

Cal Lee

Kam Woods

UNC School of Information and Library Science

November 5th, 2010

CURATION OF A DISK IMAGE COLLECTION TO SUPPORT EDUCATION



UNC
SCHOOL OF INFORMATION
AND LIBRARY SCIENCE



Who we are and why we're doing this...

Project Background

- **Raw data streams (i.e. images) extracted from storage media can play an essential role in the acquisition and management of digital collections.**
- Ensuring continued access to the underlying bits without depending on physical carriers, which may be fragile or become obsolete
- Fail-safe mechanisms when curatorial actions have made unexpected changes to data
- Proof of file integrity and chain of custody
- Data below the userspace filesystem; metadata, recoverable sectors and configuration information.

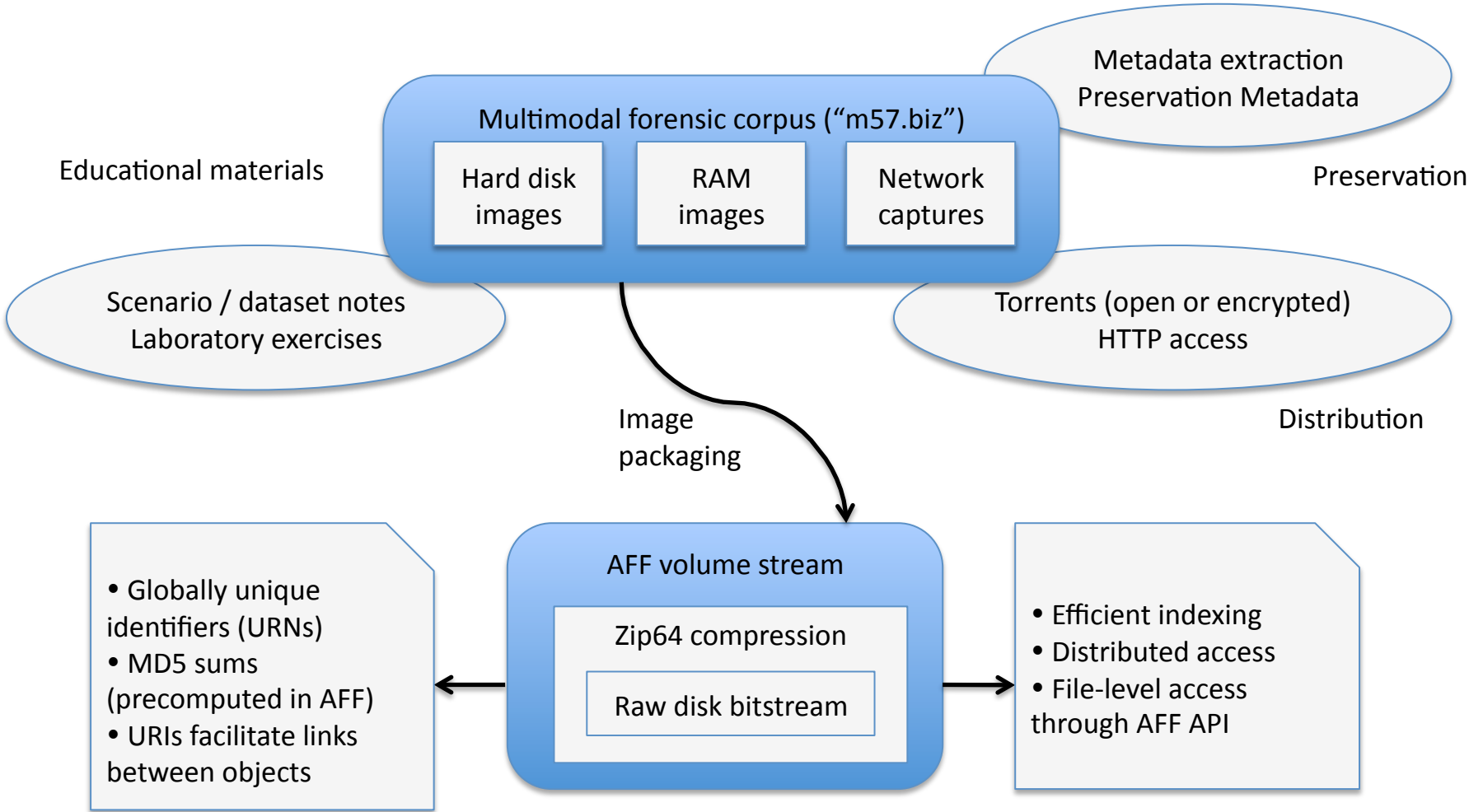
Objectives

- Explore mechanisms for archival storage and distribution of large disk image and multi-modal corpora.
- Develop educational materials (laboratory exercises) to support classroom use of a realistic forensic corpus
- Develop paths to construct preservation and descriptive metadata from metadata contained in both disk images and forensic disk image wrapper formats.

The m57.biz Realistic Corpus

- Multi-modal “realistic” forensic corpus (disk images, RAM images, network captures)
- Scenario acted out with four primary personas on dedicated hardware:
 - 5 PCs, 1 router, 4 USB drives, 1 cellphone, 17 days
- Specific activities of interest: data exfiltration, hardware theft, illegal digital materials
- Over 600GB of raw data
 - Hard drive images packaged with Advanced Forensic Format 4
 - object-oriented architecture
 - persistent identifiers for digital objects across AFF4 containers
 - fine-grained management of object attribute and relationship information
 - scales well, open source libraries for access, manipulation

Overview



Educational Materials

- Realistic corpora are sufficiently complex to support meaningful learning experiences but less “noisy” than real-world data
 - Item-oriented: file-carving, grep, tool use (FTK, EnCase, bulkextractor, SleuthKit, Volatility, TCPflow, wireshark)
 - Chain-of-event reconstruction
 - Tool development (data views, image comparison)

AFF4 Imaging and Metadata

- Drive images, acquisition

```
kamwoods@FWS309:~/Research/M57_Scenario/drives-redacted$ affinfo charlie-2009-11-12.aff
charlie-2009-11-12.aff is a AFF file
```

```
charlie-2009-11-12.aff
[skipping data segments]
```

Segment	arg	data length	data
=====	=====	=====	=====
badflag	0	512	BAD SECTOR.....bI...JI..... ..f.
badsectors	2	8	= 0 (64-bit value)
afflib_version	0	7	"3.5.6"
creator	0	9	afconvert
aff_file_type	0	3	AFF
acquisition_commandline	0	1665	afconvert charlie-2009-11-12.raw
pagesize	16777216	0	
sectorsize	512	0	
imagesize	2	8	= 10239860736 (64-bit value)
md5	0	16	0609 2DFE AA4F B183 946F 95D8 AD84 519E
sha1	0	20	4B44 F1D8 B35E 212B 9330 C89B A7D9 319C EB89 5F75
image_gid	0	16	3535 B83B 747B 8F68 859D 6159 64DE BE6A
acquisition_date	0	20	2010-05-18 14:33:46.

Additional info on acquisition environment

AFF4 Imaging and Metadata

- Drive images, file level: AFF specific DTD with DC elms

```
<!-- NTFS and attr=0x1005169b0 -->
  <libmagic>PE32 executable for MS Windows (native) Intel 80386 32-bit </libmagic>
  <byte_runs>
    <run file_offset='0' fs_offset='230068224' img_offset='230100480' len='32768' />
    <run file_offset='32768' fs_offset='190132224' img_offset='190164480' len='77824' />
    <run file_offset='110592' fs_offset='939683840' img_offset='939716096' len='15424' />
  </byte_runs>
  <hashdigest type='MD5'>32e5e7f33f6a414894ad70cacff45db6</hashdigest>
  <hashdigest type='SHA1'>47e429123ddc5c1c1c3e57b43b1bb1e014c19ce4</hashdigest>
</fileobject>
<fileobject>
  <filename>dell/drivers/R66787/Win2K/intelnic.dll</filename>
  <partition>1</partition>
  <id>108</id>
  <name_type>r</name_type>
  <filesize>24064</filesize>
...
...
```

Distribution

- Torrent packaging of disk images, memory dumps, and network captures
 - High availability as the user base grows
 - AFF images can be verified locally for integrity
 - Encryption and key support in format provides facility to distribute educational materials securely – lessons, answer keys, checklists

Implications for Digital Curation R & D

- Repository and architecture characteristics
- Metadata conventions and integration
- Use of such collections for education of digital curation professionals

Acknowledgements

- Thanks to Simson Garfinkel and the IT and systems administration professionals at the Naval Postgraduate school.
- This work is administered through a sub-grant of “Creating Realistic Forensic Corpora for Undergraduate Education and Research” (NSF Award DUE-0919593) led by Simson Garfinkel of the Naval Postgraduate School.

Questions?

- Garfinkel, S. "Digital forensics research: The next 10 years." DFRWS 2010
- M. I. Cohen, S. Garfinkel and B. Schatz. "Extending the Advanced Forensic Format to accommodate Multiple Data Sources, Logical Evidence, Arbitrary Information and Forensic Workflow." DFRWS 2009
- Elford, D., N. Del Pozo, S. Mihajlovic, D. Pearson, G. Clifton, and C. Webb. "Media Matters: Developing Processes for Preserving Digital Objects on Physical Carriers at the National Library of Australia." 74th IFLA General Conference and Council, Québec, Canada, August 10-14 2008.
- Garfinkel, Farrell, Roussev and Dinolt. "Bringing Science to Digital Forensics with Standardized Forensic Corpora" DFRWS 2009
- John, Jeremy Leighton, Ian Rowlands, Peter Williams, and Katrina Dean. "Digital Lives: Personal Digital Archives for the 21st Century >> an Initial Synthesis." 2010.
- Kirschenbaum, M. G., E. Farr, K. M. Kraus, N. L. Nelson, C. S. Peters, G. Redwine, and D. Reside. "Approaches to Managing and Collecting Born-Digital Literary Materials for Scholarly Use." College Park, MD: University of Maryland, 2009.
- Woods, K., and G. Brown. "From Imaging to Access - Effective Preservation of Legacy Removable Media." In Archiving 2009, 213-18. Springfield, VA: Society for Imaging Science and Technology, 2009.