DIGITAL FORENSICS IN LIBRARIES AND ARCHIVES

Neil Jefferies
Bodleian Libraries
University of Oxford
@NeilSJefferies
https://orcid.org/0000-0003-3311-3741
neil.Jefferies@Bodleian.ox.ac.uk
NOT MY FIRST TIME!

• A long time ago...
  • ...well early 90’s
  • Dr Solomon’s Anti-Virus Toolkit for DOS
  • A chap called Graham Cluley worked on Windows...

• Dr Solomon’s Data Recovery...
  • ...Authentec
  • ...Vogon
  • ...Ontrack

• https://www.youtube.com/watch?v=pJ0wASeHQHo
FORENSICS IN RESEARCH LIBRARIES

- Long Term Data Management is only now becoming a expectation/requirement:
  - funders, institutions, policy makers share the blame
  - ...and libraries for not being noisy enough
  - ...and probably not getting involved earlier
- When research projects ends there is no funding/time for:
  - Specifications
  - Documentation
  - Hand-over
  - ...everyone has moved on to the next project
- A box arrives...
  - If you are lucky someone is still around from the project
  - If you are really lucky, they are technical
BUT AT LEAST IT WAS WORKING RECENTLY

AND THERE IS A GOOD CHANCE SOME OF THE PEOPLE ARE STILL ALIVE
FORENSICS IN ARCHIVES

• Personal archives
  • People “papers and effects” are increasingly digital
  • The number of eminent people who get their “effects in order” is very small
  • ...and this frequently does not include digital
• Boxes of media and devices
  • Can’t easily tell if it works
  • Can’t easily tell if it’s valuable
    • Triage issue – to fix or not to fix?
  • Dates, times, labels?
• An increasing problem
  • The number and variety of memory devices is increasing
CORPORATE ARCHIVES

- Not entirely dissimilar to personal archives
  - More complex
  - The work of multiple people
    - Less consistency, fewer patterns to work with
  - Corporate systems are not designed for long term access
    - Complex proprietary formats
    - Lots of home grown software/systems
  - Corporate failures are often messy
- **Lots** of boxes of media and devices
  - Corporate storage devices can be expensive
  - Spanned media...
  - And all the problems of personal archives
More Corporate Archives

https://www.innosight.com/insight/creative-destruction/
"Amazon is not too big to fail...In fact, I predict one day Amazon will fail. Amazon will go bankrupt. If you look at large companies, their lifespans tend to be 30-plus years, not a hundred-plus years."

Jeff Bezos, Nov 2018

“ARCHIVING IN THE CLOUD”

...SHOULD BE APPROACHED WITH EXTREME CAUTION – EGRESS (COSTS AND SPEED)
SIMILAR TOOLS

• Write blockers
• Kryoflux
  • https://github.com/archivistsguidetokryoflux
• eBay (or other sources of ancient kit)
• Forensic Toolkit
• ... or BitCurator (https://bitcurator.net/)

• And sometimes we have to crack things...

But we are concerned with information extraction rather than legal admissibility...although scholarly authenticity is similar
• Signature systems age badly
• Access matters
FORSENSICS SEGUES INTO DIGITAL PRESERVATION

Need to retain access to extracted information in the long term
- Format registries (https://www.nationalarchives.gov.uk/PRONOM)
- Format conversions
  - Significant properties
  - Stable formats
  - Stable delivery (https://iiif.io)
- Emulation (for difficult encapsulated formats)
- Physical preservation (games, art installations etc.)
- Machine analytics and access increasingly important

Downstream curation and processing
- ePadd (https://library.stanford.edu/projects/epadd)
- Digital preservation tools like Archivematica
- Oxford Common File Layout (https://ocfl.io) designed for recoverability
A lot of Humanistic research includes elements of reconstructing narratives from extant physical resources (usually libraries, museums and archives) – in effect, timeline reconstruction.

Increasingly broad range of resources becoming available digitally (https://doi.org/10.17875/gup2019-1146)

Not just timelines are being reconstructed:

- Medieval Libraries of Great Britain and the Corpus of British Medieval Library Catalogues reconstructs libraries (http://mlgb3.bodleian.ox.ac.uk/)
BUT THERE’S MORE

- Forensic techniques in cultural heritage
  - Meaning and Cultural Importance are often critically dependent of context
  - The collection process is not impartial so provenance matters
  - Hyperspectral and spectroscopic analysis of materials
    - Mexican palimpsests (https://www.bodleian.ox.ac.uk/news/2016/aug-18)
  - 3D reconstructions
    - Digitally unrolling burnt and fragile manuscripts (https://www.nature.com/articles/srep27227/)
- More advanced analytics based on access to larger digital corpora
  - Handwriting analysis and recognition
    - Often enhanced by predictive linguistic models
  - Linguistic text mining approaches – Sentiment and topic analyses
  - Entity extraction and identification
  - ...and yes, Machine Learning and AI
LOOKING FOR OPPORTUNITIES TO COLLABORATE

Neil Jefferies
Bodleian Libraries
University of Oxford
@NeilSJefferies
https://orcid.org/0000-0003-3311-3741
neil.Jefferies@Bodleian.ox.ac.uk