Digital Forensics in the Archive

Research Use Cases, Archival Requirements, Opportunities and Caveats of Automation

Thorsten Ries

7 Sept 2018, Computational Archival Science Workshop
The National Archives, London.

MARIE CURIE ACTIONS

UNIVERSITY OF SUSSEX

Overview

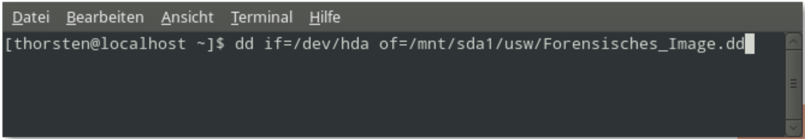
- Digital forensics in the archive
  - Why digital forensics for philology?
  - Projects, past and present
- The historical digital forensic record
  - In which way is the digital forensic record historical?
  - Complexity by the example
- Automation: opportunities, caveats
  - Where is automation of the forensic records I do research on possible / a good idea?
  - Where may it not be?

Digital forensics in the archive

UNIVERSITY OF SUSSEX

Philology, genetic critcism, and digital forensics

- My own work revolved around the digital forensic analysis of German literary authors' hard drives, inspecting them for recoverable draft stages of their work: Michael Speier, Thomas Kling, Marcel Beyer, Friedrich Kittler, etc.
- Digital forensics, forensic imaging (bitstream-preserving imaging) and forensic analytical tools (e.g. ddrescue, testdisk, file carver, sleuthkit, caine linux, etc.) became my most important instruments in digital philology and my practice of scholarly editing of born-digital texts and digital *critique génétique*.



```
[thorsten@localhost ~]$ dd if=/dev/hda of=/mnt/sda1/usw/Forensisches_Image.dd
```

Philology, genetic critcism, and digital forensics

- Next to M. Kirschenbaum's (*Mechanisms*, *Track Changes*, BitCurator etc), D. Reside's and L. Duranti's philological, theoretical and library-information science publications, the work of archivists, librarians and developers such as S. Thomas, J.L. John, A. Dappert, R. Foss, T. Owen, C. Lee etc. and multiple large scale international born-digital archiving projects (I will spare you with the list) were most influential to my own work.
- My own work on Michael Speier and Thomas Kling is now by and by being published, key theoretical concepts have been developed in *The Rationale of the Born-Digital Dossier Génetique*, published in *DSH* 2018, and a further article in *Cahier voor literatuurwetenschap* 2017.
- I am currently working on finishing a book on born-digitals by Speier and Kling (in German, as some other articles on the topic), also the first special issue of a new journal *Digital Scholar* on born-digital archives, which I am editing, will appear soon.

Extended scope: the historical humanities

The Marie Sklodowska-Curie project *Digital Forensics in the Historical Humanities* ('18-19) features exemplary case studies on forensic aspects of three archives, one literary, two of historical relevance:

- **Hanif Kureishi** personal digital archive at BL
- **Glyn Moody** private personal digital archive
- Born-digital records in the **Mass Observation Archive**

Goals: survey digital forensic analysis of born-digital historical sources across the humanities, exemplary cases of critical appraisal of born-digital primary sources.

MARIE CURIE ACTIONS

US UNIVERSITY OF SUSSEX

The historical digital forensic record

Unwieldy digital materiality: the digital forensic record

- Bitstream-preserving (forensic) imaging: preserves physical data structure of storage medium, geometry of the file system, partitions and partition table incl. unallocated space (recoverable data), specific historical features of the data / file system (e.g. fragmentation, metadata format, bad and corrupted sectors, hidden data). One-off hash authentication. Removes barriers of outdated hardware, controllers, risks of bitrot → Very good preservation format.
- Historicity and forensic perspective: forensic images freeze a physical data record, preserving it in its historically authentic form (as authentic as possible). Access vectors, formats, forensic aspects and phenomena of the forensic record are a) mostly undocumented, not standardized and b) specific to individual software and operating systems' versions c) diverse due to the expanding multitude of hard- and software platforms, cloud computing. → Challenge for automation (Garfinkel, *The next 10 years*, 2010).
- This **historicity of the forensic record** and its latent features is rooted in the historical development decisions made on software architectures designed to overcome performance constraints of hardware platforms (sensu J.F. Blanchette, *A Material History of Bits*, 2011).

US
UNIVERSITY OF SUSSEX

Unwieldy digital materiality: the digital forensic record

- **Bibliographic, archival perspective:** The digital forensic record is not a physical *ad oculos* record and, as far as forensic images are concerned, a bitstream-identical copy, but still a copy → questions of provenace principle, custodianship to secure chain of provenance, evidence (C. Rogers).
- **Forensic perspective:** The forensic record is mostly latent and forensic methods might produce false positives, so-called ›constructed traces‹ (F. Cohen). A rolling back of system states based on forensic traces is mostly impossible, due to unsystematically incomplete record and not recorded metadata (F. Cohen).

Automation: opportunities, caveats

UNIVERSITY
OF SUSSEX

Desirable automation / semi-automation

- **Acquisition, ingest, first assessment** of forensic images,
- **Categorisation** File systems, operating systems, structure breakdown, how to mount,
- **Software**: malware scans, scans for installed software that might have to be disabled due to copyright, files that might have to be deleted due to copyright and that don't belong to the archive, check against NSRL National Software Reference Library,
- **Repository** automatic bitstream hashing and checking, bitflipper.

Limits of automation / undesirable automation

- **Migration** to other formats changes the bitstream record, please don't migrate and delete. Ever.
- **Triage, masking and censoring** Workflows for archivists and heirs to mask or censor specific parts of a storage medium could be (at best) semi-automatic.
- **Automatic appraisal: forensic tools error and obsolescence, historicity** Forensic features of the historical record are undocumented, unstandardised and often result in fragmented, messy, garbled data that can only be ungarbled with the right choice of tools and context knowledge about the specific historical system. Forensic softwares catch a lot, but they also miss a lot and produce false positives. They are historical themselves and become obsolescent (incompatible with the archival working environment). Historicity of Platforms and algorithms: no forensic software to date is able to uncompress Windows 10's »Compact OS« packages. → often data is missed in preservation and appraisal due to failed automated processes [DLA example]
- **Personal rights / data protection / GDPR** automatically identifying potentially sensitive data is targeted access, processing and potentially copying, of most sensitive data without a specific reason. Also: which data is regarded as sensitive is context-dependent, therefore hard to automate (bulk_extractor, fiwalk).

Thank you for your attention!

MARIE CURIE ACTIONS

US
UNIVERSITY OF SUSSEX