Using and Developing with Open Source Digital Forensics Software in Digital Archives Programs

Mark A. Matienzo
Manuscripts and Archives, Yale University Library
2012 SAA Research Forum
August 7, 2012

Is open source digital forensics software extensible enough and well-suited to support work in the archival domain?

Digital forensics in the archival domain

- Increasing use of digital forensics tools/methodologies within the context of digital archives programs (Kirschenbaum et al. 2010)
- Technology-focused work (John 2008; Woods & Brown 2009; AIMS Work Group 2012)
- Methodology-focused work (Duranti 2009; Xie 2011)

Significant barriers to use of digital forensics in archives

- Cost (Kirschenbaum et al. 2010; Daigle 2012)
- Complexity (Kirschenbaum et al. 2010; Daigle 2012)
- Digital archives as an emerging market for forensics

Potential of open source digital forensics software

- Requires additional tool development work to be useful for archivists (Kirschenbaum et al. 2010)
- Requires additional integration work (Lee et al. 2012)

Institutional Context

- Focus on implementation of and development with open source digital forensics software at Yale University Library
- Work must support accessioning, processing, and management of born-digital archival material
- Primary focus are records received on legacy media

Design Principles

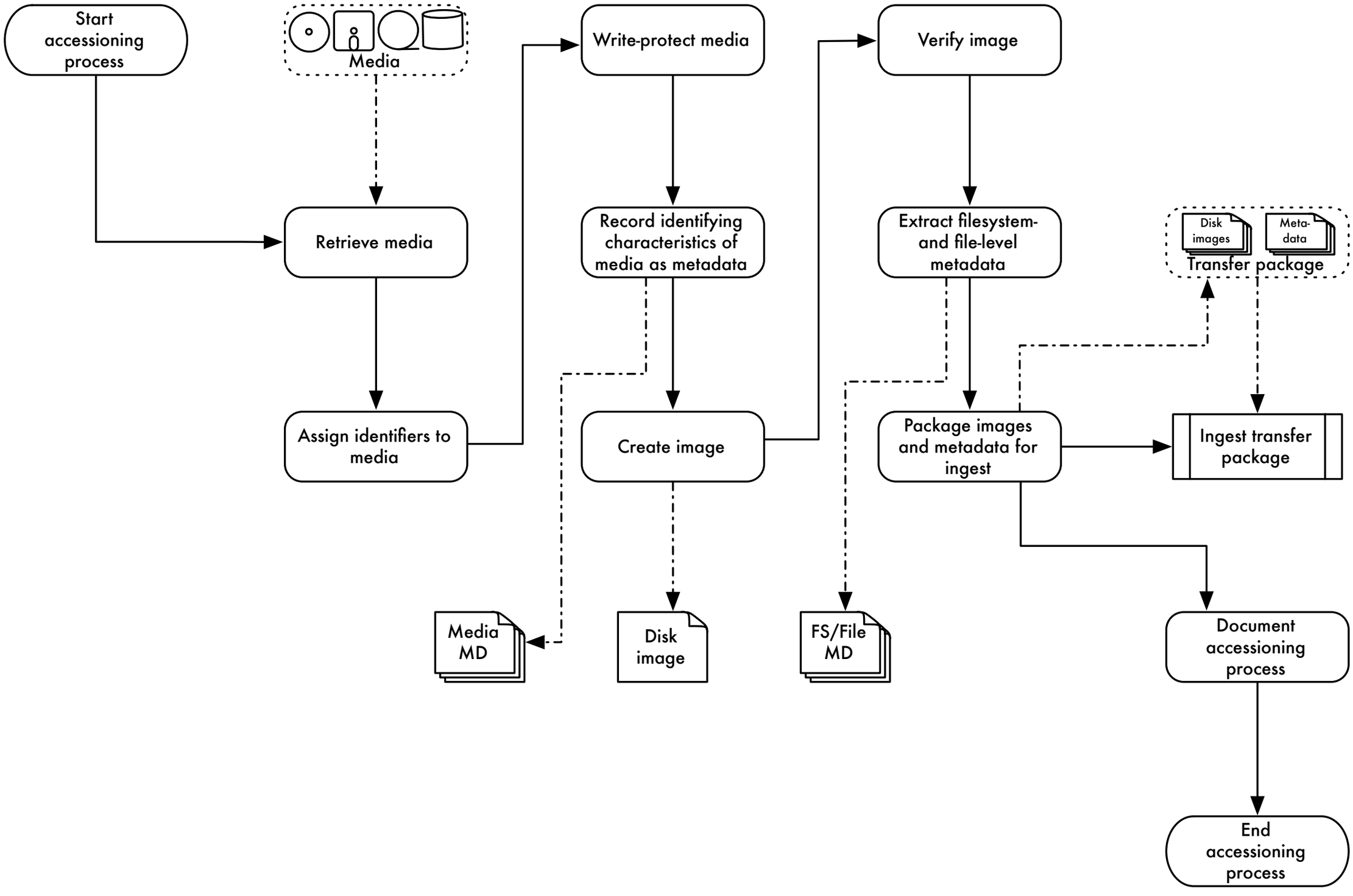
- Use and develop with open source digital forensics software to support accessioning, arrangement, and description of born-digital archival records
- Focus on first two phases (preservation and searching) of Carrier's (2005) model of digital investigation process
- Curation micro-services (Abrams, et al. 2010) as philosophical basis to guide development and implementation
- Recognition of both disk images as digital object (Woods, Lee, and Garfinkel 2011) and objects within disk images as needing management
- Intention of forensic soundness, but assume much of state is lost

Micro-services as Design Philosophy*

Principles	Preferences	Practices
• Granularity	• Small and simple over large and complex	• Define, decompose, recurse
• Orthogonality	• Minimally sufficient over feature-laden	• Top down design, bottom up implementation
• Parsimony	• Configurable over the prescribed	• Code to interfaces
• Evolution	• The proven over the merely novel	• Sufficiency through a series of incrementally necessary steps
	• Outcomes over means	

*UC Curation Center/California Digital Library, 2010

Workflow


Start accessioning process
Media
Write-protect media
Verify image
Retrieve media
Record identifying characteristics of media as metadata
Extract filesystem- and file-level metadata
Disk images
Meta-data
Transfer package
Assign identifiers to media
Create image
Package images and metadata for ingest
Ingest transfer package
Media MD
Disk image
FS/File MD
Document accessioning process
End accessioning process

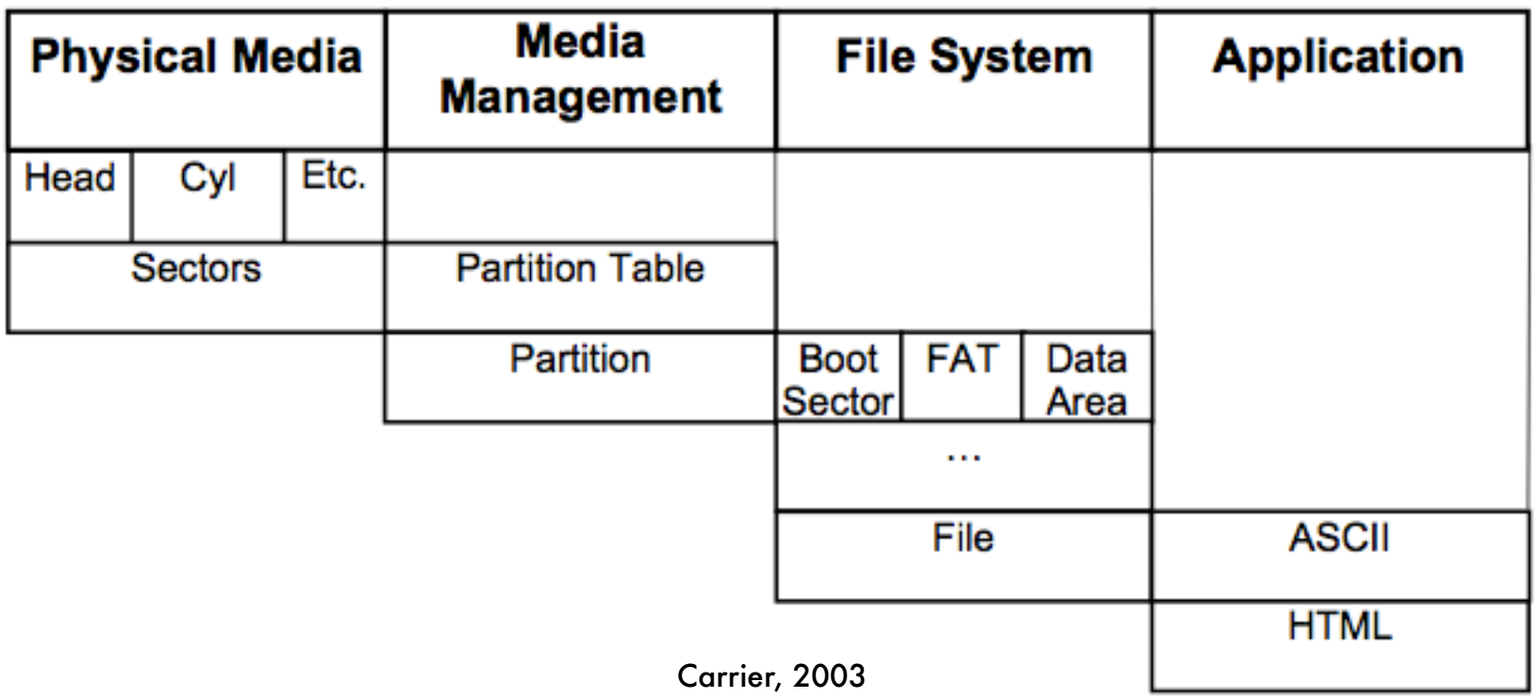
Disk Image Acquisition

- Requires a combination of hardware (drives/media readers, controller cards, write blockers) and software
- In some cases, software depends on particular hardware
- Software tested: FTK Imager (proprietary/gratis), hardware-specific solutions (FC5025 WinDIB; KryoFlux DTC/GUI; Catweasel Imagetool3)
- Goal: sector image interpretable by multiple tools


MSSA100
TABLEAU
T3458is Forensic Bridge
ASUS

Analysis Process

- Multiple levels of analysis within digital forensics based on layers of abstraction (Carrier 2003)
- Conceptual linkages with metadata extraction/analysis processes with digital curation/archival domain



Carrier, 2003

Metadata Extraction

- Use open source digital forensics software (Sleuth Kit, fiwalk) and other open source tools to characterize media, volume, file system, and file information
- Attempt to repurpose this information as descriptive, structural, and/or technical metadata to support accessioning, appraisal, and processing

The Sleuth Kit

- Open source C library, command line tools, and GUI application (Autopsy) for forensic analysis
- Supports analysis of FAT, NTFS, ISO9660, HFS+, Ext2/3, UFS1/2
- Splits tools into layers: volume system, file system, file name, metadata, data unit ("block")
- Additional utilities to sort and post-process extracted metadata

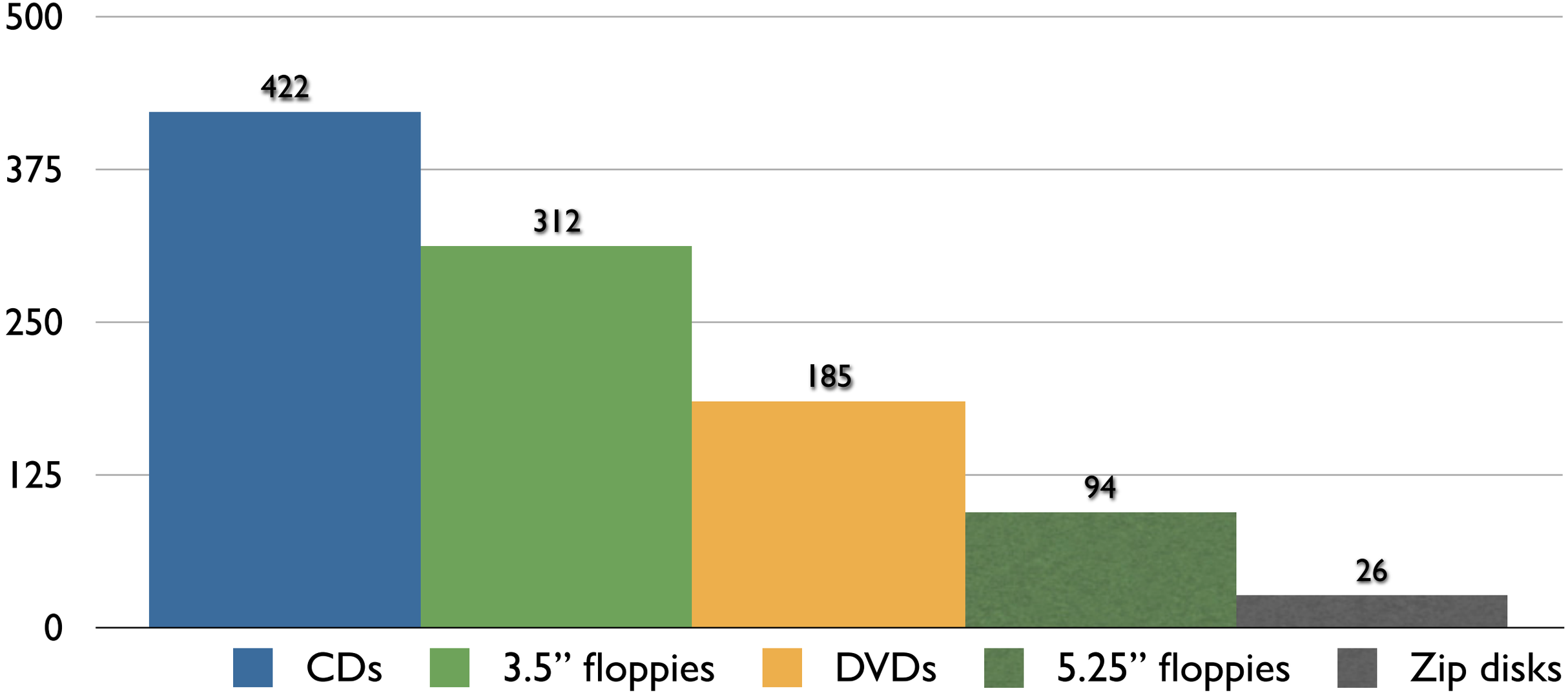
Digital Forensics XML

- Representation in XML of structured forensic information developed by Simson Garfinkel
- Produced by tools including fiwalk (Garfinkel 2012), which uses Sleuth Kit for volume, file system, file, and application-level analysis
- Easily extensible (local plugin development as focus)
- Straight forward to process

Results

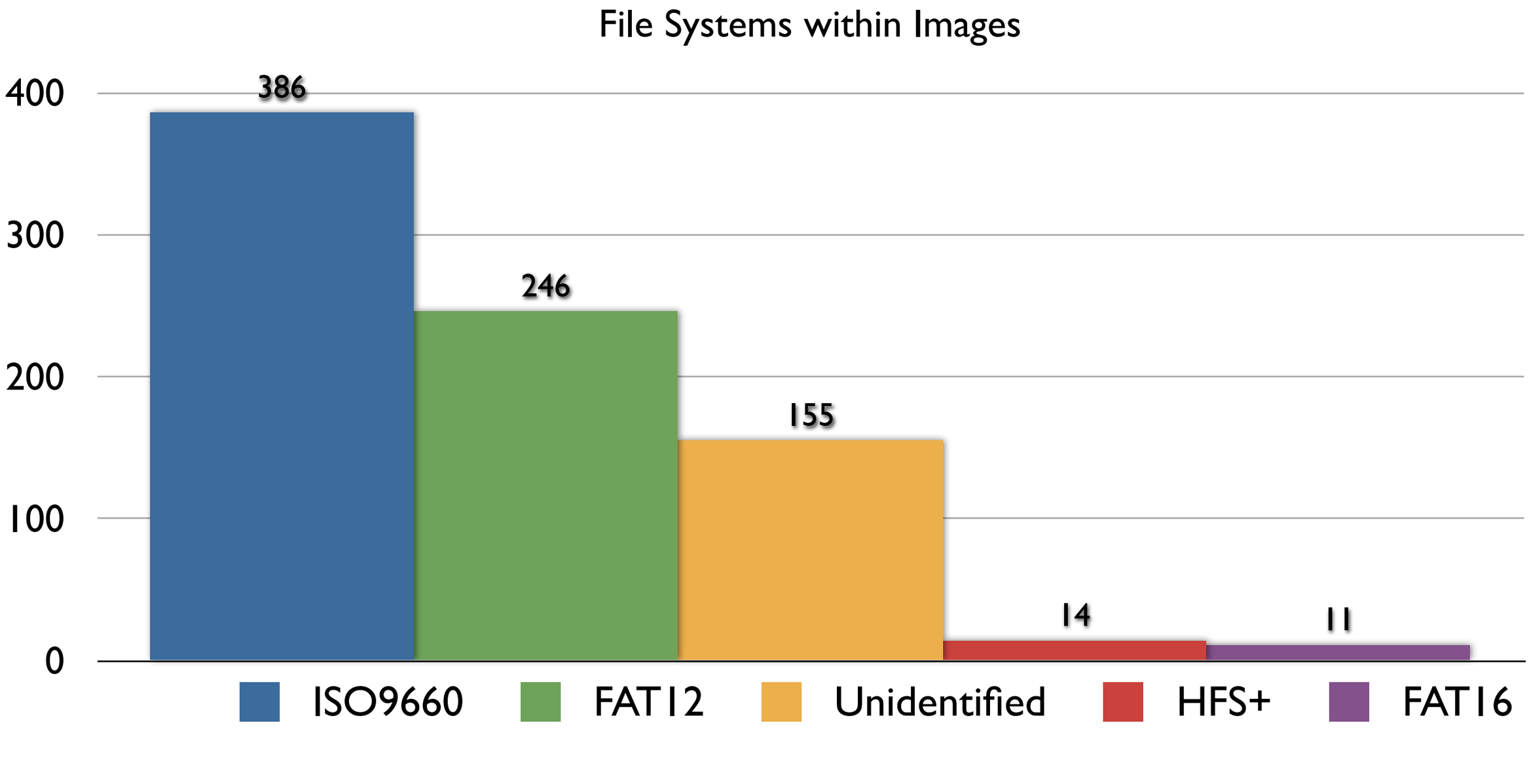
Disk Images

- Acquired 1,039 disk images from across 69 accessions at Manuscripts and Archives



Metadata Extraction

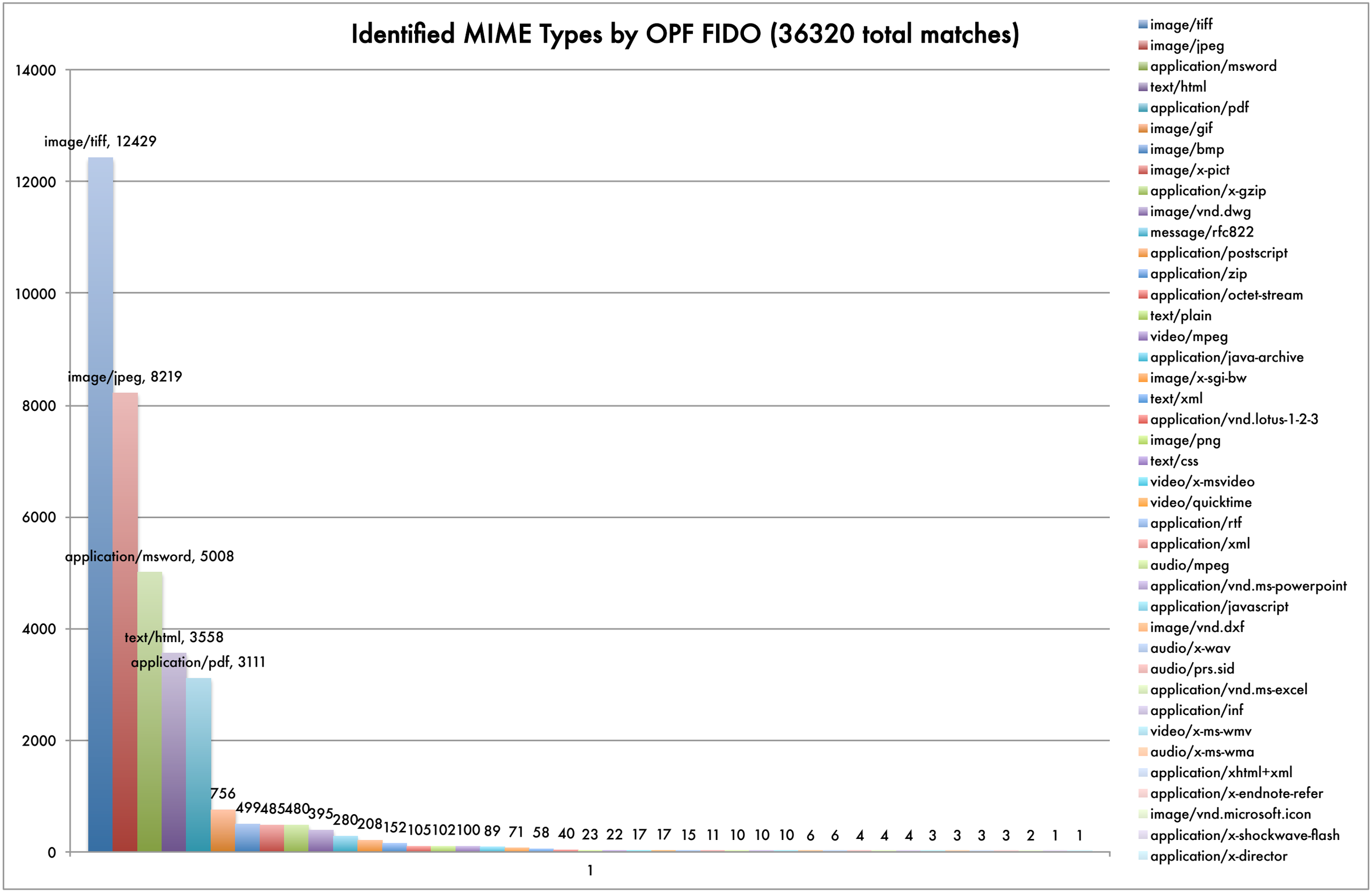
- Ran metadata extraction on 812 images



Metadata Extraction

- Ran enhanced metadata extraction on 619 images (users plugins for fiwalk developed during research)
- Performed analysis on 49,724 files within images
- Successfully identified 43,729 files (147 unique file types) against PRONOM format registry
- Identified 9 files as containing virus signatures (2 unique virus signatures)

Identified MIME Types by OPF FIDO (36320 total matches)



Software Development

- Created Fiwalk plugins to perform additional analysis and evaluation of files/bitstreams within disk images

- Virus identification plugin using ClamAV/pyclamd

- File format identification against PRONOM format registry using Open Planets Foundation's FIDO

- Code (including additional plugins) available online: https://github.com/anarchivist/fiwalk-dgi/

Gumshoe

- Prototype based on Blacklight (Ruby on Rails + Solr)
- Indexing code works with fiwalk output or directly from a disk image
- Populates Solr index with all file-level metadata from fiwalk and, optionally, text strings extracted from files
- Provides searching, sorting and faceting based on metadata extracted from filesystems and files
- Code at http://github.com/anarchivist/gumshoe

Gumshoe

Search History

Fork me on GitHub

Limit your search

Image File

ubnist1_casper_rw_gen2 (1,210)

ntfs1_gen2 (39)

Extension

Format

data (453)

empty (139)

ASCII text (112)

XML document text (58)

JPEG image data, JFIF standard 1.02 (48)

JPEG image data, JFIF standard 1.01 (34)

ASCII English text (29)

GNU dbm 1.x or ndbm database, little endian (26)

HTML document, ASCII text, with very long lines, with CRLF, LF line terminators (22)

PDF document, version 1.4 (22)

more »

Type

Regular file (793)

Directory (381)

Shadow (28)

Symbolic link (24)

Unknown type (22)

Named FIFO (1)

in All Fields Search

Displaying items **1 - 10** of **1,249**

Start over

Sort by size

Show 10 per page

« Previous 1 2 3 4 5 6 7 8 9 ... 124 125 Next »

1. /home/ubuntu/Desktop/MyStuff/SEC Documents/spch121708cc-idata.wmv

Filename	spch121708cc-idata.wmv
Full Path	/home/ubuntu/Desktop/MyStuff/SEC Documents
Image file	ubnist1_casper_rw_gen2
Type	Regular file
Size (bytes)	37887210
Inode number	15697
MD5	8e7d1611c0b870f658529d94556f9a21
Format (libmagic)	Microsoft ASF
Modification Time	2008-12-17T17:10:00Z
Access Time	2008-12-29T05:35:21Z
Change Time	2008-12-29T05:35:21Z

2. /Compressed/logfile1.txt

Filename	logfile1.txt
Full Path	/Compressed
Image file	ntfs1_gen2
Type	Regular file
Size (bytes)	21888890
Inode number	48

Advantages

- Faster (and more forensically sound) to extract metadata once rather than having to keep processing an image
- Possibility of developing better assessments during accessioning process (significance of directory structure, accuracy of timestamps)
- Integrating additional extraction processes and building supplemental tools is simple
- Performance of tools correlates to complexity of analysis

Limitations

- Use of tools limited to specific types of file systems
- Additional software (particularly to document imaging process) requires additional integration and data normalization
- DFXML is not (currently) a metadata format common within domains of archives/libraries and requires an domain-specific application profile
- Extracted metadata maybe harder to repurpose for descriptive purposes based on level of granularity

Work in Progress

- BitCurator project under development; early release available for testing: http://wiki.bitcurator.net
- The Sleuth Kit and related tools under continuing development (Autopsy, fiwalk, etc.): http://sleuthkit.org
- Additional testing, development integration under work at Yale and NYPL

Thanks!

Mark A. Matienzo
mark.matienzo@yale.edu
http://matienzo.org
@anarchivist

References

- Abrams, S., et al. (2011). "Curation Micro-Services: A Pipeline Metaphor for Repositories." *Journal of Digital Information* 12(2). http://journals.tdl.org/jodi/article/view/1605
- AIMS Work Group (2012). *AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship*. http://www2.lib.virginia.edu/aims/whitepaper/
- Carrier, B. (2003). "Defining Digital Forensic Examination and Analysis Tools Using Abstraction Layers." *International Journal of Digital Evidence* 1(4).
- Carrier, B. (2005). *File System Forensic Analysis*. Boston and London: Addison Wesley.
- Daigle, B.J. (2012). "The Digital Transformation of Special Collections." *Journal of Library Administration* 52(3-4), 244-264.
- Duranti, L. (2009). "From Digital Diplomatics to Digital Records Forensics." *Archivaria* 68, 39-66.
- Garfinkel, S. (2012). "Digital Forensics XML and the DFXML Toolset." *Digital Investigation* 8, 161-174.
- John, J.L. (2008). "Adapting Existing Technologies for Digitally Archiving Personal Lives: Digital Forensics, Ancestral Computing, and Evolutionary Perspectives and Tools." Presented at iPRES 2008. http://www.bl.uk/ipres2008/presentations_day1/09_John.pdf
- Kirschenbaum, M.G., et al. (2010). *Digital Forensics and Born-Digital Content in Cultural Heritage Collections*. Washington: Council on Library and Information Resources.
- Lee, C.A., et al. (2012). "BitCurator: Tools and Techniques for Digital Forensics in Collecting Institutions." *D-Lib Magazine* 18(5/6).
- UC Curation Center/California Digital Library (2019). "UC3 Curation Foundations." Revision 0.13. https://confluence.ucop.edu/download/attachments/13860983/UC3-Foundations-latest.pdf
- Woods, K. and Brown, G. (2009). "From Imaging to Access: Effective Preservation of Legacy Removable Media." In *Archiving 2009*. Springfield, VA: Society for Imaging Science and Technology.
- Woods, K., Lee, C.A., and Garfinkel, S. (2011). "Extending Digital Repository Architectures to Support Disk Image Preservation and Access." In *JCDL '11*.
- Xie, S.L. (2011). "Building Foundations for Digital Records Forensics: A Comparative Study of the Concept of Reproduction in Digital Records Management and Digital Forensics." *American Archivist* 74(2), 576-599.

Sleuth Kit example

```
$ fsstat -t 2004-M-088.0007.dd
fat12
```

Sleuth Kit example

```
$ fsstat -t 2004-M-088.0007.dd
fat12

$ fls -a -m A: 2004-M-088.0007.dd
0|A:/DRURY|3|r/rrwxrwxrwx|0|0|1281|1284955200|871048826|0|0
0|A:/BEARD.897|4|r/rrwxrwxrwx|0|0|2392|1284955200|871054862|0|0
0|A:/_P}WP{2 (deleted)|5|r/rrwxrwxrwx|0|0|2392|0|871054894|0|0
0|A:/$MBR|45779|v/v---------|0|0|512|0|0|0|0
0|A:/$FAT1|45780|v/v---------|0|0|4608|0|0|0|0
0|A:/$FAT2|45781|v/v---------|0|0|4608|0|0|0|0
0|A:/$OrphanFiles|45782|d/d---------|0|0|0|0|0|0|0
```

Sleuth Kit example

```
$ fsstat -t 2004-M-088.0007.dd
fat12

$ fls -a -m A: 2004-M-088.0007.dd
0|A:/DRURY|3|r/rrwxrwxrwx|0|0|1281|1284955200|871048826|0|0
0|A:/BEARD.897|4|r/rrwxrwxrwx|0|0|2392|1284955200|871054862|0|0
0|A:/_P}WP{2 (deleted)|5|r/rrwxrwxrwx|0|0|2392|0|871054894|0|0
0|A:/$MBR|45779|v/v---------|0|0|512|0|0|0|0
0|A:/$FAT1|45780|v/v---------|0|0|4608|0|0|0|0
0|A:/$FAT2|45781|v/v---------|0|0|4608|0|0|0|0
0|A:/$OrphanFiles|45782|d/d---------|0|0|0|0|0|0|0

$ icat 2004-M-088.0007.dd 4 | file -
/dev/stdin: (Corel/WP)
```

Sleuth Kit example

```
$ fsstat -t 2004-M-088.0007.dd
fat12

$ fls -a -m A: 2004-M-088.0007.dd
0|A:/DRURY|3|r/rrwxrwxrwx|0|0|1281|1284955200|871048826|0|0
0|A:/BEARD.897|4|r/rrwxrwxrwx|0|0|2392|1284955200|871054862|0|0
0|A:/_P}WP{2 (deleted)|5|r/rrwxrwxrwx|0|0|2392|0|871054894|0|0
0|A:/$MBR|45779|v/v---------|0|0|512|0|0|0|0
0|A:/$FAT1|45780|v/v---------|0|0|4608|0|0|0|0
0|A:/$FAT2|45781|v/v---------|0|0|4608|0|0|0|0
0|A:/$OrphanFiles|45782|d/d---------|0|0|0|0|0|0|0

$ icat 2004-M-088.0007.dd 4 | file -
/dev/stdin: (Corel/WP)

$ icat 2004-M-088.0007.dd 4 | strings | head -n 6
WPCN
Courier 10cpi
HP LaserJet+
HPLASERJ.PRS
Cowles Foundation for Research in Economics
Yale University
```

Sleuth Kit example

```
$ fsstat -t 2004-M-088.0007.dd
fat12

$ fls -a -m A: 2004-M-088.0007.dd
0|A:/DRURY|3|r/rrwxrwxrwx|0|0|1281|1284955200|871048826|0|0
0|A:/BEARD.897|4|r/rrwxrwxrwx|0|0|2392|1284955200|871054862|0|0
0|A:/_P}WP{2 (deleted)|5|r/rrwxrwxrwx|0|0|2392|0|871054894|0|0
0|A:/$MBR|45779|v/v---------|0|0|512|0|0|0|0
0|A:/$FAT1|45780|v/v---------|0|0|4608|0|0|0|0
0|A:/$FAT2|45781|v/v---------|0|0|4608|0|0|0|0
0|A:/$OrphanFiles|45782|d/d---------|0|0|0|0|0|0|0

$ icat 2004-M-088.0007.dd 4 | file -
/dev/stdin: (Corel/WP)

$ icat 2004-M-088.0007.dd 4 | strings | head -n 6
WPCN
Courier 10cpi
HP LaserJet+
HPLASERJ.PRS
Cowles Foundation for Research in Economics
Yale University

$ tsk_recover -a 2004-M-088.0007.dd /tmp
Files Recovered: 2
```

Sample DFXML Output

```
<?xml version='1.0' encoding='UTF-8'?>
<dfxml version='1.0'>
  <metadata
  xmlns='http://www.forensicswiki.org/wiki/Category:Digital_Forensics_XML'
  xmlns:xsi='http://www.w3.org/2001/XMLSchema-instance'
  xmlns:dc='http://purl.org/dc/elements/1.1/'>
    <dc:type>Disk Image</dc:type>
  </metadata>
  <creator version='1.0'>
    <!-- provenance information re: extraction - software used; operating system -->
  </creator>
  <source>
    <image_filename>2004-M-088.0018.dd</image_filename>
  </source>
  <volume offset='0'><!-- partitions within each disk image -->
      <fileobject><!-- files within each partition --></fileobject>
  </volume>
  <runstats><!-- performance and other statistics --></runstats>
</dfxml>
```

Sample DFXML Output

```
<fileobject>
  <filename>_ublist1.wpd</filename>
  <partition>1</partition>
  <id>1</id>
  <name_type>r</name_type>
  <filesize>202152</filesize>
  <unalloc>1</unalloc>
  <used>1</used>
  <inode>3</inode>
  <meta_type>1</meta_type>
  <mode>511</mode>
  <nlink>0</nlink>
  <uid>0</uid>
  <gid>0</gid>
  <mtime>2001-02-22T22:30:52Z</mtime>
  <atime>2001-02-22T05:00:00Z</atime>
  <crtime>2001-02-22T22:31:54Z</crtime>
  <libmagic>(Corel/WP)</libmagic>
  <byte_runs>
   <byte_run file_offset='0' fs_offset='16896' img_offset='16896' len='512'/>
  </byte_runs>
  <hashdigest type='md5'>d7bc22242c0a88fd8b68712980d5ab28</hashdigest>
  <hashdigest type='sha1'>64bf2bdf82e33fcda50158804483ac611e753db5</hashdigest>
</fileobject>
```