Sytze Van Herck

# Historians as computer users

## Organizing sources with digital asset management for a history of computing

Mass digitization and born-digital sources have changed the work of historians, archivists, and museologists. Many historians spend less time in the archives, instead photographing or scanning sources on short research visits. In other cases archivists and museologists create online repositories to make sources more accessible. But what happens to all the photographs and scans researchers collect? This chapter studies how data management and curation influence historical research with textual, (audio)visual, and material sources regarding the history of the design and use of computing devices.

Placing computers into a broader context, my research focuses on the societal, business, and labor developments that are at the basis of computing devices. My PhD dissertation deals with the emergence of new occupations and changes in existing labor structures in relation to different computer models. Computers have also had an impact on both the workspace and the workflow of the user. Lastly, marketing has created an image of the idealized and archetypal user and has influenced gender stereotypes. To illustrate the evolution of the design and use of computer devices, my dissertation is based on objects and their representation in images, audiovisual sources, texts, and other sources such as drawings.

Since my primary sources were located across Europe and the United States, the time I could spend in museums and archives was limited. Rather than combining source collection and analysis during my research visits, I instead used these visits to briefly look at the objects on display, or sources in the archive catalog, to decide whether or not to digitize. Section 1 therefore looks into the different approaches for selecting sources before, during, and after museum and archive visits. Additionally, each section of this chapter discusses one or more of

**Acknowledgments:** I would like to thank Lori Emerson (University of Colorado) for her useful and constructive feedback. Part of this chapter is based on previous blogposts, mainly: Sytze Van Herck, "Annotating Sources: Digital Asset Management for Historians," Digital History & Hermeneutics blog, May 23, 2019, accessed June 10, 2020, https://dhh.uni.lu/2019/05/23/annotating-sources-digital-asset-management-for-historians/. Other parts of the text are based on the first draft of the methodology chapter of my dissertation.


https://doi.org/10.1515/9783110723991-011

the FAIR *Guiding Principles for scientific data management*.[1] FAIR stands for findable, accessible, interoperable, and reusable, and these principles can be used to evaluate existing research datasets such as museum and archive catalogs, or as selection criteria for research data management tools. Section 1 places a particular focus on the first FAIR principles of findable and accessible data.

After spending a day in the museum or archive, I usually amassed between 200 and 700 photographs, depending on the number of objects and archival sources I selected and photographed. Seven museum and seven archive visits have resulted in a total of over 18,000 photographs. In order to remember when, where, and what was digitized, the structure and organization of notes, photographs, and/or scans was essential. The metadata needed to be rigid enough to relocate a source, yet flexible enough to reorganize and recombine sources to facilitate analysis.

Section 2 discusses the digitization process and is followed in Section 3 by an analysis of (meta)data structure and documentation practice using the Tropy data asset management tool. Section 4 focuses on the influence of the interface and algorithms on the analysis stage of research, as well as on the final FAIR principles of data being interoperable and reusable. The chapter concludes with a reflection on the notion of the "original," the consequences of curating or selecting sources, the value of digital asset management, and the human factors that influence digital historical research.

## 1 Selecting sources

Since my research covers an array of machines, users, and applications, my dissertation is composed of several historical case studies. In the first stage I selected different types of computing devices – such as punch card equipment, a mainframe, a minicomputer, a microcomputer, and a personal computer – which differed in medium, size, price, and application. After exploratory visits to several computer history museums in the United Kingdom, Germany, and the United States, I limited the selection to specific computer manufacturers and models.

The first case study centers around Remington Rand and Powers-Samas equipment for punch cards used in accounting in the 1950s. The second case study shifts to the 1960s with IBM's System/360 mainframe, aimed at small and medium-sized companies, while the third study focuses on a computer model

**1** Mark D. Wilkinson et al., "The FAIR Guiding Principles for Scientific Data Management and Stewardship," *Scientific Data* 3, no. 1 (2016): 1–9.

from the 1970s: the PDP-11 minicomputer developed by the Digital Equipment Company (DEC) and used in the aerospace industry. The final two case studies compare a microcomputer from the United Kingdom, the BBC Micro, with the Apple *II*e personal computer from the United States, both of which were popular in schools in the 1980s.[2]

The framework of life cycle studies from material culture allows a comparison of different computing devices based on Henry Glassie's three main contexts of objects, namely "creation, communication and consumption."[3]These contexts translate to the following life cycle stages for computers: design and manufacturing (creation), sales and marketing (communication), and installation and use (consumption). The refined selection of case studies above formed the basis of my preparation for archive visits.

Reviewing the policies of the archive or museum in detail before a visit is paramount, in order to check whether scans or photographs are allowed, what has already been digitized, and how the collection is organized.[4] Limiting the research question to a small number of case studies helped me select only relevant objects and sources. Ideally, a museum or archive's online catalog allowed browsing, filtering, and search. Finding aids were especially helpful to get an idea of the material an archive offered. In large collections, even specific searches sometimes generated too many results, so filtering could refine results further.[5] Reviewing the temporary and permanent exhibitions in advance also made for a more efficient museum visit.

For both museum objects and archival records, most items are not on display but located in storage, either on-site or further away. Furthermore, archives have not usually cataloged their entire holdings and certain (sensitive) information might even be withheld from the publicly available online catalog. The internal catalog of the institution often contains more extensive metadata. Metadata is "data about (digital) data or any physical or conceptual object."[6] As Leonie Hannan and Sarah Longair remind researchers in their research guide, "the composition of collections is shaped by a variety of factors including the

**2** Sytze Van Herck, "Re/constructing Computing Experiences. From 'punch girls' in the 1940s to 'computer boys' in the 1980s." (PhD diss., University of Luxembourg), xxxi.

**3** Henry Glassie, *Material Culture* (Bloomington: Indiana University Press, 1999), 48.

**4** For further guidance on preparing research visits to museums see Leonie Hannan and Sarah Longair, *History through Material Culture*, IHR Research Guides (Manchester: Manchester University Press, 2017), 95–120.

**5** See chapter 6 of Eva Andersen in this volume.

**6** Julia Damerow and Dirk Wintergrün, "The Hitchhiker's Guide to Data in the History of Science," *Isis* 110, no. 3 (2019): 514.

purpose for its foundation and the original contents; decisions made by collectors, donors, and curators; the institution's changing priorities and locality; and its resources and size."[7]

As Gerben Zaagsma puts it, "digitisation is about selection" based on a variety of criteria which are rarely made explicit and certainly influenced by costs or funding.[8] At some of the museums, the enthusiasm of volunteers and staff – and their preference for certain computer models – together with the availability of sources that provided a glimpse of users from the past, certainly influenced my selection of case studies. Additionally, my time in archives across the world was limited as I had a generous but not unlimited amount of funding – and this too impacted my selection.[9]

Unlike library catalogs, which often use the same metadata format (MARC, or machine-readable cataloging) to develop online public access catalogs (OPACs) – and for which global catalogs such as Worldcat allow researchers to search beyond a single institution – museum and archive catalogs are rarely searchable through a single global catalog due to the diverse nature of items in the collections. Many initiatives such as the Conceptual Reference Model of the International Council for Documentation (CIDOC-CRM), which is a Linked Open Data ontology to describe metadata, in combination with overarching collections like Europeana, attempt to standardize metadata across cultural heritage institutions and thus create search portals beyond single institutions.[10] However, each institution I visited used a different portal. As Fotis Lazarinis concludes after listing the advantages of OPACs over card catalogs, "encoding data using the same format promotes interoperability among digital library tools."[11]

Some catalogs accommodated browsing, whereas the interface of others was only meant for search. Browsing allows researchers to discover sources they were not necessarily looking for and provides an overview of the institution's collections and structure. Searching, however, is only useful if you know what you are looking for. The Computer History Museum in Mountain View, California made archival finding aids available as PDFs describing each collection donated to the museum and listing nearly all individual items, thus allowing me to explore its

**7** Hannan and Longair, *History through Material Culture*, 98.

**8** Gerben Zaagsma, "On Digital History," *BMGN – Low Countries Historical Review* 128, no. 4 (2013): 20–1.

**9** Supported by the Luxembourg National Research Fund (FNR) (10929115).

**10** "What is the CIDOC CRM?," accessed October 1, 2020, http://www.cidoc-crm.org/; and "About Us," Europeana, accessed October 1, 2020, https://www.europeana.eu/en/about-us.

**11** Fotis Lazarinis, "Introduction to Cataloguing Classification," in *Cataloguing and Classification*, ed. Fotis Lazarinis (Boston: Chandos Publishing, 2015), 1–17.

content before my visit.[12] Another unexpected find occurred while I was browsing the Centre for Computing History's website, where I discovered a single box filled with documents from a former Helena Rubinstein employee which provided the foundation for an entire case study.[13]

Helena Rubinstein was one of three influential female business owners in the cosmetics industry and was described by Life Magazine in 1941 as belonging to the "matriarchy of the beauty industry."[14] The company's billing departments in the United States and United Kingdom used punch card equipment and Madame Helena Rubinstein helped to promote this technology in the 1940s and 1950s. Since both advertisements and company documents have survived, comparing the changing gender composition of the beauty industry and the punch card industry brought to light some interesting parallels.

Other catalogs, such as the OPAC catalog of the Living Computers Museum + Labs, were difficult to locate on the museums' website and had been designed for searching rather than browsing.[15] Browsing through finding aids respects the original order of a collection and is thus particularly useful in the exploratory phase of a study, whereas search functionality facilitates making connections across collections, which can aid not only the selection of sources but also the analysis of the material. In any case, "the ability to formulate meaningful queries and an awareness of how these queries might influence the search results and thus the analytical outcome is essential."[16]

Before each archive or museum visit I created a spreadsheet listing the collection, box, call number, date, and short title of each item of interest to my research. Depending on the institution's policy I either contacted the archivist beforehand or when I arrived. The policy of the Computer History Museum, for instance, only allowed researchers to consult a maximum of ten boxes per day, meaning I had to pare down my original selection. Choosing only those boxes with either very important or many items per box, I managed to shorten my list. Adding simple tick boxes to my spreadsheet also helped me keep track of the

**12** "Archival Finding Aids," accessed October 1, 2020, https://www.computerhistory.org/collections/findingaids/.

**13** "Powers-Samas & ICT Documentation," accessed October 1, 2020, http:// www.computinghistory.org.uk/det/28274/Powers-Samas-ICT-Documentation/.

**14** Elaine Brown Keiffer, "Madame Rubinstein. The Little Lady from Krakow Has Made a Fabulous Success of Selling Beauty," *Life* 11, no. 3 (1941): 40.

**15** The catalog has, at the time of writing, been suspended indefinitely due to the closure of the museum during the COVID-19 pandemic. "Living Computers: Museum+Labs is Closing for Now," accessed October 5, 2020, http://livingcomputers.org/Closure.aspx.

**16** Zaagsma, "On Digital History," 25.

items I selected and photographed, meaning I could just take very short notes for each item while I was on-site.

As for any historical research, the selection of sources depends on the research question and scope of the project. The sources available in turn influence the delimitation of the research question in terms of temporal and geographical boundaries and thus refine the topic. For instance, I originally intended to cover the entire twentieth century but noticed that the case studies differed substantially after the 1980s and required more context regarding networks and the advent of the World Wide Web. Furthermore, computing devices in the first half of the twentieth century revolved around improvements in punch card machines or single-purpose installations for defense or research projects, which have already been extensively researched and had fewer users. Therefore, the temporal delimitation of my research question was reduced from the entire twentieth century to the period between the 1940s and the 1980s.

## 2 Digitizing sources

Many historians rarely mention the technical setup and equipment they use during archive visits. In this section I intend to make my process more explicit, given my interest in the use of technology both as a subject of study and as a user myself. I digitized most sources on-site. Due to the large variety of material objects, textual, and visual sources, a camera was more appropriate than a portable scanner. Although our media center provides compact digital cameras, I preferred my own camera for its SR+ setting which "automatically optimizes settings to suit the scene" and its TEXT setting that ensures clear images of text or drawings in print.[17] The lighting in archive and museum settings and the limited time frame of research visits, as well as a lack of professional photography training, resulted in acceptable rather than professional digitization. Aside from a camera, I took along an SD memory card reader and an external hard drive to back up the photographs, a ReMarkable tablet to take notes, a laptop to go over my list of items and quickly look up additional information, and chargers for each device, just in case. In museums I also used my smartphone to record short videos of functioning computers.[18]

**17** "FujiFilm Digital Camera X20 Owner's Manual," 32, 34, 38, accessed September 28, 2020, https://dl.fujifilm-x.com/support/manual/x/fujifilm_x20_manual_en.pdf.
**18** An iPhone SE from 2016.

At the Centre for Computing History in Cambridge in the United Kingdom I established a useful workflow for documenting a single machine. First, an overview photograph showed the setup of a computer, then I photographed each component or peripheral such as the screen, keyboard, or mouse, followed by the object description. In some cases, exceptional inscriptions or signs of wear were also interesting, such as the names of schools printed on some of the BBC Micro computers. Perspective was also important, to convey the dimensions of the object.

Including the folder or item number where possible, at least for the first and/or last photograph of the same source helps to distinguish photographs of archival sources later on. In addition, respecting the order of a source by photographing one page, folder, or object after another saves time in comparing notes to photographs. At the end of each day, I copied all the images including metadata into the correct folder (one for each museum or archive) on my laptop and immediately created a backup on an external hard drive. At first I made the mistake of renaming files according to the model of the computer being photographed, but quickly realized that this disrupted the order of the photographs. Adding tags to files directly, without a clear ontology or structure in place to ensure consistency was disorganized – plus the document tags were difficult to export. Because of these difficulties I switched to dedicated software to organize the photographs.

Among some especially challenging sources to digitize were slides from the 1970s in the DEC collection of the Computer History Museum.[19] Using a light box was useful for capturing the image of the slides but obscured the handwritten text on the frame. In the worst instances this meant I had trouble identifying the subject, location, or date of the slide later on. Digitizing the museum objects was challenging for two reasons. Firstly, in crowded museums, visitors were sometimes accidentally included in the photographs and needed to be cropped out later. Secondly, images of reflective surfaces such glass display cases or computer screens could not be included if they showed visitors or the photographer/researcher. Finally, when the battery of the camera ran low or my muscles became sore and the images were not sharp, some text was no longer legible and some sources became unusable.

**19** Bo Doub, Kim Hayden, and Sara Chabino Lott, "Guide to the Digital Equipment Corporation Records," Computer History Museum, 2017, accessed November 12, 2020, https://www.computerhistory.org/collections/catalog/102733963.

## 3 Documenting collections

Choosing a tool or method always depends on what works for the project. Collaborative projects need software that allows each team member to interact with documents on a shared server or in a cloud infrastructure, whereas an individual researcher dealing with sensitive or copyrighted material certainly cannot share files and needs sufficient storage on a single device. For some researchers, combining images into a single PDF and adding each to a bibliography management tool such as Zotero is the easiest and most efficient solution. Others choose to build their own dataset or turn to other digital asset management software.[20]

When selecting a tool or method it is essential to first determine the requirements, keeping in mind the budget, digital literacy, and time investment. Since annotating images is just one step in the research project, it is important to consider whether the tool fits the research ecosystem. Criteria for evaluating the suitability of a tool include, for example, the workflow, compatibility with other research tools, or integration into a website.

Tropy is one of the first research photo management tools specifically designed for historians. Digital asset management software helps researchers to annotate, structure, and recombine not only text, or photographs, but also audio, video, and even 3D renditions of objects. For now, files are stored locally which means that neither synchronization between different machines nor collaboration with others is possible. Fortunately, an expansion of the range of media that Tropy can import, and the addition of remote access and cloud storage, are currently under development. The strength of digital asset management tools lies in navigating large photo collections with powerful searching, tagging, and annotation features. Although organizing sources, adding metadata, tagging, and transcribing is time consuming, data asset management tools definitely save time in the crucial final writing phase.

In the end, Tropy suited the project best, mostly because the photographs remain in their original folders, and only a small tile of the image is uploaded. Therefore the software does not duplicate the files, thus saving precious storage space. The tool was developed by the Roy Rosenzweig Center for History and New Media, which also developed Zotero. A collection in Tropy contains items that in turn consist of one or several images.

I chose to separate photographs into projects based on the subject of each of my dissertation chapters. For the first collection in which I tested the tool I

**20** See chapter 9 of Floor Koeleman in this volume.

browsed all images from the archives and museums I visited and wrote down the file names of the images from the original folder that were selected for the collection. Because the file path needs to remain the same to display the original image in the item view, I created a designated folder that I used only for copies of the selected photographs from the external solid-state drive (SSD) while still continuing the file structure according to the name of the institution.

I changed one crucial step in this selection process for all other collections after the first trial. Rather than repeatedly browsing all photographs whenever starting a new collection, it proved more efficient to browse everything only once and to take note of the file name, the designated collection, and a short title, including the date where possible. For this selection process to be ergonomic I used a larger monitor and on one side of the 24-inch display I browsed the images, while on the other side I checked archive notes via my MacBook using the monitor's Picture in Picture (PIP) option.

After this selection process, there were two potential approaches to uploading the images to Tropy. Photographs can be added all at once through the import feature or simply dragging and dropping, and separated or grouped into items later on. Alternatively, an item and its metadata can be created first, before any photographs are added to a single item. The first method is more efficient for a large number of items containing only a small number of images, such as letters consisting of one or two images, whereas the second method is useful for a small number of items with a large number of images attached to each item – in particular for user manuals consisting of up to 300 images for a single item.

I created metadata templates using existing ontologies. Metadata are data describing data or, in the case of digital sources, data generated by the camera such as date and time, as well as a file name, accompanied by data added by the researcher including the archive or museum, the identifier, the collection number or call number, a title, the folder and box or location, the language, the manufacturer or author, the copyright holder, etcetera. Each of the optional metadata fields for an item can be selected based on different ontologies. I created a separate metadata template for each institution based on the information available in the catalog and any additional research needs. Indicating the correct institution (i.e., the museum or archive) and the rights (e.g., educational non-commercial use only) determined by the institution's policy was partially automated.

When datasets adhere to the FAIR principles, they should also be interoperable. Tim Berners-Lee introduced the idea of a Semantic Web containing Linked Open Data in 2008 as "the idea of having data on the Web defined and linked in a way that it can be used by machines not just for display purposes, but also for

automation, integration, and reuse of data across various applications."[21] The Semantic Web is based on the Resource Description Framework (RDF) built up of triples that contain a subject (e.g., the image DSCF7180), a predicate or property (e.g., has the title) and an object or value (e.g., Apple *II*e).

Each element of a triple can be a Uniform Resource Identifier (URI) which, similarly to a file path, identifies a resource on a computer network. For instance, http://purl.org/dc/elements/1.1/title is the URI referring to the property called "title" according to the Dublin Core Metadata Initiative (DCMI).[22] Thus, by searching the DCMI's website for this URI, other researchers or algorithms can find the description of the title property. The most common form of URI is a URL, such as https://www.dublincore.org/specifications/dublin-core/dcmi-terms/#http://purl.org/dc/elements/1.1/title, which links to a web page that describes the aforementioned URI.

An RDF can be expressed in different languages: Tropy uses JSON-LD.[23] The URI is defined by an ontology which describes concepts that can be used to build semantic models adding meaning to data which can be reused by others. In other words, by adhering to established standards or URIs both future researchers and search algorithms can understand the meaning of metadata. In Tropy's metadata template builder I could select concepts from an ontology such as the DCMI metadata terms mentioned earlier, as well as RDF Schema.[24] In theory I could also incorporate a cultural heritage ontology such as CIDOC-CRM but the concepts were often too detailed or complex for my research dataset.[25]

Aside from the metadata based on the catalog and the content of the source, I added tags such as *image*, *mainframe*, or *manual* to allow filtering. Filtering in turn made it easier to relocate certain types of items or compare items with the same tag. The "notes" function was very useful for transcribing any text on the image but can also be used to formulate ideas or thoughts and will show up in search results. Besides transcribing sources, Tropy makes comparisons between

**21** Vipul Kashyap, Christoph Bussler, and Matthew Moran, *The Semantic Web: Semantics for Data and Services on the Web*, Data-Centric Systems and Applications (Berlin: Springer-Verlag, 2008).

**22** "DCMI Metadata Terms," Dublin Core, accessed October 5, 2020, https://www.dublincore.org/specifications/dublin-core/dcmi-terms/.

**23** Other examples are RDF/XML and Turtle.

**24** "RDF Schema 1.1," World Wide Web Consortium, accessed October 5, 2020, https://www.w3.org/TR/rdf-schema/.

**25** This paragraph is based on Sytze Van Herck, "DHOxSS 2017 – Linked Open Data, July 2017," Centre for Contemporary Digital History, accessed October 5, 2020, https://www.c2dh.uni.lu/thinkering/dhoxss-2017-linked-open-data.

sources from different institutions easier, since tags help users discover links between sources. Nevertheless, tags also further decontextualize sources from their original order in the archive and from the collection to which the original source belonged. Tagging is always a trade-off between few but broad terms, and many but narrow terms.

An advertisement for Remington Rand's SYNCHRO-MATIC bookkeeping machine featuring Helena Rubinstein as a person, a brand, and a user of the punch card machinery illustrates the documentation process described above.[26] First I noted the file name of my photographs of the first and last pages of the 12-page advertising leaflet. I selected the photographs in the relevant folder and dragged them into Tropy. After merging them into a single item, I applied the Computer History Museum Archive metadata template from the drop-down menu. The template mimics the online catalog and includes additional information. The first photograph included the catalog number, so finding the record online was straightforward.[27]

Since the title in the catalog referenced the contents of an entire folder rather than this particular advertisement, the Tropy item title was taken from the front cover of the leaflet. I determined the date based on an example of an invoice dated 1940 inside the advertisement. The metadata for the publisher, collection, URL, catalog number, dimension, and provenance came from the Computer History Museum catalog. The description was based on the content of the advertisement and the box number came from my research visit preparation notes which were based on the finding aid.[28] Tropy automatically inserted technical metadata and I added four tags to this particular item: *applications and use*, *equipment*, *marketing*, and *punch card*. After a comparison of all items tagged *marketing*, this advertisement became the focus of my Helena Rubinstein punch card case study.[29]

**26** Computer History Museum, "Madame Helena Rubinstein: Pioneer in Beauty, Research Scientist, Astute Businesswoman," *Remington Rand*, 1940, 102683284.

**27** "Remington Rand Promotional and Sales Materials," 1940, accessed November 12, 2020, https://www.computerhistory.org/collections/catalog/102683284.

**28** Dale Jenne, "Guide to the David C. Faloon papers," 2007, Computer History Museum, accessed November 12, 2020, https://www.computerhistory.org/collections/catalog/102634421.

**29** "Powers-Samas & ICT Documentation."

## 4 Analyzing sources through the interface

The key advantage of digital asset management during the analysis phase of research lies in the easy navigation between an overview of all sources in the project view that can be filtered through tags or sorted based on metadata, and the detailed item view where each image or page can be annotated separately using notes. When describing the workflow of punch operators and the wholesale department at Helena Rubinstein's company in London in the second half of the 1950s, for example, the software allowed me to quickly switch between a transcription of a policy document, an image of several invoices, the corresponding punch cards, an order form, and personal notes of the manager of the machine room. Another example of how the interface facilitates the analysis is through the zoom function in the item view. For a visual discourse analysis of the Apple *II*e advertisement, zooming into and panning over the image of the advertisement and the corresponding text allowed me to capture small details that would be overlooked without this function.

However, one issue I mentioned earlier remains unresolved. Greater flexibility in navigating the collection results in a loss of the original order of archival sources or a violation of the core principles of the *respect de l'ordre* and the *respect des fonds* in archival science and the historical method.[30] Gerben Zaagsma states that the key issue facing digital history is the loss of context both on "the level of collections, the use of digital archives or when dealing with information retrieval strategies," and in engaging with and experiencing historical materials.[31]

Although switching between items is easy, comparing images by placing them next to each other on the screen is not possible in Tropy itself. Within the program, users cannot open two images from the same item or two items next to each other. As a temporary solution the original files can be opened from the folder via the file path and placed next to each other outside of the program window. An interesting feature for in-depth discourse analysis is the possibility to select a specific part of the photograph which is added underneath and can be combined with additional metadata about the selection. This selection feature can be useful for separating images from text in the analysis.

Finally, the last FAIR principle – stating that data should be reusable – is certainly supported by Tropy since the metadata adheres to existing ontology. Unfortunately, sharing photographs and metadata is limited because of archival

**30** Andreas Fickers, "Update für die Hermeneutik. Geschichtswissenschaft auf dem Weg zur Digitalen Forensik?," *Zeithistorische Forschungen*, author preprint (2020): 10.
**31** Zaagsma, "On Digital History," 27.

and museum policies, and heavily impeded by copyright laws and the European Union's General Data Protection Regulation (GDPR). Or, in the words of Julia Damerow and Dirk Wintergrün, "copyright practices often prohibit sharing acquired data, which significantly hampers attempts to reproduce or build on the results of a project."[32] These legal restrictions had never hindered my previous research as a medievalist, but they prevent me from sharing sources and metadata concerning contemporary history. Due to these restrictions the dataset for my project will remain internal to the institution and reusable only when permission is granted by the archive or museum and the original copyright holder.

## 5 Dissemination

A final discrepancy to acknowledge is that the published images will differ substantially from the photographs taken at the archive. The preparation of photographs in photo editing software depends on the medium of publication. For online publications the resolution of an image file is usually set to 72 dpi or the standard monitor resolution, whereas publishers of printed publications generally require a higher resolution of 300 dpi. As DiMarco explains, print requires high resolution "because it is a high-fidelity media looked at closely by the viewer," whereas "Web images are viewed on a screen, and at a distance."[33] Aside from the resolution, photo editing software can improve the lighting, straighten crooked photographs, and cut out the part of the image that is relevant. After editing an image I usually indicate the archive, call number, original image file name, and whether the reworked version is an edit or cutout of the original.

As Andreas Fickers mentions in "Update für die Hermeneutik," the notion of the original becomes problematic in any case because data is, as Lisa Gitelman and Virginia Jackson pointed out, never raw but rather the result of a process, as I have illustrated here.[34] The provenance of a digital source is influenced by the format, display, storage, and compatibility of a file. In other words, both the software and the hardware used to store, open, and display the file, influence the representation of a digital or digitized source.[35]

**32** Damerow and Wintergrün, "The Hitchhiker's Guide," 520.

**33** John DiMarco, *Digital Design for Print and Web: An Introduction to Theory, Principles, and Techniques*, (Hoboken, NJ: Wiley, 2010), 135.

**34** Lisa Gitelman and Virginia Jackson, "Introduction," *"Raw Data" Is an Oxymoron*, ed. Lisa Gitelman, Infrastructures series (Cambridge, MA: MIT Press, 2013), 2.

**35** Andreas Fickers, "Update für die Hermeneutik," 7.

As Matthew Kirschenbaum explains "One can, in a very literal sense, *never* access the 'same' electronic file twice, since each and every access constitutes a distinct instance of the file that will be addressed and stored in a unique location in computer memory. [. . .] Access is thus duplication, duplication is *preservation*, and preservation is creation – and recreation."[36] To further complicate matters, Jacob Gaboury explains how "the computer is not a visual medium," but rather "our engagement with computing technology is increasingly mediated through the interface of the screen."[37] Although we perceive an image on the screen as a representation of an original source, the file is saved on an SSD in the form of zeros and ones, bits and bytes, or colored pixels.

To illustrate the discrepancy between what is stored and what is shown, without wrecking a hard drive, I opened a PNG file with the text editor. The first few letters made some sense (*àPNG*), but the rest of the file can only be described as a gibberish of letters, numbers, and symbols. What the PNG file contains when opened with image software is an edited cutout of a Powers-Samas tabulator which processed punch cards to produce reports and was taken from a user manual.[38]

To avoid a mismatch between the gray background of the image and the white pages of my printed dissertation, I deleted the background of the image using the magic eraser feature, in combination with the lasso and pencil cutout features of PhotoScape X. I also adjusted the perspective to correct the strange angle of a picture taken from a book. Finally, the file name of the edited image (CCHa_CH28274_DSCF3374_edit-cutout.png) references the archive where the image was made (Centre for Computing History archive), the call number of the source (CH28274), the file name of the original as assigned by the camera (DSCF3374), and the transformations made in PhotoScape X (edit and cut out). The name thus ensures that, even if the document is taken out of context when uploaded to Overleaf for insertion into my dissertation, I can easily relocate the source and remember which changes were made in the photo editing software.

**36** Matthew Kirschenbaum, "The .txtual Condition: Digital Humanities, Born-Digital Archives, and the Future Literary," *Digital Humanities Quarterly* 7, no. 1 (2013).

**37** Jacob Gaboury, "Hidden Surface Problems: On the Digital Image as Material Object," *Journal of Visual Culture* 14, no. 1 (2015): 540.

**38** Centre for Computing History, "The Tabulator for 36 and 40-Column Cards. Operator's Manual," accessed November 12, 2020, http://www.computinghistory.org.uk/det/28274/Powers-Samas-ICT-Documentation/.

## 6 Conclusion

Historians increasingly rely on digitized and born-digital sources. But digitizing collections is often an expensive undertaking and nearly always a political choice which can bias historians' gravitation toward certain sources over others that are not digitized and yet form a much larger part of the overall collection. The problems of access to sources and a bias toward particular topics at the expense of others also depend on which sources survive both in analog and digital form, or how much of a museum's collection is cataloged, accessible, and can be searched, versus what remains hidden. However, for geographically dispersed archival research, "quick and dirty" digitization of sources through photography has an impact on the kinds of questions historians can ask. I argue that although digital history is primarily concerned with digital tools and methods for analysis, data entry – much like archival research – requires careful reflection and should not be taken for granted.

Aside from selecting relevant sources and accompanying metadata for the researcher's collection, categorization, and the use of established existing ontologies ensure consistency, and assigning metadata in bulk can improve the accuracy of the data. As the FAIR Guiding Principles state, "beyond proper collection, annotation, and archival, data stewardship includes the notion of 'long-term care' of valuable digital assets, with the goal that they should be discovered and reused for downstream investigations, either alone, or in combination with newly generated data."[39]

The notion of the *digital dark age* in which potentially valuable digital sources rapidly disappear into a "digital black hole" reveals several tensions that worry historians and archivists. In a constantly expanding mass of digital records, and despite the common notion that what appears online "sticks," data and files are easily deleted, or can become damaged, or incompatible with ever-changing hardware and software. What is often overlooked is the fact that behind most of the data there are people, organizations, and events that impact the sustainability of datasets. This holds true for museums and archives as well as their dynamic catalogs, and for the data management of individual researchers' projects and their affiliations with an institution.

In other words, institutions are never permanent and both collections and institutions are subject to change. As soon as stakeholders disappear, so too might data after the termination of a project. As researchers or employees of institutions move on and lose interest, a lack of transparency can result in the

**39** Wilkinson et al., "The FAIR Guiding Principles," 1.

loss of the most valuable asset of research, its sources and data. The loss of data becomes even more problematic in research funded by public institutions.

## References

Brown Keiffer, Elaine. "Madame Rubinstein. The Little Lady from Krakow Has Made a Fabulous Success of Selling Beauty." *Life* 11, 3 (1941): 37–45.

Centre for Computing History. "Powers-Samas & ICT Documentation." Accessed October 1, 2020. http://www.computinghistory.org.uk/det/28274/Powers-Samas-ICT-Documentation/.

CIDOC-CRM. "What is the CIDOC CRM?" Accessed October 1, 2020. http://www.cidoc-crm.org/.

Computer History Museum. "Archival Finding Aids." Accessed October 1, 2020. https://www.computerhistory.org/collections/findingaids/.

Computer History Museum. "Guide to the David C. Faloon papers." Dale Jenne, 2007. Accessed November 12, 2020. https://www.computerhistory.org/collections/catalog/102634421.

Computer History Museum. "Remington Rand Promotional and Sales Materials," 1940. Accessed November 12, 2020. https://www.computerhistory.org/collections/catalog/102683284.

Computer History Museum. "The Tabulator for 36 and 40-column cards. Operator's Manual." Accessed November 12, 2020. http://www.computinghistory.org.uk/det/28274/Powers-Samas-ICT-Documentation/.

Computer History Museum. "Madame Helena Rubinstein: Pioneer in Beauty, Research Scientist, Astute Businesswoman." Remington Rand, 1940, 102683284.

Damerow, Julia, and Dirk Wintergrün. "The Hitchhiker's Guide to Data in the History of Science." *Isis* 110, 3 (2019): 513–21. https://doi.org/10.1086/705497.

Dublin Core™ Metadata Initiative. "DCMI Metadata Terms." Accessed October 5, 2020. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/.

DiMarco, John. *Digital Design for Print and Web: An Introduction to Theory, Principles, and Techniques*. Hoboken, NJ: Wiley, 2010.

Doub, Bo, Kim Hayden, and Sara Chabino Lott. Computer History Museum. "Guide to the Digital Equipment Corporation records." 2017. Accessed November 12, 2020. https://www.computerhistory.org/collections/catalog/102733963.

Europeana. "About Us." Accessed October 1, 2020. https://www.europeana.eu/en/about-us.

Fickers, Andreas. " Update für die Hermeneutik. Geschichtswissenschaft auf dem Weg zur Digitalen Forensik?" *Zeithistorische Forschungen*, Author preprint (2020): 12.

Fujifilm. "FujiFilm Digital Camera X20 Owner's Manual." Accessed September 28, 2020. https://dl.fujifilm-x.com/support/manual/x/fujifilm_x20_manual_en.pdf.

Gaboury, Jacob. "Hidden Surface Problems: On the Digital Image as Material Object." *Journal of Visual Culture* 14, 1 (2015): 40–60. https://doi.org/10.1177/1470412914562270.

Gitelman, Lisa, and Virginia Jackson. "Introduction." In *"Raw data" is an Oxymoron*, edited by Lisa Gitelman, Infrastructures series, 2. Cambridge, MA: MIT Press, 2013.

Glassie, Henry. *Material Culture*. Bloomington: Indiana University Press, 1999.

Hannan, Leonie, and Sarah Longair. *History through Material Culture*. IHR Research Guides. Manchester: Manchester University Press, 2017. Accessed April 24, 2020. https://manchesteruniversitypress.co.uk/9781784991265.

Kashyap, Vipul, Christoph Bussler, and Matthew Moran. *The Semantic Web: Semantics for Data and Services on the Web*. Data-Centric Systems and Applications. Berlin: Springer-Verlag, 2008. https://doi.org/10.1007/978-3-540-76452-6.

Kirschenbaum, Matthew. "The .txtual Condition: Digital Humanities, Born-Digital Archives, and the Future Literary." *Digital Humanities Quarterly* 7, 1 (2013). Accessed June 18, 2021. http://www.digitalhumanities.org/dhq/vol/7/1/000151/000151.html

Lazarinis, Fotis. "Introduction to cataloguing classification." In *Cataloguing and Classification*, edited by Fotis Lazarinis, 1–17. Boston: Chandos Publishing, 2015. https://doi.org/10.1016/B978-0-08-100161-5.00001-4.

Living Computers Museum+Labs. "Living Computers: Museum + Labs is Closing for Now." Accessed October 5, 2020. http://livingcomputers.org/Closure.aspx.

Van Herck, Sytze. "Annotating Sources: Digital Asset Management for Historians," 2019. Accessed June 10, 2020. https://dhh.uni.lu/2019/05/23/annotating-sources-digital-asset-management-for-historians/.

Van Herck, Sytze. "DHOxSS 2017 – Linked Open Data," 2017. Accessed October 5, 2020. https://www.c2dh.uni.lu/thinkering/dhoxss-2017-linked-open-data.

Van Herck, Sytze. "Re/constructing Computing Experiences. From 'punch girls' in the 1940s to 'computer boys' in the 1980s." PhD diss., University of Luxembourg.

Wilkinson, Mark D. et al., "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3, 1 (2016): 1–9. https://doi.org/10.1038/sdata.2016.18.

World Wide Web Consortium. "RDF Schema 1.1." Accessed October 5, 2020. https://www.w3.org/TR/rdf-schema/.

Zaagsma, Gerben. " On Digital History." *BMGN – Low Countries Historical Review* 128, 4 (2013): 3–29. https://doi.org/10.18352/bmgn-lchr.9344.