# Taking Bitstreams Seriously: Digital Forensics and the BitCurator Environment

Cal Lee
School of Information and Library Science
University of North Carolina, Chapel Hill

July 23-25, 2013
Alexandria, Virginia







## Two Main Acquisition Paths

Systematic Transfer	Dealing with Whatever you Get
Close pre-coordination between Producer and Archive*	Little pre-coordination between Producer and Archive
Archive has (at least some) say in how materials are produced, packaged and transferred	Archive has relatively little say in how materials are produced, packaged and transferred
Relatively little need to engage in guess work or ad hoc description after the transfer	Substantial need to engage in guess work or ad hoc description after the transfer

<sup>\*</sup>Archive in the OAIS sense – can be library, archives, museum, data center, ...

## Examples of Systematic Transfer Developments

- Protocols and tools for transfer (e.g. BagIt)<sup>1</sup>
- Systematic and predictable ingest workflows<sup>2</sup>
- Substantial, well-documented Producer-Archive interactions (e.g. PAIMAS)<sup>3</sup>

<sup>1. &</sup>quot;BagIt: Transferring Content for Digital Preservation" (Video). Library of Congress, August 6, 2009. http://www.youtube.com/watch?v=I3p3ao\_JSfo

<sup>2.</sup> Glick, Kevin, and Eliot Wilczek. "Ingest Guide." Tufts University and Yale University, 2006. http://dca.lib.tufts.edu/features/nhprc/reports/3\_1\_draftpublic3.pdf

<sup>3.</sup> Producer-Archive Interface Methodology Abstract Standard (PAIMAS). CCSDS 651.0-M-1. Consultative Committee for Space Data Systems. May 2004.

## Sometimes things are a little messier



## Applying Digital Forensics to Digital Collections – Previous Work\*

- Ross and Gow (1999) potential relevance of advances in data recovery and digital forensics to collecting institutions
- More recently active stream of literature related to use of forensic tools and methods for digital collections, including activities at the British Library, National Library of Australia and Indiana University
- PERPOS (Georgia Tech) has applied data capture and extraction to US presidential materials
- "Computer Forensics and Born-Digital Content in Cultural Heritage Collections" symposium and report (2010)
- Born Digital Collections: An Inter-Institutional Model for Stewardship (AIMS) framework for the stewardship of born-digital materials, including digital forensics methods
- Digital Records Forensics project has articulated connections between the concepts of digital forensics and archival science
- Two Open Planets Foundation (OPF) Hackathons this year one in Copenhagen, one in Chapel Hill

<sup>\*</sup>See citations in: Lee, Christopher A. "<u>Archival Application of Digital Forensics Methods for Authenticity, Description and Access Provision</u>." In *Proceedings of the International Council on Archives Congress, Brisbane, Australia, August 20-24, 2012.* 

## BitCurater

- Funded by Andrew W. Mellon Foundation
  - Phase 1: October 1, 2011 September 30, 2013
  - Phase 2 October 1, 2013 September 30, 2014
- Partners: School of Information and Library
   Science (SILS) at UNC and Maryland Institute for
   Technology in the Humanities (MITH)

## Core BitCurator Team

- Cal Lee, Pl
- Matt Kirschenbaum, Co-Pl
- Kam Woods, Technical Lead
- Alex Chassonoff, Project Manager (UNC)
- Sunitha Misra, GA (UNC)
- Porter Olsen, GA (MITH)









### Two Groups of Advisors

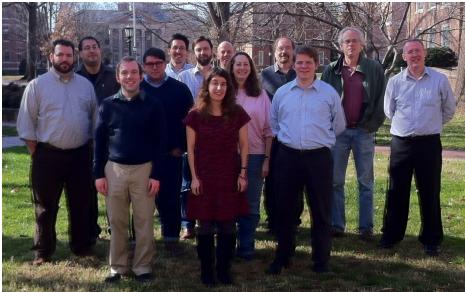
#### **Professional Experts Panel**

- Bradley Daigle, University of Virginia Library
- Erika Farr, Emory University
- Jennie Levine Knies, University of Maryland
- Jeremy Leighton John, British Library
- Leslie Johnston, Library of Congress
- Naomi Nelson, Duke University
- Erin O'Meara, Gates Archive
- Michael Olson, Stanford University Libraries
- Gabriela Redwine, Harry Ransom Center, University of Texas
- Susan Thomas, Bodleian Library, University of Oxford

#### **Development Advisory Group**

- Barbara Guttman, National Institute of Standards and Technology
- Jerome McDonough, University of Illinois
- Mark Matienzo, Yale University
- Courtney Mumma, Artefactual Systems
- David Pearson, National Library of Australia
- Doug Reside, New York Public Library
- Seth Shaw, University Archives, Duke University
- William Underwood, Georgia Tech





## **BitCurator Goals**

- Develop a system for collecting professionals that incorporates the functionality of opensource digital forensics tools
- Address two fundamental needs not usually addressed by the digital forensics industry:
  - incorporation into the workflow of archives/library ingest and collection management environments
  - provision of public access to the data

### **BitCurator Environment**

 Bundles, integrates and extends functionality (primarily data capture and reporting) of open source software: fiwalk, bulk extractor, Guymager, The Sleuth Kit, sdhash and others

#### Can be run as:

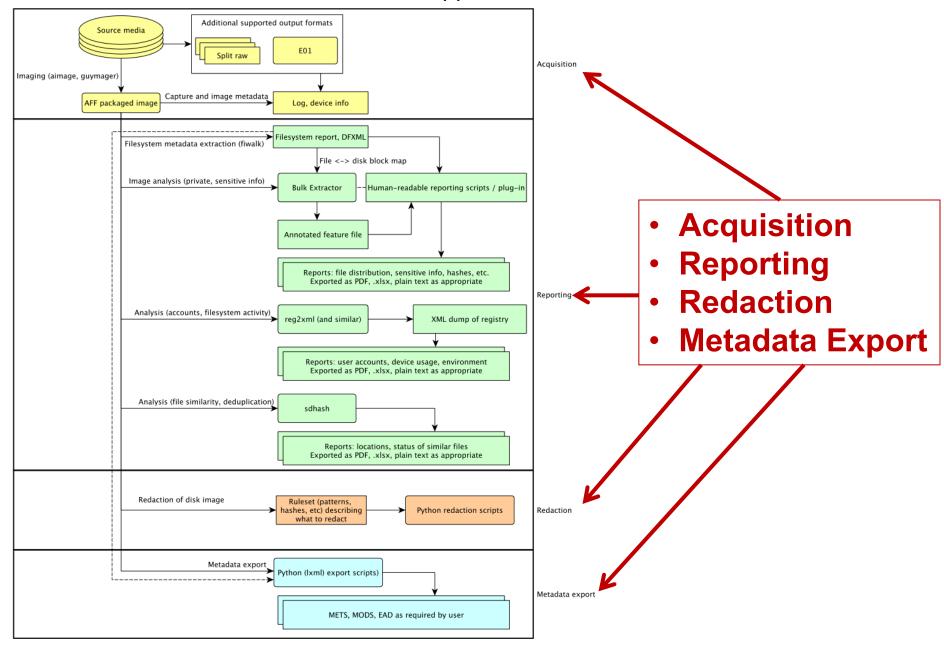
- Self-contained environment (based on Ubuntu Linux) running directly on a computer (download installation ISO)
- Self-contained Linux environment in a virtual machine using VirtualBox
- As individual components run directly in your own Linux environment or (whenever possible) Windows environment

## Acknowledgement to Simson Garfinkel

- Digital forensics scholar at Naval Postgraduate School
- Responsible for:
  - fiwalk
  - Bulk Extractor
  - Digital Forensics XML (DFXML) metadata conventions
  - forensicswiki.org
  - digitalcorpora.org

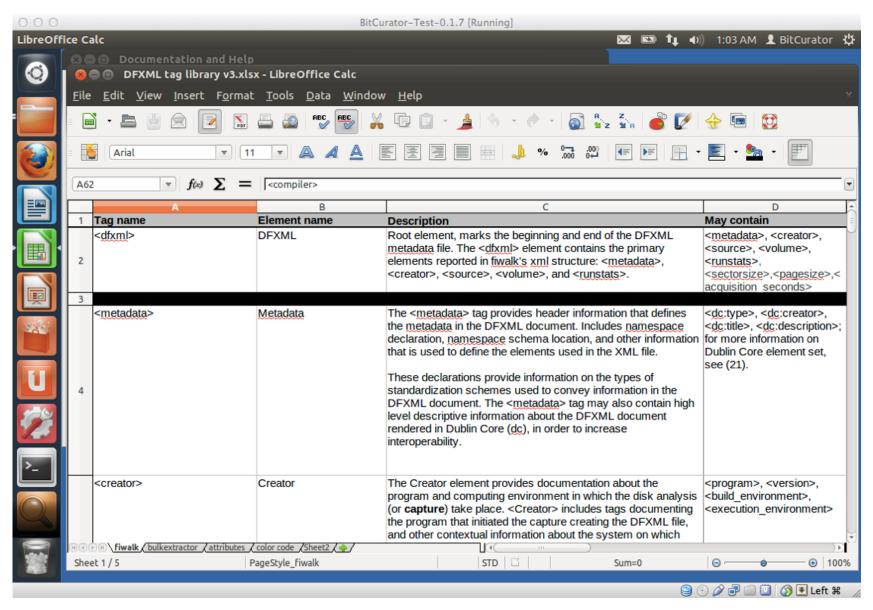


#### BitCurator-Supported Workflow



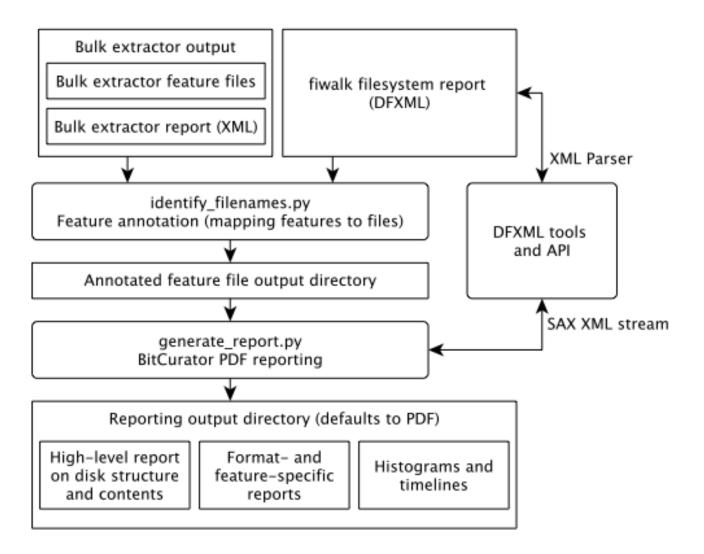
See: http://bitcurator.net

#### Metadata Conventions of the BitCurator Tools: Digital Forensics XML (DFXML)



http://www.bitcurator.net/2013/02/06/dfxml-tag-library/

## Metadata Generation and Reporting



See: Woods, Kam, Christopher Lee, and Sunitha Misra. "Automated Analysis and Visualization of Disk Images and File Systems for Preservation." In *Proceedings of Archiving 2013* (Springfield, VA: Society for Imaging Science and Technology, 2013), 239-244.





**Imaging Tools** 



















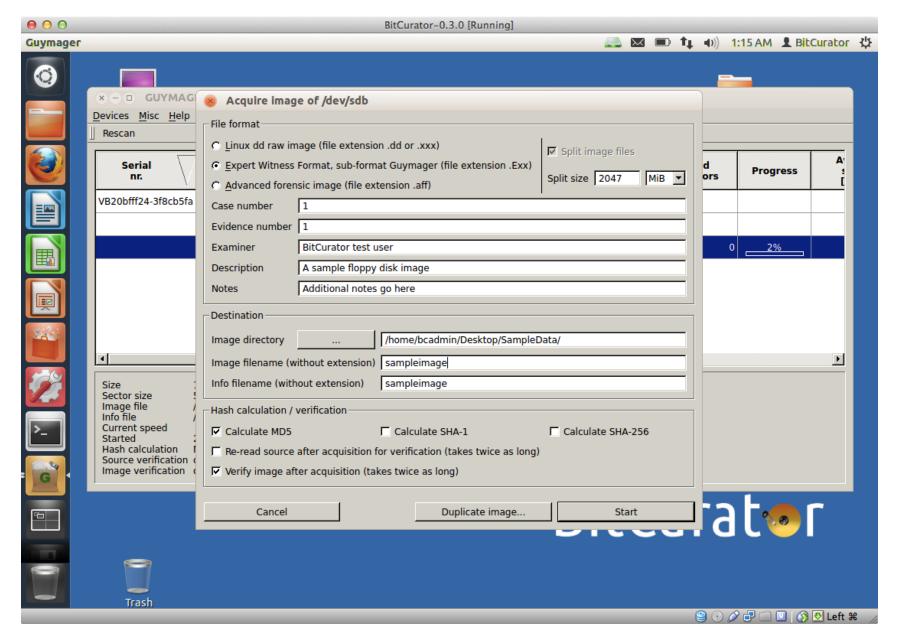




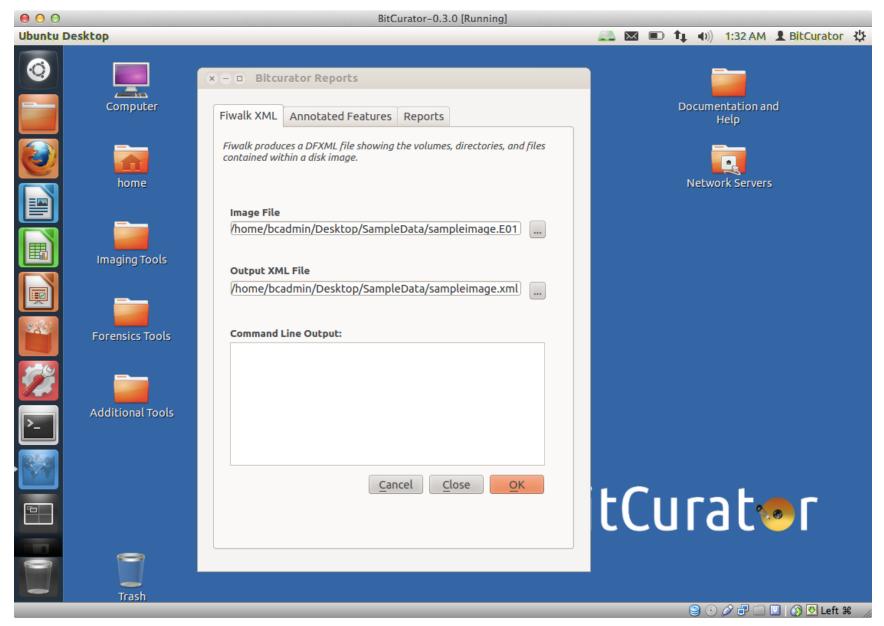




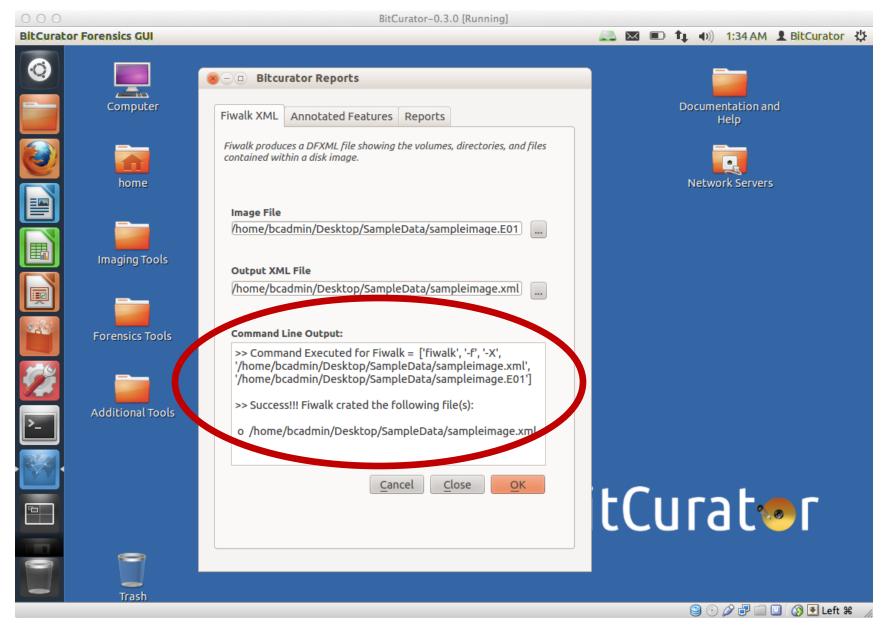
## Acquiring Disk Images with Guymager



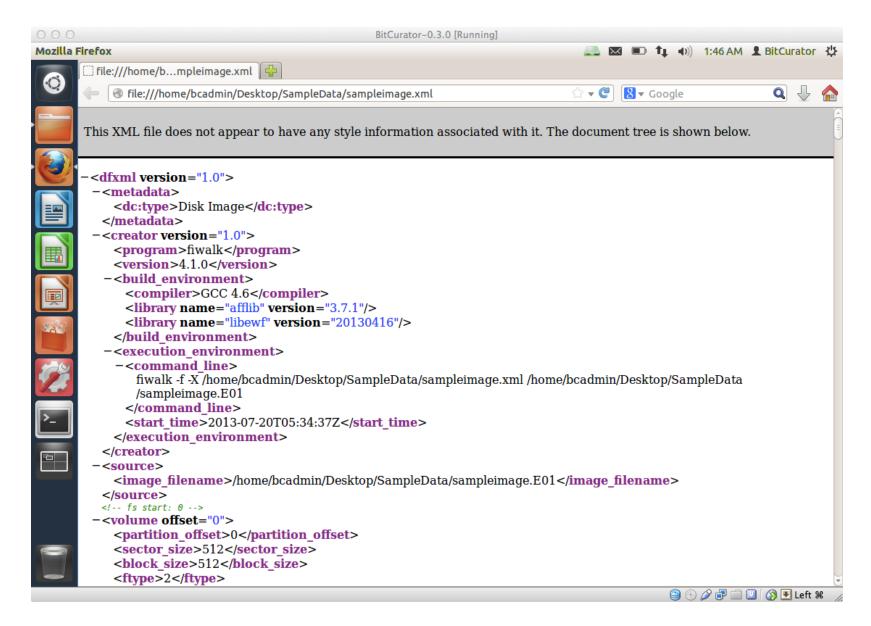
## **Exporting Filesystem Content Using fiwalk**



## Viewing the Command Line Output



## DFXML Output from fiwalk

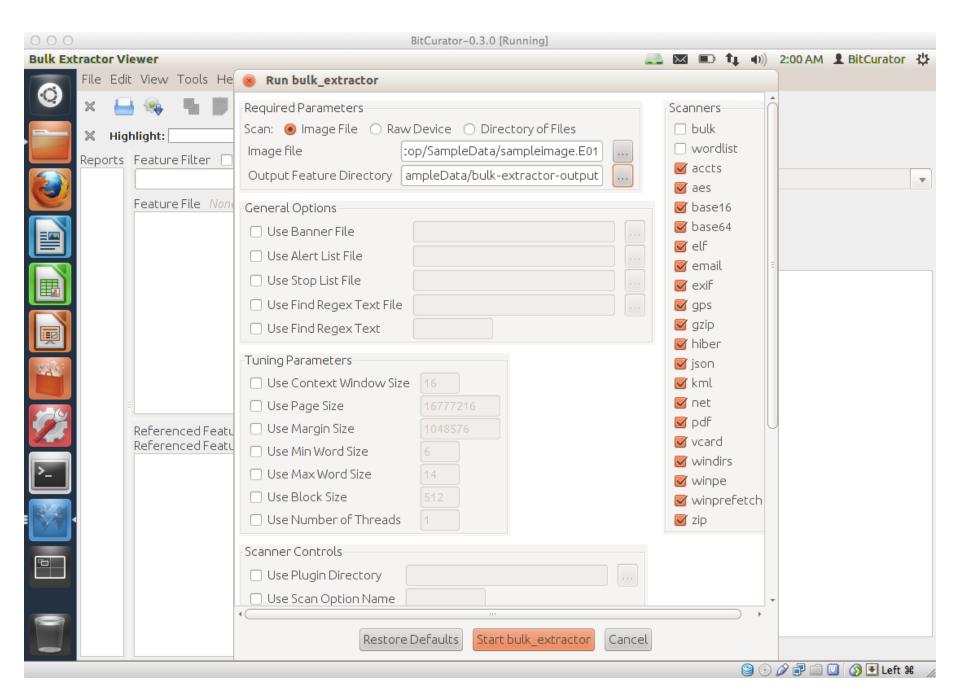


## DFXML for a Specific File

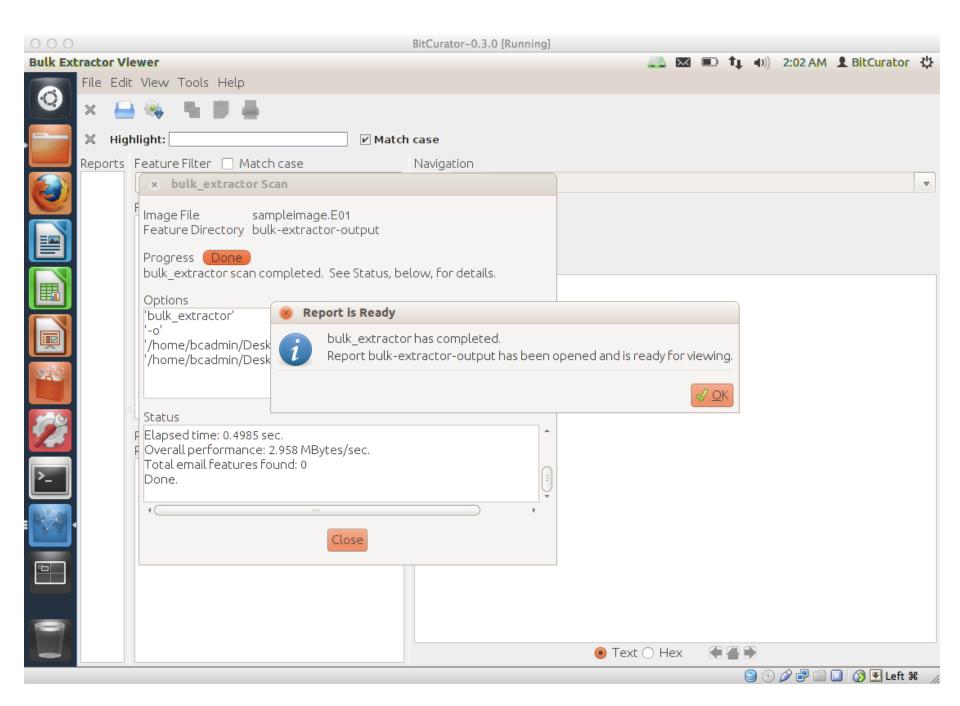
```
<fileobject>
      <filename>Documents and Settings/All Users/Documents/
                 My Pictures/Sample Pictures/Blue hills.jpg
      </filename>
      <filesize>28521</filesize>
      <alloc>1</alloc>
      <used>1</used>
      <inode>6245</inode>
      <uid>0</uid>
      <qid>0</qid>
      <mtime>1208174400</mtime>
      <ctime>1257729636</ctime>
      <atime>1257729636</atime>
      <crtime>1257729636</crtime>
      <seq>2</seq>
      libmagic>JPEG image data, JFIF standard 1.02
      <br/>
<br/>
te runs>
       <run file_offset='0' fs_offset='0' img_offset='363200512'</pre>
         len='0'/>
      </byte runs>
      <hashdigest type='MD5'>
          6fb2a38dc107eacb41cf1656e899cf70
      </hashdigest>
      <hashdigest type='SHA1'>
          4eee44b18576e84de7b163142b537d2fe6231845
      </hashdigest>
</fileobject>
```

## Identifying "Features" of Interest in Disk Images or Directories

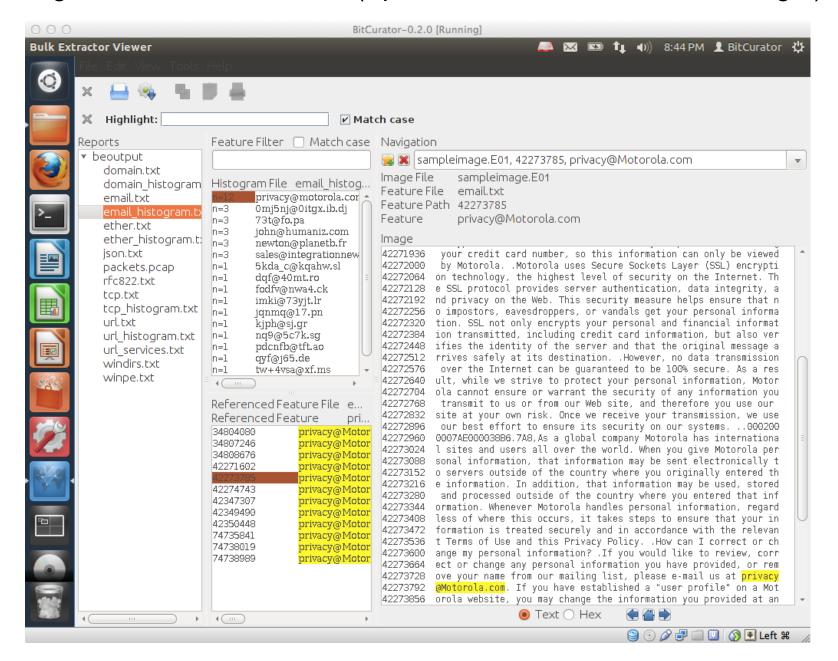
**Bulk Extractor** 



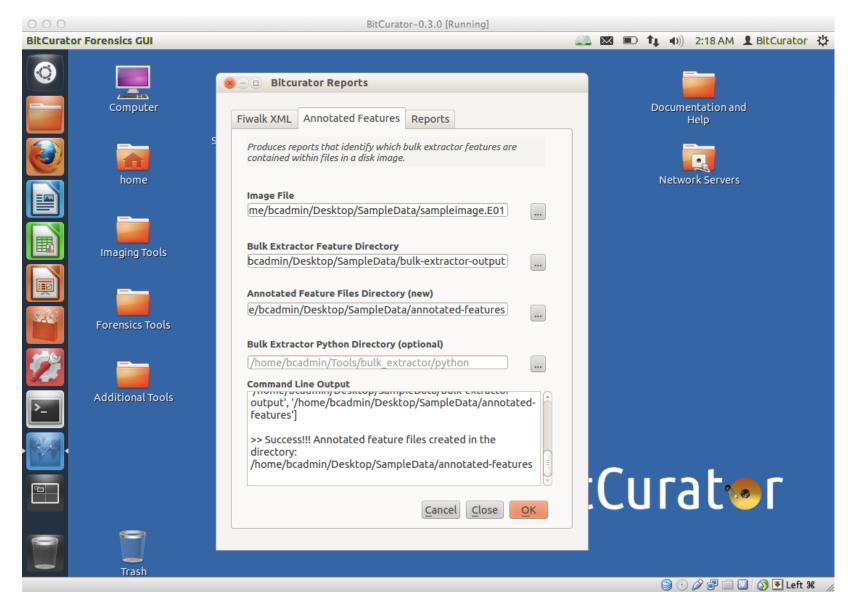
Scanner	Description	
scan-accts	Looks for phone numbers, credit card numbers, etc	② ■ ↑ (i)) 2:00 AM L BitCurator U
scan_base64	Decodes BASE64 text	Scanners bulk
scan_kml	Detects KML (Keyhole Markup Language) files – used to identify geographic locations	wordlist  ✓ accts ✓ aes
scan_gps	Detects XML from Garmin GPS devices	✓ base16 ✓ base64
scan_aes	Detects in-memory AES (Advanced Encryption Standard) keys from the key schedules	elf  i email  i exif
scan_json	Detects JavaScript Object Notation files	gps  ✓ gzip
scan_exif	Detects EXIF structures from JPEG files	<ul><li>✓ hiber</li><li>✓ ison</li></ul>
scan_zip	Detects and decompresses ZIP files and zlib streams	✓ kml ✓ net ✓ pdf
scan_gzip	Detects and decompresses GZIP files and gzip streams	<ul><li>✓ vcard</li><li>✓ windirs</li><li>✓ winpe</li></ul>
scan_pdf	Extracts text from some kinds of PDF files	<ul><li>✓ winptefetch</li><li>✓ zip</li></ul>
scan_hiber	Detects and decompresses Windows hibernation file fragments	
scan_winprefetch	Detects and extracts fields from windows prefetch fields from Windows prefetch files and file fragments	Cancel   ② ⊙ ♪ □ □ ③ • Left ₩



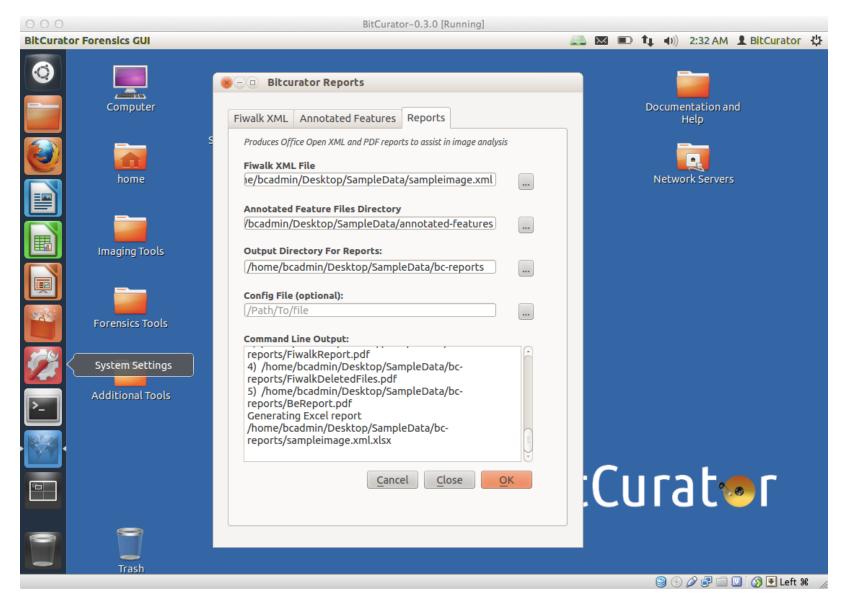
#### Histogram of Email Addresses (Specific Instances in Context on Right)

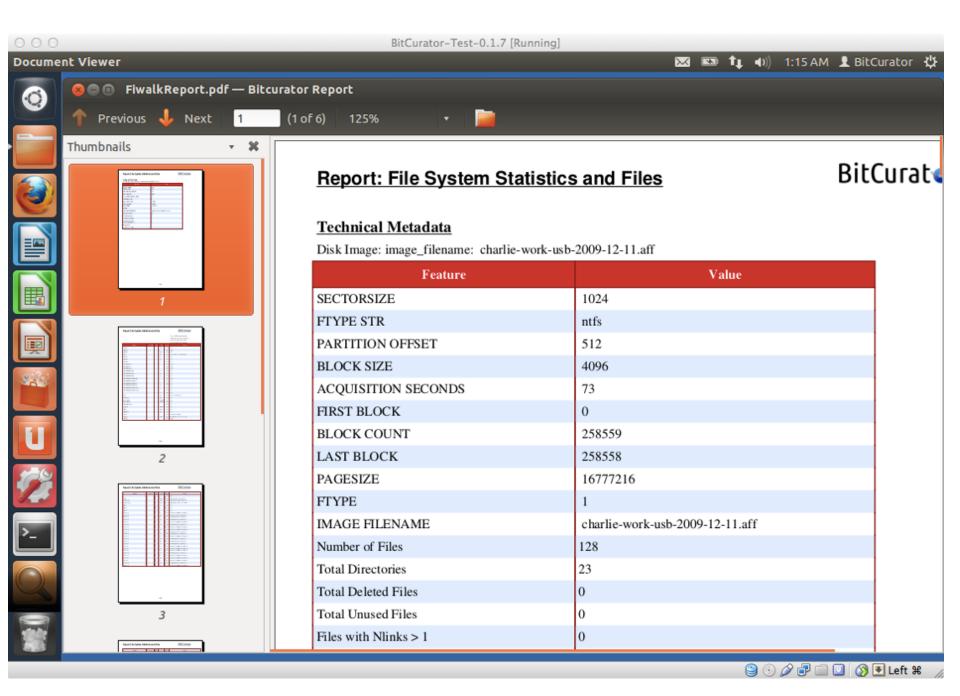


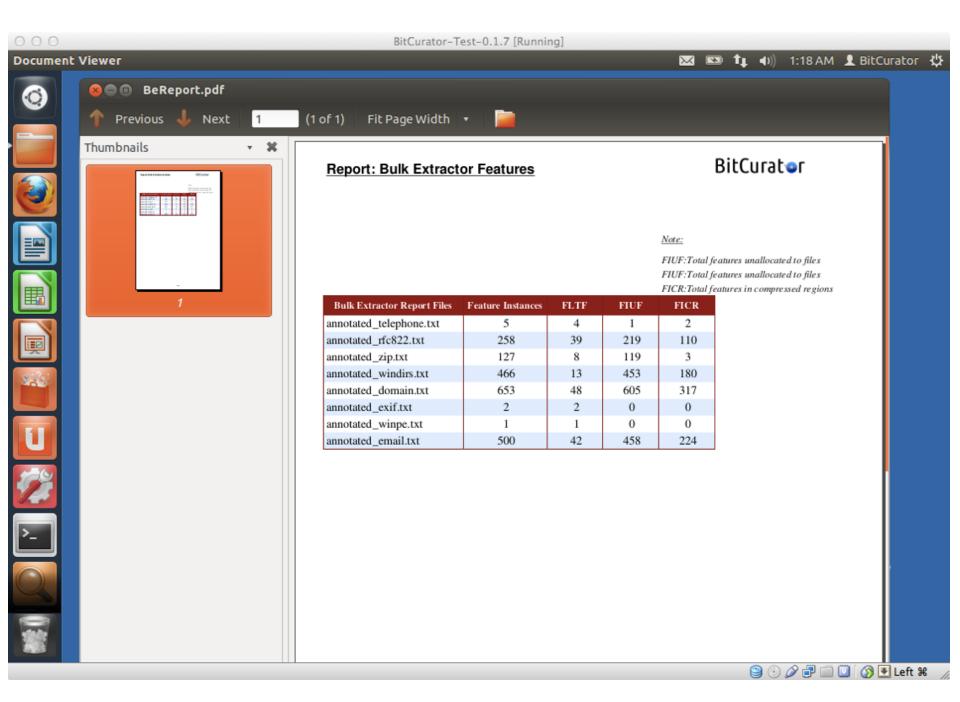
## Matching Bulk Extractor Output (Based on Byte Offsets) to fiwalk Output (Based on Filesystem Location)

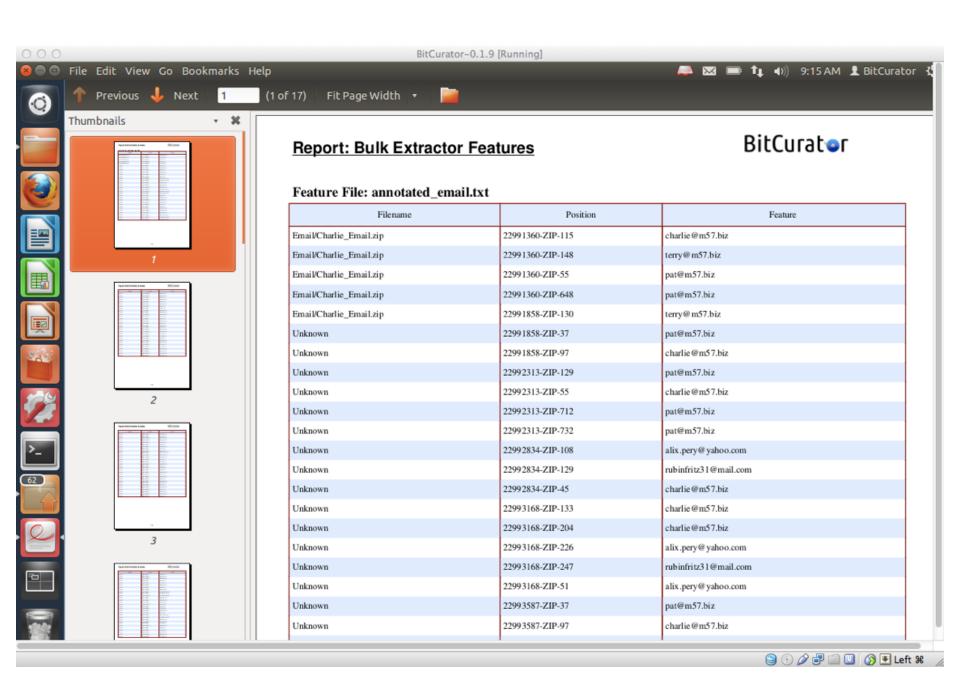


## **Generating BitCurator Reports**





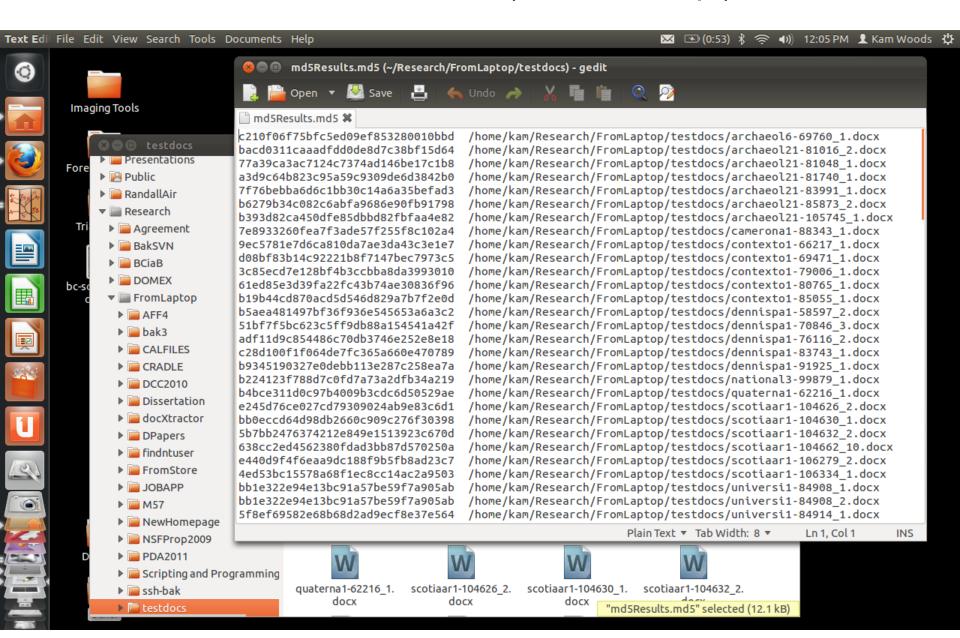




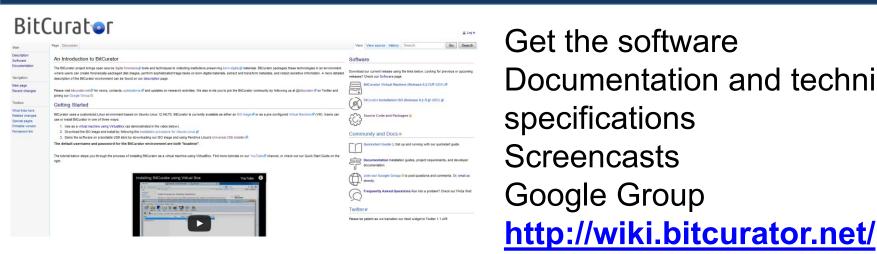
## Nautilus Scripts

- Scripts that can be run using the GNOME file manager called Nautilus (Linux analog to Windows Explorer or Mac OS X Finder)
- Can be used in the BitCurator environment or your own Linux environment

#### MD5 Hashes of Files (Nautilus Script)



#### BitCurator Tools and Further Information

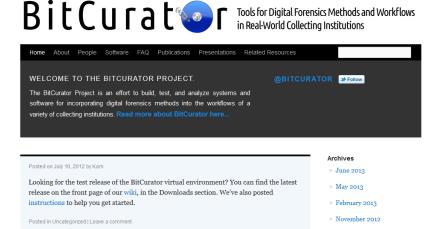


Tools for Digital Forensics Methods and Workflows

October 2012 September 2012

August 2012

Get the software Documentation and technical specifications Screencasts Google Group



Marty Gengenbach selected for National Digital Stewardship

Alliance Innovation Award

People **Project overview Publications** News http://www.bitcurator.net/

Twitter: @bitcurator