Digital Forensics for Digital Archives

Mark A. Matienzo
Manuscripts and Archives, Yale University Library

Digital Preservation 2012
Arlington, VA
July 25, 2012

Digital Archives at Yale

http://www.flickr.com/photos/johnjack/3449280414

Digital Forensics

Digital Forensics

+

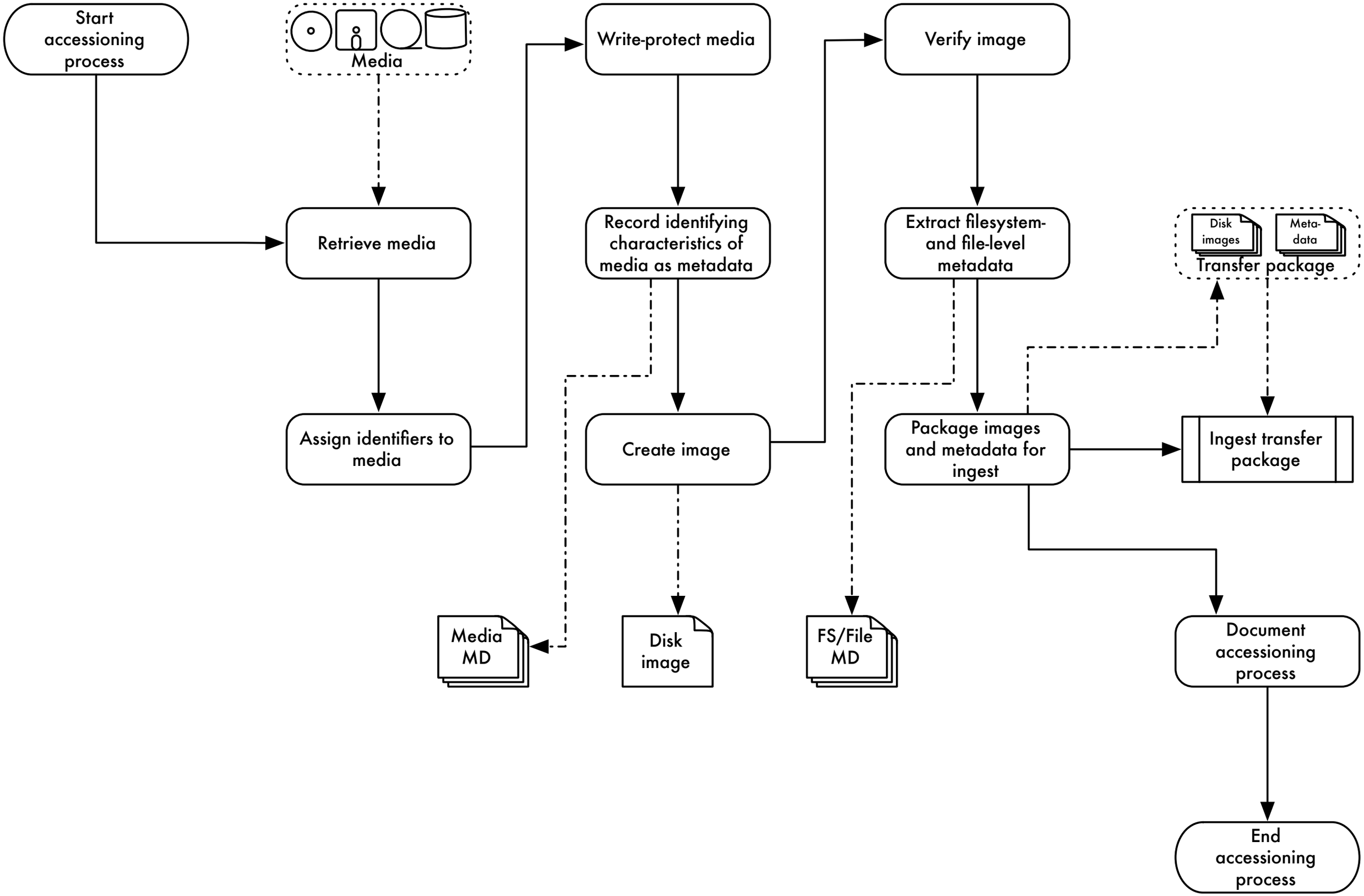
Digital Archives

Design Principles

- Use digital forensics software and methodology to support accessioning, arrangement, and description of born-digital archival records
- Mitigate risk of media deterioration and obsolescence
- Prefer open source solutions whenever possible
- Integrate into a larger, but yet-to-be-defined workflow

Applied Methodology

- Use Carrier's (2005) model of the digital investigation process: Preservation ↔ Searching ↔ Reconstruction
- Volume and file system as main areas for analysis
- Assume much of the state is already lost
- Methods should approach or intend forensic soundness


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
Disk images
Meta-data
Transfer package
Ingest transfer package
Media MD
Disk image
FS/File MD
Document accessioning process
End accessioning process

Documenting Media

- SharePoint-based list
- Unique identifiers for each piece
- Allows basic documentation of imaging process

Electronic Records on Media Accessioning Log

New | Actions | Settings | View: All Ite

Type	Media number	Media Format	Imaging Date	Imaging Successful?	Bag Created?	Metadata Extracted?	Transfer to Storage Date	Examiner	Image format	Imaging Software	Sou
	2011-M-075.0001	CD-R		No	No	No		Glick, Kevin	N/A	N/A	FAT
	2011-M-075.0002	DVD-R		Yes	No	Yes		Glick, Kevin	ISO	ImgBurn	ISO
	2011-M-075.0003	DVD-R		Yes	No	Yes		Glick, Kevin	ISO	ImgBurn	ISO (1.0
	2011-M-075.0004	DVD-R		Yes	No	Yes		Glick, Kevin	ISO	ImgBurn	ISO (1.0
	2011-M-075.0005	DVD-R		Yes	No	Yes		Glick, Kevin	ISO	ImgBurn	ISO (1.0
	2011-M-075.0006	DVD-R		Yes	No	Yes		Glick, Kevin	ISO	ImgBurn	ISO (1.0
	2011-M-075.0007	CD-R		Yes	No	No		Glick, Kevin	ISO	ImgBurn	ISO
	2011-M-075.0008	CD-R		Yes	No	No		Glick, Kevin	ISO	ImgBurn	ISO
	2011-M-075.0009	CD-R		Yes	No	Yes		Glick, Kevin	ISO	ImgBurn	ISO (1.0
	2011-M-075.0010	DVD-R		Yes	No	Yes		Glick, Kevin	ISO	ImgBurn	ISO (1.0
	2011-M-075.0011	CD-R		Yes	No	Yes		Glick, Kevin	ISO	ImgBurn	ISO
	2011-M-075.0012	CD-R		Yes	No	Yes		Glick, Kevin	ISO	ImgBurn	ISO
	2011-M-075.0013	Zip disk		Yes	No	Yes		Glick, Kevin	dd (Raw)	FTK Imager 3.0.0.1443	FAT

Electronic Records on Media Accessioning Log

New | Actions | Settings

Type	Media number	Media Format	Imaging Date	Imaging Successful?	Bag Create
	2011-M-075.0001	CD-R		No	No
	2011-M-075.0002	DVD-R		Yes	No
	2011-M-075.0003	DVD-R		Yes	No
	2011-M-075.0004	DVD-R		Yes	No
	2011-M-075.0005	DVD-R		Yes	No
	2011-M-075.0006	DVD-R		Yes	No
	2011-M-075.0007	CD-R		Yes	No
	2011-M-075.0008	CD-R		Yes	No
	2011-M-075.0009	CD-R		Yes	No
	2011-M-075.0010	DVD-R		Yes	No
	2011-M-075.0011	CD-R		Yes	No
	2011-M-075.0012	CD-R		Yes	No
	2011-M-075.0013	Zip disk		Yes	No

Electronic Records on Media Accessioning Log: 2011-M-075.0008

Close

New Item | Edit Item | Delete Item | Manage Permissions | Alert Me

Media number	2011-M-075.0008
Media Format	CD-R
Media Density (floppies only)	N/A
Interface	N/A
Label text	Osaka Monograph Final Images Aug 29 2003 Monograph Latest Files
Manufacturer	
Serial Number (hard drives only)	
Examiner	Glick, Kevin
Imaging Successful?	Yes
Imaging Date	
Image filename	2011-M-075.0008.ISO
Source File System	ISO9660, Joliet
Image format	ISO
Imaging Software	ImgBurn
Image Fixity Function	MD5
Image Fixity Value	dbca43c94690edff07329b6687550f60
Notes	mam54 04/28/2011: Could not extract metadata using fiwalk; log file from imaging process says that the block structure is Mode 2/Form 1
Metadata Extracted?	No
Bag Created?	No
Transfer to Storage Date	
Fiscal Year	2010-11

Created at 4/27/2011 9:35 AM by Glick, Kevin
Last modified at 4/28/2011 4:26 PM by Matienzo, Mark

Close

Imaging Media

- Requires a combination of hardware and software
- In some cases, software depends on particular hardware
- No single universal solution for our workflow

Imaging Hardware

- Drives (eg floppy drives, flash cardreaders)
- Interface cards (Catweasel, Kryoflux, FC5025)
- Writeblockers


MSSA100
TABLEAU
T3458is Forensic Bridge
ASUS

Imaging Software

- FTK Imager (proprietary; gratis)
- Hardware-specific imaging software for floppy interface cards
- Other software tested: dd, Guymager, etc.

AccessData FTK Imager 2.9.0.1385

File View Mode Help

Evidence Tree

- 2004-M-008.dd-0005.001
  - C910927A001 [FAT12]
    - [root]
    - [unallocated space]

Custom Content Sources

Evidence:File System|Path|File	Options

New | Edit | Remove | Remove All | Create Image

Properties | Hex Value Inter... | Custom Content...

File List

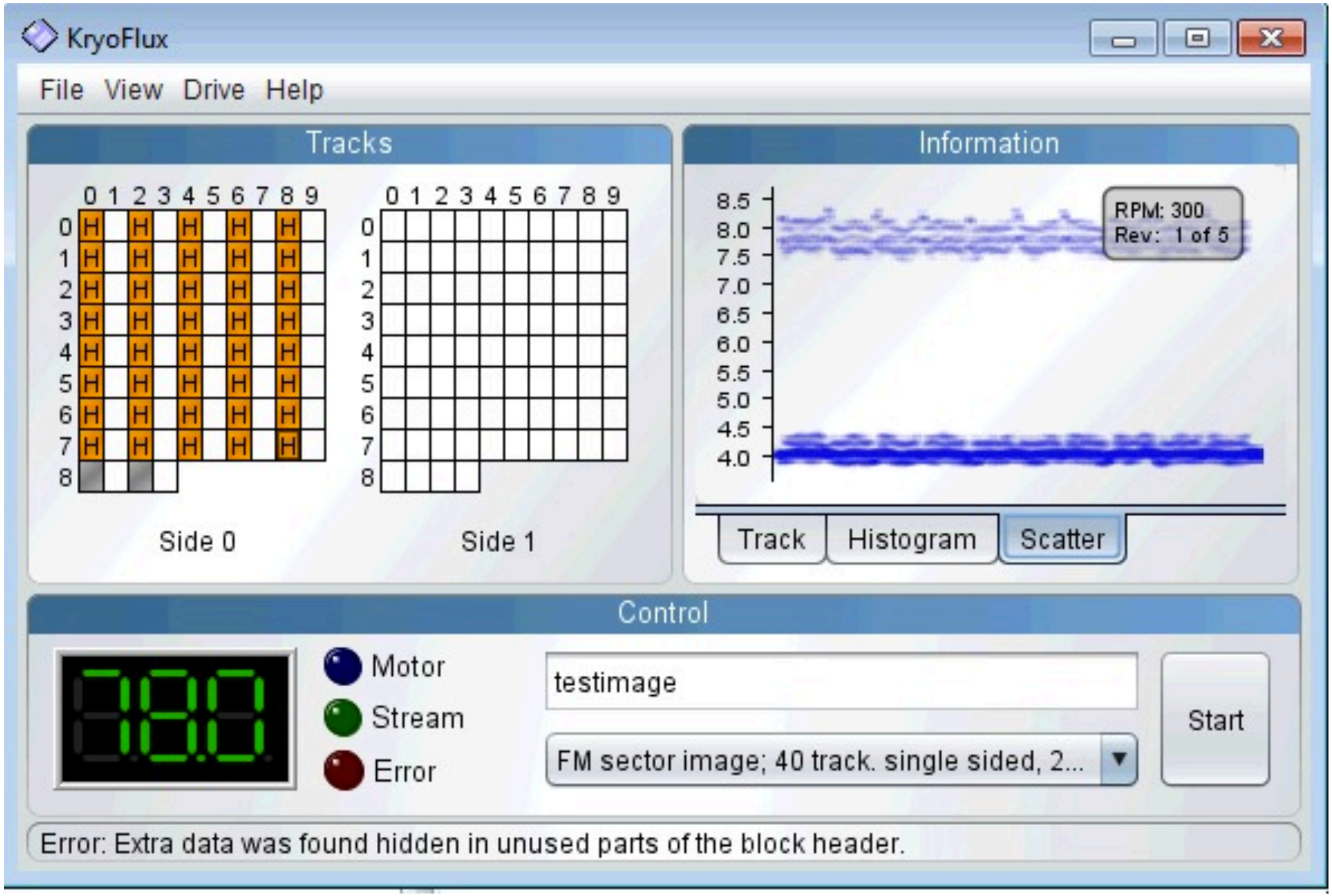
Name	Size	Type	Date Modified
!w0000	59 KB	Regular File	7/21/1997 7:58...
Hangman.bak	56 KB	Regular File	6/10/1997 5:23...
Hangman.bak	58 KB	Regular File	6/10/1997 5:36...
Hangman.bak	59 KB	Regular File	7/21/1997 7:57...
Hangman.tex	0 KB	Regular File	6/10/1997 5:23...
Hangman.tex	59 KB	Regular File	7/21/1997 7:58...
KENNEDY.WPD	41 KB	Regular File	10/17/1997 11:...
Q3.DIR	1 KB	Regular File	7/5/1999 11:18...
Q3.DIR.FileSlack	1 KB	File Slack	
QDATA.ABD	10 KB	Regular File	7/5/1999 11:12...

```
0000 43 39 31 30 39 32 37 41-30 30 31 28 00 00 00 00 C910927A001(····
0010 00 00 00 00 00 00 4C 77-1A 21 00 00 00 00 00 00 ·····Lw·!·····
0020 51 44 41 54 41 20 20 20-51 44 46 20 00 43 48 5A QDATA   QDF ·CHZ
0030 E5 26 31 3D 00 00 48 5A-E5 26 02 00 C0 7D 04 00 å&1=··HZå&··À}··
0040 51 44 41 54 41 20 20 20-51 53 44 20 00 B4 4C 5A QDATA   QSD ·´LZ
0050 E5 26 31 3D 00 00 48 5A-E5 26 92 02 60 49 00 00 å&1=··HZå&··`I··
0060 51 44 41 54 41 20 20 20-51 45 4C 20 00 66 4D 5A QDATA   QEL ·fMZ
0070 E5 26 31 3D 00 00 65 B7-B6 26 B7 02 00 3C 00 00 å&1=··e·¶&···<··
0080 51 44 41 54 41 20 20 20-41 42 44 20 00 16 4E 5A QDATA   ABD ··NZ
0090 E5 26 31 3D 00 00 8B 59-E5 26 D5 02 97 26 00 00 å&1=···Yå&Õ··&··
00a0 51 33 20 20 20 20 20 20-44 49 52 20 00 88 4E 5A Q3      DIR ··NZ
00b0 E5 26 31 3D 00 00 4E 5A-E5 26 E9 02 17 00 00 00 å&1=··NZå&é·····
00c0 E5 41 4E 47 4D 41 4E 20-42 4B 21 20 00 13 F9 85 åANGMAN BK! ··ù·
00d0 CA 22 CA 22 00 00 FA 85-CA 22 23 00 77 F7 00 00 Ê"Ê"··ú·Ê"#·w÷··
00e0 E5 41 4E 47 4D 41 4E 20-42 4B 21 20 00 2F 41 89 åANGMAN BK! ·/A·
00f0 CA 22 CA 22 00 00 44 89-CA 22 02 00 91 F7 00 00 Ê"Ê"··D·Ê"···÷··
0100 E5 41 4E 47 4D 41 4E 20-54 58 54 20 00 6B 65 89 åANGMAN TXT ·ke·
0110 CA 22 CA 22 00 00 66 89-CA 22 03 00 41 F9 00 00 Ê"Ê"··f·Ê"··Aù··
0120 E5 48 00 61 00 6E 00 67-00 6D 00 0F 00 B6 61 00 åH·a·n·g·m···¶a·
0130 6E 00 2E 00 74 00 65 00-78 00 00 00 00 00 FF FF n·.·t·e·x·····ÿÿ
0140 E5 41 4E 47 4D 41 4E 20-54 45 58 20 00 A0 F6 8A åANGMAN TEX · ö·
0150 CA 22 CA 22 00 00 F6 8A-CA 22 00 00 00 00 00 00 Ê"Ê"··ö·Ê"·····
0160 E5 57 30 30 30 30 20 20-20 20 20 20 08 1C F7 8A åW0000      ··÷·
0170 CA 22 CA 22 00 00 F9 8A-CA 22 04 00 8C DD 00 00 Ê"Ê"··ù·Ê"···Ý··
0180 E5 48 00 61 00 6E 00 67-00 6D 00 0F 00 22 61 00 åH·a·n·g·m···"a·
```

Cursor pos = 0; log sec = 19

2004-M-008.dd-0005.001/C910927A001 [FAT12]/[root]

NUM


KryoFlux
File View Drive Help
Tracks
0 1 2 3 4 5 6 7 8 9
Side 0
Side 1
Information
RPM: 300
Rev: 1 of 5
8.5
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
Track
Histogram
Scatter
Control
Motor
Stream
Error
testimage
FM sector image; 40 track. single sided, 2...
Start
Error: Extra data was found hidden in unused parts of the block header.

Metadata Extraction

- Desire to repurpose existing information as archival description and reports to other staff
- Ideal output is XML; can be packaged with disk images going into medium- or long-term storage
- Tools: Fiwalk/The Sleuth Kit; FTK Imager; testing others
- Have integrated file-format identification (using OPF's FIDO) and virus/malware recognition (using ClamAV) using Fiwalk's plugin architecture

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>
```

Analysis/Processing

- Once acquired, we can perform additional analysis or reporting to captured assets or records
- Few tools are easily useable by archivists (BitCurator toolset under development will help)
- Additional forensic tools can be used for archival arrangement and description of this information

Forensic Toolkit

- Proprietary application to analyze files, filesystems, etc.
- Provides full-text indexing, tagging, bookmarking, file presentation/viewing, and reporting
- Used at Yale, Stanford, and other institutions for archival processing of born-digital records
- Still a challenge to use given the complexity of the application

AccessData Forensic Toolkit Version: 3.4.1.34295 Database: MSSA-BRBL FTK Case: Sample Test Case (NDIIPP)

File Edit View Evidence Filter Tools Manage Help

Filter: - unfiltered - Filter Manager...

Explore Overview Email Graphics Bookmarks Live Search Index Search Volatile

Evidence Items

- Evidence
  - ntfs1-gen2.raw
    - NTFS1 [NTFS]
      - [orphan]
      - [root]
        - $BadClus
        - $Extend
        - $Secure
        - Compressed
        - Encrypted
        - RAW
        - System Volume Inf
      - [unallocated space]

File Content

Hex Text Filtered Natural

NIST Special Publication 800-88

Guidelines for Media Sanitization

Recommendations of the National Institute of Standards and Technology

Richard Kissel
Matthew Scholl
Steven Skolochenko
Xing Li

Default Media Web

File Content Properties Hex Interpreter

File List

Normal Display Time Zone: Eastern Daylight Time (From local machine)

Name	Label	Item #	Ext	Path	Category	P-Size	L-Size	MD5	SHA1	SHA256	Created	Accessed
$I30		1022		ntfs1-gen2.raw/NTFS1 ...	Index ...	4096 B	4096 B	7519D...	4F4FF9...	8A7D3...	12/31/2008 5:4...	1/5/2009 5:00:..
20076517123273.pdf		1023	pdf	ntfs1-gen2.raw/NTFS1 ...	Adobe ...	984.0 KB	978.2 KB	2E1678...	597646...	F33B00...	12/31/2008 6:1...	12/31/2008 6:1..
logfile1.txt		1024	txt	ntfs1-gen2.raw/NTFS1 ...	Text	20.88 MB	20.87 MB	BE2828...	4A97F...	450AC...	1/5/2009 5:00:...	1/5/2009 5:01:..
NIST_logo.jpg		1026	jpg	ntfs1-gen2.raw/NTFS1 ...	JPEG	8192 B	2205 B	D651F...	0687C...	9422D...	12/31/2008 6:0...	1/5/2009 4:41:..
NISTSP800-88_rev1.pdf		1025	pdf	ntfs1-gen2.raw/NTFS1 ...	Adobe ...	544.0 KB	541.1 KB	E91036...	E0B99F...	D4038...	12/31/2008 6:0...	12/31/2008 6:0..
report02-3.pdf		1027	pdf	ntfs1-gen2.raw/NTFS1 ...	Adobe ...	1392 KB	1388 KB	DEDE9...	3C078...	B2E4E...	12/31/2008 6:0...	12/31/2008 6:0..

Loaded: 6 | Filtered: 6 | Total: 6 | Highlighted: 1 | Checked: 0 | Total LSize: 23.72 MB

ntfs1-gen2.raw/NTFS1 [NTFS]/[root]/Compressed/NISTSP800-88_rev1.pdf

Ready | Explore Tab Filter: [None]

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
- Develop better assessments during accessioning process (directory structure significant? timestamps accurate?)
- Integrating additional extraction processes and building supplemental tools takes less time

Limitations

- Use of tools limited to specific types of filesystems
- Additional software requires additional integration and data normalization
- DFXML is not (currently) a metadata format common within domains of archives/libraries
- 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 development (Autopsy, fiwalk, etc.): http://sleuthkit.org
- Additional testing and integration under work at Yale, using DFXML as common schema whenever possible
- Possible development of a new media log to record media/imaging metadata and workflow status

Thanks!

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