The British Academy

Journal of the British Academy, 10: 157–185
https://doi.org/10.5871/jba/010.157
Posted 1 December 2022

Digital history and born-digital archives: the importance of forensic methods

Thorsten Ries

Abstract

This article explores the historical materiality of born-digital primary records from a digital forensic, archival and historical scholarship perspective. On a conceptual level, the article discusses the historical materiality, layeredness and complexity of born-digital records, considers the impact of technological change on their forensic materiality as well as archival and historical scholarship methodology and practice. These historical, forensic and archival perspectives will be laid out drawing on examples from Glyn Moody's personal digital archive, born-digital documents of the Mass Observation Project Archive (MOPA) and cases of cybersecurity events, including the hack of matrix.org in 2019. The discussion makes the case for further developing open and sustainable digital forensic preservation formats, workflows and practice in the archive sector for personal, institutional and web archives, as well as for the development of digital forensic methodology for critical digital source appraisal in historical scholarship.

Keywords

Digital history, digital forensics, Glyn Moody, Mass Observation Project, matrix.org, born-digital primary sources, cybersecurity.

Author

Thorsten Ries is an Assistant Professor at the Department of Germanic Studies at the University of Texas at Austin, USA. He is specialised in German literature from the 18th to the 21st century, Digital Humanities and Digital Learning, Scholarly Editing (Genetic Criticism), Digital History, Born-digital Archives and Digital Forensics. He received his PhD from Ghent University and Hamburg University. Before coming to Austin, he worked at Ghent University (Belgium), Regensburg University (Germany), Antwerp University (Belgium), the University of Sussex (UK) and Hamburg University (Germany).

Preserving, analysing and understanding born-digital records as historical evidence is one of the most important challenges of contemporary historical scholarship.[1] The history of the late 20th and 21st century will be written based on born-digital primary sources, collected in personal archives, institutional collections, web archives, cloud services and social media archives. These collections are challenging for archivists and historians alike, as they are subject to technical obsolescence, large in volume, and fragmented. They are difficult to preserve, verify, appraise and contextualise with respect to their complex digital materiality, evidential status, documentary value and historical context. Digital historical scholarship relies on the development of archival workflows and standards (acquisition, archiving and documentation formats, chain of custody, curation, access) that allow verification, critical appraisal and forensic analysis of digital primary sources according to established digital forensic standards which are both sustainable and open.

There has been substantial progress by international research projects on forensic practice in digital preservation and archiving since the publication of the key study by Matthew Kirschenbaum, *Mechanisms: New Media and the Forensic Imagination* in 2008 and the subsequent publications in information science scholarship.[2] The creation of standards for web archiving services and the development of national and international web archives are important achievements for digital historical scholarship. However, digital forensic practice, workflows, formats and access are still largely confined to large memory institutions; forensic standards and forensic queriability (that is, simple methods by which users can construct valid queries) for web archives are desiderata. Among historical scholars, digital forensics and digital source criticism are still in early conceptual development, as the discipline is only just beginning to recognise born-digital records as important historical sources in their own right,[3] and awareness of the value of analysing the historical digital materiality of such records on a forensic level is limited.[4] While an emerging network of scholars in archive and information studies and contemporary history (among them Amelia Acker, James Baker, Anne Helmond, James Hodges, Lise Jaillant, Cal Lee, Trevor Owens, Corinne Rogers, Kees Teszelszky, Jane Winters and certain investigative collectives) embrace digital forensic perspectives on born-digital archives, some historians hesitate to adopt this methodological perspective. Andreas Fickers, for instance, promotes the development of digital source and data criticism

[1] Although this article is first being published in the *Journal of the British Academy*, it has originally been prepared as a chapter in the forthcoming *Proceedings of the British Academy* volume on *Materialities of the Archive in a Digital Age*, edited by Eirini Goudarouli and Andrew Prescott. Therefore the bibliographical references appear here as they were prepared for that volume.

[2] M. Kirschenbaum, *Mechanisms. New Media and the Forensic Imagination* (Cambridge, MIT University Press, 2008). A complete list of the international initiatives and projects on born-digital preservation would be too long. I would like to recognise at least some of the most impactful current and historical projects: PREMIS (https://www.loc.gov/standards/premis/), PLANETS (planets-project.eu), the OCLC born digital initiatives, Paradigm (www.paradigm.ac.uk), LOCKSS (www.lockss.org), Wayback Machine (web.archive.org), RESAW(resaw.eu), Memento (http://timetravel.mementoweb.org/) BitCurator (bitcurator.net) and Nestor (S. Dobratz and H. Neuroth, 'Network of Expertise in Long-term STOrage of Digital Resources – A Digital Preservation Initiative for Germany', *D-Lib Magazine* 10:4 (April 2004): https://www.dlib.org/dlib/april04/dobratz/04dobratz.html) LOCKSS, Wayback Machine, RESAW, Memento, and BitCurator.

[3] J. Winters, *Humanities and the born digital: Moving from a difficult past to a promising future?* Keynote at DHBenelux 2018 Amsterdam. 7 June 2018; see also J. Winters, 'Web archives and (digital) history: a troubled past and a promising future?', in N. Brügger, I. Milligan (eds), *Sage Handbook of Web History* (Sage, London, 2018), pp. 593-606.

[4] T. Ries, 'The Rationale of the Born-digital Dossier Génétique: Digital Forensics and the Writing Process: with examples from the Thomas Kling Archive', *Digital Scholarship in the Humanities,* 33.2 (2018), 391–424; T. Ries, G. Palkó, 'Born-Digital Archives', *International Journal of Digital Humanities,* 1.1 (2019), 1–11; C.A. Romein, J. Birkholz, M. Kemmann, J. Baker, M. De Gruijter, A.M. Penuela, T. Ries, S. Scagliola, 'State of the field: Digital History', *History*, 105.365 (2020), 291–312.

as an evolved version of historical hermeneutics, but also argues that contemporary historians 'do not have to turn into digital forensics (i.e. technical data specialists)' to answer questions around digital evidentiality and data transparency.[5]

The present article argues that digital forensics will necessarily become part of the historian's professional toolkit and describes digital forensic dimensions of inquiry with reference to three different types of born-digital archives: personal, institutional and online archives, drawing on records from Glyn Moody's personal digital archive, the Mass Observation Project Archive, and ways in which online and web archives enable the digital history of malware and disinformation to be investigated. The three case studies show that the ability to secure, analyse and interpret the historical digital materiality of the born-digital record in compliance with forensic standards is the key not only to verification, provenance and maintaining the chain of custody in the archive, but also to understanding the born-digital record at its historical core: forensic evidence may reveal hidden, overwise lost data, and retains latent traces of a record's creation, its history and its technological, social and cultural historical contexts. Materially losing forensic aspects of the born-digital record, or losing the ability to understand them on a forensic level, would mean the destruction of incredible amounts of latent records and substantially decontextualise documents, while the distributed, often contextual nature of digital evidence and cultures could open avenues to inquiries into histories yet to be written: technological design histories that shape our present individually and as a society, malware histories, the history of digital information campaigns. The three exemplary case studies are preceded by a brief methodological introduction to the digital forensic perspective on born-digital historical records, the concept of their historical digital forensic materiality, and aspects of digital forensics in digital preservation.

Historical digital materiality and the archive

Understanding historical forensic materiality and complexity of born-digital records and forensic traces is essential if we are to make justifiable choices about which records and digital objects to archive, which aspects of their materiality to preserve, and which methods and formats should be chosen to maintain authenticity, fixity and a documented chain of custody. Historical scholarship and humanities research relies on the ability to verify and critically appraise sources on a forensic as well as contextual level, and trace their creation, provenance, processing and manipulation history by analysing latent material features of the digital records. Hardware and software obsolescence compel memory institutions to archive proactively and take momentous decisions on highly context-dependent digital records while the technological environment is constantly changing. The arrival in the archives of increasing volumes of data and the increasing technological variability and fluidity of digital record formats (and the applications, operating systems and hardware they depend on), along with changes in forensic access and analytical methods, complicate archival preservation as well as critical appraisal. It is necessary to deal with different generations of storage media (floppy disks, hard drives, solid state drives), operating systems,

[5] A. Fickers, 'Update für die Hermeneutik. Geschichtswissenschaft auf dem Weg zur digitalen Forensik?', *Zeithistorische Forschungen,* 17.1 (2020), 157-168.

applications and devices (computers, mobile devices, flight recorders,[6] car telemetry,[7] navigation systems and dashcams, fitbits, cloud and social media forensics).[8] The historic shift from personal and office computers connected to the internet to various types of mobile devices, cloud services, social media, embedded devices in cars, household and wearables came with an extended range of device types, softwares, platforms, and historical versions that also increased the range of aspects and potential events of interest regarding data as historical evidence (malware and intrusion detection, record forgery, disinformation detection, AI forensics and preservation). At the same time, the increased access and information complexity as well as aspects of access, format and data obsolescence and interpretability pose a challenge for forensic practitioners, archivists and historians alike.[9]

A digital forensics approach to born-digital primary records involves the analysis of features and artefacts which are often latent but serve as traces of their (former) existence, creation, provenance, processing or manipulation in a specific logical and physical system context, allowing the reconstruction of certain deleted data and system processes. This evidence consists of bits (such as logical file system objects, *types*), and has a unique physical substrate (physical *token*) as a digital object entity managed by an ensemble of hardware and its controllers, firmware, file systems, operating systems and applications. A so-called bitstream-preserving image of a storage medium or memory chip, the complete array of the physically available address space of a system, memory chip or a storage medium, can be transferred to a true diplomatic image-copy ('forensic image') that preserves certain forensically relevant physical, geometrical and filesystem-specific features, including potentially deleted data. Hashing algorithms are used to assign a kind of digital fingerprint to the image-copy.[10] As logical objects, this evidence (or representation of physical evidence) is not unique, since the function of a digital object during its processing is to be copied and manipulated, as a whole or partially, across a multitude of logical and physical system states, layers, locations, even networks and the cloud, resulting in the often fragmented 'distributed materiality' of digital forensic evidence.[11] The complex relationship

[6] P. Benzon, 'Lost in the Clouds: A Media Theory of the Flight Recorder', J. Sayers (ed.), *The Routledge Companion to Media Studies and Digital Humanities* (New York, Routledge, 2018), pp. 310-17.

[7] Telemetry is the remote collection of data in order to evaluate the performance of machines ranging from motor vehicles to oil rigs and medical equipment.

[8] J. Baker, 'Digital Forensics in the House of Lords: Six Themes relevant to Historians', *Blog of the Software Sustainability Institute* (Edinburgh, Software Sustainability Institute, 2019), part I (29/03/2019), II (05/04/2019): https://software.ac.uk/blog/2019-03-29-digital-forensics-house-lords-six-themes-relevant-historians-part-one ; https://software.ac.uk/blog/2019-04-05-digital-forensics-house-lords-six-themes-relevant-historians-part-two.

Another material challenge of born-digital archiving is the environmental impact of redundant mass data storage, sustainability and availability. See K.L. Pendergrass, W. Sampson, T. Walsh, L. Alagna, 'Toward Environmentally Sustainable Digital Preservation', *The American Archivist*, 82.1 (2019), 165–206.

[9] Baker, 'Digital Forensics in the House of Lords', part I: 'According to Jan Collie from Discovery Forensics Ltd: '[...] What I am seeing in the field is that regular police officers are trying to be digital forensic analysts because they are being given these rather whizzy magic tools that do everything, and a regular police officer, as good as he may be, is not a digital forensic analyst. [...] They will jump to conclusions about what that means [...] and they do not have the resources or the training to be able to make the right inferences from those results.'

[10] Kirschenbaum, *Mechanisms*, pp. 55-6, 85-6. Forensic images do not record all the physical features of the original, but physical shadows of earlier magnetisations are in practice hardly ever exploitable. Ries, '[R]ationale', 392-3.

[11] J. Drucker, 'Performative materiality and theoretical approaches to interface', *Digital Humanities Quarterly*, 7.1 (2013), URL: http://www.digitalhumanities.org/dhq/vol/7/1/000143/000143.html (accessed 25/06/21); J.-F. Blanchette, 'A material history of bits', *Journal of the American Society for Information Science and Technology,* 62.6 (2011), pp. 1042–57. See also Ries, '[R]ationale', p. 393-4.

between the physical storage medium or memory chip as token, the materially precise representation of this data as image, and the representation as dynamic logical object as type in a file system or during processing in an application has been subject of forensic theory as well as humanities inquiry.[12]

The distributed, fragmented nature of digital forensic evidence is the reason to adopt a multi-evidential perspective and distinguish original traces from reconstructed traces.[13] These material system mechanisms, structures, and features, as well as the methodology of their forensic analysis are subject to change, because they depend on historical design decisions by engineers and businesses on the interplay of hardware, firmware, operating systems and applications in computing architectures[14] that ultimately result in historically specific latent traces of digital objects and their processing in digital evidence – historically and evidentially specific to certain storage media, operating systems, applications, and their different versions. The historical design decisions that lead to the creation and retention of forensic artefacts have often been the result of strategies to manage problems with digital materiality: signal error correction and backup mechanisms (physically unreliable storage media and network connections, general system instability), retention of deleted data on storage media (performance cost of effective deletion, non-deletion), system logs (bugs, unreliable media), modularisation, layering (maintainability), the separation of volatile and non-volatile memory, document and operating system.[15] In another recent historical shift, the use of new types of storage media (SSD, flash drives), the emergence of pervasive computing with processors inserted in everyday objects, transmission encryption, and use of cloud computing fundamentally change the forensic structure and materiality of the born-digital record.

Digital forensics: materiality and historical change

The case studies to be presented in this article illustrate how changes of digital forensic materiality become a significant feature of the historical source itself. It is important to keep in mind that these case studies are historical and domain specific snapshots which are part of a bigger picture of historical change of the digital primary record. Classic examples of the physical aspects of digital forensic materiality illustrating their propensity to change over time may include, for instance: traces of deleted files left on hard drive platters;[16] the fast save artefacts and Globally Unique Identifier numbers (GUID) embedded in Microsoft Word .doc document metadata (which encoded

[12] M. Kirschenbaum, 'The .txtual Condition: Digital Humanities, Born-Digital Archives, and the Future Literary', *Digital Humanities Quarterly*, 7.1 (2013), par. 16, URL_ http://www.digitalhumanities.org/dhq/vol/7/1/000151/000151.html (accessed 11/14/2022); see also F. Cohen, 'Putting the science in digital forensics', *Journal of Digital Forensics, Security and Law*, 6.1 (2011), 7–14.

[13] 'Multi-evidential perspective': J.L. John, *Digital Forensics and Preservation*. DPC Technology Watch Report 12-03 November 2012. Digital Preservation Coalition. 2012, pp. 43, 45. 'Original trace' vs. 'constructed trace': Cohen, 'Putting the Science in Digital Forensics', 10.

[14] Blanchette, 'Material History', 1049-51.

[15] For instance, the existence of prototypes of non-volatile random-access memory (NVRAM) and the historical *TempleOS* (2005-2017) by Terry A. Davis show that the separation of RAM and non-volatile storage, and the separation of document, file and operating system are historical design decisions, often motivated by material constraints of legacy implementations.

[16] Ries, '[R]ationale', 392-3; see also Kirschenbaum, *Mechanisms*, p. 62.

the unique MAC address of the physical networking card of the originating computer);[17] and identifying metadata that some colour laser printers print on every sheet of paper.[18]

There are other more profound changes taking place which affect the material structure of born-digital sources. For instance, the term 'document' as a documentary, evidential and archival unit, while being the dominant desktop metaphor for work processor files (Alan Kay's 'magical paper'),[19] is becoming a more and more problematic term on the digital forensic level. Not all digital 'documents' are uniformly defined as digital file objects of a certain format, as, for instance, files created by the legacy text processor Mac Write Pro and by Google Docs documents demonstrate. While a forensic file carver (a type of file recovery software) usually assumes that all files can be reconstructed by finding their specific 'header' and 'footer', Mac Write Pro files only have a 'header'.[20] Even more complex are Google Docs, which 'have no serialized representation on the local storage [i.e. document file], and cannot ever be acquired with client-side methods',[21] but are virtual documents only dynamically assembled in an interaction between cloud server and browser from a datastream of their editing history (although this can be reconstructed).[22]

A document file consists of its digital object bitstream, its physical complement, and its metadata in the file system, which is not part of the bitstream. Backup copies and temporary files can often be found or recovered throughout a preserved system. Deleted files are often recoverable from unallocated space after partial overwriting, having been saved in fragmented form, as part of a filesystem-corruption or after disk defragmentation, but will not resemble the document file format anymore. In some cases, document fragments may be preserved because of particular features of the Windows operating system such as CHKDSK which scans and repairs system errors but may introduce corruption in its CHK files.[23] The specificity of fragmentation is itself a historical feature, as since the introduction of scheduled hard drive defragmentation in Windows 7 and the schedule change for SSDs in Windows 10, automatic defragmentation regularly overwrites large amounts of deleted data in a background process, or, for that matter, reshuffles the content of unallocated blocks.

This points to one of the most momentous technological shifts of digital materiality as far as the digital forensic record is concerned: the shift from hard drives to solid state drives. Security mechanisms in historical computing systems like temporary files or partitions of the sort created

[17] R. Chen, 'The case of the UuidCreateSequential that didn't use the MAC address', *Microsoft Dev Blog*, 20/11/2019, (Microsoft Developer, 2019), URL: https://devblogs.microsoft.com/oldnewthing/20191120-00/?p=103118 (accessed 25/06/21). Until 2005, version 1 of GUID was used; in 2005, version 2 was proposed as a new standard.

[18] Electronic Frontier Foundation, 'Investigating Machine Identification Code Technology in Color Laser Printers', EFF website, 22/07/2005, (San Francisco, Electronic Frontier Foundation, 2005), URL: https://www.eff.org/wp/investigating-machine-identification-code-technology-color-laser-printers (accessed: 25/06/21). See also the more recent work by T. Richter, S. Escher, D. Schönfeld, T. Strufe, 'Forensic Analysis and Anonymisation of Printed Documents, *Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec '18)* (New York, ACM, 2018), 127–38.

[19] A. Kay: 'User interface: A personal view', B. Laurel (ed.), *The Art of Human Computer Interface Design,* (Reading , MA, Addison-Wesley, 1990), pp. 191–207, at p. 199.

[20] T. Ries. 'Pour la reconstruction des processus d'écriture numériques de Derrida grâce à la computer forensics. Reconstruction des données et matérialité numérique historique', [Proceedings of the 50th Anniversary Connference of ITEM/ENS, Paris], 2021 [in print]. See also Laurent Alonso's parser and converter *libmwaw* for legacy Mac Write Pro files: http://sourceforge.net/p/libmwaw/_list/git (accessed: 26/06/21).

[21] V. Roussev, I. Ahmed, A. Barreto, et al. 'Cloud forensics-Tool development studies & future outlook', *Digital Investigation*, 18 (2016), 79–95 (85)

[22] Roussev et al. 'Cloud forensics-Tool'. Also Ries, '[R]ationale', 396. The editing history of Google Docs can be extracted via Google's API in JSON/XML and reconstructed with tools like kumodd or Draftback.

[23] Ries, '[R]ationale', 408.

by Microsoft's Volume Shadow Copy Service (VSS) in Windows Vista and Windows 7 and 8 reflect the fact that software had to manage physical volatile memory limitations, mitigate system instabilities, and minimise possible file system corruption errors due to physically failing storage media hardware. Due to progress in hard- and software design in the consumer market and increased software stability since the early 2000s, these mechanisms kick in less often and work differently, but still operate in the background of many systems to protect data in case of system instability. Data recovery from magnetic platter-based storage media is feasible because effective data erasure was an extremely resource-intensive task, requiring the physical overwriting of blocks with a physically moving read-write head.

This situation changed with the advent of consumer-market NAND flash and solid state drive (SSD) technology – first in mobile devices around 2007, then in hard drives after 2015 – which has different performance limitations. SSD and other flash-based drives, for instance, do not need defragmentation to increase input/output speed (in fact it is usually delayed or disabled to prevent hardware obsolescence).[24] Instead, modern SSD chips need 'wear levelling', which spread write operations evenly across memory cells, and 'garbage collection', which pre-deletes memory blocks. These SSD processes operate at controller-level and file-system compression to mitigate hardware deterioration over time, resulting in large scale permanent evidence loss when garbage collection flashes memory blocks for rewrite.[25] For some generations of flash memory, retention and recovery of deleted data is still possible in reduced volumes, but data will be found fragmented and materially in different, often difficult to access physical places.[26]

Digital forensics: archiving security, verification

The archive research community responds to the important problems of ensuring the integrity, stability and fixity of born-digital records with a variety of proven methods including: the use of pervasive hashing and recurrent integrity checks in popular repository systems such as Archivematica, DSpace or Fedora Commons; use of offsite archive repositories; deployment of advanced metadata standards for born-digital such as PREMIS and Matterhorn METS; the development of distributed ledger technology including the use of blockchain,[27] and application of the LOCKSS ('Lots of Copies Keep Stuff Safe') principles and software.[28] While these technologies are in different ways suitable to mitigate or avoid 'digital preservation risks such as media failure

[24] Ries, '[R]ationale', 392–393.

[25] Y. Gubanov, O. Afonin, 'SSD Forensics 2014. Recovering Evidence from SSD Drives. Understanding TRIM, Garbage Collection and Exclusions', (Belkasoft, 2014), URL: https://belkasoft.com/ssd-2014.

[26] The loss of deleted digital evidence on NAND, Flash-based memory and SSD is considerable, but SSD drive slack and overprovisioning areas still offer some evidence recovery opportunities. Rather than standard access, chip-off forensics or access via factory access mode is recommended.

O. Afonin: 'Life after Trim: Using Factory Access Mode for Imaging SSD Drives', Elcomsoft Blog, 16/01/2019 (Elcomsoft, 2019), https://blog.elcomsoft.com/2019/01/life-after-trim-using-factory-access-mode-for-imaging-ssd-drives/

[27] M. Bell, J. Sheridan, J. Collomosse, T. Bui, A. Brown, J. Fawcett, O. Thereaux, J. Tennison: 'Using blockchain to engender trust in public digital archives'. *Proceedings iPRES 2018,* (iPres, OSF, 2018) 1-7; T. Bui, D. Cooper, J. Collomosse, M. Bell, A. Green, J. Sheridan, J. Higgins, A. Das, J. Keller, O. Thereaux, A. Brown, 'ARCHANGEL: Tamper-proofing Video Archives using Temporal Content Hashes on the Blockchain', *Proceedings IEEE CVPR Workshop on Computer Vision AI and Blockchain,* (Arxiv, 2019), URL: https://arxiv.org/abs/1904.12059 (accessed: 26/06/21).

[28] D.S.H. Rosenthal, 'What Could Possibly Go Wrong?', *Bibliothek Forschung und Praxis*, 39.2 (2015), 180–188. See also resources at https://www.lockss.org/ (Stanford University, accessed: 26/06/21).

and format obsolescence [...], as well as [...] factors such as human error, [direct] malicious attack, and organizational failure', when it comes to forensics-standard archiving,[29] there are other structural and historical threats to born-digital archives from a forensic point of view.

For instance, the unavoidable technical migration and decontextualisation of highly context-sensitive and fragile born-digital records in archiving systems, by separating the digital evidence from its physical carrier, file system and operating system context, may alter or erase forensically important characteristics and context, affect metadata, and render any forensic analysis impossible.[30] Niels Brügger has pointed out that web archives such as the Internet Archive are not authentic representations of archived webpages, because they recreate historically inconsistent page views from WebARChive (WARC) format files, which are composites of asynchronous partial crawls.[31] Following Brügger's argument, these recreated records are neither reflections of the original pageview, nor reliable, as they lack fixity, integrity and completeness, and would therefore not satisfy forensic standards.

Web archives are especially vulnerable to malicious content injection. For instance, not all web archives have transparent, explicit policies for the preservation and disabling of malware (or links to malware) that may have been embedded in the original – the UK web archive is an exception.[32] Forensic inquiries into web history were therefore limited to overt and (presumably relatively) unchanged elements of archived webpages such as web bugs (also known as 'pixel tags' and 'invisible gifs'), used for such purposes as tracking third-party use of a web site, and the spread and functionality of specific versions of GIFs.[33] Amelia Acker and Mitch Chaiet demonstrated that certain campaigns abused the 'Save Page Now' function of the Internet Archive to inject disinformation online in a persistent format, circumvent content moderation, and exploit the reputation of the Internet Archive to gain credibility.[34]

While online information campaigning and cyberwarfare date back as far as to the Yugoslav Wars (1991- 2001), recent use of social media for disinformation campaigns, the development of microtargeting, and the growth of 'deep fakes' generated by the type of AI known as Generative

[29] LOCKSS initiative: 'Preservation Principles', https://www.lockss.org/about/preservation-principles (Stanford University Libraries), (accessed: 26/06/21).

[30] Winters and Prescott are pointing to the isolation problem when a researcher uses a search engine – they might be 'missing the empire' (citing J. Assange) by finding keywords. I would extend their point and argue that this problem starts already on the digital preservation level, when files are separated from their originating technical context, e.g. by ingesting into a repository as file without context. J. Winters, A. Prescott, 'Negotiating the born-digital: a problem of search', *Archives and Manuscripts*, Special Issue: After the Digital Revolution, 47 (2019), 391-403.

[31] N. Brügger: 'Historical Network Analysis of the Web', *Social Science Computer Review*, 31.3 (2013), 306-21; N. Brügger, 'When the Present Web is Later the Past: Web Historiography, Digital History and Internet Studies', *Historical Social Research*, 37.4 (2012), 102–117. The Memento protocol may offer a partial solution. J. Winters, 'Coda: Web Archives for Humanities Research – Some Reflections', in N. Brügger, R. Schroeder (eds), *The Web as History: Using Web Archives to Understand the Past and the Present* (London, UCL Press, 2017), pp. 238-48.

[32] R.G. Coram, 'Viral Content in the UK Domain', *The British Library UK Web Archive blog*, 05/08/2015, URL: https://blogs.bl.uk/webarchive/2015/08/viral-content-in-the-uk-domain.html.

[33] J.A. Hodges, 'Forensic Approaches to Evaluating Primary Sources in Internet History Research: Reconstructing Early Web-based Archival Work (1989–1996)', *Internet Histories Digital Technology, Culture and Society*, 5.2 (2021), 119-34; T. Owens, G. H. Thomas, 'The Invention and Dissemination of the Spacer GIF: Implications for the Future of Access and Use of Web Archives', *International Journal of Digital Humanities*, 1 (2019), 71–84; A. Helmond, 'Historical Website Ecology. Analyzing Past States of the Web Using Archived Source Code', in N. Brügger (ed.), *Web 25: Histories from the first 25 Years of the World Wide Web* (New York: Peter Lang), pp. 139–55.

[34] A. Acker, M. Chaiet, 'The Weaponization of Web Archives: Data Craft and COVID-19 Publics', *Harvard Kennedy School (HKS) Misinformation Review*, 1.3 (2020), 1-11.

Adversarial Networks (GAN) have given public prominence to concern about the authenticity and verifiability of born-digital primary sources. The identification of 'manual' image forgery and deep fake images and videos is always based on bitstream-precise copies of the evidence. 'Manual' image manipulation can be identified by analysing artefacts created as a result of the use of image software. Features that can assist in identification of such 'manual' image manipulation include: investigation of the image metadata; changed colour quantisation tables; implausible photo sensor noise patterns that do not match each other or the camera used; scaling, rotation or splicing of artefacts, as well as more obvious indications of image manipulation. Deep fake detection is still in development, and is mostly based on training AI networks to detect the work of GANs by identifying co-occurrence matrices of tampered data in order to classify other tampered images by hidden features.[35] In such work, forensic archiving standards are essential: most of the digital material qualities that forensic tampering detection is based on either entirely disappear or are significantly diminished when a source is migrated to a different format or resampled in the archive.

The detection of online disinformation campaigns, propagation of specific information, news and narratives by inserting them into online ads, by means of bots and 'sock puppet' online personae or hijacked accounts on social media, is an increasingly complex task, as campaigns are often working collaboratively with networks of authentic individuals and media outlets across online platforms[36] and are at present adopting AI technology. From an archival and historical point of view, it needs to be stressed that an isolated tweet, Facebook or Instagram post cannot be proven to be part of a manipulative information campaign by inherent features unless deliberate content forgery, for example by the appearance of embedded files, can be established. Current approaches to forensic detection of disinformation campaigns are based on source-external factors and pattern recognition: either observing and analysing such features as the coordinated behaviour patterns of accounts, the appearance of typical artefacts of AI content generation, and evidently deceptive behaviour.[37] Twitter and Meta publish details of their monitoring of disinformation campaigns, and Twitter makes available datasets of disinformation campaigns to facilitate research and AI training to enable manipulation to be more easily detected.[38]

Preserving, archiving and documenting digital forensic investigation results is challenging in general. Take the example of the Citizen Lab at the University of Toronto which has undertaken numerous major investigations of such cases as the use of spyware against the Thailand pro-

[35] This is an extremely dynamic field. For an exemplary introduction to what is possible, see L. Nataraj, M. Goebel, T. M. Mohammed, S. Chandrasekaran, B. S. Manjunath, 'Holistic Image Manipulation Detection using Pixel Co-occurrence Matrices': URL: https://arxiv.org/pdf/2104.05693.pdf.

[36] T. Wilson, K. Starbird, 'Cross-platform Disinformation Campaigns: Lessons Learned and Next Steps', Harvard Kennedy School (HKS) Misinformation Review, 1-11 DOI: https://doi.org/10.37016/mr-2020-002; K. Starbird, A. Arif, T. Wilson, 'Disinformation as collaborative work: Surfacing the Participatory Nature of Strategic Information Operations'. *Proceedings of the ACM on Human-Computer Interaction*, 3, CSCW, Article 127 (2019), 1–26.

[37] Twitter: 'Platform Manipulation and Spam Policy', URL: 'https://help.twitter.com/en/rules-and-policies/platform-manipulation.

[38] Twitter blog: 'Enabling Further Research of Information Operations on Twitter', 17/10/2018 URL: https://blog.twitter.com/en_us/topics/company/2018/enabling-further-research-of-information-operations-on-twitter; 'Disclosing New Data to our Archive of Information Operations', 20/09/2019, URL: https://blog.twitter.com/en_us/topics/company/2019/info-ops-disclosure-data-september-2019.

For current research on Twitter disinformation corpora (and the specific problems with the IRA corpora released by Twitter, see section *Method*, *Data* and footnote 2): Y. Zhang, J. Lukito, M.-H. Su, J. Suk, Y. Xia, S.J. Kim, L. Doroshenko, C. Wells: 'Assembling the Networks and Audiences of Disinformation: How Successful Russian IRA Twitter Accounts Built Their Followings, 2015–2017', *Journal of Communication*, 71.2 (2021), 305–331.

democracy movement or the enforced mass collection of DNA by the Chinese government in Tibet and the Xinjiang Uyghur Autonomous Region. Imagine how the digital evidence of the Citizen Lab case studies would have to be archived in a way that is immutable and beyond reasonable doubt. Many of their cases involve analysing victim handsets or computers, which would have to be forensically imaged and archived as evidence.[39] In the *Tainted Leaks* case, for instance, where a contributor of the investigative community Bellingcat became the victim of a phishing and then of a tainted leaks disinformation campaign, the Citizen Lab used forensic investigation methods to prove the attack and the campaign. Toronto Citizen Lab exploited the URL numbering scheme of the tiny.cc URL shortener service to track down a coordinated series of phishing emails and used the PassiveTotal service's historical Domain Name System (DNS) resolution to gather further information about the phishing page domain.[40] These methods were methodologically valid, but the findings are not long-term reproducible – which would mean that any approach to archiving this research in a verifiable way would have to involve independent peer-review, documented reproduction and verification. Evidence gathered in open-source intelligence (OSINT), if archived by a memory institution, has to be preserved according to forensic standards, as historians will later have to be able to verify whether among the OSINT findings were any altered media or information campaign content.[41]

Glyn Moody's personal archive

The personal archive of the London-based technology journalist, book author and open source digital rights advocate Glyn Moody is kept at the Science Museum in London. It contains not only analogue paper files in boxes but also a digital archive – according to the catalogue stored on DVD optical disc[42] – featuring research material, drafts, digital gallery proofs and digital published versions of his journalistic work for numerous newspapers and technical journals as well as for his books *Rebel Code: The Inside Story of Linux and the Open Source Revolution* (2001) and *Digital*

[39] E.g. B. Marczak, J. Scott-Railton, N. Al-Jizawi, S. Anstis, R. Deibert: 'The Great iPwn Journalists Hacked with Suspected NSO Group iMessage 'Zero-Click' Exploit', *Toronto Citizen Lab Research Blog*, 20/12/2020 (Toronto Citizen Lab, 2020), URL: https://citizenlab.ca/2020/12/the-great-ipwn-journalists-hacked-with-suspected-nso-group-imessage-zero-click-exploit/ (accessed: 26/06/21).

[40] A. Hulcoop, J. Scott-Railton, P. Tanchak, M. Brooks, R. Deibert, 'Tainted Leaks Disinformation and Phishing With a Russian Nexus', *Toronto Citizen Lab Research Blog*, 25/05/2017: URL: https://citizenlab.ca/2017/05/tainted-leaks-disinformation-phish/. The Domain Name System (DNS) could be thought of as the 'phonebook of the Internet', which connects domain names with the current IP addresses of servers.

[41] Bellingcat has developed several methods and trains investigators on research archiving. The application Hunch.ly was developed to preserve website views for later analysis and use as evidence. Aric Toler: 'How to Archive Open Source Materials', *Bellingcat website*, 22/02/2018 (Bellingcat, 2018), URL: https://www.bellingcat.com/resources/how-tos/2018/02/22/archive-open-source-materials/ (accessed: 26/06/21).

The necessity for securing and archiving evidence reliably is clear: Bellingcat is regularly being publicly accused of manipulating digital images themselves. e.g. see material under the title 'MH17: Fake photo was used to falsely claim there was a Russian Buk [an air defence system] in Ukraine': https://web.archive.org/web/20151026072035/https://energia.su/mh17/fake_buk.html.

See also OSR4Rights, 'Putting Principles into Practice: Mock Admissibility Hearing on Open Source Evidence', Feb 2021, (OSR4Rights), URL: https://osr4rights.org/mock-admissibility-hearing/.

[42] Science Museum Archive Reference: MOOD/H, series title: Glyn Moody's digital archive; extent 1 DVD, 'The DVD comprises 9 folders of electronic files which relate to the early days of the Internet, and complement the physical archive.'

Code of Life: How Bioinformatics is Revolutionizing Science, Medicine, and Business (2004).[43] The collection includes transcripts of interviews with early Linux developers, email correspondence and locally saved websites. The collection was self-curated by the author.[44] A look at the folder structure and the file arrangement hints at aspects of Moody's overall workflow structure. Journalistic production is collected in the [1work] folder containing articles, material and drafts since September 2004 and folders by publication channel, only a few articles have their own folder. The books have dedicated project folders. The timestamp metadata of the folders suggest that the content of some folders has been redacted in the course of the archive's preparation for release to research while still on the source file system, presumably (to judge from the microsecond granularity of the modified timestamp) Microsoft's NTFS file system, while the content of the book project folders remained unchanged from 14 February 2013 (for folder [rebel code], as is shown by the timestamp granularity of an EXT3 file system, which Moody used after migrating from Windows) and 12 November 2016 (for folder [digital code of life], which has the timestamp granularity of an EXT4 file system).

Extension statistics of the archive reveal that Moody over time used several versions of Microsoft Office (e.g. doc, docx, rtf), OpenOffice (e.g. swx, odt, rtf), plain text editors (probably EMACS) and the free form database application askSam (ask) for his writing projects. The extension

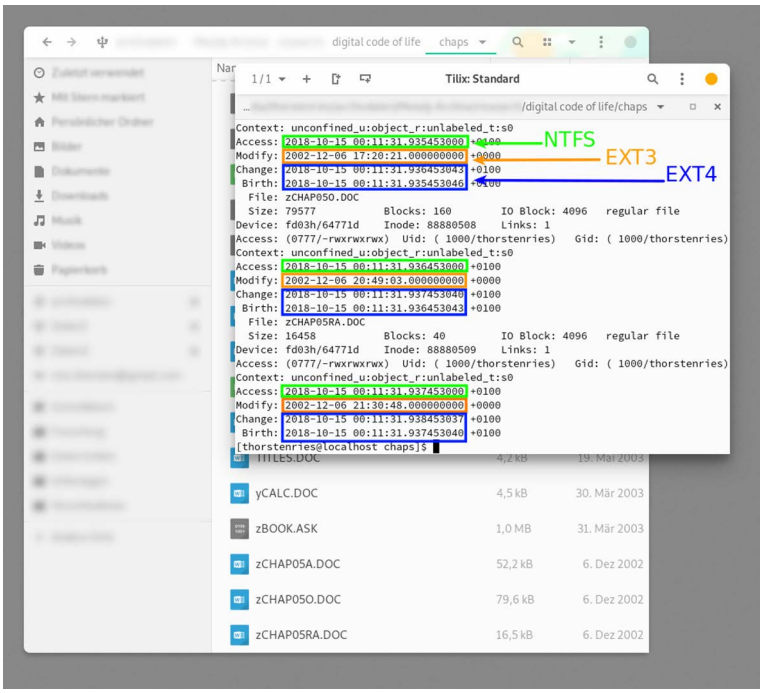


Figure 1. Glyn Moody, personal archive, files in book project folder Digital Code of Life [digital code of life], timestamp granularity.

[43] Glyn Moody, *Rebel Code. The Inside Story of Linux and the Open Source Revolution* (Cambridge, Mass., Perseus Publishing, 2001), Glyn Moody, *Digital Code of Life. How Bioinformatics is Revolutionizing Science, Medicine, and Business* (New York, John Wiley & Sons, 2004).

[44] The author had access to a copy of the archive transferred on a thumbdrive prepared by Glyn Moody, which enabled the analysis of timestamps.

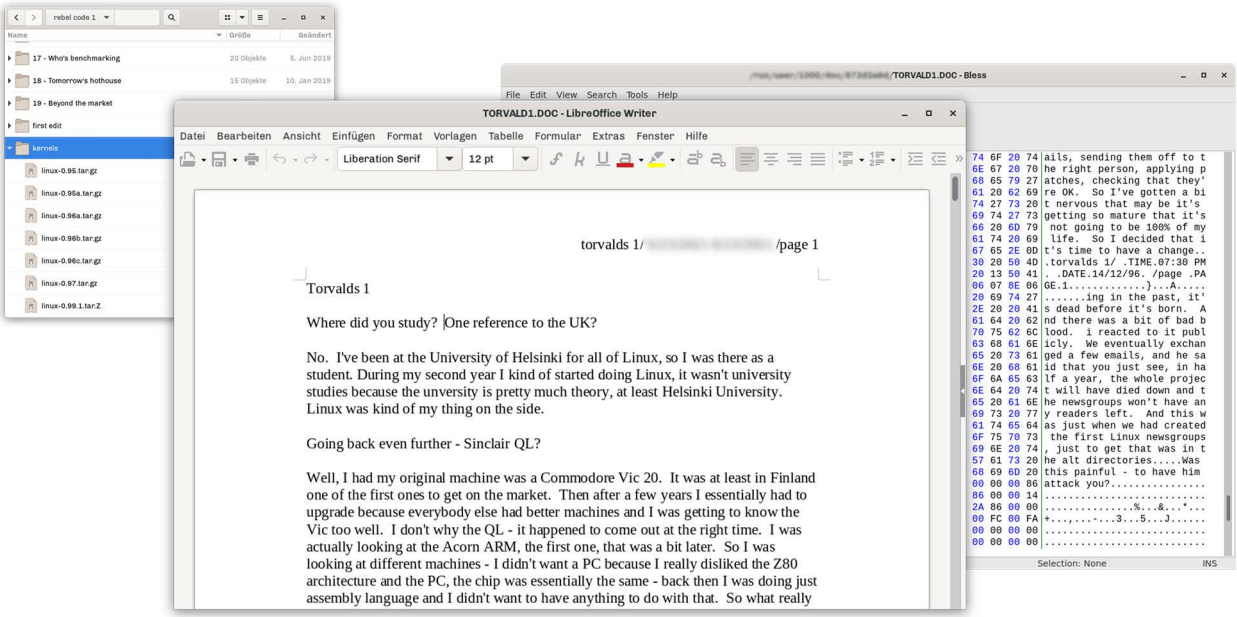


Figure 2. Glyn Moody, personal archive, files of book project Rebel Code [TORVALD1.DOC], folder [kernels].

statistics also reveal that some temporary and backup data of Microsoft Word (e.g. tmp, bak) and the plain text editor (plain text ending on '~') has been copied with the archive, which may contain draft text material.

A closer look at the folder [rebel code] shows that Moody separated the book production folders from those concerned with reviews and talks that he gave in the wake of the book. Within the book production part of the folder tree, it is most fascinating to see that he was quite organised and had separate folders for each chapter ([00 - Haunted cloisters], [01 - The coolest year], [02 - What's GNU], etc.) subdivided into two subfolders [rebel code 1] and [rebel code 2], where the former tree contains the chapter texts (sometimes in several versions), research and interview material, the latter only saved websites that were part of the research for the particular chapters. The [rebel code] folder contains, as well as a number of gallery proofs (per-chapter, and two edits), general research material and drafts,[45] transcripts of the interviews with Linus Torvalds. [TORVALD1.DOC] is about the earliest development phase of the Linux kernel and contains a few fast save artefacts documenting revision stages of the text. The [rebel code 1] folder contains a subfolder [kernels] with several of the actual Linux kernels from version 0.01 to 1.00 referred to in the first part of the interview.

Moody's archive contains several drafting process traces, some in the temporary and backup files created by the version of Microsoft Word that Moody used earlier (different versions, 1992 to 2001), others embedded in the documents themselves as track changes revisions or fast save artefacts in the datastream. In some cases, fast save artefacts show draft notes deleted from a note-taking document which illustrate, for instance, Moody thinking about a passage in the course of writing or his plans to do further research on certain aspects such as the origins of SCO group's claim to patents over the Unix system.

[45] Such as the frontispice file [FRONTIS.DOC] of 18 Dec 1999, which features the preliminary title 'REBEL CODE[.] How Linus Torvalds, GNU/Linux and Open Source are creating the post-Microsoft era.' Moody digital archive.

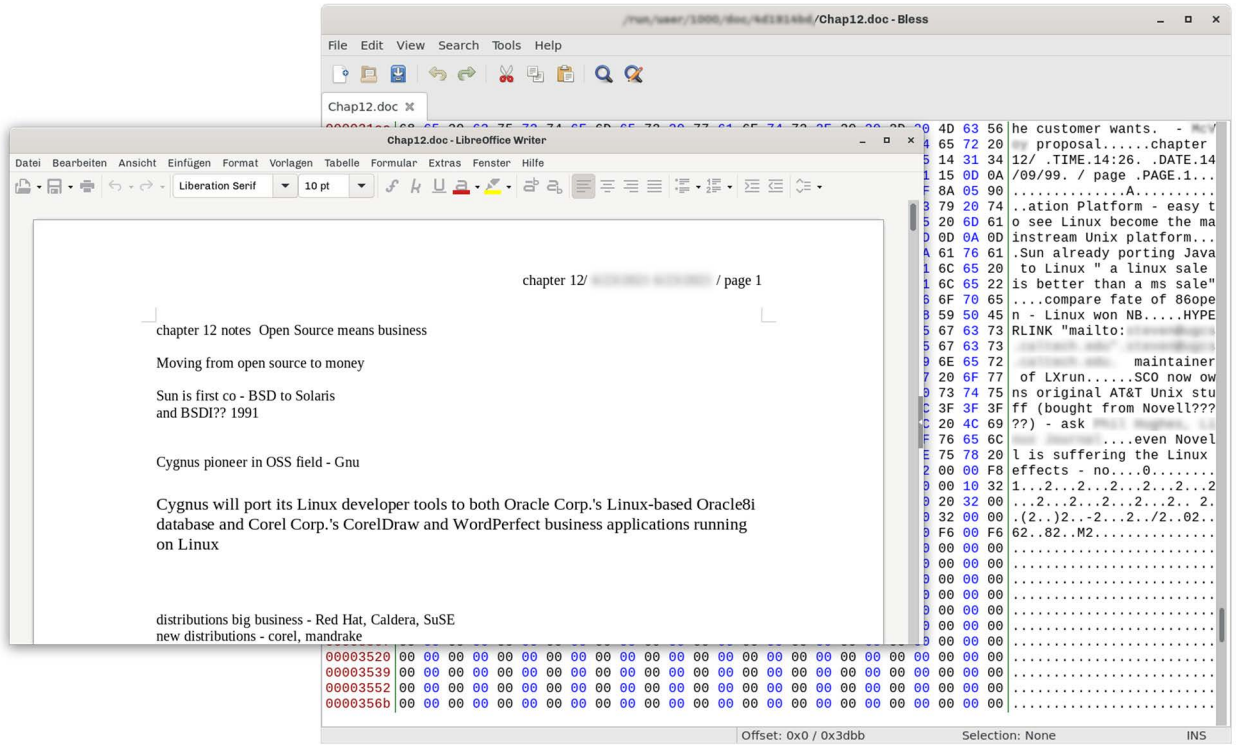


Figure 3. Glyn Moody, personal archive, deleted passages, fastsave artefacts, in [Chap12.doc].

Occasionally, temporary backup copies typically created by the plain text editor EMACS are found. These forensic traces of the writing process are relatively infrequent philological features in this archive, due to the fact that there is no forensic image of the complete hard drive.[46] Yet, these traces are historically significant and specific, as Microsoft's text processor left fast save features in documents only until around 2000.

Another historically significant find are multiple askSam database files (ask). Moody used this 'freeform' database, at that time based on Windows, as a text processor, next to Word and OpenOffice, for his book projects. askSam (1985-2008) was designed as a tool for scientific research and knowledge organisation, 'combining the ease of word processing and the power of database queries'.[47] Opening database files of legacy software retired since 2008 is not straightforward, but it is possible to use a hex editor to take a peek into the file [TRCED.ASK] which contains almost the whole book *Rebel Code*. This reveals that the askSam files contain revision artefacts and multiple copies of the database entry.

Additionally, Moody had the habit of saving multiple file versions of his working database, resulting in numerous revision stages of his working process on *The Digital Code of Life* being recoverable from the askSam databases.

The materiality of the files in Glyn Moody's personal digital archive also informs the researcher about the operating systems used, a significant issue given Moody's prominence in promoting Linux as an open operating system. The last modification timestamp resolutions of the *Rebel Code*

[46] And probably cannot be, due to the sensitivity of journalist source protection.
[47] D.J. Pfeifer, 'askSam ver 3.0', *Journal of the Association for History and Computing* 2.2 (1999), URL: http://hdl.handle.net/2027/spo.3310410.0002.215.

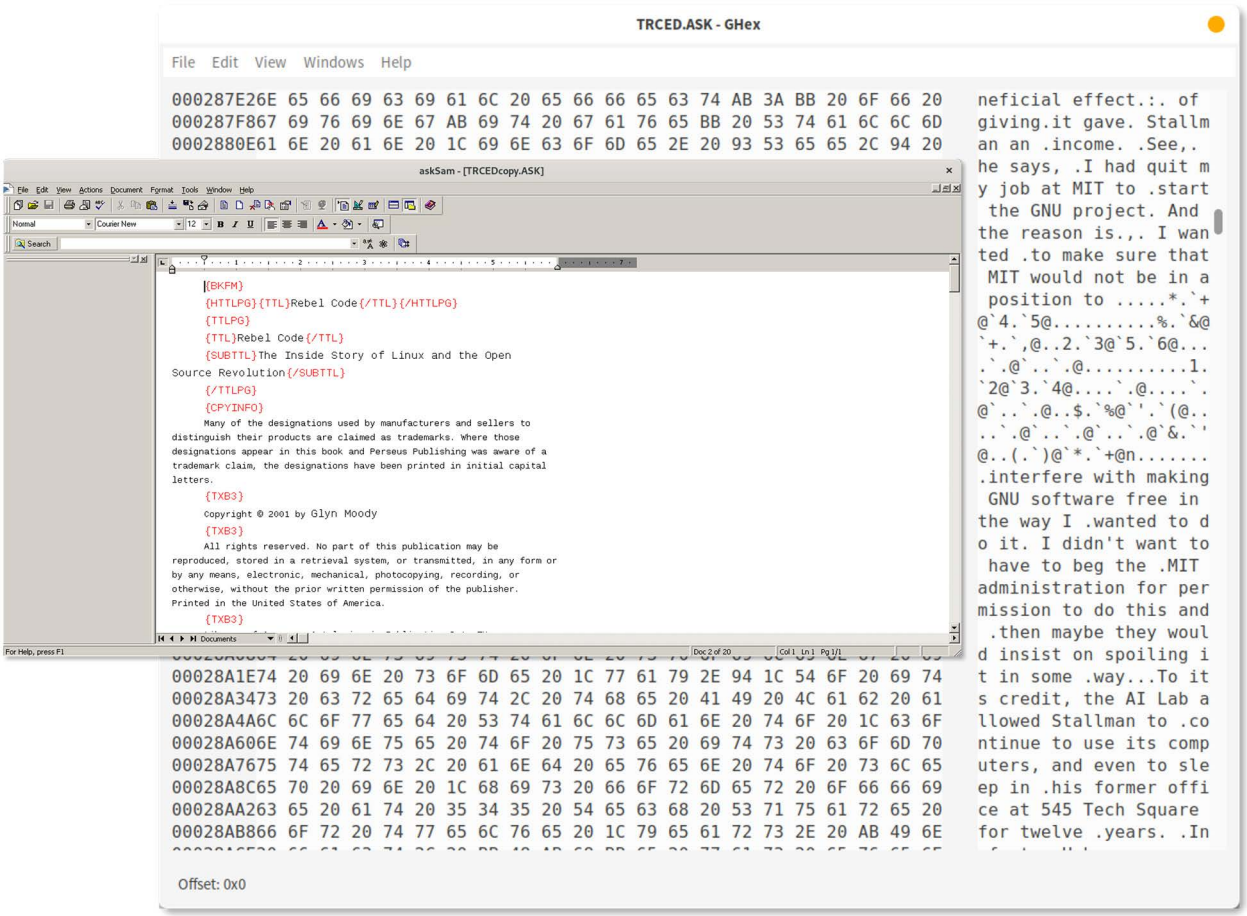


Figure 4. Glyn Moody, personal archive, AskSam file [TRCED:ASK], opened in legacy AskSam in a Win emulator, in a hex editor.

chapter files, which have a granularity of two seconds, indicates that Moody wrote this book on a Windows computer with a FAT32 file system. The timestamp resolution of the last access timestamps are 100 nanoseconds and indicate that Moody's personal archive storage was formatted with another Microsoft Windows product, NTFS, although the changed and birth timestamps were created by the Linux EXT4 filesystem Moody used when copying the files for research use. The timestamp resolution of the *Digital Code of Life* chapters (doc, ask) is somewhat mystifying at first, as their last modification date granularity is one second, unlike FAT32 (2 s), exFAT (10 ms) or NTFS (100 ns). Given this happened in 2003 the most likely candidates are the Linux file systems ReiserFS 3, EXT2/3, and Apple's HFS Plus file system (all 1 s).[48]

These findings are consistent with the scenario that Moody migrated his production system to Linux in late 2002 or 2003, and could have run either a dual-boot system with Windows and Linux or askSam via Wine, which enabled Windows applications to be run on Linux and offered 'platinum' support for askSam since version 1.0. It is relevant to note that if an archive is transferred via

[48] Some of the subfolders in the [digital code of life] folder show erroneous modification timestamps from the year '2036', which may have been caused by a synchronisation of the system with a temporarily misconfigured 32-bit NTP server or local client.

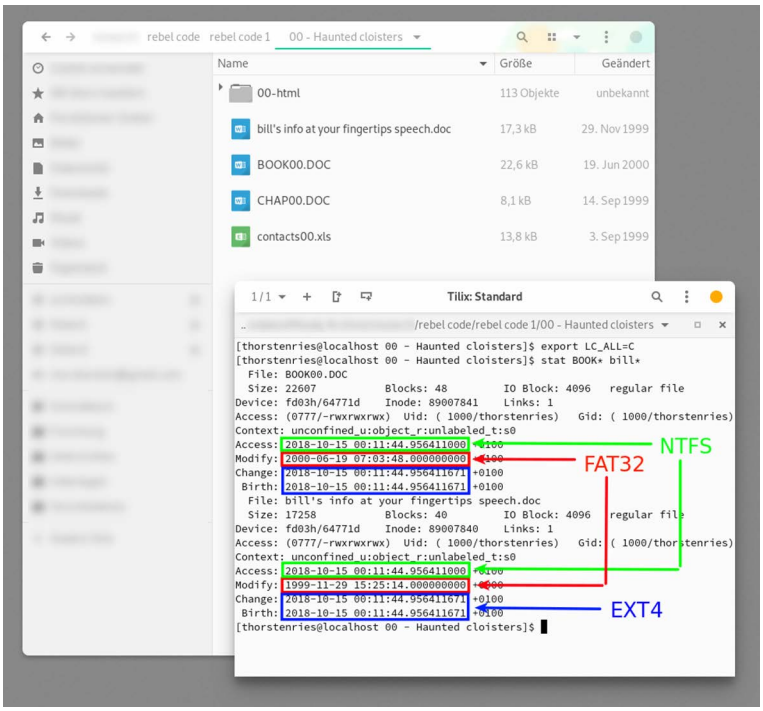


Figure 5. Glyn Moody, personal archive, files in book project folder [rebel code], timestamp granularity.

DVD, as was the case with Moody's archive, it probably is copied to a UFS file system, which has a 1 ms (1000 ns) resolution, thereby limiting the resolution of timestamps, and potentially changing the materiality of the files.

Mass Observation Project Archive

The Mass Observation Project Archive collection (MOPA) at The Keep, University of Sussex, documents everyday life in the UK in the 20th and 21st centuries. As desktop computers started to enter British homes and workplaces, MOPA started receiving born-digital submissions from the project's 'observers', creating a material archive of the UK's digital transformation. Since 1987, MOPA has received born-digital records in hardcopy form (catalogued as 'word-processed'). Since 1996, submissions were also 'received by email' as born-digital records in digital form. MOPA continuously adapted to the digital habits of the observers, and has worked hard to ensure that these records are properly archived according to archival best practice and professional standards. Since 2006, MOP has sent its directives via email; since 2015, responses to the *12th May Directives* on the social media service Twitter have been encouraged. Since 1984, MOP has sent out directives asking observers to report about their experiences with digital aspects of their lives, such as electronic banking (1984), email (1996), letters (2004), and online lives (2015). Baker and Geiringer analysed these directives in order to trace how the introduction of personal computers on the UK market changed personal lives and homes, and how it blurred workplace and private life ('Unmaking homes'): 'I always write in my study and use my Amstrad word processor', one

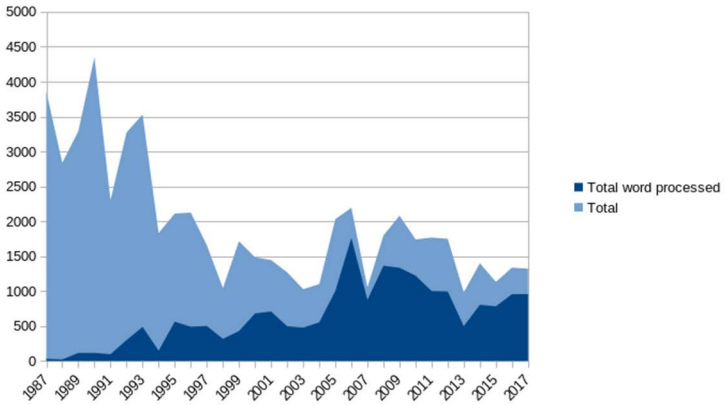


Figure 6. Mass Observation Project Archive, observer submissions on paper, word processed.

observer wrote.[49] Since 1996, the fraction of born-digital directive responses has been continuously growing.

Because MOPA accepted submissions in whatever digital formats observers were most comfortable with, it has created a timeline of word processing formats since 1996.[50] Prior to 1996, it is necessary to infer which word processors were used by observers from the non-distinctive material characteristics of the printouts. 2014 marked a shift in archival practice, resulting in a material change of the MOPA records. In order to support sustainability of the records and to comply with MOPA's obligation to assure the observers' anonymity, MOPA began – with few exceptions – to convert the submitted records to PDF/A format. This, for instance, applied even to the records that were directly written in an email client, 'Sent from my iPhone' or 'from my iPad', presumably written up directly in an email client, which was converted to PDF/A in the archive. The *12th May Directives* (2010-today), where observers record an average day, were not subject to this change in archival processing practice, delivering a more accurate view on the development of digital formats used for submissions until today. Figure 7 shows how the share of submissions in the docx format consistently grew after 2013, while pdfs decreased, and odt, pages and other alternative formats remained a stable minority.

It is often easy to distinguish a typescript produced with a mechanical typewriter from one written with an electrical typewriter by the typical imperfections of typing mechanisms, individual type characteristics and correcting methods, as can be seen from Figure 8.

[49] Mass Observation Archive, University of Sussex: Spring Directive 1991, B1989, see J. Baker, D. Geiringer, 'Space, Text and Selfhood: Encounters with the Personal Computer in the Mass Observation Project Archive, 1991–2004', *Contemporary British History* 33.3 (2019), 293-312, at 303.

[50] From 1996 on, the distribution of formats reflects the then present consumer market dominance of Microsoft (88% doc, 12 % txt, i.e. CP/M, MS-DOS txt, ASCII, ANSI, delimited) that followed the beginnings of an earlier digital transformation in the UK, which was marked by the prevalence of the early British personal computers such as the BBC Micro, Sinclair PCs (e.g. ZX 80, 81, Spectrum), Amstrad PCW and Acorn Archimedes. In 2005, formats of submissions become more diverse, with rtf, wps (i.e. Microsoft Works), pdf and, occasionally, html being submitted. The years 2007 and 2008 mark the shift towards XML-based document formats, with Microsoft's introduction of the docx format in 2007. Since 2008, docx and odt, the open source format implemented by OpenOffice, and pdf have been consistently growing in numbers relative to the doc format, which was phased out. Occasionally, in the time frame until 2013, submissions in the Mac Pages format (pages) and even txt plain text formatted files can be found, as well as submissions where the observers were technologically creative and changed the file extension to a custom, more semantic suffix (2003, 2007, 2009). This change seems to indicate observers' insecurity about doc as a universal and compatible document exchange format, first leading to an increase of rtf submissions until 2009, then of pdf until 2013.

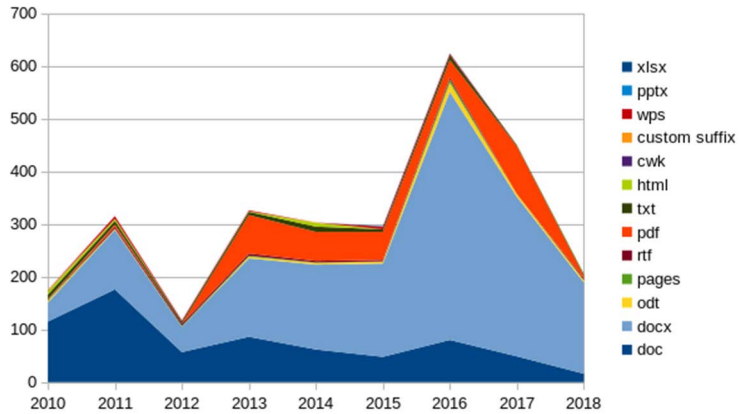


Figure 7. Mass Observation Project Archive, digital observer submissions for the 12th May Directives, by file format.

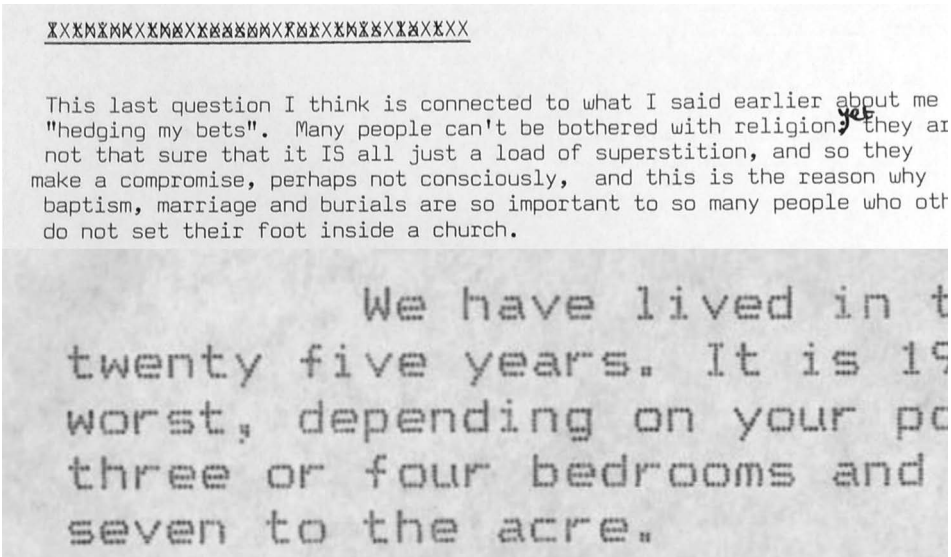
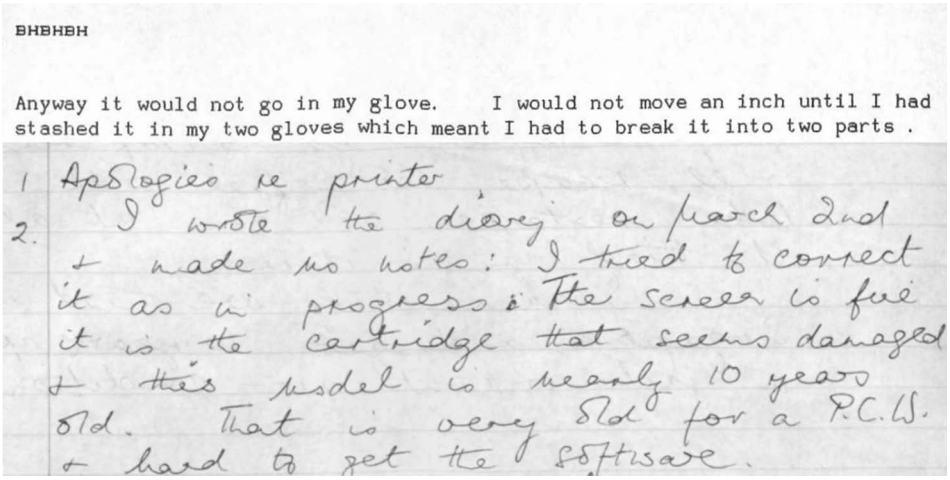


Figure 8. Mass Observation Project Archive, electric typewriter and handwritten revisions, MOPA K313, Spring Directive 1985, p. 3; matrix printer, MOPA D2206, Spring Directive 1990, p. 1.

However, telling an electric typewriter typescript from the printout from a software processor via a daisywheel printer would require accurate type samples and will often not reveal the software text processor.[51] In some cases, as well as BBC Micro, Amstrad, Sinclair and Acorn computers,[52] imported systems such as the Olivetti, Osborne or the Kaypro brands with different

[51] This is because the printout largely depended on the hardware printer, its capabilities and built-in fonts. Justification is not a secure indicator of a software word processor, as some electric typewriters, e.g. the Selectric Composer, already mastered this technique with proportional fonts, as did the IBM MT/ST, which marks the transition to word processing with its tape memory and support for bold, italics and several styles.

[52] The observer's responses to the MOP directives at MOPA feature many accounts of the UK's unique early personal computing development. For instance, observer S4743 remembers their beginnings with the BBC Micro: 'I first acquired a computer in about 1983 or 1984 at home [...]. We had a BBC micro which we shared – I remember my husband did his PhD on it (submitted 1985) and I typed my MA thesis on it in 1987-8. [...] they were much harder to operate then than today's devices. Even word-processing required some coding. And the language was very strange and over-emphatic. I remember trying to do something on a networked system in the mid 80s and a message coming up saying the procedure had been "aborted". I was quite shocked!' (S4743, Summer Directive 2015). On the same directive, H1541 remembered about their


BHBHBH

Anyway it would not go in my glove. I would not move an inch until I had stashed it in my two gloves which meant I had to break it into two parts .

1 Apologies re printer.
2. I wrote the diary on March 2nd + made no notes: I tried to correct it as in progress: The screen is fine it is the cartridge that seems damaged + this model is nearly 10 years old. That is very old for a P.C.W. + hard to get the software.

Figure 9. Mass Observation Project Archive, Amstrad PCW printer, handwritten note, MOPA A1292, Spring Directive 1995, p. 3, 7.

word processors could have been the writing tools behind the documents. However, in others the hard- and software is quite clear: In the document partly reproduced in Figure 9, a page header accidentally repeated on every page (middle, 'BHBHBH') gives away the word processing nature of the document, and the printer failure forced the observer to continue their report in longhand, mentioning their [Amstrad] PCW, which was a combined computer and word processor (Spring Directive, 1995, A1292).

In other cases, observers make explicit reference to the software, as in the case of the software developer who referenced his usage of WordPerfect on his separate work and private computer, also in Spring 1995.[53] While hard copies produced with a matrix, ink or laser printer clearly indicate a software text processor as the source, while identification of the software text processor used by typographic features, fonts and layout nevertheless remains insecure.

The statistics presented above are significant with regard to tracing the adoption of digital word processors for private writing in UK homes during the early phase of home desktop computing, but hey should be read with caveats, as the MOPA sample is statistically not entirely representative.[54]

Amstrad: 'I was an "early adopter" of the first Amstrad word-processors [...]. I do remember that the word-processing programme – Locoscript – was surprisingly sophisticated and efficient.' (H1541, Summer Directive 2015). The BBC Micro and the Amstrad were quite formative experiences for the observers, e.g. for the later Acorn developer learning to program BBC Basic (V3767, Summer Directive 2015), or observer S2207 who shares a six pages long biographical timeline, which mentions for '1993 – Got an Amstrad word processor!' as this year's most important event (S2207, Spring 1994).

[53] '[My partner] had agreed to type up an application for a friend so we spent a good hour searching the attic for the daisy-wheels for her typewriter. I spent a long time trying to copy Wordperfect from my own notebook computer onto the one I have from work. But had a lot of trouble with faulty disks or disks of the wrong size'. (W2720, Spring Directive 1995). Otherwise, no other WordPerfect formatted born-digital documents were found in MOPA. The same observer from New Zealand also remembered starting their programming carreer with punch cards: 'We wrote our programs on coding sheets, pieces of paper, which were sent off in the internal mail to the data entry section who typed them into the computer and compiled them. We got back a set of punched cards and a printout of the program. To make a correction to the program, if it was not a big correction, I would punch out the cards needed on a little hand punch. Actually I enjoyed using the hand punch, it required a little manual dexterity and had a pleasent feel when it cut the little rectangular holes in the card.' (W2720, Summer Directive 1997).

[54] The observers seem to follow technological format changes quite quickly and seem to be relatively early adopters. Possibly due to demographic skew of MOP observers ('the project's sample is weighted in favour of middle class women aged over 50 from the south of England', Baker, Geringer, 'Space, text and selfhood', p. 297). Digitally and word-processed documents have been submitted as hard copies until 1996. Observers may have submitted converted versions of their documents for compatibility reasons (e.g. txt, rtf, pdf). Some records have been edited or converted during archival processing.

Evidence of the material origins of born-digital records is often distributed across multiple levels of the digital record. As noted above, the encoding type of legacy delimited txt files tells us something about the system and word processor with which the file was edited (structural level). Equally, an automatically added line 'Sent from my iPhone' or 'Sent from my iPad' overtly indicates the original text processing context (at content level).[55] The Mass Observation Project guarantees anonymisation to the observers, and archivists need to edit and anonymise the records in order to ensure the trust relationship with the observers. In some cases, this meant conversion of an email to Word doc format, changing its digital materiality and forensic features. In a number of cases, the datastream of the records contains fastsave artefacts, which show the archivist's editing process separate from the user's, the archivist's user account name and the laser printer that was connected to the computer in the archive (see fig. 10). Fast save artefacts may show separate portions of text added to the document by the archivist, such as standard metadata about the observer or an instruction about how to record email headers, which has later been deleted (see Figure 10).

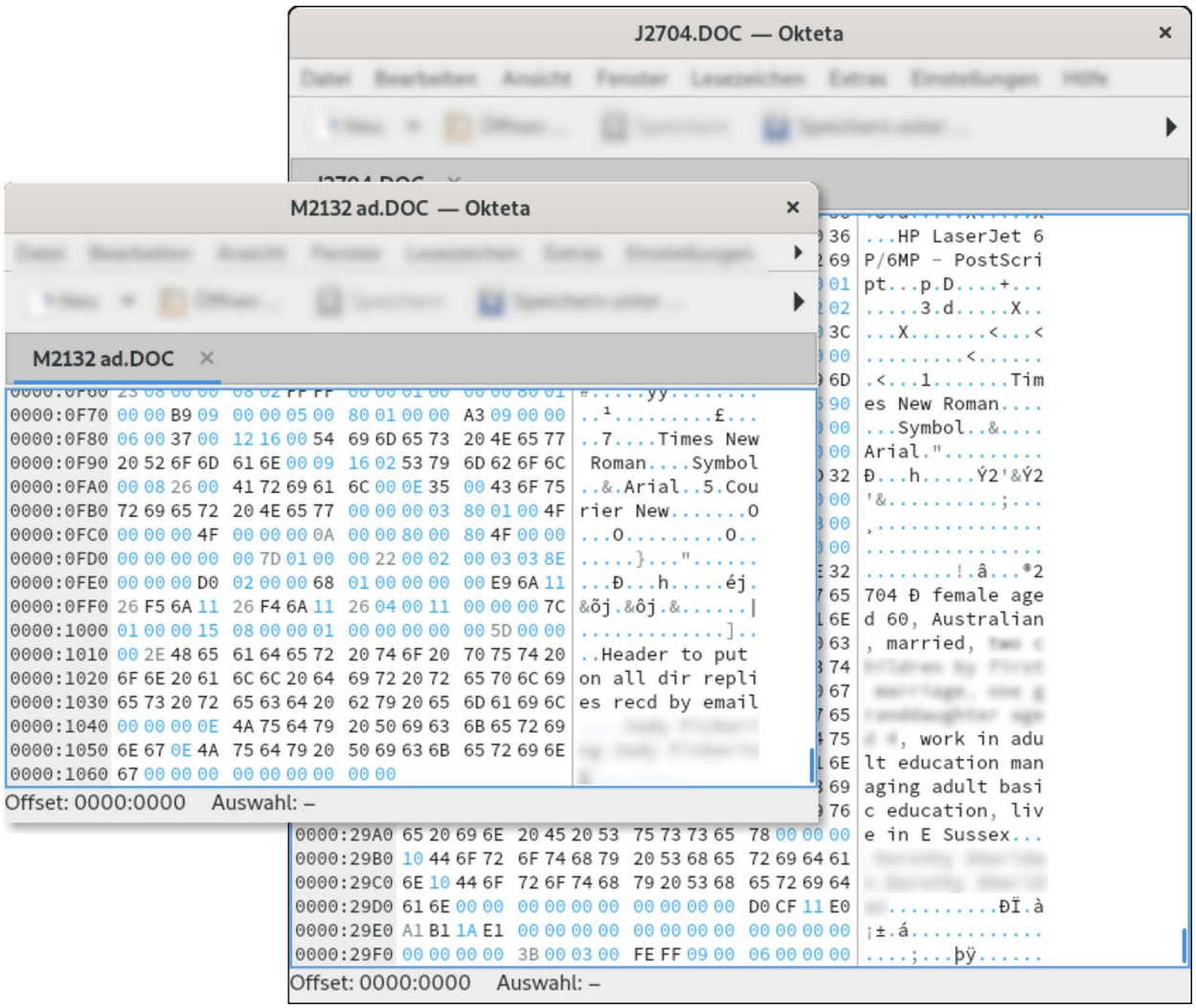


Figure 10. Mass Observation Project Archive, Word for Windows document files, [M2132 ad.DOC], [J2704.DOC]; M2132, Winter Directive 1996; MOPA J2704, Spring Directive 1998.

[55] For instance, 'Sent from my iPad' (Summer Directive 2014, MO no. 03436), 'Sent from my iPhone' (Autumn Directive 2014, G4466).

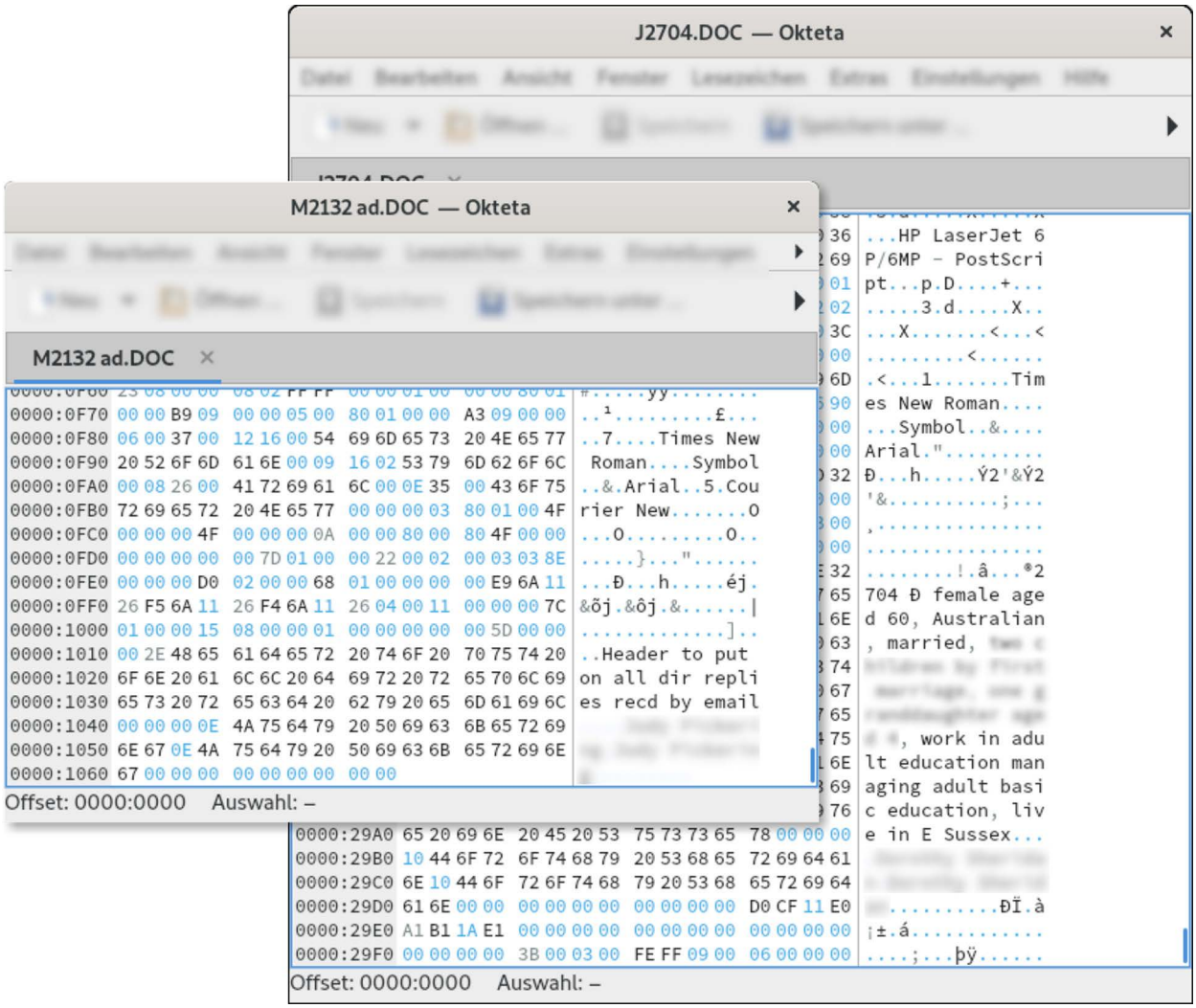


Figure 10. Mass Observation Project Archive, Word for Windows document files, [M2132 ad.DOC], [J2704.DOC]; M2132, Winter Directive 1996; MOPA J2704, Spring Directive 1998.

At the same time, the archival processing of these reports erased most forensic traces of the observer's editing process. For instance, apart from the deletion of metadata and traces which could be used to identify the observer, potentially relevant information about the original word processor, system context, revision phases and editing time was also overwritten, and draft text version fragments embedded in the datastream of the files have been erased during this process. As a result of this necessary archival editing and processing, the resulting digital record preserves more evidence of the archivists' activities than the observer's textual variants.

With Microsoft Office 2007, the Office Open XML format (docx) became the prevailing de facto industry standard for digital text documents. By 2012, half of the observers' submissions to MOPA were in docx format. This switch is significant from a historical digital forensic perspective, as documents in the docx format do not contain fastsave artefacts (on a hard drive one will still find temporary data), but the embedded ZIP-compressed XML file [document.xml] contains Revision Identifier for Style Definition (RSID) tags if the document was edited in MS Word. Other applications do not use RSID tags due to forensic identification risks.

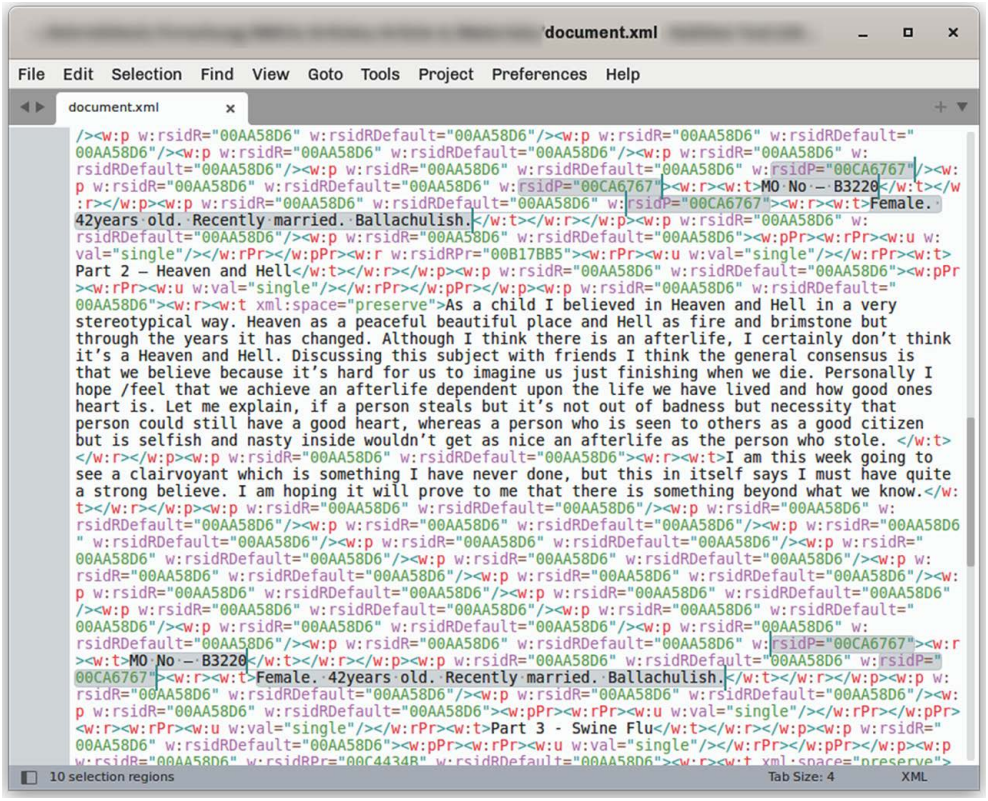


Figure 11. Mass Observation Project Archive, Word for Windows document file, [B3220abc.docx], extracted [document.xml]; B3220, Summer Directive 2009.

For instance, a record by observer B3220 for the Summer 2009 Directive on Animals was archived as docx. The application metadata tag shows it was saved by a MS Word application, version 12.0. In fact, the observer wrote this document with a different application, possibly OpenOffice or LibreOffice. The embedded RSID tags show that the archivist added metadata about the observer at two paragraphs in the document. The archivist's revisions have the RSID revision identifier [00CA6767], while all paragraphs written by the observer have the RSID tag [00AA58D6].

As MS Word alters the RSID tag it assigns to content every ten minutes, it is not plausible that the observer's text has just one RSID tag per paragraph and for the whole document – it would mean that they wrote the whole text in one go within ten minutes or copied it into the document. It can be concluded that most likely the observer wrote the text in OpenOffice or LibreOffice and saved it from this application to docx, which creates a document without any RSID tags. The document was submitted to MOPA, where the archivist opened the document with MS Word 12.0 and added the metadata, which is recorded by the RSID tags [00CA6767], but in order to make these RSID tags meaningful, it also added the [00AA58D6] tag to all observer text, and changed the application tag in the metadata (as well as the 'owner' to 'University of Sussex').

RSID tags also reveal relevant information about the writing process itself, if recorded during the writing process by MS Word. The response to the Spring 2013 Directive on Human Rights, Immigration, and the Legal System by observer C4131 shows an expected pattern of scattered distribution of RSID tags throughout the text.

Figure 12. Mass Observation Project Archive, Word for Windows document file, [C4131abc.docx], extracted [document.xml]; C4131, Spring Directive 2013.

In fact, the high density of different RSID tags in the more emotional, potentially politically and ethically sensitive sections of the text indicate frequent revisions while the author kept coming back to certain passages during different revision cycles. The sentence 'This is quite a topical subject for me at the moment as one of my close friends is going through a very difficult situation which has really brought home to me the complexities and deep unfairness involved in the immigration system' is marked by two different RSID tags and is interrupted by multiple '<w:...>' tags.

Web archives

Forensic aspects of web archives

Web history is an important strand of digital history, and web archives currently are the largest open and freely available born-digital archives. Important scholarly and archival work is being done to investigate biases and identify material of doubtful authenticity in web archives as well as search engines.[56] Archivists and scholars are doing crucial work to close gaps in the historical web record by means of web archaeology, in order to document more fully such phenomena as conspiracy websites and dark archives.[57] A relatively recent development are historical forensic inquiries into the code structure of websites, for example, tracing the way in which GIF image files are used in the design and presentation the history and distribution of culturally significant web sites and historical patterns in the distribution of web trackers and hyperlinks.[58] Malware as a cultural phenomenon and archival subject has been a significant research topic,[59] and procedures for archival processing and quarantining malware in web archives has been a notable recent research and development topic.[60] A digital forensic perspective on web history and web archives has to review web archiving standards and consider which formats and aspects of born-digital sources and evidence need to be secured in order to allow in-depth investigation, verification and critical source appraisal of historical web records. In fact, documentation of the archiving process, procedures (e.g. snapshot frequency and choice of snapshot time), and (compund) formats is essential to maintain a verifyable record.

As noted above, the material consistency and authenticity of web archive records depend on archiving format, and particularly the way in which the automated crawlers used to create archives are implemented and configured. Different browsers, settings and locations all result in variations in the way web pages are rendered and accessed. It is impossible to replicate all the possible settings that may affect the display of a web page if automated web crawlers are used for archiving. As long as web archives capture mechanisms rely on the static crawler concept, large parts of websites that dynamically serve content will remain 'dark' within web archives. For forensic verification and in-depth investigation, all elements of a served website need to be preserved as bit-precise copy, including original timestamp metadata of static website assets, external assets and aspects of the original implementation (spacer GIFs, the cascading style sheets which control display, and dynamic elements). Forensic traces of content manipulation and malware infection,

[56] Ian Milligan, *History in the Age of Abundance?: How the Web Is Transforming Historical Research* (Montreal et al., McGill, 2019); Brügger, 'Historical Analysis of the Web'; Winters, Prescott, 'Negotiating the born-digital'; S. U. Noble, *Algorithms of Oppression. How Search Engines Reinforce Racism*, (New York, NYU Press, 2018).

[57] K. Teszelszky, 'How to Harvest Born Digital Conspiracy Theories: Webarchiving Dutch Digital Culture in the Post-Truth Era', Presentation at the 2018 International Internet Preservation Consortium General Assembly and Web Archiving Conference, Wellington, NZ, slides at URL: https://digital.library.unt.edu/ark:/67531/metadc1477126/; A.W. Xavier (pseudonym), R. Amour *et al.*, 'Where in the World is Q? Clues from Image Metadata', *Bellingcat website*, 10/05/2021, URL: https://www.bellingcat.com/news/rest-of-world/2021/05/10/wat-in-the-world-is-q-clues-from-image-metadata/ (accessed: 26/06/21); L. Jaillant, 'After the Digital Revolution: Working with Emails and Born-digital Records in Literary and Publishers' Archives', *Archives and Manuscripts*, 47.3 (2019), 285-304.

[58] Owens, Thomas, 'The Invention and Dissemination of the Spacer GIF'; Hodges, 'Forensic Approaches to Evaluating Primary Sources'; Helmond, 'Historical Website Ecology'.

[59] J. Parikka, *Digital Contagions. A Media Archaeology of Computer Viruses* (New York, Peter Lang, 2007), J. Farbowitz, *More Than Digital Dirt. Preserving Malware in Archives, Museums, and Libraries*, Master Thesis NYU, 2016, URL: https://archive.org/details/16sThesisFarbowitzFinal.

[60] Coram, 'Viral Content in the UK Domain'.

embedded malware and trackers are aspects of the authenticity of a website that may be relevant evidence in a historical analysis. While many server malware infections and trackers leave traces on websites, modern malware such as members of the CDorked family are designed to avoid detection on the server itself and leave no detectable traces in the appearance of the static website.[61] Preservation formats need to be chosen that enable later critical appraisal, including checks for possible forgery. This is particularly important in the case of open source intelligence investigations. Web archives are always confronted by the problems caused when popular proprietary web applications are discontinued or blocked (usually for security reasons). Examples of discontinued closed source web applications that were at different times widely used on the web are RealPlayer, Adobe Flash and Unity WebPlayer.

Where was that server?

Contextual data of a website crawl, such as WHOIS-data, which is at present not recorded in any curated or historically systematic way by public archives, would also be essential for a representation of a web archive entry that is sufficiently detailed to allow for rigorous analysis as a historical source. WHOIS is the name of a query protocol to access databases that store information about the registered users and servers behind an internet resource including its IP and geolocation data of the server provider. While WHOIS data standards are subject to change due to changing data protection policies, such as the EU's General Data Protection Regulation (GDPR), preservation of certain metadata can be crucially important to web historical research because such forensic metadata provides crucial information about the origins and context of the data.

For instance, Anat Ben-David's 2019 essay 'National Web Histories at the Fringe of the Web: Palestine, Kosovo and the Quest for Online Self-Determination' had in part to rely on present live web WHOIS data, as opposed to historical data about servers used during the Yugoslav civil wars, in trying to reconstruct geolocations of Kosovo. Ben-David identifies this ahistorical data connection as a crucial methodological problem.[62] Indeed, although historical DNS data is archived by private digital security companies, there is currently no historical DNS archive that reaches back far enough to reliably cover the period 1998/99.[63] Ben-David's study points to a systematic lacuna in the WARC format and web archives from a forensic point of view: they did and do not record available WHOIS-data with the record, and no archive institution stores the historical connections

[61] Descriptions of how malware of the CDDorked family worked are distributed, fragmented and often closed in databases. For an introduction, see: D. Goodin, 'Attack hitting Apache sites goes mainstream, hacks nginx, Lighttpd, too. Linux/Cdorked backdoor exposes 100,000 Web visitors to potent Blackhole exploits', *ArsTechnica*, 05/07/2013: URL: https://arstechnica.com/information-technology/2013/05/attack-hitting-apache-sites-goes-mainstream-hacks-nginx-lighttpd-too/ (accessed: 26/06/21).

[62] A. Ben-David, 'National Web Histories at the Fringe of the Web. Palestine, Kosovo, and the quest for Online Self-Determination', in N. Brügger, D. Laursen (eds), *The Historical Web and Digital Humanities* (London, New Yourk, Routledge, 2019), pp. 89–109 and particularly at pp. 98-101; see also A. Ben-David, 'What does the Web Remember of its Deleted Past? An Archival Reconstruction of the former Yugoslav Top Level Domain'. *New Media & Society*, 18.7 (2016), 1103–1119.

[63] Ben-David, 'National Web Histories', pp. 100-101. These are all privately owned, commercial, closed access services. Research in these databases showed that only one service claimed to have DNS records on alb-net.com reaching back to 2001, mostly, these private archives reached only back to 2010. For Ben-David's research, the records would have been reaching back far enough. But the even bigger problem with respect to web history research is that these records are not professionally curated or governed with respect to privacy and data protection compliance by any authority. Their source and reliablity is often unclear, the archives are not sustainable.

between certain IP address ranges and geographical regions for research. As a result, we do not have reliable data as to location of the servers which created archived web pages. When country code top level domains (ccTLD, like .uk, .de, .be, .yu, etc) shift or are retired, DNS address domains change. Historical scholarship relies on historical metadata to contextualise the IP-addresses that are usually recorded within the WARC format.[64]

However, it should be borne in mind that, while such region-indicating geodata would be useful for web historical research, concrete ethical and legal data protection and privacy issues need to be considered and balanced with legitimate research interests.[65]

Malware

The high profile cases in recent years of malware, hacking and disinformation illustrate that these topics deserve the attention of historical scholarship.[66] Digital forensic analysts and scientists, cybersecurity specialists and information security businesses are developing detection methods, documentation and preservation standards for incidents and structures ranging from individual hacking, disinformation campaigns and digital mass surveillance to cyberwarfare events. From a historian's point of view, the evidenciary basis of this layer of digital history has to be regarded as precarious. Historical documentation of online threats, cybersecurity incidents and campaigns is mainly technical and kept in private databases of digital security and online threat response firms. Only small parts of this information is being shared in open source intelligence exchange formats, which are not professionally archived or curated.[67]

As an example for how difficult historical malware infection detection is when using web archives, let's have a look at the Browsealoud-Hack of UK government websites including the NHS in February 2018, which spread cryptominers onto website visitors' browsers via a redirection to an infected version of the accessibility plugin.[68] The cryptominer was not included in the UK government websites themselves, the 'dropper' was an obfuscated code sequence in the Browsealoud loader (ba.js) that linked to the manipulated version of the plugin.[69] Whether a historian will be

[64] See definition at Library of Congress, https://www.loc.gov/preservation/digital/formats/fdd/fdd000236.shtml.

[65] At the time of writing this chapter, ICANN is working on a solution to make the WHOIS service compliant with European GDPR regulations. For this reason, '[o]n 17 May 2018, the ICANN Board of Directors (ICANN Board) adopted by resolution the Temporary Specification for gTLD Registration Data.', https://www.icann.org/dataprotectionprivacy.

[66] For a general perspective, see also A. Mauro, *Hacking in the Humanities. Cybersecurity, Speculative Fiction, and Navigating a Digital Future*, (London, New York, *et al.*, Bloomsbury, 2022).

[67] Open Source Threat Intelligence (OSINT), information about potentially exploited vulnerabilities and known cybersecurity events, is available through many so-called OSINT feeds. A unifying format for shared intelligence is at this point missing, the aggregation of this information is a business model. One open source exception is the European MISP platform, which comes with its own standardised format. URL: https://www.misp-project.org/feeds/.

[68] P. Greenfield, 'Government websites hit by cryptocurrency mining malware', *The Guardian*, 11/02/2018, URL: https://www.theguardian.com/technology/2018/feb/11/government-websites-hit-by-cryptocurrency-mining-malware; J. Finkle et al., 'U.S., UK government websites infected with crypto-mining malware – report', *Reuters*, 11/02/2018, URL: https://www.reuters.com/article/uk-bitcoin-cyber/u-s-uk-government-websites-infected-with-crypto-mining-malware-report-idUKKBN1FV0W0?edition-redirect=uk; C. Williams, 'UK ICO, USCourts.gov... Thousands of websites hijacked by hidden crypto-mining code after popular plugin pwned', *The Register*, 11/02/2018, URL: https://www.tatgister.com/2018/02/11/browsealoud_compromised_coinhive/ (accessed: 26/06/21).

[69] For comparison, Internet Archive snapshots without and with obfuscated dropper code: [ba.js file without dropper code] URL: https://web.archive.org/web/20180211025804/https://www.browsealoud.com/plus/scripts/ba.js [ba.js file infected with obfuscated dropper code:] URL: https://web.archive.org/web/20180211142243/ https://www.browsealoud.com/plus/scripts/ba.js.

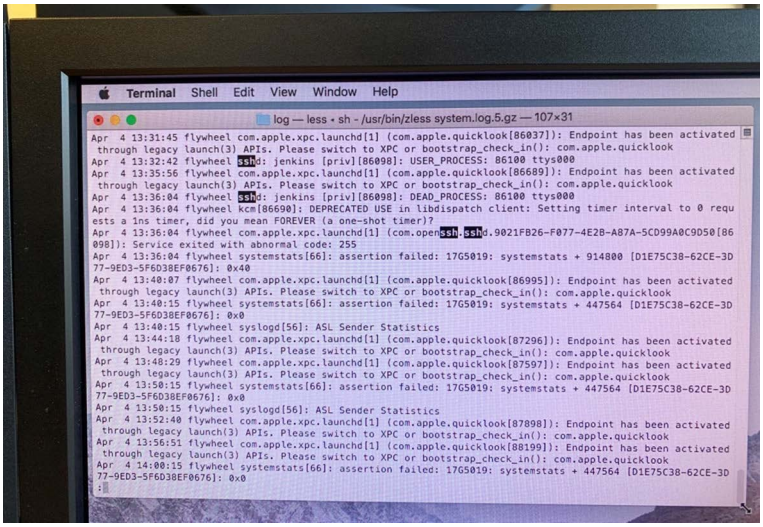


Figure 13. Moment of SSH agent being hijacked 4 April 2019; photo taken and provided by matrix.org development team.

able to spot this infection at all will depend on whether a snapshot has been taken at the time of the infection and whether the manipulated loader is included in the crawl, which is the case in the Internet Archive (e.g. for http://www.ouh.nhs.uk/)[70], but not in the UK Web Archive, where the last snapshot is dated 6 July 2017, the next 15 September 2018.[71]

The importance of historical research into such cybersecurity events, as well as the archiving issues, are illustrated by a recent unusually well-documented incident: the hack of matrix.org in March and April 2019. Matrix.org is a distributed, open source chat and instant messaging platform. During the workshop *Born-Digital Archives and Digital Forensics – Where are We Now?* on 15 March 2019 in London,[72] which invited the audience to take part in the conversation via matrix.org, the otherwise exceptionally secure online chat platform was in the process of being compromised by a hacker. The hack started on 13 March with the compromise of the vulnerable version of the Jenkins continuous integration server, which is part of the matrix.org infrastructure.[73] On 4 April, the attacker gained access to the production infrastructure by hijacking a forwarded SSH agent. An image taken at matrix.org headquarters in London captured this moment (Figure 13).

On 9 April, a private security researcher informed the matrix.org team via a Twitter direct message about a vulnerability in their version of the Jenkins server. On 10 and 11 April,[74] the Matrix team removed Jenkins and the attacker's access to compromised machines, rebuilt the

[70] [Snapshot Internet Archive 11 Feb 2018] URL: https://web.archive.org/web/20180211154439/http://www.ouh.nhs.uk/.
[71] [Snapshot UK Web Archive 6 Jul 2017] URL: https://www.webarchive.org.uk/wayback/en/archive/20170706221031/http://ouh.nhs.uk/ (accessed: 26/06/21).
[Snapshot UK Web Archive 15 Sept 2018] URL: https://www.webarchive.org.uk/wayback/en/archive/20180915121537/https://www.ouh.nhs.uk/.
[72] Event: *Born-Digital Archives and Digital Forensics – Wat are We Now?*, 15 March 2019, School of Advanced Study, University of London, organised by T. Ries, J. Baker, J. Winters, URL: https://www.sas.ac.uk/events/event/19289.
[73] Matthew Hodgson, 'Post-Mortem and Remediations for Apr 11 Security Incident', *matrix.org blog*, 08/05/2019, URL: https://matrix.org/blog/2019/05/08/post-mortem-and-remediations-for-apr-11-security-incident.
[74] Security notice by Matrix.org Team, 'We have discovered and addressed a security breach. (Updated 2019-04-12)', *matrix.org blog*, 11/04/2019: URL: https://matrix.org/blog/2019/04/11/we-have-discovered-and-addressed-a-security-breach-updated-2019-04-12.

infrastructure and restored the matrix.org home server. During 12 April, the matrix.org home page was defaced using a cloudflare API (application programming interface) key exfiltrated during the hack, which made clear that the hacker gained access to encrypted passwords, triggering matrix.org's response to force users to change their passwords. According to matrix.org, the attackers tried to cover most of their tracks, but not all.

Several things are remarkable about this chain of incidents from a historical point of view. The matrix.org team gave an example of a swift, effective and, above all, unusually transparent response to the attack with timely information and an exhaustive post-mortem report. At the same time, the forensic evidence and detailed findings of the team's investigation will remain secret, and therefore inaccessible for historical research. Second, the attackers themselves (and potentially their audience) used several social media and forum accounts to publish their own public post-mortem report about the hack on GitHub,[75] which is probably the most unusual aspect of this incident. Third, even though the attack itself has not been attributed and the exact motives behind the hack remain unclear, the documentation of its forensic analysis and the context of the hack add a historical perspective. During its development phase, matrix.org became known as a promising decentralised, optionally self-hosted and encrypted open source instant messaging and chat alternative to closed source social media platforms and messaging services. matrix.org was at the time of the attack just about to reach version 1.0 and become a standard solution for large organisations with high security standards. The attack used a short window of opportunity (an unpatched component) to hack the system and publicly expose the vulnerability rather than exploiting it, thereby potentially influencing the conversation about the security of open source software and services.

The attack on matrix.org was targeted and relatively limited in scope, but in its publication strategy seemed to target the discourse around the software and its concept. The web history to be written will have to include the history of its technical infrastructure, its structural development, and also the attacks committed on it. The currently unwritten history of the web, for instance, would put the 2013, 2019 and 2021 attacks on the backbone internet technology PHP in a broader web historical context. The term PHP originally stood for Personal Home Page but now stands for PHP: Hypertext Preprocessor. PHP was originally created by the Danish-Canadian programmer Rasmus Lerdorf in 1994 and is one of the most popular script languages, mostly executed on web servers, driving large parts of world wide web and web-based applications. The community website php.net is the centre of its development, and is connected to the community as well as code repositories.

The October 2013 attack targeted php.net, and made the website spread malware to visitors, mostly PHP developers.[76] The malware was not detectable by any website code, which gave rise to initial speculations on security channels that a version of the aforementioned CDorked family had been involved[77] – in fact, later reports indicate it was DGA.Changer, a malware that avoids detection

[75] The original post (URL: https://web.archive.org/web/20190412142447/https://github.com/matrix-org/matrix.org/issues/created_by/matrixnotorg) has been removed from GitHub, the content has been reposted here: URL: https://github.com/matrix-org/matrix.org/issues/371.

[76] D. Goodin, 'Hackers compromise official PHP website, infect visitors with malware (updated)', *ArsTechnica*, 25/10/2013, URL: https://arstechnica.com/information-technology/2013/10/hackers-compromise-official-php-website-infect-visitors-with-malware/ (accessed: 27/06/21).

[77] Goodin, 'Hackers compromise official PHP website': 'Ars has covered several varieties of malware that target webservers and are extremely hard to detect.', with links to CDorked reporting. The internet archive did not snapshot the php.net website during the hack, the Snapshots are before the intrusion and after the repair (attack 22-24 Oct 2013, snapshots on 20 and 25). If this had been a CDorked infection, it would have been detectable only by a specific response test.

by the virtual machines known as 'sandboxes' used to identify malware.[78] The January 2019 attack targeted the community-driven framework and distribution system PHP Extension and Application Repository (PEAR). The attackers replaced PEAR's package manager, which is widely used by web application programmers, leaving the site infected, PHP programmers and their applications vulnerable for several months.[79] The most detailed information about this incident was to be found on online threat open source intelligence channels, such as the European MISP – the problem is, of course, that these OSINT feeds are not archives. Another supply chain attack on php.net and its GIT code repository (git.php.net) was attempted in late March 2021, malicious commits to the repository were made (although they supposedly did not end up in consumer code), possibly the user database was leaked.[80]

A web history that includes online threats would have to analyse incident, vulnerability and software analysis reports at scale in order to assess the historical dynamic. The examples given, Browsealoud, matrix.org and php.net, may at larger historical scale appear less impactful than other incidents such as the emergence of Stuxnet or the 2018 cyberattack on the Winter Olympics in South Korea.[81] However, it is the vast amount of lesser known cybersecurity incident and vulnerability reports, forum communications that reveal the history of malwares as part of the material history of internet infrastructure. The supply chain attacks on php.net are part of a technology-historical thread of the history of the 2020 United States federal government data breach (03-12/2020), which – as a recent example of sophisticated supply chain attacks – reportedly involved a combination of exploits and breaches of multiple software providers' products (e.g. Microsoft, SolarWinds). The fact that the last two paragraphs relied on journalism sources is in itself symptom of a gap in the archive: reconstructing this history relies either on forensic evidence or the analysis of archived, citable OSINT online threat feed reports, which could not be referenced in this case.

Conclusions

The present article outlined challenges and opportunities of digital forensic perspectives on born-digital archives for archivists and historical scholarship. If historians are to critically appraise primary sources and establish the circumstances of their creation, provenance, processing history, so as to facilitate the identification of forgeries, fakes and disinformation, it is essential to explore the forensic history of the material creation of these records. The precondition for this type of research is preservation according to digital forensic standards, which enables critical forensic

[78] D. Goodin, 'Hackers who breached php.net exposed visitors to highly unusual malware', *ArsTechnica*, 18/12/2013, URL: https://arstechnica.com/information-technology/2013/12/hackers-who-breached-php-net-exposed-users-to-highly-unusual-malware/ (accessed: 27/06/21). The promised full post mortem of this incident does not seem to be public (URL: https://www.php.net/archive/2013.php, accessed: 27/06/21).
[79] D. Goodin, 'If you installed PEAR PHP in the last 6 months, you may be infected', *ArsTechnica*, 23/01/2019, URL: https://arstechnica.com/information-technology/2019/01/pear-php-site-breach-lets-hackers-slip-malware-into-official-download/.
[80] D. Goodin, 'Hackers backdoor PHP source code after breaching internal git server', *ArsTechnica*, 29/03/2021, URL: https://arstechnica.com/gadgets/2021/03/hackers-backdoor-php-source-code-after-breaching-internal-git-server/, see also URL: https://externals.io/message/113981. As a result, php development announced to move their main repository to GitHub.
[81] A. Greenberg, 'The Untold Story of the 2018 Olympics Cyberattack, the Most Deceptive Hack in History', *Wired*, 17/10/2019, URL: https://www.wired.com/story/untold-story-2018-olympics-destroyer-cyberattack/.

analysis of digital material traces of a record's history, provenance, functionality and the way it was used, as well as analysis of latent evidential features and forensic recovery, based on a verifiable chain of custody. Not only preservation, but also the ability to access, analyse, understand and appraise born-digital records in the context of their historical digital materiality critically depends on a coordinated, cross-sector and interdisciplinary research and development process of the archival sector, the information science, digital forensic, history, and culture studies departments, as well as the software and digital forensic industry. The challenges of increasing variability, diversity and fluidity of hardware, applications, operating systems, platforms and formats, for digital preservation and maintaining forensic tools, capabilities and knowledge – as well as ethical and ecological questions - are substantial, but the history of the digital age cannot be written without an accurate understanding, verifiability, critical analysis and appraisal of these records on a forensic level of their historical digital materiality. Additionally, there are histories to be written that cannot be traced without in-depth knowledge about the historical digital forensic materiality of our records (and that they are worth preserving), from the social history of early personal computing to hacking, disinformation and AI histories of our present.

Acknowledgements

The research for this article was funded by the Marie Skłodowska-Curie actions (MSCA, Horizon 2020) program, project *Digital Forensics in the Historical Humanities: Hanif Kureishi, The Mass Observation Archive, Glyn Moody*. Additional funding was provided by the Research Foundation Flanders, project *Hard Drive Philology / Source Code Philology: Tracing the digital writing and coding process in German literature*.

The author wishes to thank especially Glyn Moody, Jessica Scantlebury, Fiona Courage (Mass Observation Project Archive, University of Sussex Libraries) and Matthew Hogson (matrix.org) for making this chapter possible by opening the(ir) archives and being helpful on so many levels to enable this research. Furthermore, he would like to thank James Baker, Felix Freiling, Matthias Vallentin and *Control-F*, as well as the editors (Eirini Goudarouli and Andrew Prescott) of the forthcoming *Proceedings of the British Academy* volume on *Materialities of the Archive in a Digital Age*, for invaluable input, review and feedback.

To cite the article: Ries, T. (2022), 'Digital history and born-digital archives: the importance of forensic methods', *Journal of the British Academy*, 10: 157–185. https://doi.org/10.5871/jba/010.157

Journal of the British Academy (ISSN 2052–7217) is published by
The British Academy, 10–11 Carlton House Terrace, London, SW1Y 5AH
www.thebritishacademy.ac.uk

The Editors welcome 'Responses' to articles published in the Journal of the British Academy. Offers of 'Responses' should be sent to journal@thebritishacademy.ac.uk