International Journal of Digital Humanities (2019) 1:71–84
https://doi.org/10.1007/s42803-019-00006-8

RESEARCH ARTICLE

The invention and dissemination of the spacer gif: implications for the future of access and use of web archives

Trevor Owens[1] · Grace Helen Thomas[1]

Published online: 5 April 2019

Abstract

Over the last two decades publishing and distributing content on the Web has become a core part of society. This ephemeral content has rapidly become an essential component of the human record. Writing histories of the late 20th and early 21st century will require engaging with web archives. The scale of web content and of web archives presents significant challenges for how research can access and engage with this material. Digital humanities scholars are advancing computational methods to work with corpora of millions of digitized resources, but to fully engage with the growing content of two decades of web archives, we now require methods to approach and examine billions, ultimately trillions, of incongruous resources. This article approaches one seemingly insignificant, but fundamental, aspect in web design history: the use of tiny transparent images as a tool for layout design, and surfaces how traces of these files can illustrate future paths for engaging with web archives. This case study offers implications for future methods allowing scholars to engage with web archives. It also prompts considerations for librarians and archivists in thinking about web archives as data and the development of systems, qualitative and quantitative, through which to make this material available.

Keywords Web archiving · Computational scholarship · Cryptographic hash · Digital history

'The Web Is Ruined and I ruined it.' This is the title of author and Web Designer David Siegel's 1997 post to XML.com (Siegel 1997). Siegel, the author of the book *Creating*

The following research represents the opinions, perspectives and ideas of the authors. It does not necessarily represent the perspectives of any institutions with which they are affiliated.

✉ Grace Helen Thomas
grth@loc.gov

Trevor Owens
trow@loc.gov

[1] U.S. Library of Congress, Washington, DC, USA

Killer Websites (Siegel, 1996), went on to explain his role in what he describes as 'The Roots of HTML Terrorism.' (Siegel 1997) Specifically, he contends that 'The hacks I've espoused, especially the single-pixel GIF, and using frames and tables to do layout, are the duct tape of the Web.' All of these elements of design went out of fashion. As he explains, 'I ruined the Web by mixing chocolate and peanut butter so they could never become unmixed. I committed the hangable offense of mixing structure with presentation.' In particular, he advocated the use of these single-pixel, clear GIF files as a way of building page layouts. These kinds of technical discussions of design practices in web history are invaluable resources for understanding the records of the web (Owens 2015). One of his self-proclaimed offenses, 'the single-pixel GIF,' became a subject of analysis and study by digital artist and folklorist Olia Lialina in a 2013 online exhibit (Lialina 2013).

As part of an ongoing effort to explore and explain the early history of the web, Lialina produced the online exhibit illustrated below. This presentation, *clear.gif*, shows a series of transparent GIFs wrapped in elaborate frames. Widely referred to as 'spacer' GIFs, these single-pixel, transparent GIFs were used first and foremost as a way of controlling the placement and presentation of content on a website. They were invisible, or rather transparent, i.e. whatever was behind them showed through. However, they still took up space. So a designer could encode into their HTML document any number of spacer GIFs to appear in a row in order to control the placement of any given element on a page. This provided a means of controlling exactly where visual elements would appear on a given web page. As is evident in Fig. 1, they only become visible when broken, when the link to the image file no longer resolves.

These tiny files, the presence of which is only conspicuous when they are no longer present, are invaluable aids which help us understand the history of the web. Simultaneously, exploration of The study of these files, furthermore, offers insight toward the future of enabling scholarly research on the history of the web. In our explanation of the findings of this investigation, we identify key ways of working with records of the web, and born-digital collections more broadly, which can inform our future understanding of our digital past. The single-pixel GIF is an element of design, invisible like so many other aspects of design on the web, but still encoded in highly structured ways.

In an interview about her ongoing work to explore and understand the early web, in particular the Geocities archive, Lialina explains, 'I remember, everybody who made pages in the 1990s had cgif, maybe it was called clear gif, some people would call it

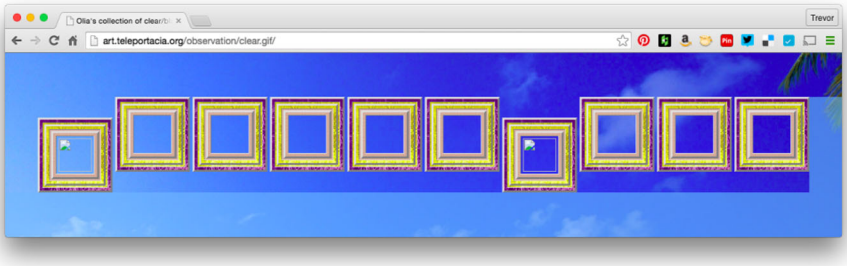

Fig. 1 Screenshot of *clear.gif* online exhibit

zero-dot-gif, but it was this transparent one that would help you to make layouts.' (Johnson 2011). Her exhibit functions as a way of drawing attention to this practice, but it also provides a point of entry to begin to explore the form and function of the history of these images in the history of web design.

In 2006, Jesper Rønn-Jensen, asked exactly this kind of question as a blog post: *Who Invented the Spacer Gif* (Rønn-Jensen 2006). Rønn-Jensen is an early web developer who has remained passionate and outspoken about the history of web design and development. In an update to the post, Rønn-Jensen notes that Siegel claimed credit in personal email correspondence with him. Specifically, Siegel claimed 'I invented it all by myself in my living room.' But at that point, another designer, software developer Joe Kleinberg, chimed in and claimed that he was really the one who had invented it (Rønn-Jensen 2007). What answers do web archives and other born-digital archives offer to such questions? Furthermore and in some ways more interestingly, in what ways might we be able to track the emergence and decline of something like the single-pixel GIF?

Cultural heritage organizations such as the Internet Archive, the British Library, the Library of Congress, and hundreds of others across the globe are working to collect and preserve the web. Many of these institutions now have significant holdings documenting more than two decades of the web's history. In what follows, we approach these collections as a means of exploring the ways in which we can ask and answer such questions concerning web archives.

Before diving into specific questions regarding single-pixel GIFs, we contextualize this work in ongoing discussions about the future of access and use of digital collections. Cultural heritage institutions are increasingly exploring ways of thinking about enabling computational scholarship to think of their collections as data. Much of these conversations are about digitized collection materials, but we now have access to massive corpora of born-digital material, These born digital collections are functionaly born computable for digital scholarship.

Within that section, we briefly introduce computational scholarship and how approaching digital collections as data sets results in new kinds of research. We then provide examples of ongoing projects which focus on applying computational scholarship to web archives as a model of treating web archives collections as data to support new and evolving kinds of research.

Next, we present the findings of our efforts to trace the history of single-pixel GIFs as far back as the first instances appearing in the Internet Archive and Library of Congress Web Archives. Then, we share the findings of the use of computational scholarship, more specifically distant reading, on the UK Web Archive, headquartered at the British Library, to map the patterns of single-pixel GIFs over a 15-year period of web harvesting. Finally, using our methods as a case study, we discuss the findings of an approach based on tracing tiny files through terabytes of messy web archives data and the implications of these findings for researchers and digital library practitioners.

1 Situating web archives in trends in online collections

Without realizing it, humanists have been using computational methods to carry out their research for decades by using full-text search to explore electronic databases

(Underwood 2014) and, prior to this, with the advent of the computer, grappling with how to integrate computational analysis into historical inquiry, if at all (Anderson 2008). In other words, much of current scholarship is already computational, but many people are unaware of the role that computation plays in their research and discovery process. Over the course of the last twenty years, a more sophisticated approach to computational research has developed for humanists who are working with cultural heritage collections and imposing pattern and relevance algorithms directly onto the contents they are studying.

'Distant reading' has evolved into its own methodology of studying texts at scale (Jockers 2013), especially for text-based collections. Letting a computer 'read' hundreds of thousands of novels in seconds has significantly expanded the types of questions we can ask about collections, beyond keyword and word co-occurrence patterns. For example, text mining can identify linguistic patterns, highlight and map named entities (Finkel et al. 2005), compare authors' styles, create connected network graphs, and generate interrelated topics (Blei et al. 2003) over a collection or corpus. These methods have been applied to a collection of twenty thousand novels to predict trends in the literary world (Archer and Jockers 2016) and to thousands of articles from eighteenth-century (Newman and Block 2006) and nineteenth-century (Smith et al. 2013) newspapers to discover trends in news coverage and reprinting over time and geographic location.

The work has continued with specifically non-text-based collections. Scholars have used similar distantly-consumptive analytic methods on their recorded sound (Clement et al. 2016), image (Lorang et al. 2015), audio-visual, visual, and crowdsourced collections, whether the content in the collection began as digital items or had been digitized. Indeed, the expansion of these methods has itself resulted in the need for libraries, archives, and museums increasingly to rethink the modes of access they provide to collections. Computational scholarship is powered by corpus level engagement with works and artifacts as data.

The Library of Congress Collections as Data events and the related Always Already Computational initiative have stimulated conversation concerning access for digital collections and helped articulate visions for multi-modal access to digital collections (Mears 2017). The series brought together experts and practitioners creating digital collections and using digital collections in an effort to highlight common themes throughout the process. Major takeaways included a need for iterative processes with the goal of providing digital collections with better access, form, and quality (Padilla 2017).

To date, much of the work on broad access to digital collections has focused on digitized content. However, work on web archives is one significant exception. The Wayback Machine, the platform developed by the Internet Archive to provide access to web archives, has long been the primary means of entry to viewing web archives content. Alternatively, archives may use other, similar playback software, such as the community-driven open-source OpenWayback[1] or pywb,[2] a version of Wayback written in the programming language Python. It is important to note that the Wayback

[1] See the wiki for OpenWayback at https://github.com/iipc/openwayback/wiki
[2] See the documentation for pywb at https://pywb.readthedocs.io/en/latest/manual/apps.html#wayback-pywb.

Machine and other, similar efforts are not *the* archive. Rather, as software, the Wayback Machine, OpenWayback, and pywb provide windows onto the resources stored in any web archive.

With basic computer and internet literacy, one is able to navigate through archived web content on the Wayback Machine much like browsing the live web. However, as web archives have grown exponentially from gigabytes to petabytes, clicking through weekly captures of one section of one website gives users only a tiny fraction of the archive's content and even of that particular website over time. The sheer amount of web archive data now necessitates computational methods to detect patterns across the archived web and highlight areas of the archive in which to dig deeper.

In the autumn of 2016, the Library of Congress commissioned a pilot project simulating a potential researcher using LC web archives (Gallinger and Chudnov 2016). The LC web archiving team provided more than five terabytes of web archives content by means of a secure cloud platform to enable bulk use and analysis. The Web ARChive file format,[3] or WARC, is the standard aggregate file for harvested web content. It combines multiple resources as content blocks within each WARC, as well as associated metadata for each resource. WARC files are well suited for use in a playback mechanism like the Wayback Machine, but the structure and scale of these files is often challenging for researchers to work with directly.

Utilizing the cloud infrastructure and distributed computing provided by the third-party service, the contractors generated derivatives of the WARC files: Web Archive Transformation (WAT) files. WAT files are a slimmed version of WARC files which consist only of metadata for each resource contained in a WARC file, excluding the resource itself. This metadata includes the referring URI, the resource URI, MIME type, a timestamp of harvest, and the size of the resource. WAT files are a lightweight option for dealing with web archive resource metadata, taking up less than 20 % of the space of a WARC file.[4] For the pilot project, the contractors ultimately used the referring URIs and resource URIs to create link analysis visualizations in order to map how each website domain in the collection linked externally to other website domains.

Network analysis is a common way for researchers to explore web archives and for institutions practicing web archiving to begin understanding the breadth of their own collections[5] or perform quality review and completeness checks. This type of analysis over web archives provides a snapshot in time, i.e. a high-level view of a subset of the archive.

In order to arrive at a deeper understanding of researchers' needs, the British Library's UK Web Archiving Team hosted ten researchers on campus in 2014 under the Big UK Domain Data for the Arts and Humanities (BUDDAH) project. These researchers aimed to complete case studies while collaborating with the UK Web Archiving Team as a long term project. As a result, the case studies highlighted ways in which communication between the Web Archiving Team, project managers, and

[3] See the file format description at https://www.loc.gov/preservation/digital/formats/fdd/fdd000236.shtml.
[4] See the Internet Archive documentation at https://webarchive.jira.com/wiki/spaces/ARS/pages/90997503/WAT+Overview+and+Technical+Details.
[5] See the UK Web Archive Link Analysis visualization https://www.webarchive.org.uk/ukwa/visualisation/ukwa.ds.2/linkage and the ongoing Web Archives for Longitudinal Knowledge (WALK) Project by partners at the University of Waterloo, the University of Alberta, and York University http://webarchives.ca/ for more information.

researchers would be improved and more intuitive interfaces and datasets could be created for the researchers.[6]

To this end, there have been efforts to lower the barrier of entry to WARCs and analysis of web archives content. The Mellon-funded Archives Unleashed Toolkit (AUT),[7] which grew out of Warcbase (Lin et al. 2017), is currently the most robust system providing streamlined access to web archives data for researchers. AUT consists of web archives data loaded onto a high-performance computing platform, with data analysis interfaces at the ready. Similarly, Web Archiving Systems API, or WASAPI (Bailey and Taylor 2017), is an effort funded by the Institute of Museum and Library Services (IMLS), which seeks to map an interoperable API-based model for access to web archives data.

The existence and evolution of these efforts gesture toward a future in which we move increasingly away from one-at-a-time views of rendered web pages toward a model of treating web archives as digital corpora. It took tremendous effort to make something like the Google Ngram viewer to make sense of the noise in digitized texts. In contrast, libraries, archives and museums have billions of born-digital files in their web archives which, as born-digital objects, are born ready for computational scholarship.

Having provided this context and background, we return now to the questions raised at the beginning of this essay. Traces of the single-pixel GIF in web archives will offer some insights into the potentials of this mode of engaging with web archives.

2 Explorations in the history of the single-pixel GIF

What can we understand about the history of the single-pixel GIF when we begin by approaching web archives computationally? Part of the initial impulse to conduct this research was Lialina's online exhibit of single-pixel GIFs. If we take these hand-picked and curated examples of single-pixel GIFs as an initial source, we can begin to characterize them and, in turn, use that characterization to query web archives.

Lialina's exhibition links to a series of live manifestations of these images, presented in the list below. Of particular note, these are each specific locations on the web where one can find, or could once find, a copy of a spacer GIF. After the last forward slash in each of the URLs, we find the filename and extension. One of the exhibited works comes directly from Siegel's site (killersites.com), but in each of them, even just at the filename level, we can see the different names these files take on:

http://www.geocities.com/clipart/pbi/c.gif
http://pic.geocities.com/images/pixel.gif
http://www.google.com/clear.gif
http://killersites.com/killerSites/resources/dot_clear.gif
http://visit.geocities.yahoo.com/visit.gif
http://blingee.com/images/spaceball.gif
http://www-cdr.stanford.edu/~petrie/blank.gif
http://img.artlebedev.ru/;-)/n.gif
https://mail.google.com/mail/images/cleardot.gif
http://www.google.com/images/cleardot.gif

[6] For final reports from the BUDDAH project, see the blog https://buddah.projects.history.ac.uk/2016/04/.
[7] http://archivesunleashed.org/about-project/.

2.1 Characterizing/identifying files

Below we have characterized each of the files using two methods. First, by querying their instances on the Wayback Machine, we have identified the earliest date for which the Internet Archive and the Library of Congress have captures of each respective resource in the specified location. Second, we have computed a SHA-1 cryptographic hash for each file. A cryptographic hash function is an algorithm which takes a given set of data (such as a file) and computes a sequence of characters which can then serve as a unique identifier for that data. Even changing a single bit in a file will result in a different sequence of characters. For a sense of just how high that confidence can be, it is worth noting that a cryptographic hash offers more confidence as a characterizer of individualization than a DNA test does for uniquely identifying a person (Kruse II and Heiser 2001, p. 89).

URL	Earliest LC	Earliest IA	SHA-1	Match
http://www.geocities.com/clipart/pbi/c.gif	2/8/02	10/13/99	356F32DA60A0387E36ED94 B0CB3D0A0394D90B60	
http://pic.geocities.com/images/pixel.gif	2/23/02	3/2/00	328E472721A93345801E D5533240EAC2D1F8498C	1
http://www.google.com/images/cleardot.gif	7/2/02	5/10/00	56D45F8A17F5078A20A F9962C992CA4678450765	2
http://www.google.com/clear.gif	2/22/02	8/5/00	317496A096D6C86486A71 D4521994BCD171A6BB3	
http://killersites.com/killerSites/resources/dot_clear.gif	8/5/09	6/20/03	328E472721A93345801E D5533240EAC2D1F8498C	1
https://mail.google.com/mail/images/cleardot.gif	1/28/08	4/5/06	56D45F8A17F5078A20A F9962C992CA4678450765	2
http://visit.geocities.yahoo.com/visit.gif	none	7/3/06	FAA81452F0C19B304B89 F0086F85A2941A57C32D	
http://blingee.com/images/spaceball.gif	10/3/07	1/18/07	2DAEAA8B5F19F0BC209 D976C02BD6ACB51B00B0A	3
http://www-cdr.stanford.edu/~petrie/blank.gif	6/16/09	6/12/07	9D01CC5DC8E042C0D4AD6 CFB8B3AC38E84A5EF9F	
http://img.artlebedev.ru/;-)/n.gif	none	12/5/13	2DAEAA8B5F19F0BC209 D976C02BD6ACB51B00B0A	3

Of these, the earliest recorded capture of any of the single-pixel GIFs is the Geocities Clipart link. With that noted, this only tells us when that file was acquired by respective institutions, not necessarily when it was created. This is a recurring pattern which we will encounter as we work through our analysis. A central challenge in interpreting the contents of web archives is retaining a certain level of skepticism: to what extent are any research findings mapping trends in web history, versus trends in how the web was collected? This is a topic, we futher explore later.

Significantly, by hashing the files, we have found seven distinct files out of the original ten. The chart above is coded to show three sets of duplicate files (coded '1,' '2,' and '3' in the 'Match' column) and four unique files. The files within each duplicate set are bit-for-bit identical (i.e. the file coded with '1' is identical to the other

file coded with '1'). In most cases where this occurred, one could deduce that the files with identical hash values are themselves historically related. In other words, one file is likely a later, identical copy of the original. However, in this unique case, given the miniscule file size, we cannot assume any interrelation of identical files. A tiny transparent image file does not lend much to the original maker's unique creativity, and it is possible that several users created identical files using identical processes.

2.2 Single-pixel GIF trends across corpora

Given that we have distinct, digital fingerprints for each of these single-pixel GIFs in the form of their SHA-1 hash values, it becomes possible to query an entire corpus of a web archive to determine where and when files with the same hash value were collected.

To date, the UK Web Archiving program remains unique in that it stores a copy of all the content it has collected in a high-performance distributed computing system. As a result, it is possible to run queries across the entirety of the content of their web archive. Andrew Jackson, the technical leader for the UK Web Archives, generously scanned the UK Web Archive for appearances of these seven hash values. Jackson then published the scripts and data resulting from this query (Jackson 2015).

The charts below display the number of times each of the seven distinct single-pixel GIFs from the Geocities data set appeared in the UK Web Archive collections over time. The first initial pass at the findings shows that there are three extant examples of GIFs in the archive dating from 1996: two instances of blank.gif, three instances of pixel.gif, and 46 instances of spaceball.gif. Hence, we can conclude that spaceball.gif was the earliest widely used or at least widely collected example of single-pixel GIFs. This year is significantly earlier than the first instance of each GIF from the Geocities data set previously discussed (Fig. 2).

Each of the seven unique GIFs studied here existed in the UK Web Archive by 1997. Yet, as the charts show, they made their way across the web and through time in strikingly varied ways. Cleardot.gif (a category documented in two distinct, original Google URLs) emerges as the most widely collected GIF out of the seven. In 2008, the British Library collected and documented the presence of more than one million copies of cleardot.gif (1,062,943 copies). This collection results in a fascinating spike, while the other six GIFs nearly vanish from the archive after having had a large presence in 2006 and 2007. Clear.gif had the earliest significant spike in 1998, and the usage of dot_clear.gif/pixel.gif shot up to nearly 200,000 entries (combined total) in 2004. Blank.gif resurfaced in 2010 and all seven GIFs have low representation in 2009. To begin understanding the trends of single-pixel GIFs over time, it is important to consider whether the GIFs themselves had distinct histories and to examine the details of those histories, separately from collection practices.

Exploration of the histories of each of these individual files through independent searching reveals the varied ways in which these files have been developed and used. As a post by Martin Brinkmann from 2007 documents, spaceball.gif was used by Flickr, the community-driven website launched in 2004 hosting photographs and images, to prohibit easy download of the image files by individuals or crawlers. When a user would attempt to right click and download an image file, they would instead be

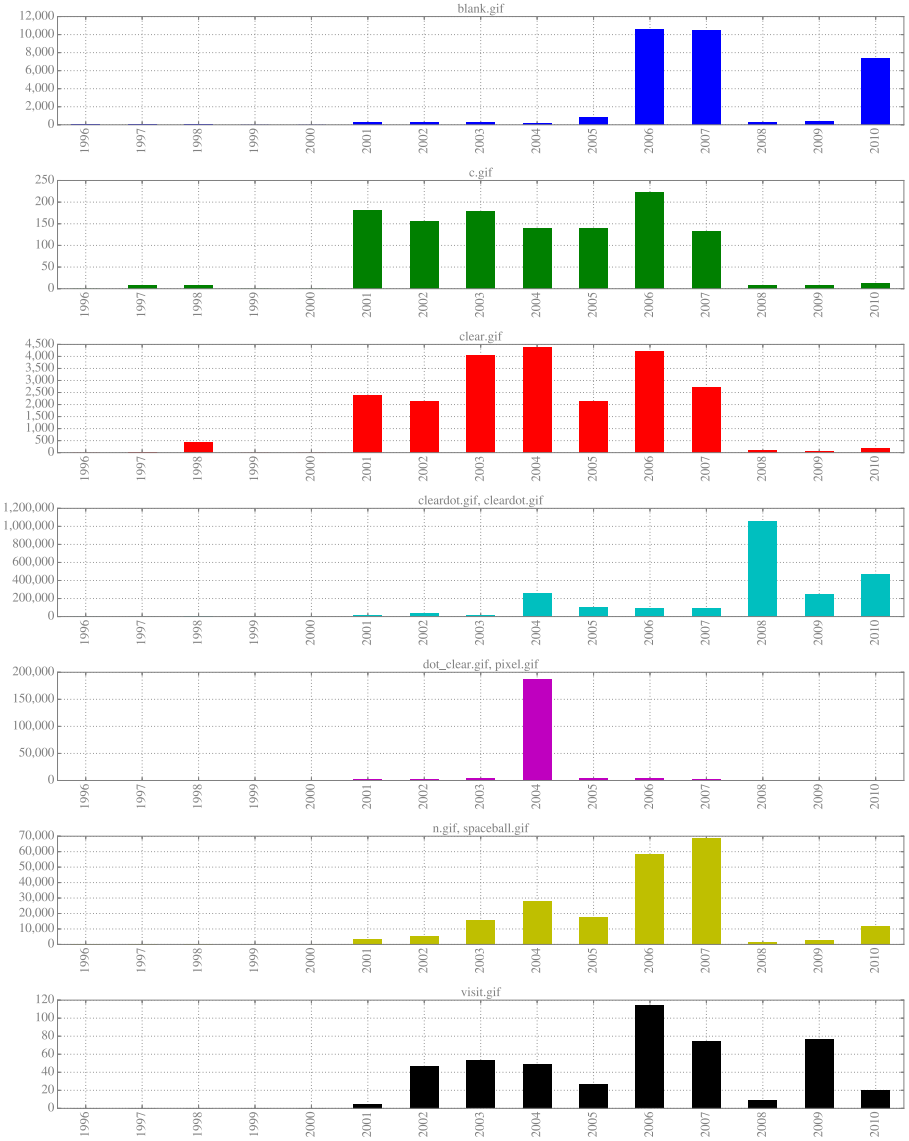

Fig. 2 Appearances of the seven distinct GIFs in the UK Web Archive from 1996 to 2010

tricked into downloading a tiny, transparent GIF which had been invisibly masking the underlying displayed image (Brinkmann 2007).

Similarly, cleardot.gif (much like spaceball.gif) appears to serve a distinctly different purpose from a spacer GIF solely used for formatting. Often referred to as 'web beacons' or 'web bugs,' these files are widely known to be used as a means of surveillance and tracking. Specifically, their tiny size and invisibility means that they load quickly, without being detected. Each time one of these files loads, it results in a ping back to the source. Indeed, the URL https://mail.google.com/mail/u/0/images/cleardot.gif is an example of this (pabouk 2013). Critiques of these methods go back to at least late 1999, when sites for companies including

Quicken, FedEx, Metamucil, Oil of Olay, and StatMarket were identified as using this technique (Smith 1999).

These histories present interesting and challenging issues, admittedly beyond the scope of the current study: given the range of functions of single-pixel, transparent GIFs, how are we to understand their presence in different locations over time? To what extent can we take the presence of a single-pixel transparent GIF as serving a formatting function when the same file has been used for other purposes, such as blocking the download of other image files? Using the data, is it possible to identify which uses of the single-pixel, transparent GIF predate other uses? If we were to zero in on that early year, we might well be able to pinpoint the URL that each of these images first appeared at in the archive and the day they first appeared, which would constitute a possible next step for this kind of study.

3 Discussion: what invisible files let us see

There are millions of copies of single-pixel, transparent GIFs in the world's web archives. Each one is a trace of a practice and method of presenting information on the web. Some are traces of changes in web design. Some are traces of methods of surveillance. By working back and forth between the URLs for these tiny, functionally invisible images and their hash values, we have begun to map some of this history. The findings of this preliminary mapping offer a range of considerations for the future of access and use of web archives and the history of the web. They suggest requirements for a better understanding of crawling and collecting practices, new methods for characterizing and indexing files, and issues for the interpretation of born-digital collection data.

3.1 Seeing web history or web archiving history?

A web crawler whose job is to archive particular websites makes appraisal decisions in a different way than a human archivist processing a donated collection. Both processes include having all documents in front of the archivist and the crawler, and both must decide which to keep and which to pass over. However, all of the rules for a crawler must be set before the crawl starts. It is possible to change the crawler behavior during the crawl, but this change takes a significant amount of effort and ongoing quality review. To avoiding crawling the entire Internet every time, the rules tell a crawler what to archive and what to avoid. Restricted areas can include entire domains or a regular expression for all URLs with the string 'login,' for example.

For this study, it is possible that any dramatic drop in GIF appearances, such as in 1999 and 2000, could reflect the choice of a web archivist to exclude single-pixel, transparent GIFs from the crawl entirely. This decision may have been made for any number of reasons, including space constraints or a simple belief that single-pixel, transparent GIFs were unnecessary to store in the archival record. It is also possible that the program stopped archiving a site or many sites which contained a large number of these single-pixel, transparent GIFs. Collateral content, or superfluous content the crawler ends up harvesting during a crawl, is unavoidable given the nature of the web. If most of the single-pixel GIFs were crawled as collateral content, the exclusion of certain websites may have caused a reduction in their appearances.

3.2 Approaching web archives as data corpora

It is imperative for libraries and archives to consider the end *data* utilized by researchers in the future when building digital *collections* in the present. Part of this practice requires web archivists to create scope and content notes and keep records of crawl decisions as they are made and as crawls are performed. Content processing done by the web archivists to understand their own collections as data can help with this. If an archivist saw these dramatic drops in appearances of single-pixel, transparent GIFs as a result of crawling practices, the archivist could file the information and share it with a researcher attempting to understand the collection in the future.

This study looks at transparent GIFs appearing in two specific collections, Olia Lialina's exhibit of transparent GIFs from the Geocities archive and the UK Web Archive. These two collections make up a small percentage of content in web archives throughout the world, web archives which have had varying crawl practices over time (Milliganet al. 2016). We took a look at the history of seven transparent GIFs in data resulting from harvesting done by the UK Web Archiving Team. We have not looked at the complete history of all single-pixel GIFs as they appeared on the live web over time (Brügger 2017).

With appropriate technical infrastructure, this same study could be completed on any organization's web archives. Since each one of these entities will have different crawl practices, multiple web archiving initiatives collecting the same websites is invaluable to researchers studying the web. As the crawl becomes more comprehensive, we can begin to see how the findings of case studies like these are influenced by crawling practices (crawl frequency, crawl depth, deduplication, etc.) and whether the findings are indicative of web usage trends throughout time. Decoupling these concepts is essential for an understanding of the practice of web archiving and the history of the web, respectively, and can only be done through multiple archives.

When we approach each institution's web archives as corpora it becomes increasingly clear that there is significant value in having a range of organizations engaged in web archiving Ideally, they are engaging in these practices with a range of tools. The trends in the appearance of these files raise all kinds of questions. For instance, what conclusions do we reach when we apply similar methods to different kinds of files? In other words, what do trends in identical copies of files themselves tell about the movement, dissemination, and popularity of practices and approaches? There is informational content in the files, but the history of the appearance of a given file in a given place also has potential informational value.

3.3 Characterizing files as key to future modes of access

Knowing the specific URLs at which files exist is also invaluable to the study of web history. The case of single-pixel GIFs illustrates the significant value of modes of characterizing and identifying files using other methods. The ability to hash a file and use that digital fingerprint to see where else it, or files created through identical processes, exists in web archives is immensely powerful. Who would have imagined there were millions of copies of one of these tiny files captured in the UK Web Archive in one particular year? When we discover that two URLs held identical files at a particular date, we can start to track and trace the replication and movement of

information. Importantly, this is all derivative information about the content. Even in a situation in which archives can't offer global access to the content itself, non-consumptive hashes could very well be provided for this kind of work.

While hashes are exciting, it is important to remember that there are many other ways of characterizing similarity. An alternative approach to this kind of research could involve simply identifying all the '.gif' files in a web archive that are particularly small and visually inspecting them to identify potential other candidates for different, unique single-pixel GIFs. When one moves further into hash-based approaches to the study of files, it will be critical to remember that minor changes in a file are going to give it a new hash. With that noted, this only further points to the need to root the future of the study of web archives in the ability to compute against the files in these corpora.

3.4 Implications for digital library infrastructure

Access issues highlighted in the computational scholarship are a sobering reminder that 'digital' or 'digitized' doesn't not necessarily mean immediately ready for computational scholarship. Different kinds of questions require data to be prepared, processed, and made accessible in a number of ways. While digital material, rather than analog, is one step closer to becoming data, there is still work to be done to strategically arrange the content for a future of computational scholarship. Furthermore, there are specific necessary affordances in technical architecture in order to enable researchers to compute against a corpus.

As the Library of Congress pilot project showed, cracking open complex WARC files to perform high-level analyses of the archive takes computing power that many researchers, and even institutions, do not always have at their disposal. The present study was, in large part, possible because a copy of the UK Web Archive is maintained and managed on a high-performance distributed computer system and because its archivist was willing to field a request to search across this web archive corpus to answer this particular question. Most web archives are not currently configured in a manner which enables researchers to compute against their content as a corpus.

In order for this kind of research to become more of a reality, library institutions will first have to explore having compute-on-demand capabilities for their entire corpus of web archives and, more broadly, other large, born-digital and digitized collections. This has significant implications for the future of infrastructure. It largely requires either establishing local high-performance computing environments or a shift to approaching access systems that rely on cloud computing environments for access copies of content. Models that involve caching portions of content and working across multiple levels of tiered storage media simply will not be able to facilitate this kind of data corpus use of querying collections.

4 Conclusion: researchers and web archivists embracing distant Reading

The single-pixel, transparent GIF seems to exemplify the essence of insignificance. The files are tiny and invisible. However, the history of these files reveals a great deal about the history of web design, tracking, and surveillance. Sometimes they are spacer GIFs,

sometimes they are web bugs, and sometimes they are web beacons. While we have not offered conclusive answers to any of the questions about their history, we have explored single-pixel, transparent GIFs as a case study to shed light on future methods of studying the history of the web through born-digital web archives collections.

The future of the study of the web and the future of collecting the web are intertwined. When we step back and see the patterns that emerge by looking at the hashes of a small set of files in the UK Web Archive, we immediately are prompted to raise two questions: What does this tell us about the history of the web? What does this tell us about the history of web archiving practices? Researchers, now and in the future, will want to approach web archives collections by pivoting between distant reading and close reading. The pairing of distant and close reading as a method of studying the archived web is the only way of conceptualizing the sheer scale of the archived web and performing meaningful research.

However, these methods will also help iteratively to build better, more comprehensive, and more curated web archives throughout the world. The scale of a web archive is also a challenge for the archivists charged with curating and maintaining it. Yet, the same tools used by researchers can be used by web archivists and practitioners in the field to understand their archives or, sometimes more importantly, what is missing from their archives. As practitioners come to understand their archives in greater detail, this knowledge will inform future preservation practices and will provide immediate assistance in provenance for researchers utilizing the data.

Since the scale of web archives does not lend itself to traditional page-through reading and distant reading will become a necessity of close reading, the burden is on digital librarians to rethink the nature and structure of digital libraries, digital content, and web archives infrastructure. This could mean putting more resources into development of tools outside of web page rendering mechanisms, such as streamlined creation and delivery of data sets or web archives content derivatives. Overall, detailed collection notes, especially crawling, scoping, and other specific decisions made over time, are crucial to improving the system and furthering research.

References

Anderson, I. (2008). History and computing. *Making History*. Retrieved from http://www.history.ac.uk/makinghistory/resources/articles/history_and_computing.html.

Archer, J., & Jockers, M. L. (2016). *The bestseller code: Anatomy of the blockbuster novel*. New York: St. Martin's Press.

Bailey, J., & Taylor, N. (2017). *Web Archiving Systems APIs (WASAPI) for systems interoperability and collaborative technical development*. Paper presented at the CNI Fall 2017, Washington DC, US.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*(Jan), 993–1022.

Brinkmann, M. (2007). How to avoid saving spaceball.gif at Flickr. *gHacks Tech News*. Retrieved April 25, 2018 from https://www.ghacks.net/2007/09/29/how-to-avoid-saving-spaceballgif-at-flickr/.

Brügger, N. (2017). The archived website and website philology. *Nordicom Review, 29*(2), 155–175. https://doi.org/10.1515/nor-2017-0183.

Clement, T. E., Auvil, L., & Tcheng, D. (2016). *High performance sound technologies for access and scholarship*. Retrieved from http://hdl.handle.net/2152/33295.

Finkel, J. R., Grenager, T., & Manning, C. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *ACL-05 - 43rd Annual Meeting of the Association for*

Computational Linguistics, Proceedings of the Conference (pp. 363–370). Michigan: AnnArbor. https://doi.org/10.3115/1219840.1219885.

Gallinger, M., & Chudnov, D. (2016). *Library of Congress lab: Library of Congress digital scholars lab pilot project report*. Washington, DC: The Library of Congress Retrieved from http://digitalpreservation.gov/meetings/dcs16/DChudnov-MGallinger_LCLabReport.pdf.

Jackson, A. (2015). Tracing clear.gif: Jupyter Notebook. UK Web Archive Github Repository. https://nbviewer.jupyter.org/github/ukwa/halflife/blob/master/clear/tracingclear.gif.ipynb.

Jockers, M. L. (2013). *Macroanalysis: Digital methods and literary history*. Urbana: University of Illinois Press.

Johnson, P. (2011). Digital folklore with Olia Lialina & Dragan Espenschied: The transcript. Retrieved from http://artfcity.com/2011/05/13/digital-folklore-with-olia-lialina-dragan-espenschied-the-transcript/.

Kruse, W. G., II, & Heiser, J. G. (2001). *Computer forensics: Incident response essentials*. Boston: Addison–Wesley Professional.

Lialina, O**.** (2013). Olia's collection of clear/blanc/0/transparent/cover/beacon GIFs. Retrieved from http://www.collection.evan-roth.com/olia_lialina/clear.gif/.

Lin, J., Milligan, I., Wiebe, J., & Zhou, A. (2017). Warcbase: Scalable analytics infrastructure for exploring web archives. *Journal on Computing and Cultural Heritage (JOCCH), 10*(4), 22.

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. *D-Lib Magazine, 21*(7/8). https://doi.org/10.1045/july2015-lorang.

Mears, J. (2017). Read collections as data report summary. Retrieved April 25, 2018 from https://blogs.loc.gov/thesignal/2017/02/read-collections-as-data-report-summary/.

Milligan, I., Ruest, N., & Lin, J. (2016). Content Selection and Curation for Web Archiving: The Gatekeepers vs. The Masses. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries* (pp. 107–110). New York, NY, USA: ACM. https://doi.org/10.1145/2910896.2910913

Newman, D. J., & Block, S. (2006). Probabilistic topic decomposition of an eighteenth-century American newspaper. *Journal of the Association for Information Science and Technology, 57*(6), 753–767.

Owens, T. (2015). *Designing online communities: How designers, developers, community managers, and software structure discourse and knowledge production on the Web*. New York: Peter Lang.

pabouk. (2013). How does Google's cleardot.gif track email recipients with a generic URL? *Super User*. Retrieved April 25, 2018 from https://superuser.com/questions/658098/how-does-googles-cleardot-gif-track-email-recipients-with-a-generic-url.

Padilla, T. (2017). On a collections as data imperative. Retrieved April 25, 2018 from http://digitalpreservation.gov/meetings/dcs16/tpadilla_OnaCollectionsasDataImperative_final.pdf.

Rønn-Jensen, J. (2006). Who invented the spacer.gif? Retrieved from http://justaddwater.dk/2006/03/03/who-invented-the-spacergif/.

Rønn-Jensen, J. (2007). Who invented the spacer.gif (Part 2). Retrieved from http://justaddwater.dk/2007/02/11/who-invented-the-spacergif-part-2/.

Siegel, D. (1997). The Web is ruined and I ruined it. *XML.Com*. Retrieved from https://www.xml.com/pub/a/w3j/s1.people.html.

Smith, R. M. (1999). The Web Bug FAQ. Retrieved April 25, 2018 from https://w2.eff.org/Privacy/Marketing/web_bug.html.

Smith, D. A.**,** Cordell, R., & Dillon, E. M. (2013). Infectious texts: Modeling text reuse in nineteenth-century newspapers. In *Big Data, 2013 IEEE International Conference on* (pp. 86–94). IEEE.

Underwood, T. (2014). Theorizing research practices we forgot to theorize twenty years ago. *Representations, 127*(1), 64–72. https://doi.org/10.1525/rep.2014.127.1.64.