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

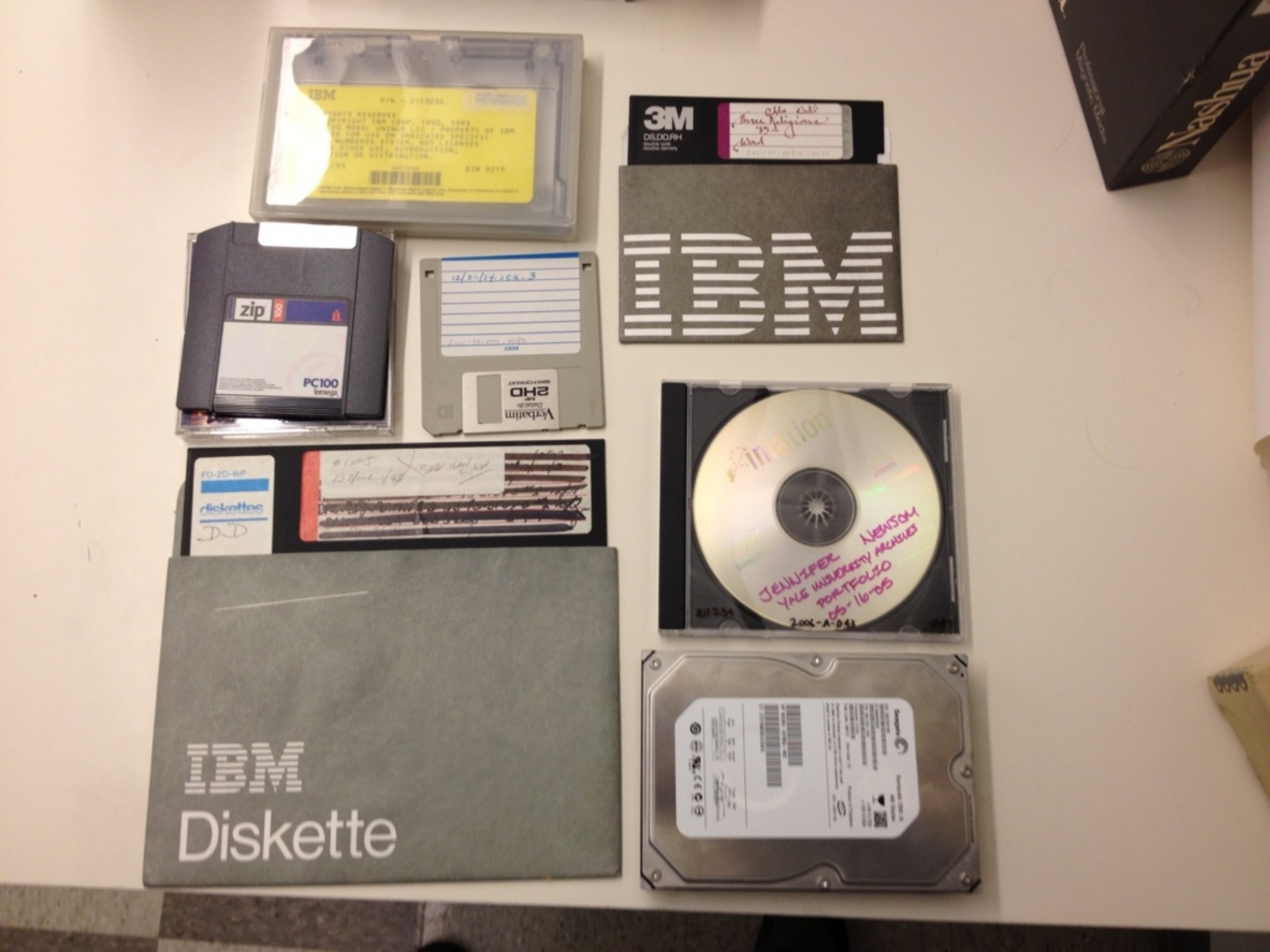
Mark A. Matienzo, Yale University Library

Open Source Digital Forensics Conference
Chantilly, VA
October 3, 2012

About Me

- I am an archivist
- Occasionally I develop software
- I am not a digital forensics “expert”

Digital Archives at Yale


IBM
P/N 21F3036
COPYRIGHT IBM CORP. 1990, 1993
3M
DS,DD,RH
double side
double density
Three Religions
IBM
zip 100
PC100
Iomega
12/31/14 sca. 3
Verbatim
DataLife
MF
2HD
FD-2D-WP
diskettes
DD
IBM
Diskette
imation
JENNIFER NEWSOM
YALE UNIVERSITY ARCHIVES
PORTFOLIO
05-16-05
RU 254
2006-A-041
Seagate
Nashua

Digital Forensics in the Archival Domain

- Increasing use of digital forensics tools/methodologies within the context of digital archives programs (Kirschenbaum et al. 2010)
- Barriers to adoption: cost, complexity, need for additional tool development (Kirschenbaum et al. 2010; Daigle 2012; Lee et al. 2012)
- BitCurator project: http://bitcurator.net

Initial Goals

- Focus on implementation of and development with open source digital forensics software at Yale University Library
- Work must support accessioning, arrangement, description, and management of born-digital archival material
- Material received on physical media as primary focus

Design Principles

- Digital objects needing management are both disk images themselves (Woods, Lee, and Garfinkel 2011) and bitstreams that they contain

- Intention of forensic soundness, but assumption that much of the state is lost

- Curation micro-services (Abrams, et al. 2010) as philosophical basis to guide our thinking

Micro-services as Design Philosophy*

Principles

- Granularity
- Orthogonality
- Parsimony
- Evolution

Preferences

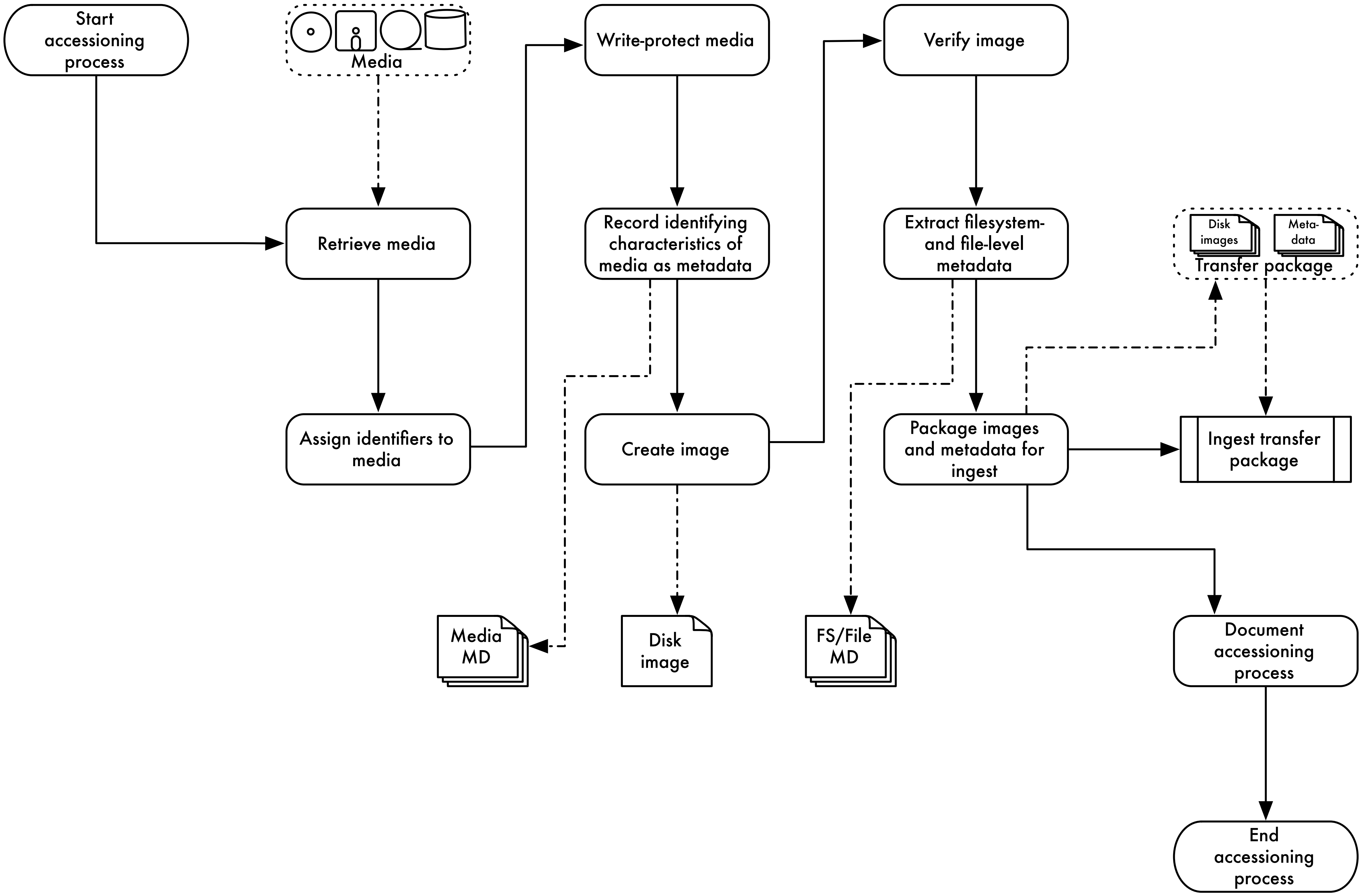
- Small and simple over large and complex
- Minimally sufficient over feature-laden
- Configurable over the prescribed
- The proven over the merely novel
- Outcomes over means

Practices

- Define, decompose, recurse
- Top down design, bottom up implementation
- Code to interfaces
- Sufficiency through a series of incrementally necessary steps

*UC Curation Center/California Digital Library, 2010

Workflow



Start accessioning process

Media

Retrieve media

Assign identifiers to media

Write-protect media

Record identifying characteristics of media as metadata

Create image

Verify image

Extract filesystem- and file-level metadata

Package images and metadata for ingest

Ingest transfer package

Document accessioning process

End accessioning process

Media MD

Disk image

FS/File MD

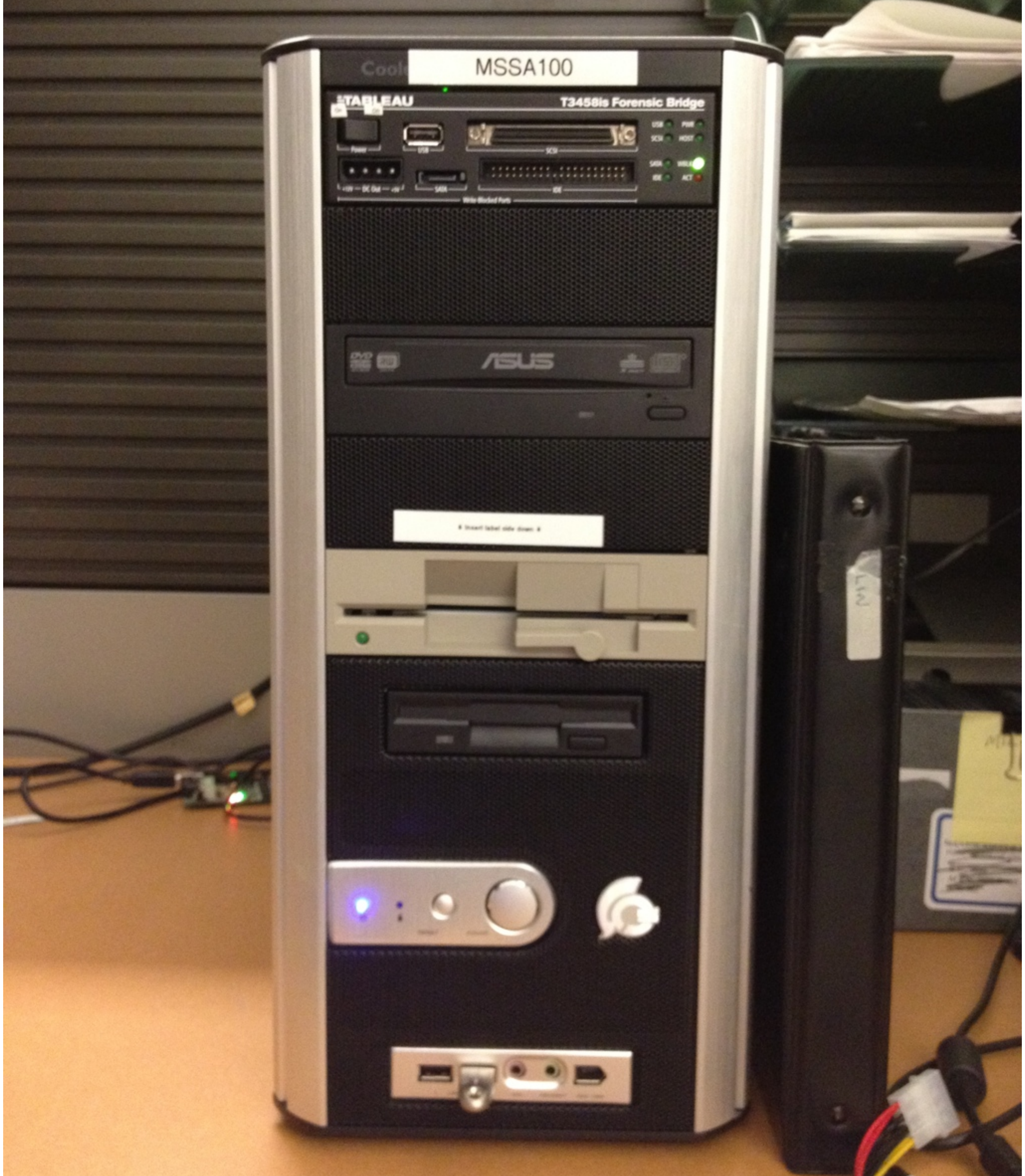
Disk images

Meta-data

Transfer package

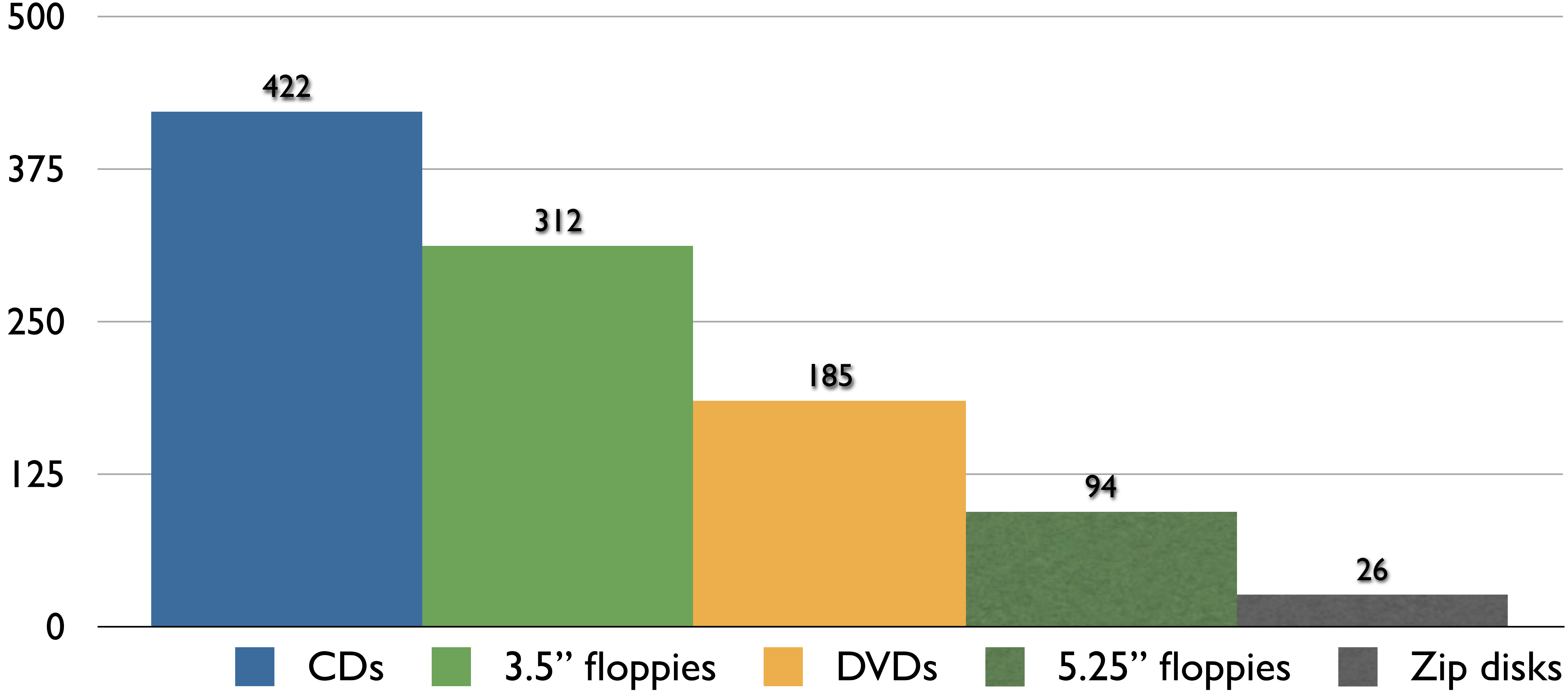
Disk Image Acquisition

- Requires a combination of hardware (drives/media readers, controller cards, write blockers) and software
- In some cases, hardware requires specific software (e.g. floppy disk controller cards that sample magnetic flux transitions)
- Goal: sector image interpretable by multiple tools


MSSA100
TABLEAU
T3458is Forensic Bridge
Power
USB
SCSI
DC Out
SATA
IDE
Write Blocked Ports
USB
PWR
SCSI
HOST
SATA
WBLK
IDE
ACT
ASUS
# Insert label side down #

Disk Images

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



Initial Work with Disk Images

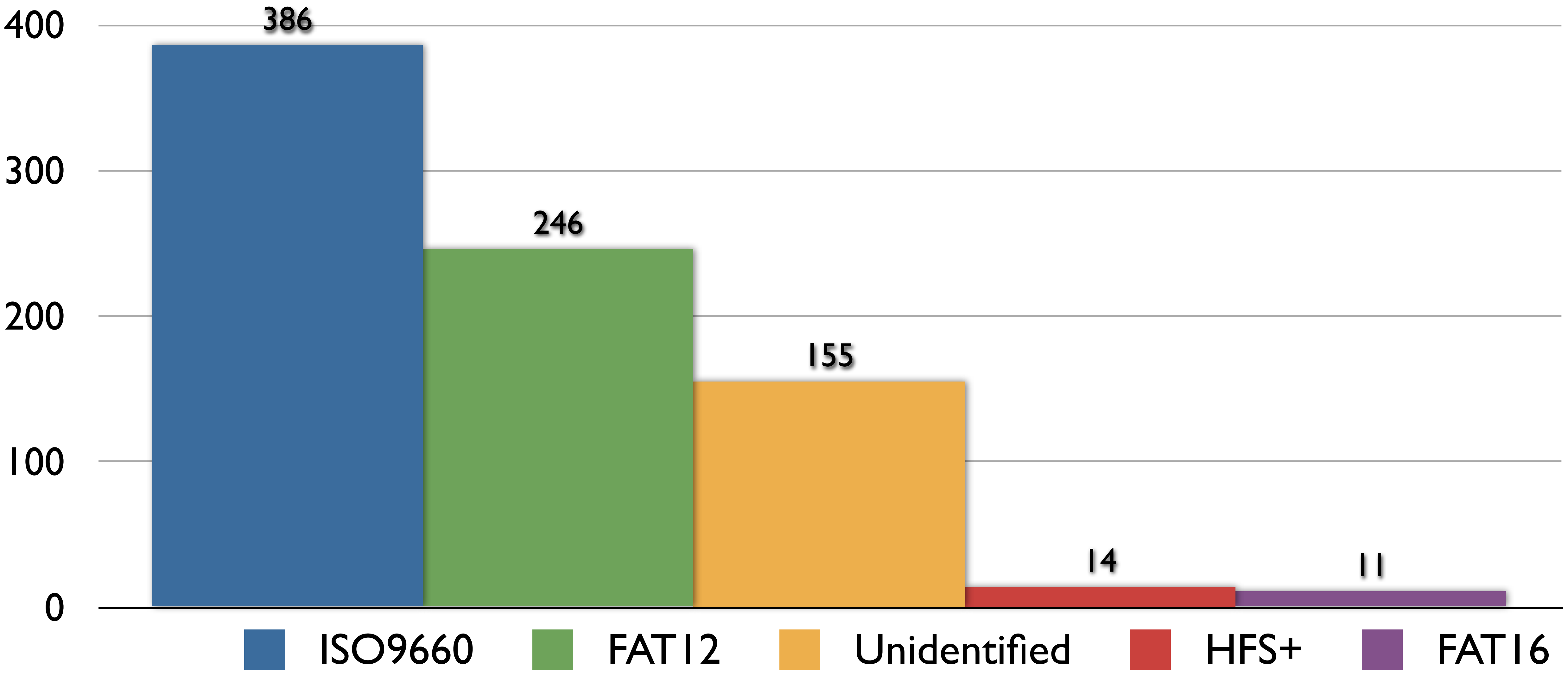
- Experimentation with various tools: The Sleuth Kit (3.1+), Autopsy, Pyflag, bulk_extractor ...
- Basic integration/processing with shell scripts or Python
- Discovering fiwalk was my “eureka” moment

Metadata Extraction

- Used 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
- Extracted metadata expressed in Digital Forensics XML
- Easily extensible and straightforward to process

File Systems

- Ran metadata extraction on 812 images



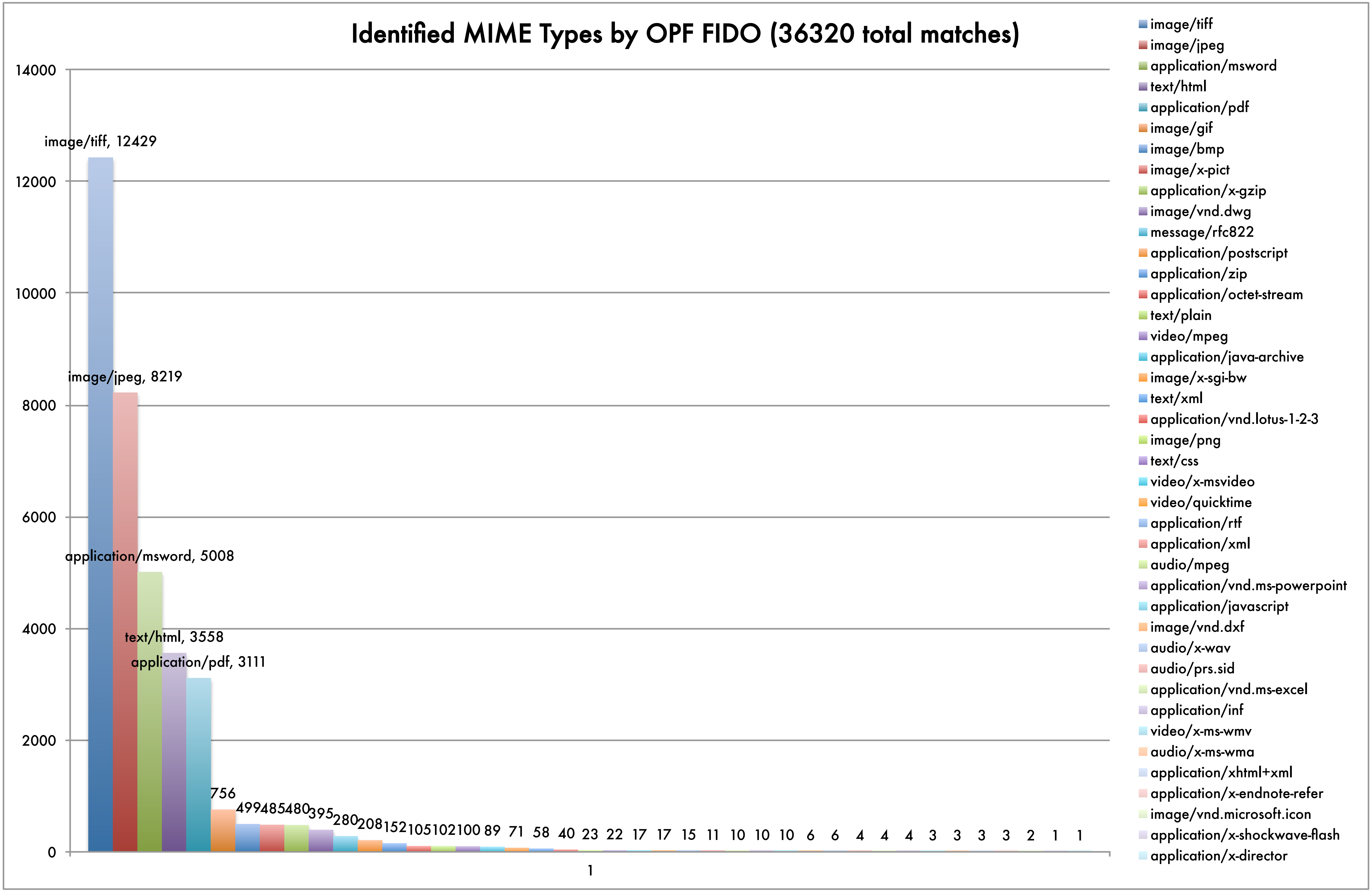
Extraction Plugins

- 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/

File Analysis

- Ran enhanced metadata extraction on 619 images (using our plugins)
- 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)



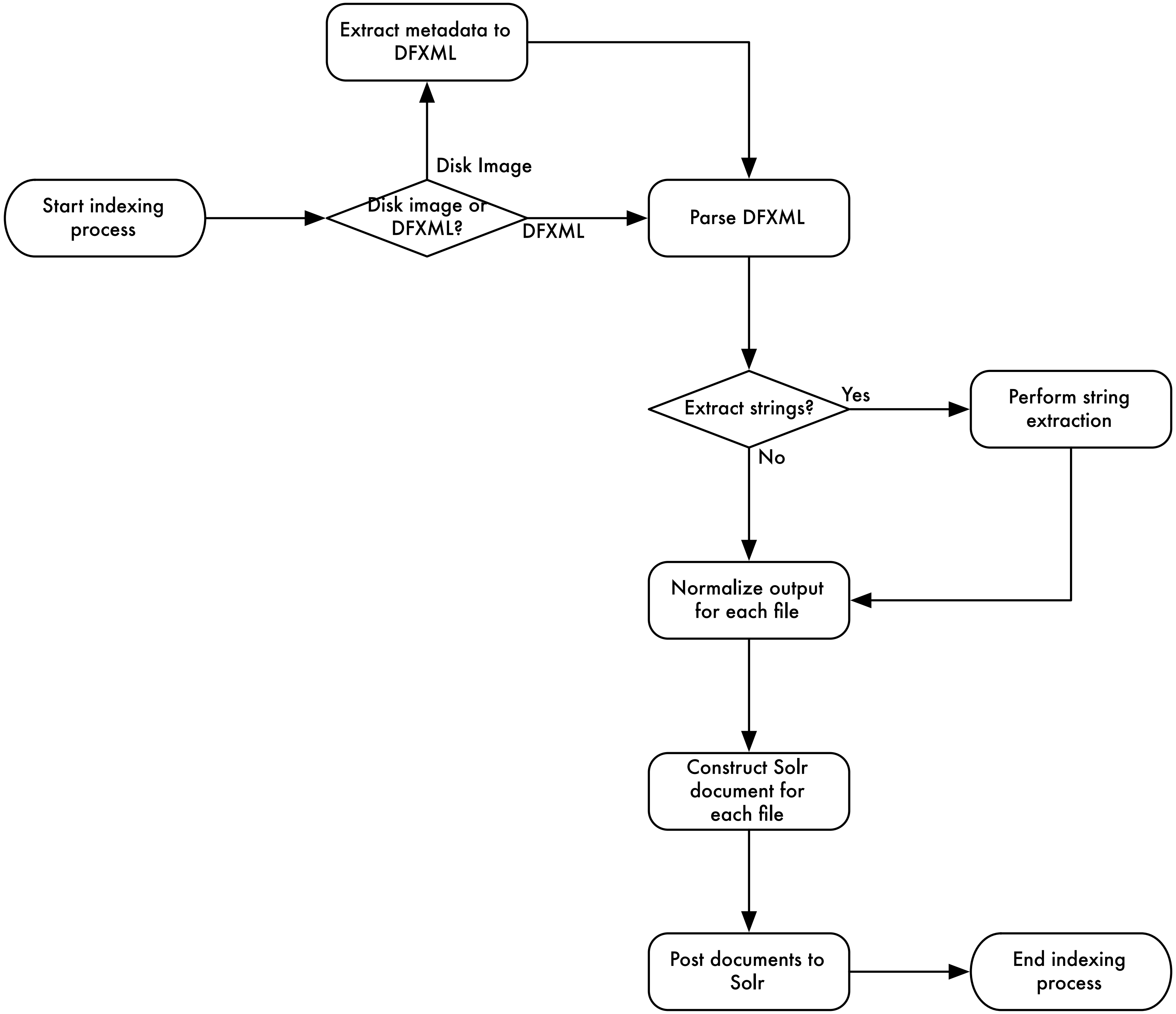
Gumshoe

- Prototype web application to provide search/browse interface to metadata extracted from disk images
- Built as a Ruby on Rails application using Blacklight
- http://github.com/anarchivist/gumshoe

Blacklight

- http://projectblacklight.org
- Ruby gem for use in Rails applications
- Provides discovery layer over Solr indexes, with support for faceting, bookmarking, etc.
- Use is fairly common in library community
- Implementers include Stanford, Columbia, NC State, UVA, WGBH, National Agricultural Library (AGNIC) ...

Indexing Process

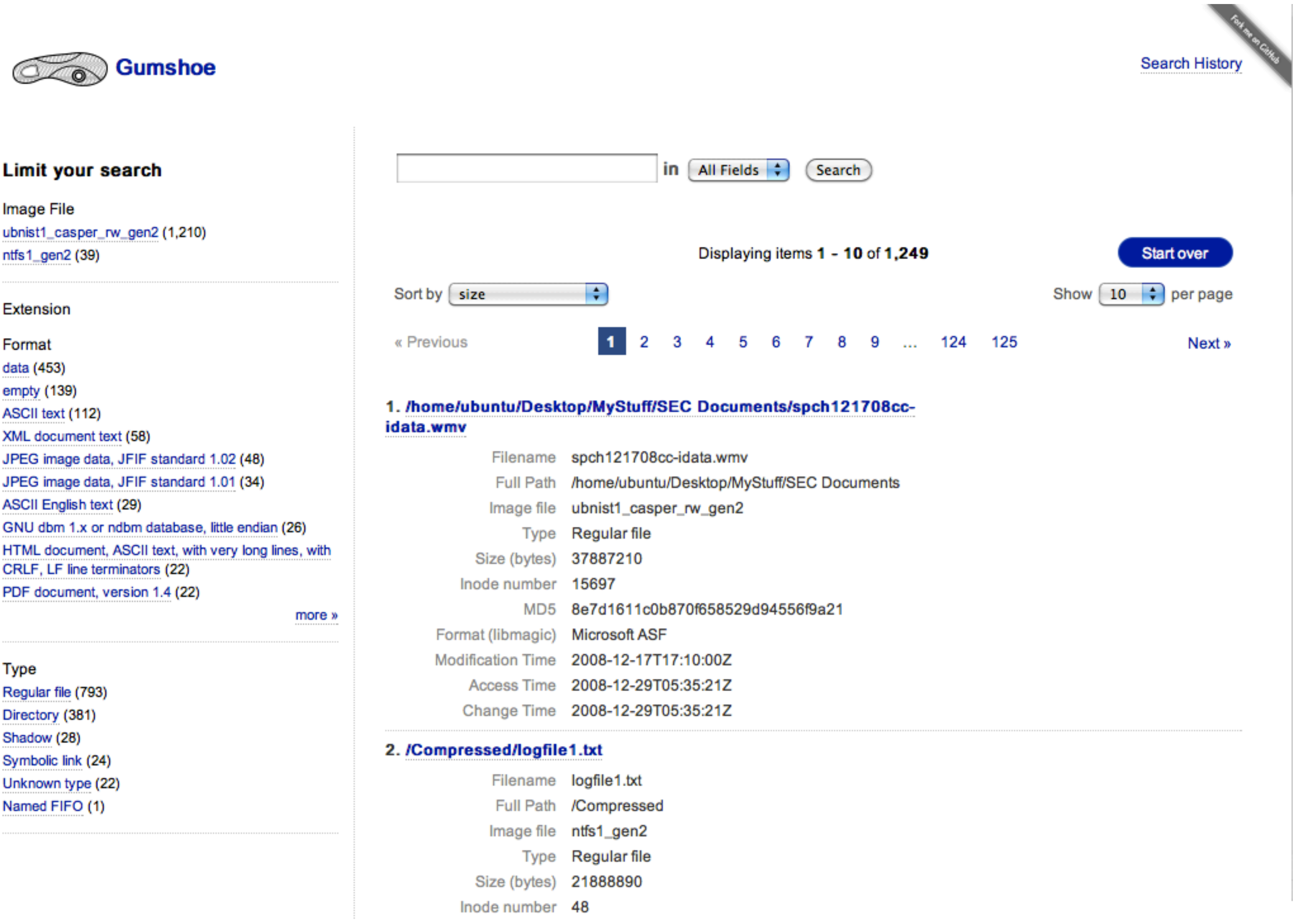


Data Normalization

- Depends on DFXML gem
- Translate metadata-layer data to more easily searchable or human-readable version (e.g. file type/file system codes to text labels; certain flags to booleans)
- Data type conversion (e.g. integers-as-strings to integers)
- Prepend full path data to filename
- Transform timestamps to ISO8601

Features

- Basic browse view, with sorting by filename, size, modification/access/creation times
- Faceting by disk image, extension, file format, file type
- Basic bookmarking
- Searching based on metadata values (e.g. checksums), file content (still under development; somewhat slow)



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)

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

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

Limitations

- Use of tools limited to specific types of file systems
- Requires additional integration and data normalization to work with additional tools
- DFXML is not (currently) a metadata format common within domains of archives/libraries; somewhat in flux
- Extracted metadata harder for archivists to repurpose in some cases based on level of granularity
- Still struggling with how to best present data to archivists

BitCurator

- http://bitcurator.net
- Currently under development; preview releases available
- Provides unified environment (VM) with tools for disk imaging, data triage, PII identification, metadata extraction, etc.
- Uses familiar tools: Sleuth Kit, Guymager, fiwalk, sdhash ...

Thanks!

Mark A. Matienzo
mark@matienzo.org
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*.