Accessioning-Based Metadata Extraction and Iterative Processing: Notes From the Field

Mark A. Matienzo, Yale University Library

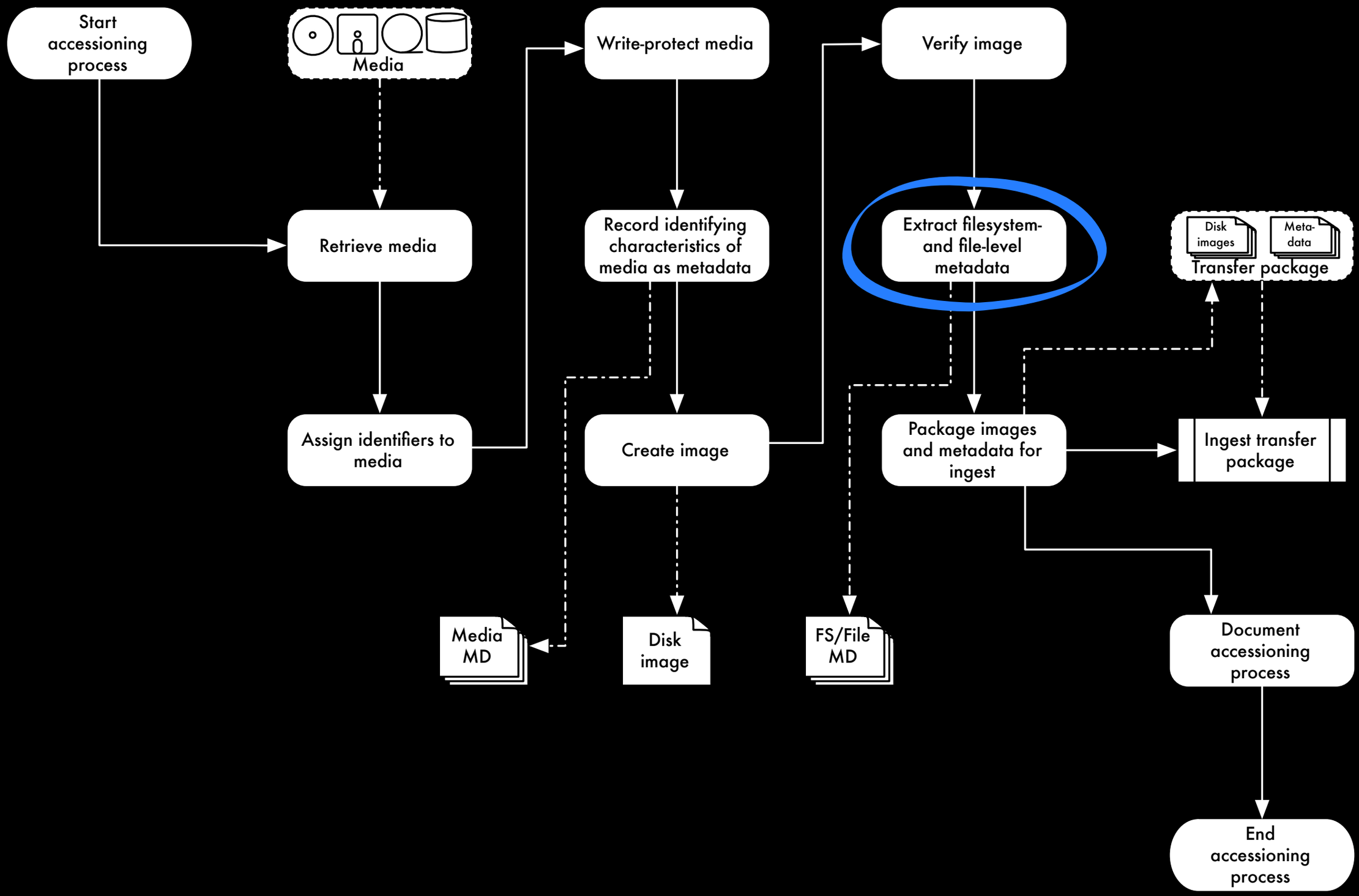
CurateGear: Enabling the Curation of Digital Collections

January 6, 2012

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

Digital Archives at Yale

Accessioning Workflow



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/Sleuthkit; FTK Imager; testing others

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>
```

Current 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?)
- Building supplemental tools takes less time

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

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

Current Limitations

- Use of fiwalk 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

Thank You

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