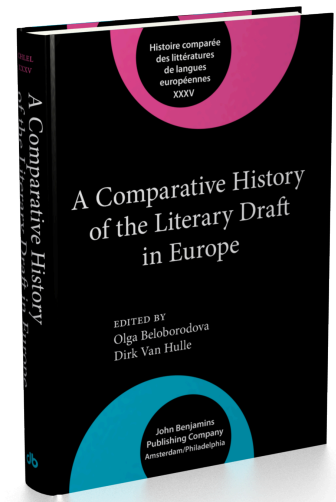
2.1.2

Born-digital documents

Isabelle Van Ongeval

https://doi.org/10.1075/chlel.xxxv.28van



Pages 410–416 of
A Comparative History of the Literary Draft in Europe
Edited by Olga Beloborodova and Dirk Van Hulle
[**Comparative History of Literatures in European Languages**, XXXV]
2024. xiv, 550 pp.

2.1.2 Born-digital documents

Isabelle Van Ongeval

These are exciting times for literary archivists. Nowadays, a writer's legacy presents itself as large chunks of hybrid and disrupted data, partly analogue and partly digital. This chapter reflects on the challenges faced by literary archivists in acquiring, managing, and unlocking born-digital archives of writers, publishers, and literary organisations. There is a real threat of gaps emerging in collections of literary archives, because of the hybrid way a writer writes in the twenty-first century, as well as the unpreparedness of archival institutions. Literary archives are in need of technical skills for dealing with born-digital content in many forms, from obsolete carriers to online content. Finally, they need to work on the writers' awareness of the fragility of their digital content. Overall, there is a strong need for more and structural collaboration with IT professionals, academics, and the makers of literary archives in order to secure, manage, and unlock born-digital literary archives.

Keywords: literary archives, born-digital texts, digital forensics, cross-disciplinary research, digital curation practices

Almost everyone is confronted in everyday life with the rapid technological developments that are felt throughout society. The impact of technology is also palpable in the literary field. The making and experiencing of literature is gradually shifting to digital spaces. Technology influences what and how authors create, preserve and disseminate their ideas and work. Today, almost every published work of literature is made digitally. Texts are written with a word processor, stored on a hard disk or other storage medium and can be accessed via a number of software programs on a PC, tablet or another device. No doubt, texts will still be written by hand or perhaps even on a traditional typewriter, but at some point those texts will be processed via a word processor before the publishing process. This hybrid and complex genesis context of literary work complicates the preservation and future research of literary heritage.

Ideally, literary archives should reflect the complete work and life of a writer: preparatory notes, annotated books, literary drafts, correspondence with publishers and authors, journals, photos and videos, scrapbooks and so forth. It's up to the archivist to make sense of the variety of documents and the wide range of formats, objects and carriers in order to appraise and preserve the writer's legacy and make it accessible through archival descriptions and indexes. The advent of personal computers and above all the increase in online activity by writers and readers have highlighted just how fragmentary and incomplete a literary archive is and remains. The digital traces left on computers or other data carriers and the online presence and literary activities of a writer cannot be captured instantly and entirely. The digital working method also influences the integrity and completeness of an author's archive. The precursors to literary products take many different forms and are stored on many different data carriers; some

https://doi.org/10.1075/chlel.xxxv.28van

may not even survive the creative process. The number of digital variants of individual digital literary texts in the Letterenhuis collection is disappointingly low.[1] There is a possibility that there are still some print copies of digital variants available in the paper archive. This fragmentation of the literary archive across various folders, locations, data carriers and applications presents an enormous challenge for both archivists and researchers. The working method is a hybrid one: paper, cloud, external hard drives, USB sticks, laptops, tablets, smartphones and so on. Communication and networking between writers, literary organisations and (communities of) readers are also gradually shifting online (writers' collectives, blogs, forums, online communities).

Though it may sound contradictory, the exponential growth in the number of data carriers and channels for literary creation gives the impression that everything is recordable, that everything has the potential to be archived and saved for later. In future, digital heritage will be like today's incidentally uncovered fragment of hieroglyphics or piece of graffiti on the remnants of a sixteenth-century wall in Venice. Technology creates the illusion that we have at our fingertips an entire body of literary heritage, but in spite of this technology we are still beholden to chance, the individual choices made by writers and the ravages of time.

Whereas under ideal (that is, dry and fire-free) conditions paper archives can easily endure for centuries, the situation is not quite so simple for born-digital heritage. It all starts with writers, who must have some level of technical knowledge to maintain control over the digital material they create: online management, version management, structured saving of files and regular back-ups, consistent file naming, and so on. Moreover, the digital creative writing process does not occur in a vacuum. Equipment, operating systems, software and servers enable writers to produce digital works, but the choice of applications and equipment also limits the future storage and accessibility of their literary legacy.

The complex digital environment in which literary creations come to life results in "black boxes": whereas in the past, archival institutions would once receive cardboard boxes filled with sheets of paper covered in handwriting, they are now faced with computers, floppy disks, USB sticks and hard drives that are not immediately recognisable, readable or identifiable. Instead of acid-free paper and boxes and the usual training, archivists now require technical skills, equipment and software to be able to read this "new" born-digital archival material and make it accessible. If they do not have the expertise or the appropriate skills in-house, archivists tend to classify digital archives as not accessible (Dean and Tuomala 2014: 149). It is not only these technical constraints, but also concerns about privacy and copyright restrictions that cause archival institutions to assess their digital archive as a closed collection. The primary concern for archivists is providing a secure and appropriate environment for long-term storage of born-digital literary heritage.

1. The Letterenhuis is the largest literature archive in Flanders (located in Antwerp, Belgium), whose mission is to safeguard the Flemish literary heritage. For more information, see https://letterenhuis.be/en/page/letterenhuis-nutshell.

The limited understanding of creative work in a digital context and a noticeable underappreciation of literary heritage in digital form likewise pose a threat to the long-term preservation of that literary heritage. Often, writers assume that the creative work they undertake in a digital context is not eligible for the collections of literary heritage organisations. As a result, many computers, floppy disks and USB sticks are simply lost, as is a great deal of online content. Following the final transfer of the literary archive of the Belgian author Ivo Michiels (1923–2012) in 2019, for example, it appeared that the writer's family had taken his computer and floppy disks to the rubbish tip. A lack of interest as well as limited understanding of digital management and storage among writers, publishers and literary organisations shows that literary archivists should intensify contacts to raise more awareness of the fragility of their digital content. There is also a need for awareness-raising and technical up-skilling among archivists and researchers so that they can recognise, appraise, preserve and provide access to born-digital collections.

The born-digital collection at the Letterenhuis started off in 1998 with the donation of the literary archive of poet, columnist and founder of the literary journal *Nieuw Wereldtijdschrift* Herman de Coninck (1944–1997). The archive contains handwritten and typewritten documents, correspondence, journals and notebooks, photos, and 218 floppy disks (3.5" and 5.25" disks). At that time, the curation practices and workflows in the Letterenhuis were geared solely for processing analogue or physical artefacts. Back then, the digital collection at the Letterenhuis consisted purely of digital reproductions or derivatives of tangible archival material for presentation and consultation online. Thus, 1998 marked the first time the Letterenhuis received an archive that was created in a hybrid environment. The technical knowledge and infrastructure required to read obsolete digital data carriers was limited at the time. Since there just so happened to be a few computers in use that had built-in 3.5" disk drives, the material on these data carriers alone was accessed. Unlike the handwritten works and letters, the disks were not described or catalogued. They were considered as inferior to the clearly recognisable handwritten documents, which instilled awe and a sense of history. The digital content was retrieved from the obsolete carrier, which was no longer of any relevance to the archive, and stored on a fileserver. Since then, de Coninck's papers have been fully catalogued, without appraising the digital files.

In the recent decade, the Letterenhuis has experimented and developed a digital program thanks to use cases from fellow archival institutions and the evolution of emerging technologies (Colavizza et al 2021). Old equipment and computers were actively collected, and a host of freeware and open-source tools were evaluated for analysis and identification of text files. The existing workflows for acquiring and processing archival material were altered to process born-digital archives. The Letterenhuis now boasts a small forensic lab for reading, retrieving and analysing data from obsolete carriers. Over the past few years, policymakers have also begun turning their attention to digital developments, vacancies have opened up for IT-related or technical positions (for which there is a shortage of candidates in the heritage sector), and (modest) budgets have been made available for creating an infrastructure to handle the influx of digital archival material and ensure that it is stored permanently.

Digital archival material is acquired in a number of ways, be that in the form of a random collection of disks in archive boxes, the donation of a laptop owned by a deceased author, explicit digital transmission of a publication file (for instance, via Wetransfer) or the copying of a hard disk or a mailfile (a file that stores e-mails) directly from a computer or laptop that is on temporary loan from the writer. Modern technology makes it possible to transfer archives more quickly and more easily. When an archive is transferred, the writer keeps their material, and can even opt to transfer only selected material.

After the transfer of a digital archive, the first dilemma presents itself: (how) can we reconstruct and capture born-digital material in its original form and at the same time remove as many barriers to access as possible, all without damaging the integrity of the archive? Technological advancements are rarely driven by historical awareness. When software, apps and devices are designed, usage or data history are not usually considered important factors. This is in sheer contrast with the core mission of archival institutions, which is to preserve archival heritage in its most authentic form, regardless of the medium, and to document the context in which it was created and stored. In a digital context, applying the archival principle of "respect des fonds"[2] can be a challenge (Thibodeau 2016). The fragility and volatility of digital information results in unstable archival material, whereby the smallest intervention or manipulation of data may be irreversible. Thanks to the technique of disk imaging, we can now create a snapshot of the data in its original form. Creating an image allows us not only to produce a back-up copy, but also gives us additional options for accessing problematic data carriers at bit level.

Data on certain digital carriers cannot be accessed and transferred without special equipment. Floppy disk drives or even CD drives are no longer installed on our laptops or PCs. Moreover, today's (commercial) operating systems are no longer able to read files stored on a disk. The infamous Windows error message "You need to format the disk in drive A: before you can use it. Do you want to format it now?" is misleading: it makes it seem as if the data carrier is no longer accessible as a result of corrupt data or damaged sectors. So the floppy disk is put into storage, or sometimes even destroyed. However, in most cases the data can still be retrieved from the disk. For example, when using a computer with the open-source operating system Linux installed, it is possible to read and transfer data stored on obsolete data carriers (whether internally or externally via USB). Thanks to the growing popularity of retro video games, a host of tools and devices have been developed for reanimating the games stored on old floppy disks and vintage computers. For instance, the KryoFlux controller and software make it possible to connect disk drives via a USB port and transfer the data regardless of the computer's operating system.[3]

2. "Respect des fonds" (or the principle of provenance and sanctity of original order) has been a guiding principle in archival science since the end of the eighteenth century. It implies that an archival fonds should maintain the original arrangement as the creator had intended and that records (documents) from different creators should be kept separate from each other. See https://dictionary.archivists.org/entry/respect-des-fonds.html.

3. https://webstore.kryoflux.com/catalog/index.php?cPath=1

Sometimes, a laptop, personal computer or hard drive belonging to a writer is handed over to the Letterenhuis, as was the case with the laptop belonging to Kamiel Vanhole (1954–2008), or the somewhat older AMSTRAD PCW 8512 (Schneider Joyce computer) belonging to Walter van den Broeck (1941–), with integrated floppy disk drive and matrix printer. In such cases, advanced forensics tools and peripheral devices like write blockers[4] or KryoFlux are not essential in order to view or capture the data. The context in which texts and correspondence were created (software and hardware) are contained together with the content in a single complex object, the computer. Thanks to imaging tools like FTK Imager,[5] it is possible to create a single file that includes all files, system files, software and the operating system on the computer. In addition to the mythical value of these devices, comparable to the aura emanating from a fountain pen or typewriter that once belonged to a famous author, the research value thereof should not be underestimated. Folder structures, downloads, hidden system files, log files and email files provide an insight into the author's digital workspace (and online activities). For researchers, this is the ideal infrastructure for analysing the methodology of a writer. Since Matthew Kirschenbaum's ground-breaking work *Mechanisms: New Media and the Forensic Imagination* was published in 2008, digital workflows have been developed and experiments conducted with forensic tools, virtual machines and innovative techniques for identifying and analysing the digital traces left by creators. One of the most important initiatives is the development of the BitCurator Environment (2011–2014), an Ubuntu Linux package spearheaded by a number of major American universities (including the technology faculty of the University of Maryland, MITH).[6] The BitCurator is a software environment that includes a collection of open-source digital forensics tools and instruments for data analysis that help with the assessment and appraisal of digital archives. For many archival organisations, the arrival of the BitCurator marked an enormous step forward, as this package made it possible to bring together diverging scripts and tools from different fields of application in one environment. This includes, for example, the Sleuth Kit, a collection of command line tools for analysing images and recovering deleted files.[7] Working with the files at bit level, that is, not through the intermediary of operating systems, makes it possible to restore damaged disks. The ddrescue recovery tool can be used to restore damaged sectors on a data carrier, thus allowing some of the data to be recovered.[8] In addition, the BitCurator offers tools for detecting duplicated or similar files and even for scanning images, for example to search for identical words or names. It is also possible to run a privacy scan on images to identify any sensitive information, which is particularly useful when processing large volumes of data. Email files are especially large, and creators are often unaware of the presence of privacy-sensitive information about them-

4. A write blocker is "any tool that permits read-only access to data storage devices without compromising the integrity of the data" (https://www.cru-inc.com/data-protection-topics/write-blockers/).
5. https://www.exterro.com/ftk-imager
6. https://bitcurator.net/bitcurator/
7. https://www.sleuthkit.org/
8. https://www.gnu.org/software/ddrescue/

selves and others. Detecting this type of data is a complex and labour-intensive task. However, this work can now be automated by feeding algorithms with certain structures (such as mobile phone numbers, bank account numbers, and so on). This allows only the sensitive material to be protected without having to close off the entire archive. Finally, the BitCurator also helps archivists to tackle massive volumes of data using data mining and triage tools, which allows prioritising archival processing.

Technology allows the extraction of digital content with minimal intervention or manipulation. The archivist can analyse hidden, corrupt and deleted data, and at the same time detect creation, revision, conversion and transmission of texts or images and compare files qualitatively and quantitatively. Code and data *an sich* are not tangible, but they become tangible due to analysis of data carriers. Without interfering with the authenticity of digital archives, data can be displayed, opened and interpreted (see Kirschenbaum 2008 and Cleary 2019). The challenge of preserving and making born-digital archives discoverable has caused an immense dynamic in traditional archival practice. Archivists are learning to use code and are finding their way into open-source communities.

Despite the many opportunities for making digital archives accessible, there remains a tension between, on the one hand, preserving and documenting, and on the other hand, making accessible, visualising and researching born-digital material. This happens partly due to the learning curve in the humanities, but technical and ethical barriers also mean that accessibility and study of digital literary archives is not yet a reality.

To achieve for example the impressive emulation environment created by Emory University so that researchers can scroll and browse through Salman Rushdie's files on his old computers and laptops, structural dialogue between literature researchers, archivists and IT professionals is absolutely essential. Strategies or visions for displaying all this valuable data, for creating multiple views on source material for researchers and future users, for interfacing the data, are lacking at present among digital archivists (Drucker 2013).

The issue of digitally created material requires structural collaboration across reading room walls, both with researchers and IT professionals. Documenting and visualising traces of use on digital material and the interaction between data and users in an interface offer a treasure trove of new perspectives and "new" information. The time has come to "format" and to reshape the traditional alliances between heritage and research (Jaillant 2022).

References

Cleary, Thomas. 2019. "Visualizing Archives and Library Collections." *MAC Newsletter* 46 (3): 30–32.

Colavizza, Giovanna, Tobias Blanke, Charles Jeurgens, and Julia Noordegraaf. 2021. "Archives and AI: An Overview of Current Debates and Future Perspectives." *Journal on Computing and Cultural Heritage* 15 (6): 1–15.

Dean, Jackie, and Meg Tuomala. 2014. "Business as Usual: Integrating Born-Digital Materials into Regular Workflows Description." In *Innovative Practices for Archives and Special Collections*, ed. by Kate Theimer, 149–161. Lanham: Rowman and Littlefield.

Drucker, Johanna. 2013. "Performative materiality and theoretical approaches to Interface." *DHQ* 7 (1): http://www.digitalhumanities.org/dhq/vol/7/1/000143/000143.html

Jaillant, Lise. 2022. "Design Thinking, UX and Born-digital Archives: Solving the Problem of Dark Archives Closed to Users." In *Archives, Access and Artificial Intelligence. Working with Born-digital and Digitized Archival Collections*, ed. by Lise Jaillant, 83–130. Digital Humanities Research. Bielefeld: Bielefeld University Press.

Kirschenbaum, Matthew. 2008. *Mechanisms: New Media and the Forensic Imagination*. Cambridge, MA and London: MIT University Press.

Thibodeau, Kenneth. 2016. "Breaking Down the Invisible Wall to Enrich Archival Science and Practice." IEEE International Conference on Big Data, 3277–3282. https://ai-collaboratory.net/wp-content/uploads/2020/04/6.pdf.