International Journal of Digital Humanities (2021) 1:325–341
https://doi.org/10.1007/s42803-020-00028-7

RESEARCH ARTICLE

Digital sources and digital archives: historical evidence in the digital age

Trevor Owens [1] · Thomas Padilla [2]

Received: 6 July 2019 / Accepted: 22 April 2020/ Published online: 4 May 2020

Abstract

As the cultural record becomes increasingly digital the evidentiary basis of history expands and shifts. How must historical scholarship change when the evidentiary basis shifts toward the digital? Through explorations of a series of born digital and digitized sources, we identify and discuss key issues relating to humanities scholars ability to develop claims and arguments grounded in digital sources and digital archives. In exploring these issues in digital source criticism, we work to provide practical guidance for scholars on key issues and questions to consider when working with born digital and digitized primary sources.

Keywords Digital history · Historigraphy · Research methods · Collections as data · Source criticism · Digitization · Archives

The world is full of potential primary sources. Almost anything can be a source. The rings of a tree testify to weather conditions and changes in climate (Cronon 1983). Probate records document the material goods individuals held at the end of their lives (Bushman 1992). Court proceedings offer insight into the experiences of the oppressed (Pagan 2003). Just as any kind of physical object might serve as a source, so does a digital source. As societies increasingly express themselves using digital means, the evidentiary bases of history expands.

The historians' ability to study the past is largely 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 national institutions, small local historical societies, and manuscript collections at research libraries to name a few

✉ Trevor Owens
trevor.johnowens@gmail.com

Thomas Padilla
tgpadillajr@gmail.com

[1] University of Maryland, College Park, MD, USA

[2] University of Nevada, Las Vegas, NV, USA

examples. Increasingly these collections are digital. This change of state is the product of decades of digitization effort commingling with the collection of contemporary culture that begins its life in digital form – think email, word documents, photos from mobile phones, websites, software, code, and social media data.

How does Historical scholarship change when the evidentiary basis shifts toward the digital? How is interacting with digital archives different or similar? What does it even mean to have a "digital archive"? What follows is an attempt to identify and discuss implications of these questions relative to the Historian's' ability to develop claims. We work toward practical advice that bears on some the of issues and questions one should consider when working with digital sources.

1 What are digital sources?

As Martha Howell and Walter Prevenier explain in their introduction to historical analysis of sources, "… to make wise choices among potential sources, historians must … 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" (Howell and Prevenier 2001, p. 28). It is essential to ask these same kinds of questions of digital sources. This is particularly challenging given that pace of change in communications technologies and media continues to accelerate.

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 an e-mail message?

Making sense of a source in order to establish a defensible claim requires context. Knowing that 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. 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 a Gmail account they would need to know about many of the automated rules a user set that "mark as read" emails from a range of individuals and organizations and in some cases that are set to "skip" the inbox entirely. Without knowing about those rules one could end up making all kinds of problematic inferences about what a user had or had not read based on the parameters of their email system. As a result, the future of sound historical interpretation is going to be dependent on cultural histories of platforms like email systems.

As this frame of consideration is expanded we must also contend with the implications that algorithmically tailored digital environments pose for historical interpretation. Simply put, participants in digital environments may not have been aware of the forces shaping their behavior. For example, email commonly includes paid advertising and content is 'promoted' to users of social media platforms like Twitter and Facebook. In order to interpret sources whose production was algorithmically facilitated Historians must have access to documentation that speaks to technical infrastructure that

subjectively shapes human interactions – things like versioned application programming interfaces and standards governing the exchange of data across computational systems.

2 Digitized primary sources

In working with digitized sources it is essential to ask a range of questions about the production of the digitized copy. We breifly detail examples of key questions to ask of digitized sources.

2.1 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. The same question 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. One must be able to contextualize a source and understand why they have it at hand and as such it is important to think through the kinds of limitations on claims relative to what you know about the policies of a library, archive, or museum.

For example, because of copyright restrictions many institutions in the United States are focusing efforts on digitizing materials from before 1923. Or an archive might have the rights cleared to digitize one particular collection, or the writings of one person instead of another. Indeed, as public private partnerships increasingly drive digitization efforts in libraries historians need to be considering how those pressures are shaping access to particular sets of sources (Thylstrup 2019; Kriesberg 2015) In each case 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.

2.2 Is this copy of significant quality for my purpose?

All digitized objects are surrogates for and representations of the originals (Jones 2014). That is 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."(Tyrrell 2005, p. 38). 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 qualities (Fleischhauer 2011). 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 (Rudy 2010). 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 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 source (Werner 2012). 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.

2.3 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 before (Leary 2005). Searching across millions of sources at once is transforming many historians' methods for research and scholarship (Ramsay 2014; Putnam 2016). At the same time, full text search presents a whole new set of challenges for reasoning from and interpreting sources (Gibbs and Owens 2013).

Where in the past one would develop an explicit sampling 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. It's important for historians to begin to understand and document how digitization practices and how the affordances of particular sources, like those with typewritten text, produce unevenness in the discoverability and usability of collections (Wright 2019).

3 Born digital sources

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 twenty-first 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. At this point, archivists, librarians, curators and other cultural heritage professionals have been collecting, managing, preserving and providing access to born digital primary sources for more than half a century (Owens 2018). In this context, it's critical for historians to continue to develop forms of source criticism for born digital records. What follows is an initial exploration of some key source criticism questions to ask of born digital sources.

3.1 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 performance (Kirschenbaum 2007; Arcangel 2014). The actual digital object is a sequence of data values registered on a medium. Hard drives, CDs, flash drives, etc. are all things that register sequences of values 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 typically 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 appear (Montfort 2004). 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 created (Reside 2011). 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.

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 editor (Kirschenbaum 2008). This ROM had a copy of an early video game called Mystery House. A Hex editor renders the hexadecimal notation, a calculation 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 files system asserts 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.

3.2 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 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 it (Fino-Radin 2012).

3.3 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 communicates (Carr and Stefaniak 2012). First off, that the sender sent an email from a 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 are 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 messages (Henderson and Srinivasan 2011). 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 them (Marshall 2011). 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 well (Trachtenberg 1989). In the last few years more photographs have been taken 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 Instagram (Good 2011). 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.

3.4 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 media (Bailey 2013). From the perspective of an end user, there really isn't a first row in a database (Manovich 2002). 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 (Lynch 2017). 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 "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 wants 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 photos (Owens 2015). While historians have a wealth of work exploring how texts and objects have been received, significantly, in the context of forms of social media like Flickr photos, reception and use of digital content both change how that content is presented to other users and are simultaneously documented in metadata within these systems.

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 (Kalman and Ravid 2015).

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 we 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.

4 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, we 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.

4.1 Collections of aggregated digitized primary sources

When historians and other 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.

Historian Joshua Sternfeld has suggested considering calling these kinds of projects a genre of "digital historical representations" (Sternfeld 2011). 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 Archive (McGann 1996). The Rossetti Archive presents a dizzying array of sources related to nineteenth 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 individual (McGann 1996). 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.

4.2 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 from the finding aid for the physical collection works just as well for contextualizing the digital surrogates of these sources. To this end, something like the Clara Barton papers is functionally a digital surrogate of an entire manuscript collection.

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 customers (Kriesberg 2015). 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.

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. While it may seem straightforward to separate informational and artifactual aspects of objects, in practices, there is a nearly infinite amount of information which can exist in the artifactual qualities of objects. 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.

4.3 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 laptops (Kirschenbaum et al. 2009). 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. 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 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 continue to develop and evolve as technology and media continue to evolve. 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 eighteenth 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.

4.4 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 re-rendered 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. In some cases this is individual collections, in the cases of many national libraries it is associated with legal mandates to harvest as much as possible of a national web domain, like .fr or .uk. 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.

4.5 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 attacks (Rosenzweig 2003 and Haskins 2007). 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 Rita (Brennan and Kelly 2009).

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.

5 Sources as data

To this point, we have used the concept of the "source" to set the initial context for interactions with digital content. A growing movement asks many of the same questions we have explored in this article but begins a bit differently (Padilla 2017). Rather than ask "How can I make an argument with this source?", the question becomes, "How can I make an argument with this source as data?" The shift in perspective is meant to help assess and engage the meaning making capacity of a source in a manner that is attuned to the nature of its material conditions - information expressed as data stored on a medium with computer as mediator. Attention to the affordances of data relative to analog materials opens up additional modes of interaction. Where there is no analog precursor for the data, as is the case with much contemporary knowledge production this shift in perspective becomes all the more important. Given the nature of their form, data are amenable to questions at macroscopic and microscopic scales. At the macroscopic scale it becomes possible to use methods like text mining, data mining, machine learning, and or image analysis to ask questions across hundreds, thousands, and even millions of sources (Lorang et al. 2015). At the microscopic scale working with sources as data also allows for more granular engagements (Froehlich 2018). At its base a source as data framing is meant to be generative. That is to say, seeing the source as data should allow for the possibility of more questions to be asked.

A sources as data framing also has imperative connotations. It is imperative in the sense that it helps align historical scholarship with a wider field of contemporary theory and practice wherein data is a primary rather than secondary consideration. Engagement with a wider field is vital to helping historians sort out how to give their disciplinary training purchase on a contemporary knowledge environment. Consider how a historian might evaluate Twitter data as a source 40 years from now. Seemingly staccato statements in the form of "tweets" are issued 24 h a day from all corners of the globe. It is not readily apparent what exists beyond the text of the statements without digging into the standards and systems that give each tweet their context. Opacity on this point is a design decision by a for-profit company. In actuality, every tweet is governed by something called a metadata standard. For each tweet, the text of the message it contains is a minority of information that is transmitted. Contextual information like date of transmission, geocoding, and a wide array of relational information (number of followers, number of retweets, links to multimedia resources, etc.) is packaged along with it. The metadata standard provides grist for the computational

mill, allowing programmers around the world to use an application programming interface to develop custom applications and "bots" that constantly recontextualize user interaction with Twitter. Standardized capture of contextual information and the systematic interaction that it supports is the lifeblood of Twitter as company and a vector for "fake news" manipulation (Meyer 2018). Twitter hides much of this complexity from the user.

A sources as data orientation is an investment in development of a critically oriented research practice. Without engagement with digital sources as data a Historian runs the risk of becoming complicit with systems designed to monitor, extract, and sell information about human activity. Notably, historians have significant experience in reading presences and silences in archival records in areas like colonial history (Trouillot 1995 and Pagan 2003). That kind of research is only possible through a thorough understanding of the structures and systems of oppression and that similarly holds true for approaching the structures and systems that produce and manage data. Building on the Twitter example, consider the uptick in academic study of Twitter use. The uptick makes sense given the company's role in providing a platform that many people have used to organize large scale protests. Black Lives Matter is a prime example. In the wake of a number of tragedies and with clear signals that Twitter was a key component of social activism, archival projects like *Documenting the Now* arose, provisioning tools to enable large scale capture of social media data (Jules 2016). Running alongside this effort Bergis Jules noted the rise of for profit entities that captured Twitter data, packaged it, and sold it in the form of threat profiles that law enforcement agencies could use to target leaders like DeRay McKesson. It is an open question the extent to which research with data of these kind is directly or indirectly feeding into private sector efforts that run counter to ethical commitments. Institutional Review Boards generally exempt themselves from an example of this kind. Without an understanding of these sources as data the Historian's ability to consider the harm that their research could inflict on a community in the first, second, or third instance is severely attenuated.

While Twitter may not persist long into the future there will surely be more like it around the corner. In a world turning toward predominant born digital knowledge production a naturalized disposition toward data is required. It is required in order to establish arguments that have purchase on the complexity and scale of the data traces we leave behind. What might we gain by thinking of a source as data? It is a seemingly simple distinction that a growing community of scholars and cultural heritage professionals are finding value in.

6 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 we'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. Throughout, we have shared some examples of the ways that historians are beginning to develop this kind of digital source criticism. As digital

sources become more and more central to historical scholarship it is imperative that this kind of scholarship becomes a key area of scholarly concern.

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 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.

References

Arcangel, C. (2014). The Warhol files: Andy Warhol's long-lost computer graphics. *Artforum*, (summer). Retrieved from https://artforum.com/inprint/issue=201406&id=46874

Bailey, J. (2013). Disrespect des Fonds: Rethinking arrangement and description in born-digital archives - archive journal issue 3. Archive Journal, (3). Retrieved from http://www.archivejournal.net/issue/3/archives-remixed/disrespect-des-fonds-rethinking-arrangement-and-description-in-born-digital-archives/

Brennan, S., & Kelly, M. (2009). *Why Collecting History Online is Web 1.5* (Center for History and new Media, case study). Fairfax, VA. Retrieved from https://rrchnm.org/essay/why-collecting-history-online-is-web-1-5/

Bushman, R. L. (1992). *The refinement of America: Persons, houses, cities*. New York: Knopf.

Carr, C. T., & Stefaniak, C. (2012). Sent from my iPhone: The medium and message as cues of sender professionalism in Mobile telephony. *Journal of Applied Communication Research, 40*(4), 403–424. https://doi.org/10.1080/00909882.2012.712707.

Cronon, W. (1983). *Changes in the land: Indians, colonists, and the ecology of New England*. New York: Hill and Wang.

Fino-Radin, B. (2012). *Rhizome ArtBase: Preserving Born Digital Works of Art. Presented at the digital preservation*. Virginia: Arlington Retrieved from http://digitalpreservation.gov/meetings/documents/ndiipp12/DigitalCulture_fino-radin_DP12.pdf.

Fleischhauer, C. (2011). Information or Artifact: Digitizing a Book, Part 1 | The Signal: Digital Preservation [webpage]. Retrieved August 7, 2015, from http://blogs.loc.gov/digitalpreservation/2011/10/information-or-artifact-digitizing-a-book-part-1/

Froehlich, H. (2018, June). Distance-reading the feminine landscapes of the awakening. Retrieved December 4, 2018, from https://blog.bham.ac.uk/clic-dickens/2018/06/29/distance-reading-the-feminine-landscapes-of-the-awakening/

Gibbs, F., & Owens, T. (2013). The hermeneutics of data and historical writing. In K. Nawrotzki & J. Dougherty (Eds.), *Writing history in the digital age*. University of Michigan Press Retrieved from http://hdl.handle.net/2027/spo.12230987.0001.001.

Good, J. (2011, September 15). How many photos have ever been taken? Retrieved August 8, 2015, from https://web.archive.org/web/20120305055510/http://1000memories.com/blog/94-number-of-photos-ever-taken-digital-and-analog-in-shoebox

Haskins, E. (2007). Between archive and participation: Public memory in a digital age. *Rhetoric Society Quarterly, 37*(4), 401–422. https://doi.org/10.1080/02773940601086794.

Henderson, S., & Srinivasan, A. (2011). Filing, piling & structuring: Strategies for personal document management. In *System Sciences (HICSS), 2011 44th Hawaii International Conference on* (pp. 1–10). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5718470

Howell, M. C. & Prevenier, W. (2001). *From Reliable Sources:* An Introduction to Historical Methods. Ithaca, N.Y: Cornell University Press.

Jones, S. (2014). *The emergence of the digital humanities*. New York: Routledge.

Jules, B. (2016). Surveillance and social media archiving. *Documenting the Now.* https://news.docnow.io/surveillance-and-social-media-archiving-7ea21b77b807

Kalman, Y., & Ravid, G. (2015). Filing, piling, and everything in between: The dynamics of E-mail inbox management. *Journal of the Association of Information Science and Technology., 66*, 2540–2552. https://doi.org/10.1002/asi.23337.

Kirschenbaum, M. (2007). *Mechanisms: New media and the forensic imagination*. Cambridge, Mass: MIT Press.
Kirschenbaum, M. (2008). *Mechanisms: New media and the forensic imagination*. Cambridge: MIT Press.
Kirschenbaum, M., Farr, E. L., Kraus, K. M., Nelson, N., Peters, C. S., Redwine, G., & Reside, D. (2009). Digital Materiality: Preserving Access to Computers as Complete Environments. Retrieved from https://escholarship.org/uc/item/7d3465vg
Kriesberg, A. M. (2015). *The Changing Landscape of Digital Access: Public-Private Partnerships in US State and Territorial Archives*. Retrieved from http://deepblue.lib.umich.edu/handle/2027.42/111584
Leary, P. (2005). Googeling the Victorians. *Journal of Victorian Culture, 10*(1), 72–86.
Lorang, E. M., Soh, L.-K., Datla, M. V., & Kulwicki, S. (2015). Developing an image-based classifier for detecting poetic content in historic newspaper collections, Developing an Image-Based Classifier for Detecting Poetic Content in Historic Newspaper Collections.
Lynch, C. (2017). Stewardship in the "age of algorithms. First Monday 22 (12). https://firstmonday.org/article/view/8097/6583
Manovich, L. (2002). *The language of new media*. Cambridge: MIT Press.
Marshall, C. C. (2011). Digital copies and a distributed notion of reference in personal archives. In M. A. Winget & W. Aspray (Eds.), *Digital media: Technological and social challenges of the interactive world* (pp. 89–115). Lanham: Scarecrow Press.
McGann, J. J. (1996). The rationale of hyper text. *Text, 9*, 11–32.
Meyer, R. (2018). The grim conclusions of the largest-ever study of fake news. Retrieved December 4, 2018, from https://www.theatlantic.com/technology/archive/2018/03/largest-study-ever-fake-news-mit-twitter/555104/
Montfort, N. (2004). Continuous Paper: MLA. Retrieved February 3, 2017, from http://nickm.com/writing/essays/continuous_paper_mla.html
Owens, T. (2015). Lego, handcraft, and costumed zombies: What zombies do on Flickr. *New Directions in Folklore, 12*(2), 3–25.
Owens, T. (2018). *The theory and craft of digital preservation*. Baltimore: Johns Hopkins University Press.
Padilla, T. (2017). On a collections as data imperative.
Pagan, J. R. (2003). *Anne Orthwood's bastard: Sex and law in early Virginia*. New York: Oxford University Press.
Putnam, L. (2016). The transnational and the text-searchable: Digitized sources and the shadows they cast. *The American Historical Review., 121*(2), 377–402. https://doi.org/10.1093/ahr/121.2.377.
Ramsay, S. (2014). The hermeneutics of screwing around; or what you do with a million books. In K. Kee (Ed.), Pastplay: Teaching and learning history with technology. University of Michigan Press. Retrieved from http://hdl.handle.net/2027/spo.12544152.0001.001
Reside, D. (2011) "'No Day But Today': A Look at Jonathan Larson's Word Files," New York Public Library Blog. http://www.nypl.org/blog/2011/04/22/no-day-today-look-jonathan-larsons-word-files.
Rosenzweig, R. (2003). Scarcity or abundance? Preserving the past in a digital era. *The American Historical Review, 108*(3), 735–762.
Rudy, K. (2010). Dirty books: Quantifying patterns of use in medieval manuscripts using a densitometer. Journal of Historians of Netherlandish Art. 2 (1). https://doi.org/10.5092/jhna.2010.2.1.1.
Sternfeld, J. (2011). Archival theory and digital historiography: Selection, search, and metadata as archival processes for assessing historical contextualization. *The American Archivist, 74*(2), 544–575.
Thylstrup, N. (2019). *The politics of mass digitization*. Cambridge: MIT Press.
Trachtenberg, A. (1989). *Reading American photographs: Images as history, Mathew Brady to Walker Evans* (1st ed.). New York: Hill and Wang.
Trouillot, M. (1995). *Silencing the past: Power and the production of history*. Boston: Beacon Press.
Tyrrell, I. R. (2005). *Historians in public: The practice of American history, 1890–1970*. Chicago: University of Chicago Press. Retrieved from http://www.loc.gov/catdir/toc/ecip058/2005003459.html.
Werner, S. (2012). Where material book culture meets digital humanities. Journal of Digital Humanities, 1(3). Retrieved from http://journalofdigitalhumanities.org/1-3/where-material-book-culture-meets-digital-humanities-by-sarah-werner/
Wright, R. (2019). Typewriting mass observation online: Media imprints on the digital archive. *History Workshop Journal., 87*(Spring), 118–138.