



Using Forensic Software to Assign Metadata to Born Digital Archives

Peter Chan, Digital Archivist

Metadata and Digital Object Roundtable (MDOR) Meeting
2011, SAA, Chicago

Challenge – File Formats Identification

- JHOVE and other similar tools designed for digital preservation can only identify very limited file formats.
- DROID cannot recognize the 3,574 WordPerfect files in the Robert Creeley Collection (333,557 files total).

Challenge – View File

- Can we assign descriptive metadata without knowing the contents? (similar to a letter in an unopened envelope).
- Many files in born digital collections were created years ago with obsolete software (e.g. WordPerfect, Lotus 1-2-3, etc.)
- QuickView Plus

Challenge – Annotate

- How we going to record the metadata we assigned to files (database, spreadsheet, xml, etc.)?
- No annotation function in QuickView Plus

Challenge – Search

- MS Search 4.0 (recognize MS Office documents, PDF, ASCII files; Not WordPerfect, Lotus 1-2-3)
- No pattern search (social security number, credit card number, etc.)

Challenge – Report

- Reports, if any, from several independent software to interface with the delivery and preservation platform.

Need

- An integrated tool to perform
 - Identify hundreds of file formats
 - View Files with obsolete file format
 - Annotate metadata to individual/group of files
 - Search for sensitive information in files with obsolete file format
 - Report to access & delivery and preservation platforms

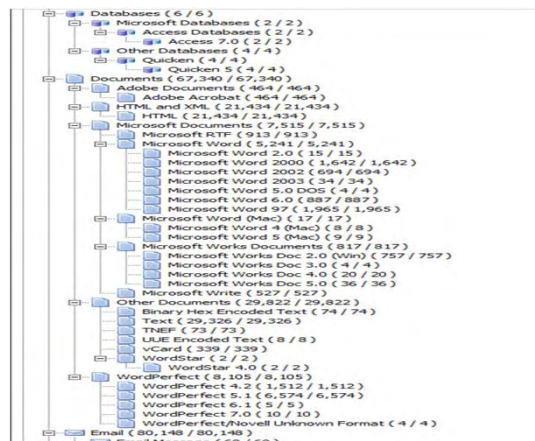
AccessData FTK

- Identify ~300 file formats
 - <http://accessdata.com/downloads/media/Recognized%20File%20Types%20FTK2%20%207-28-08.pdf>
- Embedded viewer.
- Bookmark and label functions.
- Full text and pattern search.

AccessData FTK

- The XML report from FTK is in fact a XSL-FO document containing all the formatting scripts for a XSL formatter to output the file to PDF.
- A XSLT program was written to transform the XSL-FO output from FTK to XML content document. A Ruby program was written to ingest the XML content document, original files, and the display derivatives to Fedora.

File Format Identification



Bookmark Function

AccessData Forensic Toolkit Version: 3.2.0.32216 Database: localhost Case: M1437 Stephen Jay Gould

File Edit View Evidence Filter Tools Manage Help

Filter: -unfiltered-

Explore Overview Email Graphics Bookmarks Live Search Index Search Volatile

Evidence Items

Fig. 1-1
the Museum
for February
largest animal
Anomalocaris
(the large crea
Fig. 1-2
Whittington
by two spec
on sponges a
Wiwaxia gra

1940. This original illustration in reconstruction in National Geography groups. Above *Sidneyia*, the belong to the unique creature the rear end of a bivalved arthropod a 1985 article by Briggs and right, and the figure is dominated odd organisms. Three *Aysheeha* feed from just left of *Aysheeha*. Two

File Content

File Category: All Highlighted All Checked All Listed

1 Item selected

FISICAP CH004.ad1\A\A\FISICAP

File Comment

File Content

Allow include

Embed attachments Parent Email

Bookmark Selection in File

Select Existing Bookmarks

- Series VI: Stephen Jay Gould Born-Digital Materials
 - 01 Subseries II: Article
 - 02 Subseries III: Abstracts, Reviews, Letters, etc.
 - 03 Subseries III: Natural History Column
 - 04 Subseries IV: Books
 - 05 Subseries V: Bibliography & CV
 - 06 Subseries VI: Teaching
 - 07 Subseries VII: SSG Rare Books
 - 08 Subseries VIII: Punch Cards
 - 09 Subseries IX: Misc.
 - 10 Subseries X: Computer Media Photos
 - Restricted Files

File List

Name	Path	Item #	Ext	Display	Category	P-Size	L-Size	MDS	SHA1	SHA256	Created	Accessed	Modified
BUSINESS	CH004.ad1\A\A\BUSINESS	30009	emsgmpt	Unknown	n/a	6.0 KB	05545...	49383...	90CF...	12/31/1979 11...	12/31/1979 11...	5/19/1980 9:36...	
FISICAP	CH004.ad1\A\A\FISICAP	30009	emsgmpt	Unknown	n/a	44.36 KB	A9270...	0123F...	39D58...	12/31/1979 11...	12/31/1979 11...	6/9/1980 7:54...	
INSP	CH004.ad1\A\A\INSP	30009	emsgmpt	WordPerfect 4.2	n/a	44.41 KB	F481B...	A219F...	878D4...	12/31/1979 11...	9/14/1980 8:22...	6/9/1980 8:22...	
TASCAN	CH004.ad1\A\A\TASCAN	40212	emsgmpt	Unknown	n/a	17.71 KB	50584...	6119B...	27399...	12/31/1979 11...	12/31/1979 11...	4/20/1980 5:4...	
WAXIA	CH004.ad1\A\A\WAXIA	40212	emsgmpt	Unknown	n/a	40212	E37EE...	5B13F...	48A9D...	12/31/1979 11...	12/31/1979 11...	9/5/1980 5:4...	

AccessData Forensic Toolkit Version: 3.2.0.32216 Database: localhost Case: M1437 Stephen Jay Gould

File Edit View Evidence Filter Tools Manage Help

Filter: -unfiltered-

Explore Overview Email Graphics Bookmarks Live Search Index Search Volatile

Case Overview

Bookmarks

- Series VI: Stephen Jay Gould Born-Digital Materials
 - 01 Subseries II: Abstracts, Reviews, Letters, etc.
 - 02 Subseries III: Natural History Column
 - 03 Subseries IV: Books
 - 04 Subseries V: Bibliography & CV
 - 05 Subseries VI: Teaching
 - 06 Subseries VII: SSG Rare Books
 - 07 Subseries VIII: Punch Cards
 - 08 Subseries IX: Misc.
 - 09 Subseries X: Computer Media Photos
 - Restricted Files

File List

Name	Label	Item #	Ext	Path	Category	P-Size	L-Size	MDS	SHA1	SHA256	Created	Accessed	Modified
ACCERC3	A.CH.S...	30009	emsgmpt	CH004.ad1\A\A\ACCERC3	Unknown	n/a	6.0 KB	05545...	49383...	90CF...	12/31/1979 11...	12/31/1979 11...	5/19/1980 9:36...
ALICMDC.BK	A.CH.S...	72002	bkf	CH070.ad1\A\A\ALICMDC.BK	WordP...	n/a	44.36 KB	A9270...	0123F...	39D58...	12/31/1979 11...	12/31/1979 11...	6/9/1980 7:54...
ALICMDC.WPD	A.CH.S...	72003	wpd	CH070.ad1\A\A\ALICMDC.WPD	WordP...	n/a	44.41 KB	F481B...	A219F...	878D4...	12/31/1979 11...	9/14/1980 8:22...	6/9/1980 8:22...
ANDERSON.WSD	A.CH.S...	37012	wsd	CH070.ad1\A\A\ANDERSON.WSD	WordP...	n/a	30.83 KB	2C9F5...	4794E...	D007F...	12/31/1979 11...	12/31/1979 11...	12/31/1981 3:5...
BOALINDICA	A.CH.S...	75014	emsgmpt	CH066.ad1\A\A\BOALINDICA	WordP...	n/a	23.35 KB	83C62...	83367...	035E9...	12/31/1979 11...	12/31/1979 11...	12/31/1981 5:1...
BOALINDICA	A.CH.S...	155014	emsgmpt	CH066.ad1\A\A\BOALINDICA	WordP...	n/a	23.35 KB	CAD86...	F9469...	D4F03...	12/31/1979 11...	12/31/1979 11...	12/9/1980 8:41...
BOULBAW.BK	A.CH.S...	77018	bkf	CH066.ad1\A\A\BOULBAW.BK	WordP...	n/a	26.07 KB	3442E...	96919...	83C09...	12/31/1979 11...	12/31/1979 11...	8/20/1980 9:8...
BOULBAW.WPD	A.CH.S...	77020	wpd	CH066.ad1\A\A\BOULBAW.WPD	WordP...	n/a	63.33 KB	21AD7...	68766...	58413...	12/31/1979 11...	9/16/1980 12:0...	10/25/1981 11...
CEP3ADST	A.CH.S...	16300	emsgmpt	CH008.ad1\A\A\CEP3ADST	Unknown	n/a	2229 B	7983C...	AC4E3...	40CF...	12/31/1979 11...	12/31/1979 11...	5/14/1980 7:46...
CEP3ADST	A.CH.S...	16300	emsgmpt	CH008.ad1\A\A\CEP3ADST	Unknown	n/a	45.62 KB	9630A...	4948E...	6E20E...	12/31/1979 11...	12/31/1979 11...	11/16/1981 11...
CEP3ADST	A.CH.S...	30303	emsgmpt	CH008.ad1\A\A\CEP3ADST	WordP...	n/a	35.31 KB	87A1C...	C0923...	3F9E5...	12/31/1979 11...	12/31/1979 11...	5/14/1980 7:46...
CEP3ADST.WPD	A.CH.S...	77021	wpd	CH066.ad1\A\A\CEP3ADST.WPD	WordP...	n/a	38.50 KB	09F02...	29E70...	0930C...	8/21/1980 2:58...	9/16/1980 12:0...	8/21/1980 2:58...
COMHERST	A.CH.S...	40101	emsgmpt	CH055.ad1\A\A\COMHERST	WordP...	n/a	26.96 KB	8C3C6...	EE8C2...	70000...	12/31/1979 11...	12/31/1979 11...	9/5/1980 5:56...
COMHERST	A.CH.S...	156002	emsgmpt	CH055.ad1\A\A\COMHERST	WordP...	n/a	26.96 KB	8C3C6...	EE8C2...	70000...	12/31/1979 11...	12/31/1979 11...	9/5/1980 5:56...
COMHERST	A.CH.S...	156002	emsgmpt	CH055.ad1\A\A\COMHERST	WordP...	n/a	26.96 KB	8C3C6...	EE8C2...	70000...	12/31/1979 11...	12/31/1979 11...	9/5/1980 5:56...
COMHERST	A.CH.S...	153003	emsgmpt	CH025.ad1\A\A\COMHERST	WordP...	n/a	76.30 KB	301F3...	C7709...	207FC...	12/31/1979 11...	12/31/1979 11...	8/20/1980 9:46...
CONSTRAT	A.CH.S...	146003	emsgmpt	CH010.ad1\A\A\CONSTRAT	WordP...	n/a	74.25 KB	0F4E3...	0E21F...	03211...	12/31/1979 11...	12/31/1979 11...	3/23/1980 7:35...
CONSTRAT	A.CH.S...	146003	emsgmpt	CH010.ad1\A\A\CONSTRAT	WordP...	n/a	47.39 KB	E786F...	A3E3D...	40E79...	12/31/1979 11...	12/31/1979 11...	1/25/1980 12:3...
PROGTEST.WPD	A.CH.S...	80003	wpd	CH004.ad1\A\A\PROGTEST.WPD	WordP...	n/a	26.96 KB	8C3C6...	EE8C2...	70000...	12/31/1979 11...	12/31/1979 11...	9/5/1980 5:56...

SU LAIR

Assigning Administrative and Descriptive Metadata Using Label Function

- Use "Labels" to represent access restrictions, document types, computer media, and subject headings. When files are exported to access repository (Hypatia), they will carry the labels with them.

Access restrictions:
 AR:Owner
 AR:Archivist
 AR:Invited person
 AR:Public
 AR:Reading room

File types:
 FT:Document
 FT:Spreadsheet
 FT:Email
 FT:Image
 FT:Video
 FT:Sound

Computer media:
 CM:5.25 Floppy 5.25 inch. floppy diskettes
 CM:3.5 Floppy 3.5 inch. floppy diskettes
 CM:Punch Card Punch cards
 CM:Tape Open reel (0.5, 0.25, 0.125, etc.); cartridge ; DAT, etc. tape

All Subject Headings start with "S:".

SU LAIR

AccessData Forensic Toolkit Version: 3.2.0.32216 Database: localhost Case: M1437 Stephen Jay Gould

Labels dialog box:

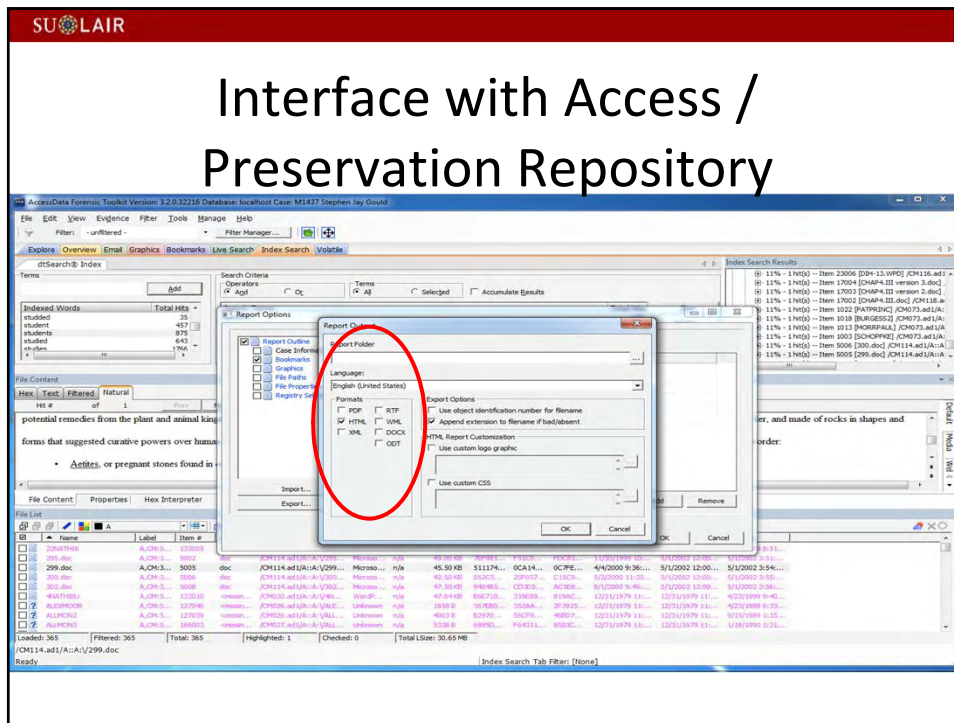
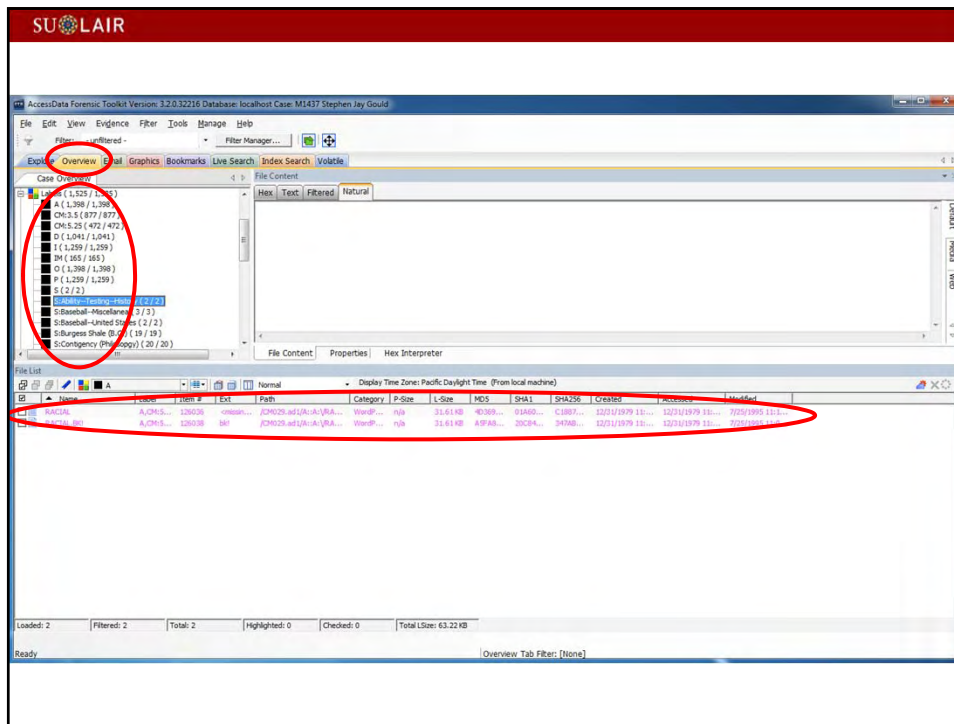
Note: To use the labels assigned to each item, the Label column must be selected on Column Settings

All Highlighted
 All Checked
 All Locked
 Extend labels to associated (Family) files

Name	Color
S:01	
S:02	
S:03	
S:04	
S:05	
S:Ability-Testing-History	

File List:

Name	Path	Ext	Category	Size	MDS	Created	Accessed	Modified	Duplicate File	Flags...
BD01-2	[CH004.ad1]A:\A\BD01-2	missing	Unknown	4127B	0864016AACC3E2F735A07E909F9C203	12/31/1979 ...	12/31/1979 ...	11/11/1988 ...	False	
BD01-3	[CH004.ad1]A:\A\BD01-3	missing	Unknown	9.81 KB	348CE808BDC67317AAD37FACF81249	12/31/1979 ...	12/31/1979 ...	11/29/1998 ...	False	
BD01-6	[CH004.ad1]A:\A\BD01-6	missing	Unknown	16.28 KB	C2695452D4865396531846C70212126	12/31/1979 ...	12/31/1979 ...	11/29/1998 ...	False	
FIGCAP	[CH004.ad1]A:\A\FIGCAP	missing	WordPerfect 4.2	29.07 KB	D791992949FC7F38085230E19A525CD	12/31/1979 ...	12/31/1979 ...	12/20/1988 ...	False	



XML-FO

```

1 <?xml version="1.0" encoding="UTF-8"?>
2 <?root xmlns:fo="http://www.w3.org/2001/XMLSchema" font-family="arial unicode ms" font-size="10pt"
3   xmlns:fo="http://www.w3.org/1999/XSL/Format">
4   <fo:layout-master-set>
5     <fo:simple-page-master master-name="IOC">
6       <fo:region-body region-name="body" margin-top="0.5in" margin-bottom="1in"
7         margin-left="0.5in" margin-right="0.5in"/>
8       <fo:region-before extent="0.5in"/>
9       <fo:region-after extent="0.5in"/>
10    </fo:simple-page-master>
11    <fo:simple-page-master master-name="caseInfoPage">
12      <fo:region-body region-name="body" margin-top="0.5in" margin-bottom="1in"
13        margin-left="0.5in" margin-right="0.5in"/>
14      <fo:region-before extent="0.5in"/>
15      <fo:region-after extent="0.5in"/>
16    </fo:simple-page-master>
17    <fo:simple-page-master master-name="bookmarkPage">
18      <fo:region-body region-name="body" margin-top="0.5in" margin-bottom="1in"
19        margin-left="0.5in" margin-right="0.5in"/>
20      <fo:region-before extent="0.5in"/>
21      <fo:region-after extent="0.5in"/>
22    </fo:simple-page-master>
23    <fo:simple-page-master master-name="index">
24      <fo:region-body region-name="body" margin-top="0.5in" margin-bottom="1in"
25        margin-left="0.5in" margin-right="0.5in" column-count="2" column-gap="1pc"/>
26      <fo:region-before extent="0.5in"/>
27      <fo:region-after extent="0.5in"/>
28    </fo:simple-page-master>
29  </fo:layout-master-set>
30  <fo:page-sequence master-reference="IOC">
31    <fo:flow flow-name="body">
32      <fo:block text-align="left" justify="space-between"12pt"><fo:basic-link color="blue"
33        font-weight="bold" text-decoration="underline"
34        internal-destination="CaseSummaryMainSection" font-size="14pt">Case
35        Summary</fo:leader leader-pattern="space" leader-alignment="reference-area"
36        ><fo:page-number-citation ref-id="CaseSummaryMainSection"
37        /></fo:leader</fo:basic-link</fo:block>

```

XSLT Transformation – XML Content

```

1 <?xml version="1.0" encoding="UTF-8"?>
2 <?xml-stylesheet href="http://www.w3.org/1999/XSL/Format" type="text/xsl" />
3 <series>Series 13: Computer Files</series>
4 <collection_title>Robert Creeley papers</collection_title>
5 <callnumber>M0662</callnumber>
6 <did>
7   <unittitle>creeley.wmv</unittitle>
8   <unitid>209002</unitid>
9   <filepath>M0662 CM050.ad1/D:/D:\creeley.wmv</filepath>
10  <num>CM050</num>
11  <extent>90738025 B</extent>
12  <unitdate>4/3/2002 4:06:28 AM (2002-04-03 12:06:28 UTC)</unitdate>
13  <accessrestrict>False</accessrestrict>
14  <scopecontent>[event]Poetry Reading at Albright-Know Gallery</scopecontent>
15  <physdesc>Windows Media Video</physdesc>
16  <export_path>files\creeley.wmv</export_path>
17 </did>
18 <did>
19   <unittitle>Arakawa.JPG</unittitle>
20   <unitid>212167</unitid>
21   <filepath>M0662 CM048.ad1/D:/D:\Photo\arakawa\Arakawa.JPG</filepath>
22   <num>CM048</num>
23   <extent>195103 B</extent>
24   <unitdate>11/29/2000 9:45:44 AM (2000-11-29 17:45:44 UTC)</unitdate>
25   <accessrestrict>False</accessrestrict>
26   <scopecontent>[subject]Family Photos</scopecontent>
27   <physdesc>JPEG EXIF</physdesc>
28   <export_path>files\Arakawa.JPG</export_path>
29 </did>

```