Parametric Curation in Digital Archives: Concept and Potential Benefits

Cal Lee

School of Information and Library Science

University of North Carolina, Chapel Hill

Society of American Archivists Research Forum

August 13, 2013

New Orleans, Louisiana

BitCurator

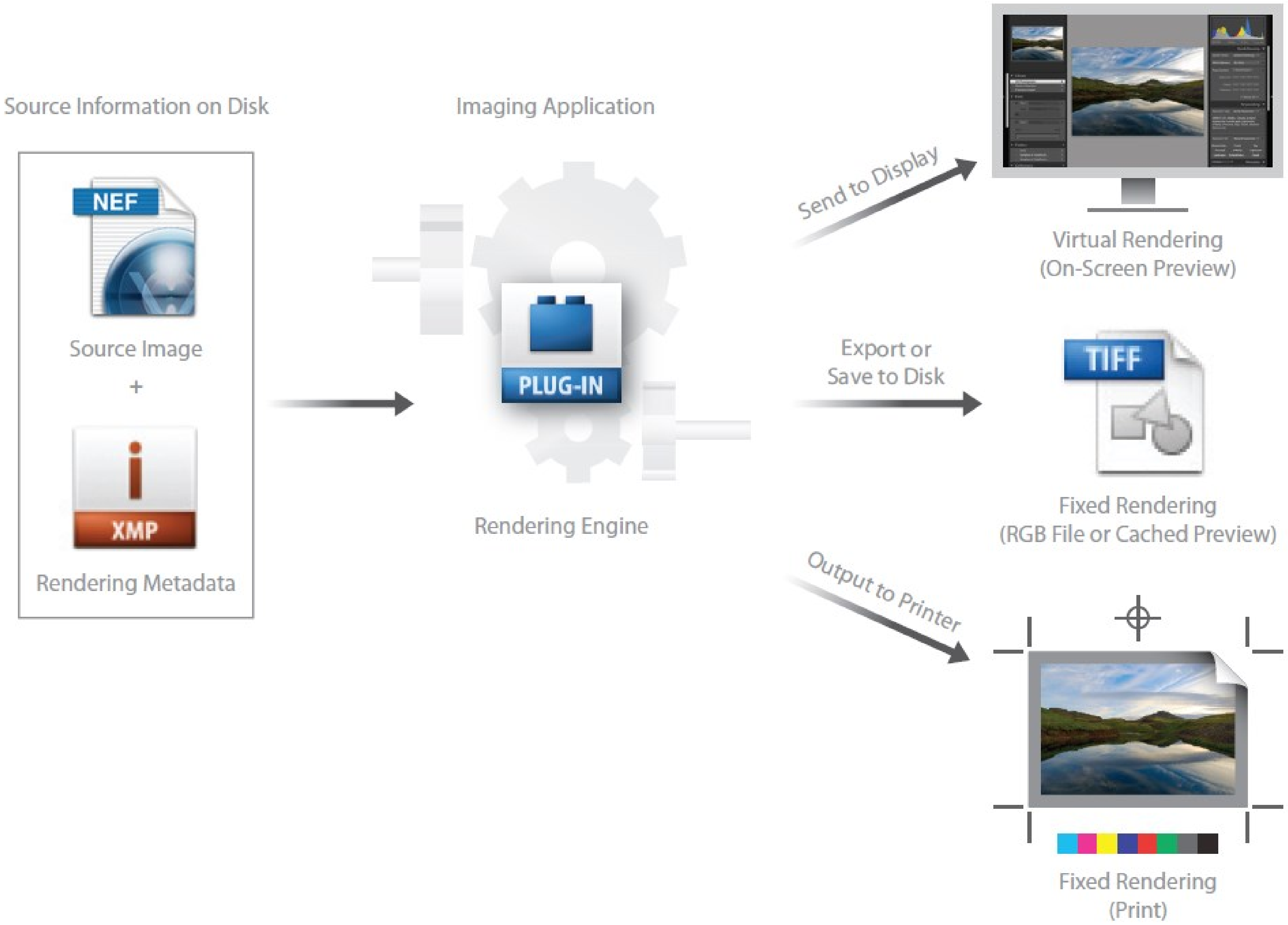
UNC SCHOOL OF INFORMATION AND LIBRARY SCIENCE

Acknowledging My Co-Authors

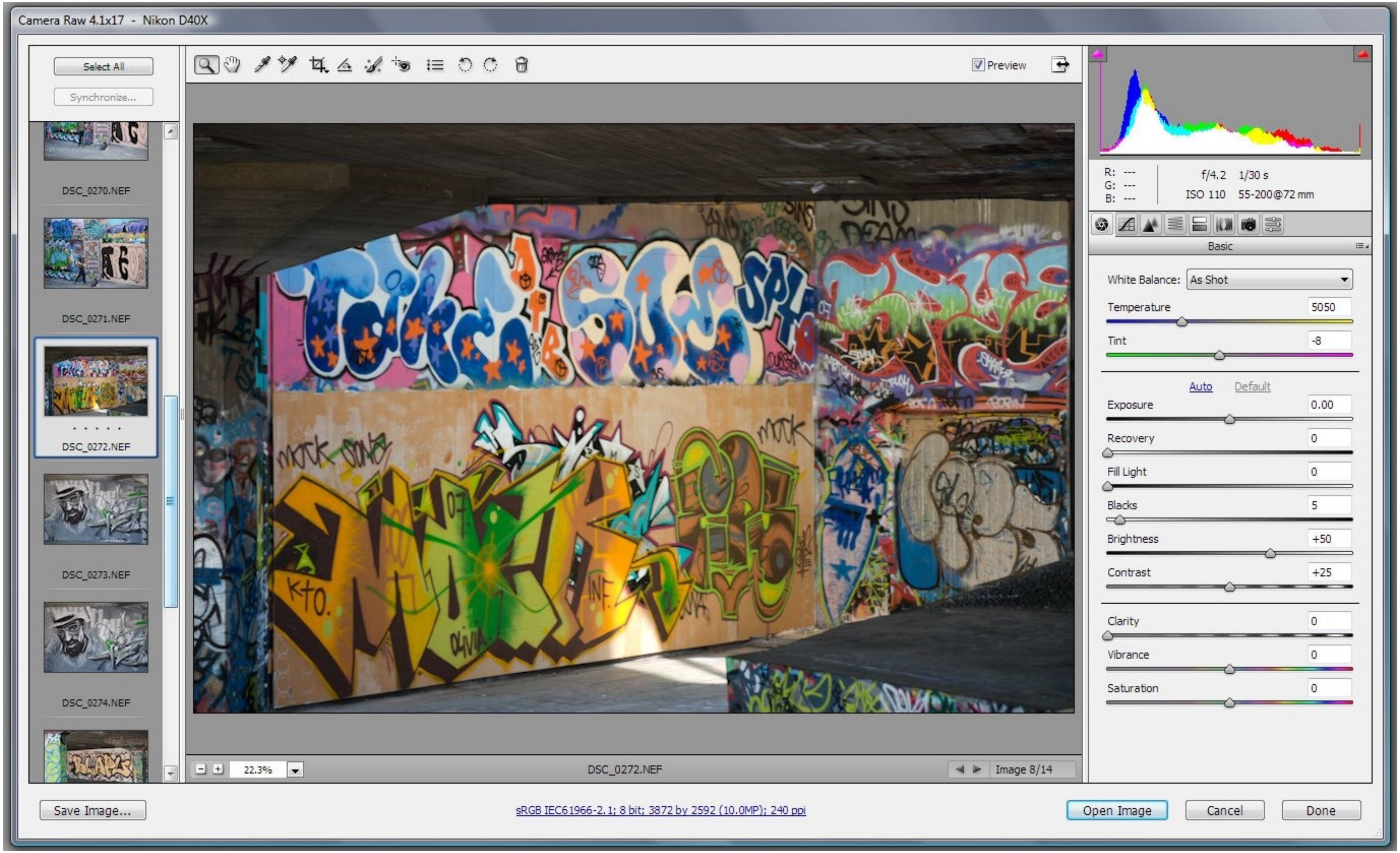
- Jeremy Leighton John, British Library - introduced the idea of parametric curation
- Kam Woods, University of North Carolina

Parametric Image Editing (PIE)

- Photographers used to produce numerous variant files from a given master as they experimented with editing options (e.g. white balance, cropping, tweaked color profiles).
- Now common practice for professional and amateur photographers to use “raw” formats – editing is handled through the recording of metadata about changes made to a single master file.
- Parametric image editing (PIE) allows for nondestructive and reversible editing of digital photographs - avoids excessive and inefficient management and storage of files.
- If an earlier version of a file is required, it is simply restored using the metadata.



Krogh, Peter. "Non-Destructive Imaging:An Evolution of Rendering Technology." San Jose, CA: Adobe Systems Incorporated, 2007. Figure 9.



Camera Raw 4.1x17 - Nikon D40X

R: ---
G: ---
B: ---

f/4.2 1/30 s
ISO 110 55-200@72 mm

White Balance: As Shot

Temperature 5050

Tint -8

Exposure 0.00

Recovery 0

Fill Light 0

Blacks 5

Brightness +50

Contrast +25

Clarity 0

Vibrance 0

Saturation 0

22.3%

DSC_0272.NEF

Image 8/14

sRGB IEC61966-2.1; 8 bit; 3872 by 2592 (10.0MP); 240 ppi

Ron G, “Adobe Camera Raw for Windows 7.1 RC released,” May 7, 2012, http://www.winbeta.org/news/adobe-camera-raw-windows-71-rc-released

Parametric Curation

- Use metadata to record changes made rather than:
  - Making irreversible changes to the underlying data
  - Unduly replicating identical or similar information
- Transformations and views into data can reflect current needs (e.g. migration on request[1])
- Can apply "permission overlays" to implement appropriate access permissions[2] and redact as needed
- Consider e.g. versions of a document in Subversion (SVN) – stored as a single original file along with all of the "diffs" between that original and later versions

[1] Mellor, Phil, Paul Wheatley and Derek Sergeant. "Migration on Request, a Practical Technique for Preservation." In *Research and Advanced Technology for Digital Libraries: 6th European Conference, ECDL 2002, Rome, Italy, September 16-18, 2002: Proceedings, edited by Maristella Agosti and C. Thanos, 2458, 516-526. Berlin: Springer, 2002.*

[2] Thanks to *Geoffrey Brown (Indiana University) for this idea*

One Approach to Parametric Curation – Forensic Disk Images

- Capture and retention of forensic disk images, which provide exact copies of all sectors on storage media.
- Files within the disk image may be retained *in situ* for future export as and when necessary.
- Disk image could be considered the primary preservation object (in an AIP) and/or extracted files can treated as preservation objects
- Redaction, reorganization and description could be represented with metadata, rather than changing the disk image itself

Getting below the File System – Low-Level Copying

- Getting an “image” of a storage medium involves working at a level below the file system
    - Can get at file attributes and deleted files not visible through higher-level copy operations
- Most commonly used tool is dd (or variant) - UNIX program for low-level copying and conversion of data from a storage device
- More specialized tools for creating forensic images include:
    - FTK Imager
    - Guymager
    - Imaging utilities in commercial applications (including EnCase and FTK)

BitCurator

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

BitCurator Goals

- Develop a system for collecting professionals that incorporates the functionality of open-source 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

Core BitCurator Team

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

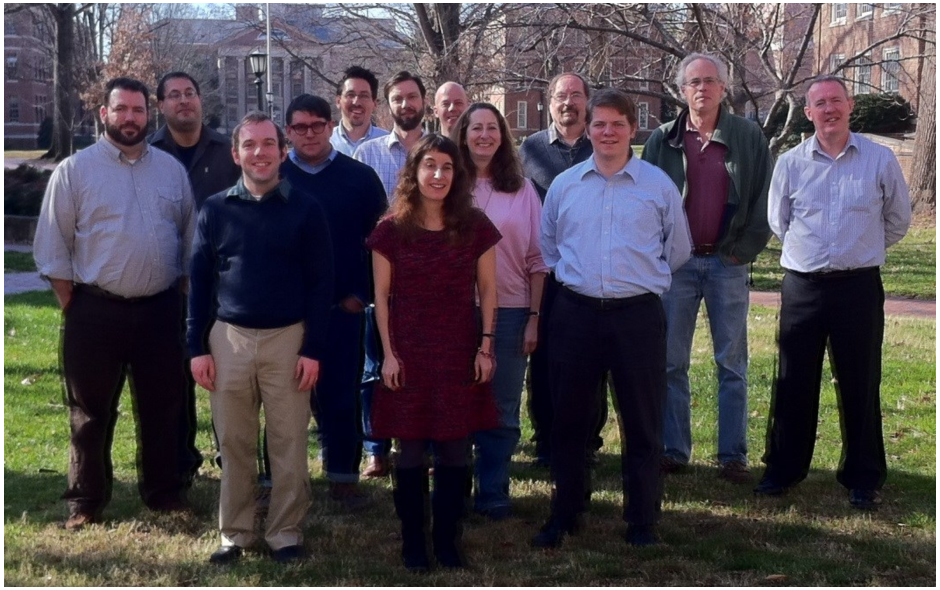
Two Groups of Advisors

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

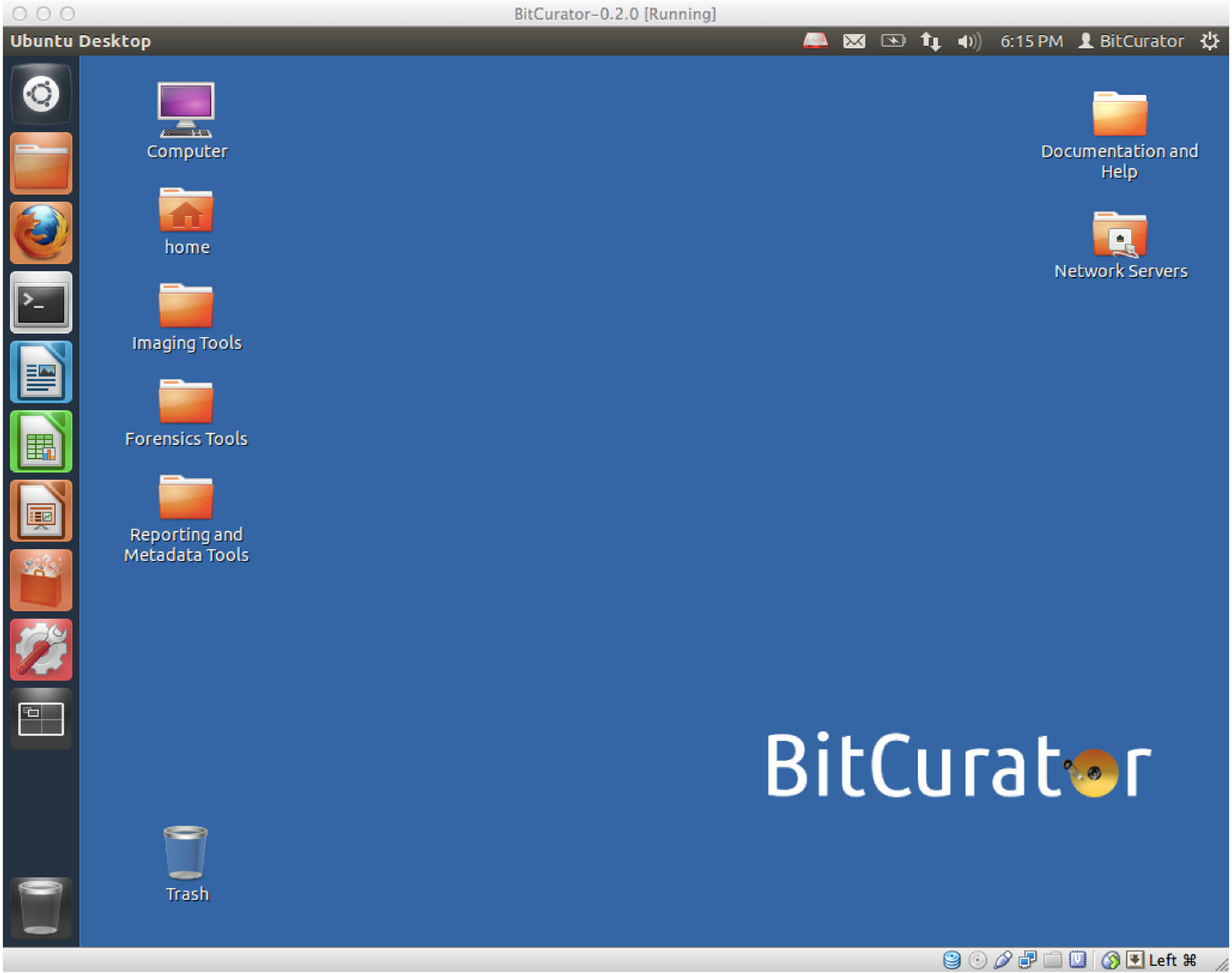
- Barbara Guttman, National Institute of Standards and Technology
- Jerome McDonough, University of Illinois
- Mark Matienzo, Yale University
- Courtney Mumma, Artefactual Systems
- David Pearson, National Library of Australia
- Doug Reside, New York Public Library
- Seth Shaw, University Archives, Duke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



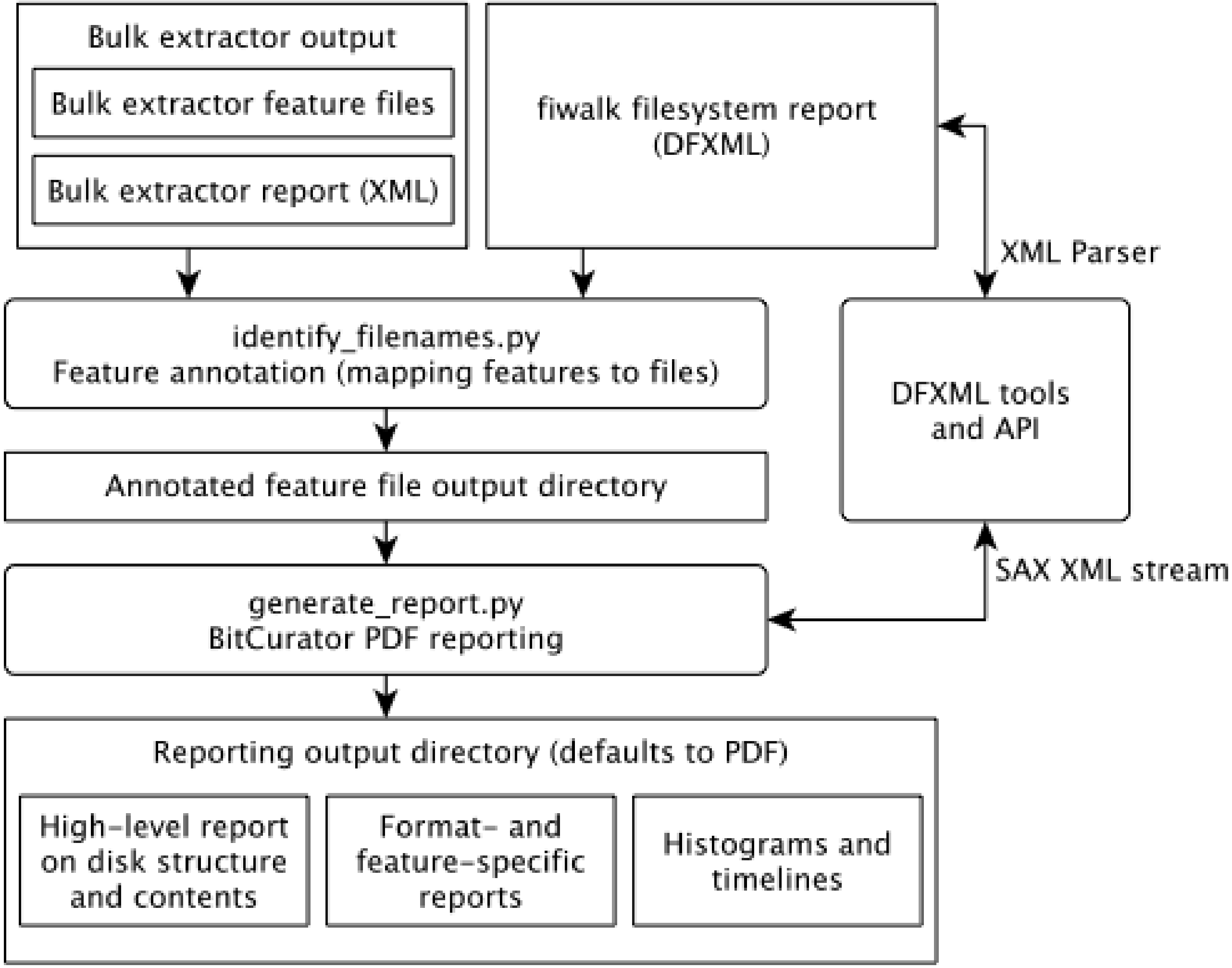
BitCurator Environment*

- 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 e.g. Virtual Box or VMWare
  - As individual components run directly in your own Linux environment or (whenever possible) Windows environment

*To read about and download the environment, see: http://wiki.bitcurator.net/

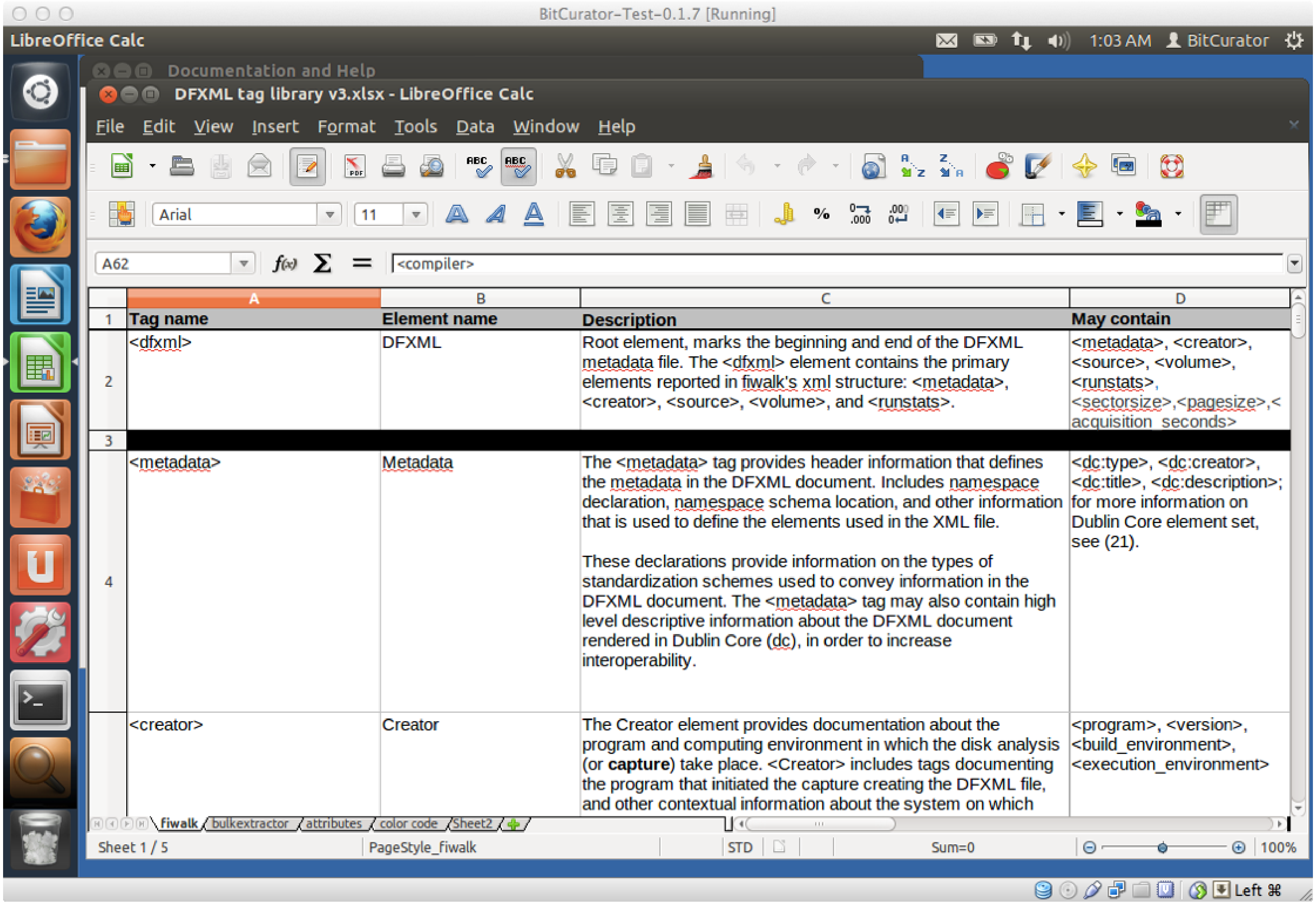

BitCurator-0.2.0 [Running]
Ubuntu Desktop
6:15 PM
BitCurator
Computer
home
Imaging Tools
Forensics Tools
Reporting and Metadata Tools
Trash
Documentation and Help
Network Servers
BitCurator
Left ⌘

High-Level view of Metadata Generation and Reporting


Bulk extractor output
Bulk extractor feature files
Bulk extractor report (XML)
fiwalk filesystem report (DFXML)
XML Parser
identify_filenames.py
Feature annotation (mapping features to files)
DFXML tools and API
Annotated feature file output directory
SAX XML stream
generate_report.py
BitCurator PDF reporting
Reporting output directory (defaults to PDF)
High-level report on disk structure and contents
Format- and feature-specific reports
Histograms and timelines

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.

Documentation of Digital Forensics XML (DFXML) Elements



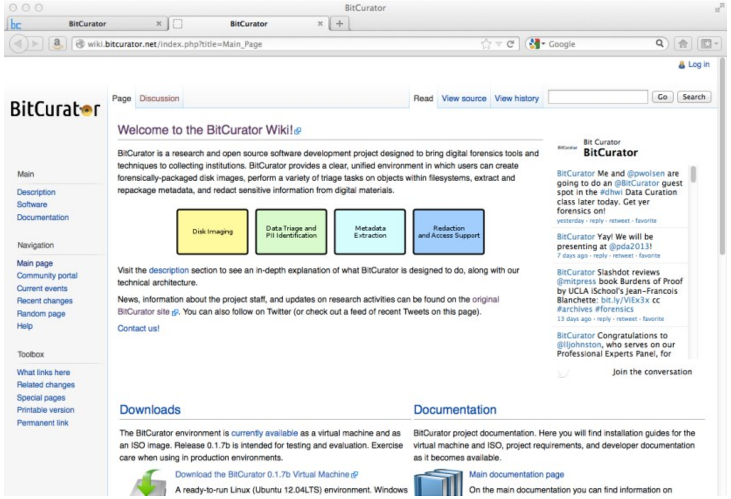
	A	B	C	D
1	**Tag name**	**Element name**	**Description**	**May contain**
2	<dfxml>	DFXML	Root element, marks the beginning and end of the DFXML metadata file. The <dfxml> element contains the primary elements reported in fiwalk's xml structure: <metadata>, <creator>, <source>, <volume>, and <runstats>.	<metadata>, <creator>, <source>, <volume>, <runstats>, <sectorsize>,<pagesize>,<acquisition_seconds>
3				
4	<metadata>	Metadata	The <metadata> tag provides header information that defines the metadata in the DFXML document. Includes namespace declaration, namespace schema location, and other information that is used to define the elements used in the XML file. These declarations provide information on the types of standardization schemes used to convey information in the DFXML document. The <metadata> tag may also contain high level descriptive information about the DFXML document rendered in Dublin Core (dc), in order to increase interoperability.	<dc:type>, <dc:creator>, <dc:title>, <dc:description>; for more information on Dublin Core element set, see (21).
	<creator>	Creator	The Creator element provides documentation about the program and computing environment in which the disk analysis (or **capture**) take place. <Creator> includes tags documenting the program that initiated the capture creating the DFXML file, and other contextual information about the system on which	<program>, <version>, <build_environment>, <execution_environment>

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

Conclusions

- **Not** claiming that the full vision of parametric curation has been realized by the current BitCurator environment
- **Not** claiming that disk images must always be retained when acquiring born-digital materials
- We **are** claiming that a conceptual shift to parametric curation (with PIE as the motivating analogy) has great potential to ensure future use of materials and improve many archival workflows

Thank You!



Welcome to the BitCurator Wiki!

BitCurator is a research and open source software development project designed to bring digital forensics tools and techniques to collecting institutions. BitCurator provides a clear, unified environment in which users can create forensically-packaged disk images, perform a variety of triage tasks on objects within filesystems, extract and repackage metadata, and redact sensitive information from digital materials.

Visit the description section to see an in-depth explanation of what BitCurator is designed to do, along with our technical architecture.

News, information about the project staff, and updates on research activities can be found on the original BitCurator site. You can also follow on Twitter (or check out a feed of recent Tweets on this page).

Contact us!

Bit Curator
BitCurator

BitCurator Me and @pwolsen are going to do an @BitCurator guest spot in the #dhwi Data Curation class later today. Get yer forensics on!
yesterday · reply · retweet · favorite

BitCurator Yay! We will be presenting at @pda2013!
7 days ago · reply · retweet · favorite

BitCurator Slashdot reviews @mitpress book Burdens of Proof by UCLA iSchool's Jean-Francois Blanchette: bit.ly/ViEx3x cc #archives #forensics
13 days ago · reply · retweet · favorite

BitCurator Congratulations to @lljohnston, who serves on our Professional Experts Panel, for

Downloads

The BitCurator environment is currently available as a virtual machine and as an ISO image. Release 0.1.7b is intended for testing and evaluation. Exercise care when using in production environments.

Download the BitCurator 0.1.7b Virtual Machine

A ready-to-run Linux (Ubuntu 12.04LTS) environment. Windows

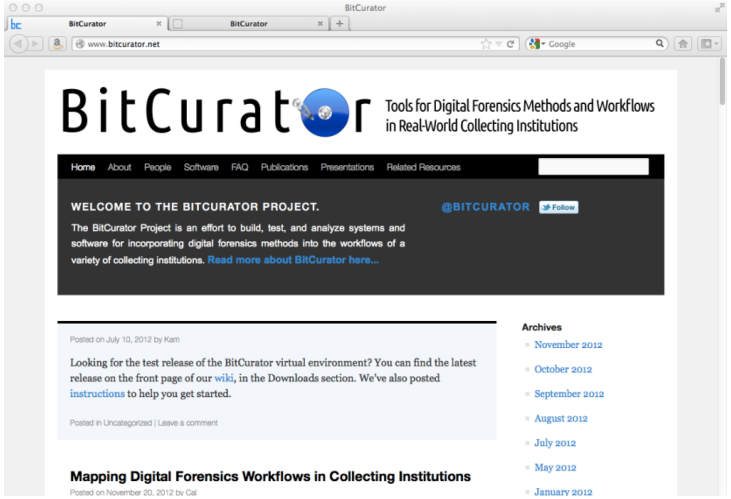
Documentation

BitCurator project documentation. Here you will find installation guides for the virtual machine and ISO, project requirements, and developer documentation as it becomes available.

Main documentation page

On the main documentation you can find information on

Get the software
Documentation and technical specifications
Screencasts
Google Group
http://wiki.bitcurator.net/



WELCOME TO THE BITCURATOR PROJECT.

The BitCurator Project is an effort to build, test, and analyze systems and software for incorporating digital forensics methods into the workflows of a variety of collecting institutions. Read more about BitCurator here...

Posted on July 10, 2012 by Kam

Looking for the test release of the BitCurator virtual environment? You can find the latest release on the front page of our wiki, in the Downloads section. We've also posted instructions to help you get started.

Posted in Uncategorized | Leave a comment

Mapping Digital Forensics Workflows in Collecting Institutions

Posted on November 20, 2012 by Cal

Archives
- November 2012
- October 2012
- September 2012
- August 2012
- July 2012
- May 2012
- January 2012

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

Twitter: @bitcurator