Born-digital Literary Archives

Forensic Perspectives on an Expanding Field, and Current Challenges for Preservation

Thorsten Ries, The University of Texas at Austin
Feb 26, 2024 at 1:00pm

Online DPC event: Digital forensics and digital preservation: Investigating good practice
Best Practices

• What are they?
• Defined by actual forensic practice? (ISO 27050, DEFSOP)
• Effects of Historical Digital Materiality
• Aspects of (my own) practice.

• Preservation
• Discovery, Analysis, Interpretation
• AI?
Challenges → General, with Baker

→ Time and obsolescence (physical degradation, software obsolescence, hardware support, forensic hardware support for historical systems)
→ Volume / Storage capacity vs Gaps
→ Data curation, security management, migration – is a long-term task
→ Variety, Diversification, Fluidity of technology dev changes preservation and analytical methods (SSD’s).
→ Ethics, Data Security, Policies and Solutions for Safe Research Access – see GDPR.
→ Environmental – carbon footprint long term storage.

Challenges → New

→ Change of programming, legal, and forensic paradigms and frameworks (e.g. cloud computing)
→ Encryption.
→ AI integrations in standard software thwart forensic approaches, mostly leave no trace, and change the idea of authorship and agency (whodunnit, this may not be a problem).
→ Long-term interpretability of historical forensic phenomena, gaps and fragmentation of record, and forensic software artefacts (false positives, C-traces).

James Baker: Digital Forensics in the House of Lords: six themes relevant to historians (Part One / Two), 2019 [Link].
Preservation
Preservation Formats and Strategies – Forensic Methodology as Approach

→ **Securing evidence** – Bitstream-preserving images (»forensic images«, keeps deleted material and original of the system context)
→ **Formats** – Sustainable, open formats (e.g. dd vs EnCase), and system context documentation.
→ **Authentication** – Hashing (encrypted checksums, MD5, SHA256), alternative: blockchain (sustainable?)
→ **Lots of copies keep stuff safe** (LOCKSS, distributed), Repositories with enhanced security mechanisms (bit-flipper, etc).
→ **Digital original vs. migration / conversion.**
Highlights

→ Tool documentation (version)! (effects of tools, e.g. dd, ddrescue, recuperabit)
→ Careful with image compression, encryption, proprietary formats!
→ Hardware documentation (SSD: garbage collection, overprovisioning areas, SSD slack)
→ We need system context. Evidence often does not come in expected formats.
Tool documentation: Imaging tools

- dd
- ddrescue
- dd_rescue
- recuperabit (experimental)
- Advanced Forensic Format (sustainable?)
- FTKimager (proprietary)
- EnCase (proprietary)
Hardware: Solid State Drives

Recovery Problems

• Wear leveling.
• TRIM (operating system level)
• Garbage collection (controller level)
• Major challenges for forensics, often deleted data is forever lost after a few minutes to a day, special hardware is required for capture – garbage collection starts automatically, controller-driven, when the drive is connected, data scrambling, firmware-compression, potentially in the future: encryption.

Recovery Opportunities / Solutions

• Wear leveling.
• Implementation is not standard, sometimes TRIM is not active due to bugs, faulty implementation or incompatibilities, operating systems etc support TRIM.
• Drive slack, overprovisioning, incompatibilities (USB), nonstandardized builds and bugs. – Overprovisioning areas accessible only with very specialized methods / hardware.
• Chip-off forensics
• Factory access mode to access overprovisioning

This is how you block access to NAND chips on a Crucial BX100:

Next, we’ll connect the SSD to ACELab PC-3000 in order to switch it into factory access mode. Factory access mode is activated by sending a service command (widely different per OEM/controller) to the SSD controller. Then, we’ll upload the LDR code into the controller’s RAM.
C M Taylor on ‘keystroke logging project’ with British Library

a guest blog by Craig Taylor, whose latest novel, Staying On, is published by Duckworth in 2018. In 2014 he began a project with the British Library to document the creative process of writing the book, using key-logging software. You can reach Craig on Twitter at @CMTaylorStory.

Re-entering the academic world after starting work as an Associate Lecturer on the Publishing degree at Oxford Brookes University, I began speculating about writer’s archives. Did previous scholars have access to more hand-written and typed drafts of works in progress - actual objects showing the shaping of works of art - but with the normalisation of computerised authorship, were these discrete drafts abolished in the rolling palimpsest of write and digital re-write?

Plus, I was considering a new novel myself, but as I have written elsewhere, emotionally I was daunted by the long-haul loneliness of novel writing, a process I considered in my most despairing moments as like wallpapering a dungeon.

I spoke to my friend Mark about these two things - the lost drafts and the loneliness - and in a flash
### Context II: Keylogger

#### Keystrokes Typed

**Application:** SoftIce

**Start:** 15/11/2016 08:33:40

**Title:** More Em November 16.odt - OpenOffice.org Writer

<table>
<thead>
<tr>
<th>Keystrokes Typed</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;8:33&gt;</td>
</tr>
<tr>
<td>3&lt;PGUP&gt;3&lt;DOWN&gt;25&lt;UP&gt;3&lt;RIGHT&gt;25&lt;silly&gt;DOWN&gt;LEFT&gt;8&lt;RIGHT&gt;DOWN&gt;3&lt;UP&gt;4&lt;RIGHT&gt;DOWN&gt;136&lt;UP&gt;3&lt;RIGHT&gt;DOWN&gt;25&lt;UP&gt;3&lt;RIGHT&gt;DOWN&gt;57&lt;UP&gt;2&lt;LEFT&gt;4&lt;APART&gt;DELETE&gt;18&lt;DOWN&gt;LEFT&gt;4&lt;DOWN&gt;57&lt;UP&gt;2&lt;LEFT&gt;14&lt;A&gt;SPACE&gt;</td>
</tr>
<tr>
<td>3&lt;LEFT&gt;12&lt;DOWN&gt;7&lt;RIGHT&gt;10&lt;LEFT&gt;O&lt;DOWN&gt;2&lt;Well&gt;SPACE&gt;a&lt;DOWN&gt;15&lt;UP&gt;UP&gt;3&lt;DOWN&gt;61&lt;UP&gt;11&lt;UP&gt;DOWN&gt;5&lt;LEFT&gt;</td>
</tr>
</tbody>
</table>
The Reconstruction of the Author’s Movement Through the Text, or How to Encode Keystroke Logged Writing Processes in TEI-XML

Lamyk Bekius

p. 3-43

https://doi.org/10.4000/variants.1245
writing processes that take place in a word processing environment that the
author is already familiar with: Microsoft Word (Leijten and Van Waes 2013). Each
time an author activates Inputlog to start a new writing session, the Word
document in which the author is working is saved in the background, in a
folder that contains the session's date and number. Subsequently, the Word
document is saved again when the author ends the writing session by de-
activating Inputlog. This results in a session-version of the text for each
session, which shows the text's gradual expansion. But Inputlog does not just
save Word documents. When the program is running, every keystroke and
mouse movement is recorded with a timestamp (Leijten and Van Waes 2013).
While writing, authors retain control of the process; they can start and stop the
logging when they choose, and the data is stored on their local PC or laptop.

Although Inputlog is developed for textual and cognitive study of writing, the
data output from the writing process of Bogart's Roosevelt, generated in
Inputlog, is not immediately suitable for literary textual research. While
Inputlog provides a video replay of the recorded writing session, some issues
emerge when replaying Bogart's writing process. Short writing sessions
comprising linear text production are replayed accurately, but as soon as larger
segments of text are relocated or deleted, when the writing is characterized
with non-linearity, or when the logged session is of considerable length, the
replay mode is affected and represents the revisions and text production at the
wrong location in the text. Moreover, relying solely on a video replay of the
writing session for text genetic analysis also seems undesirable, as one would
need to watch a writing session of, say, two hours in its entirety, while
constantly pausing to analyse the effect of the revisions. A static
reconstruction of the writing session — whether or not in combination with a
video replay, as in Dirk Van Hulle's proposal for a "Dynamic Facsimile" in the
present issue (Van Hulle 2021) — is favoured to ensure adequate analysis.
Hence, in order to be able to study the revisions (contained in the keystroke
logging data) in their textual context, the twofold output of Inputlog — the
Word document and the keystroke logging data — requires some reassembly.

Since TEI-conformant XML is widely used to create a digital form of humanities
data — texts, manuscripts, archival documents and so on — I opted to encode
the keystroke logging data in TEI-XML to visualize and analyse revisions in their
textual context (Barnard 2014). For me, these transcriptions function as a tool
to gain more insight into the textual genesis. They could eventually be used for
visualizations of the writing process, but a proper discussion of the latter lies
outside the scope of this article. In order to reflect on how keystroke logging
Context III: System crashes

Think also about system files and areas like pagefile.sys, hiberfil, VSS shadow partitions, file history feature storage ...

Context III: System crashes
James Somers: http://features.jsomers.net/how-i-reverse-engineered-google-docs/

Draftback plugin for Chrome.
Discovery, Analysis, Interpretation
Highlights

→ Document every tool (with version)
→ Know the tool’s C-trace potential
→ Test historic software behavior
Impact of the tool: The case of Photorec 7
Impact of the tool: foremost and MacWrite Pro
MacWrite Pro 1.0 artefacts

- MacWrite Pro 1.0 file viewed in the word processor / hex editor.
- The file exists in four versions spread over all harddrives (also Syquest!). → Synopsis.
- Two files dated 28 Jan 2000 are text-identical, two further are dated 27 November 2001, 8 September 2002.

[in Print:] Thorsten Ries. Pour la reconstruction des processus d’écriture numériques de Derrida grâce à la computer forensics : reconstruction des données et matérialité numérique historique ... 2021 [in print] [preprint available (HAL)].
Impact of the tool: foremost and MacWrite Pro

- MacWrite Pro 1.5 compressed datastream.
- Diff view of fastsave artefact from decompressed SECRETA datastream compared to doc text.
- The decompression algorithm was reengineered.

[in Print:] Thorsten Ries. Pour la reconstruction des processus d’écriture numériques de Derrida grâce à la computer forensics : reconstruction des données et matérialité numérique historique ... 2021 [in print] [preprint available (HAL)].
Impact of the tool: foremost and MacWrite Pro

Foremost File Carving

- The MacWrite Pro format(s) usually not part of any file carver, but could be easily added to the foremost heuristics.
- Fun software historical fact: MacWrite Pro files did not have a footer, their end was defined in the document as length value. Therefore all recovered files had to have the same defined length (C-trace!).

Recovery from Unallocated Space

- Raw datastream comparison of „non-punched“ drive data (left) vs „punched“ drive data (right), indicating in white the start of a deleted file.
- BLKLS does not work with the old HFS. Therefore I had to punch out the »existing« files from the image with Bleachbit. But no modern OS mounts HFS writeable!
- But Parrot Linux does (don’t ask me why)!

[in Print:] Thorsten Ries. Pour la reconstruction des processus d’écriture numériques de Derrida grâce à la computer forensics : reconstruction des données et matérialité numérique historique … 2021 [in print] [preprint available (HAL)].
Testing needed! The case of MS Word GUID’s

<table>
<thead>
<tr>
<th>0000CB20</th>
<th>00 00 00 06 00 00 00 54 69 74 65 6C 00 03 00 00</th>
<th>........Title....</th>
</tr>
</thead>
<tbody>
<tr>
<td>0000CB30</td>
<td>00 01 00 00 00 00 00 98 00 00 00 03 00 00 00</td>
<td>..............6...</td>
</tr>
<tr>
<td>0000CB40</td>
<td>00 00 00 00 20 00 00 00 01 00 00 00 36 00 00 00</td>
<td>...........&gt;......</td>
</tr>
<tr>
<td>0000CB50</td>
<td>02 00 00 00 3E 00 00 00 01 00 00 00 02 00 00 00</td>
<td>...._PID_GUID...</td>
</tr>
<tr>
<td>0000CB60</td>
<td>0A 00 00 00 5F 50 49 44 5F 47 55 49 44 00 02 00</td>
<td>...........A...N...{</td>
</tr>
<tr>
<td>0000CB70</td>
<td>00 00 E4 04 00 00 41 00 00 00 4E 00 00 00 7B 00</td>
<td>6.9.E.B.8.D.A.0.</td>
</tr>
<tr>
<td>0000CB80</td>
<td>36 00 39 00 45 00 42 00 38 00 44 00 41 00 30 00</td>
<td>-.C.5.5.3.-.1.1.</td>
</tr>
<tr>
<td>0000CB90</td>
<td>2D 00 43 00 35 00 35 00 33 00 2D 00 31 00 31 00</td>
<td>D.5.-.B.C.5.C.-.</td>
</tr>
<tr>
<td>0000CBA0</td>
<td>44 00 35 00 2D 00 42 00 43 00 35 00 43 00 2D 00</td>
<td>3.4.8.1.0.B.C.1.</td>
</tr>
<tr>
<td>0000CBB0</td>
<td>33 00 34 00 38 00 31 00 30 00 42 00 43 00 31 00</td>
<td>0.0.0.0.}.........</td>
</tr>
<tr>
<td>0000CBC0</td>
<td>30 00 30 00 30 00 30 00 7D 00 00 00 00 00 00 00</td>
<td></td>
</tr>
</tbody>
</table>
Projekt Vorzeitbelebung, Herodot WRL, WBK, WRS / Synopsis

1 Herodot
Ein grosser Verschweiger unter den Geschichtscheibern und [[†frühen|17]|†].
| 214861335 | Herodot |
| 214861343 | Ein grosser Verschwieger unter den Geschichtsschreibern und Ethnologen ist Herodot. Und Unterst |
| 214861439 | tzung in Informationsunterdr |
| 214861468 | ckung fand und findet er bei zahlreichen seiner Kommentatoren. Man wundert sich ein ums andere Mal, was Herodotlesern alles nicht zugemutet werden soll. Oft genug stellt sich dann dieses bl |
| 214861599 | de Gef |
| 214861606 | hl ein, das sich jedes Mal breit macht, f |
| 214861708 | hlt man sich man wieder fussnotenm |
| 214861743 | ssig genasf |
| 214861755 | hrt, da |
| 214861764 | keine Ausnahme! - nicht ernstgenommen oder auf ganz und gar irrelevante, l |
| 214861840 | cherliche Pfade geleitet. Man hat den Eindruck bei Herodot, eine Berufskrankheit der Historiker: das Nichtinteressante, das Ger |
| 214861968 | ll, die tauben N |
| 214861985 | sse werden die Prachtalleen der Histoire entlanggerollt. Und mit Verve nichttransportiert und nichtbereitgehalten, noch nicht einmal in den staubigen Gr |
| 214862140 | berfeldern die den Fussnoten reserviert sind wird allzu oft was an vermeintlich Unser |
| 214862227 | sem, an vorgeblich beziehungsweise noch Nichtgesichertem, vor allem: was an Nicht-Sicherungsw |
| 214862321 | rdigem von vorne herein als haneb |
| 214862355 | chen-nichtdiskurs |
| 214862374 | rdig aussen vor gelassen werden kann. |
| 214862413 | Ein Beispiel nur. Besonders darauf verlassen kann man sich, dass abgewiekt werden wird, wendet sich notgedrungen das Thema kultischen Drogen zuwendet. Dann wird der Leser gewiss mit einem |
| 214862604 | alten Abschreibefehler |
| 214862626 | sediert oder alles wird, schwuppdwupp, zu einem |
| 214862677 | Missverst |
| 214862687 | ndn |
| 214862693 | runtererk |
| 214862705 | rt - wie im Fall der nomadisierenden Budiner, Anrainern der Skythen in der Dnepr-Region, die sich durch den Verzehr von Fichtenzapfen auszeichnen. So Herodot. Der Herodotkommentar i |

System crashes can lead to preservation!

MOPA: C4131, Spring 2013 Dir: Human rights, Immigration, ...
Old and new types of sources, formats, and: AI
Challenges → General, with Baker

→ **Time and obsolescence** (physical degradation, software obsolescence, hardware support, forensic hardware support for historical systems)

→ **Volume / Storage capacity vs Gaps**

→ **Data curation, security management, migration** – is a long-term task

→ **Variety, Diversification, Fluidity of technology dev changes preservation and analytical methods (SSD’s).**

→ **Ethics, Data Security, Policies and Solutions for Safe Research Access** – see GDPR.

→ **Environmental** – carbon footprint long term storage.

Challenges → New

→ Change of **programming, legal, and forensic paradigms and frameworks** (e.g. cloud computing)

→ **Encryption**.

→ **AI integrations** in standard software thwart forensic approaches, mostly leave no trace, and change the idea of authorship and agency (whodunnit, this may not be a problem).

→ **Long-term interpretability** of historical forensic phenomena, gaps and fragmentation of record, and forensic software artefacts (false positives, C-traces).

James Baker: *Digital Forensics in the House of Lords: six themes relevant to historians (Part One / Two)*, 2019 [Link].
Types of Born-digital Archives – Continuous Diversification and Development

→ **Personal Archives, Institutional Archives** (floppy disks, hard drives, CD, DVD, tape)
→ **Distributed, Networked Archives** – Web Archives, Cloud Archives, Databases, Platforms, Social Media
→ **Software, Hardware-Software ensembles.**
→ **Other** AI models (?), hacking archives, digital art installations.
### Types of Born-digital Archives – Continuous Diversification and Development

- **Personal Archives, Institutional Archives** (floppy disks, hard drives, CD, DVD, tape)
- **Distributed, Networked Archives** – Web Archives, Cloud Archives, Databases, Platforms, Social Media
- **Software, Hardware-Software ensembles.**
- **Other** AI models (?), hacking archives, digital art installations.

### Historical Digital Materiality – These Archives are Material and Not

- **Double character of digital materiality** – *immaterial* (*not-unique* digital objects: lossless copying btw states/ media, multiple physical instances of same digital object, storage as processing instead shelf-»existence«, MK: »formal materiality«) vs. *material* (*physically unique*: logical-material, context-dependent, MK: »forensic materiality«)
- **Historic specificity** – formats, and their designed dependencies on historical softwares, file- and operating systems, hardware, etc.
- **Digital Object / Context** – Layered, Distributed aspects (filesystem metadata) and processing traces (e.g. temp files, spool files, drive slack, VSS shadow partitions, etc).
- **Changing historical concepts** – »document« / »file« vs. cloud datastream object, clusters; »memory« as RAM, storage vs. NVRAM. (Blanchette 2011: »design histories«)
- **Gaps / Loss / Threat** – the digital forensic record is per definitionem incomplete, discontinuous (snapshot character), and contains gaps, is not reversible.
Curveship.js

AUTOMATIC NARRATIVE VARIATION

THANKS: PUBLICATIONS - CURVESHIP:PY, LINKS: CURVESHIP:JS, LINKS: GITLAB

Curveship is a system for automatic narrative variation. It can create the same underlying events in different ways, responding to the same content as different narrative discourses. Curveship enables users to test ideas and work with feedback. It can generate different characters, the narrator or narrator changes in the user to alter the narrative is told, and for scenarios, content to change from a different setting, changing or altering events. The system is open-source and can be used for various research and creative projects. It is an ongoing project that is changing and developing. The system is focused on generating and creating narratives. The system is designed to be flexible and adaptable to different narrative structures. The system is developing further to support different narrative forms. The system is open-source and can be used for research and creative projects. The system is ongoing and developing. The system is open-source and can be used for various research and creative projects. The system is ongoing and developing.

Curveship.js is the web version representation of a curveship ad network. Through HTML, CSS, and JavaScript, the system is designed to visually represent the underlying structures and trends of a narrative. It enables users to experiment with different narrative structures and visualize how they affect the narrative.

Curveship.js is a tool for generating and testing narrative structures. It is an open-source system that can be used for various research and creative projects. The system is ongoing and developing.

Curveship.js is a tool for generating and testing narrative structures. It is an open-source system that can be used for various research and creative projects. The system is ongoing and developing.

NARRATIVE VARIATION

Narrative variation means realizing the same underlying events in different ways. It is a core idea behind the system. The system is designed to automatically generate different narrative structures from the same underlying events. Curveship can apply different narrative structures, automatically, in a way that lets an author or author determine and programs Curveship can be an extensible and flexible system.

A simple interface makes a complex idea easy to understand. The system is designed to be easy to use and intuitive. The system is designed to be flexible and adaptable to different narrative structures. The system is ongoing and developing.

Contributors: Nick Montfort, Lauren R. Labrecque, Daniel K. Dewan

Languages: English

Open Source: Yes

Thanks to:

Nick Montfort

Curveship.js is a system for automatic narrative variation. It can create the same underlying events in different ways, responding to the same content as different narrative discourses. Curveship.js is an ongoing open-source system that is developing. It can be used to test ideas and work with feedback. It can create different characters, the narrator or narrator changes in the user to alter the narrative is told, and for scenarios, content to change from a different setting, changing or altering events. The system is open-source and can be used for various research and creative projects. It is an ongoing project that is changing and developing. The system is focused on generating and creating narratives. The system is developing further to support different narrative forms. The system is open-source and can be used for research and creative projects. It is an ongoing project that is changing and developing.
Write an ironic love poem to Dante Alighieri in the style of Ann Cotten, in Italian!

Oh Dante, mio amore infinito,
Il tuo Inferno brucia nel mio cuore,
Le tue parole scivolano, come vino,
Lasciandomi avvolta da un dolce torpore.

Come il tuo amor per Beatrice,
Così è il mio per te, o poeta celebre,
Ma la mia passione è solo un capriccio,
Un omaggio ironico, un amore da intendere.

Le tue rime diventano un gioco,
Nel tuo Purgatorio, mi perdo e mi ritrovo,
Ma non so se il tuo Paradiso
Possa donarmi un amore vero e nuovo.
GPT-4 generating a (simple) digital poem in LUA (and Italian)

Create a LUA code program that generates an endless stream of varying distichons that praise Dante Alighieri

```lua
local function generate_random_adjective()
    local adjectives = {
        "glorioso", "sublime", "magnifico",
        "immenso", "geniale", "divino"
    }
    return adjectives[math.random(#adjectives)]
end

local function generate_random_noun_phrase()
    -- function to generate random noun phrases
end
```

Due to high demand free GPT-4 usage is currently limited to 5 messages per day.

Type a message...
These challenges are really an opportunity for Research and Development in the Humanities, History, Social and Political Sciences to cooperate with Information Science to maintain long-lasting and indispensable impact in society, including social justice, culture, and archives.

Professional change – added skillsets as data accession, safeguarding and curation specialists, historical digital forensics, and specialists in critical data verification for humanists as well as information scientists create a unique data science profile for humanities professions that offers transferable skills reaching in the data security and legal sector.

Specialists in Curation, Chain of Custody / Evidence, Safeguarding of Information with open standards will be even more important as our verification / falsification methods continue to fail in the light of AI.

Our basic research concepts in the humanities, e.g. »human authorship« and »human experience«, will continue to shift as AI progresses to be integrated into everyday workflows (writing). Is this really a problem, if digital forensics does not properly distinguish human and AI contribution? (AI forensics, membership inference attacks, XAI)

Specialism in forensic analysis and AI verification.
New Corpora

→ Tech (Sub-)Culture Archives (e.g. national cyberfeminism archives, queer online history archives, etc).
→ Disinformation Archives.
→ Hacking Culture Archives,
→ Hacking History, Online Threat Event, Vulnerability Archives.
AI in forensic analysis?

- AI is already the subject of forensic inquiry (membership attacks, etc).
- AI is already being used in forensic investigations to find data more quickly on evidence images.
- Caveat I: non-reproducible; it is often unclear what it finds (and what it doesn’t).
- Caveat II: the use scenario in an archive is much different from a forensic lab, which processes a lot of very similar devices, whereas in an archive the historical variation differentiates every device.
Thank you for your attention!