Special Issue: Critical Technical Practice(s) in Digital Research

CONVERGENCE

Convergence: The International
Journal of Research into
New Media Technologies
2023, Vol. 0(0) 1–23

DOI: 10.1177/13548565231164759
journals.sagepub.com/home/con
SAGE

Know(ing) Infrastructure: The Wayback Machine as object and instrument of digital research

Jessica Ogden
University of Bristol, UK

Edward Summers
Stanford University, USA

Shawn Walker
Arizona State University, USA

Abstract

From documenting human rights abuses to studying online advertising, web archives are increasingly positioned as critical resources for a broad range of scholarly Internet research agendas. In this article, we reflect on the motivations and methodological challenges of investigating the world's largest web archive, the Internet Archive's Wayback Machine (IAWM). Using a mixed methods approach, we report on a pilot project centred around documenting the inner workings of 'Save Page Now' (SPN) – an Internet Archive tool that allows users to initiate the creation and storage of 'snapshots' of web resources. By improving our understanding of SPN and its role in shaping the IAWM, this work examines how the public tool is being used to 'save the Web' and highlights the challenges of operationalising a study of the dynamic sociotechnical processes supporting this knowledge infrastructure. Inspired by existing Science and Technology Studies (STS) approaches, the paper charts our development of methodological interventions to support an interdisciplinary investigation of SPN, including: ethnographic methods, 'experimental blackbox tactics', data tracing, modelling and documentary research. We discuss the opportunities and limitations of our methodology when interfacing with issues associated with temporality, scale and visibility, as well as critically engage with our own positionality in the research process (in terms of expertise and access). We conclude with reflections on the implications of digital STS approaches for 'knowing infrastructure', where the use of these infrastructures is unavoidably intertwined with our ability to study the situated and material arrangements of their creation.

Corresponding author:
Jessica Ogden, School of Sociology, Politics and International Studies, 11 Priory Road, Bristol BS8 1QU, UK.
Email: jessica.ogden@bristol.ac.uk

Keywords

critical technical practice, digital history, digital methods, internet archive wayback machine, infrastructure studies, save page now, science and technology studies, web archives

Introduction

The Web presents a continually changing target for researchers. Due to the dynamic nature of the Web and its inhabitants, many have advocated for the collection and use of *web archives* with the aim of stabilising and rendering online objects of investigation legible for future analysis. Providing the underlying data for web-based research (and a range of other uses), web archives can be seen as a form of knowledge infrastructure that supports the production and dissemination of knowledge about the Web through time. However, as partial and contingent representations of the Web, these knowledge infrastructures also necessitate critical engagement with the terms of their collection and use. But how can we come to know the infrastructures that support web archives in practice?

In this article, we reflect on the motivations and methodological challenges of investigating the world's largest public web archive, the Internet Archive's Wayback Machine (IAWM). Since 1996, the IA has acted as an online repository for film, television, books, music, games and other digital media, as well as collected, maintained and provided access to 'snapshots' of the Web through the IAWM.[1] With the addition of the 'Save Page Now' tool (SPN) in 2013, the IA enables anyone with a web browser and an Internet connection to save a web resource directly to the Wayback Machine. The IAWM is widely recognised as the largest web archive in the world, and according to recent estimates, over 100 URLs per second are added to the archive via the SPN API and browser-based tool.[2] Despite its prolific use, very little is known about the sociotechnical processes supporting SPN, the ways it is used, the motivations driving use, and indeed what is being saved. Our work is motivated by the observation that despite the positioning of IAWM as a critical resource for a broad range of scholarly Internet research agendas (Karpf, 2012; Rogers, 2013) and its widespread use as a tool for evidence-based accountability online (Hackett and Parmanto, 2005; Murphy et al., 2008; Quarles and Crudo, 2014), relatively little has been written about how to engage with the study of web archiving as a form of critical technical practice that is actively shaping future knowledge about the Web. And whilst there is an emerging body of scholarship that documents substantive and methodological considerations for *using* web archives in digital research (Brügger, 2018; Milligan, 2019), here we diverge to discuss the challenges of studying the infrastructure of web archiving *itself*.

In light of this, we designed a pilot project to examine, document and 'reverse engineer' the processes that create web archives in the context of SPN in order to better understand how this public service mediates our knowledge about the Web's past. In this article, we reflect on the methodological opportunities and challenges for Science and Technology Studies (STS) approaches to studying emergent knowledge infrastructures. Building upon existing digital STS approaches, we chart our development and application of methods for studying SPN, including a mix of ethnographic methods, experimental blackbox tactics, data tracing, modelling and documentary research. Using the web archives themselves, we sketch a sociotechnical history of SPN and reverse engineer the role of various infrastructural components, as well as their implications for the web archives they produce. We discuss challenges encountered when interfacing with issues associated with access, temporality, scale and visibility, and reflect on the opportunities and limitations of our methodology as a form of 'critical technical practice' (Agre, 1997). We also critically engage with our own positionality in the research process and raise questions about who can study these types of infrastructures at scale – including what forms of expertise and access are privileged in their examination.

We conclude with reflections on the implications of digital STS approaches for 'knowing infrastructure', where the use of these infrastructures is unavoidably intertwined with our ability to study the situated and material arrangements of their creation. In addition to examining how we come to *know infrastructure*, the article recognises the ways that web archives act as a type of *knowing infrastructure* that also accumulates knowledge about the Web and its users over time (Ben-David and Amram, 2018). As web archives are increasingly amassed as 'algorithmic fuel' for computing machine learning models, which get operationalised in a wide variety of everyday applications (Jo and Gebru, 2020), we point towards the critical need for more work that further explores the relationship between how collections of archived web content create, shape and delimit new knowledge, as well as the new methodologies needed to understand them.

Background

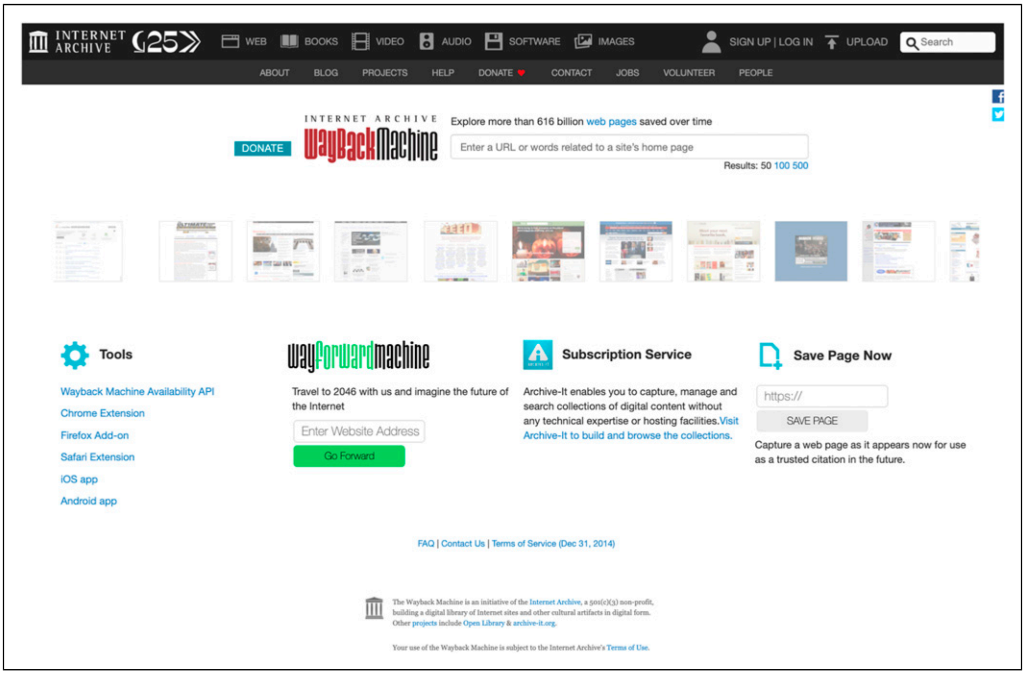
Web archiving and the internet archive

The practice of web archiving extends preservation and access activities developed for physical documents and media to content that is available on the World Wide Web. At a high-level, the infrastructures that support web archiving typically consist of two main components: *crawling* technologies for collecting and storing content from the Web and *playback* technologies for making that content viewable again through the web browser. Brügger (2016) argues that web archives are *reborn digital* in the sense that content is first *born digital* somewhere on the Web, then collected, processed and made available in a web archive elsewhere on the Web where they get *born again*. This framing underscores the ways that web archives are not merely copies or static mirror images of the Web, but rather, how web archives should be seen as partial and dynamic representations that are contingent upon the sociotechnical processes (people, labour, software, algorithms, bots, etc.) that underpin their creation and use.

Whilst web archiving practices are diverse and employed by many different types of organisations and actors today, the Internet Archive is a central figure in the field through their development of web archiving standards, practices and the large-scale data archives they produce. With its current home page listing 616 billion web pages archived, the IAWM is routinely engaged in a range of selective, massive and iterative web archiving activities, as well as providing services and software tools for other organisations to save web resources into the Wayback Machine, such as their own subscription service Archive-It. While it is not the only archive of its kind; its size, collection scope and commitment to public access has made IAWM an important source for scholarship about the Web's past (Brügger, 2018; Milligan, 2019) and a critical node for the Web as a knowledge infrastructure.

Save page now: A public service for web archiving

Since launching the service in 2013, the Internet Archive has promoted SPN as an open access tool for 'saving the web pages [users] care about most' (Rossi, 2017). In its most basic sense, SPN is a web-based tool that allows users to capture and preserve web content in the IAWM. The tool manifests as a simple web form on the home page of the Wayback Machine (Figure 1), as well as browser extensions and plugins that 'bring the power of the web archive to a regular user' (Milligan, 2019: 216). In addition, the capture of sites can be automated via the command line (eg using curl



Explore more than 616 billion web pages saved over time

Enter a URL or words related to a site's home page

Travel to 2046 with us and imagine the future of the Internet

Enter Website Address

Archive-It enables you to capture, manage and search collections of digital content without any technical expertise or hosting facilities. Visit Archive-It to build and browse the collections.

https://

Capture a web page as it appears now for use as a trusted citation in the future.

The Wayback Machine is an initiative of the Internet Archive, a 501(c)(3) non-profit, building a digital library of Internet sites and other cultural artifacts in digital form. Other projects include Open Library & archive-it.org.

Your use of the Wayback Machine is subject to the Internet Archive's Terms of Use.

Figure 1. The home page of the internet archive wayback machine displaying the 'save page now' tool (bottom right).

and custom python tools), enabling users to crawl sites at scale and regularly timed intervals to capture snapshots of web resources over time.

Elsewhere SPN has been acknowledged as one of several mechanisms by which the Internet Archive is attempting to 'leverage the power and labour of the crowd' in order to diversify the selection of domains and content for the Wayback Machine (Ogden et al., 2017: 306). Ben-David and Amram (2018) position the IAWM as a networked assemblage of (temporally and geographically) situated practices and identifies SPN as one of many distributed 'epistemic processes' that shapes how and what of the Web is saved. And as Milligan (2019: 216–217) describes, SPN can be seen as part of a broader effort to 'democratise' access to the tools of web archiving that have historically been limited to large memory institutions with the capacity to build and store web archives at scale. SPN is regularly positioned by IA and others as a critical tool for citation, transparency and accountability online, enabling the preservation of social media, news media, government websites and other forms of web content by 'citizen archivists' for a myriad of future uses (Beis et al., 2021). However, despite these claims and prolific use, recent research has also observed that SPN is being used to (for example) capture and circulate health misinformation (Acker and Chaiet, 2020; Donovan, 2020) and terrorist propaganda online (Kelion, 2018; Littman, 2018). These observations underscore the role and significance of SPN in both the IAWM and the wider circulation of content online and open the door to further questions about how to investigate this knowledge infrastructure at scale.

Studying web archiving as critical technical practice

With this context in mind, we designed a pilot project with the aim of investigating SPN. In this article, we focus on the challenges of operationalising a sociotechnical study of SPN, the IAWM and web archiving practices as a form of knowledge infrastructure. Traditional methods focusing on a singular approach (i.e. quantitative vs qualitative vs so-called 'big data') cannot fully help researchers explore the relationship between how these infrastructures create, shape and delimit new knowledge. The article responds and contributes to the special issue on critical technical practice (Agre, 1997) by centring our experience of reverse engineering the tools of web archiving and their implications for the data archives they produce. Our project takes an experimental approach to probing and methodically 'attacking' the black box (McMillan Cottom, 2017) to lay the groundwork for future research that builds on these tactics to study participatory web archiving infrastructure.

Fundamentally, the pilot project presented the opportunity to document and critically engage with the methods and access required to study knowledge infrastructures at scale – an issue of long-standing interest of STS scholars.[3] This positioning is supported by STS and infrastructure studies scholarship that has illuminated the ways that processes, practices, classifications and standards, software, data and people shape the nature of how information is produced and circulated (Bowker et al., 2010; Bowker and Star, 1999; Edwards et al., 2009; Shankar et al., 2016). When considering web archives, we actively draw on Edwards' (2013: 17) definition of knowledge infrastructures that takes into account the 'robust networks of people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds'. Referring to scientific knowledge practice, Edwards argues that the 'language of infrastructure' works to recognise the everyday ways that knowledge production is underpinned by sociotechnical 'supports' that are often taken for granted but fundamentally make knowledge (and the scientific practices that produce knowledge) endure. Here, web archives act as a particular form of knowledge infrastructure that is increasingly intertwined with the Web itself,[4] whilst also providing persistent, citable and standalone data archives of the Web through time.

These archives are shaped by web archiving practices themselves (i.e. what and how the Web is collected, maintained and preserved for future use), as well as actively conditioning how we will come to know the Web in future (Ben-David and Amram, 2018; Ogden et al., 2017). Therefore in order to make use of web archives, a form of critical technical practice (Agre, 1997) is required to make visible the contingent nature of these data archives and the tools used to produce them. Rooted in this premise, recent work in and on web archives illustrates the importance of studying the processes and practices of web archiving (Maemura et al., 2018; Milligan et al., 2016; Ogden et al., 2017; Summers and Punzalan, 2017), where these archival interventions are framed as a relatively under-examined, yet an increasingly embedded component of the Web's architecture. Underscoring the need for revealing the 'knowledge-production mechanisms' that enable the IAWM, our project is bolstered by the work of Ben-David and Amram (2018) who argue against the idea of web archives as *passive* agents or so-called 'lobster traps' (Karpf, 2012: 648–649) which sit quietly collecting web content amidst 'a sea of big data'. This passive framing of web archives draws on a much longer arc of thinking about archives that are shaped by natural (Jenkinson, 1948), organic (Eastwood, 1994) or sedimenting processes (Caravaca, 2017). STS provides a critical frame for unpacking these *natural* archive metaphors in order to make visible the *active* infrastructural processes that work to facilitate access to the Web's past.

Designing an approach: Cobbling, scavenging and opening the black box

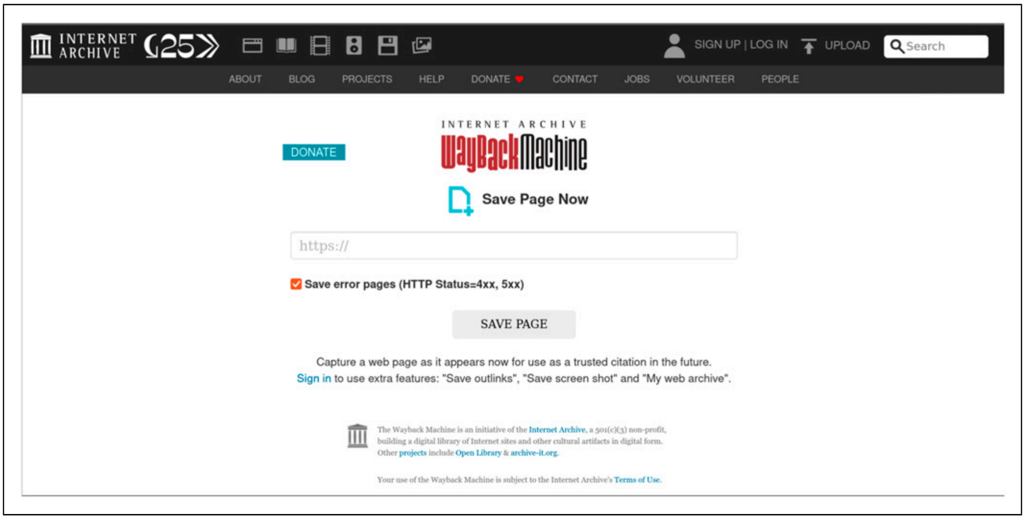
To address our research aims we implemented a methodological strategy that drew inspiration from several sociomaterial and computational approaches to digital infrastructure studies. This involved looking at tactics that attempt to circumvent barriers or limitations of access, including scavenging (Seaver, 2017), and 'covert cobbling' (McMillan Cottom, 2017) inspiring the team to develop experimental techniques for probing SPN. In particular, our approach draws on critical reverse engineering (Gehl, 2017), providing a genealogical method for documenting and 'tracing' how software is situated within a network of relations that shapes the 'conditions of possibility' that make software a reality. Akin to Star's (1999) 'ethnography of infrastructure', this requires examining the organisations that create software, how software users are imagined and configured, as well as tracking the evolution of software over time (eg preceding versions).

In this way, our approach can be seen as aligned with the 'forensic social sciences' (McFarland et al., 2016) methods used by Ben-David and Amram (2018) to study the collection practices that led to the deposition of archival snapshots of (typically elusive) websites from the North Korean top-level domain to the IAWM. Whereas Ben-David and Amram (2018) focus on questions pertaining to the epistemic processes and mechanisms that shape how the IAWM 'knows' the Web, here we extend this work by 'zooming in' to the SPN service specifically, to develop probing tactics for revealing its role in shaping IAWM content over time.[5] To complement this approach, we drew on a 'theory-methods' package advocated by Nicolini (2012) for organisational studies which places emphasis on a 'zoom in/zoom out' approach to investigating and conceptualising the connections between localised practices and the products and implications of practice – in our case, the users and web archives generated by SPN. Significance is placed on iteratively zooming in and out of the digital traces produced by SPN, as well as the broader organisational context within which the infrastructure is supported over time in order to follow the 'trails of connections between practices and their products' (Nicolini, 2012: 219).

The seeds of the pilot project originally emerged from previous ethnographic work at the Internet Archive, which provided both the organisational context of SPN and a lens through which the project was initially approached (Ogden, 2020). During this previous work, SPN was identified as a key component of IA's web archiving infrastructure, offering one of the few publicly accessible services for users to create, store and subsequently share archived versions of web-based content. Ethnographic observations and interviews with IA staff, combined with additional documentary work, provided both the motivations for extending the study to SPN, but also supplied further qualitative context for some of the inner workings of both the organisation and the sociotechnical infrastructure that are often obscured through methodologies solely reliant on computational or 'distant reading' approaches to infrastructure studies. Interviews with staff revealed both anecdotal accounts of SPN use and a desire by key informants to know more about how SPN was actually being used 'in the wild'. For example, IA staff speculated that the majority of SPN traffic was contributed by automated agents or 'bots' using the tool at scale. They also suspected that third parties used SPN to ingest their own web content into the Wayback Machine, thus turning the IA into a de-facto web hosting service. Insights into practice gleaned during ethnographic field work, as well as the where, when and who of the wider IA infrastructure and organisation were critically important for framing the investigation of SPN.

In support of our methods/strategies (described below), during the core phases of the project we also conducted weekly video conference calls to discuss the state of the project, new results and possible approaches to tackling emerging questions and problems encountered. Given the distributed and interdisciplinary nature of the team, we used Google Docs, Slack, GitLab and Python/Jupyter notebooks to communicate, share and document our work. Strategies for collaborative working have been an essential element of our methodological design.

In the next section we provide 'notes from the field' which document a reflective account of our chosen design and investigative strategies for illuminating how SPN works in practice. Rather than



☑ Save error pages (HTTP Status=4xx, 5xx)

Capture a web page as it appears now for use as a trusted citation in the future.
Sign in to use extra features: "Save outlinks", "Save screen shot" and "My web archive".

The Wayback Machine is an initiative of the Internet Archive, a 501(c)(3) non-profit, building a digital library of Internet sites and other cultural artifacts in digital form. Other projects include Open Library & archive-it.org.

Your use of the Wayback Machine is subject to the Internet Archive's Terms of Use.

Figure 2. The 'white box' of save page now.

communicate particular substantive findings about SPN, the intention is to foreground the evolution and *process* of critical technical analysis to understand SPN as knowledge infrastructure.

Black boxing SPN and crafting entrée

Our attempts to study SPN were initially framed by an assumption that access to the archived web data that SPN generates would yield insights into what was being archived and what users were attempting to achieve by using this service. SPN was thus cast as a *black box*, a topic of enduring interest to STS, especially in academic and popular conceptions of algorithmic processes (Bucher, 2017; Geiger, 2017; Pasquale, 2015). At first encounter, SPN invites the black box framing as its interface presents a single input field (a white box), where a user enters a URL to be archived (Figure 2). Once submitted, the software performs its archiving function, deposits the data in the Live Web Proxy Crawls collection and renders the results for playback in the Wayback Machine.[6] Since 2008, the Internet Archive has captured and stored web archives in the WARC file format (Web ARChive), an international standard (ISO 28500:2017) that grew from their early experiments to serialise the HTTP requests and responses that transpire when viewing a web page (*The WARC format 1.1*, 2017). This WARC data (and its many different 'derivative' file formats) provides the underlying basis for the Wayback Machine, as well as our analysis presented here.

However, as we discuss further below, once granted access to the full dataset, we quickly realised that SPN resembled less of a black box (or singular piece of software or set of algorithms), but rather more like a series of black boxes containing their own internal logics, technical affordances and hidden relations that shape the nature of web archive records. The need to understand SPN as an artefact that had been 'black boxed' or made opaque by its successful articulation as a simple web form, was replaced by the recognition that we required an approach that addressed the relational nature of SPN and the data it produced.

As described above, contacts established during (Ogden's, 2020) ethnographic work at the IA were the foundation for the study of SPN presented here. These initial connections provided inroads for understanding the relational nature of SPN, for example, knowing who to talk to, how to interpret what was said, as well as what documentation to look for and how to contextualize it within the organisational norms and cultural politics of the IA. However, it was also apparent that our individual informants also had partial knowledge of how SPN fit into the larger assemblage of the IAWM. After 20 years of operation, IA has developed an array of often ad-hoc, but continually evolving archiving practices in response to changes and growth in the scale of the Web, especially as staff have come and gone. What may first appear to be a monolithic crawling process, quickly resolved into a complex tapestry of curation practices, content moderation and feedback loops, once we were granted even a partial view from the inside (Ogden et al., 2017). SPN is a singular (if also, multi-faceted) part of this larger architecture because it is one of the few public-facing components by which members of the 'the public' can add web archives to the IA, unimpeded by subscription fees or insider knowledge of how web archiving tools work. In this way, SPN stands out, like exposed piping, or an intake valve onto the larger Web.

We observed a type of paradox in IA's public mission to provide 'universal access to all knowledge' using an infrastructure that was largely opaque in terms of its operations. The IAWM software infrastructure performs a critical gate-keeping function on the crawled web data. On the one hand, SPN data are listed openly in the IA's object storage as WARC files, but they are not fully accessible like other media objects which typically can be browsed and downloaded. While they are visible, the files in the Live Web collection[7] themselves are not available for download. Building the ethnographic work, we therefore established a research agreement with the IAWM team which granted the project access to the SPN WARC data contained within the Live Web collection.[8]

As we discovered when we were given access to the data and began analysis, the decision-making activities around what was worth archiving from the Web retreated from view even as we were granted entrée to view the SPN operations closer and analyse the WARC data written by their web crawling operations. SPN provides a portal into this infrastructure for the public to shape IA's collections, but the mechanics of how this process operates remained largely opaque. Or as Seaver (2013) argues:

> The use of phrases like "the Google algorithm" or "the Facebook algorithm" should not fool us into thinking that our objects are simple, deterministic black boxes that need only to be opened. These algorithmic systems are not standalone little boxes, but massive, networked ones with hundreds of hands reaching into them, tweaking and tuning, swapping out parts and experimenting with new arrangements (Seaver, 2013: 10).

In the case of SPN these hands included not only users, designers and engineers behind the SPN service, but also those that had fashioned other parts of the IA, such as the storage infrastructure used to house the data, and the IAWM used to reconstitute and replay archived web content. The further the network expanded, the less useful the metaphor of the black box became, but what remained was an abiding interest in the network of relations that constitute SPN and the so-called 'big data' it generates.

The limitations of access

However, while we were granted complete access to this data through the IA's API, we ran into practical limits in operationalising this access at scale. As a starting point and prior to the creation of our data corpus, we focused on first mapping the SPN collection's metadata as a mechanism for detecting the tool's use and the collection's rate of growth. Using and building on the `internetarchive` Python library, we queried the IA's API and plotted the item-level metadata

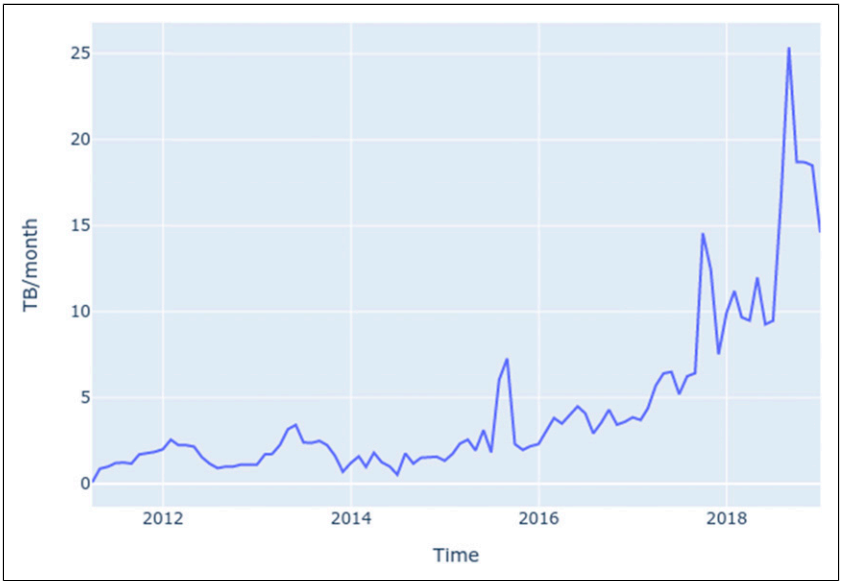


Figure 3. Save page now data ingest rate in TB/month.

associated with SPN captures between 2013 and 2018 (Figure 3).[9] We observed that the collection itself was extremely large (almost half a petabyte and accelerating in size). It became increasingly clear that we would not be able to analyse the entire corpus without moving our computation closer to the data.

Nonetheless, our agreement with IA did not include the use of in-house computational resources, outside of what was required to transfer the WARC data out of their infrastructure. However, we received a start-up grant from the NSF XSEDE program (xsede.org), which provides Internet2 connectivity (10 Gb/s), as well as a highly scalable storage and analysis platform using their Apache SPARK cluster (247 TB of RAM) and Jupyter Notebooks. However, we were unable to fully exercise the XSEDE infrastructure because we were ultimately bound by the Internet Archive’s outbound traffic rate of 280 kB/s - meaning it would have taken *50 years* to download the full dataset. Rather than framing this purely as a limitation of the IA infrastructure, this constraint is emblematic of many types of ‘data frictions’ that are encountered when performing data analysis in today’s distributed computing environments (Edwards et al., 2011). While the metaphor of ‘the cloud’ encourages us to think of data as instantly transportable and untethered from physical environments, this was a reminder that data has locality and material contingencies (Hu, 2015).

The intractability of localising the entire SPN dataset freed us from a purely quantitative analysis geared towards the production of *representative* claims about the SPN dataset as a whole. Instead, the network constraint or data friction invited us to play and experiment with the WARC data as a means to critically reflect on the ‘black boxing’ of SPN as *infrastructure*, and not simply as a singular black box (Bucher, 2017). To further this new mode of enquiry we chose to sample the same day (October 25) for each year that the SPN had been in operation (2013–2018). The date was chosen somewhat arbitrarily, but also to roughly coincide with the initial opening of the SPN service, which we discovered when analysing the IA forums (anigbrowl, 2013). The resulting dataset took 1 month to collect and store on the XSEDE platform, where we were able to run data analysis experiments with the recognition that we were only looking at a proportionately small amount of data (relative to the size of the collection). With this in mind, we began speculating and conceptualising about how SPN had been assembled, rather than attempting to make claims about the SPN data as a whole. Our investigation focused on opening up the network of relations that SPN was a part of, and as we unravelled some of these relations it was this mirroring process, or rather, how SPN *sees* the Web that took centre stage.

Seeing like a web archive: Investigating SPN as infrastructure

As a starting point, examining the SPN WARC data required us to first acknowledge that web pages are dynamic *hypermedia* objects.

> The logic of hypermediacy acknowledges multiple acts of representation and makes them visible [… and] offers a heterogeneous space, in which representation is conceived of not as a window onto the world, but rather as "windowed" itself--with windows that open onto other representations or other media (Bolter and Grusin, 1996).

The experience of viewing a web page is in fact a visualisation that is assembled by web browsers that request and mediate interactions between web resources (including interlinked HTML, CSS stylesheets, images, video and JavaScript source code) and user interactions with the Web. Consequently, a request to archive a single web page could result in writing tens, hundreds or thousands of resources to a WARC file. The number and types of these representations depend not only on how the web page is published, but on how the resource is viewed by the browser.

By this logic, the SPN data in the Live Web collection is not simply the output of a well-understood downloading process. Instead, it is the result of a situated *way of seeing* the Web that is encoded into the SPN algorithm and run by the IAWM's server infrastructure. Whilst the same could be said of other web archiving tools, as we will discuss further, SPN diverges *sociotechnically* both through its reliance on crowdsourced users from outside the organisation, and the myriad ways in which they make use of the SPN tools in practice. Additional complexities are introduced through the multiple routes that SPN crawls can be triggered, and the technical implications this has for how the crawler 'sees', crawls and archives the Web.

Due to this complexity, we were confronted with the challenge of how to read the SPN data, specifically with the purpose of making legible the power relationships embedded in 'how it works' and 'who it works for' (Galloway, 2004: xiii). Furthermore, this way of seeing the Web is not an esoteric concern since it directly impacts the fidelity of the archive record. For example, if parts of the hypermedia object were not collected by SPN, then the 'replayed' web page could have a broken layout (Brunelle et al., 2014), present a temporally shifted resource (Ainsworth et al., 2014), or even be intentionally manipulated or 'hacked' to misrepresent the past (Lerner et al., 2017**.** Understanding the shape of SPN as a knowledge infrastructure required a nuanced reading of the WARC data it produced with the aim of ascertaining how SPN itself viewed the Web.

With this in mind, the following describes our approach to 'seeing like a web archive' by iteratively zooming in and out of the SPN algorithm and the data it produces. The next sections describe how we moved from using conventional 'distant reading' (Moretti, 2013) tactics to the use of an 'experimental probe', designed to trace and reveal the operation of SPN in practice and at different scales.

Zooming out: Probing

Our initial plan was to use a set of analytic tools known as the Archives Unleashed Toolkit (AUT) to analyse the SPN WARC data (Ruest et al., 2020). AUT was specifically designed to help scholars conduct research with web archives, enabling users to batch process and 'filter, extract, aggregate and visualize' data using network diagrams and other reporting tools (Ruest et al., 2020). Optimised for distant reading (Moretti, 2013), the toolkit was helpful for characterising aspects of the SPN sample data, getting to grips with the SPARK programming language and working with data at scale. However, whilst AUT provided a useful starting place, the toolkit was not designed to address probing questions that require a close reading of the inner workings of WARC data. In many ways AUT forced us to recognise that in order to understand the mediating role of the SPN algorithm we needed a different strategy.

We therefore changed tack and embarked on an *experimental* approach to the SPN data, while documenting our conversations in meeting notes, Jupyter Notebooks and Slack. Here, we shifted from a data science or 'big data' approach towards adopting Bucher's (2016a) stance of the *technographer* where algorithmic systems are understood to be 'complex, diffuse and messy', but where, despite their opacity, black boxes can still be made to talk. For Bucher (2016a: 86), *technography* is 'a way of describing and observing technology in order to examine the interplay between a diverse set of actors (both human and non-human)'. We attempted to reverse engineer SPN software by analysing how the tools responded to various stimuli, in order to better understand how the SPN data came to be.

We first used our own subjective experience as SPN and IAWM users, to hypothesise that there were (at least) three avenues by which users could instruct SPN to archive specific web content: directly through the web form, using the IAWM browser plugin, and using SPN as an API from a programme or bot. We then devised a *probing* experiment to exercise SPN using each of these three modes of access (web form, browser plugin and bot) employing a uniquely encoded URL that:

1. pointed at a web server *hostname* within our purview, so we could examine the web server logs to see how and what resources SPN requested
2. referenced a web page that contained a JavaScript *application* that fetched additional resources when executed
3. contained a query parameter with a unique 'cache busting' *timestamp* (t) that (we hoped) would ensure that SPN actually fetched the URL instead of returning a previously cached snapshot
4. contained a query parameter (ua) that recorded the *mode of access* used (ie web form, browser extension or bot)

So, for example:

http://www.example.com/research/?ua=spn-probe&t=20181024140003

Each WARC file in the Live Web collection bore a filename which included a timestamp (eg live-20180313221948-wwwb-app4.us.archive.org.warc.gz). But it was unclear if these timestamps were when a) the specific datetime a request was received by SPN; b) the request was actually processed; c) the response data was collected into the file or d) something else altogether. To accommodate for this uncertainty, our experiment scheduled probes to run the day before, the day of, and the day after 25 October 2018 so we could increase the chances of seeing the results in our sample dataset. We used a spreadsheet to plan the experiment and log each action, its corresponding URL and any observations. A programme was also written to perform the automated API testing on a schedule (Summers et al., 2021b).

Zooming in: Tracing

In contrast to the distant reading approach, the results of our SPN probe enabled a close reading of a small number of WARC records. We moved from an analytical mode of 'big data', where the identification of patterns and broad trends were the focus, to one of ‘small data’ where the actors, context and network of relations were of primary interest (Abreu and Acker (2013)). A critical component of this close reading was achieved by correlating the WARC data with the SPN request traces in our web server logs. Each request to SPN generated an entry in the logs, for example:

```
207.241.225.226 - - [ 25/Oct/2018:23:10:59 +0000] "GET http://www.example.
com/research/?ua=firefox&t=20181025230000 HTTP/1.1" 200 86976 "https://web.
archive.org/" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10.14; rv:61.0)
Gecko/20100101 Firefox/61.0"
```

Log messages detailed how a SPN request was made including

- the requester’s IP address,
- the timestamp when the request was received,
- the URL requested,
- the referring URL (the URL the client was viewing prior to the request)
- and the User-Agent string that identifies the client (in this case, the Firefox browser that was used to conduct this probe).

Here, the single log message for requesting the HTML web page was followed by 116 other log messages to fetch the CSS, JavaScript and image files used to render the HTML.

In addition, the corresponding request could be found in the SPN WARC data that we downloaded for 25 October 2018. While many analytic tools for web archives are geared towards analysing the retrieved content stored in *response* records (eg the web resources being archived), the WARC standard also allows for the crawl *requests* themselves to be stored as discrete records. This allowed us to follow who and how the request for the crawl was made, providing details on what web clients requested and how this differed between bots and browsers. For example, using the Tracery Jupyter Notebook (Summers et al., 2021a), we used XSEDE to locate the ‘needle in the haystack’ (our single request in the 1354 gigabytes of WARC data collected on October 25). It looked like this:

```
GET /research/?ua=firefox&t=20181025230000 HTTP/1.1Accept:text/html,ap-
plication/xhtml+xml,application/xml;q=0.9,*/*;q=0.8 Accept-Language: en-
US,en;q=0.5 User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10.14; rv:
61.0) Gecko/20100101 Firefox/61.0 Via: HTTP/1.0 web.archive.org (Wayback
Save Page) Referer:https://web.archive.org/ Connection: close Host:www.example.com
Accept-Encoding: gzip,deflate
```

This revealed some important details that were not tracked in our server logs. These HTTP headers list and rank a client’s preferences for how to receive each request, but they are important because they shape how the server responds through a process known as Content Negotiation (Fielding and Reschke, 2014). For example, a web server may respond with a PDF file instead of an HTML web page if a client prefers ‘application/pdf’ to ‘text/html’. The significance of Content Negotiation as a way of ‘seeing the Web’ came into focus when we compared the traces between the browser and bot probes. The web server log for the bot probe looked like this:

```
207.241.227.104 - - [ 24/Oct/2018:14:10:15 +0000] "GET http://www.example.com/
research/?ua=spn-probe&t=20181024140003 HTTP/1.1" 200 86944 "-" "spn-probe"
```

With the following corresponding WARC request record:

```
GET /research/?ua=spn-probe&t=20181025040002 HTTP/1.1Accept: */* User-Agent: spn-probe Via: HTTP/1.0 web.archive.org (Wayback Save Page) Connection: close Host:www.example.com Accept-Encoding: gzip,deflate
```

We noted that the Accept header sent by SPN in the bot probe was '*/*' which indicates 'no preference' for a particular media type. This difference allowed us to surmise that the Accept header is not 'hard coded' into the SPN application itself, instead SPN simply relayed the client's Accept header preferences to the server. This indicates that how SPN *sees* and *archives* the Web is dependent on *how* a particular web client or browser sends the request, which in turn has significant implications for how a given web resource is represented in the IAWM.

The consequences for the fidelity of SPN web archives were further compounded by the observation that different probes also generated different numbers of requests in our web server logs. Our browser-based probe (which used the SPN web form) generated 116 requests to fetch and archive the various resources (the CSS, images, etc.) needed to view the web page, whereas the bot probe only generated *a single request*. This means that the IAWM could contain a very low fidelity web archive for bot-driven SPN requests, since it may lack the images, CSS and JavaScript needed to fully render the page. Therefore, the rendering of archived resources requested by a browser could be staggeringly different from those requested by a bot - which has potentially significant implications for the fidelity of SPN-archived resources and their use in future. With this in mind and returning to our framing of web archives as a form of *knowledge infrastructure*, this finding is significant as it points to the technical complexities of web archiving, but also disrupts commonly held 'epistemic assumptions' about how the IAWM and SPN work in practice, and potentially unsettles the trust placed in this technical architecture by its end-users (Ben-David and Amram, 2018: 183). Our findings suggest that here (and in this version of SPN), the entanglement of non-human actors or bots in the selection and preservation of web resources ultimately has detrimental consequences for the 'representativeness' or quality of the web archive in relation to the original resource, and reinforces the observation that both *how* the Web is archived and *who* is doing the archiving has consequences for the ways these archives are used to produce knowledge about the world in future (Ogden, 2022).

Zooming out again: modelling

With this new *situated* knowledge of how SPN translated archive requests into seeing and archiving web resources, it was necessary to zoom back out again to speculate about some broader trends at work. One especially salient feature of the WARC data was the footprint left by User-Agent strings. When doing a close reading of the 25 October 2018 WARC data, we observed that User-Agents were passed to the web server logs and not 'hard coded' into SPN. We then hypothesised that the User-Agent data could be used to investigate the different types of users (bots or browsers) making SPN requests across each year.

In the User-Agents Notebook (Summers et al., 2021a) we extracted all the User-Agent strings from the 1354 gigabytes of WARC data and then counted them. For the 30,235,173 HTTP requests that were sent for the 6 days of our sample, we found 93,843 distinct User-Agent strings. While the top 50 User-Agents per year accounted for the majority of the requests, there was a significant longtail of other User-Agents in the data (Figure 4). Classifying thousands of User-Agents by whether they were a browser or a bot was simply not feasible.

Whilst spot checking the top 50 User-Agents for each day, we noticed some unusual patterns. In 2013, there was only one User-Agent used that uniquely identified the SPN service. During 2014–2015 the User-Agents of different browsers appeared to be used with ‘(via Wayback Save Page)’ inserted at the end. And finally, in 2016–2018 the User-Agent assumed the form that we had observed in our own probing of the service. This shifting behaviour suggested that the SPN software was not static; SPN’s way of seeing the Web had changed over time, and therefore the way it archived WARC data had presumably changed too.

Continuing the experiment a bit further, we used the ua-parser project’s database and Python extension[10] to parse these User-Agent strings in order to extract the *family name* of each agent. For example, the family name of this User-Agent is ‘Safari’:

```
Mozilla/5.0 (Macintosh; Intel Mac OS X 10_13_6) AppleWebKit/605.1.15
(KHTML, like Gecko) Version/12.0 Safari/605.1.15
```

Counting the families of User-Agents yielded two orders of magnitude less family names (n = 934) (Figure 5). Fortunately, the top 50 User-Agents for each year accounted for almost all of the requests, with only 103 distinct families in the aggregate of each year. We then easily scanned the results to find obvious bots (e.g. PHP/5.4, Python-urllib, curl) and reran the analysis to observe how many probable bot requests there were (Figure 6). Bots are often engineered to pretend to be browsers to escape detection by changing the User-Agent HTTP header that is sent as part of the request. However, we hypothesised that the reverse (where a browser is set up to appear like a bot) would be an unlikely scenario. In other words, this is a very conservative estimate of the prevalence of bot requests in our sample data – there could be many more. What stands out in these results is not

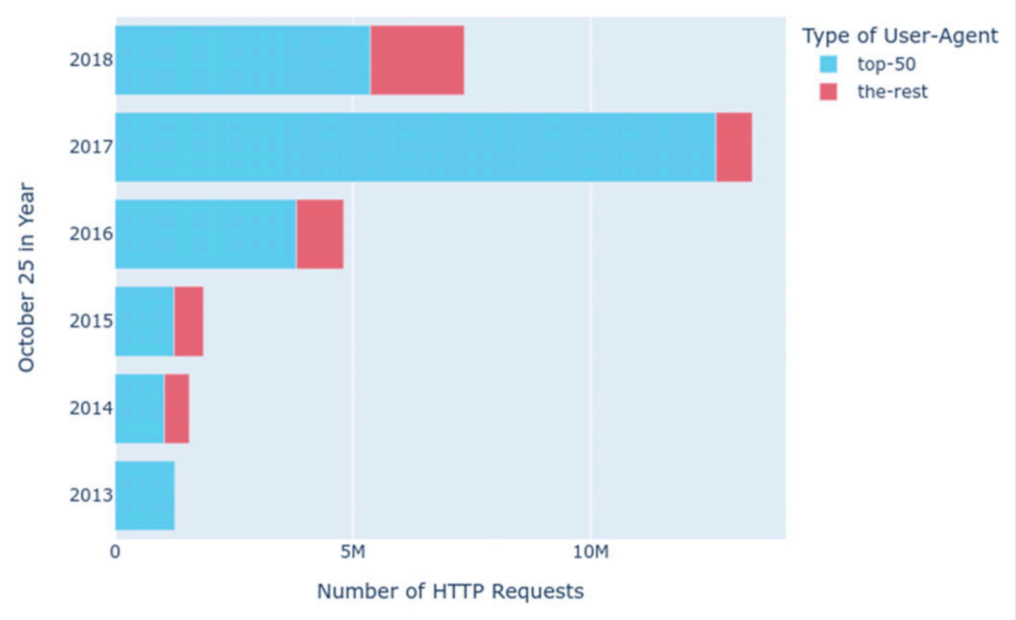


Figure 4. User-agent diversity.

only the significant presence of bots (particularly in 2017), but that based on what we learned in the probe tracing, bot requests are unlikely to result in high fidelity archival snapshots.

It's worth emphasising again that these are only provisional speculations based on a single day of SPN data over 6 years. It is possible that this approach would benefit from a more extensive or indeed random sampling of the SPN data to be able to say more about the prevalence of obvious bots in the SPN collection. It is also worth noting that this was just one example of the type of speculative modelling that we could have highlighted. The probe and process of zooming in and back out again was a generative exercise that spawned further questions about the nature and significance of the SPN collection for the IAWM as knowledge infrastructure. We also planned further mechanisms for characterising the collection, such as examining (for example) whether or not the requested content had been collected before, or if the archived URL was still available on the Web - findings that we plan to highlight in a future publication. These types of questions and methods for framing the contribution of SPN have clear implications for understanding the role of SPN in the diversity, range and fidelity of archived content in the Wayback Machine.

Discussion: Web archiving as critical technical practice

The details in the preceding section outlined a particular path of enquiry or process for understanding who or what was using the SPN service, and what, if anything, this meant for knowing SPN as a knowledge infrastructure. In this article we focus on SPN as a case study for experimenting with methods for investigating how this web archive operates in practice. A key part of this trajectory was a process of iterative *zooming out and zooming in* that unpacked SPN at macro and micro scales in order to trace the network of relations that comprise and envelop SPN. To summarise, starting with a zoomed out view of the SPN data corpus as a whole, we zoomed in on a single day (October 25) for

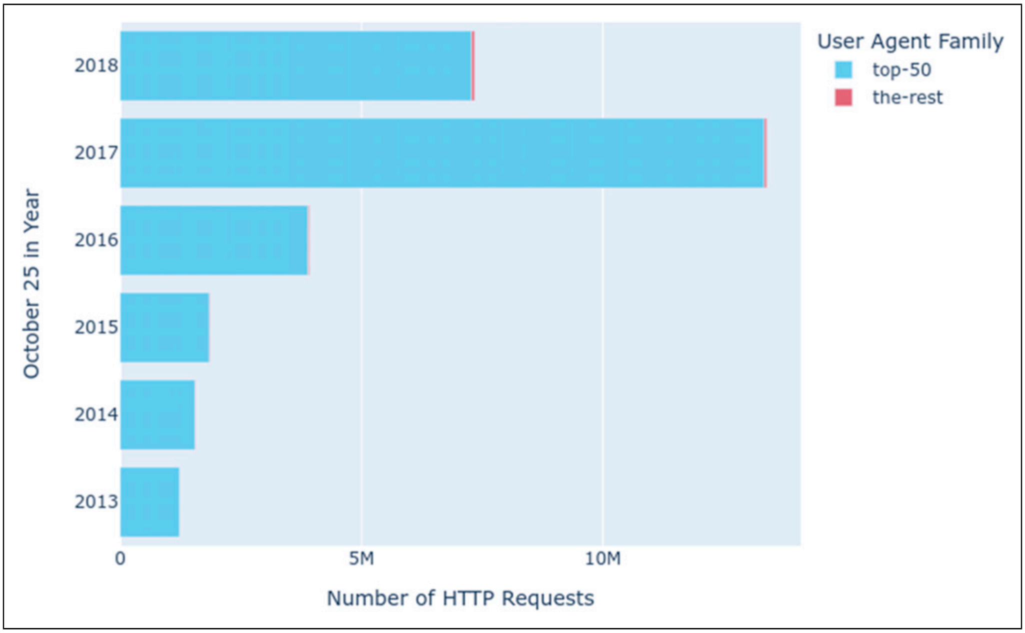


Figure 5. User-Agent family diversity.

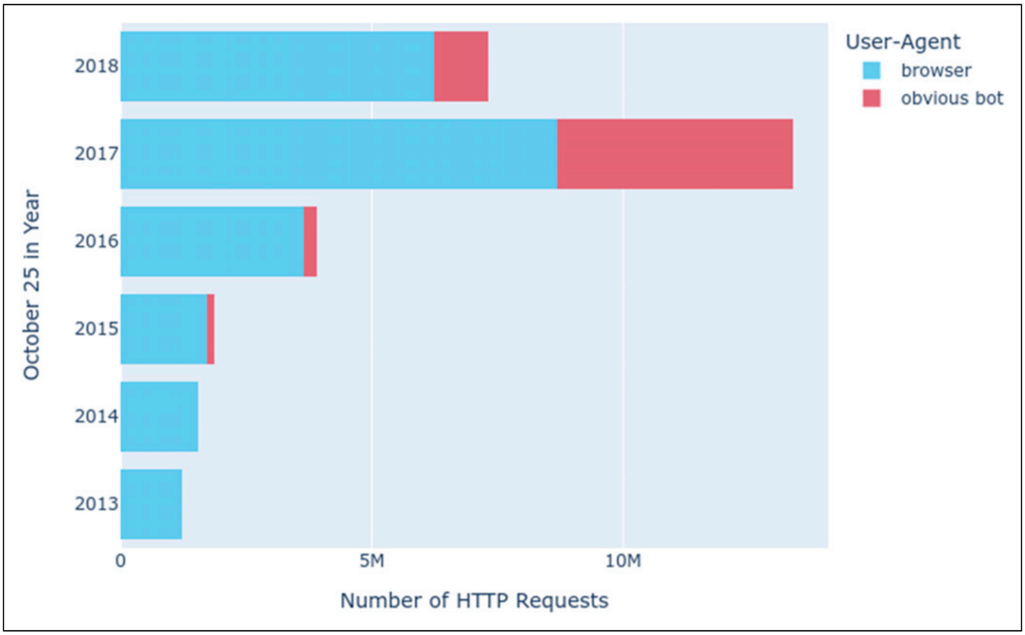


Figure 6. Obvious bots versus browser-based SPN requests[11].

six different years; then zoomed out using the AUT reports to present broad trends in the data; which prompted a zooming in to examine how SPN inscribed WARC records in response to various probes; followed by a zooming out again to make legible the browser-based and automated agents that were using SPN. While it would be tidy and convenient to present this process as ‘part of the plan’ from the beginning, it really only surfaced in the analysis of our field notes, documentation and computational notebooks as we struggled to understand SPN through the data it generated. It also became clear that the zooming out/in happened at critical junctures in the project where our attempts to understand SPN as an infrastructure were challenged.

Challenges

The project highlighted several challenges encountered in our attempt to clarify the hidden operations that underlie the SPN infrastructure. The project brought to the fore a complex layering of data and findings (black boxes inside black boxes) and alerted us to the ways that data abstraction often creates more questions and uncertainty. In general, we encountered three major challenges for studying SPN that we believe can be generalised to other forms of large-scale knowledge infrastructures that STS researchers must engage with. These challenges, or questions to ask of infrastructure, are not meant to supplant other theories or methods. Instead, they decentre the prioritisation given to the black box as a metaphor and as something which must be opened to understand *how* it works and *what* it is (Bucher, 2016a). These challenges and questions are meant to be generative, to foster a deepened engagement with critical technical practice, and enable a ‘critical understanding of the mechanisms and operational logic of software’; but they also prompt us to ask - as Bucher (2016a: 86–87) suggests – *who* these logics work for.

Critical technical practice also encourages reflection on who has access to study the inner workings of SPN and the implications this has for critically engaging the knowledge derived from using the Wayback Machine in practice. As discussed, our project was predicated on both the ethnographic work and a research agreement for accessing the WARC data, and whilst this was necessary for enabling our project it was also not sufficient for fully realising our original research aims. This said, researcher access to web archives data at scale is a long-standing and complex challenge within the field, and recent efforts by the IA and other web archiving institutions are making strides towards 'democratising' access. Nonetheless, examining the entirety of SPN is well beyond the capabilities of all but a handful of researchers due to the size of the dataset and the computational resources required to process data of this scale. At its current size of roughly 2.4 petabytes (Summers et al., 2021a) the SPN collection requires extremely large amounts of compute time to analyse. More significantly, the computation almost certainly must be brought to the data, rather than bringing the data to the computation.

As noted by Karasti et al. (2016) knowledge infrastructures like SPN and IA are geographically dispersed across multiple locations. In the case of SPN these locations included the many sites making archive requests, the site of SPN itself where the data is gathered and stored, the sites (or websites) that are being archived, and finally the site of analysis - which in our case was the XSEDE platform. This geographic distribution proved to be a barrier to understanding the relations which make up SPN as a knowledge infrastructure. Indeed, seeing these sites and the relations that connected them as networked, instantaneous and virtual worked as a form of *medial ideology* that obscured their actual material relations (Kirschenbaum, 2008). Being prepared for the complexity of answers that accompany the question of *where is infrastructure* is important for critical technical practice which must engage with the significance and legacy of digital knowledge infrastructures.

The second challenge we encountered was the lack of visibility into the processes that comprise the SPN data pipeline. The pipeline from submission, to collection, to archiving, to playback involved software and hardware components that we either had limited ability to inspect (in the case of SPN) or were so abundant and heterogeneous that they swamped us with complexity (for example, when considering the many types of agents interacting with SPN). Contacts established during the first author's ethnographic work at the IA proved critical for learning who to talk to, how to interpret what was said, what supporting documentation to look for, and ultimately how to craft entrée into the SPN data archive itself. Ethnographic data gathering, analysis and conceptualising were essential for addressing the 'double challenge' of understanding both technologies and their application domain. They also provided the context needed for our experimental probes that enabled insights into the otherwise opaque operations of the data pipeline.[12] Keeping an open and enquiring mind about *who is infrastructure* is a key element of critical technical practice.

A third challenge, as evidenced by our study, is that SPN is not a fixed entity, but rather it is a sociotechnical assemblage that is under regular revision and in constant motion. These shifting sands challenged our own attempts (and those of our ethnographic interlocutors) to understand SPN as knowledge infrastructure. Our close reading and tracing of User-Agents made it clear that SPN had been through several software incarnations which altered how it both saw and archived the Web. As we were completing our data gathering activities in 2019, we learned that yet a new version of SPN was in the process of being released (Graham, 2019). This new version altered yet again how SPN sees the Web since it had been engineered to use a server-side 'headless' browser to improve the fidelity and consistency of archived content. As the service came online and the use of the Live Web collection was subsumed by a new collection (called 'save-page-now'), the shift in how much data SPN collected was apparent (Figure 7). In the 2 years since the updated release SPN has collected more data from the Web than in all the previous 6 years of existence. Not only did the

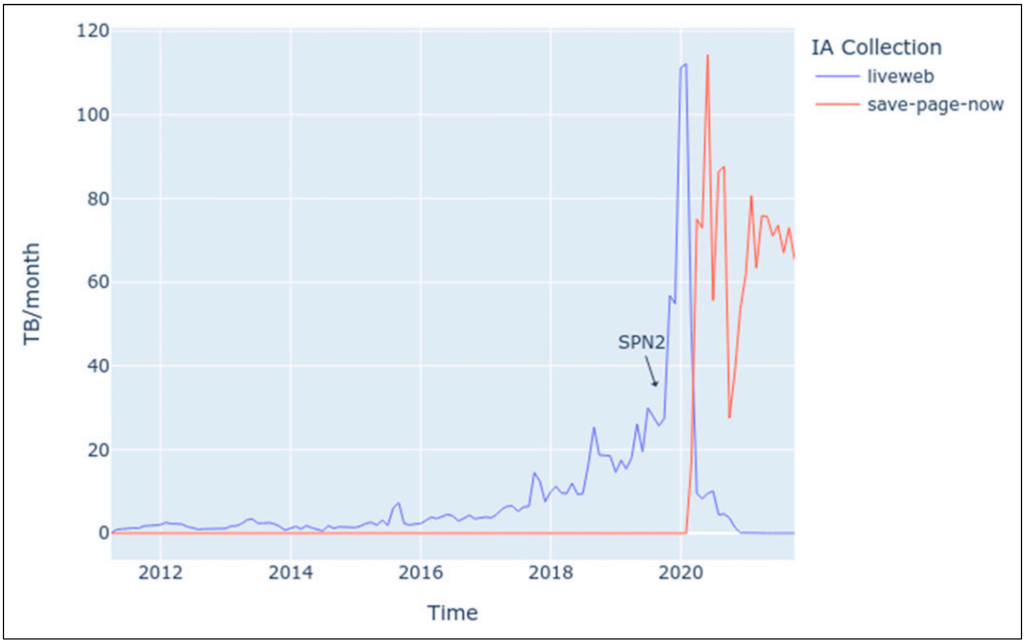


Figure 7. Data collected by SPN and SPN2 split across two internet archive collections over time.

underlying software and tools change over time, as our research demonstrates, so did the construction and fidelity of archival captures.

And finally, as an emerging body of work in the field of web archiving illustrates, the proliferation and embeddedness of web archives beyond the confines of the Internet Archive (Acker and Chaiet, 2020; Donovan, 2020; Kelion, 2018; Littman, 2018) necessitates new methods for studying and observing the implications of these archives for the circulation of 'historical facts' online (Ben-David and Amram, 2018). Whilst this article has focused on the methodological challenges of studying SPN, further work is needed to connect our initial findings on the operations of SPN to the wider workings of IAWM as knowledge infrastructure, and the specific types of knowledge(s) that it produces or precludes.

Conclusions

In conclusion, these challenges all beckon back to Star and Ruhleder's (1996) provocation to consider not just *what* but *when is infrastructure*? With the constant submission of new content by users and continuous development of SPN by the Internet Archive, SPN and large-scale web archival infrastructures, in general, can be said to be in a perpetual 'process of becoming' (Barad, 2003). And yet, despite the various ways these infrastructures elude observation, the situated and material arrangements of their creation are central to our understanding of both archival context and the innumerable versions of the Web's past which Save Page Now creates and re-presents through the Wayback Machine.

It is also useful to return to the title of the article and acknowledge the dual purpose that it serves. Although we have focused on the methodological design challenges of studying a particular form of knowledge infrastructure – or in other words, how we come to know infrastructures that often allude observation – it also bears reflection on the ways that these types of ‘knowing infrastructures’ also produce new and particular forms of knowledge. As the archived Web more generally is being used to underpin new machine learning algorithms (Birhane et al., 2021; Jo and Gebru, 2020), generate evidence for holding governments and public figures to account (Donovan and Lim, 2021; Rinberg et al., 2018; Taylor, 2014), and more – we invite future work that further outlines the relationship between how these infrastructures create new knowledge about the world, their implications and the new methodologies needed to understand them going forward.

Acknowledgement

This study would not have been possible without the support and data access provided by the Internet Archive Wayback Machine team. The authors would like to also thank the Convergence Special Issue editors and two anonymous reviewers for their extensive engagement with the article and constructive feedback.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. Specifically, it used the Bridges system through allocation ECS180012, which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center (PSC). Further support was provided to the first author to produce this article by the UK Economic and Social Research Council under *The Social Life of Web Archives* project, Grant No. ES/VO12177/1.

Data Access Statement

This study used third party data made available under a data sharing agreement with the Internet Archive which specified that the authors do not have permission to reshare this data. Requests to access SPN data should be directed to the Internet Archive at info@archive.org. We have made the code, notebooks and derivative data (from the probe experiment) associated with our analysis available through Zenodo in two open access repositories (Summers et al., 2021a, 2021b).

ORCID iDs

Jessica Ogden https://orcid.org/0000-0003-4696-7340
Edward Summers https://orcid.org/0000-0001-7320-8150
Shawn Walker https://orcid.org/0000-0002-7052-5705

Notes

1. https://web.archive.org/
2. https://twitter.com/brewster_kahle/status/994380510011928578 (Accessed: 16 October 2021)
3. See Karasti et al. (2016) for discussion.
4. For example, this can be seen through the proliferation of web archive links embedded in Wikipedia and web-based documents to reduce ‘link rot’ and unresolvable links on the Web (AlNoamany et al., 2014; Finley, 2019).
5. See also Ogden at al. (2017) for a discussion of the knowledge work which frames the inner workings of IAWM archiving practices.
6. https://archive.org/details/liveweb (henceforth referred to as the Live Web collection)
7. https://archive.org/details/liveweb

8. It is our understanding and experience that the research agreement was largely modelled on a standard research agreement template used by the IA for all web archive data projects involving researchers outside the organisation. Additional terms of the agreement were provided that were specific to our SPN project, defined by the aims and types of analyses we aimed to undertake, and specifying that: we would not replicate or make data available elsewhere, attempt to identify people or SPN users, and that we would credit and cite the IA in publications and adhere to their Terms of Use.
9. https://archive.org/services/docs/api/internetarchive/
10. https://github.com/ua-parser
11. Whilst we cannot fully interpret the meaning behind this finding within this scope of this article, we have speculated that the surge in bot requests in 2017 could be in relation to an uptick in SPN use as part of a range of grassroots web archiving initiatives that emerged in 2016–2017 in the US, including, for example, the Environmental Data and Governance Initiative's Data Rescue events that occurred during this time (Lamdan, 2018; Walker et al., 2018). However, this remains purely speculative at present and further research would be needed to substantiate and correlate this observation.
12. The need for oral histories and engagement with the humans behind decision-making processes in web archiving is something that has been noted elsewhere in studies of web archiving practices (Ben-David and Amram, 2018; Ogden et al., 2017; Summers and Punzalan, 2017).

References

Acker A and Chaiet M (2020) The weaponization of web archives: data craft and COVID-19 publics. *Harvard Kennedy School Misinformation Review* 1(3). DOI: 10.37016/mr-2020-41

Agre P (1997) *Computation and Human Experience*. Cambridge University Press.

Ainsworth SG, Nelson ML, and Van de Sompel H (2014) A framework for evaluation of composite memento temporal coherence. *arXiv:1402.0928 [cs]*. Available at: http://arxiv.org/abs/1402.0928 (accessed 9 October 2021).

AlNoamany Y, AlSum A, Weigle MC, et al. (2014) Who and what links to the internet archive. *International Journal on Digital Libraries* 14(3): 101–115. DOI: 10.1007/s00799-014-0111-5

anigbrowl (2013) Wayback machine gets a facelift, new features | Hacker News. Hacker News. Available at: https://news.ycombinator.com/item?id=6670546 (accessed 16 October 2021).

Barad K (2003) Posthumanist performativity: toward an understanding of how matter comes to matter. *Signs* 28(3): 801–831. Available at: http://www.jstor.org/stable/10.1086/345321

Beis C, Harris K, and Shreffler S (2021) The internet archive has been fighting for 25 years to keep what's on the web from disappearing – and you can help. *The Conversation*, 13 August. Available at: http://theconversation.com/the-internet-archive-has-been-fighting-for-25-years-to-keep-whats-on-the-web-from-disappearing-and-you-can-help-163867 (accessed 10 October 2021).

Ben-David A and Amram A (2018) The internet archive and the socio-technical construction of historical facts. *Internet Histories* 2(1–2): 179–201. DOI: 10.1080/24701475.2018.1455412

Birhane A, Prabhu VU, and Kahembwe E (2021) Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv:2110.01963 [cs]*. Available at: http://arxiv.org/abs/2110.01963 (accessed 11 October 2021).

Bolter JD and Grusin R (1996) Remediation. *Configurations* 4(3): 311–358.

Bowker GC and Star SL (1999) *Sorting Things Out: Classification and its Consequences*. Paperback. Boston, MA: MIT Press.

Bowker GC, Baker K, Millerand F, et al. (2010) Toward information infrastructure studies: ways of knowing in a networked environment. In: J Hunsinger, L Klastrup, and M Allen (eds) *International Handbook of Internet Research*. Dordrecht: Springer Netherlands, pp. 97–117. DOI: 10.1007/978-1-4020-9789-8_5

Brügger N (2016) Digital humanities in the 21st century. *DHQ: Digital Humanities Quarterly* 10(2).

Brügger N (2018) *The Archived Web: Doing History in the Digital Age*. Cambridge, MA; London, England: MIT Press.

Brunelle JF, Kelly M, SalahEldeen H et al (2014) Not all mementos are created equal: measuring the impact of missing resources. In: Proceedings of the 14th ACM/IEEE-CS joint conference on digital libraries, September 8-12, 2014, London, UK, 2014, pp. 321–330.

Bucher T (2016a) Neither black nor box: ways of knowing algorithms. In: S Kubitschko and A Kaun (eds) *Innovative Methods in Media and Communication Research*. Cham: Springer International Publishing, pp. 81–98. DOI: 10.1007/978-3-319-40700-5_5

Bucher T (2017) The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms', Information, Communication & Society, 20(1), pp. 30–44. Available at: https://doi.org/10.1080/1369118X.2016.1154086.

Caravaca MM (2017) The concept of archival 'sedimentation': its meaning and use in the Italian context. *Archival Science* 17(2): 113–124.

Donovan J (2020) Covid hoaxes are using a loophole to stay alive—even after content is deleted. Available at: https://www.technologyreview.com/2020/04/30/1000881/covid-hoaxes-zombie-content-wayback-machine-disinformation/ (accessed 28 June 2021).

Donovan J and Lim G (2021) The Internet is a Crime Scene. *Foreign Policy*. Available at: https://foreignpolicy.com/2021/01/20/internet-crime-scene-capitol-riot-data-information-governance/ (accessed 21 January 2021).

Eastwood T (1994) What is archival theory and why is it important? *Archivaria* 37: 122–130.

Edwards P, Bowker G, Jackson S, et al. (2009) Introduction: an agenda for infrastructure studies. *Journal of the Association for Information Systems* 10(5): 374. Available at: https://aisel.aisnet.org/jais/vol10/iss5/6

Edwards P, Mayernik MS, Batcheller Aet al. (2011) Science friction: data, metadata, and collaboration. *Social Studies of Science* 41(5): 667–690. doi: 10.1177/0306312711413314.

Edwards PN (2013) *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Paperback. Cambridge, MA; London, England: MIT Press.

Fielding R and Reschke J (2014) *Hypertext Transfer Protocol (HTTP/1.1): Semantics and Content*. 7231. Internet Engineering Task Force. Available at: https://datatracker.ietf.org/doc/html/rfc7231

Finley K (2019) The internet archive is making Wikipedia more reliable. Available at: https://www.wired.com/story/internet-archive-wikipedia-more-reliable/

Galloway AR (2004) *Protocol: How Control Exists After Decentralization*. Leonardo. Cambridge, MA: MIT Press.

Gehl RW (2017) (Critical) Reverse engineering and genealogy. *Le Foucaldien* 3: 4(1). DOI: 10.16995/lefou.26

Geiger RS (2017) Beyond opening up the black box: investigating the role of algorithmic systems in Wikipedian organizational culture. *Big Data and Society* 4(2).

Graham M (2019) The Wayback machine's save page now is new and improved. *Internet Archive Blogs*. Available at: http://blog.archive.org/2019/10/23/the-wayback-machines-save-page-now-is-new-and-improved/ (accessed 10 October 2021).

Hackett S and Parmanto B (2005) A longitudinal evaluation of accessibility: higher education web sites. *Internet Research* 15(3): 281–294.

Hu T-H (2015) *A Prehistory of the Cloud*. MIT Press.

Jenkinson H (1948) *The English Archivist: A New Profession*. London: H. K. Lewis.

Jo ES and Gebru T (2020) Lessons from archives: strategies for collecting sociocultural data in machine learning. In: *Proceedings of the ACM 2020 Conference on Fairness, Accountability, and Transparency* (ACM FAT*), 27-30, January 2020, Barcelona, Spain.

Karasti H, Millerand F, Hine CM, et al. (2016) Knowledge infrastructures: part I. *Science and Technology Studies* 29(1): 1–12. DOI: 10.23987/sts.55406

Karpf D (2012) Social science research methods in internet time. *Information, Communication and Society* 15(5): 639–661. DOI: 10.1080/1369118X.2012.665468

Kelion L (2018) IS propaganda 'hidden on Internet Archive'. *BBC News*, 15 May. Available at: https://www.bbc.com/news/technology-44112431 (accessed 11 October 2021).

Kirschenbaum MG (2008) *Mechanisms: New Media and the Forensic Imagination*. MIT Press.

Lamdan S (2018) Lessons from datarescue: the limits of grassroots climate change data preservation and the need for federal records law reform. *Law Review Online* 166(231): 231–248.

Lerner A, Kohno T, and Roesner F (2017) Rewriting history: changing the archived web from the present. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, Dallas Texas USA, 30 October 2017, pp. 1741–1755. ACM. DOI: 10.1145/3133956.3134042

Littman J (2018) Islamic state extremists are using the internet archive to deliver propaganda. Available at: https://medium.com/on-archivy/islamic-state-extremists-are-using-the-internet-archive-to-deliver-propaganda-a132597dd16 (accessed 12 November 2019).

Maemura E, Worby N, Milligan I, et al. (2018) If these crawls could talk: studying and documenting web archives provenance. *Journal of the Association for Information Science and Technology* 69(10): 1223–1233. DOI: 10.1002/asi.24048

McFarland DA, Lewis K, and Goldberg A (2016) Sociology in the era of big data: the ascent of forensic social science. *The American Sociologist* 47(1): 12–35. DOI: 10.1007/s12108-015-9291-8

McMillan Cottom T (2017) Covert (Ethical) Cobbling. Conference. Boston, MA. Available at: https://www.slideshare.net/tressiemcphd/covert-cobbling-2 (accessed 27 September 2019).

Milligan I (2019) *History in the Age of Abundance*. Montreal and Kingston; London; Chicago: McGill-Queen's University Press.

Milligan I, Ruest N, and Lin J (2016) Content selection and curation for web archiving: the gatekeepers vs. the masses. In: Proceedings of the 16th ACM/IEEE-CS on … (Query date: 2021-05-24 14:15:55). dl.acm.org. Available at: https://dl.acm.org/doi/abs/10.1145/2910896.2910913

Moretti F (2013) *Distant Reading*. London; NY: Verso.

Murphy J, Hashim NH, and O'Connor P (2008) Take me back: validating the wayback machine. *Journal of Computer-Mediated Communication* 13(1): 60–75. DOI: 10.1111/j.1083-6101.2007.00386.x

Nicolini D (2012) *Practice Theory, Work, and Organization: An Introduction*. First. Oxford: Oxford University Press.

Ogden J. (2020) *Saving the Web: Facets of Web Archiving in Everyday Practice*. PhD Thesis. University of Southampton. Available at: http://eprints.soton.ac.uk/id/eprint/447624.

Ogden J (2022) "Everything on the internet can be saved": archive team, Tumblr and the cultural significance of web archiving. *Internet Histories* 6(1–2): 113–132. DOI: 10.1080/24701475.2021.1985835

Ogden J, Halford S, and Carr L (2017) Observing web archives: the case for an ethnographic study of web archiving. In: Proceedings of WebSci'17, Troy, NY, USA, June 25–28, 2017, 2017, pp. 299–308. ACM. DOI: 10.1145/3091478.3091506

Pasquale F (2015) *The Black Box Society: The Secret algorithms That Control Money and Information*. Harvard University Press.

Quarles JL III and Crudo RA (2014) [Way]Back to the future: using the wayback machine in patent litigation. *Landslide* 6(3). Available at: https://www.americanbar.org/groups/intellectual_property_law/publications/landslide/2013-14/january-february/wayback-future/

Rinberg T, Anjur-Dietrich M, Beck M, et al. (2018) *Changing the Digital Climate: How Climate Change Web Content is Being Censored Under the Trump Administration*. 100 Days and Counting, January. Environmental Data and Governance Initiative. Available at: https://envirodatagov.org/wp-content/uploads/2018/01/Part-3-Changing-the-Digital-Climate.pdf (accessed 31 July 2019).

Rogers R (2013) *Digital Methods*. Cambridge, MA: MIT Press.

Rossi A (2017) If you see something, save something - 6 Ways to Save Pages in the Wayback Machine. *Internet Archive Blogs*. Available at: https://blog.archive.org/2017/01/25/see-something-save-something/ (accessed 10 October 2021).

Ruest N, Lin J, Milligan I, et al. (2020) The archives unleashed project: technology, process, and community to improve scholarly access to web archives. In: JCDL'20: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, Wuhan, Hubei, China, 2020, pp. 157–166. ACM. Available at: DOI: 10.1145/3383583.3398513

Seaver N (2013) Knowing algorithms. *Media in transition 8: Public media, private media*, May 2013. Available at: http://nickseaver.net/s/seaverMiT8.pdf

Seaver N (2017) Algorithms as culture: some tactics for the ethnography of algorithmic systems. *Big Data and Society* 4(2).

Shankar K, Eschenfelder KR, and Downey G (2016) Studying the history of social science data archives as knowledge infrastructure. *Science and Technology Studies* 29(2): 62–73. DOI: 10.23987/sts.55691

Star SL (1999) The ethnography of infrastructure. *American Behavioral Scientist* 43(3): 377–391. DOI: 10.1177/00027649921955326

Star SL and Ruhleder K (1996) Steps toward an ecology of infrastructure: design and access for large information spaces. *Information Systems Research* 7(1): 111–134.

Summers E and Punzalan R (2017) Bots, seeds and people: web archives as infrastructure. *The Computing Research Repository* abs/1611.02493. Available at: http://arxiv.org/abs/1611.02493 (accessed 5 December 2016).

Summers E, Walker S, and Ogden J (2021a) Save Page Now (SPN) Activity. Zenodo. DOI: 10.5281/ZENODO.5529672.

Summers E, Walker S, and Ogden J (2021b) Save Page Now (SPN) Probe. Zenodo. DOI: 10.5281/ZENODO.5529699.

Taylor N (2014) The MH17 crash and selective web archiving. Available at: https://blogs.loc.gov/thesignal/2014/07/21503/ (accessed 8 March 2019).

The WARC format 1.1 (2017) ISO 28500:2017. International Organization for Standardization. Available at: https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1/

Walker D, Nost E, Lemelin A, et al. (2018) Practicing environmental data justice: from DataRescue to data together. *Geo: Geography and Environment* 5(2): 1–14. DOI: 10.1002/geo2.61.

Abreu A and Acker A (2013) Context and collection: a research agenda for small data. In: iConference proceedings, February 12-15, 2013, Fort Wurth, Texas USA, 2013, pp. 549–554.