Automated Forensics Research at NPS

Simson L. Garfinkel
Associate Professor, Naval Postgraduate School

May 12, 2010

https://domex.nps.edu/deep/

NPS is the Navy's Research University.

Location: Monterey, CA

Campus Size: 627 acres

Students: 1500

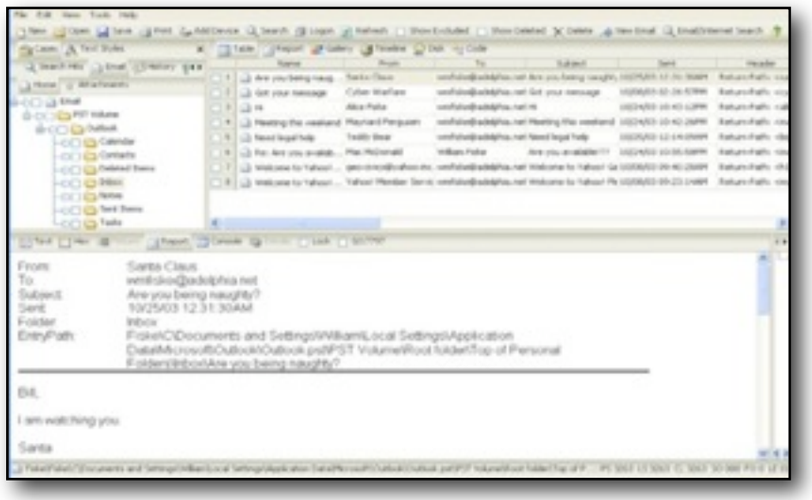
- US Military (All 5 services)
- US Civilian (Scholarship for Service & SMART)
- Foreign Military (30 countries)

Digital Evaluation and Exploitation:

- *Research* computer forensics.
- *Develop* "corpora" for use in research & education.
- *Identify* limitations of current tools & opportunities for improvement.
- http://domex.nps.edu/deep/

"The views expressed in this presentation are those of the author and do not necessarily reflect those of the Department of Defense or the US Government."

Digital Forensics is at a turning point.
Yesterday's work was primarily *reverse engineering*.



Key technical challenges:

- Evidence preservation.
- File recovery (file system support); Undeleting files
- Encryption cracking.
- Keyword search.

Digital Forensics is at a turning point. Today's work is increasingly *scientific.*

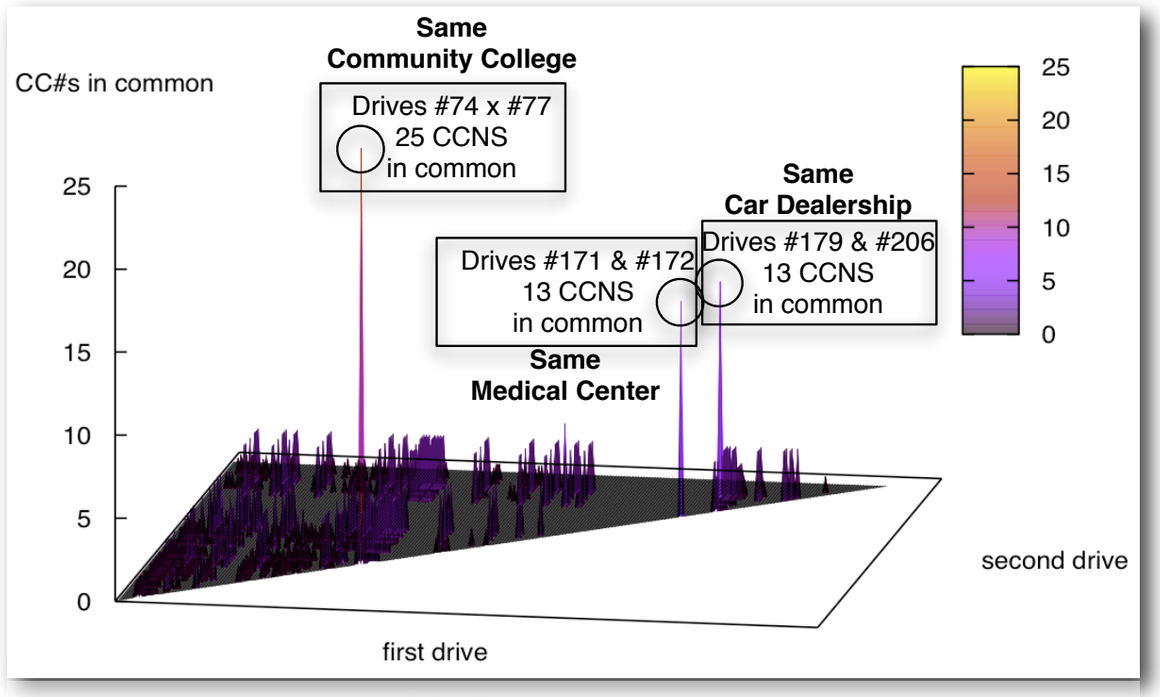
Evidence Reconstruction

- Files (fragment recovery carving)
- Timelines (visualization)

Clustering and data mining

Social network analysis

Sense-making



Science requires the *scientific process.*

Hallmarks of Science:

- Controlled and repeatable experiments.
- No privileged observers.

Why repeat some other scientist's experiment?

- Validate that an algorithm is properly implemented.
- Determine if ***your*** new algorithm is better than ***someone else's*** old one.

We can't do this today.

- Bob's tool can identify 70% of the data in the windows registry.
  - —*He publishes a paper.*
- Alice writes her own tool and can only identify 60%.
  - —*She writes Bob and asks for his data.*
  - —*Bob can't share the data because of copyright & privacy issues.*

To address this problem, we are creating releasable corpora.

This talk discusses four projects of interest to the Digital Archivist Community

Realistic Data

- For research and training
- 1.5 TB available today

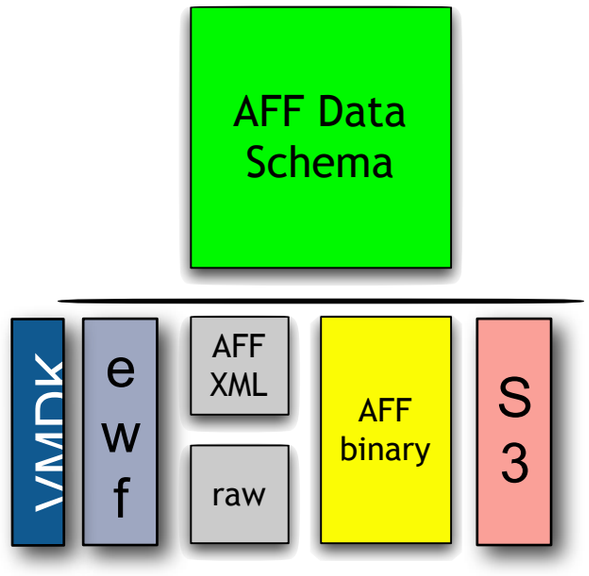
Real Data Corpus

- Lots of PII for research
- 10-20 TB; IRB approval required

bulk_extractor

- Rapid triage

AFF & Digital Forensics XML



Standardized Forensic Corpora

We have created dozens of disk images, packet captures, and memory dumps.

Test Images:

- nps-2009-hfstest1 (HFS+)
- nps-2009-ntfs1 (NTFS)

Realistic Images:

- nps-2009-canon2 (FAT32)
- nps-2009-UBNIST1 (FAT32)
- nps-2009-casper-rw (embedded EXT3)
- nps-2009-domexusers (NTFS)

Scenarios:

- M57 startup — spear phishing attack
- M57 patents — small business victim of internal hacking
- Nitroba University — Harassment case solved through network forensics

http://digitalcorpora.org/

NPS-govdocs1: 1 Million files available *now*

1 million documents downloaded from US Government web servers

- Specifically for file identification, data & metadata extraction.
- Found by random word searches on Google & Yahoo
- DOC, DOCX, HTML, ASCII, SWF, etc.

Free to use; Free to redistribute

- No copyright issues — US Government work is not copyrightable.
- Other files have simply been moved from one USG webserver to another.
- No PII issues — These files were already released.

Distribution format: ZIP files

- 1000 ZIP files with 1000 files each.
- 10 “threads” of 1000 randomly chosen files for student projects.
- Full provenance for every file (how found; when downloaded; SHA1; etc.)

http://domex.nps.edu/corp/files/

The Real Data Corpus: "Real Data from Real People."

Most forensic work is based on “realistic” data created in a lab.

We get real data from CN, IN, IL, MX, and other countries.

Real data provides:

- Real-world experience with data management problems.
- Unpredictable OS, software, & content
- Unanticipated faults

We have multiple corpora:

- Non-US Persons Corpus
- US Persons Corpus (@Harvard)
- Releasable Real Corpus
- Realistic Corpus

Real Data Corpus: Current Status

Corpus	HDs	Flash	CDs	GB
US*	1258			2939
BA	7			38
CA	46	1		420
CN	26	568	98	999
DE	37	1		765
GR	10			6
IL	152	4		964
IN		66		29
MX	156			571
NZ	1			4
TH	1	3		13
* Not available to USG	1694	643	98	6748

Note: IRB Approval is Mandatory!

To use the Real Data corpus, you need IRB approval

US Law requires approval of an Institutional Review Board to work with human subject data [45 CFR 46]

- Clearly describe what you want to do.
- Get submit your protocol to your local IRB.
- Provide your protocol and your approval.
- Approval must be renewed every year.

No approval required for “Realistic” data.

- http://www.digitalcorpora.org/

```
/Users/simsong/domex/src/bulk_extractor/src/r0:
total used in directory 140432 available 37283372
drwxr-xr-x  12 simsong  simsong        408 May 11 22:19 .
drwxr-xr-x  67 simsong  simsong       2278 May 13 21:19 ..
-rw-r--r--   1 simsong  simsong         48 May 11 22:22 _thread0.stat
-rw-r--r--   1 simsong  simsong         48 May 11 22:22 ccn.txt
-rw-r--r--   1 simsong  simsong        102 May 11 22:18 config.cfg
-rw-r--r--   1 simsong  simsong    3184451 May 11 22:22 domain.txt
-rw-r--r--   1 simsong  simsong    4197524 May 11 22:22 email.txt
-rw-r--r--   1 simsong  simsong         63 May 11 22:22 report.txt
-rw-r--r--   1 simsong  simsong        839 May 11 22:22 rfc822.txt
-rw-r--r--   1 simsong  simsong          0 May 11 22:18 tcp.txt
-rw-r--r--   1 simsong  simsong    1086578 May 11 22:22 url.txt
-rw-r--r--   1 simsong  simsong  135305057 May 11 22:22 wordlist.txt
```

bulk_extractor

bulk_extractor finds and reports interesting strings

Currently bulk_extractor finds:

- email addresses, email Subject: lines, dates
- URLs
- Credit Card Numbers
- Other interesting information

bulk_extractor:

- Ignores file systems & file formats
- Is multi_threaded; can process bulk data very fast
- summarizes what it finds in easy to read reports.

Input formats:

- raw (dd) disk images
- Database files
- EnCase E01 files
- AFF files

sample output on nps-2009-ubnist1/ubnist1.gen3.raw

email.txt:

```
68728516        sy@f.Lr
70878552        cristy@mystic.es.dupont.com
262257627       PH@O.Jp
317697633       micke@imendio.com
317697671       alexl@redhat.com
317697704       shaunm@gnome.org
324391670       ari@debian.org
324391814       tugrul@galatali.com
328471989       hp@redhat.com
335671479       mike@markley.org
625382493       QU@X.tn
738975536       astarikovskiy@suse.de
739006965       tiwai@suse.de
739006998       perex@perex.cz
739052440       perex@perex.cz
```

email_histogram.txt:

```
n=27640 ubuntu-users@lists.ubuntu.com
n=17133 ubuntu-motu@lists.ubuntu.com
n=12883 ubuntu-devel-discuss@lists.ubuntu.com
n=4032  language-packs@ubuntu.com
n=1966  ubuntu-desktop@lists.ubuntu.com
n=1484  debian-x@lists.debian.org
n=1006  pkg-perl-maintainers@lists.alioth.debian.org
n=878   debian-qt-kde@lists.debian.org
```

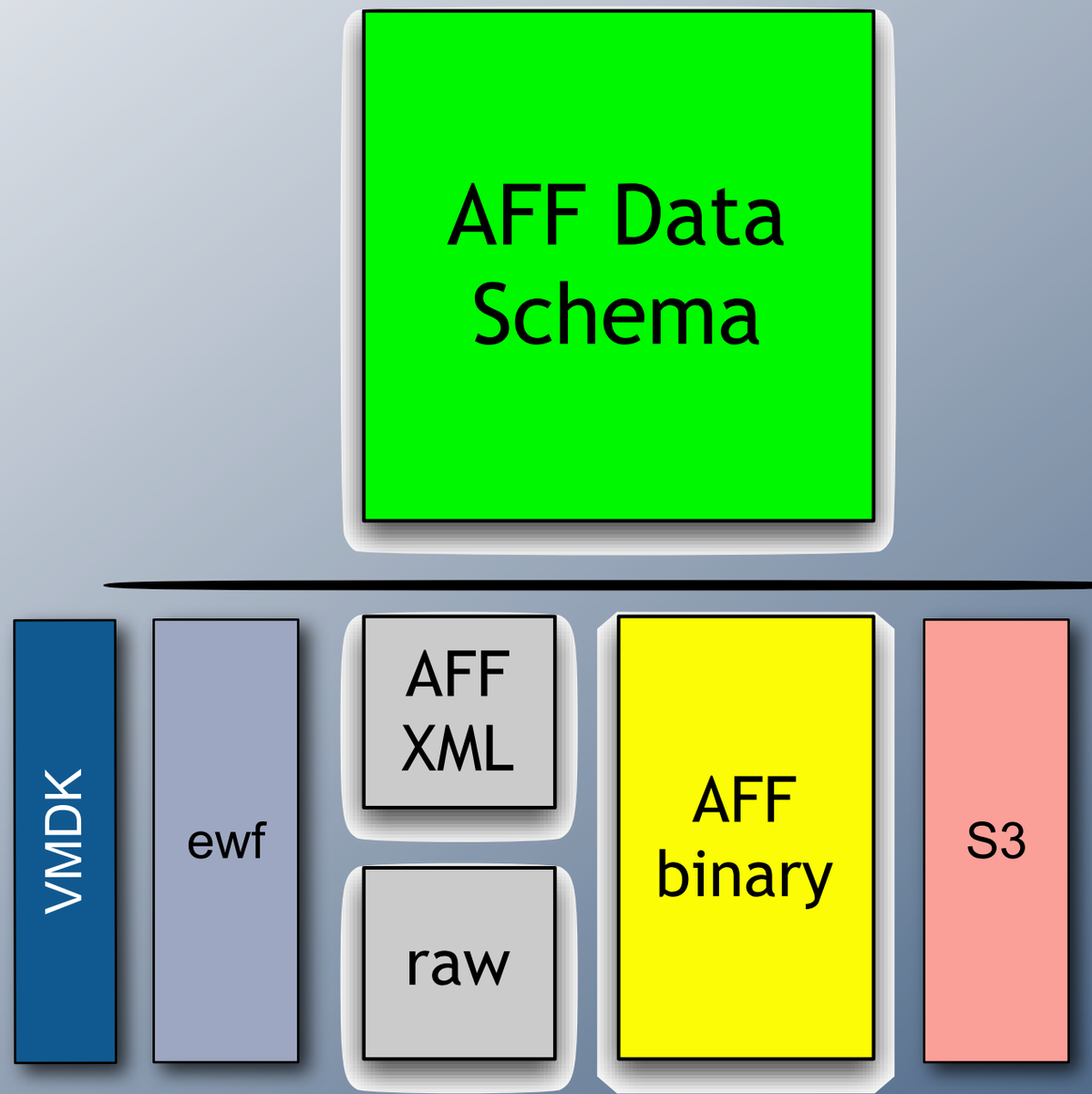
bulk_extractor is a flexible tool

Possible uses:

- Rapid characterization of newly ingested media
- Identification of PII
- Determine who used a computer
- Social network analysis

Availability:

- Source code compiles on MacOS, Linux & Windows
- Pre-compiled Windows application available.



AFF & Digital Forensics XML

Digital Forensics Lacks Standards and Abstractions

Today's standards are limited to:

- Disk images (EnCase E01 format)
- Data bundling (ZIP)

We have two standardization efforts:

- AFF — Advanced Forensics Format
- Digital Forensics XML

Advanced Forensic Format (AFF)

AFF extensible schema stores:

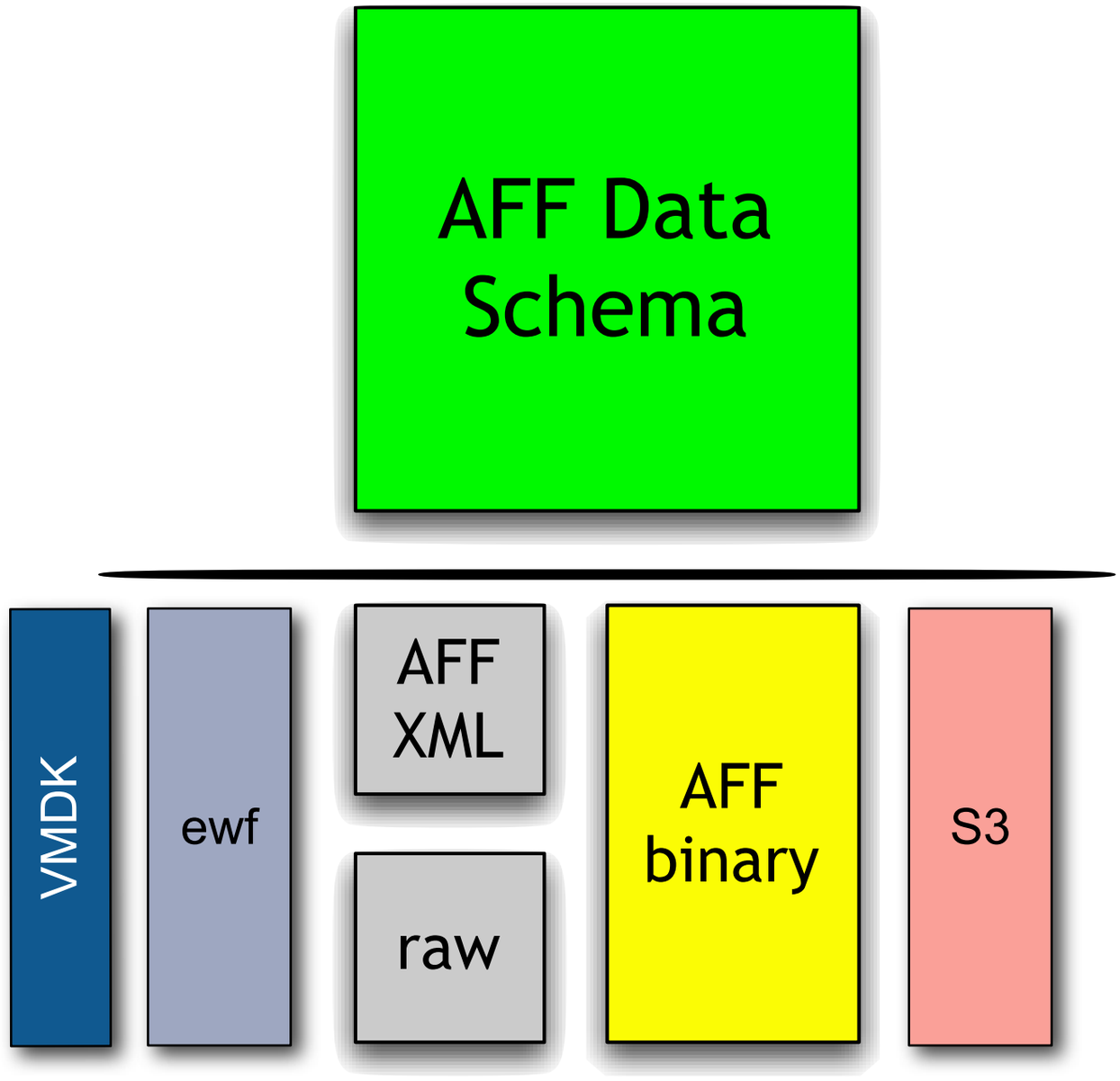
- Data copied from disk
- Metadata (SN, date of image)
- Chain-of-custody information

AFF supports:

- Compression
- Public Key Encryption & Digital Signatures
- Incremental file transfer

AFF Adoption:

- Sleuth Kit
- Black Bag (Mac Forensics)
- Others have expressed interest



AFF is designed for automating media exploitation

AFF: Can I use it today?

AFF3 — Downloadable today at http://afflib.org/

- Designed for archiving
- affuse — mounts an AFF disk as a “raw” image.

AFF4 — Under development by Michael Cohen

- Designed for archiving, high-performance & workflow automation
- Will be supported by SleuthKit this summer
- AFF3 disk images can be converted to AFF4.
- AFF4 programs can read AFF3 disk images.

Digital forensics XML enables automated processing of forensic disk images.

1

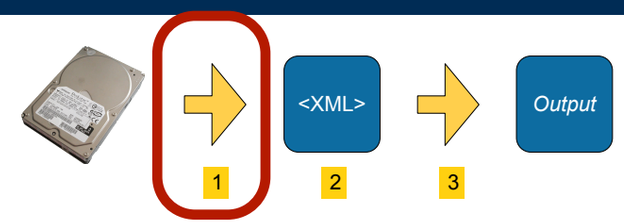
<XML>

2

3

Output

fiwalk extracts metadata from disk images.

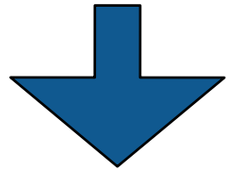


fiwalk is a C++ program built on top of SleuthKit

```
$ fiwalk [options] -X file.xml imagefile
```

Features:

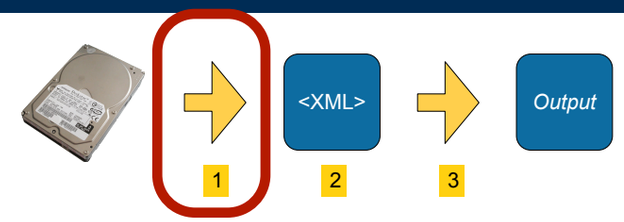
- Finds all partitions & automatically processes each.
- Handles file systems on raw device (partition-less).
- Creates a *single output file* with forensic data data from all.



Single program has multiple output formats:

- XML (for automated processing)
- ARFF (for data mining with Weka)
- "walk" format (easy debugging)
- SleuthKit Body File (for legacy timeline tools)
- CSV (for spreadsheets)*

fiwalk provides limited control over extraction.



Include/Exclude criteria:

- Presence/Absence of file SHA1 in a Bloom Filter
- File name matching.

```
fiwalk -n .jpeg /dev/sda          # just extract the .jpeg files
```

File System Metdata:

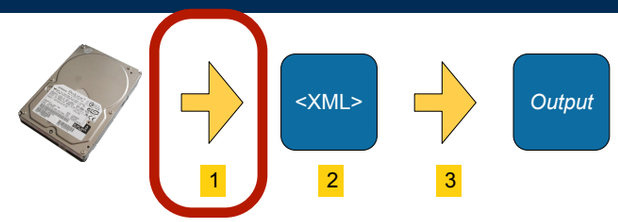
- -g — Report position of all file fragments
- -O — Do not report orphan or unallocated files

Full Content Options:

- -m — Report the MD5 of every file
- -1 — Report the SHA1 of every file
- -s *dir* — Save files to *dir*

fiwalk has a plugable metadata extraction system.

Configuration file specifies Metadata extractors:



- *Currently the extractor is chosen by the file extension.*

```
*.jpg    dgi      ../plugins/jpeg_extract
*.pdf    dgi      java -classpath plugins.jar Libextract_plugin
*.doc    dgi      java -classpath ../plugins/plugins.jar word_extract
```

- *Plugins are run in a different process for safety.*
- *We have designed a native JVM interface which uses IPC and 1 process.*

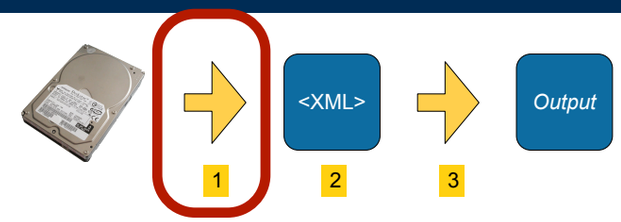
Metadata extractors produce name:value pairs on STDOUT

```
Manufacturer: SONY
Model: CYBERSHOT
Orientation: top - left
```

Extracted metadata is automatically incorporated into output.

XML is ideally suited for representing forensic data.

Forensic data is tree-structured.

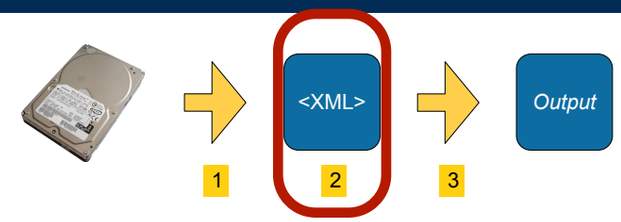


- Case > Devices > Partitions > Directories > Files

- Files
  - *file system metadata*
  - *file meta data*
  - *file content*

- Container Files (ZIP, tar, CAB)
  - *We can exactly represent the container structure*
  - *PyFlag does this with “virtual files”*
  - *No easy way to do this with the current TSK/EnCase/FTK structure*
  - *(Note: Container files not currently implemented.)*

fiwalk produces three kinds of XML tags.



Per-Image tags

```
<fiwalk> – outer tag
<fiwalk_version>0.4</fiwalk_version>
<Start_time>Mon Oct 13 19:12:09 2008</Start_time>
<Imagefile>dosfs.dmg</Imagefile>
<volume startsector="512">
```

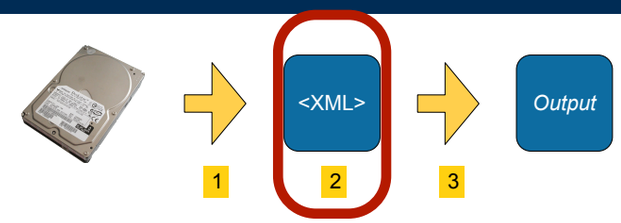
Per <volume> tags:

```
<Partition_Offset>512</Partition_Offset>
<block_size>512</block_size>
<ftype>4</ftype>
<ftype_str>fat16</ftype_str>
<block_count>81982</block_count>
```

Per <fileobject> tags:

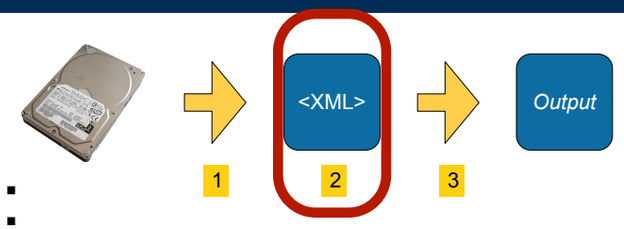
```
<filesize>4096</filesize>
<partition>1</partition>
<filename>linedash.gif</filename>
<libmagic>GIF image data, version 89a, 410 x 143</libmagic>
```

fiwalk XML example




```
<fileobject>
<filename>WINDOWS/system32/config/systemprofile/「开始」菜单/程序/附件/_rf55.tmp</
filename>
<filesize>1391</filesize>
<unalloc>1</unalloc>
<used>1</used>
<mtime>1150873922</mtime>
<ctime>1160927826</ctime>
<atime>1160884800</atime>
<fragments>0</fragments>
<md5>d41d8cd98f00b204e9800998ecf8427e</md5>
<sha1>da39a3ee5e6b4b0d3255bfef95601890afd80709</sha1>
<partition>1</partition>
<byte_runs type='resident'>
  <run file_offset='0' len='65536'
       fs_offset='871588864' img_offset='871621120'/>
  <run file_offset='65536' len='25920'
       fs_offset='871748608' img_offset='871780864'/>
</byte_runs>
</fileobject>
```

XML incorporates the extracted metadata.



fiwalk metadata extractors produce name:value pairs:

```
Manufacturer: SONY
Model: CYBERSHOT
Orientation: top - left
```

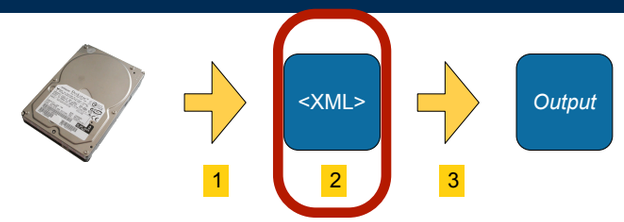
These are incorporated into XML:

```
<fileobject>
...
<Manufacturer>SONY</Manufacturer>
<Model>CYBERSHOT</Model>
<Orientation>top - left</Orientation>
...
</fileobject>
```

— *Special characters are automatically escaped.*

Resulting XML files can be distributed with images.

The XML file provides a key to the disk image:

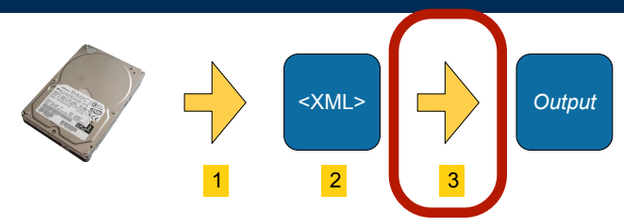



```
$ ls -l /corp/images/nps/nps-2009-domexusers/

-rw-r--r--  1 simsong  admin  4238912226 Jan 20 13:16 nps-2009-realistic.aff
-rw-r--r--  1 simsong  admin    38251423 May 10 23:58 nps-2009-realistic.xml

$
```

XML files:

- Range from 10K — 100MB.
  - *Depending on the complexity of the disk image.*
- Only have files & orphans that are identified by SleuthKit
  - *You can easily implement a "smart carver" that only carves unallocated sectors.*

fiwalk.py: a Python module for automated forensics.



Key Features:

- Automatically runs fiwalk with correct options if given a disk image
- Reads XML file if present (faster than regenerating)
- Creates **fileobject** objects.

Multiple interfaces:

- SAX callback interface

```
fiwalk_using_sax(imagefile, xmlfile, flags, callback)
```

  — *Very fast and minimal memory footprint*

- SAX procedural interface

```
objs = fileobjects_using_sax(imagefile, xmlfile, flags)
```

  — *Reasonably fast; returns a list of all file objects with XML in dictionary*

- DOM procedural interface

```
(doc,objs) = fileobjects_using_dom(imagefile, xmlfile, flags)
```

  — *Allows modification of XML that's returned.*

We have created several applications with the framework.

imap.py

- reads a disk image or XML file and prints a “map” of a disk image.

igroundtruth.py

- reads multiple disk images (different generations of the same disk)
- uses earlier images as “maps” for later images.
- Outputs new XML file

iverify.py

- Reads an image file and XML file.
- Reports which files in the XML file are actually resident in the image.

iredact.py

- reads a disk image (or XML file) and a “redaction file”
- Produces new disk image.

The redaction language is flexible.

Language: {CONDITION} {ACTION}

Conditions:

- FILENAME *filename*
- FILEPAT *file*name*
- DIRNAME dirname/
- MD5 *d41d8cd98f00b204e9800998ecf8427e*
- SHA1 *da39a3ee5e6b4b0d3255bfef95601890afd80709*
- FILE CONTAINS *user@company.com*
- SECTOR CONTAINS *user@company.com*

Actions:

- FILL *0x44*
- ENCRYPT
- FUZZ *(changes instructions but not strings)*

In Summary...

Download bulk_extractor from

- http://afflib.org/

Use our corpora for tool testing and training.

- http://digitalcorpora.org/

Push vendors to adopt standards.

Migrate to open source software.