SU LAIR

Stanford University LIBRARIES & ACADEMIC INFORMATION RESOURCES

Using Forensic Software to Assign Metadata to Born Digital Archives

Peter Chan, Digital Archivist

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

SU LAIR

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).

SU LAIR

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

SU LAIR

Challenge – Annotate

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

SU LAIR

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.)

SU LAIR

Challenge – Report

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

SU LAIR

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

SU LAIR

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.

SU LAIR

AccessData FTK

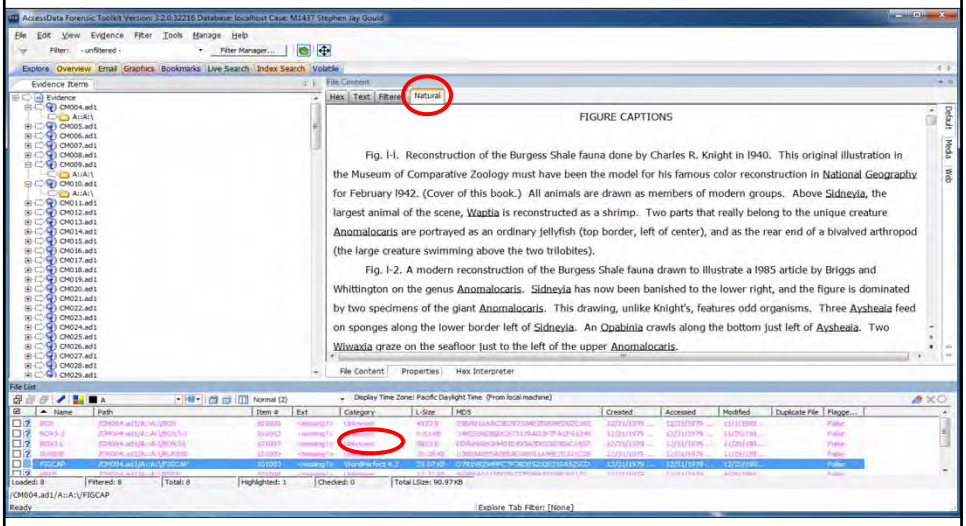
- The XML report from FTK is in fact a XSL-FO document containing all the formatting scripts for a XSL formatter to output the file to PDF.
- A XSLT program was written to transform the XSL-FO output from FTK to XML content document. A Ruby program was written to ingest the XML content document, original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

SU LAIR

File Format Identification

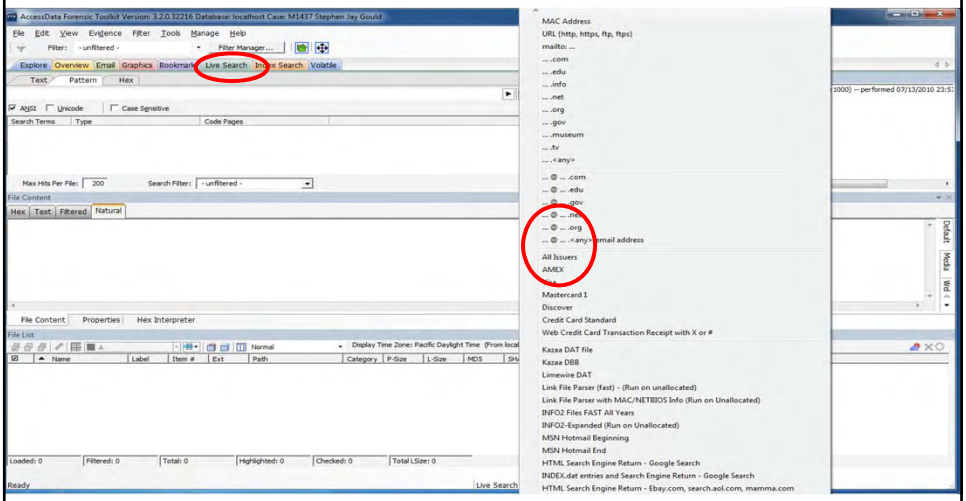
SUMURI

View Files in “obsolete” File Formats



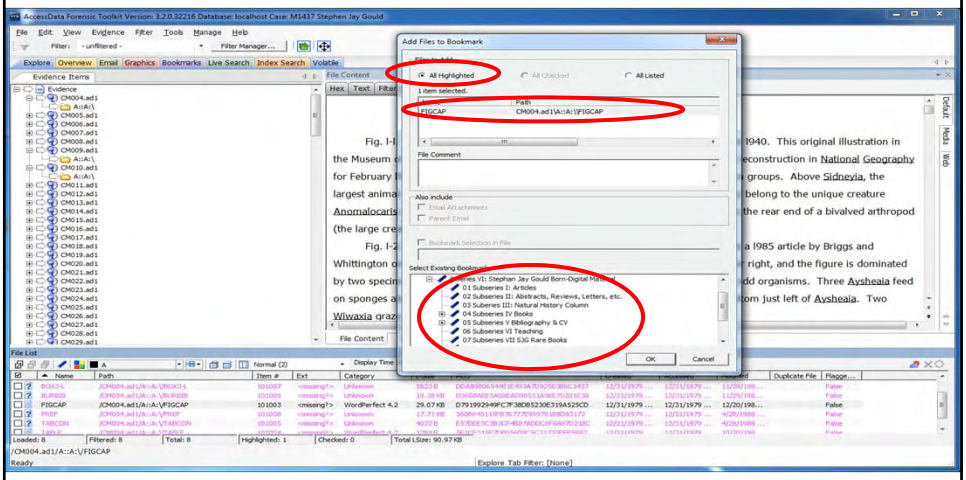
SUMURI

Pattern Search

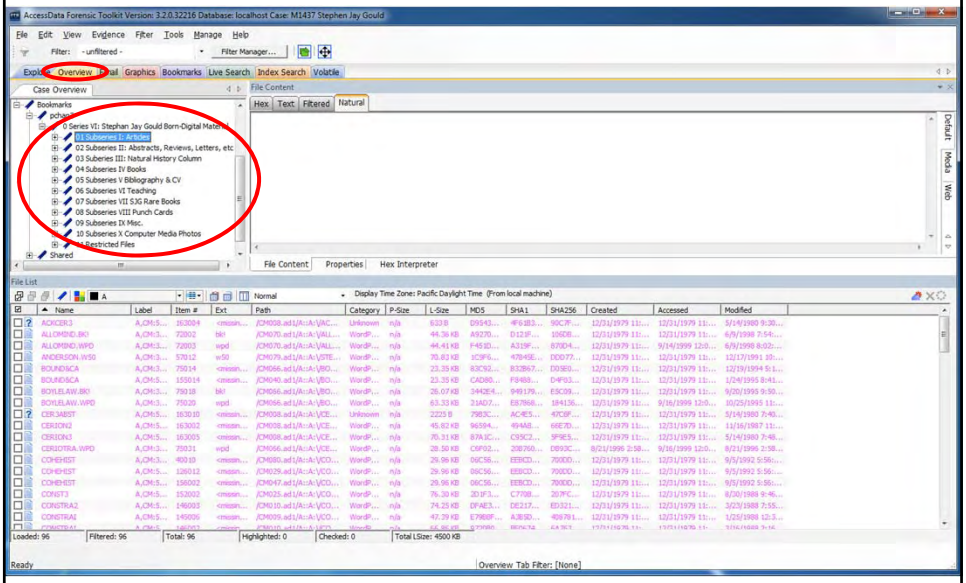


SUNLAIR

Bookmark Function



SUNLAIR



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

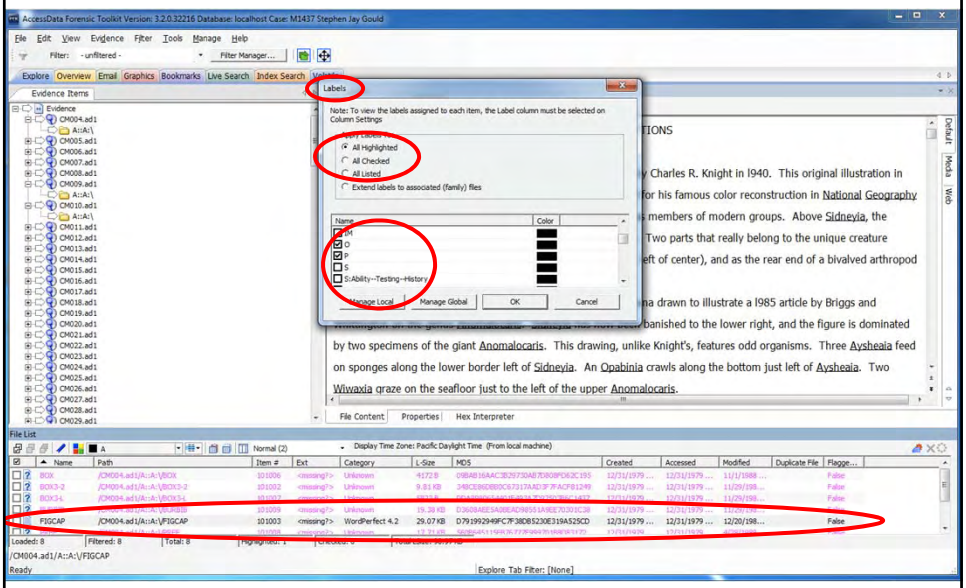
File types:
FT:Document
FT:Spreadsheet
FT:Email
FT:Image
FT:Video
FT:Sound

Computer media:

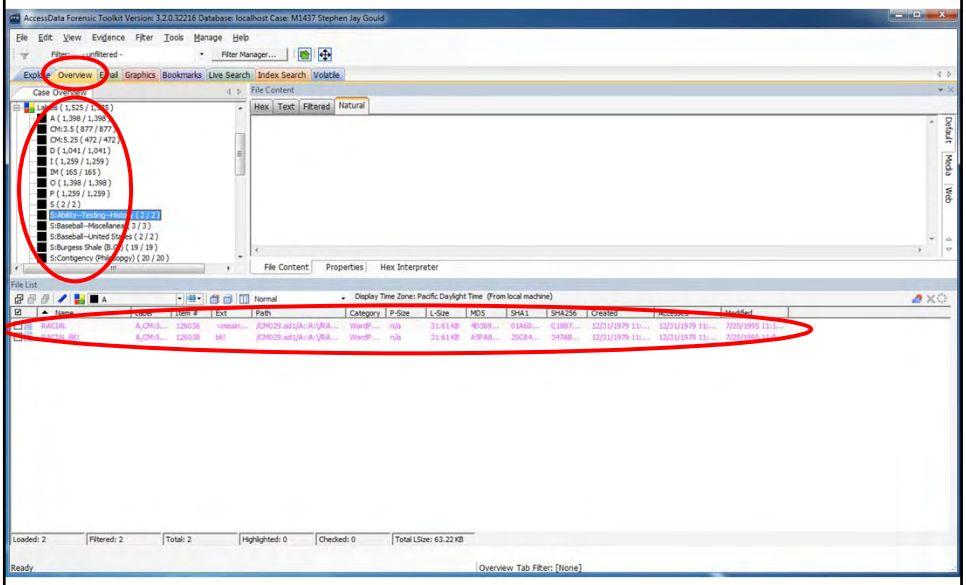
CM:5.25 Floppy	5.25 inch. floppy diskettes
CM:3.5 Floppy	3.5 inch. floppy diskettes
CM:Punch Card	Punch cards
CM:Tape	Open reel (0.5, 0.25, 0.125, etc.); cartridge ; DAT, etc. tape

All Subject Headings start with "S:".

SU LAIR

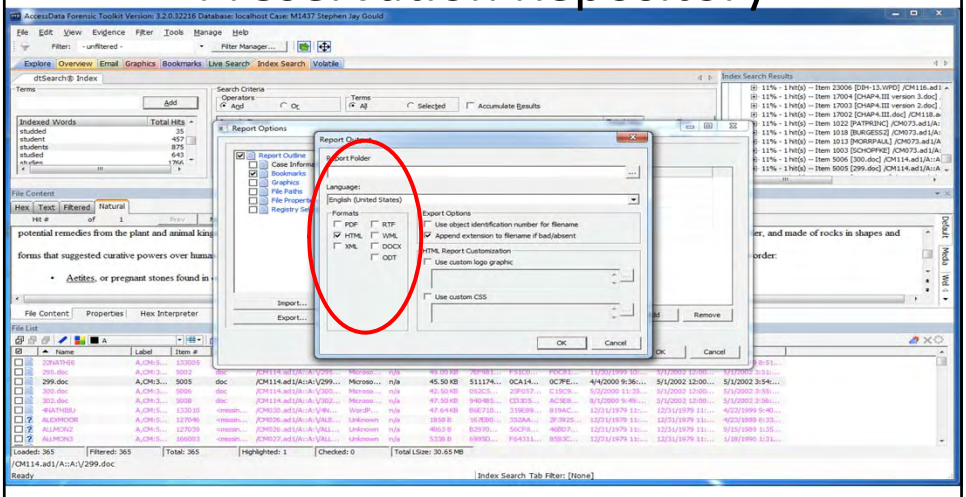


SUOLAIR



SUOLAIR

Interface with Access / Preservation Repository



SUOLAIR

XML-FO

```
<?xml version="1.0" encoding="UTF-8"?>
<fo:root xmlns:xs="http://www.w3.org/2001/XMLSchema" font-family="arial unicode ms" font-size="10pt"
    xmlns:fo="http://www.w3.org/1999/XSL/Format">
    <fo:layout-master-set>
        <fo:simple-page-master master-name="TOC">
            <fo:region-body region-name="body" margin-top="0.5in" margin-bottom="1in"
                margin-left="0.5in" margin-right="0.5in"/>
            <fo:region-before extent="0.5in"/>
            <fo:region-after extent="0.5in"/>
        </fo:simple-page-master>
        <fo:simple-page-master master-name="caseInfoPage">
            <fo:region-body region-name="body" margin-top="0.5in" margin-bottom="1in"
                margin-left="0.5in" margin-right="0.5in"/>
            <fo:region-before extent="0.5in"/>
            <fo:region-after extent="0.5in"/>
        </fo:simple-page-master>
        <fo:simple-page-master master-name="bookmarksPage">
            <fo:region-body region-name="body" margin-top="0.5in" margin-bottom="1in"
                margin-left="0.5in" margin-right="0.5in"/>
            <fo:region-before extent="0.5in"/>
            <fo:region-after extent="0.5in"/>
        </fo:simple-page-master>
        <fo:simple-page-master master-name="index">
            <fo:region-body region-name="body" margin-top="0.5in" margin-bottom="1in"
                margin-left="0.5in" margin-right="0.5in" column-count="2" column-gap="2pc"/>
            <fo:region-before extent="0.5in"/>
            <fo:region-after extent="0.5in"/>
        </fo:simple-page-master>
    </fo:layout-master-set>
    <fo:page-sequence master-reference="TOC">
        <fo:flow flow-name="body">
            <fo:block text-align-last="justify" space-before="12pt"><fo:basic-link color="blue"
                    font-weight="bold" text-decoration="underline"
                    internal-destination="CaseSummaryMainSection" font-size="14pt">Case
                        Summary<fo:leader leader-pattern="space" leader-alignment="reference-area"
                            /><fo:page-number-citation ref-id="CaseSummaryMainSection"
                    /></fo:leader></fo:basic-link></fo:block>
```

SUOLAIR

XSLT Transformation – XML Content

```
<?xml version="1.0" encoding="UTF-8"?>
<ftk_report xmlns:fo="http://www.w3.org/1999/XSL/Format">
    <series>Series 13: Computer Files</series>
    <collection_title>Robert Creeley papers</collection_title>
    <callnumber>M0662</callnumber>
    <did>
        <unittitle>creeley.wmv</unittitle>
        <unitid>209002</unitid>
        <filepath>M0662 CM050.ad1/D::D:\/creeley.wmv</filepath>
        <num>CM050</num>
        <extent>90738025 B</extent>
        <unitdate>4/3/2002 4:06:28 AM (2002-04-03 12:06:28 UTC)</unitdate>
        <accessrestrict>False</accessrestrict>
        <scopecontent>[event]Poetry Reading at Albright-Know Gallery</scopecontent>
        <physdesc>Windows Media Video</physdesc>
        <export_path>files\creeley.wmv</export_path>
    </did>
    <did>
        <unittitle>Arakawa.JPG</unittitle>
        <unitid>212167</unitid>
        <filepath>M0662 CM048.ad1/D::D:\/Photo/arakawa/Arakawa.JPG</filepath>
        <num>CM048</num>
        <extent>195103 B</extent>
        <unitdate>11/29/2000 9:45:44 AM (2000-11-29 17:45:44 UTC)</unitdate>
        <accessrestrict>False</accessrestrict>
        <scopecontent>[subject]Family Photos</scopecontent>
        <physdesc>JPEG EXIF</physdesc>
        <export_path>files\Arakawa.JPG</export_path>
    </did>
```