Archival Science (2021) 21:43–57
https://doi.org/10.1007/s10502-020-09345-w

ORIGINAL PAPER

The Wayback Machine: notes on a re-enchantment

Surya Bowyer[1]

Published online: 8 July 2020

Abstract
The Internet Archive's Wayback Machine holds over 424 billion webpages, making it the largest publicly accessible archive in the world. Thus far, much of the research on the Machine has approached the technology using computational thinking. This type of thinking treats technology operationally, as something that we can use to do jobs for us. This article takes a different approach. It steps back from computational thinking to consider the language we use to apprehend technology. It argues that the metaphors we use actually obfuscate, rather than merely describe, the operations of the Machine. By making explicit the workings of these metaphors, the article draws attention to, and thus counteracts, this obfuscation. In so doing, these notes on the Wayback Machine point more widely towards the usefulness of a language-oriented approach to other technologies.

Keywords Internet Archive · Michel Foucault · Web crawler · Photography · Death · James Bridle

You are exploring Donald Trump's website, http://donaldjtrump.com. (Let us imagine, for a moment.) You read the quotes, you look at the images, you click on hyperlink after hyperlink. But then you take a wrong turn. The page you were hoping to find is not there; instead, you are met with a page with the words "Oops! This is awkward. You're looking for something that doesn't exist…" (Wayback Machine 2019). Next to the words, a photograph. Hillary Clinton stands at a podium emblazoned with the presidential seal. Three numbers are inscribed on the browser tab: 404.

404 error pages signal that a URL cannot be found. They often occur when a hyperlink is outdated, or when one mistypes a URL. They are perhaps the most commonly encountered error pages on the internet, and are an unavoidable part of the web as it has been constructed. When flicking through an article's bibliography, there is always a risk that a citation's URL will return a 404 page. The problem

✉ Surya Bowyer
sbowyer@qebarnet.co.uk

[1] Queen's Library & Collections, Queen Elizabeth's School, Barnet, London, UK

persists across all disciplines (Kumar et al. 2015; Russell and Kane 2008; Evangelou et al. 2005; Hester et al. 2004; Crichlow and Winbush 2004; Dellavalle et al. 2003). The older a citation URL is, the more likely it is not to work (Sampath Kumar and Prithviraj 2015). The problem is likely to only get worse as the proportion of citations to online sources increases.

In 1996, the non-profit company The Internet Archive began trying to solve the problem of 404s by archiving webpages. Since then, the company has begun archiving other media, currently holding over 20 million books, 4.5 million audio recordings, 4 million videos, and 3 million images. But the web archive is by far its most extensive collection. Named the Wayback Machine, it now holds over 424 billion webpages, making it the largest publicly accessible archive in the world (Wayback Machine 2020a).

The Wayback Machine recalls a passage in Michel Foucault's "Of Other Spaces:"

> the idea of accumulating everything, of establishing a sort of general archive, the will to enclose in one place all times, all epochs, all forms, all tastes, the idea of constituting a place of all times that is itself outside of time and inaccessible to its ravages, the project of organizing in this way a sort of perpetual and indefinite accumulation of time in an immobile place, this whole idea belongs to our modernity (Foucault and Miskowiec 1986, p. 26).

The Wayback Machine cannot actually accumulate everything. But even if the reality of absolute accumulation remains untenable, the idea appears to be the Machine's guiding principle. Unlike national web archives, which attempt to archive all websites from a certain nation's top-level domain, the Wayback Machine ignores borders. It is a "general archive" in the sense that there appears no overriding theme guiding its acquisitions, and indeed any user can choose to add a webpage, at a particular time, to its holdings. Foucault locates the episteme of absolute accumulation in the "western culture of the nineteenth century" (Foucault and Miskowiec 1986, p. 26). Yet the genealogy of the episteme stretches back at least as far as the mid-sixteenth century, when the Swiss scholar Conrad Gessner created the Bibliotheca Universalis, an alphabetical bibliography that attempted to list all known books printed in Greek, Latin, and Hebrew. Stretching the other way, the Wayback Machine suggests that the episteme remains relevant in the present.

A note on terminology: various other web archives use an open-source version of the Wayback Machine software that is now referred to as the Open Source Wayback Machine or OpenWayback. The two versions of the software, the Internet Archive's Wayback Machine and OpenWayback, remain separate—there is cross-pollination but not total feature parity between the original Machine and other open-source implementations. From a technical viewpoint, the Wayback Machine software is not an archive in itself, but rather an access mechanism. However, in practice the Internet Archive refers to websites being archived *into* the Wayback Machine. Given this, and to avoid confusion, I use the Internet Archive to refer to the non-profit organisation, and the Wayback Machine to refer to the Internet Archive's web archive.

Thus far, much of the work on the Wayback Machine has taken one of two forms. The first is a computer science approach, assessing the Machine's technological problems and/or suggesting possible improvements (see for example: Kanhabua et al. 2016; Sampath Kumar and Prithviraj 2015; Al Noamany et al. 2014). The second is an investigation into how the Machine will affect research methodologies in the discipline of history (see for example: Rogers 2017; Belovari 2017; Milligan 2016; Kaur 2015). Both these strands have produced valuable and varied research. Nonetheless, computational thinking underlies both approaches. This type of thinking treats technology operationally, as something that we can use to do jobs for us. We may encounter problems, such as the Wayback Machine lacking a comprehensively indexed keyword search function (Kanhabua et al. 2016), or the question of how to present a website's history (Rogers 2017), or indeed simply the insurmountable amount of archived information in the Machine (Milligan 2016). Computational thinking assures us that these problems have operational solutions: that to solve them we must manipulate the way we use our technology. In fact, the problems themselves derive from computational thinking, in that they are concerned primarily with technology in an operational sense. This article takes a different approach. It steps back from computational thinking to consider the language we use to apprehend technology. In the first section, "The making of metaphors," I outline my methodological approach, which takes its cues from James Bridle's proposal that we re-enchant our tools through thoughtfulness. In the second, "The wild frontier," I explore the importance of spatial metaphors to our understanding of how the Wayback Machine archives the web. In the third, "We define," I show how our understanding of the Machine is reliant on a photographic metaphor, and then I explore problems with the Internet Archive's definition of its holdings. In the fourth section, "Words, words, words," I explore the metaphoric and literal importance of death to the Machine, and I ask what happens when we try to write the history of the web. Throughout, I show how the metaphors we use actually obfuscate the operations of the Machine. By making explicit the workings of these metaphors, I hope to draw attention to, and thus counteract, this obfuscation. In so doing, it is my hope that these notes on the Wayback Machine point more widely towards the usefulness of a language-oriented approach to other technologies.

The making of metaphors

> Let us not begin at the beginning, nor even at the archive. But rather at the word "archive" — and with the archive of so familiar a word (Derrida 1996, p. 1).

> By this term [archive] I do not mean the sum of all the texts that a culture has kept upon its person as documents attesting to its own past, or as evidence of a continuing identity; nor do I mean the institutions, which, in a given society, make it possible to record and preserve those discourses that one wishes to remember and keep in circulation (Foucault 2002, p. 145).

The "archive" of the archival turn was predominantly a conceptual one; or, as Carolyn Steedman puts it, "an idea rather than a place" (Steedman 2011, p. 321). One of the notable shifts in more recent work is a (re)turn to approaching archives as physical entities located in space. Steedman herself has focused on dust, that irksome guest in archive buildings (2002). In his account of the birth of the archive in the early modern period, Markus Friedrich argues more widely that the "material and physical dimensions" of an archive are key because the "social significance" of an archive "is constituted precisely by means of physical and spatial interaction" (2018, p. 15). Belinda Battley has argued that the importance of place extends deeper than archives themselves, to the individual records within them (2019). Indeed, the place of records is central to the question recently posed by James Lowry of what to do with displaced archives, in that displacement necessitates an archive first having a place from which to be displaced (2019).

The internet relies on places: "a physical infrastructure consisting," in the words of James Bridle, "of phone lines, fibre optics, satellites, cables on the ocean floor, and vast warehouses filled with computers, which consume huge amounts of water and energy and reside within national and legal jurisdictions" (2018, p. 7). But it is not reducible to any single one of these places, and nor could the internet (as we interact with it) be said to be located at these places.

Bridle notes that the cloud has become "the central metaphor of the internet." In his estimation "it is a very bad metaphor" because of the amount of physical infrastructure the internet is predicated on. Yet Bridle speaks of clouds as "noumenal […] numinous […] almost impossible to grasp" (2018, p. 7), and this surely sums up most users' experiences with the internet. We seldom think of the wires, the warehouses, the servers. Indeed, Bridle's "almost impossible" leaves a space for the possible: while we do not usually experience the internet's physicality, we cannot say that this physicality does not exist. We generally experience both terms of the metaphor as non-physical entities, but this is not to say that they are non-physical entities. Clouds may not be easy to visit, but this does not mean we cannot visit them; much the same can be said for the internet, at least insofar as its physical infrastructure is concerned.

Regardless of whether or not the cloud is a bad metaphor, it is the central metaphor. For Bridle, metaphors help us make sense of new technology. He argues that technology is not simply the making and use of tools, but also "the making of metaphors" (2018, p. 13). By creating a tool or technology, we make concrete a certain understanding of the world. However, we then dissociate this understanding from our agency, offloading it into the technology itself, so that the understanding "no longer needs thinking to activate." Bridle urges that: "To think again or anew, we need to re-enchant our tools […] such a re-enchantment, [is] an attempt to rethink our tools — not a repurposing or a redefinition, necessarily, but a thoughtfulness of them" (2018, p. 13).

What might "a thoughtfulness" look like, in practice? How exactly do we go about re-enchanting our tools? By turning to Michel Foucault's understanding of language as constitutive of our reality, I would like to suggest one method of re-enchantment. Foucault speaks of:

> A task that consists of not — of no longer — treating discourses as groups of signs (signifying elements referring to contents or representations) but as

> practices that systematically form the objects of which they speak. Of course, discourses are composed of signs; but what they do is more than use these signs to designate things. It is this *more* that renders them irreducible to the language *(langue)* and to speech. It is this "more" that we must reveal and describe (2002, p. 54).

In Bridle's words I hear an echo of Foucault. For Foucault, language forms the object of which it speaks; for Bridle, our objects are saturated with our own ideas and understandings. If we take ideas to be predicated on language—or, if this is a controversial view, if we take ideas to be inextricably tied up with language—the parallel becomes clear. For both Foucault and Bridle, language is not merely descriptive but also productive. This is what gives it the "more" of which Foucault speaks.

In the remaining sections of this article, I turn my focus squarely to the Wayback Machine. In the metaphors we use to apprehend it, we have offloaded a certain way of thinking. By not taking these metaphors for granted, I hope to make clear the ways in which they obfuscate, rather than simply describe, the workings of the Machine. In this way, thoughtfulness can show the ways that our technology exceeds the language we use to apprehend it. It is in acknowledging this gap between language and tool that we re-enchant the latter.

The wild frontier

> The present epoch will perhaps be above all the epoch of space (Foucault and Miskowiec 1986, p. 22).

The Internet Archive uses web crawlers to gather archival material for the Wayback Machine. Historically, much of their material has come from the for-profit company Alexa Internet, which passes information collected by its crawler to the Internet Archive. But for at least the last decade, the Internet Archive has also deployed its own crawler, named Heritrix. Like the Machine, this piece of technology has been made available open-source (see Mohr et al. 2004). The Internet Archive also accepts crawl data from other donors. A crawler begins with a webpage, and then follows each hyperlink on that webpage to reach new webpages. On each of the new webpages, the crawler repeats the process. The image is thus that of a web of "nodes and links," to borrow James Gleick's terminology (2012, p. 423). Each webpage leads to multiple other webpages, in theory ad infinitum. Crawlers, also referred to as spiders, travel along this web.

At least two things are notable about this terminology. One is the zoomorphism: this is not simply a piece of software, it is a spider or crawler. The other is movement, made clear of course in the name—crawler—but also in our understanding of what the software is doing (crawling from one page to another). The metaphor is that of movement through space.

We do not think of the internet as reducible to the places on which it relies: the warehouses full of servers, with whirring fans keeping everything cool and operational. Yet space is the key metaphor in our terminology, not just for crawlers but for the web and internet more widely. A collection of pages is called a site. A page that is designed to attract and direct visitors to a certain website is referred to as a gateway. When a human acts akin to a crawler, going from one page to another using hyperlinks, we say they are surfing the web. "On the Nature of the Internet," a report by the Global Commission on Internet Governance, refers to "boundaries" not only in the literal "geographic sense" of the word, but also "in a network sense" (Daigle 2016, p. 18). We speak of the internet spatially. There is room for an analysis of the internet which examines space and place, and how they might be different—perhaps along Lefebvrian lines. However, in this article I use place in the sense of location/site, and space in the sense of the physical relationship between places.

How does one deal with such an unruly, ever-changing, ever-growing space? As soon as the spider has moved along the web, leaving one page for another, the previous page is liable to change. John McDonald's description of the difficulties of record-keeping in a modern office environment reads almost as well in relation to the internet more widely:

> From a record-keeping perspective, the modern office is like the wild frontier. Office workers can create and send electronic messages and documents to whomever they wish. They can store them according to their own individual needs and then delete them without turning to anyone else for approval. There are no rules of the road. The autonomy of the individual reigns supreme! (1995, p. 70)

McDonald's words would fit even better if we substituted "store" with "publish." On this point, Jeffrey Macintyre (2012) has argued that it is beneficial to view the internet as first and foremost a publishing medium. There is also another link. When a crawler has identified all the hyperlinks on a given page, it adds these URLs to a list. The crawler uses the list to keep track of which URLs it is yet to visit. This list is referred to as the crawl frontier. A distinction is thus demarcated: on the one side, the land already settled; on the other, the as-yet unexplored wilderness that will soon be settled by the crawler as it roams across the web. In this way, something as seemingly mundane as a list of URLs is, by the language we use to refer to it, made terrestrial and territorial. Discourse constitutes it as an uncharted wilderness ripe for exploration. The logic is that of the cartographer, or colonialist.

The crawling process is therefore figured in the following terms: the crawler crawls along the web, discovering sites, delineating a constantly expanding frontier of unexplored territory which guides its subsequent movements. This network of metaphors—crawler; crawl; web; frontier—produces a sense of agency in the crawler. My own account in the above paragraphs does the same. But do crawlers have agency? No. Every crawl is an automated process that is reliant, directly or indirectly, on the agency of another: a human. By referring to the crawler *as* a crawler—and thereby instilling it with agency—we erase from view the human agency involved in the crawling process. This erasure becomes particularly pertinent (and troublesome) given the Wayback Machine does not—and thus far cannot—archive the internet in its entirety. Humans must make conscious decisions about what is crawled, shaping algorithms accordingly. The language with which we apprehend the crawling process offloads this agency onto the crawler itself, and in so doing it eradicates the human.

The Internet Archive has recently implemented measures to re-introduce the human. A new feature titled “Collections” allows users “to learn,” in the words of an Internet Archive blog post, “about why a given URL has been archived into the Wayback Machine” (Wayback Machine 2020d). The adverb is key. When a user is picking which dated snapshot of a URL to look at, the adverb’s three letters pop up again, followed by a colon and a list of hyperlinks to collections of which that snapshot is a part—for example “why: focused_crawls, top_domains.” Collection information is also included in the dropdown “About this capture” menu in the top right of a snapshot. Following either link will lead to the collection’s page, which in its “About” section includes the name and photograph of the person who created the collection, as well as a link to their user account. The human is back!

Except, not always. Sometimes the “why:” is followed by three non-hyperlinked words: “No Collection Info.” Moreover, even when there is collection information, the website design prioritises the archived object over the collection. A useful point of comparison here is the Library of Congress’s web archiving program. If one visits the program homepage, and then navigates to “Archived Websites,” one is met with a list of collections (Wayback Machine 2020f). These collections are made up of descriptive records of archived websites. As not all websites have been fully described yet, the Library of Congress also allows users to navigate to its holdings via URL, and in fact this is the older method of discovery on the Library of Congress website (Wayback Machine 2020g). This navigate by URL feature provides a Wayback Machine-like user experience, though functionality is much more limited. However, browsing by collection is the primary method of discovery promoted on the website. The Library of Congress has thus implemented collections in a manner which first prioritises collection and then archived webpage. The Wayback Machine’s design inverts this order of priority. You first search by URL or keyword. Only once you have chosen an archived webpage does the option to view collections information appear.

There are also differences in what constitutes a collection in these two web archives. The Library of Congress’s collections are based on themes and/or events. In the Wayback Machine, collections are equated to crawls. Each collection is formed of archived webpages from a “specific web crawl” (Wayback Machine 2020e). This is important because collections are not neutral methods of organising holdings. Rather, they provide a reason, and answer a question, as to why a particular holding has been archived. They provide a trace of agency. This difference in approach links into a wider debate surrounding how we should archive the web. The two opposing paradigms are that of the librarian-archivist versus the computer scientist. The question is whether web archives should include only carefully selected websites or should try to include everything. Lyman (2002) provides a good summary of this debate. Through its thematic and event-based collections, the Library of Congress web archive expresses a human judgement as to which websites have potential historical significance. The Wayback Machine, by conflating collection with crawl, ties its collections to the agency of the crawler. In the words of Wolfgang Ernst, “archives actively define what is at all archivable, insofar as they determine as well what is allowed to be forgotten" (2013, p. 139). Or, indeed, in the words of Arlette Farge, whom Ernst quotes: "The question is to know what to keep and what to abandon" (Ernst 2013, p. 139; Farge 1989, p. 87). The Library of Congress's web archive collections put this human decision at the forefront of the user experience. The Wayback Machine's website design, coupled with the language with which we comprehend the web crawling process, obfuscates the human agency at the heart of the crawl: the very agency which decides what is crawled, and what is allowed to be forgotten.

We define

What does the crawler do when it comes upon a new land? It photographs. The Internet Archive states it holds 424 billion webpages, but when you look at any one of these webpages the Wayback Machine refers to it as a "snapshot" or "capture." These terms are rooted in the language of photography and, in the case of "capture," more particularly in the discourse of archival digitisation. If an archive was looking to outsource the digitisation of some old documents in its collection, it might well be quoted a price per number of captures. (We might also talk of a screen capture when trying to make a visual copy of our screen at a given moment, though in this context screenshot is by far the more common term.) By contrast, when duplicating a digital object in order to archive it, the archival norm is to refer to the resultant objects as "copies," for instance the "preservation copy" and "display copy" created by digital preservation software. These digital objects are treated as identical even though they will likely be in different file types and thus contain quite radically different data. The language used by the Wayback Machine thus aligns more closely with that used in the digitisation of physical objects than it does the archiving of digital objects.

If the Wayback Machine can be said to photograph, where is its camera? The difficulty one faces in trying to answer this question suggests a strangeness. Photography without a camera. Compare this to Ariella Azoulay's description of our typical encounters with photography:

> photography is an event that is not conditioned by the eventual production of a photograph. Considered in relation to the camera or the photographed persons, this sounds obvious […] The photographed persons will not necessarily view the photographs taken at the photographic event of which they were a part, but this does not obliterate the fact that it took place (2010, p. 12).

The apparatus is key here. To encounter a photographic event, we need not see the photograph itself. Instead, encountering the in-use photographic apparatus is enough to signal to us that a photographic event has taken place. Picture yourself at a party in the 1990s: someone produces a disposable camera, you blink from its flash. You may never see the developed photograph, yet you have certainly encountered photography. Conversely, our encounter with the Wayback Machine's photography is always by way of its photographs. The Machine's apparatus is intangible, largely invisible, and basically unencounterable. Yet for Azoulay this is actually a symptom of our contemporary relationship with photography. "[A]t a time in which nearly everyone possesses photographic tools," the presence of the apparatus is no longer needed to signal a photographic event: "photography has become a potential event even when there is no camera visible" (2010, p. 12). In this way, the Machine's photography is distinctly contemporary in that it lacks a visible apparatus.

The photographic language suggests that the Wayback Machine's holdings are copies of webpages, rather than webpages themselves. The user experience works against this idea by being broadly analogous to the internet more widely: a user types a URL into a bar, and then a webpage appears. (A necessary intermediary step breaks the illusion slightly: the user must pick from a list of snapshots arranged in a calendar view by the date and time of capture.) A relatively recent development is the Machine's search feature, which removes the need for users to know the precise URL of a page to access the Wayback Machine's holdings (for more on this development, see: Ben-David and Huurdeman 2014). In this regard, the Machine's development mirrors the early years of the internet, when the rise of popular search engines such as Yahoo and Google altered the way users reached websites, allowing them to search by keyword. Again, the Library of Congress is a useful point of comparison. Its website encourages users to browse through a vertical list of collections. Alternatively, users can search archived websites via a bar in the top right corner of the webpage (Wayback Machine 2020f). The placement of this search bar is a common feature in the visual grammar of modern websites. Compare this placement to the Wayback Machine's search bar, which takes pride of place in the middle of the webpage (Wayback Machine 2020a). This is atypical of most modern websites, and approximates the search bar position of a particular type of website: web search engines. Because of this, users of the Wayback Machine get the subconscious yet uncanny feeling that what they are searching is not an archive of copies but rather the web itself.

To add to the confusion, the photographic language simplifies the act of archiving webpages into a straightforward capture or snapshot. In reality, when a user is looking at a "snapshot" in the Wayback Machine, what they are really looking at is a complicated interplay of web objects which often have all been crawled at different times. The Internet Archive does well to expose this stitched-together quality via timestamps provided in the "About this capture" menu attached to the top of each snapshot. These timestamps are "of all page elements compared to the date and time of the base URL of a page. This means that users can see, for instance, that an image displayed on a page was captured X days before the URL of the page or Y hours after it" (Wayback Machine 2020h). This links into the wider question of how to define a webpage. A single webpage will often have links to other pages or objects, as well as sourced objects such as images or sounds. From a technical standpoint, "the boundaries of the digital object are ambiguous" (Lyman 2002, p. 41).

The Internet Archive's explanation of what constitutes a webpage highlights these ambiguous boundaries while at the same time denying them:

> The Internet Archive has been archiving the web for 20 years and has preserved billions of webpages from millions of websites. These webpages are

> often made up of, and link to, many images, videos, style sheets, scripts and other web objects. Over the years, the Archive has saved over 510 billion such time-stamped web objects, which we term web captures.
>
> We define a webpage as a valid web capture that is an HTML document, a plain text document, or a PDF (Wayback Machine 2020i).

The Internet Archive blog post from which this passage is taken is also accessible via the homepage of the Wayback Machine. Above the Machine's search bar are the words: "Explore more than 424 billion web pages saved over time." The phrase "web pages" is hyperlinked to the blog post. On the Internet Archive's homepage, which also includes a Wayback Machine search bar in its centre, the text is slightly different ("Search the history of over 424 billion web pages on the Internet") but "web pages" again links to the blog post. This passage acknowledges the "many images, videos, style sheets, scripts and other web objects." In so doing, it traces the ambiguous technical boundaries of this puzzling thing, the "webpage." It then outlines a definition of "a webpage" which ignores these ambiguous boundaries. In this definition we see a slippage from snapshot/capture to webpage itself—from representation/copy to original. This is where the photographic metaphor, however flawed, is useful. We do not mistake the photograph for the photographed object.

The slippage becomes particularly important in light of how most users encounter the web. Regardless of the technical ambiguities surrounding the webpage's boundaries, we usually experience webpages as singular, discrete objects:

> From a user's point of view, a Web page is the image called forth by placing a URL address into a Web reader. This operational definition is necessary but not sufficient, for an archive also must be sure that the document is *translated* in an authentic manner. In this case, authenticity means that the document must both include the context and evoke the experience of the original (Lyman 2002, p. 41).

The Internet Archive's definition of "a webpage" hinges on the 2xx class of HTTP status codes: the word "valid" in the blog post is hyperlinked to the "2xx Success" section of the Wikipedia page which lists these status codes (Wayback Machine 2020j). The issue here is that the return of a 2xx code from the server does not always ensure that a document has been "*translated* in an authentic manner." The technical limits of web archiving mean that an archived "webpage" is sometimes not synonymous, in an operational sense, with the webpage a user would have experienced if they had loaded the URL contemporaneously in their browser. One of the more palpable examples here is when a "valid" archived webpage, as viewed in the Wayback Machine, lacks some or all of its images. The user is met with a white box where an image would have been. In the top left corner of the box: an icon depicting a sheet of paper ripped in half.

This problem will likely get worse. More and more of the web's architecture is moving away from static pages, in favour of webpages that have been generated on the fly. Webpages that we encounter in our day to day use of the web, whether static or otherwise, are part of the surface web. Yet the much larger portion of the web is the deep web, which is not indexed by web search engines. Due to its structure, the deep web itself is very difficult to crawl, though some limited progress has been made on this front. The deep web also includes huge data sources and the software code responsible for generating surface webpages on the fly. If you were to navigate to Amazon's website, you would—assuming cookies stored on your computer logged you in automatically—be met with a personalised homepage that had been created on demand using databases in the deep web. To put this another way: museum curator Doron Swade's (1998) observation that software develops only in the course of its execution gains new resonance in relation to contemporary webpages. Or another way: "The deep Web is the information architecture that produces what we read on the surface; the surface itself exists only as long as a reader is using it" (Lyman 2002, p. 41). There's the rub. If a webpage only exists as long as it is executed in a browser, and if that execution relies on each individual user, the Wayback Machine has little chance. With our surface web experience becoming increasingly personalised, it becomes increasingly difficult—and in many cases, currently impossible—to archive. The webpage that you see now, loaded in your browser, is allowed to be forgotten.

Words, words, words

Let us return to Foucault's "Of Other Spaces," to a moment where he outlines the paradox at the heart of heterotopias:

> the curious property of being in relation with all the other sites, but in such a way as to suspect, neutralize, or invert the set of relations that they [heterotopias] happen to designate, mirror, or reflect. These spaces, as it were, which are linked with all the others, which however contradict all the other sites (Foucault and Miskowiec 1986, p. 24).

Whenever I read this part of Foucault's essay, a question arises in my mind: how might one thing appear to mirror another thing, while at the same time contradicting that other thing? Each of the Machine's snapshots mirrors a certain webpage at a certain time in the past, even if sometimes this mirroring is imperfect. The Machine only becomes directly useful when a webpage is either altered or no longer extant. Thus, the Machine might be said to mirror the web (as it was at one time) while simultaneously contradicting the web (as it is currently). While one can technically view a recent snapshot of a webpage that has remained extant and unchanged, the Internet Archive implicitly stresses that changed or lost pages provide the Machine's raison d'être. For one, the organisation encourages the celebration of "404 Day" on the fourth of April. It has also made much of a recent partnership with the Brave web browser, which now offers one-click access to snapshots when a 404 page is discovered. In an Internet Archive blog post announcing the partnership, a delightfully unambiguous pseudo-mathematical equation combines the Machine's logo with the Brave logo and the universal prohibition symbol to express that Wayback Machine + Brave browser = No 404s (Wayback Machine 2020k).

Here are some of the terms used when talking about 404s:

> *Decay* (see for example: Hennessey and Ge 2013; Sampath Kumar and Manoj Kumar 2012; Russell and Kane 2008).
>
> *Rot* (see for example: Wayback Machine 2020b, c; Davis 2016; Ince 2013).
>
> *Dead* (see for example: Wayback Machine 2020c; Sampath Kumar and Prithviraj 2015).

These terms have become commonplace, and are worth contemplating. They make the process they describe sound natural, agentless, and inevitable, when it is often anything but. 404s are usually the result of human actions—or inaction. The relationship between a hyperlink and the webpage to which it points does not degrade gradually like decaying timber or a rotting corpse. Rather, the connection between the two is immediately and abruptly severed when a webpage is deleted or moved, or when someone forgets to renew the domain. Here, much like with the crawler-crawl-frontier metaphors, language conceals human agency.

Death is relevant to the Machine in another way. As the web grows older, more and more of it will be made up of traces of the dead. But some traces will not be so lucky. Failure to renew a domain may bring an abrupt end, though this end may not be totally final if a spectre of the page is archived in the Machine. As Caroline Steedman puts it in relation to archives more generally, it is not only the literal holdings which are significant, but also "what is not actually there, with the dead who are not really present in the whispering galleries, with the past that does not, in fact live in the record office, but is rather, *gone* (that is its point; that is what the past is for)" (2002, p. 81). This *goneness* is foundational to the creation of history. As Christina Riggs puts it: "The uncanny dead that Steedman had in mind were not corpses or ghosts, but those individuals whose traces survive in the documents, letters, and ledger books of the archive, which she construes as the precursor of history writing" (2017, p. 127). Wolfgang Ernst outlines the link in more active terms, contending that history itself is "not simply given, but rather is first produced through the medium of the archive" (2003, p. 554)—I quote Markus Friedrich's translation from the German (2018, p. 10). For Ernst, the archive is not simply descriptive, but productive. He is not alone in this view. Peter Fritzsche, for instance, speaks of "archival production" (2005, p. 16). This understanding of the role of the archive has a particular bearing on the Wayback Machine. When Anat Ben-David asks the question "What does the Web remember of its deleted past?" (2016), the implicit answer is *nothing*: the internet is an everlasting present that is constantly being over-written. It has no history in and of itself. To use Ernst's own phrase: "cyberspace has no memory" (Ernst 2013, p. 138).

It is easy to think of language in purely communicative terms—that the words we use merely describe the phenomena we experience. But whenever we speak of a "crawler," or of link "decay," or of the Wayback Machine's holdings as "webpages," we are thinking through very specific metaphors. As this essay has explored, each of these metaphors not only describes but also obfuscates. It may be that, for most people most of the time, the thinking behind the metaphor is offloaded into language itself. But that does not mean the thinking ceases to exist? With the relentless passage of time, the web's past gets more remote. On the Internet Archive's homepage, twelve words are inscribed above the Wayback Machine's search box: "Search the history of over 424 billion web pages on the Internet" (Wayback Machine 2020a). If we are to use the Machine to remember the web's forgotten past, we must also remember that the language we use to comprehend the Machine has its shortcomings. We must be thoughtful of the human agency behind the web's past as we see it in the Machine, and thoughtful of the gaps in this past.

Unfamiliar technology can at first seem impenetrable. Metaphors provide a way in. By invoking the known, we familiarise the unknown. The more we then use this language, the less we think about it. New words, through repeated use, become commonplace. The understanding that brought the metaphor into being can no longer be found in the minds of the people using the metaphor. Some metaphors become so ingrained that we forget they are metaphors. Language itself thus becomes the trace of the understanding. In this way, a language-oriented approach to technology attempts to re-inscribe that which is lost in the course of familiarisation. We apprehend technology through language, so by better understanding the workings of this language we better understand our relationship to technology—and we are reminded of the fact that technology always, to some degree, escapes language. Our thoughtfulness re-enchants our tools.

References

Al Noamany Y, Al Sum A, Weigle MC, Nelson ML (2014) Who and what links to the Internet Archive. Int J Digit Libr 14(3–4):101–115

Azoulay A (2010) What is a photograph? What is photography? Philos Photogr 1(1):9–13

Battley B (2019) Archives as places, places as archives: doors to privilege, places of connection or haunted sarcophagi of crumbling skeletons? Arch Sci 19(1):1–26

Belovari S (2017) Historians and web archives. Archivaria 83:59–79

Ben-David A (2016) What does the Web remember of its deleted past? New Media Soc 18(7):1–17

Ben-David A, Huurdeman H (2014) Web archive search as research. Alexandria 25(1–2):93–111

Bridle J (2018) New dark age: technology and the end of the future. Verso, London

Crichlow R, Winbush N (2004) Accessibility and accuracy of web page references in 5 major medical journals. J Am Med Assoc 292:2723–2724

Daigle L (2016) On the nature of the internet. A universal internet in a bordered world: research on fragmentation, openness and interoperability. Global Commission on Internet Governance. Centre for International Governance Innovation, Waterloo

Davis RC (2016) The future of web citation practices. Behav Soc Sci Libr 35(3):128–134

Dellavalle RP, Hester EJ, Heilig LF, Drake AL, Kuntzman JW, Graber M, Schilling LM (2003) Going, going, gone: lost internet references. Science 302:787–788

Derrida J (1996) Archive fever: a Freudian impression. Prenowitz E (trans). University of Chicago Press, Chicago (Original work published in French 1995)

Ernst W (2003) Im Namen von Geschichte. Fink, Munich

Ernst W (2013) Digital memory and the archive. University of Minnesota Press, Minneapolis

Evangelou E, Trikalinos TA, Ioannidis JPA (2005) Unavailability of online supplementary scientific information from articles published in major journals. FASEB J 19(14):1943–1944

Farge A (1989) Le gout de l'archive. Seuil, Paris

Foucault M (2002) The archaeology of knowledge. Sheridan Smith AM (trans). Routledge, London and New York (Original work published in French 1969)

Foucault M, Miskowiec J (1986) Of other spaces. Diacritics 16(1):22–27

Friedrich M (2018) The birth of the archive: a history of knowledge. Dillon JN (trans). University of Michigan Press, Ann Arbor (Original work published in German 2013)

Fritzsche P (2005) The archive. Hist Mem 17(1–2):13–44

Gleick J (2012) The information. Fourth Estate, London

Hennessey J, Ge SX (2013) A cross disciplinary study of link decay and the effectiveness of mitigation techniques. BMC Bioinform 14(Supplement 14):S5

Hester EJ, Heilig LF, Drake AL, Johnson KR, Vu CT, Schilling LM, Dellavalle RP (2004) Internet citations in oncology journals: a vanishing resource? J Natl Cancer Inst 96(12):969–971

Ince D (2013) link rot. A Dictionary of the internet, 3rd edn. Oxford University Press, Oxford

Kanhabua N, Kemkes P, Nejdl W, Nguyen TN, Reis F, Tran NK (2016) How to search the Internet Archive without indexing it. In: Fuhr N, Kovács L, Risse T, Nejdl W (eds) Research and advanced technology for digital libraries. 20th international conference on theory and practice of digital libraries. Proceedings. Springer, Basel, pp 147–160

Kaur R (2015) Writing history in a paperless world: archives of the future. Hist Workshop J 79(1):242–253

Kumar BTS, Vinay Kumar D, Prithviraj K (2015) Wayback machine: reincarnation to vanished online citations. Program 49(2):205–223

Lowry J (2019) "Displaced archives": proposing a research agenda. Arch Sci 19(4):349–358

Lyman P (2002) Archiving the world wide web. Building a National Strategy for Digital Preservation: Issues in Digital Media Archiving. Council on Library and Information Resources and Library of Congress, Washington D.C., pp 38–51

MacIntyre J (2012) Milad Doueihi: digital cultures. Publ Res Q 28:147–149

McDonald J (1995) Managing records in the modern office: taming the wild frontier. Archivaria 39:70–79

Milligan I (2016) Lost in the infinite archive: the promise and pitfalls of web archives. Int J Hum Arts Comput 10(1):78–94

Mohr G, Stack M, Ranitovic I, Avery D, Kimpton M (2004) An introduction to Heritrix: an open source archival quality web crawler. In: Proceedings of the 4th international web archiving workshop IWAW'04, July 2004, Bath, UK

Riggs C (2017) The body in the box: archiving the Egyptian mummy. Arch Sci 17:125–150

Rogers R (2017) Doing web history with the Internet Archive: screencast documentaries. Internet Hist 1(1–2):160–172

Russell E, Kane J (2008) The missing link: assessing the reliability of internet citations in history journals. Tech Cult 49(2):420–429

Sampath Kumar BT, Manoj Kumar KS (2012) Decay and half-life period of online citations cited in open access journals. Int Inf Libr Rev 44(4):202–211

Sampath Kumar BT, Prithviraj KR (2015) Bringing life to dead: role of Wayback Machine in retrieving vanished URLs. J Inf Sci 41(1):71–81

Steedman C (2002) Dust: the archive and cultural history. Rutgers University Press, New Brunswick

Steedman C (2011) After the archive. Comp Crit Stud 8(2–3):321–340

Swade D (1998) Preserving software in an object-centred culture. In: Higgs E (ed) History and electronic artefacts. Clarendon, Oxford, pp 195–206

Wayback Machine (2019) Oops! This is awkward (donaldjtrump.com). https://web.archive.org/web/20190717030514/https://www.donaldjtrump.com/404. Accessed 10 Jan 2020

Wayback Machine (2020a) Internet Archive (archive.org). https://web.archive.org/web/20200109234109/https://archive.org. Accessed 9 Apr 2020

Wayback Machine (2020b) In Supreme Court Opinions, Web Links to Nowhere (nytimes.com). https://web.archive.org/web/20200220110811/https://www.nytimes.com/2013/09/24/us/politics/in-supreme-court-opinions-clicks-that-lead-nowhere.html. Accessed 9 Apr 2020

Wayback Machine (2020c) Link rot (wikipedia.org). https://web.archive.org/web/20200104002147/https://en.wikipedia.org/wiki/Link_rot. Accessed 9 Apr 2020

Wayback Machine (2020d) The Wayback Machine: Fighting Digital Extinction in New Ways (archive.org). https://web.archive.org/web/20200108102319/https://blog.archive.org/2019/10/18/the-wayback-machine-fighting-digital-extinction-in-new-ways/. Accessed 9 Apr 2020

Wayback Machine (2020e) FAQs for some new features available in the Beta Wayback Machine (archive.org). https://web.archive.org/web/20200102122915/http://blog.archive.org/2016/10/24/faqs-for-some-new-features-available-in-the-beta-wayback-machine/. Accessed 9 Apr 2020

Wayback Machine (2020f) Archived Websites (loc.gov). https://web.archive.org/web/20200412111941/https://www.loc.gov/programs/web-archiving/archived-websites/. Accessed 12 Apr 2020

Wayback Machine (2020g) Web Archives (loc.gov). https://web.archive.org/web/20200310022744/http://webarchive.loc.gov/. Accessed 9 Apr 2020

Wayback Machine (2020h) Wayback Machine Playback… now with Timestamps! https://web.archive.org/web/20200310040802/https://blog.archive.org/2017/10/05/wayback-machine-playback-now-with-timestamps/. Accessed 9 Apr 2020

Wayback Machine (2020i) Defining Web pages, Web sites and Web captures (archive.org). https://web.archive.org/web/20200406010236/https://blog.archive.org/2016/10/23/defining-web-pages-web-sites-and-web-captures/. Accessed 9 Apr 2020

Wayback Machine (2020j) List of HTTP status codes (wikipedia.org). https://web.archive.org/web/20200408161025/https://en.wikipedia.org/wiki/List_of_HTTP_status_codes/. Accessed 9 Apr 2020

Wayback Machine (2020k) Brave Browser and the Wayback Machine: working together to help make the Web more useful. (archive.org). https://web.archive.org/web/20200402091540/https://blog.archive.org/2020/02/25/brave-browser-and-the-wayback-machine-working-together-to-help-make-the-web-more-useful-and-reliable/. Accessed 9 Apr 2020

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Surya Bowyer is Head of Library Services at Queen Elizabeth's School, Barnet. Previously, he taught at Sorbonne University's Faculté des Lettres. His research interests include alternative archives, the relationship between word and image, and the history and theory of media. He is currently working on a digital archive at Queen Elizabeth's.