Developing a Born-Digital Preservation Workflow

Jack Kearney & Bill Donovan

Audiovisual Archives Assistant
John J. Burns Library

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

Boston College Libraries

April 8, 2014

Goals

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

Who is Mary O’Hara?

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.

The Mary O'Hara (MOH) papers include promotional materials, business correspondence, recordings, sheet music, books,..., and a hard drive used to transfer files from her personal computer.

Institutional Context

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

Since the Mary O'Hara papers were already being processed, also begin evaluating the electronic records on the hard drive.

Digital preservation much discussed at meetings attended by BC staff from Digital Libraries, Archives, and the Irish Music Center

We anticipated needing to manage many more born-digital collections

Opportunity for collaborative project involving all 3 groups

Focus: digital preservation of MOH electronic records

Chain of Custody

August 2009:

- Mary O'Hara's electronic files were copied to a 750 GB Seagate FreeAgent Pro external hard drive at her home in England
- The hard drive was shipped to the Burns Library at Boston College, where the files were copied for backup
- The hard drive itself was then stored in a secure, climate-controlled location within the Burns Library

January 2014:

- The hard drive was transported to the O'Neill Library DigLab to use as a test-case in developing a born-digital preservation workflow

Where does forensics come in?

"Digital forensics focuses on the use of hardware and software tools to collect, analyze, interpret, and present information from digital sources, and ensuring that the collected information has not been altered in the process."

Gengenbach, M.J. (2012) "The Way We Do it Here": Mapping Digital Forensics Workflows in Collecting Institutions. A Master's Paper for the M.S. in Library Science degree, University of North Carolina at Chapel Hill.

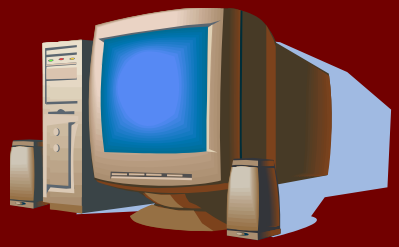
How to not alter the information

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

USB

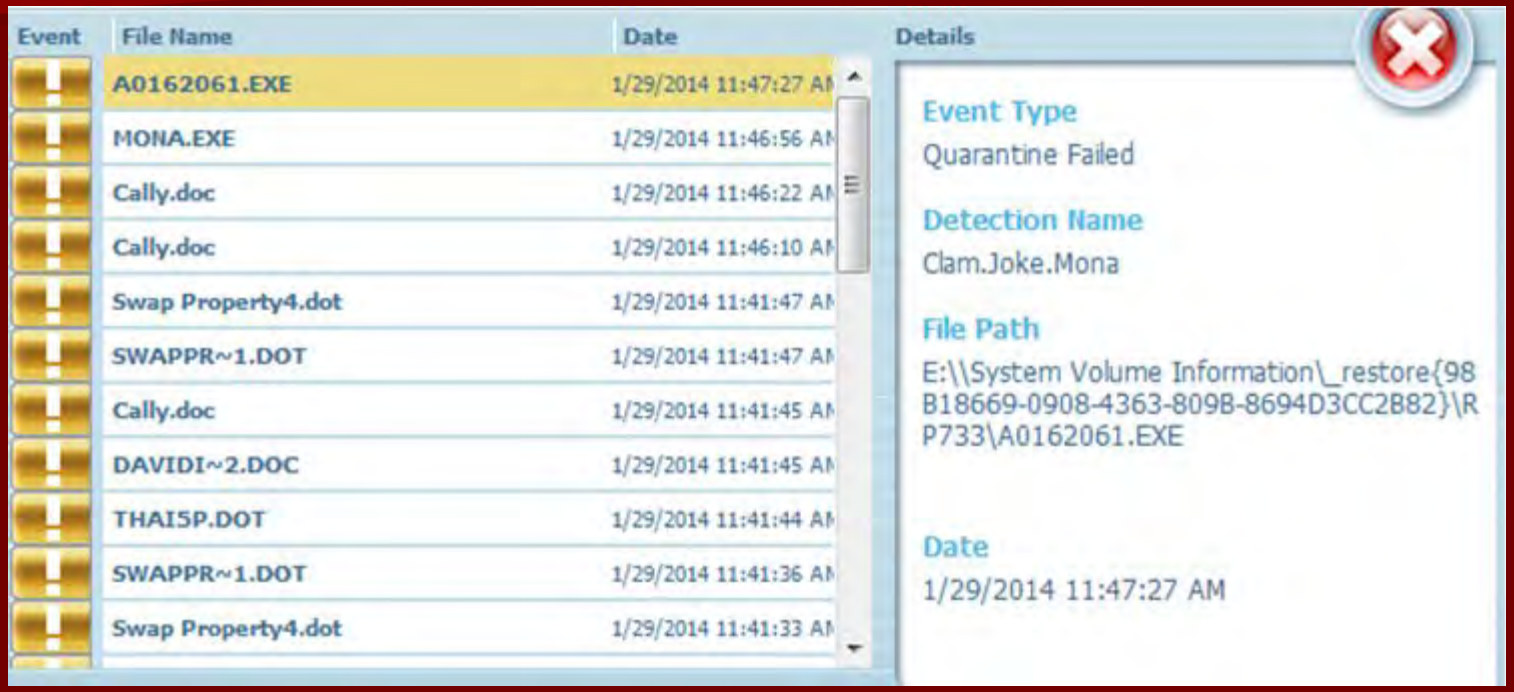
USB

DP Workstation



(offline)

Is the information virus-free?



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?

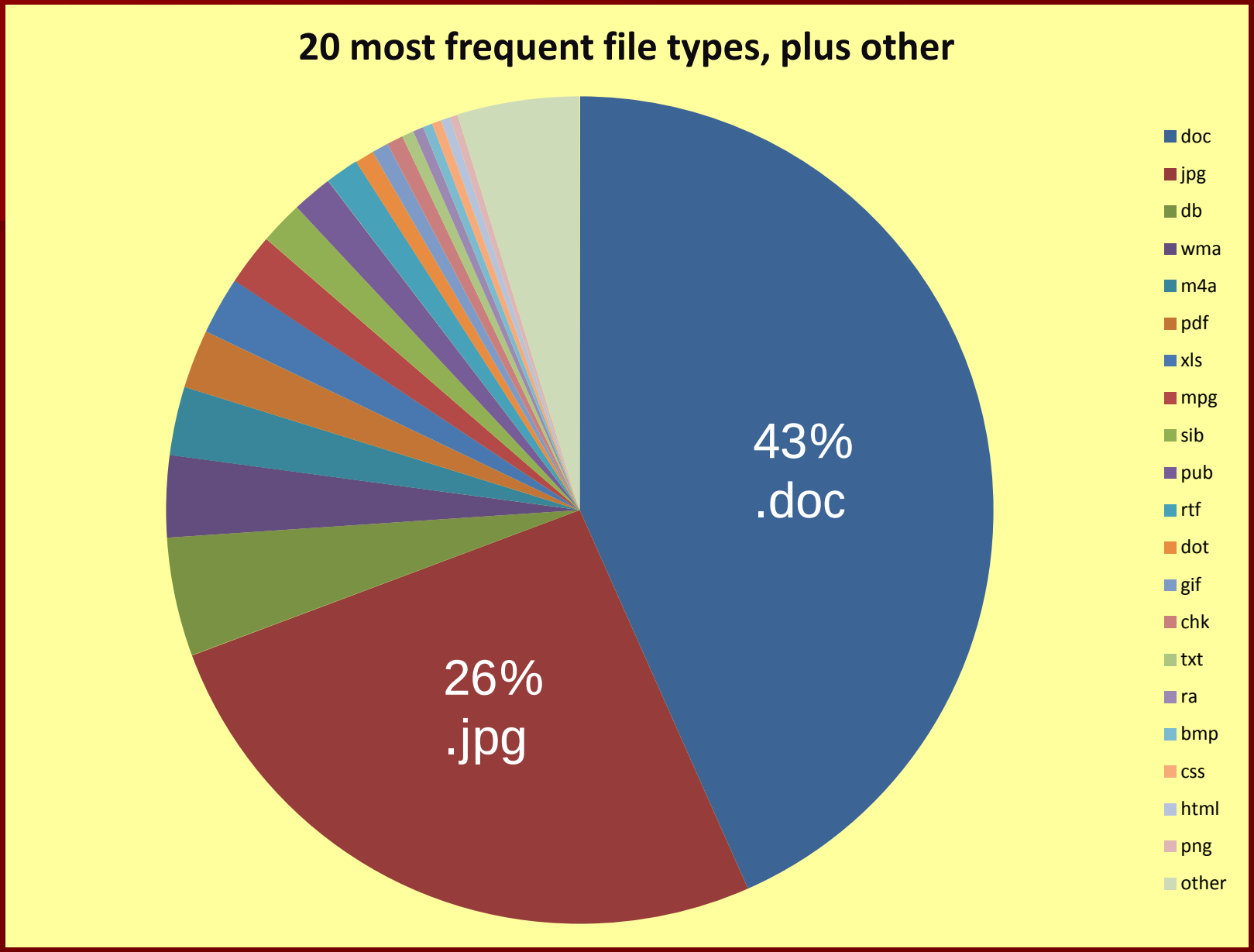
Inventory taken with this Unix command:

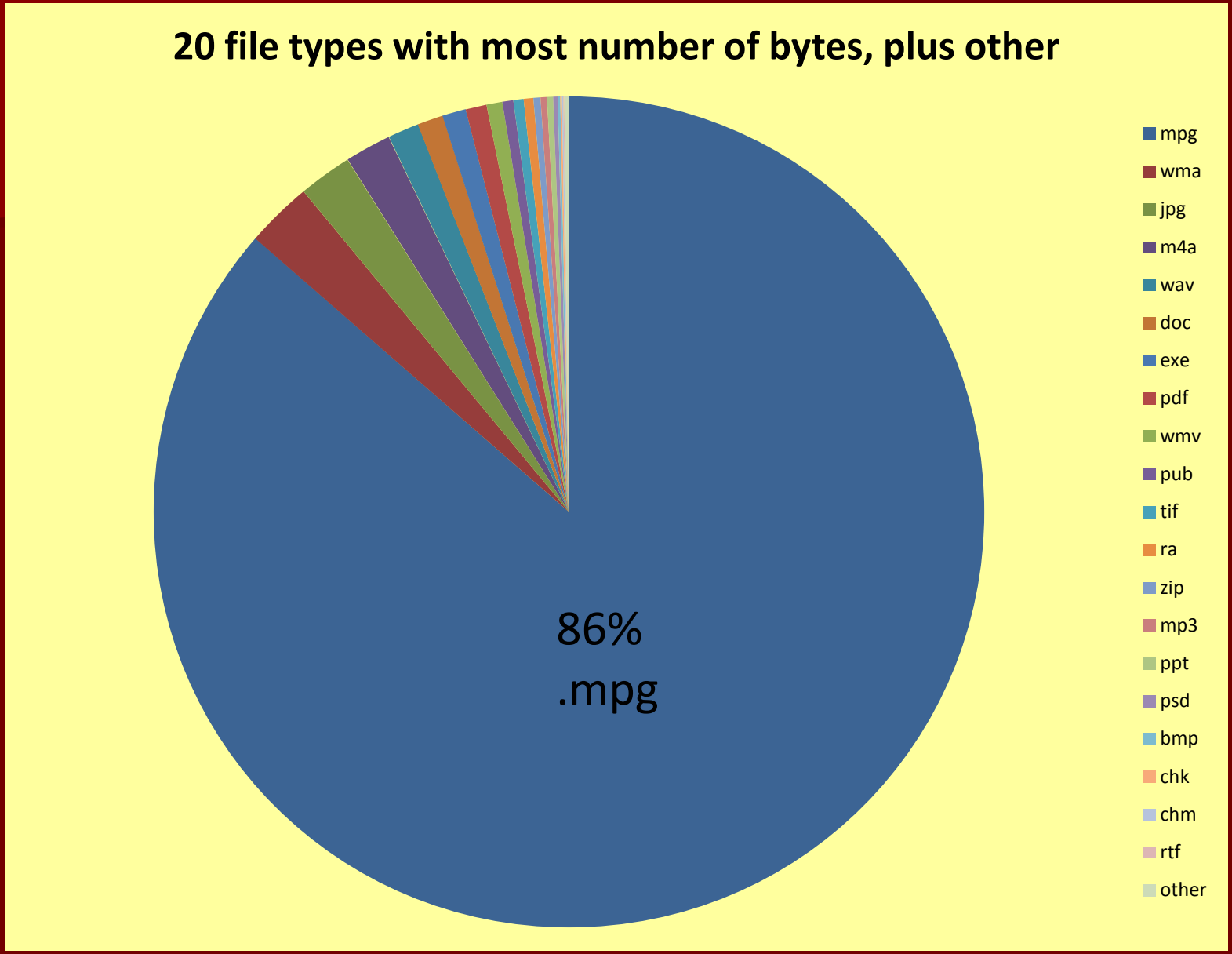
find *directory-name* -type f -exec ls -l {} ; >c:\data\MOH\inventory.txt

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

If this were a "working" drive there may also be additional information besides these files, e.g. deleted files that have not been overwritten

Decision: In this case we decided that we will just transfer the files to our DP workstation, rather than creating a disk image of the entire hard drive

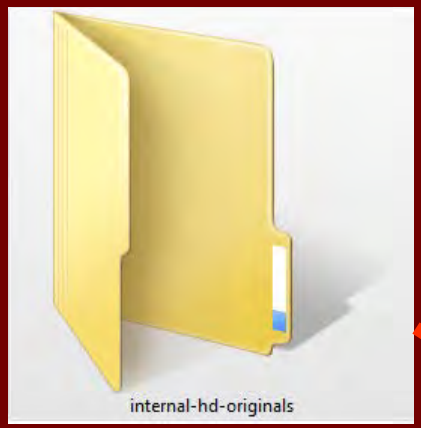

20 most frequent file types, plus other
43%
.doc
26%
.jpg
doc
jpg
db
wma
m4a
pdf
xls
mpg
sib
pub
rtf
dot
gif
chk
txt
ra
bmp
css
html
png
other


20 file types with most number of bytes, plus other
86%
.mpg
mpg
wma
jpg
m4a
wav
doc
exe
pdf
wmv
pub
tif
ra
zip
mp3
ppt
psd
bmp
chk
chm
rtf
other

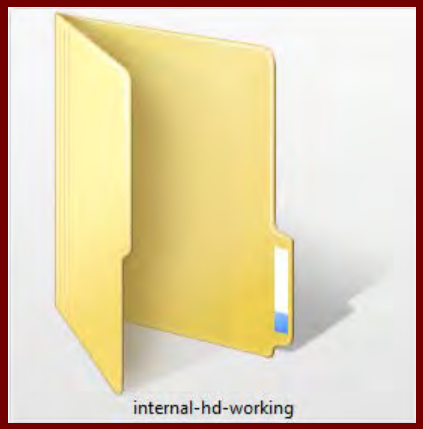
Transfer of files

MOH hard drive
with original files

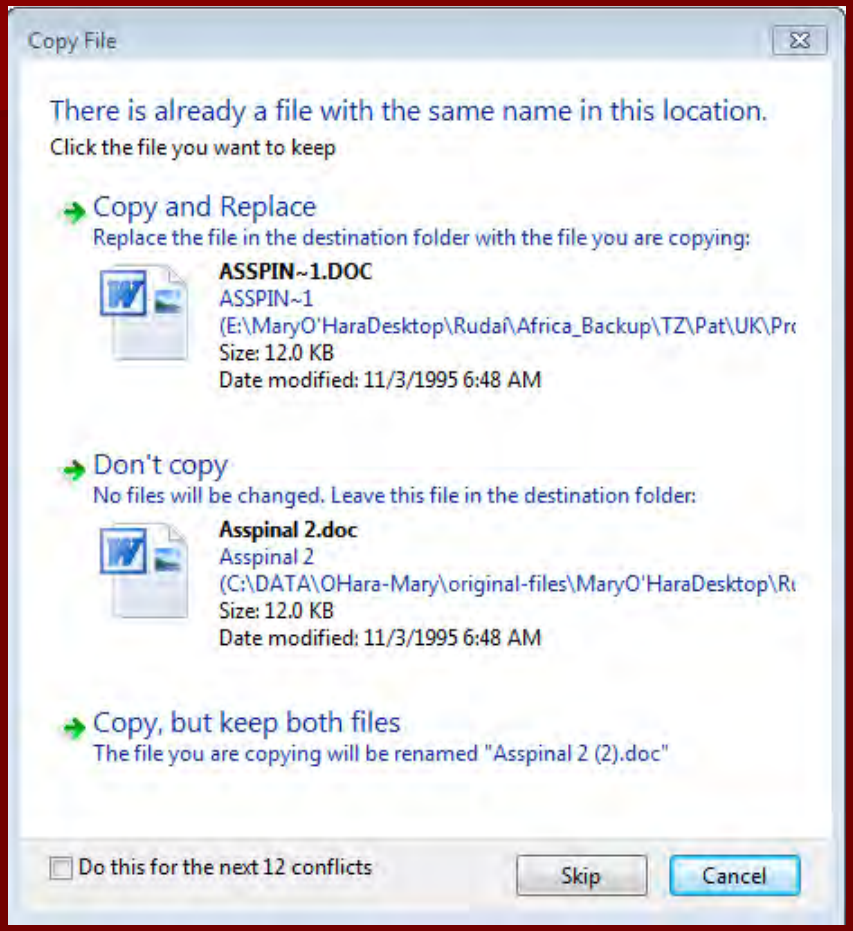
DP Workstation's
internal hard drive:
"Original files"



DP Workstation's
internal hard drive:
"Working files"



Discovered duplicate files



"8.3" constraint

Also, employ de-duping software

Decision: keep only the copy with the more complete filename

Computed initial checksums

What is a“checksum” (aka hash) ?

We computed initial checksums for each set of files

We compared checksums to verify original files and copies identical

For long-term fixity-checking, use automated app “Fixity”

	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

File/folder names

Idiosyncratic folder names on the external hard drive. Had to “escape” problematic characters in folder name when issuing Unix commands.

Local conventions regarding naming files and folders:

- Use letters of the English alphabet and the numerals 0 thru 9.
- Avoid punctuation marks other than underscores or hyphens.
- Do not use spaces.
- Limit file/folder names to 31 characters, including the 3 digit file extension – strive for shorter names.

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

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 value
  - Thumbs.db

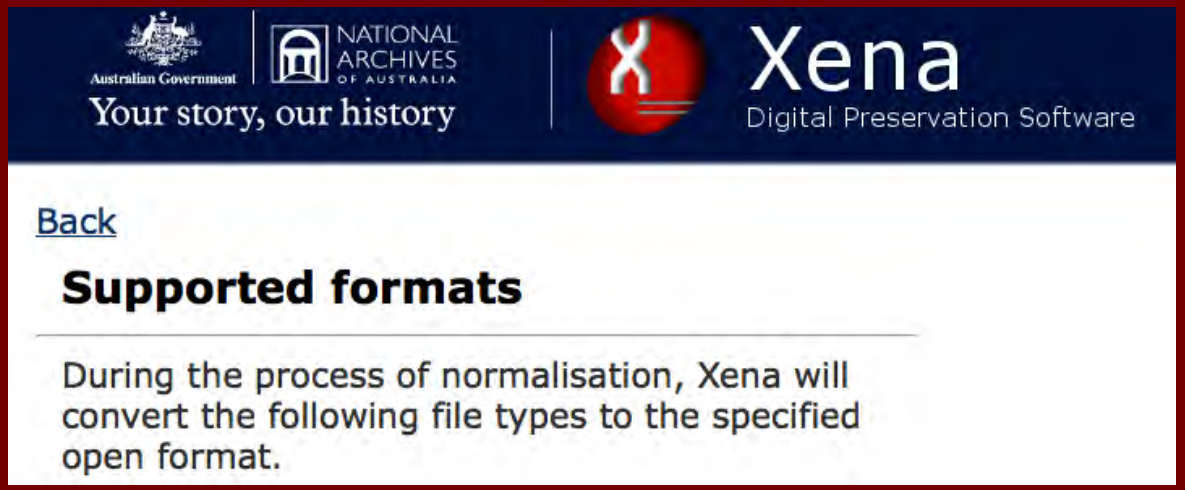
Personally Identifiable Information (PII)

Policy decisions to be made:

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

Any proprietary file formats?

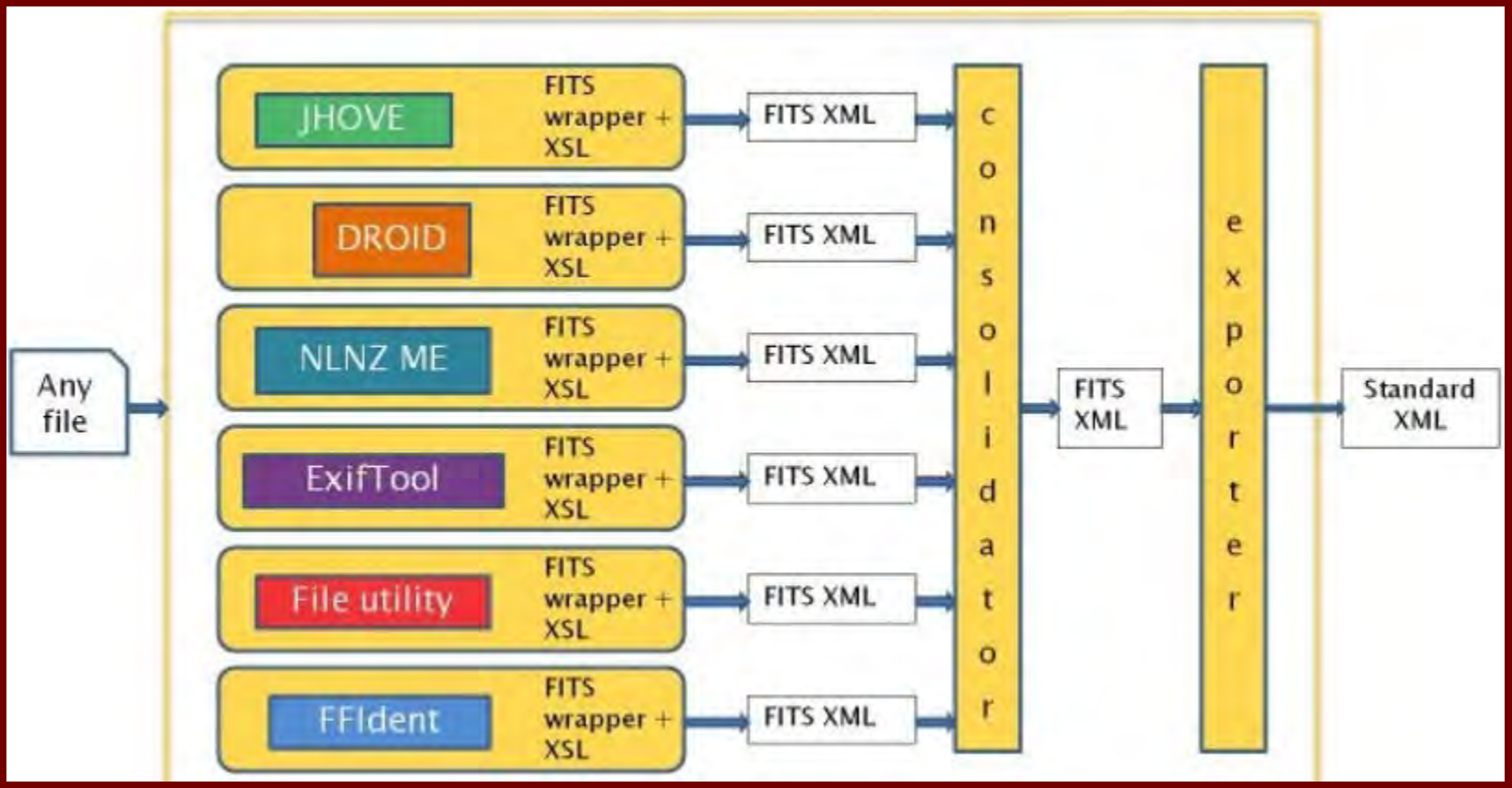
Normalize using Xena ?



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



Output of FITS --- identification

...

<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>

...

Local Archival Copies

Copy remediated files to DP file server.

Compute checksums for new copies.

Verify checksums match. (using “diff” command?)

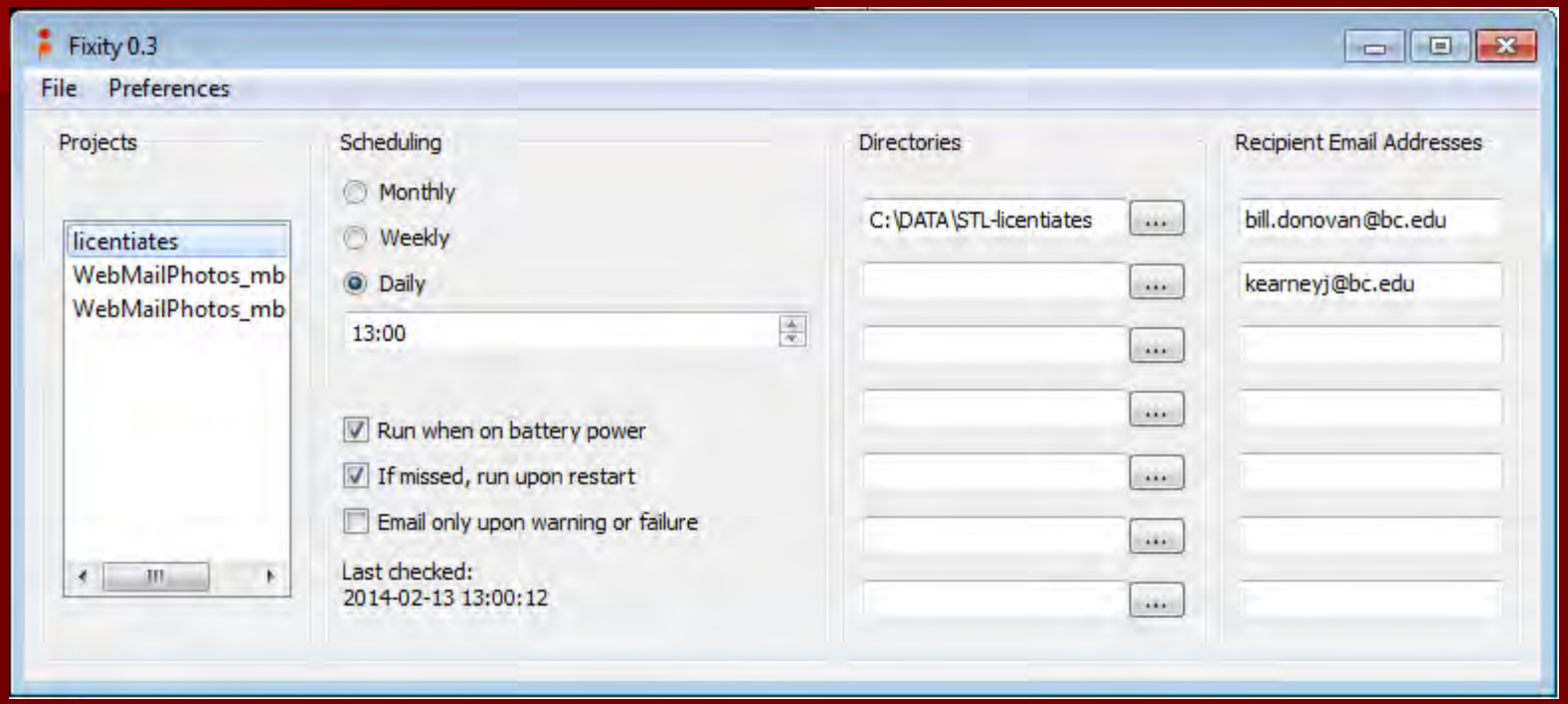
Local long-term fixity-checking

Key: make it simple and automated

Using “Fixity” (available for free)

- Verifies checksums on a schedule
- Per “project” monitor up to 7 folders, recursively
- Looks for...
    - Confirmed Files
    - Moved or Renamed Files
    - New Files
    - Changed Files
    - Removed Files
- Sends emails to notify up to 7 stakeholders

Fixity

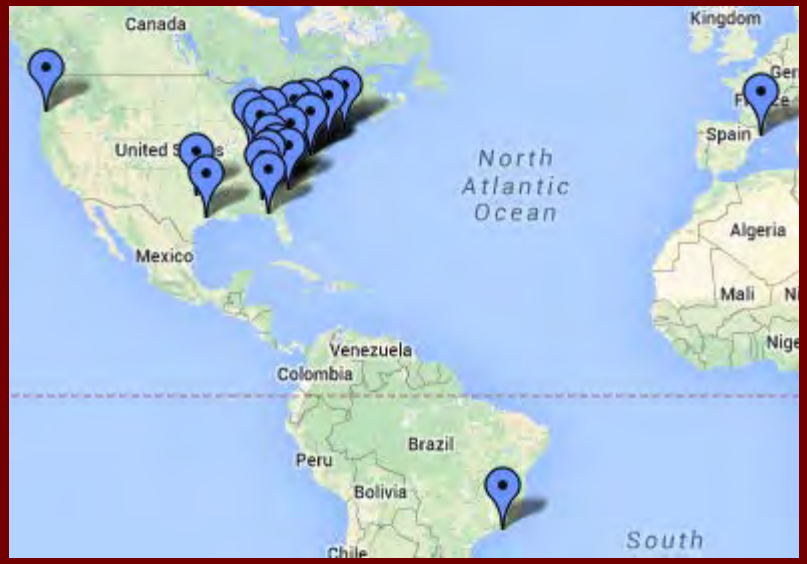


http://www.avpreserve.com/news/fixity-v0-3-released/

Distributed Digital Preservation

LOCKSS-based MetaArchive Cooperative

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



Queue up ingest into MetaArchive Cooperative or equivalent

What if file(s) corrupted?

If checksum discrepancies are discovered....

Or the checksums themselves are corrupted....

From multiple back-up copies, including multiple sets of checksums,

replace corrupted files with pristine copies

Keep track of DP actions

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

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

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