Cite as Trevor Owens, *Digital Sources & Digital Archives: The Evidentiary Basis of Digital History* in *Companion to Digital History*. Edited by David Staley.

Digital Sources & Digital Archives: The Evidentiary Basis of Digital History

In an early draft of my undergraduate thesis I wrote that a source “spoke for itself.” My advisor crossed that out and wrote in the margin something like “sources almost never speak for themselves, you have to explicate what the source means for your argument and justify your interpretation.” I imagine this sort of experience is how many individuals learn the ropes of historical research and writing. The task of the historian is to interpret sources.

The world is full of objects, archives, records and texts which historians can study and interrogate to develop and refine our understanding of the past. These are the primary sources of history; materials, relics, and texts, that testify and provide traces of the past. Almost anything could be a primary source. The rings of a tree testify to weather conditions and changes in climate. Probate records document the material goods individuals held at the end of their lives. Court proceedings offer insight into the experiences of the oppressed through the moments they are dragged in front of the justice systems that control and marginalize them¹. Just as any kind of physical object might serve as a source, as society increasingly produces digital relics, documents, artifacts and other objects the evidentiary basis of history will become increasingly digital.

While things like the rings of a tree have their own value as historical sources, the bulk of historical work continues to be anchored in archives. Historians’ ability to study the past is largely directly indebted to archivists and the range of individuals involved in the production and management of historical records. Archives come in all shapes and sizes; massive federal agencies, small local historical societies, manuscript collections at research libraries to name a few examples. The same digital shift occurring in sources is occurring in archives.

At this point, historians have access to an ever-expanding wealth of digitized versions, or digital surrogates, of a selection of primary sources through online collections. At the same time, an explosion of born-digital materials is being produced and collected at unprecedented scale (websites, the contents of a hard drives, collections of emails, digital video and photos, etc.). While these new forms of sources are emerging so too are notions of digital archives. Organizations like the Internet Archive, and projects like the September 11th Digital Archive, and the Rossetti Digital Archive have emerged with the archive name attached. However, each of these varieties of digital archives represents a somewhat different vision of the nature of the concept of an archive.

So, what happens to history when the basis of its sources and evidence becomes increasingly digital? Similarly, what happens to history when it’s archives become digital? Backing up a bit, given how the very form of archives as institution is anchored in the management of paper documents, what does it even mean to have a “digital

¹ For examples of tree rings, see. William Cronon, *Changes in the Land: Indians, Colonists, and the Ecology of New England* (New York: Hill and Wang, 1983). For examples of the perpetual value of probate records see Bushman, Richard L. *The Refinement of America: Persons, Houses, Cities*. New York: Knopf, 1992. For examples of using court proceedings see Pagan, John Ruston. *Anne Orthwood’s Bastard: Sex and Law in Early Virginia*. New York: Oxford University Press, 2003.

archive”? What follows is an attempt to identify and discuss issues in the evidentiary basis of history that arise as the materials and systems that manage those materials become digital. In looking at different kinds of sources and archives I work to suggest practical advice on the kinds of issues and questions one should ask when working to interpret, to find out what one can say, based on digital sources and digital archives.

What are Digital Sources?

When you hold a letter in your hand and read the words on it you can imagine what it was like when the recipient of that letter held it in their hands in the past. As an interpreter of the record, you can think about what it must have been like to receive it and follow a chain of correspondence to understand the exchange of thoughts and ideas. How does this interaction change when you have a digitized copy of a letter? Similarly, how does it change when you are looking at the text of an e-mail message?

Making sense of a source and making a defensible inference based on the content of a source requires context. That is, knowing a letter was sent from one individual to another and that you found it in the papers of the recipient you can likely infer that it represents a perspective that the author wanted to communicate to the recipient and you likely have reason to assume that the recipient read it. In contrast, if a historian of the future had access to an archived copy of my Gmail account they would need to know a bit about many of the automated rules I’ve set up that “mark as read” emails from a range of individuals and organizations and in some cases are set to “skip my inbox” entirely. So without knowing about those rules one could end up making all kinds of problematic inferences about what I had or had not read based on what was in my email. Understanding my email thus requires an understanding of how people like me used email at a particular period in time and the set of features and functionality that different email clients came with.

As Martha Howell and Walter Prevenier explain in their introduction to historical analysis of sources “to make wise choices among potential sources, historians must thus consider the ways a given source was created, why and how it was preserved, and why it has been stored in an archive, museum, library or any such research site.”² The same kinds of questions need to be asked of digital sources. This is particularly challenging given that pace of change in the mediums and context of communications technologies seems to continue to accelerate. Historians need to develop an understanding of digital source criticism and provenance.

Digital Source Criticism & Provenance

Given the range of digital sources and the complexities of their production and use the future of historiography will require a good bit of work in digital source criticism.³ German historian Johann Gustav Droysen’s 1867 book *Outline of the Principles of History* explains the concept and importance of source criticism as a part of

² Howell, Martha C., and Walter Prevenier. *From Reliable Sources: An Introduction to Historical Methods*. Ithaca, N.Y: Cornell University Press, 2001, p 28.

³ For further discussion of digital source criticism see Hering, Katharina. “Provenance Meets Source Criticism.” *Journal of Digital Humanities*, August 4, 2014. <http://journalofdigitalhumanities.org/3-2/provenance-meets-source-criticism/>.

historical practice. “The task of Criticism is to determine what relation the material still before us bears to the acts of will whereof it testifies. The forms of the criticism are determined by the relation which the material to be investigated bears to those acts of which will gave it shape.⁴” That is, a key part of historical research and writing involves not simply identifying sources of history but working to understand the context in which they were produced.

To this end, working with digital sources prompts the historian to ask the same kinds of questions they have long asked of sources. What is a source’s provenance? How was it created and stored? Why does it persist today? These kinds of questions are essential for interpreting a source. This is not simply an issue for those studying society after the advent of computing technology. There are a range of key source criticism questions to should ask of both digitized primary sources and born digital sources. What follows is an exploration of some of the key issues for consideration related to both of these kinds of sources.

Digitized Primary Sources

For anyone studying the world before the emergence of digital media the primary role that digital media will play is as a transmitter of digital surrogates. Libraries, archives and museums have now been actively digitizing sources for thirty years and the result is that one can find millions of digital surrogates of books, maps, photographs and manuscripts in a range of online digital collections. In working from these sources there are a few critical questions to ask from the perspective of provenance and source criticism.

Why was this digitized and not something else?

It has always been important for historians to ask why a particular source has been preserved. It is critical to think through why we have access to some kinds of sources and not others and this is a key part of that reasoning exercise. The same kind of selection questions needs to be asked of any digitized source.

In some cases, archives have digitized full runs of materials; in other cases they have digitized highlights or selections. Generally, libraries, archives and museums have only digitized a sliver of their entire holdings. It is not enough to find a source, one must be able to contextualize it and understand why they have it at hand and as such it is important to think through the kinds of limitations on inferences one can make from something based on what you know about the digitization policies of a given organization.

For example, because of copyright restrictions many institutions in the United States are focusing efforts on digitizing materials from before 1923. Or similarly, an archive might have the rights cleared to digitize one particular collection, or the writings of one person instead of another. In each of cases if one wants to work primarily from digitized materials it is critical to think through how the selection policies for what was digitized can shape and limit one’s ability to make inferences based on those materials.

⁴ Droysen, Johann Gustav Bernhard. *Outline of the Principles of History: (Grundriss Der Historik)*. Translated by Elisha Benjamin Andrews. Boston: Ginn & company, 1897. <https://archive.org/stream/outlineofprincip00droy#page/22/mode/2up>.

Is this copy of significant quality for my purpose?

All digitized objects are surrogates for the originals. That's fine. Historians have a long tradition of working from surrogates. In many cases, the only access historians have to extant historical materials is through copies of reprintings, and copies of copies created through the manuscript tradition. Similarly, when microfilm technology developed in the 1930s historians were thrilled with the prospect of reproductions of sources. Public historian Ian Tyrrell used the same rhetoric often used regarding digitization and the web to describe microfilm in the 30s. In his words, microfilm "democratized access to primary sources by the 1960s and so put a premium on original research and monographic approaches."⁵ The reproduction of sources played a key part in historians' increased focus on working from primary sources. In this vein, it's worth remembering that the development of the technologies that provide access to sources will continue to play a role in shaping the norms and expectations of the composition of history. So, surrogates are nothing new, in many ways they are the norm for many areas of historical practice. With that said, it's always critical to ask if the surrogate is good enough for the questions a historian is asking.

Historians often want to do straightforward things with a source. So if one wants to be able to say an individual wrote a particular thing in a particular document then as long as you can make out the words in a digitized copy of something that is likely enough. In this case, it is worth differentiating the informational qualities of a source from its artifactual qualities⁶. The informational qualities of a source are generally the words inscribed on it. The artifactual qualities of a source can consist of any number of different features one might study. As historians have become increasingly interested in sources as part of material culture the need to consider artifactual qualities has become increasingly important. Every physical object contains a nearly infinite amount of information in its artifactual qualities. For example, beyond the legibility of words on an object, characteristics of handwriting, fingerprints, watermarks, the chemical composition of inks or of paper or vellum can all be interrogated to provide valuable information. All of that information is anchored in the artifactual qualities of the source.

As an example, you can find some rather ugly looking, but for the most part legible, copies of Hamlet in Early English Books Online. They are black and white images created from scans of old microfilm. You can also find much nicer looking copies of the same work in the Folger Shakespeare Library's online collections. If what you care about is the text of the work, you are mostly fine in either case. With that said, researchers have used high quality full color scans, like those Folger provides, to study the placement of dirt on the margins of the page. The dirt on the pages, which comes from people handling the books, attests to the use of the books over time. That is, there are material traces of use of the books left on them that can be studied. Most interestingly, it can actually *only* be study when high quality scans of the book are created. That is, aspects of the source only become available for analysis through the

⁵ Tyrrell, Ian R. *Historians in Public: The Practice of American History, 1890-1970*. Chicago: University of Chicago Press, 2005, p. 38.

⁶ For further exploration of discussion of informational versus artifactual qualities of digitized sources see Fleischhauer, Carl. "Information or Artifact: Digitizing a Book, Part 1 | The Signal: Digital Preservation." Webpage, October 17, 2011. <http://blogs.loc.gov/digitalpreservation/2011/10/information-or-artifact-digitizing-a-book-part-1/>

production of a very high quality digital surrogate. To that end, the better quality the scans the more potential there is to examine traces of other physical properties of a source⁷. The question for someone working from a digitized surrogate of a source is thus are the significant properties of the source necessary for the sorts of questions you are interested in asking present? Similarly, it is important to consider how some aspect of the quality of a source might be obfuscated in how it was digitized or provided.

How did I find it and how does that effect what I can say about it?

At this point one can visit the Library of Congress, the Digital Public Library of America, Europeana or Google Books on the web and plug in some obscure search terms and find digital surrogates of records, artifacts and a variety of other primary sources. This is amazing. You can find things that you would never have been able to find before⁸. Searching across millions of sources at once is transforming many historians' methods for research and scholarship⁹. At the same time, full text search presents a whole new set of challenges for reasoning from and interpreting sources.

Where in the past one would develop an explicit strategy to explore a given collection or archive, or to systematically look at all the newspapers from a given date range, search encourages researchers to stumble around and find something that looks interesting. This is all fine if all one wants to do is make an existence proof argument. That is, if one just wants to make the case that something was said at a particular point in time. However, this is a rather low bar for historical argumentation. The extent to which something is representative of a particular moment in time, or a particular community or place is tied explicitly to a range of contextual questions.

To be able to make broader claims based on a given source it is important to work to contextualize it after it is discovered through search. Feel free to search for idiosyncratic terms, to as Stephen Ramsey suggests, “screw around” in searching through digitized sources. However, it then becomes necessary to do the legwork required to understand the original context from which that source emerged and think through the limitations that come from why that source was digitized and not something else. To do this, it is necessary to work backward from a digitized source to understand where it came from and the extent to which it is or isn't representative of the collection it comes from.

Born Digital Sources

⁷ For a more extensive exploration of this example, see Sarah Werner Where Material Book Culture Meets Digital Humanities, from the Journal of the Digital Humanities, Vol. 1, No. 3 Summer 2012

⁸ For an excellent example of the way that searches for obscure terms have made it possible for historians to discover things that would have been nearly impossible in the past see Leary, Patrick. “Googeling the Victorians.” *Journal of Victorian Culture* 10, no. 1 (Spring 2005): 72–86.

⁹ For an exploration of how searching through millions of books is changing research processes in the humanities see Ramsay, Stephen. “The Hermeneutics of Screwing Around; or What You Do with a Million Books.” In *Pastplay: Teaching and Learning History with Technology*, edited by Kevin Kee. University of Michigan Press, 2014. <http://hdl.handle.net/2027/spo.12544152.0001.001>. For further exploration on the way that searching through massive amounts of sources suggests the need for changes in how historical writing is framed see Gibbs, Fred, and Trevor Owens. “The Hermeneutics of Data and Historical Writing.” In *Writing History in the Digital Age*, edited by Kristen Nawrotzki and Jack Dougherty. University of Michigan Press, 2013. <http://hdl.handle.net/2027/spo.12230987.0001.001>.

Born digital is the rather clumsy term we have to discuss sources that started off digital; email messages, digital photographs, websites, databases, etc. Going forward, the bulk of the primary sources historians will work with to understand the world in the 21st century are going to be things that started off digital. This is not to suggest that we will ever get away from paper sources, but it is to note that much of that paper source material will have started out as digital as well. In those cases, the paper will often be a surrogate for the digital. While archivists and historians are still only just figuring out how to collect, preserve and provide access to born digital primary sources there are already a set of emerging key questions to ask of such sources. What follows is an initial exploration of some key source criticism questions to ask of born digital sources.

What are you not seeing on the screen?

When working with digital objects it's essential to remember that what they look like on the screen is a performance¹⁰. The actual digital object is a sequence of markings registered on a medium. Hard drives, CDs, flash drives, etc. are all things that register sequences of markings (bits) that are read by software to show up on a computer screen. In any digital file and any digital file system there is additional encoded information that one could be looking at and reading.

In contrast to looking at a hand written letter, where you can see how hard someone pressed and get a feel for their handwriting, when one looks at an email message on a screen all you see is the words. However, if you poke around in the email headers, or in the metadata associated with a message you can find a wealth of information that isn't rendered on the screen. New media scholar Nick Montfort has deemed the focus on what things look like on the screen "screen essentialism" and a growing body of work is emerging to provide basic tools and approaches for getting beyond simply taking things as they appear¹¹. Two examples of working with particular primary sources will help underscore what historians have to gain by getting beyond screen essentialism.

When curator Doug Reside first opened a file he found on a floppy disk in playwright Jonathan Larson's papers at the Library of Congress he must have been shocked. Right there on the screen was a different set of text for a famous song from one of the musicals Larson had created¹². What was it that he was looking at? Was this an alternative version of the song? As Reside dug deeper, and came to understand the nature of the word processing software that Larson had used and the software that Reside was using to render the text with he came to understand exactly what had happened. The word processing software that Larson had used would save a record of changes in the text inside the file. So an individual word-processing file would actually contain a record of the edits to a file over time.

¹⁰ For further exploration on the theme of digital objects as performance in the context of a digital art manuscript collection see Arcangel, Cory. "The Warhol Files: Andy Warhol's Long-Lost Computer Graphics." *Artforum*, Summer (2014). <https://artforum.com/inprint/issue=201406&id=46874>.

¹¹ For more on screen essentialism see Montfort, Nick. "Continuous Paper: The Early Materiality and Workings of Electronic Literature." Philadelphia, 2004. http://nickm.com/writing/essays/continuous_paper_mla.html

¹² For further detail on Reside's work with these files see Reside, Doug. "'No Day But Today': A Look at Jonathan Larson's Word Files," April 22, 2011. <http://www.nypl.org/blog/2011/04/22/no-day-today-look-jonathan-larsons-word-files>.

The only way Reside could interpret what he saw on the screen was to learn a bit more about the software that was used to write it and the software he was using to render it. Ultimately, this is a rather fascinating result; works written in this particular word-processing application have within them records of their creation and editing.

The implications of this kind of work extend beyond the structure of individual files. In working to understand the material properties of digital objects, digital humanities scholar Matthew Kirschenbaum opened up a ROM (a copy of a floppy disk) in a Hex editor¹³. This ROM had a copy of an early video game called Mystery House. A Hex editor renders the hexadecimal notation, a recording of each byte on the medium. So the Hex editor showed how the information in the ROM was laid out on the original floppy disk it was saved on. As he explored the disk he found something intriguing, a sequence of text that did not appear in the game he was studying. What had he found? Was this hidden text in the game that wasn't used? After goggling the text he was able to identify that the text came from a completely different game. From this, he was able to infer that the disk the ROM had been created from had a copy of the other game that had been overwritten by the second game. Kirschenbaum downloaded a copy of a game and was able to figure out what had been on the original disk before the game was saved on it.

Understanding how this happened requires background on how floppy disks and hard drives function. When a file is deleted it generally really isn't deleted. Instead, a computer marks the space that the file is stored as available to be overwritten. The result is that if you poke around in what is actually written on a computer disk you will find that all sorts of areas on it that the operating system will tell you are empty spaces that actually contain readable information. As a result, as archives increasingly begin accessioning this kind of born digital material they are making decisions on if they want to create forensic copies of this kind of media (that is copies that will contain all that information, including information that is hidden to the user) or if they want to create logical copies of disks and drives that will only contain what the operating system thinks is there. In either event, this suggests a whole new set of skills for interpreting primary sources that historians are going to need to become adept with. When working with born digital sources it is important to understand them beyond what they look like on the screen. It is critical to move past the performance of a file or a file system and to understand the additional information that may not be immediately revealed. The performance of digital content similarly opens a set of questions about the set of technologies used to interpret it.

What is lost in how it was/is rendered?

When files are rendered on a computer screen a user witnesses something akin to the performance of a play. The underlying data in a file is interpreted and rendered through software for a user to interact with in much the same way that the script of a play is interpreted and performed by a cast on a stage. In each case, while the underlying script or files remains the same, a given performance of a file or a play is going to look and

¹³ Kirschenbaum, Matthew G. *Mechanisms: New Media and the Forensic Imagination*. Cambridge, Mass: MIT Press, 2008, pp. 111-159.

sound different. For some kinds of research questions those differences do not matter, however, it is necessary in either case to be aware of the differences.

Archived websites offer a key case to explore how this plays out in the interpretation of a born digital primary source. At this point, many organizations are using a range of different tools to archive websites. They use a few different kinds of tools to harvest copies of what content was available at a particular URL at a given moment and then use another set of tools to be able to render that content for you to view. For example, you can go to the Internet Archive and type in the URL for www.loc.gov and you will find an interface that lets you see what the homepage of the Library of Congress website looked like at different points in time when the Internet Archive saved a copy of it. With that said, it is important to realize that when you look at a copy of the site in the Internet Archive's Wayback Machine you are not really seeing what the site looked like at that point in time because a range of characteristics of the way the site looked then are not being replicated.

One views a website through a web browser, and any given browser will render things slightly different. This is particularly true for older sites. Similarly, when one looks at a website from ten or twenty years ago those sites were designed for computers that had smaller screen resolutions, that had different processors, that ran different operating systems. Each intermediary layer of software (the browser, the operating system etc.) and the implied assumptions about computer hardware baked into that software (screen resolution, processor speed, etc.) function as part of the sequence of interpreters that perform a webpage.

When asking questions about what is lost in how a digital object or set of digital objects is rendered it is important to recognize that different elements are more likely susceptible to issues. The distinctions between the informational and artifactual elements of sources previously discussed are similarly relevant in this context. For example, if all one is focused on is how something was written in text on a page, in most cases how it is rendered isn't likely to be too much of a problem. However, in cases like the presentation of digital art created for the web or in situations where the aesthetics, design and user experience of a web page matter it is very likely that issues in how something is rendered will play a significant role in one's ability to interpret it¹⁴.

How was this created, managed and used and how does that impact what one can say about it?

To be able to accurately interpret a source it is essential to understand the context in which it was created, managed and used. This is particularly challenging in the context of born digital source materials, as there is a rapid and continual churn in the underlying technology and formats that interact with shifting behaviors and social contexts for interpreting the meaning of those behaviors.

As an example, consider what the email signature "Sent from my iPhone" at the bottom of a message communicates¹⁵. First off, that the sender sent an email from a

¹⁴ For a series of examples of how different browser rendering can dramatically affect the appearance of a born digital work of art see Fino-Radin, Ben. "Rhizome Artbase: Preserving Born Digital Works of Art." Washington, D.C, 2012.

http://www.digitalpreservation.gov/meetings/documents/ndiipp12/DigitalCulture_fino-radin_DP12.pdf.

¹⁵ For discussion of how email signatures like "sent from my iPhone" effect how messages are interpreted see Carr, Caleb T., and Chad Stefaniak. "Sent from My iPhone: The Medium and Message as Cues of Sender

mobile device which likely explains why there might be typos or it might be brief because of the limits of a smaller interface. At the same time, it tells us that the user didn't care to change the default signature that Apple added to their messages. So email's aren't just emails. The conventions and forms of the medium have developed and changed over time and what it means to send and receive an email has changed too. Part of understanding and interpreting a particular email is going to involve understanding the context through which it was created and the social conventions around email at a given point in time.

Continuing in the case of email, the way that individuals manage their email and how that email is acquired and processed is going to be an important part of interpreting archives of email. Some email users keep complex folder structures for managing email. In some cases organizations restrict the total size of storage space for users to keep email, so individuals end up managing their email by deleting emails to make space for new ones. At the same time, the development of services like Gmail have encouraged a different set of behaviors where individuals are increasingly keeping all of their email and simply using search to work their way through their messages¹⁶. To this end, developing an understanding of what an individual's practices and or an organization's practices were around email will be a key part of making sense of any given set of emails.

To illustrate another area of born digital content that has these issues consider the way that people take, manage and work with digital photographs. One of the primary characteristics of digital objects is that it is generally trivial to make exact copies, or seemingly exact, copies of them. As a result, when it comes to digital photographs, people will often have an assortment of copies of an image with varying amounts of metadata associated with them¹⁷. There is the original file from a camera or a phone, a copy downloaded to a hard drive that might be edited and a range of derivative copies created for sharing on Facebook or a series of photos using different filters. While the original might be the highest resolution, the derivative files are likely seen more and it's likely that the metadata and descriptive information about each copy can be different. As a result, there isn't really a master file or copy, so much as there is a constellation of different versions of the photo that each can be studied to understand a personal digital media ecology of an individual or organization.

It is also worth underscoring that what a photo means in a given moment is itself historically contingent as well¹⁸. In the last few years more photographs have been taken

Professionalism in Mobile Telephony." *Journal of Applied Communication Research* 40, no. 4 (November 1, 2012): 403–24. doi:10.1080/00909882.2012.712707.

¹⁶ A growing body of research on how people manage digital information will likely be invaluable for future historians in contextualizing the strategies that individuals used to organize and manage their digital information. For example see, Henderson, Sarah, and Ananth Srinivasan. "Filing, Piling & Structuring: Strategies for Personal Document Management." In *System Sciences (HICSS), 2011 44th Hawaii International Conference on*, 1–10. IEEE, 2011.

http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5718470.

¹⁷ For an exploration of the various reasons individuals copy, edit and describe a range of derivative copies of digital photos see Marshall, Catherine C. "Digital Copies and a Distributed Notion of Reference in Personal Archives." In *Digital Media: Technological and Social Challenges of the Interactive World*, edited by Megan Alicia Winget and William Aspray, 89–115. Lanham, Md: Scarecrow Press, 2011.

¹⁸ For documentation of the historically contingent nature of photographs and an exploration of issues in interpreting photos from different historical contexts see Trachtenberg, Alan. *Reading American*

then in the two hundred or so years since the camera was invented. At this point, there are more than 6 billion photos on Flickr, and hundreds of millions of photos on Facebook and Instagram¹⁹. The combination of camera phones and sites like Flickr, Instagram & Facebook have created a set of practices and social norms where all kinds of people take sequences of photos throughout their day and share them. Similarly, the fact that camera phones quickly began to have two cameras, one in the front and one in the back, illustrates the shift toward the emergence of the selfie as a key use of photographs. In this vein, photos increasingly play a role in the presentation of self in everyday life.

With this noted, digital photos increasingly come with a considerable amount of technical metadata embedded inside them that will be increasingly useful for historians studying these objects. Again, what is shown on the screen is only part of the story with digital objects. With a range of simple tools, it is possible to read the text information encoded through standards like EXIF which can document information about when a photo was originally taken, what software has been used to edit it, and the kind of camera that was used to originally take the photo. The result is that there exist inside many digital photographs records of the provenance of their creation and management that can be used to help contextualize and understand how they were in fact created.

What role did search play in the original experience of content?

The idea of original order, that the order materials are organized in by their creators and managers contains important value for contextualizing records, is somewhat at odds with the basic nature of digital media²⁰. From the perspective of an end user, there really isn't a first row in a database²¹. Instead, a user enters a query and the results of the query come in their own order. As a result, when content is preserved without preserving the interfaces to that content historians are going to be left needing to do a lot of reasoning and theorizing based on how they think those interfaces worked. This poses a key question to ask of born digital primary sources. What role did search interfaces and algorithms play in how users interacted with and made sense of content and what limitations on interpretation does likely not having that information impose? A few examples will illustrate this issue.

One of the biggest challenges facing web archives is that it is very unlikely that anyone is going to be able to recreate the central mode through which web content is accessed and understood. It is unlikely that there will be a historical Google search. While it is possible to find archived copies of many webpages at particular moments in time there won't be a way to figure out what someone in Washington D.C. who googled

Photographs: Images As History, Mathew Brady to Walker Evans. 1st ed. New York, N.Y.: Hill and Wang, 1989.

¹⁹ For an exploration of some trends in the history of numbers of photographs taken see Good, Jonathan. "How Many Photos Have Ever Been Taken?" *1000memories*, September 15, 2011. <https://web.archive.org/web/20120305055510/http://1000memories.com/blog/94-number-of-photos-ever-taken-digital-and-analog-in-shoebox>.

²⁰ Bailey, Jefferson. "Disrespect Des Fonds: Rethinking Arrangement and Description in Born-Digital Archives - Archive Journal Issue 3." *Archive Journal*, no. 3 (2013). <http://www.archivejournal.net/issue/3/archives-remixed/disrespect-des-fonds-rethinking-arrangement-and-description-in-born-digital-archives/>.

²¹ For an exploration of the logic, structure and assumptions of databases see Manovich, Lev. *The Language of New Media*. Cambridge, Mass: MIT Press, 2002 pp. 212-236.

“Benghazi” in March of 2015 would have seen in the search results. Given that search is the primary mode through which web content is found and accessed that means it won’t be easy to figure out what it is likely that people will have come across.

As a related example, consider if someone want’s to study visual representations of any given topic in the 6 billion photos on Flickr. Even if there is an archived copy of all those photos, it would be challenging to figure out what photos someone might have seen if they searched the site at a given point in time. From that archived copy of the photos and their metadata it would be possible to study what kinds of photos people created and shared and through the metadata the relative popularity of given images. However, if one wanted to know what someone would find when they visited Flickr and searched for something you would also need to have a copy of Flickr’s proprietary “interestingness” algorithm which is used to sort out what photos are shown based on a series of weights assigned to different characteristics of photos²².

Examples of the role of search in the use of digital media are everywhere. The capability of search is itself increasingly shifting how people manage their information, from a “filing” mentality to “piling,” and the result is that knowing how search worked in Gmail, or in the Mac operating system, is going to be increasingly important for making sense of born digital primary sources.

These various questions asked of digitized and born digital sources connect directly to a broader set of issues in how aggregations and collections of these materials are established and described. In this area many different kinds of projects have started to be described as digital archives. In what follows I will briefly explore some of the ways the term is used and discuss the issues that arise in terms of interpreting the various kinds of sources in these different kinds of digital archives.

What are Digital Archives?

When archivists, historians and digital humanists use the term “digital archive” they often mean different and overlapping things. I’m not so much interested in trying to decide whose use of the term is right or wrong, but in clarifying what the term means in different contexts. In each case below, I have provided an example or two of this type of usage and worked to connect the kind of usage back to the questions one needs to ask of the digital primary sources contained in them.

Collections of Aggregated Digitized Primary Sources

When digital humanities scholars use the term digital archive, they are often describing aggregated collections of digitized primary sources. For example, the Shelly Godwin Archive brings together digitized copies of primary source manuscript collections from a range of different archives around the world to create a single place to access the papers of a particular family.

²² For an example of working through a set of search results on Flickr as a primary source see Owens, Trevor. “Lego, Handcraft, and Costumed Zombies: What Zombies Do on Flickr.” *New Directions in Folklore* 12, no. 2 (2015): 3–25.

Historian Joshua Sternfeld has suggested considering calling these kinds of projects a genre of “digital historical representations”.²³ Sternfeld uses that term to talk more broadly about the diverse range of products historians are now creating from digitized sources, including visualizations and databases, but included these kinds of digital archives under this umbrella. He included these in this category as they tend to be more expansive in what they bring together than what archives have generally focused on.

The origin of this usage is anchored in Jerome McGann’s work on the Rossetti Archive²⁴. The Rossetti Archive presents a dizzying array of sources related to 19th century poet, illustrator and painter Dante Gabriel Rossetti. It contains much of what one might find in an archive, like copies of manuscripts and correspondence. However it also includes copies of published works like books and poems as well as a range of visual works by other artists, contemporary periodicals and other related texts. The site provides a wealth of resources and a mixture of interpretation and exhibition of those sources. However, it is often challenging to parse exactly what the scope is of what one is looking at in the site.

The idea behind the Rossetti Archive, and a related idea in the William Blake Archive, was to develop a sort of ever growing hypertext aggregation of related digital copies of sources anchored around an individual²⁵. In this vein, it has much more of a hybrid of a critical edition with the idea of providing the breadth of resources one might find in a literary archive.

When working with sources in this kind of digital archive it is essential to understand the context from which the original source materials were taken. In this case, the site is likely presenting materials from a range of different provenance and as such it is important to identify where something is coming from and then think through the kinds of questions one considers about why a particular object persists and others don’t related to the history of a given source.

Digitized Copies of Entire Archival Collections

In some cases, the term digital archive is used to refer to a digitized copy of the entire contents of an archival collection. For example, the Clara Barton Papers at the Library of Congress are available in full online. It’s not just the contents of the collection that was digitized but the folders they are contained in as well.

Presented online according to the boxes and folders they can be found in at the physical collection in Washington D.C. this kind of presentation of sources provides transparent access to the collection as it was arranged and described by archivists. In this vein, the scope and context note in the same finding aid that one would use to contextualize sources and understand how selection and arrangement decisions were made is useful for working with the digitized collection. To this end, something like the Clara Barton papers is functionally a digital surrogate of an entire manuscript collection.

²³ Sternfeld, Joshua. “Archival Theory and Digital Historiography: Selection, Search, and Metadata as Archival Processes for Assessing Historical Contextualization.” *The American Archivist* 74, no. 2 (October 1, 2011): 544–75.

²⁴ McGann, Jerome J., ed. *The Complete Writings and Pictures of Dante Gabriel Rossetti*. Accessed August 8, 2015. <http://www.rossettiarchive.org/>.

²⁵ McGann, Jerome J. “The Rationale of Hyper Text.” *Text* 9 (January 1, 1996): 11–32.

In a case like the Barton papers, the provenance of a given collection is much clearer and easier to parse than in the case of the previously discussed aggregations of digitized sources. With that noted, it is worth considering why a particular archive is digitized and not another as that itself represents its own selection/appraisal like decision. In the case of collections at most archives it will be a mixture of legal issues (generally focusing on digitizing older collections that are much less likely to involve a range of copyright and other rights issues), issues of what is thought to be most popular, and what is easiest to digitize.

As another example of where this kind of selection issues is raised, many state archives and historical societies are entering into contracts with companies like Ancestry.com to digitize large parts of their collections. In these cases, companies are generally deciding what collections to digitize based on what they deem to be the most useful to the genealogists who are their customers²⁶. To this end, it is worth considering why a particular collection is available and the extent to which the selection of that collection over another for digitization might change the direction of your research and writing. With that said, this is a much less significant issue than in other cases where individual documents have been cherry picked from an archival collection and digitized in that you have a sense of the structure and content of a whole coherent archival collection.

Aside from issues of selection, it is also important to think through considerations of the quality of a given set of reproductions of sources for your purpose. In the case of the Clara Barton papers, part of why they were digitized in full is that the entire collection was already microfilmed. So instead of doing high quality digital captures of the original documents it was much less expensive to simply digitize the black and white microfilm. For most purposes those digitized copies of the microfilm are perfectly serviceable. However, as the cases from the EEBO Shakespeare folios illustrated, higher quality color images of the documents would likely enable access to a much broader range of the potentially significant properties of those documents. So it's still important to consider if the quality of a digital reproduction of an object is good enough for the purpose one intends to use it for.

Born Digital Archival Collections

When archives acquire born digital materials and process those collections the results are often called digital archives, or born digital archives, as well. For example, Emory University acquired Salman Rushdie's papers that came with a series of his laptops²⁷. Disk images were created of those laptops and at this point it is possible for researchers to login and study the contents and environment he worked in. In this case, researchers can engage directly with an emulated version of his whole computer.

In this case, the digital archive is generally a subset or a hybrid component of an analog archival collection. Often these kinds of materials are described as part of a

²⁶ For a discussion of how digitization selections are made in public private partnerships see Kriesberg, Adam M. *The Changing Landscape of Digital Access: Public-Private Partnerships in US State and Territorial Archives.*, 2015. <http://deepblue.lib.umich.edu/handle/2027.42/111584> pp. 122-125.

²⁷ For further background on the Salman Rushdie digital archive see Emory University. *Rushdie Researcher Workstation Tutorial*, 2011. https://www.youtube.com/watch?v=oiqHv_SofNo.

finding aid and as such it is relatively easy to ascertain their provenance and understand why a particular set of digital objects exists and how decisions have been made in terms of their processing, arrangement and description. With that noted, the standards and practices for collecting, processing and preserving born digital archival material are still developing and evolving. So the quality and consistency of how born digital materials are described and made available varies widely across different repositories.

All of the questions and issues raised earlier about born digital primary sources are important to consider when working with these kinds of collections. In much the same way that a historian who studies 18th century documents needs to learn to read various kinds of handwriting scripts to develop an ability to read and decipher those texts, historians are going to need to develop sophisticated understandings of how digital media systems functioned at particular points in time and how different kinds of users used them. For example, understanding how different people organize their desktops, or how they name their files, and how conventions around those sorts of things have changed over time will be an important part of interpreting born digital archives.

Web Archives

Web Archives represent another genre of born digital archives that are both significant and different enough to warrant their own consideration. The Internet Archive, a range of National Libraries, and a host of smaller archives and libraries are engaged in work to collect and preserve websites and webpages and these collections are going to be of critical importance for future research. With that said, Web Archives represent a rather different approach to collecting and organizing sources.

The various organizations that archive the web use tools like Heritrix, an open source web crawler, that are sent out to grab all of the rendered content of a webpage they can get ahold of and, within defined parameters, the other pages that link to it and all their associated files. As part of this collection process, the tools log information about the date and time that the data was collected. At this point, tools store that content in WARC files, or Web Archive files, which can then be played back via tools like the Wayback Machine. So there is a lot of information in here that can be used to assert the authenticity of the data, how a particular URL presented itself to Heritrix and how Heritrix interpreted it at a particular moment in time.

There are a few key points for interpreting and studying web archives. First, web archives are consciously created. That is, an organization has a selection policy and works to collect sites that fit with that policy. So understanding those policies and the scope of a given collection is a key part of interpreting it. In that vein, it is also important to understand how a given repository works, that is many organizations require permission from content creators to collect particular kinds of sites, so in those cases, the scope of a given collection is only going to contain content from site owners that were OK with having their content collected and preserved.

Along with that, a given archived website is actually a copy of how the content of a given URL presented itself to the web crawler at a given moment in time. So, for example, if a site reconfigures how it displays itself based on the IP address of a site visitor then that will be reflected in the archived copy. There are various ways that web crawling technologies can miss some of the content provided as well. So it is important to remember that web archives are not exact and pristine copies of the content of a

particular URL at a moment in time but instead copies of how that content appeared to the crawler at that point in time.

Collections of User Generated Born Digital Primary Sources

One of the biggest affordances of the World Wide Web is the ability for users to respond; to comment, to upload and “share”. This has not been lost on historians and archivists. Projects like the September 11 Digital Archive illustrate the possibility to “crowdsource” an archive and create a collection of born digital materials around a particular issue or topic.

Shortly after the September 11th attacks, the American Social History Project at the City University of New York Graduate Center and the Roy Rosenzweig Center for History and New Media launched a site that allowed anyone to upload records and reflections related to the attacks²⁸. It contains copies of email messages, digital photographs, and a range of first hand accounts which a range of site visitors have provided over time. This sort of archive has been similarly developed around other incidents, like the Hurricane Digital Memory Bank created to generate a digital record of Hurricanes Katrina and Rita²⁹.

Where archival collections, like the papers of an individual or the records of an organization, accrue over time and have a clear and central connection to the individual or organization as the basis of their provenance these crowdsourced collections have a different kind of cohesion. Something like the September 11th digital archive can't be understood as being a representative sample of individual's reactions. It is a partial collection made up of who decided to participate at any given time. To that end, the individual reflections and objects in the collection are invaluable as records of individual experience but making sense of them as a whole is going to be challenging. Ideally, as researchers work with these kinds of collections in the future they will focus on understanding the kinds of voices that are represented in the collections as much as they work to interpret those voices. To that end, records of how these sites prompted users to participate and how those prompts developed and changed over time and how decisions were made about how to set up a site are going to be invaluable for helping researchers understand the scope and content of these collections.

Going Forward

Sources don't speak for themselves. To that end, historians have developed and deployed techniques for interrogating and understanding sources based on their properties and the context of their creation, use and management. In this essay I've worked to explicate some of the work necessary for historians to continue to be as rigorous in working with digital sources and archives as they have been with their analog counterparts.

The key questions of source criticism are the same irrespective of if a source is digital or not. However, given the rapid pace of change around digital technology it is

²⁸ For further exploration of the September 11th digital archive see Roy Rosenzweig Scarcity or Abundance? Preserving the Past in a Digital Era *American Historical Review* 108, 3 (June 2003): 735-762 as well as Between archive and participation: Public memory in a digital age *E Haskins* Fall 2007 37, 4

²⁹ For more background on this see Why Collecting History Online is Web 1.5 Sheila A. Brennan and T. Mills Kelly Center for History and New Media, Case Study <http://chnm.gmu.edu/essays-on-history-new-media/essays/?essayid=47>

likely that historians are going to need to increasingly focus on establishing and sharing techniques for working with different kinds of digital sources. As information ecologies continually shift it is going to be critical for historians to show their work in making sense of the stratigraphy of digital sources.

Acknowledgements:

Thanks to Bertram Lyons, Porter Olson, for providing thoughtful and helpful comments on drafts of this essay.