

KING'S College LONDON

Digital Forensics in the Archive

Using open source & free software to capture and curate archival digital records

Gareth Knight, FIDO Project Manager

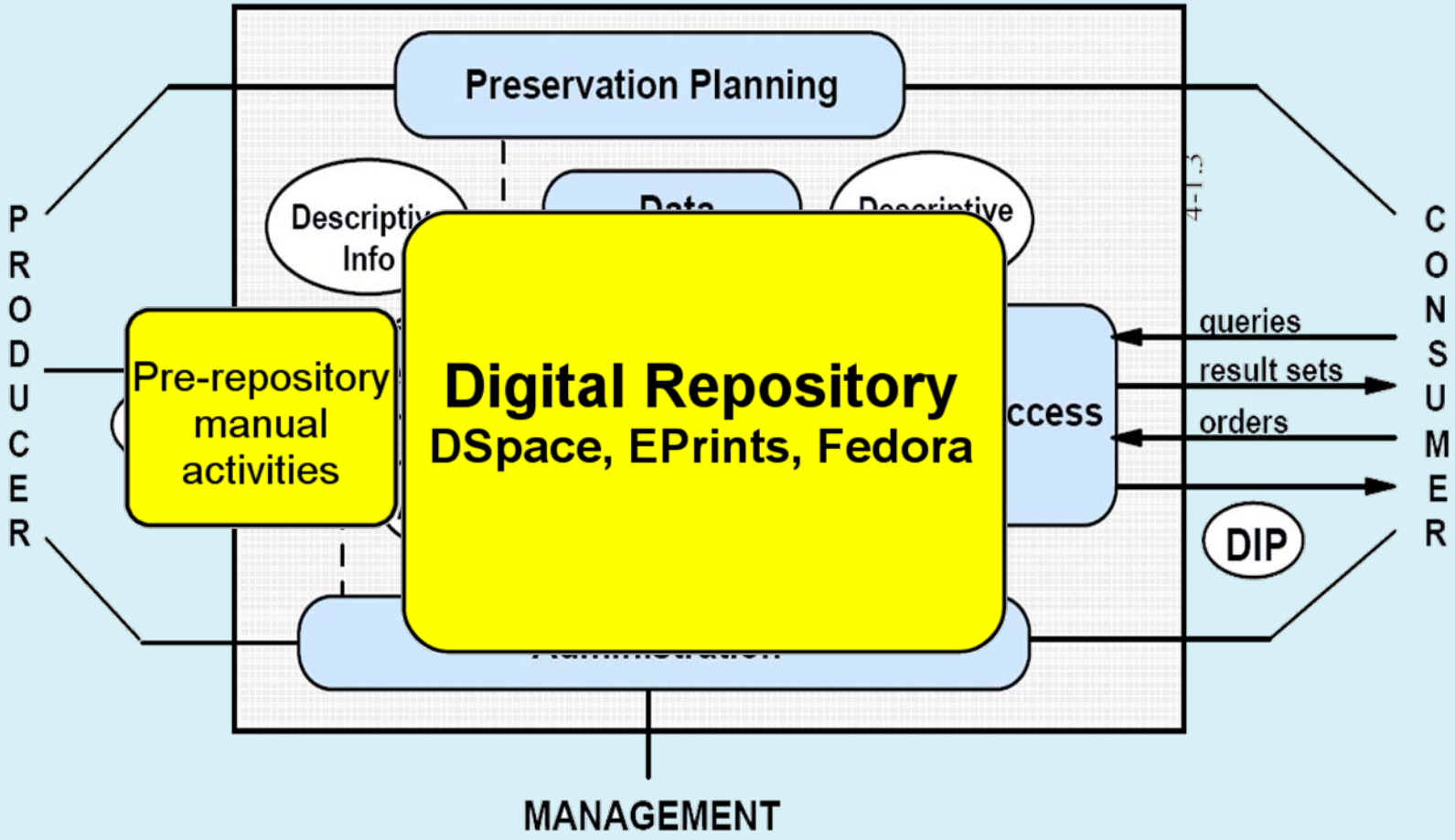
DPC: Digital Forensics for Preservation,

The Oxford Centre, Oxford, 28th June 2011

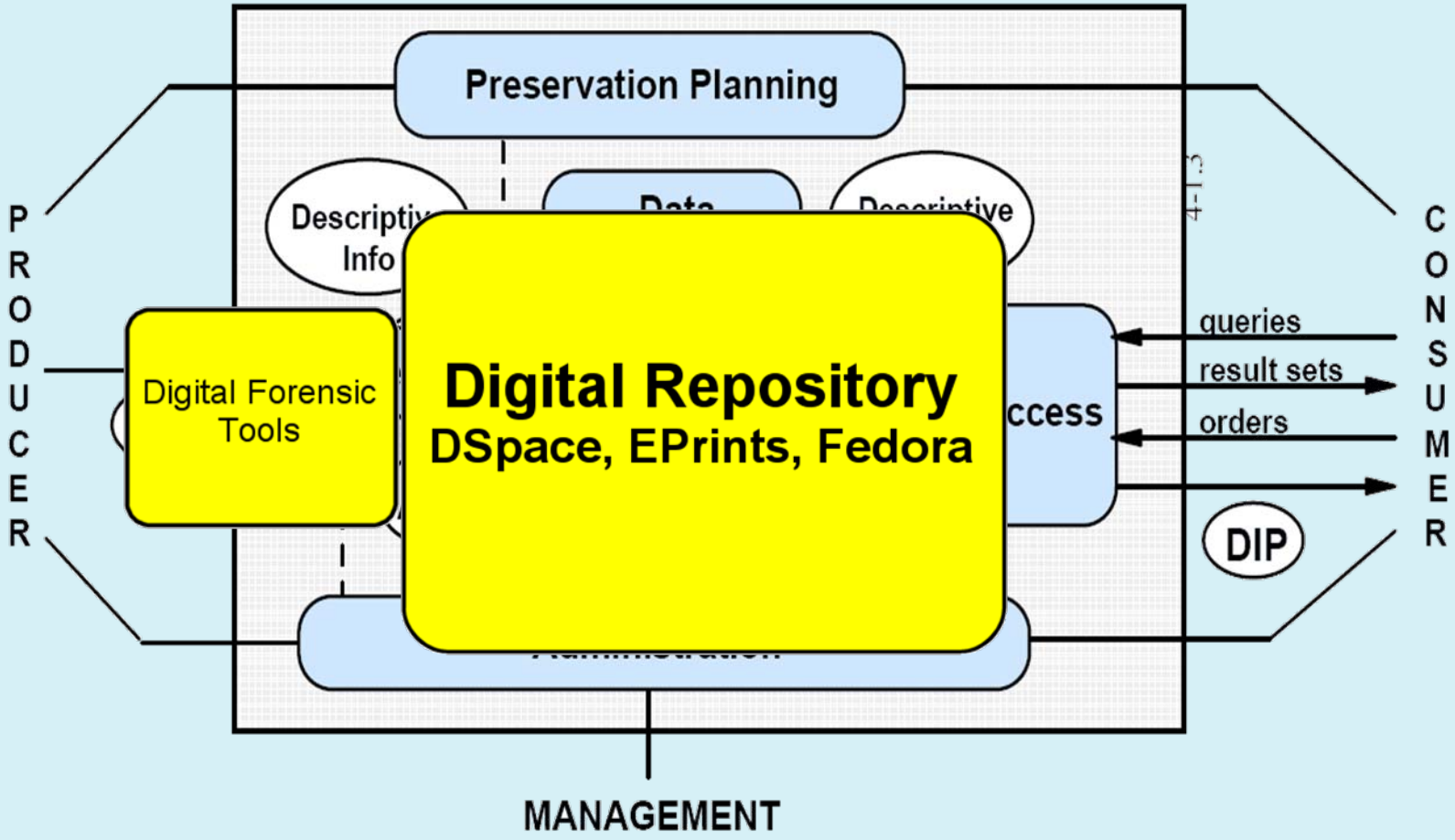
Overview

- Current practices and need for forensics
  - Gaps in the management workflow
  - Scenarios
- Forensic investigation activities
  - Decisions & factors to consider
  - Media imaging
  - File identification using hash sets
  - Data carving
- Challenges faced by forensic investigators and curators
- Summary & conclusions

Forensic tools in the archive



Forensic tools in the archive



Forensic Investigation of Digital Objects (FIDO)

- *Project team*: Centre for e-Research, working with Archives & Information Management (AIM) service
- *Funder:* JISC, Preservation Tools strand
- *Funding period:* February – July 2011
- Objectives:
  1. Evaluate the suitability of digital forensic principles and practices to enable HE archives to meet organisational commitments and legal requirements for maintaining digital records;
  2. Assess the effectiveness of using the chosen digital forensic tools set to identify, acquire, and analyse digital information held on digital media and computer systems in an archival environment;
  3. Seek to embed digital forensics tools & techniques into the working practices of the KCL Archives & Information Management (AIM);

The Daubert standard & Open Source

A judge makes a decision on whether the evidence must be relevant and reliable to be admissible in a US court.

Carrier (2002) applies the approach to DF software:

1. Testing: Can/has the procedure been verified? Does it produce false negatives or false positives?
2. Error rate: Are there known errors that arise from ‘tool implementation errors’ (buggy code, use of wrong spec) or ‘abstraction errors’ (decisions that are not 100% certain)
3. Publication: Has the procedure been published & peer reviewed?
4. Acceptance: Is the procedure general accepted as valid in the relevant domain, e.g. preservation field.

Source code may be examined to validate procedures to produce digital evidence

Open Source Digital Forensics Tools: The Legal Argument (http://www.digital-evidence.org/papers/opensrc_legal.pdf)

Archival Scenarios

Scenario 1: Donor (e.g. college alumni or their estate) contact archives to donate their research:

a. Donor provide data to be archived on digital media (floppy disk, optical media, solid state devices, internal/external hard disk)
b. Scenario 2: Donor submit system to archive for analysis, e.g. Windows PC, Apple Mac

Scenario 2: Staff working within the institution:

a. Staff have their laptops appraised to identify data of archival value not held elsewhere (e.g. college Dean)

Staff have their machine appraised prior to leaving institution

Broad issues to consider

1. *What is the working environment?*
   - Location of data capture, hardware to be used
   - Hardware/software appropriate to the environment
2. *Who will be performing the investigative work?*
   - What knowledge & expertise do your archivists/curators have?
   - What training will they require?
3. *How do you communicate intent to your user community?*
   - Ethical issues related to the retrieval of deleted and scraps of data – how do you communicate this in your donor agreement?
   - First rule of Forensic Club is: you do not mention forensics

Forensic hardware

Standard Intel/AMD system

- 1TB hard disk, 4GB memory, etc.

Connectivity

- Reader for solid state devices
- 3.1/2 & 5/14 floppy drive
- Disk controllers
  - Individual Computers Catweasel MK4 PCI or KryoFlux USB
- USB drive enclosure for IDE/SATA disks

Do not (currently) possess H/W write blocker – mount media as read-only

Data Acquisition

Act of obtaining possession of digital data for subsequent analysis. Commonly achieved through creation a disk image or clone that provides a bit copy of disk.

1. *Who will use the software?*
   - Archivist, end-user?
2. *What environment will acquisition be performed in?*
   - User computer at their workplace/home
   - User computer donated to college
   - Digital media connected to forensic machine
3. *What hardware will you be using? What media are you attempting to capture?*
   - Floppy hard disk, optical media, solid state – Mac, Windows, Unix
4. *Where will the acquired image be stored?*
   - External USB disk, Network device over Ethernet/Serial, etc.
5. What disk image format do you wish/are able to use?

Data Imaging formats/types

Formats

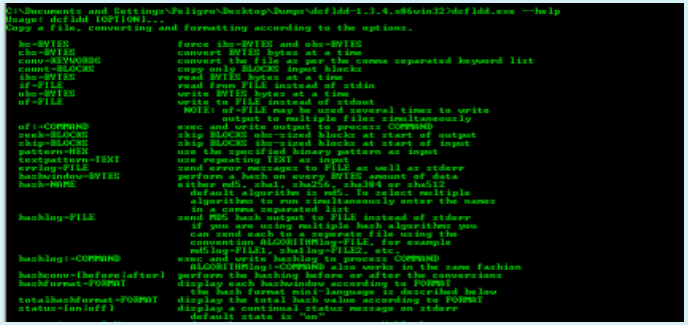
- *Raw/DD 'format':* Widespread support in range of forensic, virtualisation, and other tools. Lacks support for embedded MD & fixity, but can store MD as separate file.
- *Advanced Forensic Format (AFF):* Extensible open format comprised of Data-Storage (data) & Disk Representation (MD & other info using RDF) layer. Less support than Raw or Encase.
- *Encase Evidence format:* De-facto standard supported by EnCase & OSS (via LibEWF library). 2GB max file size, but can be split. Supports block-by-block checksums enabling the investigator to determine the sector that has been corrupted.

Choosing an appropriate format:

•FIDO built on file formats assessment criteria (Todd, 2009) for choosing disk formats

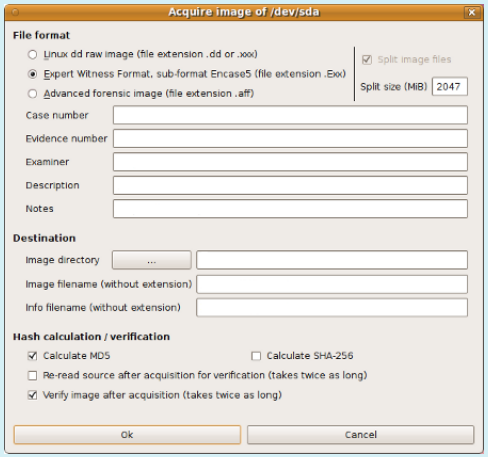
- Assessment criteria requires refinement to improve accuracy:
- AFF and EWF both scored highly.
- Raw/DD – Widely adopted & software independent, but relies upon 3rd party for metadata support, disk spanning and compression. N/A was recorded for disclosure & licence.

Acquisition tools



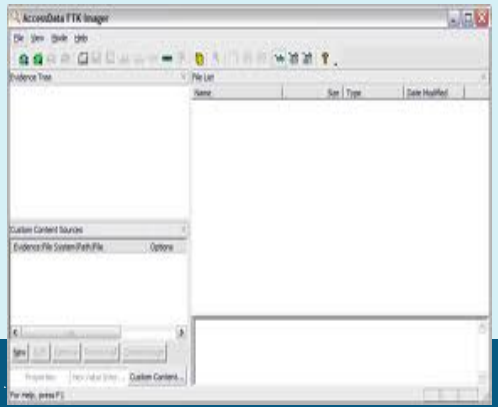
Booting from floppy

Dc3dd and dcfldd (if booting from floppy disk, wish to create Raw images, & unafraid of CLI)



Booting from CD/USB

OSFClone Guymager (pronounced: GUI-mager) and Automatic Image & Restore (Raw only)



Windows-based tools

FTK Imager & OSForensics are free commercial tools that may also be used

What type of data do you wish to retrieve?

Type of data to be captured

- *User data:* Documents, images, sound, emails, etc.
- *Software:* OS, software applications and other code
- *Log data*: Browser cache, cookies, registry entries

Does the log data support understanding of the academic user?

Level of analysis:

- *Active data:* Information readily available as normally seen by an OS
- *Inactive/residual data:* Information that has been deleted or modified
  - Deleted files located in unallocated space that have yet to be overwritten (retrieved using undelete application)
  - Data fragments that contains information from a partially deleted file (retrieved through carving)
- Inactive data useful, but need to consider ethics

Identifying origin of data files

Hashsets may be used to identify the origin and purpose of one or more files, e.g. filename, creator, magic number and fixity value

- *known good'* - Files that perform a legitimate purpose, e.g. Operating System, application.
- *'known bad'* - Files that denote viruses, Trojans, cracker's tools, or other malicious files

Information sources:

- *NIST National Software Reference Library (NSRL):* Hashset of legitimate files generated from software products obtained through purchase/donation. Stores 10,000+ software files. Reference Data Set published every 3 months & available through 3rd parties, such as Find-a-Hash
- *HashKeeper:* Maintained by the National Drug Intelligence Center. Repository contains information captured through criminal investigation. Academic (and other) institutions must file a FoI request to gain access to software and database.
- *Online File Signature Database (OFSDB):* Subscription based system dependent upon user contribution. Full access available through subscription of 25 USD per year

Currently being used by curators/archivists to distinguish between known third-party and potential user created files.

Data Carving

- The ‘carving’ of data from a larger data file for analysis by identifying header and searching for a corresponding footer.
- Equivalent to archival process of identifying paper fragments to other artefacts
- Variety of methods – different levels of success
  - Header/Footer, Block-Based, Statistical, file structure, Semantic Carving, In-place, smart
- Tools:
  - Foremost, Magic Rescue (both effective), PhotoRec, Scalpel
- Challenge: Id of files can be difficult if format uses short/no header & footer (e.g. ascii, JPEG vs. PNG)
- Produces false positives: Incomplete files, large concatenated files, extracts embedded bitstreams from complex objects

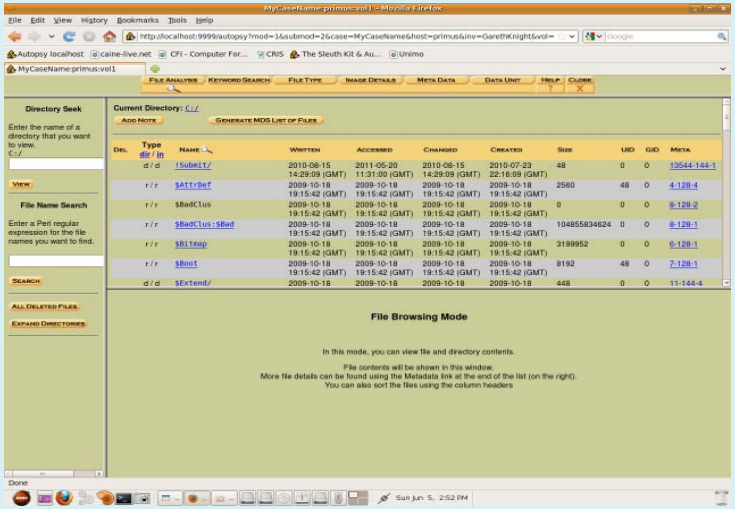
Img source: http://www.flickr.com/photos/jwthompson2/160835456/

Data Carving Examples

Imaged a disk containing 20 deleted files - 5 100k text files, 5 5Mb JPEGs, 5 90MB WMV videos and 5 300 MB AVI videos (approx file size)

- *PhotoRec* recovered all texts and JPGs. 3 AVIs were recovered in entirety, 2 were incomplete.
- *Scalpel* – Recovered all JPGs and 3 incomplete AVIs. Did not extract WMV or txt
- *MagicRescue* – Only recovers files it has a 'recipe' for (JPG, AVI, but not txt or WMV) – recovered JPGs, but not AVI. Did not attempt other formats.
- *Foremost* unable to recover any files

Integrated Forensic toolsets (1)

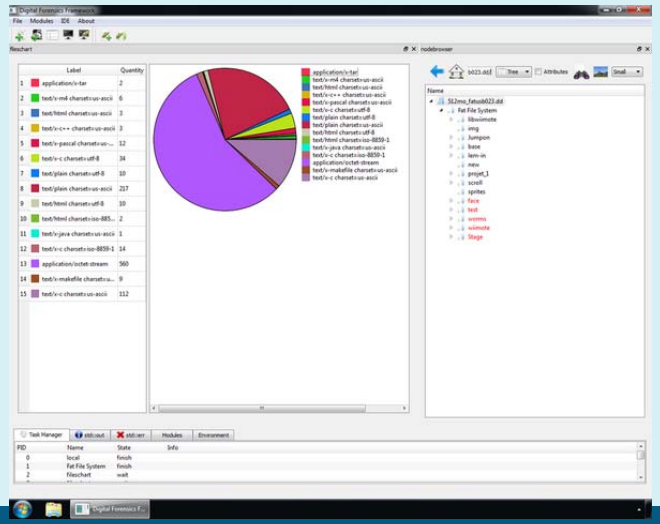


Sleuthkit & Autopsy (or PTK)

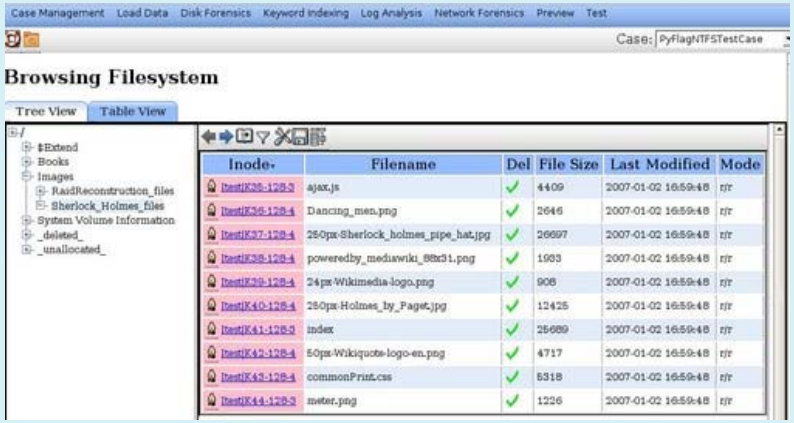
Set of command line tools for identifying file systems, performing file/keyword search, hash generation and look-up (via NSRL, HashKeeper, etc) and timeline mapping. Web client interface via Autopsy (free) or PTK (ajax-based, commercial)

Digital Forensic Framework

Cross-platform QT/Python tool. Modular design through plug-ins. Supports Raw/DD and EWF. Support for FAT, NTFS, EXTFS 2/3/4 file systems. Hash generation & check of selected files & comparison to NSRL hash dataset. However, data carving can be slow & does not begin to extract files until it has analysed entire disk



Integrated Forensic toolsets (2)



PyFlag

Web-based framework written in Python dev. by Australian Department of Defence. Supports raw, sgzip, AFF, EnCase, etc. Support for keyword/file search of active/inactive files, timelines, hash and compare using Hashkeeper.

OSForensic

Commercial, but free at moment. Mount range of formats (Raw, AFF, EWF, SMART, IMG, ISO, BIN). File/keyword search, hash generation and look-up (via NSRL, HashKeeper, etc) and timeline mapping.

Also: OSS distributions, including SIFT Workstation, BackTrack, Penguin Sleuth, DEFT Linux, CAINE and others

Current/future challenges for the forensic curator

- *Multi-user systems*
  - Distinguishing between data created by multiple users on same machine is time-consuming - requires analysis of timestamps and other features.
- *Archiving data on 3rd party services:*
  - Ethical issues associated with accessing & archiving user data on mail servers, second life, and cloud providers etc.
- *Diverse device & media types:*
  - Solid State devices subject to ‘wear levelling’ which purges inactive data
    (http://www.jdfsl.org/subscriptions/abstracts/abstract-v5n3-bell.htm)
  - Use of portable (personal/work) devices in the workplace, e.g. iPad, iPhone, Android devices – what is the master copy?

Conclusions

- Digital forensics has considerable value to archivist & digital curator
- Functionality offered by Open Source Forensic tools is often comparable to commercial equivalents
- No single tool is appropriate – require a combination of different ones
- Terminology is influenced by development in law enforcement community. Must map concepts to understandable archival equivalent & modify tools to reflect these terms
- Many OSS tools require command line interaction – further work is necessary to integrate results and provide user interfaces for non-techical users
- Hashsets provide useful method of identifying data files - academic community should contribute to development of hash sets and integrate tools into preservation workflow
- Forthcoming Bitcurator project (Matthew Kirschenbaum): may help to refactor OSS forensics tools for use in archival context

Thank You!

Questions

Gareth Knight
Centre for e-Research, King's College London
gareth.knight@kcl.ac.uk @gknight2000
020 7848 1979

http://fido.cerch.kcl.ac.uk/ @jiscfido