fiwalk With Me:

Building Emergent Pre-Ingest Workflows for Digital Archival Records using Open Source Forensic Software

Mark A. Matienzo, Yale University Library

Code4lib 2011

mark@matienzo.org http://matienzo.org/ @anarchivist

Disclaimer

The following presentation expresses opinions of my own and not of my employer, my coworkers, etc.

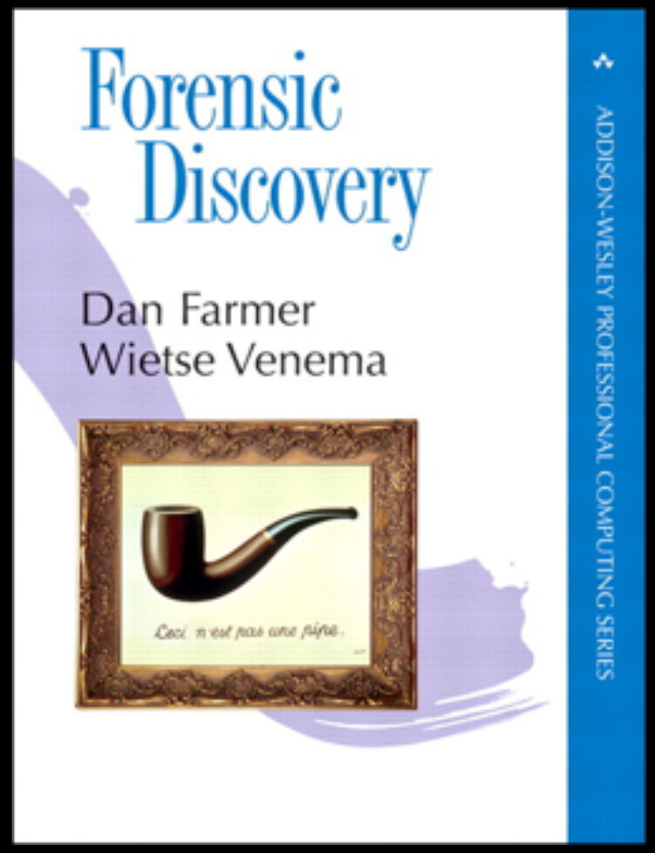
Digital forensics?

http://www.flickr.com/photos/freeparking/480863346/

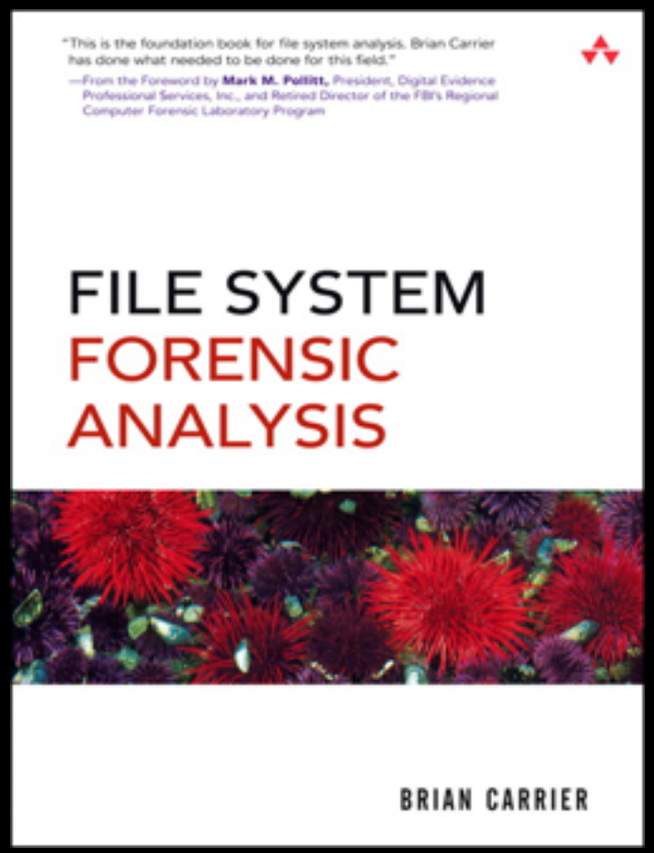
http://www.flickr.com/photos/johnjack/3449280414

Locard's exchange principle

Key Works



urn:isbn:978-0201634976



urn:isbn:978-0321268174

urn:isbn:978-1932326376

RULES

VS.

PRINCIPLES

http://www.flickr.com/photos/bjornmeansbear/4662232392/

Design Principles

- Use digital forensics software and methodology to support accessioning 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
- Use curation micro-services a guiding philosophy for implementation and further analysis

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

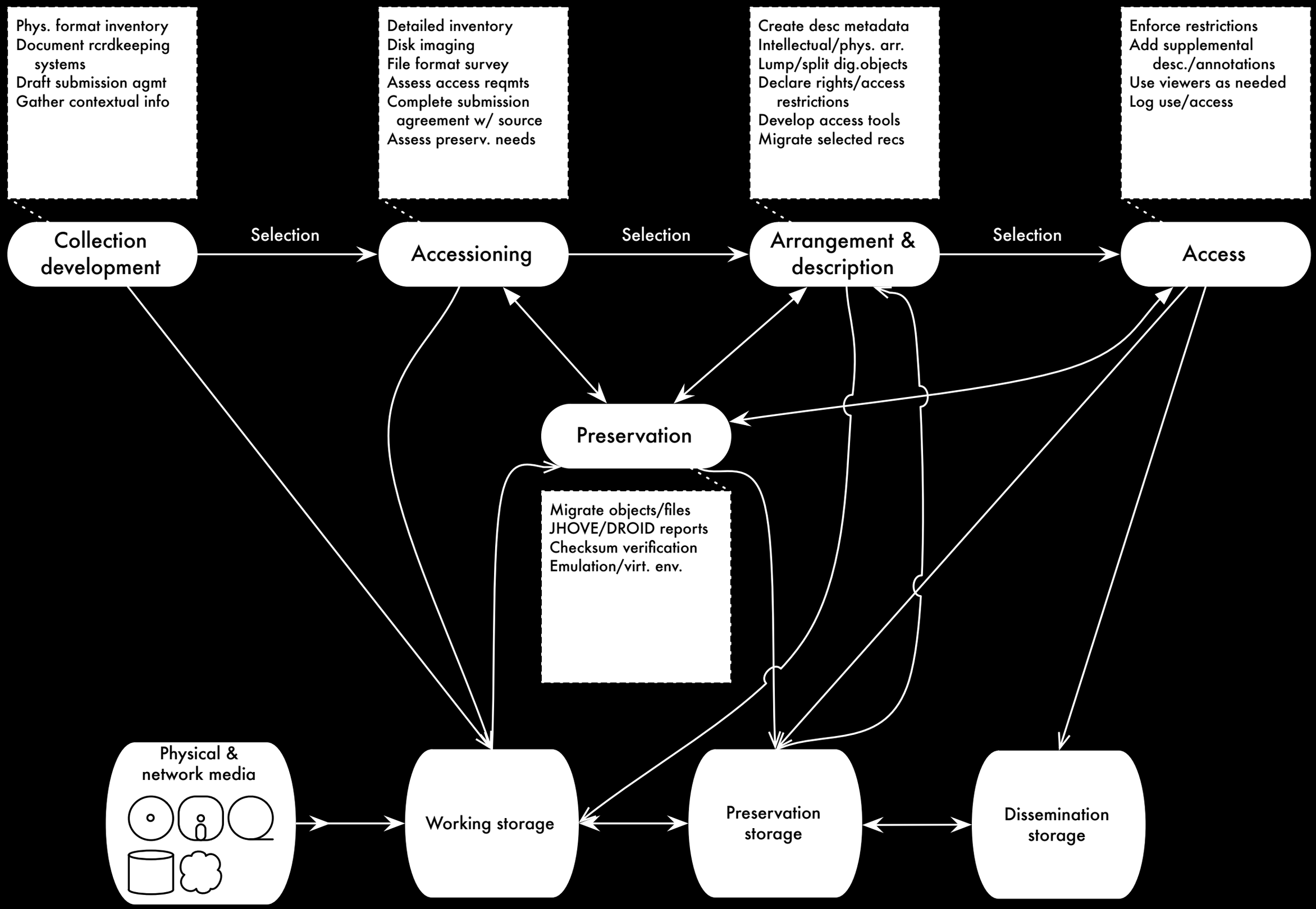
- Methods should approach or intend forensic soundness

- Ethical issues (as raised in CLIR report) are out of scope

Mitigate Risk

http://www.flickr.com/photos/moparx/4013824025/

A Larger Workflow



Open Source Forensics

- Digital forensics is a high-stakes market
- Proprietary forensics software is not easily extensible
- Proprietary forensics software is often platform-specific
- Cultural heritage institutions are still an emerging market for digital forensics
- Our needs are different and still being defined

Microservices as Philosophy

http://www.flickr.com/photos/gregmote/2797330534

Principles

- Granularity
- Orthogonality
- Parsimony
- Evolution

Preferences

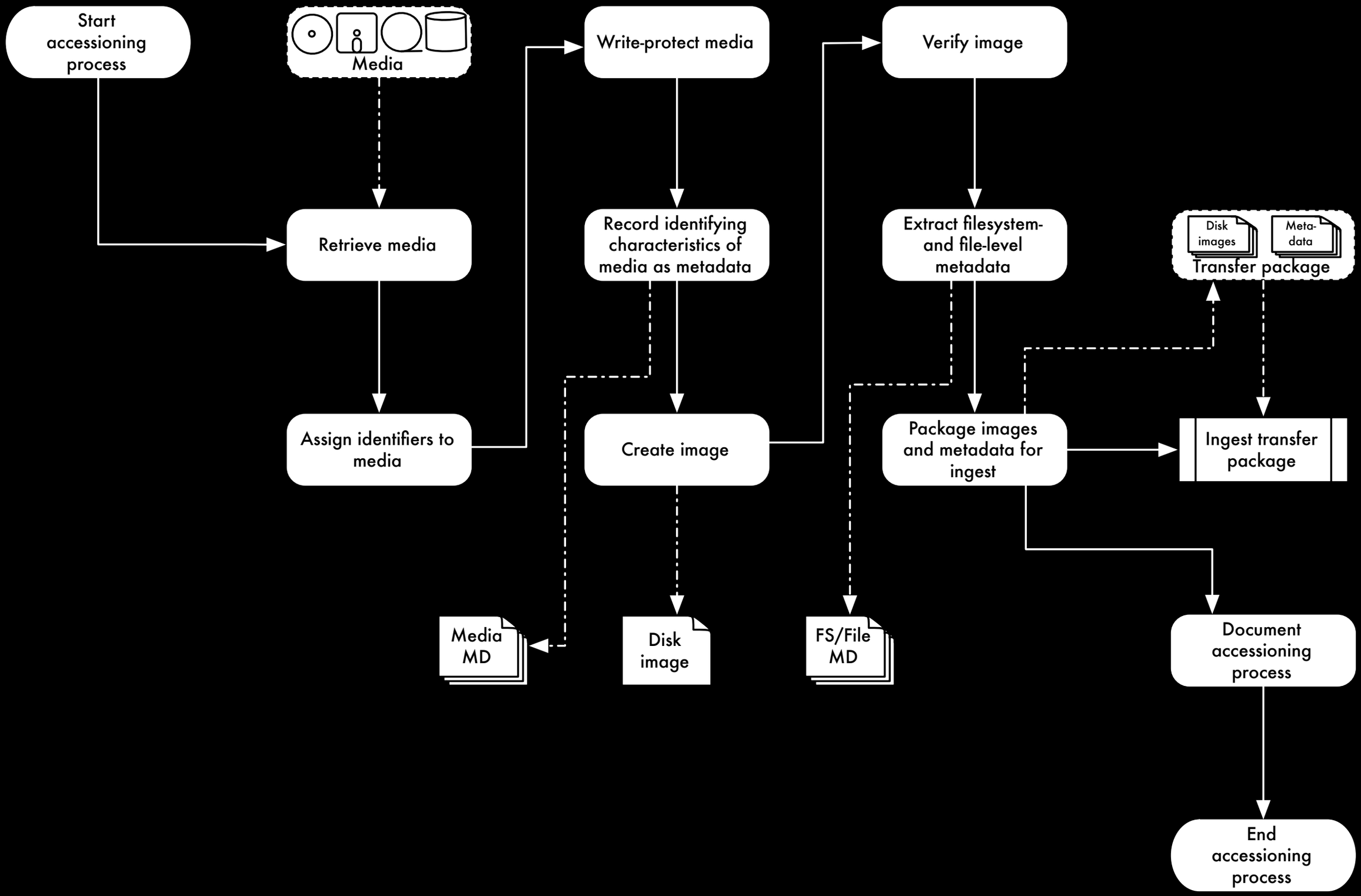
- Small and simple over large and complex
- Minimally sufficient over feature-laden
- Configurable over the prescribed
- The proven over the merely novel
- Outcomes over means

Practices

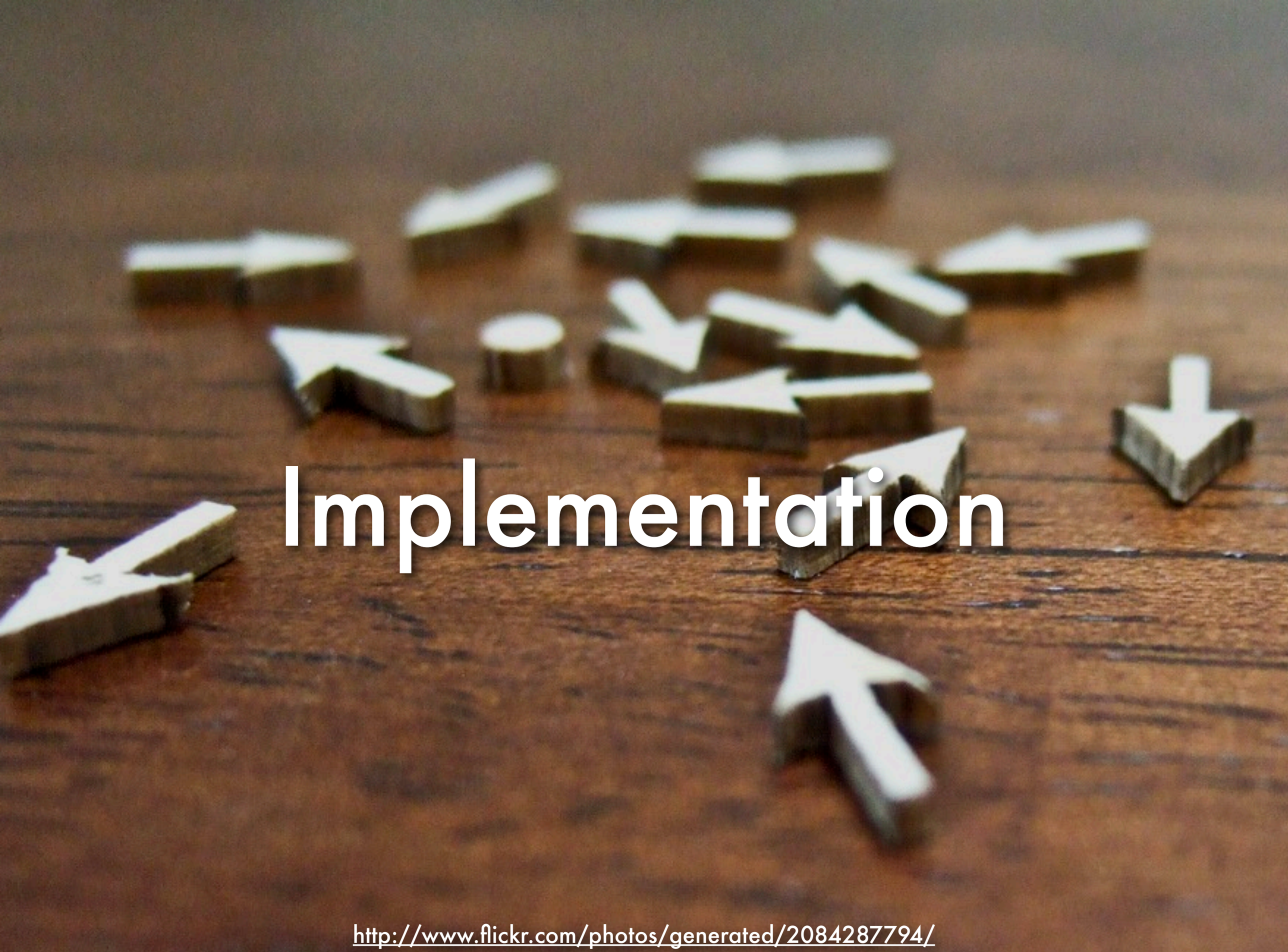
- Define, decompose, recurse
- Top down design, bottom up implementation
- Code to interfaces
- Sufficiency through a series of incrementally necessary steps

UC Curation Center/California Digital Library, “UC3 Curation Foundations”
https://confluence.ucop.edu/download/attachments/13860983/UC3-Foundations-latest.pdf

Accessioning Workflow



Implementation

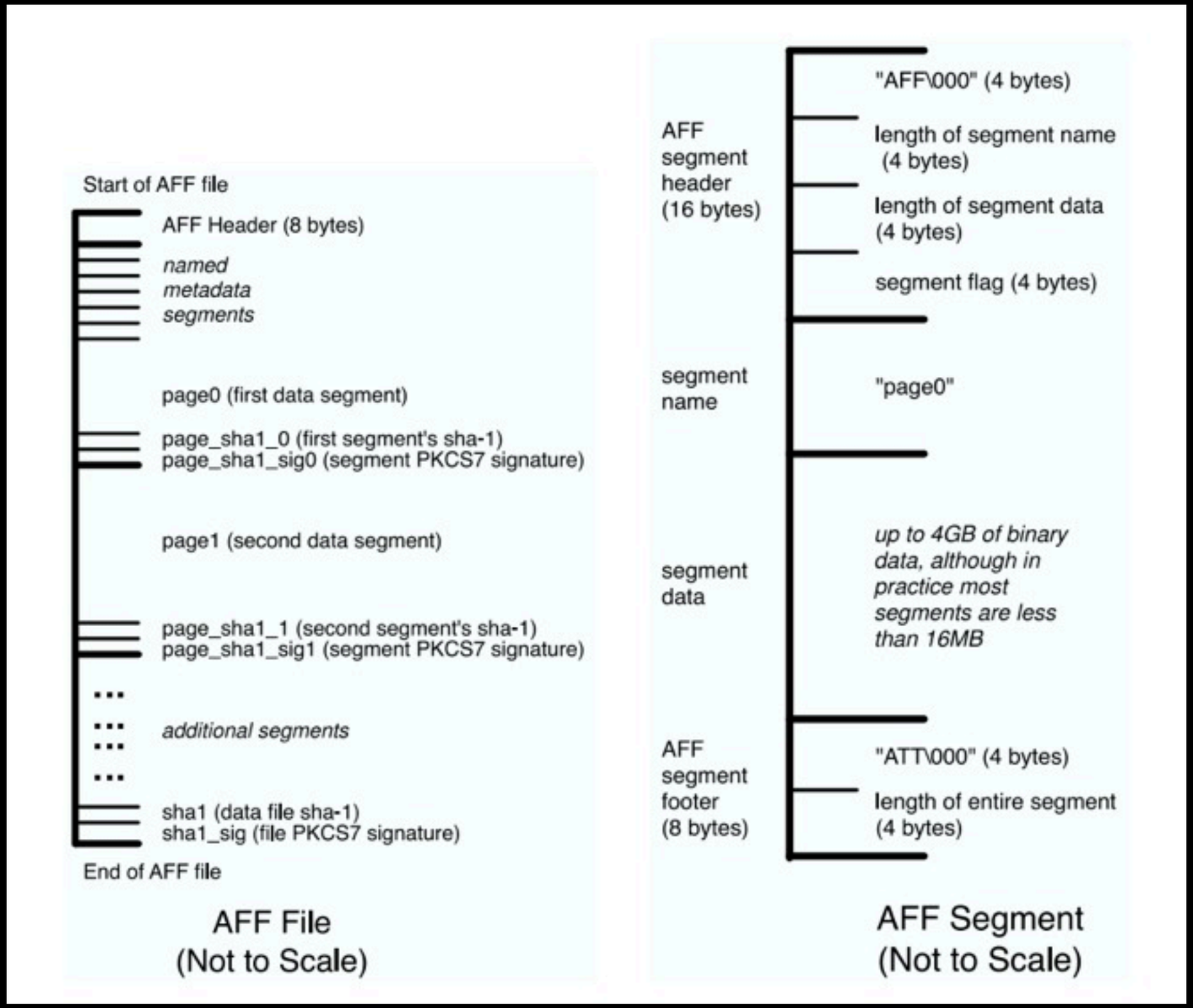


http://www.flickr.com/photos/generated/2084287794/

Disk Imaging

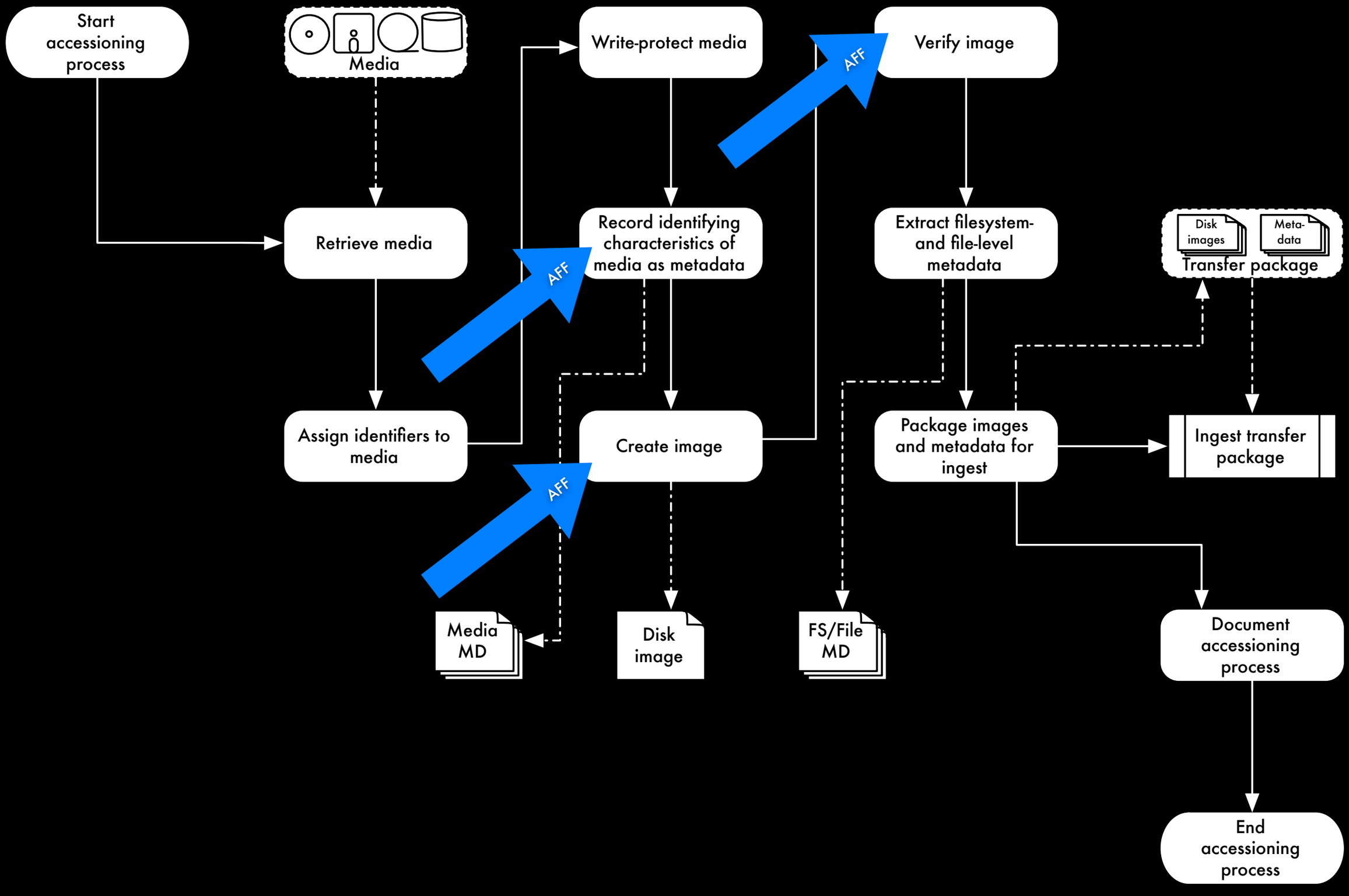
- dd: creates raw images
  - Related implementations: dc3dd, dcfldd, dd_rescue
    - Fast, but no mechanism to store imaging metadata
- Advanced Forensic Format (AFF)/AFFLib
  - Cross platform, reasonably fast
  - Can store arbitrary metadata
  - Plenty of GUI-based imaging tools

AFF File Structure



Garfinkel et al. 2006 (http://nrs.harvard.edu/urn-3:HUL.InstRepos:2829932), p. 27

Accessioning Workflow



The Sleuth Kit

- Open source C library, command line tools, and browser-based Perl application (Autopsy) for forensic analysis
- Supports analysis of NTFS, FAT, 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

Extracting Metadata & Files

```
$ fls -m A: -a -f fat 2004-M-008.0018.aff
0|A:/_ublist1.wpd (deleted)|3|r/rrwxrwxrwx|0|0|202152|982818000|982881052|0|982881114
0|A:/publist.wpd|4|r/rrwxrwxrwx|0|0|202119|1285041600|982963054|0|982963226
0|A:/publistwkg.wpd|7|r/rrwxrwxrwx|0|0|205607|1285041600|982963038|0|982963237
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

$ fls -m A: -a -f fat 2004-M-008.0018.aff | mactime
Wed Dec 31 1969 19:00:00   202152 ..c. r/rrwxrwxrwx 0        0        3        A:/_ublist1.wpd (deleted)
                           202119 ..c. r/rrwxrwxrwx 0        0        4        A:/publist.wpd
                           205607 ..c. r/rrwxrwxrwx 0        0        7        A:/publistwkg.wpd
Thu Feb 22 2001 00:00:00   202152 .a.. r/rrwxrwxrwx 0        0        3        A:/_ublist1.wpd (deleted)
Thu Feb 22 2001 17:30:52   202152 m... r/rrwxrwxrwx 0        0        3        A:/_ublist1.wpd (deleted)
Thu Feb 22 2001 17:31:54   202152 ...b r/rrwxrwxrwx 0        0        3        A:/_ublist1.wpd (deleted)
Fri Feb 23 2001 16:17:18   205607 m... r/rrwxrwxrwx 0        0        7        A:/publistwkg.wpd
Fri Feb 23 2001 16:17:34   202119 m... r/rrwxrwxrwx 0        0        4        A:/publist.wpd
Fri Feb 23 2001 16:20:26   202119 ...b r/rrwxrwxrwx 0        0        4        A:/publist.wpd
Fri Feb 23 2001 16:20:37   205607 ...b r/rrwxrwxrwx 0        0        7        A:/publistwkg.wpd
Tue Sep 21 2010 00:00:00   202119 .a.. r/rrwxrwxrwx 0        0        4        A:/publist.wpd
                           205607 .a.. r/rrwxrwxrwx 0        0        7        A:/publistwkg.wpd

$ icat 2004-M-008.0018.aff 4 | fido.sh -
OK,110,x-fmt/44,WordPerfect for MS-DOS/Windows Document,WordPerfect for Windows 6.x - 12,202119,"STDIN"

$ icat 2004-M-008.018.aff 4 > publist.wpd
```

Extracting Metadata & Files

```
$ fls -m A: -a -f fat 2004-M-008.0018.aff
0|A:/_ublist1.wpd (deleted)|3|r/rrwxrwxrwx|0|0|202152|982818000|982881052|0|982881114
0|A:/publist.wpd|4|r/rrwxrwxrwx|0|0|202119|1285041600|982963054|0|982963226
0|A:/publistwkg.wpd|7|r/rrwxrwxrwx|0|0|205607|1285041600|982963038|0|982963237
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
```

```
$ fls -m A: -a -f fat ~/Desktop/2004-M-008/data/2004-M-008.0018.aff | mactime
Wed Dec 31 1969 19:00:00   202152 ..c. r/rrwxrwxrwx 0        0        3        A:/_ublist1.wpd (deleted)
                           202119 ..c. r/rrwxrwxrwx 0        0        4        A:/publist.wpd
                           205607 ..c. r/rrwxrwxrwx 0        0        7        A:/publistwkg.wpd
Thu Feb 22 2001 00:00:00   202152 .a.. r/rrwxrwxrwx 0        0        3        A:/_ublist1.wpd (deleted)
Thu Feb 22 2001 17:30:52   202152 m... r/rrwxrwxrwx 0        0        3        A:/_ublist1.wpd (deleted)
Thu Feb 22 2001 17:31:54   202152 ...b r/rrwxrwxrwx 0        0        3        A:/_ublist1.wpd (deleted)
Fri Feb 23 2001 16:17:18   205607 m... r/rrwxrwxrwx 0        0        7        A:/publistwkg.wpd
Fri Feb 23 2001 16:17:34   202119 m... r/rrwxrwxrwx 0        0        4        A:/publist.wpd
Fri Feb 23 2001 16:20:26   202119 ...b r/rrwxrwxrwx 0        0        4        A:/publist.wpd
Fri Feb 23 2001 16:20:37   205607 ...b r/rrwxrwxrwx 0        0        7        A:/publistwkg.wpd
Tue Sep 21 2010 00:00:00   202119 .a.. r/rrwxrwxrwx 0        0        4        A:/publist.wpd
                           205607 .a.. r/rrwxrwxrwx 0        0        7        A:/publistwkg.wpd

$ icat 2004-M-008.0018.aff 4 | fido.sh -
OK,110,x-fmt/44,WordPerfect for MS-DOS/Windows Document,WordPerfect for Windows 6.x - 12,202119,"STDIN"

$ icat 2004-M-008.018.aff 4 > publist.wpd
```

Extracting Metadata & Files

```
$ fls -m A: -a -f fat 2004-M-008.0018.aff
0|A:/_ublist1.wpd (deleted)|3|r/rrwxrwxrwx|0|0|202152|982818000|982881052|0|982881114
0|A:/publist.wpd|4|r/rrwxrwxrwx|0|0|202119|1285041600|982963054|0|982963226
0|A:/publistwkg.wpd|7|r/rrwxrwxrwx|0|0|205607|1285041600|982963038|0|982963237
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
```

```
$ fls -m A: -a -f fat ~/Desktop/2004-M-008/data/2004-M-008.0018.aff | mactime
Wed Dec 31 1969 19:00:00   202152 ..c. r/rrwxrwxrwx 0        0        3        A:/_ublist1.wpd (deleted)
                           202119 ..c. r/rrwxrwxrwx 0        0        4        A:/publist.wpd
                           205607 ..c. r/rrwxrwxrwx 0        0        7        A:/publistwkg.wpd
Thu Feb 22 2001 00:00:00   202152 .a.. r/rrwxrwxrwx 0        0        3        A:/_ublist1.wpd (deleted)
Thu Feb 22 2001 17:30:52   202152 m... r/rrwxrwxrwx 0        0        3        A:/_ublist1.wpd (deleted)
Thu Feb 22 2001 17:31:54   202152 ...b r/rrwxrwxrwx 0        0        3        A:/_ublist1.wpd (deleted)
Fri Feb 23 2001 16:17:18   205607 m... r/rrwxrwxrwx 0        0        7        A:/publistwkg.wpd
Fri Feb 23 2001 16:17:34   202119 m... r/rrwxrwxrwx 0        0        4        A:/publist.wpd
Fri Feb 23 2001 16:20:26   202119 ...b r/rrwxrwxrwx 0        0        4        A:/publist.wpd
Fri Feb 23 2001 16:20:37   205607 ...b r/rrwxrwxrwx 0        0        7        A:/publistwkg.wpd
Tue Sep 21 2010 00:00:00   202119 .a.. r/rrwxrwxrwx 0        0        4        A:/publist.wpd
                           205607 .a.. r/rrwxrwxrwx 0        0        7        A:/publistwkg.wpd
```

```
$ icat 2004-M-008.0018.aff 4 | fido.sh -
OK,110,x-fmt/44,WordPerfect for MS-DOS/Windows Document,WordPerfect for Windows 6.x - 12,202119,"STDIN"

$ icat 2004-M-008.018.aff 4 > publist.wpd
```

Extracting Metadata & Files

```
$ fls -m A: -a -f fat 2004-M-008.0018.aff
0|A:/_ublist1.wpd (deleted)|3|r/rrwxrwxrwx|0|0|202152|982818000|982881052|0|982881114
0|A:/publist.wpd|4|r/rrwxrwxrwx|0|0|202119|1285041600|982963054|0|982963226
0|A:/publistwkg.wpd|7|r/rrwxrwxrwx|0|0|205607|1285041600|982963038|0|982963237
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

$ fls -m A: -a -f fat ~/Desktop/2004-M-008/data/2004-M-008.0018.aff | mactime
Wed Dec 31 1969 19:00:00   202152 ..c. r/rrwxrwxrwx 0        0        3        A:/_ublist1.wpd (deleted)
                           202119 ..c. r/rrwxrwxrwx 0        0        4        A:/publist.wpd
                           205607 ..c. r/rrwxrwxrwx 0        0        7        A:/publistwkg.wpd
Thu Feb 22 2001 00:00:00   202152 .a.. r/rrwxrwxrwx 0        0        3        A:/_ublist1.wpd (deleted)
Thu Feb 22 2001 17:30:52   202152 m... r/rrwxrwxrwx 0        0        3        A:/_ublist1.wpd (deleted)
Thu Feb 22 2001 17:31:54   202152 ...b r/rrwxrwxrwx 0        0        3        A:/_ublist1.wpd (deleted)
Fri Feb 23 2001 16:17:18   205607 m... r/rrwxrwxrwx 0        0        7        A:/publistwkg.wpd
Fri Feb 23 2001 16:17:34   202119 m... r/rrwxrwxrwx 0        0        4        A:/publist.wpd
Fri Feb 23 2001 16:20:26   202119 ...b r/rrwxrwxrwx 0        0        4        A:/publist.wpd
Fri Feb 23 2001 16:20:37   205607 ...b r/rrwxrwxrwx 0        0        7        A:/publistwkg.wpd
Tue Sep 21 2010 00:00:00   202119 .a.. r/rrwxrwxrwx 0        0        4        A:/publist.wpd
                           205607 .a.. r/rrwxrwxrwx 0        0        7        A:/publistwkg.wpd
```

```
$ icat 2004-M-008.0018.aff 4 | fido.sh -
OK,110,x-fmt/44,WordPerfect for MS-DOS/Windows Document,WordPerfect for Windows 6.x - 12,202119,"STDIN"
```

```
$ icat 2004-M-008.018.aff 4 > publist.wpd
```

Extracting Metadata & Files

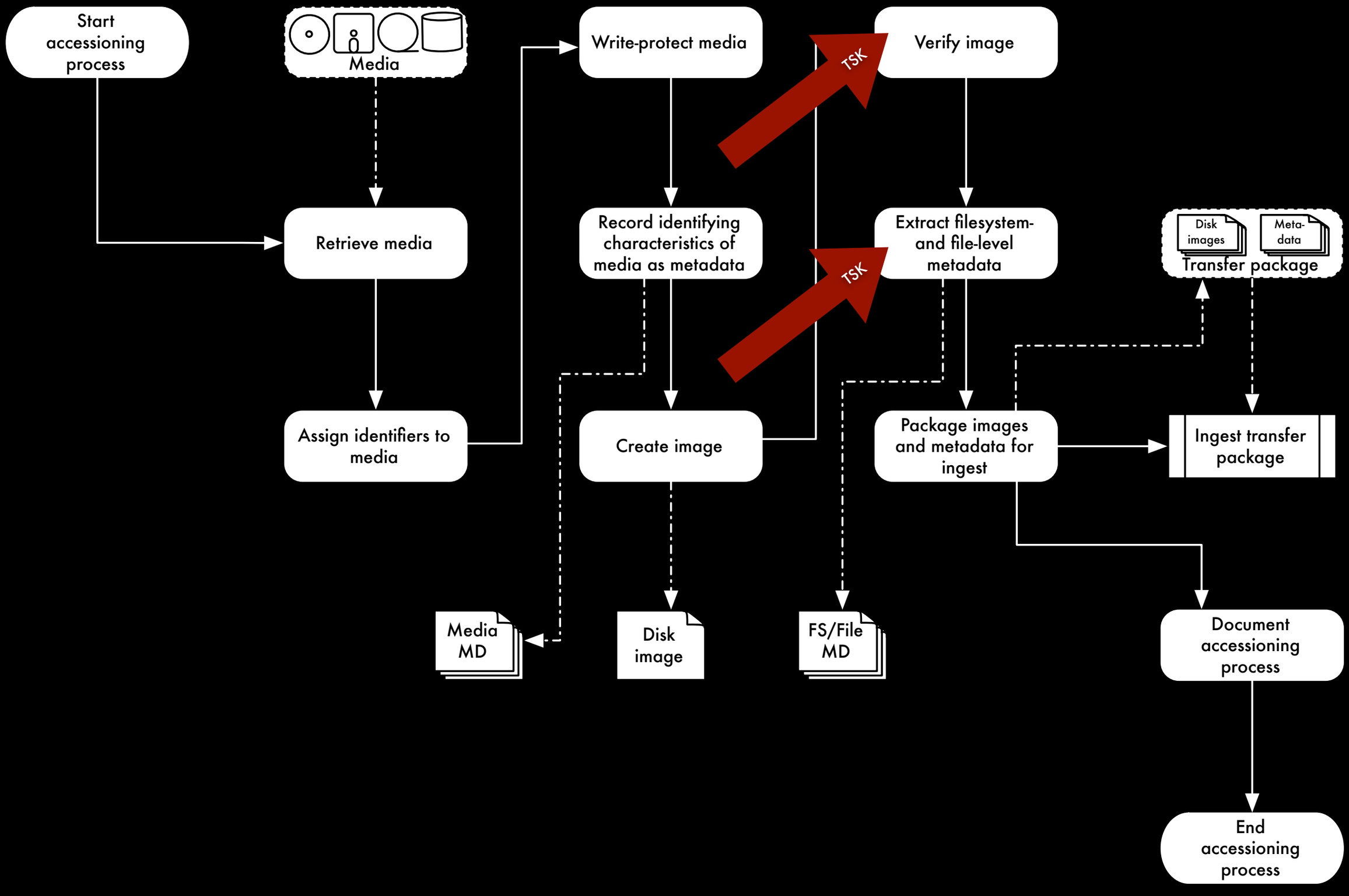
```
$ fls -m A: -a -f fat 2004-M-008.0018.aff
0|A:/_ublist1.wpd (deleted)|3|r/rrwxrwxrwx|0|0|202152|982818000|982881052|0|982881114
0|A:/publist.wpd|4|r/rrwxrwxrwx|0|0|202119|1285041600|982963054|0|982963226
0|A:/publistwkg.wpd|7|r/rrwxrwxrwx|0|0|205607|1285041600|982963038|0|982963237
0|A:/$MBR|45779|v/v---------|0|0|512|0|0|0|0
0|A:/$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

$ fls -m A: -a -f fat ~/Desktop/2004-M-008/data/2004-M-008.0018.aff | mactime
Wed Dec 31 1969 19:00:00   202152 ..c. r/rrwxrwxrwx 0        0        3        A:/_ublist1.wpd (deleted)
                           202119 ..c. r/rrwxrwxrwx 0        0        4        A:/publist.wpd
                           205607 ..c. r/rrwxrwxrwx 0        0        7        A:/publistwkg.wpd
Thu Feb 22 2001 00:00:00   202152 .a.. r/rrwxrwxrwx 0        0        3        A:/_ublist1.wpd (deleted)
Thu Feb 22 2001 17:30:52   202152 m... r/rrwxrwxrwx 0        0        3        A:/_ublist1.wpd (deleted)
Thu Feb 22 2001 17:31:54   202152 ...b r/rrwxrwxrwx 0        0        3        A:/_ublist1.wpd (deleted)
Fri Feb 23 2001 16:17:18   205607 m... r/rrwxrwxrwx 0        0        7        A:/publistwkg.wpd
Fri Feb 23 2001 16:17:34   202119 m... r/rrwxrwxrwx 0        0        4        A:/publist.wpd
Fri Feb 23 2001 16:20:26   202119 ...b r/rrwxrwxrwx 0        0        4        A:/publist.wpd
Fri Feb 23 2001 16:20:37   205607 ...b r/rrwxrwxrwx 0        0        7        A:/publistwkg.wpd
Tue Sep 21 2010 00:00:00   202119 .a.. r/rrwxrwxrwx 0        0        4        A:/publist.wpd
                           205607 .a.. r/rrwxrwxrwx 0        0        7        A:/publistwkg.wpd

$ icat 2004-M-008.0018.aff 4 | fido.sh -
OK,110,x-fmt/44,WordPerfect for MS-DOS/Windows Document,WordPerfect for Windows 6.x - 12,202119,"STDIN"
```

```
$ icat 2004-M-008.018.aff 4 > publist.wpd
```

Accessioning Workflow



fiwalk

- C++ program with Python module for processing images
- Outputs results in plain text key/value pairs, XML, CSV, or ARFF (for Weka data mining software)
- Developed to support automated forensic processing by breaking it into three steps: extract, represent, process
- Pluggable file-level metadata extraction (expects key/value pairs)
- Makes development easy and fast

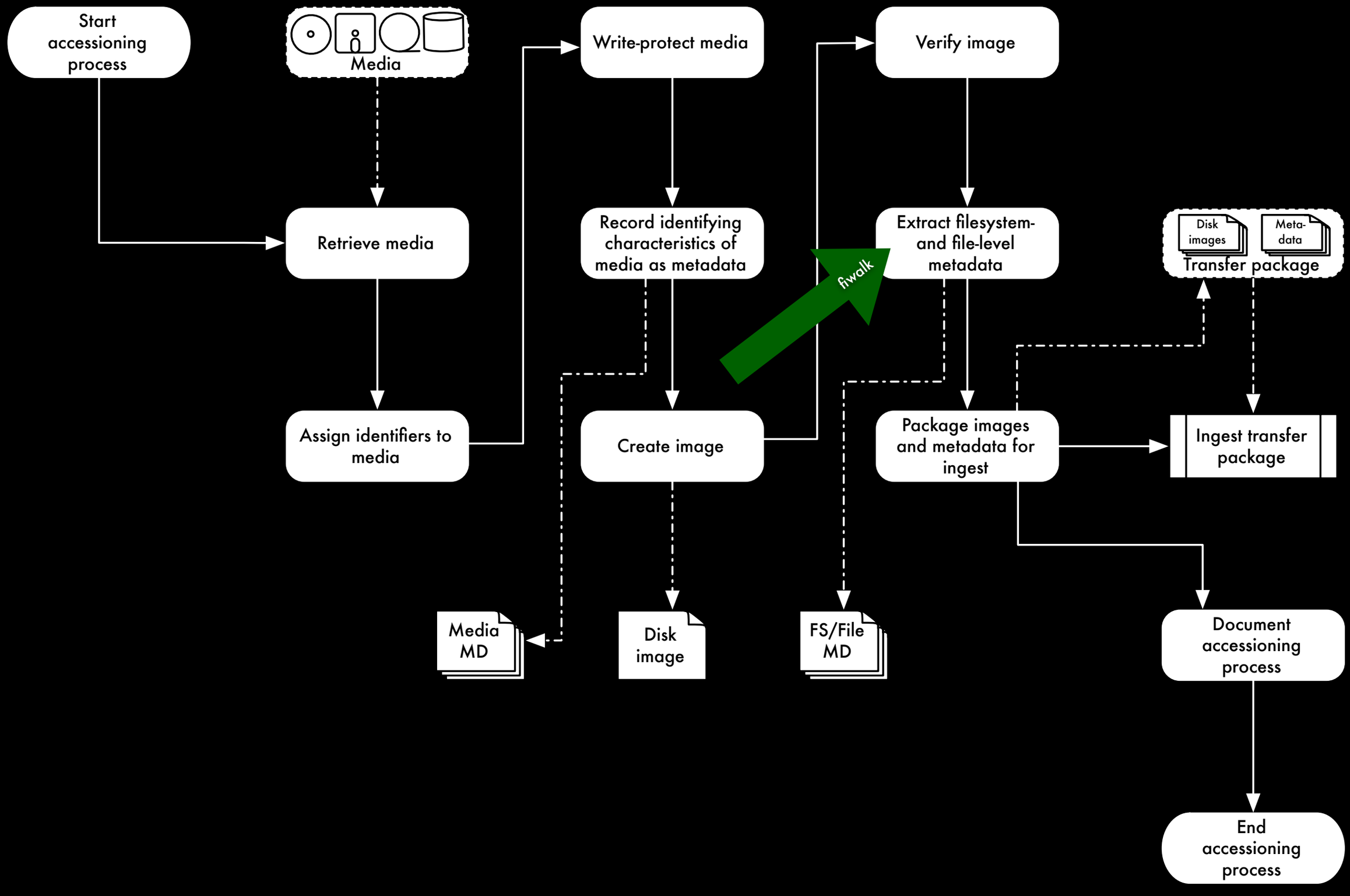
Sample fiwalk Output

```
<?xml version='1.0' encoding='UTF-8'?>
<fiwalk xmloutputversion='0.3'>
  <metadata> <!-- metadata about the disk image -->
  <creator> <!-- fiwalk provenance metadata (runtime environment, etc.) --></creator>
  <source>
    <image_filename>2004-M-008.dd-0018.001</image_filename>
  </source>
<!-- fs start: 0 -->
  <volume offset='0'>
    <!-- volume metadata -->
    <fileobject>
      <filename>_ublist1.wpd</filename>
      <!-- more metadata about specific files within the image -->
    </fileobject>
    <fileobject/><!-- one for each file -->
  </volume>
  <runstats>
    <!-- runtime statistics -->
  </runstats>
</fiwalk>
```

Sample fiwalk 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>982881052</mtime>
  <atime>982818000</atime>
  <crtime>982881114</crtime>
  <libmagic>(Corel/WP)</libmagic>
  <byte_runs>
   <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>
```

Accessioning Workflow



Why fiwalk?

- 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?)
- fiwalk's output is something like a METS structMap
- Building non-invasive assessment tools takes less time

Ballantine Books 02534-2-095

GUMSHOE

95¢

Neville Smith

A MAJOR COLUMBIA FILM STARRING
ALBERT FINNEY

The Private Eye will never be the same again.

Gumshoe

- Prototype application
- Blacklight (Ruby on Rails + Solr) & Python indexing code
- Indexing code works with fiwalk output or directly over a disk image (using fiwalk's Python bindings)
- Populates Solr index with all file-level metadata from fiwalk and text strings extracted from files
- Code at http://github.com/anarchivist/gumshoe
- Demo at http://xgumshoex.heroku.com/

Future Directions

http://www.flickr.com/photos/87913776@N00/5129662279/

AFF4

- Emerging format, with tools still under development
- Better for a distributed environment
- Friendlier to micro-services philosophy
- Clearer object and metadata model (RDF-based)
- Container format is Zip64
- Cohen and Schatz 2010 (doi:10.1016/j.diin.2010.05.015) show hash based imaging as a more efficient alternative

Sample AFF4 Metadata

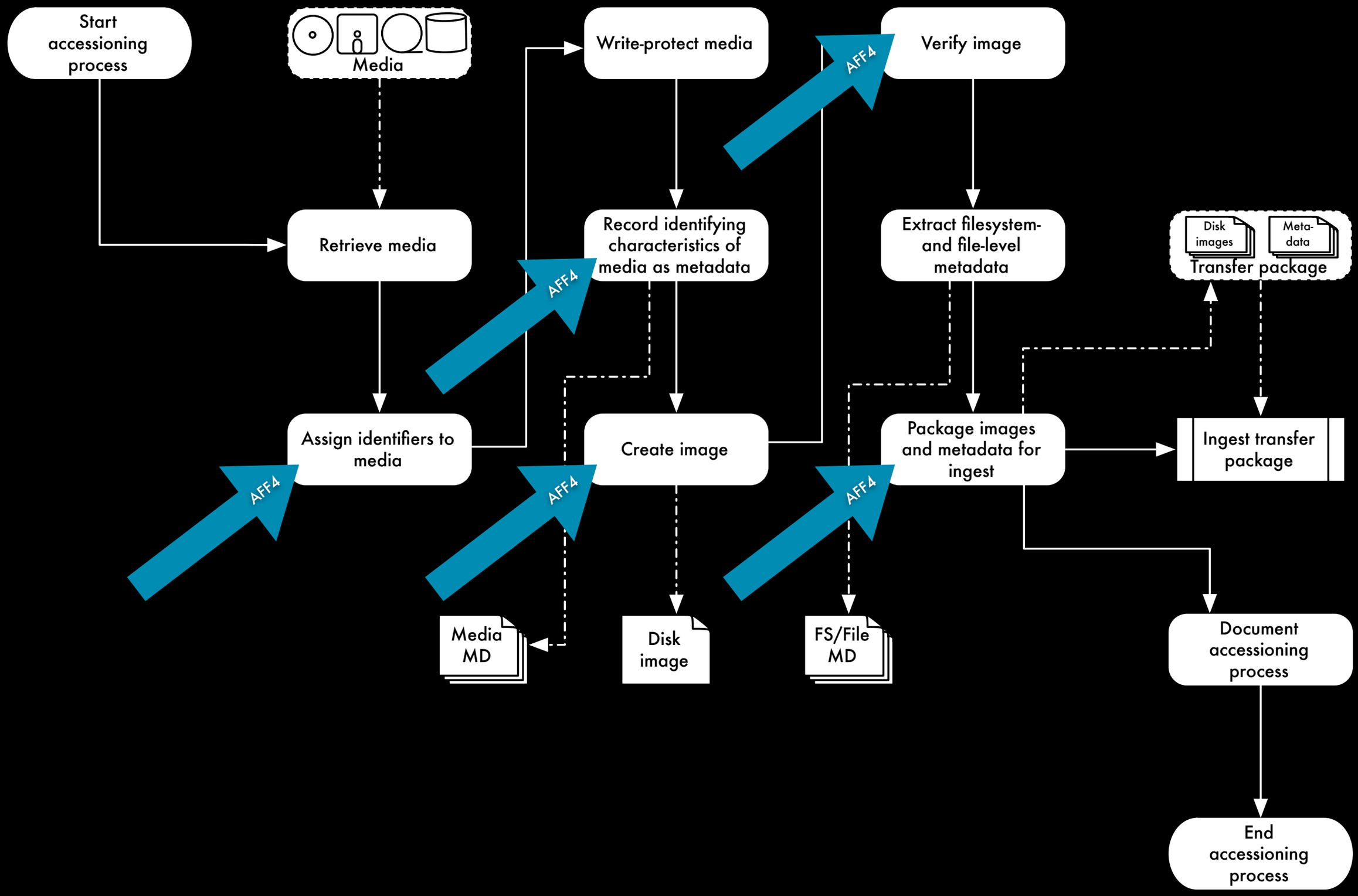
```
@prefix D1: <urn:aff4:652e4027-27fab2941>
@prefix G1: <urn:aff4:19857a87-a190b2f87>
@prefix G2: <urn:aff4:0a1fc78a-927bfacef>
@prefix S1: <urn:aff4:652e4027-ffff01199>
@prefix I1: <urn:aff4:9003027a-11199ffff>
@prefix aff4: <http://afflib.org/2009/aff4/#>
@prefix dcterms: <http://purl.org/dc/terms/>

G1: {
    D1: aff4:serialNumber "zx322o91"
    D1: aff4:hash "3897450fa18094b13"^^aff4:md5
}

G2: {
    S1: aff4:name "aff4imager"
    S1: aff4:vendor <http://aff.org/>
    S1: aff4:asserts G1:
    S1: aff4:type aff4:AcquisitionTool.
    S1: aff4:version "0.2"
    I1: aff4:type aff4:Image.
    I1: aff4:hash "3897450fa18094b13"^^aff4:md5
    I1: dcterms:creator S1:
}
```

Adapted from Schatz and Cohen 2010 (doi:10.1007/978-3-642-15506-2_16), pp. 234-236

Accessioning Workflow



Further Toolsets

http://www.flickr.com/photos/oskay/5369749968/

Arrangement and Description Tool

Home > Collections > Jimmy Olson papers (MS 201)

Accessions

2001.99 | 2005.12 | 2009.35

- A::\ (2005.12.floppy1.aff)
  - SFBG0122.doc
  - SFBG0214.doc
- c:\ (2005.12.hdd1.aff4)
  - Article drafts
    - BOLLFINAL.pdf
    - bollweevils1204.doc
  - E-mail from Barley
  - Research
    - Cabbage Foundation
    - Cider making

search | Load in Viewer | Download | Delete

Collection View

- MS 201: Jimmy Olson papers
  - Series I. Personal correspondence
    - Babbitt, Barley
    - Cabbage, Carly
  - Series II. Subject Files
    - Apple cider making
    - Boll weevil hunting
  - Series III. Published writings
    - LA Times
    - NY Times
    - SF Bay Guardian
  - Series IV. Consulting records
    - Financial records
    - Reports

Add Child | Add Sibling | search | Delete

File viewer

Hunting the Boll Weevil

by Jimmy Olson
February 1, 2000

You know, many people these days have problems with boll weevils. Cras nec enim non purus tempor consequat vel quis lorem. Ut sem turpis, blandit sed

Full Screen | Zoom

Technical Metadata

Field	Value
Format	PDF
PUID	fmt/20
Size	32.6 KB
ctime	2000-01-31 17:39
mtime	2000-01-31 18:03
Media	Hard disk
Image	2005.12.hdd1.aff4
Etc	etc

Basic Description | Access Points | Notes | Permissions

PID hdl:10079/0d34fa9 Unit ID

Level Subseries Other

Title SF Bay Guardian

Date expression 2000-2005

Inclusive Begin / / End / /

Bulk Begin / / End / /

Language English Extent 30 KB

☐ Internal Only

...

Associated Files

Filename	Accession	Image file
BOLLFINAL.pdf	2005.12	2005.12.hdd1.aff4
SFBG0122.doc	2005.12	2005.12.floppy1.aff4

Load in Viewer | Download | Remove Association

Thank You

mark@matienzo.org
http://matienzo.org/
twitter: @anarchivist