserl.org/intro-dp-2013/

OCLC Research “Demystifying Born Digital” Reports (2012-13)
http://www.oclc.org/research/publications/library/2012/2012-06r.html

BitCurator Project White Paper: “Bitstreams to Heritage: Putting Digital Forensics into Practice in Collecting Institutions” (September 2013)
http://www.bitcurator.net/docs/bitstreams-to-heritage.pdf

ARL SPEC Kit 329: Managing Born-Digital Special Collections and Archival Materials (August 2012)
http://publications.arl.org/Managing-Born-Digital-Special-Collections-and-Archival-Materials-SPEC-Kit-329

Q & A

Bill Donovan	bill.donovan@bc.edu
Jack Kearney	kearneyj@bc.edu

Archival policy questions

- Preserve just the digital files or the entire disk image?
- Delete the virus-infected files?
- Save the duplicate 8.3 files?
- Delete extraneous files? (And, define extraneous)
- Decide fate of PII files:
    - Credit cards
    - Bank accounts
    - Social Security
- Normalize file formats?
- Which files to preserve?
    - External hard drive
    - As-is copies of original files
    - Remediated copies (And, define remediated)
- What sorts of documentation to preserve and how?

DP Workstation

hardware:

- PC desktop computer with 64-bit Windows 7
- 4 GB RAM and a 465 GB internal hard drive
- UltraKit III + FireWire Write-Blocker

software:

- Cygwin 64 Terminal (md5sum)
- Immunet 3 powered by ClamAV
- Identity Finder
- FITS
- Fixity
- Access Data FTK Imager
- dBpoweramp Music Converter
- HxD Hex Editor

security measures:

- authorized personnel only room
- security cable for workstation
- need-to-know only username/password