This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.

The Internet Archive and the socio-technical construction of historical facts

Anat Ben-David

Sociology, Political Science and Communication, Open University of Israel, Ra'anana, Israel

1 University Road, P.O.Box 808 Ra'anana 43537, Israel.

+972-9-7781147

ORCHiD: https://orcid.org/0000-0003-4510-5634

Twitter: @anatbd

Corresponding author: anatbd@openu.ac.il

Anat Ben-David is a senior lecturer in the department of Sociology, Political Science and Communication, and head of the Open Media and Information Lab at the Open University of Israel. Her research focuses on national web studies and digital sovereignty, web history and web archive research, and the politics of online platforms. Methodologically, her work specializes in developing and applying digital and computational methods for web research.

Adam Amram

Open Media and Information Lab, Open University of Israel, Ra'anana, Israel

1 University Road, P.O.Box 808 Ra'anana 43537, Israel.

Adam Amram is a scientific programmer. He holds an MSc from the department of Information and Knowledge Management at Haifa University. His research focuses on developing computational tools for web research.

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.

The Internet Archive and the socio-technical construction of historical facts

This article analyses the socio-technical epistemic processes behind the construction of historical facts by the Internet Archive Wayback Machine (IAWM). Grounded in theoretical debates in Science and Technology Studies about digital and algorithmic platforms as "black boxes", this article uses provenance information and other data traces provided by the IAWM to uncover specific epistemic processes embedded at its back-end, through a case study on the archiving of the North Korean web. In 2016, an error in the configuration of one of North Korea's name servers revealed that it contains 28 websites. However, the IAWM has snapshots of the majority of the .kp websites have been archived from as early as 2010. How did the IAWM accumulate knowledge about the .kp websites that are generally hidden to the world? Through our findings we argue that historical knowledge on the IAWM is generated by an entangled and iterative system comprised of proactive human contributions, routinely operated crawls, and a reification of external, crowd-sourced knowledge devices. These turn the IAWM into a repository who's knowing of the past is potentially surplus – harbouring information which was unknown to each of the contributing actors at the time and place of archiving.

Keywords: Internet Archive; Wayback Machine; North Korea; black box; provenance; appraisal; censorship;

WORD COUNT: 9548

Introduction

After years of stabilization in the reputation of the web as a reliable source of knowledge, recent events around the 2016 presidential elections in the United States have brought with them new questions about the epistemology and ontology of online materials: how do we know how to trust an online source? What tools do we have to

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.
distinguish between fact and fake? What are the knowledge processes behind the generation of what we see on the screen?

While these questions were typical of the web in its early years (Rogers, 2002; Jones 1999 ; Thelwall, Vaughan & Björneborn, 2005), their return since 2016 is an indication of the deterioration of the epistemic perception of the web as a primary source of knowledge. Questions about the epistemology of primary sources, however, are the foundations of source criticism in historical research (McNeill, 1986). Historians of all epochs and eras face similar questions when dealing with primary sources: how are the sources authenticated? What do we know about their authors? Is the source original or a copy?

While the majority of the complaints about fake online data primarily relate to content produced and disseminated via corporate social media platforms and search engines (Harsin, 2015; Kutcharsky, 2016), two of the web's non-commercial knowledge devices remain thus far relatively uncontested: Wikipedia and the Internet Archive (Brügger & Schroeder, 2017; Rosenzweig, 2006; Winters, 2017). Despite critique about certain bias in the editing of controversial entries, or about hidden power relations in the content management structure of the crowdsource Wikipedia (Ford & Wajcman, 2017), the fact that the history of all entries is both documented and transparent turn Wikipedia into both an encyclopaedia and an archive (Rogers, 2017). The Internet Archive, on its part, remains ideologically committed to the idea of free, universal and open access to all digital knowledge, the Web being among it (Kahle, 2007). However, compared to Wikipedia, the Internet Archive's specific content-creation and distribution processes are less transparent. Since 1996, the Internet Archive crawls and archives snapshots of large proportions of the open web and makes these available for viewing from anywhere in the world on the Internet Archive Wayback Machine (IAWM), which was introduced

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.
in 2001. Archived snapshots on the IAWM serve as legal evidence in judicial processes (Eltgrowth, 2009; Howell, 2006) and in digital journalism (Ryfe & Kelley, 2015).

Both Wikipedia and the Internet Archive are not monolithic epistemic devices. Previous studies have shown the complex socio-technical processes involving both the editing and governing of Wikipedia by human and non-human actors (Ford et al., 2013; Geiger, 2014; Niederer & Van Dijck, 2010). In this paper, we join with recent literature on the socio-technical epistemology of web archives (Schafer, Musiani, & Borelli, 2016; Summers & Punzalan, 2017) and identify the socio-technical epistemic processes behind the construction of the archived website on the IAWM as a digital evidence, or a historical fact.

In traditional archives – whether they are analogue, digital, or born-digital – archival practices ensure that the materials are organized and stored by their provenance and order of arrival in the archive (Featherstone, 2006). Anyone consulting the archive would not necessarily know the exact content of a box or a folder, but the box would be named after the name of the person or institution to which it belonged before being archived; The archivists would then add relevant metadata and describe the contents of the box or folder, according to what they think is relevant; and finally, the boxes or folders would be stored next to one another by the order of their arrival to the archive. Historians consulting archives, not knowing what is in a box, first ask themselves where to begin, by thinking about the personality or institution that would probably contain the relevant information they are seeking for their research (Duff & Johnson, 2002). Then, when finding specific documents, historians have authoritative documentation not only about the contents of the document, but also about the circumstances of its creation, the identity of their authors, and the exact dates of their arrival to the archive.

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.

Compared to the traditional archives and traditional archival sciences and practices, the IAWM does not behave as an archive. Niels Brügger (2016) coined the term "webrary", to refer to the hybridity of web archives as both an archive and a library. Yet until recently, when viewing an archived snapshot on the IAWM, the user knew very little about the circumstances that led to the preservation of that snapshot at that particular point in time. For example, has it been archived by a non-human agent (a crawler, or a bot), or by a human? And if it is a human contribution, what were this person's motivations? The circumstances could be endless: from an accidental archiving of a crawler that fetched a deep link or an external link from another website, through a website owner submitting their website for archiving, to people who use the IAWM's "save this page now" feature for various reasons. Lacking sufficient circumstantial context, we would argue, makes it difficult to ground an archived snapshot as historical fact. For although two snapshots can be identical in terms of their content (albeit archived at different timestamps), the knowledge production process behind their archiving might tell a very different story and might be initiated either by human or non-human actors.

In early 2016, the IAWM launched a new feature that provides new provenance information about archived snapshots. Under the "about this capture" button at the top of the screen, the viewer can learn about "the organization" behind the archiving, and the "collection" in which it is stored. In this paper, we use the IAWM's provenance feature at scale, and apply it on the case study of the national web of North Korea, hosted under the Country Code Top Level Domain (ccTLD) .kp. In mapping the various "organizations" and "collections" that contribute content about the rare national domain to the IAWM, we identify three intertwined socio-technical epistemic processes that generate archived snapshots as historical facts: (1) Proactive human contributions, (2)

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.
Routinely operated crawls, (3) Reification of external, crowd-sources knowledge devices. Through our findings, we argue that not only one cannot relate to the IAWM as a monolithic entity, but that the various and entangled ways in which human and non-human actors produce historical evidence is greater than its parts. That is, we argue that the Internet Archive's iterative system of tens of parallel crawls, joined by distributed automated and manual contribution by experts and lay-people, turn the IAWM into a repository who's knowing of the past is potentially surplus – harbouring information accumulated, reified and verified through entangled processes, which are impossible to generate individually at the specific point in time and geographic location of any source contribution to the IAWM. At the same time, we point at the setbacks and limits of the ability to completely drill down into specific circumstantial data about the archiving process, and the vulnerability of the IAWM itself to geopolitical constraints that are otherwise impossible to detect.

Theoretically, our work is grounded in debates in Science and Technology Studies about socio-technical knowledge production, and about the extent to which complex digital systems (such as algorithms or platforms) are considered black boxes worthwhile opening. We adopt a reflexive approach, trying to position ourselves somewhere in between those who argue that black boxes can and should be opened in order to unravel complex, human and non-human knowledge production processes behind such systems, and those who argue that instead of trying to open black boxes, epistemic processes can be studied through alternative methods. Our reflexive approach and narrative describe our attempts to understand a puzzling fact: while North Korea's ccTLD was delegated in 2007, until 2016 no one could estimate how many web addresses are hosted in the national web. In 2016, a DNS leak resulted in a surprising

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.
discovery that there are only 28 websites registered in the .kp domain, and their addresses had been exposed.

However, an examination of archived snapshots of the 28 North Korean websites on the IAWM reveals that as a knowledge repository, the IAWM had already "known" about the volume of the .kp domain, as 24 of the 28 websites had already been archived from as early as 2010. Our analytical inquiry therefore began with the question: how did the IAWM "know" about the existence and scope of the North Korean web, years before the DNS leak? How was this knowledge created and accumulated? And who are responsible for introducing the rare snapshots of the North Korean websites to the IAWM? As we will soon discuss, while we do not have definitive answers to these questions, our analytical journey has uncovered a complex knowledge production process which heavily relies on the proactive intervention of social actors, working in tandem with various automated and iterative crawlers, and which is highly dependent on larger geopolitical circumstances related to the physical location of the IAWM's servers.

The following sections of this article accordingly begin with a theoretical debate about the treatment of the IAWM as a black box, and whether or not "unboxing" the socio-technical epistemic processes that generate archived snapshots as historical facts is worthwhile. Subsequently, we introduce findings from our various attempts to use a "forensic social sciences approach" to extend our knowledge (to the best of our technical ability and skills) about the construction of snapshots of the North Korean websites on the IAWM. We conclude with an epistemological discussion about the ways with which deconstructing the IAWM's monolithic status strengthens the evidentiary and historical value of its archived contents.

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.

Is the Internet Archive a Black Box?

While initially developed as a concept for understanding scientific knowledge production (Knorr-Cetina, 1999; Pinch, 1992; Winner, 1993) in recent years the terms "black box" and "black boxing" have become central in debates in media studies and science and technology studies discussing the very possibility of knowing and understanding complex digital systems and the knowledge that they produce (Bucher, 2016; Kitchin, 2017; Latour, 1999 ; McFarland, Lewis & Goldberg, 2016; Paßmann & Boersma, 2017).

Overwhelmed by the complexity of algorithmic interfaces, code and data, which are often proprietary and therefore inaccessible, these debates present various solutions to knowing despite the existence of unopenable black boxes. Paßmann and Boersma summarize the positions in these debates by distinguishing between two concepts of transparency in dealing with (algorithmic) black boxes: Formalized transparency (the attempt to gain more "positive knowledge" on "the content" of a black box), and Practical transparency: the attempt to develop skills without raising the issue of openability (2017, p. 140).

Representing the "formalized transparency" side of the debate, authors such as Rieder and Röhle (2012) and Kitchin (2017) posit that the problem of "black boxing" can and should be solved (or at least mitigated) by using different analytical tools, examining code, or reverse-engineering algorithms to examine whether or not they render similar results. This, in return, may shed further light on the interpretive imperatives and the epistemological propositions embedded in black-boxed digital systems. With regards to the ways of understanding data produced by (black-boxed) digital platforms, McFarland et al put forward the notion of "Forensic Social Sciences"

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.
as a new research paradigm in Sociology, one that uses data produced by digital apparatuses not so much for building predictive-models, but instead for identifying patterns in the data and then tracing them back to meaningful analytical constructs (2016, p. 31).

On the other side of the debate, authors such Bucher (2016) represent the Practical transparency approach, by challenging the notion that the impossibility of opening the black box is an epistemological problem. Instead of attempting to open the black box, Bucher calls for alternative methods such as "technography", involving the inspection of the assemblage of documents, patent applications, blog posts and engineering conferences around a specific digital technology or platform (2016, p. 87). A second alternative that Bucher proposes to is to locate and examine the occasions offered by accidents, breakdowns, and controversies, as is done in studies of entangled socio-technical systems and infrastructures (Latour 2005; Marres 2012; Star, 1999).

Since the recent literature debating black boxes and "black boxing" of digital infrastructures, platforms, and systems primarily relates to proprietary algorithms, code, and the data that they produce, it might seem odd to examine the IAWM through a similar prism, especially given the vision of Brewster Kahle, founder of the Internet Archive to provide "universal access to all knowledge" (Kahle, 2007).

Indeed, compared to Wikipedia, which, next to the Internet Archive functions as one of the last non-commercial knowledge-organizing devices of the World Wide Web, the Internet Archive and the Wayback Machine are rarely subject to scholarly attention from a Science and Technology Studies perspective as an entangled socio-technical system. Perhaps it is Wikipedia's reliance on crowdsourced knowledge that accounts for the considerable scholarly attention paid to the entangled interaction between human

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.
and non-human actors in generating knowledge and facts (Ford & Wajcman, 2017; Ford et al., 2013; Geiger, 2014, Niederer & Van Dijck, 2010).

By contrast, and perhaps indicative of the widespread resonance of Brewster Kahle's enthusiastic ethos, the Internet Archive is often referred to as a monolithic entity, praised for its value as a unique resource for research (AlNoamany, AlSum, Weigle & Nelson, 2014; Niu, 2012; Witt, 2015). The projected agency assigned to the IAWM as a service, or as a valuable tool responsible for preserving websites is exemplified in a beautiful analogy by Dave Karpf:

> We can think of the Wayback Machine as a "lobster trap," of sorts. Lobster traps sit passively in the ocean, placed in areas of strategic interest. From time to time, one can check the traps and see if anything interesting has come up. The Internet is similarly awash in data that may be of interest to researchers. We often want to make across-time comparisons. But without lobster traps, we are bound to go hungry, so to speak (2012, p. 648-649)

Karpf's "lobster trap" metaphor harbours assumptions about the IAWM as a passive, rather than an active epistemic agent. It is passively "placed" in areas that other determine as strategic; and it is the role of others to determine whether or not it had "caught" something of interest. While we attempt to debunk such assumptions about the IAWM in the subsequent sections of this article, the point we wish to make here is that some of the epistemic assumptions about the IAWM reflect a certain level of trust in its technical architecture. This is not to say that critical literature on the Internet Archive does not exist. Some of the critical studies point at biases in archival coverage (Thelwall and Vaughn, 2004); Others point at epistemological challenges posed by the IAWM's interface, such as the difference between the live website and the static website in its archived environment, the structuring of the web's past through the prism of a single

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.

URL, or the fact that to date it is not searchable (Ben-David and Huurdeman, 2014; Brügger, 2009; Milligan, 2016; Rogers, 2013). Among the first to address Web archives as black boxes are Schafer, Muisiani and Borelli, who argue that "understanding a web archive implies opening several black boxes, the first being that of its collection, so as to understand the human and technological decisions which lead to its constitution, as well as the creation of this source which is never an exact copy of the original" (2016, p. 3). Along similar lines, Summers and Punzalan note that the practice of appraisal of web archives – of how content is created by a growing number of human and non-human actors – is currently understudied, and that automated agents such as crawl modalities, information structures and tools often serve as collaborators that act in concert with archivists and that play a significant role in selection decisions (2017, p. 3-5).

In the following sections of this article, we undertake a similar approach to that of Schafer et al. (2016) and of Summers and Punzalan (2017), yet we propose a different methodology for understanding the complex socio-technical content-creation processes of the IAWM, and for assessing the kinds of knowledge they produce. Our methodology is positioned between the "formalized" and "pragmatic" approaches to transparency and to the very possibility of opening black boxes. On one hand, we acknowledge our desire to "get to the source", and to enlighten parts of the knowledge-production mechanisms of the IAWM which are currently unknown to its end-users. In doing so, our methodology, which analyses and identifies patterns and traces in data produced by the IAWM, is more similar to the "forensic social sciences" approach (McFarland et al., 2016) than it is to the work of Schafer et al. (2016) and Summers and Punzalan (2017), whose findings are based on interviews with web archiving practitioners and users. On the other hand, as we discuss below, we acknowledge that

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.
the forensic approach is futile to a certain extent, for despite our efforts, many of our questions remained unanswered. The fact that provenance information and other metadata provided by the IAWM did not allow us to completely "reverse engineer" the IAWM's knowledge-production process in the case of the .kp domain, brings us closer to authors such as Bucher (2016) in reflecting on whether or not such "unboxing" is necessary to establish the ontological status of archived websites as evidence and facts.

The Wayback Machine and North Korea's Mosquito Network

Before the 2016 DNS leak, relatively little was known about the Internet in North Korea. One of the few published studies on information technology infrastructures and Internet policies in North Korea describe the country's efforts to establish a closely monitored information technology infrastructure, comprised of a strictly monitored intranet and a few propaganda websites, which were primarily intended to "leapfrog" the country's economic development (Chen, Ko & Lee , 2010, p. 3). Despite that, access to information technologies in North Korea is permitted only to government officials and members of social elites (Warf, 2014).

Until the delegation of the ccTLD .kp to North Korea in 2007 (IANA, 2007), the country hosted propaganda websites in Japan (i.e. kcna.co.jp). However, it was estimated that after the delegation of the .kp cctTLD most of the allocated IPs have not been used (Warf, 2014). In 2010, however, information about some websites, such as the North Korean Airlines (http://www.airkoryo.com.kp), started to show up. One source of this information was a leaked top-secret presentation of the United States National Security Agency, (NSA) from late 2014, which revealed Internet traffic volumes to North Korea, as well as the top 20 North Korean websites that have been tracked in a one-month period during 2012 (Koop, 2014).

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.

North Korea is not disconnected from the World Wide Web, but it restricts and monitors its network activity. A former North Korean who worked in the IT sector called North Korea's approach a "mosquito net", which lets foreign investments in, but keeps foreign culture and political ideas out (Bruce, 2012). It also allows North Korea to earn money while acquiring a channel to monitor those who are making it for them and reaps the gains from increased domestic efficiency without sacrificing social stability (Greitens, 2013).

Despite the strictness and the monitoring of the network traffic in North Korea, errors occur. On September 2016, one of North Korea's Domain Name Servers (DNS) was misconfigured. Even though the most common use of DNS is to translate user-friendly domain names into an Internet Protocol (IP) addresses, it also defines a mechanism to replicate the database between different DNS servers, known as DNS zone transfer. The misconfigured DNS server was detected by the TLDR (Top Level Domain Record) project (Bryant, 2016a). TLDR is an attempt to keep record of zone files for various TLDs and to monitor how these zones change over time. On 20 September 2016, the TLDR code performed the usual zone transfer request to the North Korean country's ns2.kptc.kp name server, but instead of getting the usual "Transfer failed" error, the name server answered with the national TLD registry of 28 websites (Bryant, 2016b). Mainstream news media, bloggers and discussion fora quickly began describing the North Korean web, translating and explaining the content of each website. Most of the reports commented on the outdated, brochure-like or "web 1.0" design of the North Korean Web. (AskReddit, n.d; Reddit.com, n.d.; Taylor, 2016). About two days after the DNS leak, the North Korean name server was reconfigured, and some of the discovered websites were no longer accessible.

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.

Where does the Internet Archive Get its Sources From?

When the news broke about the TLD leak, we checked if the IAWM has captures of the 28 North Korean websites, and to our surprise we saw that 18 websites have already been captured from as early as 2010 (see Figure 1).

Figure 1. The number of times each North Korean website was archived by the Internet Archive from 2010 to 2016.

It should be noted that in addition to the 18 archived websites, three additional websites are indicated as archived but their captures are inaccessible (silibank.net.kp, star.edu.kp, and sdprk.org.kp). Rep.kp, the website of the North Korean official newspaper of the Central Committee of the Workers' Party of Korea, mirrors three other URLs which appear on the archived websites list (http://www.rodong.rep.kp, http://www.vok.rep.kp, and http://www.gnu.rep.kp). The

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.
remaining four URLs, all ISP-related, are the only URLs that have no trace on the IAWM: http://star-di.net.kp, http://portal.net.kp, http://rcc.net.kp and http://star.net.kp. (See Table 1 in the Appendix.)

How did the IAWM "know" about the North Korean websites before the DNS leak? Our first step to answer this question is based on an assumption similar to Karpf's "Lobster Trap" metaphor. Surely, we assumed, they must have been accidentally captured by the IAWM's crawlers, which had discovered them by following links from other websites. Therefore, our first "go-to" method to answer this question was to map the hyperlink network between the discovered North Korean websites, to trace the possible hyperlink-following trail which may have led the IAWM's crawlers from one website to the others. To create the hyperlink network, we fetched every snapshot of every web page of every North Korean website from the IAWM and extracted all the links that point to different North Korean websites (see Figure 2). This method brought us to a dead end. It seems that most of the North Korean websites do not contain hyperlinks, and thus if a crawl fetches one website, it will not lead to finding more websites. From the "mosquito net" point of view, the fact that the North Korean websites do not contain hyperlinks seems like a tactic decision, so that if one website is made accessible to the World Wide Web, it will expose the other websites.

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.

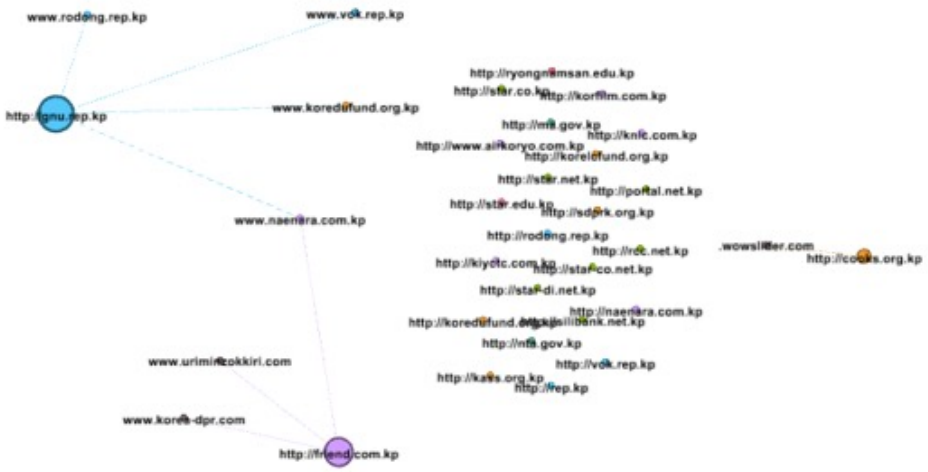


Figure 2. A Network of outbound links between the archived North Korean websites on the Internet Archives

If not through hyperlinks, how then has each North Korean website found its way to the IAWM? To answer this question, we used a third "forensic" method, this time using the IAWM's new provenance feature, which informs users about the "organization" behind the archived snapshot, and the "collection" in which it is stored. We scraped the provenance information of every snapshot of every .kp website and mapped all the organizations and collections that contributed their snapshots (see Figure 3). We also checked which organization introduced the first snapshot of each website to the IAWM (see Figure 4). The information presented in Figures 3 and 4 has helped us uncover and display the complex socio-technical processes behind what is eventually perceived as a single archived snapshot on the IAWM. Similar to Summers and Punzalan (2017), our findings clearly show that next to the IAWM's crawlers, proactive human contribution plays a significant role. As is evident in Figure 3, seven

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.

"organizations" contributed North Korean URLs to the IAWM, and the overall captures of these URLs are assigned to 31 different collections.

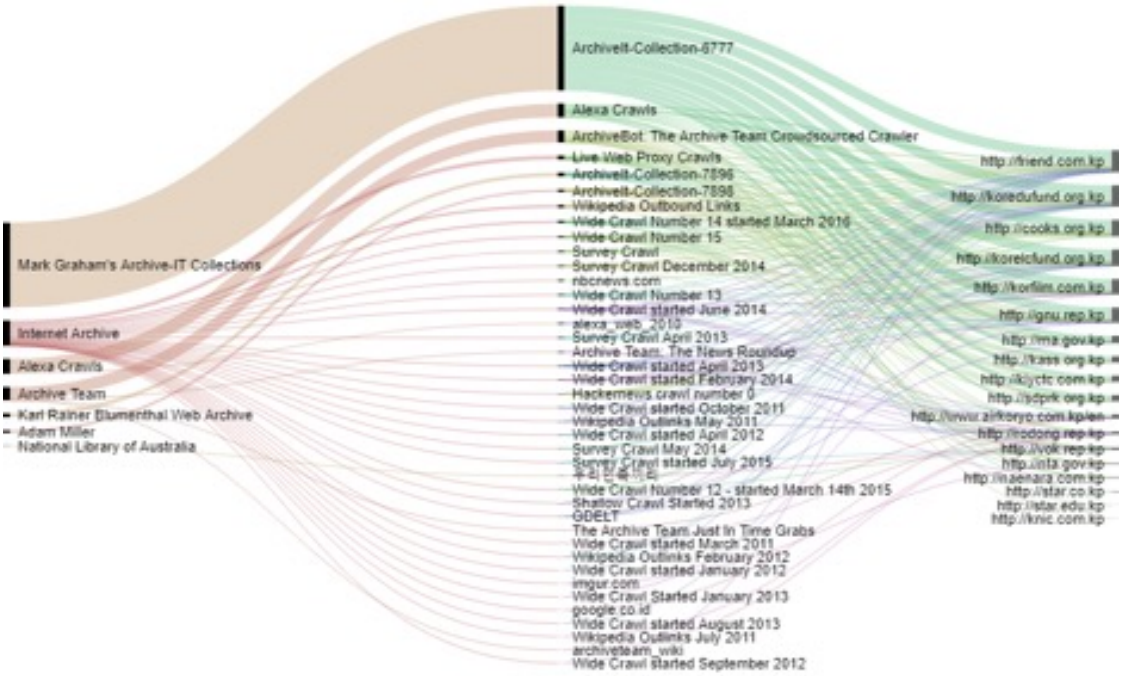
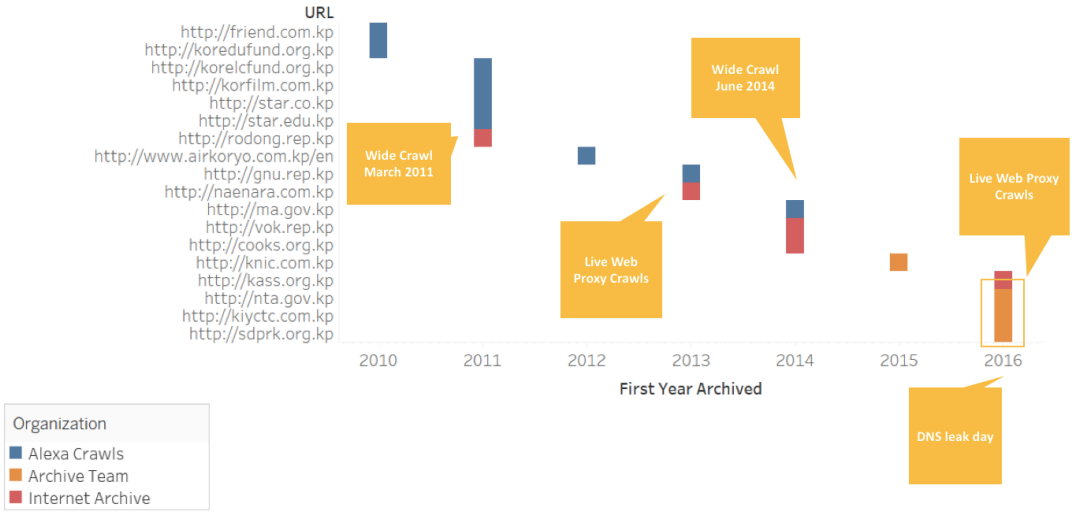


Figure 3. Distribution of organizations that contributed the North Korean websites to the IAWM, and their respective collections



This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.
Figure 4. The first snapshot of each North Korean website and their respective contributing organization.

A closer look at the specific organizations and collections behind the North Korean portion of the IAWM reveals a fascinating and complex epistemic system comprised both of automated procedures, individual expert knowledge, crowdsourced knowledge as well as activist interventions. Interestingly, Mark Graham, director of the IAWM, is the "organization" that contributed the most captures. The Internet Archive and Alexa Crawls are listed as two other organizations, but each refers to different data sets used by the IAWM for its iterative crawls. The Alexa Crawls refers to the donation of the index of Alexa Internet to the Internet Archive, which started in 1996 after Brewster Kahle sold Alexa Internet to Amazon and founded the Internet Archive (Roush, 2005). The information about the Alexa Crawls on the IAWM's website is rather cryptic, noting that these data flow in every day, after an (unspecified) embargo period. Yet it is important to note here that the Alexa Crawls are not simply copies of the search-engine index, for they also include pages users saved to the IAWM through the Alexa Toolbar (Rogers, 2017). Interestingly, the National Library of Australia is listed as an organization that contributed one capture. Two other "organizations" are named after individual web archivists: Adam Miller, an employee at the Internet Archive, and Mark Rainer Blumenthal, a Web Archivist at the Archive-It service associated with the Internet Archive. Finally, the "Archive Team" is a collective founded by Jason Scott in 2009, comprised of programmers, archivists, writers and activists dedicated to preserving digital history. Thus at least quantitatively, the contribution of .kp URLs to the IAWM by human experts, trained archivists and

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.
activists is far greater than the contribution of automated crawls based on initial seed lists. However, the source-contribution process is not a one-off event, as is evident in the numerous collections that host the various captures of .kp websites. Since the IAWM is an incremental repository, once a website is introduced to the archive, it becomes listed in several seed-lists, and is subsequently spread to several collections and re-captured by many crawlers. Of the 31 collections displayed in Figure 4, 12 are the Internet Archive's Wide Crawls which were started at different years. The IAWM's website further explains in general terms what wide crawlers are:

> Since September 10th, 2010, the Internet Archive has been running Worldwide Web Crawls of the global web, capturing web elements, pages, sites and parts of sites. Each Worldwide Web Crawl was initiated from one or more lists of URLs that are known as 'Seed Lists'. Descriptions of the Seed Lists associated with each crawl may be provided as part of the metadata for each Crawl. Worldwide Web Crawls are run using the Heritrix software.
>
> In addition, various rules are also applied to the logic of each crawl. Those rules define things like the depth the crawler will try to reach for each host (website) it finds. In general terms the crawling software will identify all the URLs on each page it captures, follow those links, attempt to capture those pages, identify new URLs, follow those links, etc., till the crawl is stopped or pre-set conditions like site depth limits are reached. For the most part a given host will only be captured once per Worldwide Web Crawl, however it might be captured more frequently (e.g. once per hour for various news sites) via other crawls (Rossi, 2010)

Although this information provides crucial data for understanding the IAWM as a knowledge device, it is interesting to note that important information is missing from the description. Specific details are replaced by general ones ("for the most part", "one or more lists", "description… may be provided"). As noted above, the description of each Wide Crawl does not provide neither the seed list nor the crawler's setting, which

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.
are crucial to understanding why a specific website has or has not been archived at a specific point in time. Moreover, it is worthwhile noting the epistemic terminology ("logic", "rules") used by the Internet Archive to describe the automated crawls. We know that the Wide Crawls use logic and rules, but these are at the moment unknown, or even "black boxed" to a certain extent. The collections associated with the Archive Team collective have particularly creative names, from the Archivebot (which is a crowd-sourced crawler), to the "News Roundup" and the "Just in Time Grabs" collections – all strongly indicating a crowd-sourced, proactive and interventional archiving mechanism. Finally, and perhaps most importantly, we find Wikipedia's back door into the IAWM. At the outset of the paper we compared Wikipedia to the IAWM as the two last non-commercial knowledge devices of the World Wide Web. In October 2016, the Internet Archive announced that in collaboration with the Wikimedia foundation it had monitored and automatically captured all new, and edited, outbound links from English Wikipedia for three years, which led to fixing 1 million broken URLs on Wikipedia (Graham, 2016). The continuing collaboration entails that Wikipedia's crowd-sourced epistemic culture, relying on authentication of knowledge and claims by references, is reified on the back end of the IAWM. The reification is reciprocal, for the IAWM is the 7th most cited source on Wikipedia, accounting for 0.45% of all outbound links (Ford et al., 2013).

The aggregation of the list of collections containing captures of the .kp websites harbours further revelations and dead-ends. Next to the Wide Crawls, there are Shallow Crawls that collect content 1 level deep, for which the data is not publicly available (Rossi, 2013); Survey Crawls, which are run twice a year and attempt to capture the content of the front page of every web host since 1996 (Rossi, 2012); and Live Proxy Crawls, which host content crawled from the IAWM's Save Page Now feature, a point

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.
to which we return (Rossi, 2011). The list reveals further collaborations between the Internet Archive and other Web-based knowledge repositories such as the online image sharing community imgur.com and the news website nbcnews.com.However, there is no description of these collections on the IAWM's website.

Finally, looking at the right column on Figure 4, another reification process is evident since the majority of the archived .kp URLs have multiple captures from various organizations and in different collections. The most frequently archived .kp website is friend.com.kp, the website of the Committee for Cultural Relations with Foreign Countries; The rarest capture at the time of study is knic.com.kp, the website of the national insurance corporation in DPRK.

Although the mapping of the various organizations and collection behind the contribution of the .kp websites to the Internet archive has revealed the collaboration between organizations, individuals, experts, users, and crawlers and external web based knowledge devices such as Alexa Internet and Wikipedia, each with their own epistemic logic and rules in all their richness and complexity, we were still unable to drill down the metadata to answer our appraisal question: who was the first to introduce each .kp website to the IAWM? To get closer to answering this question, we decided to look at the first capture of each .kp URL and note their contributing organization. As is evident in Figure 4, Alexa Web Crawl (containing both the search engine index as well as the pages users save via the Alexa Toolbar) is responsible for introducing the early snapshots of the .kp websites to the IAWM. However, we are unable to determine whether or not each first contribution was from the search engine's index, or from an individual user's contribution. Figure 4 also reveals the evolution of the IAWM's epistemic culture, after the introduction of the IAWM's "Save Page Now" feature, which was introduced in 2014 (Price, 2014), and the contribution of the Archive Team

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.
collective's collections to the IAWM. It is evident that from 2014 onwards, the volume of crowd-sourced users' contribution of new content to the IAWM increases, and culminates in 2016, especially around the DNS leak day. Thus, despite the complex, distributed archive comprised of various crawlers, seed lists and external sources, we can ascertain that proactive human intervention is responsible for introducing new content to the IAWM, at valuable and strategic points in time.

Our argument is further strengthened after an e-mail interview with Mark Graham, director of the IAWM. We mentioned above that Graham's Archive-IT collection makes the largest contribution of .kp websites to the IAWM. Since the provenance metadata provided by the current version of the IAWM does not answer the specific "how" question necessary for archival appraisal, we asked Mark Graham whether there is any documentation on the IAWM staff's decisions to increase the capturing pace of specific websites at a given point in time, which may explain surge in the number of captures around the leak day (see Figure 5). His answer was "No. In this the case the decision was me simply deciding to do it. I updated the seed list for an AIT-based crawl I had started in Jan 2016: https://www.archive-it.org/collections/6777." We further asked about his collection, and what were his sources for learning about the existence of the .kp websites. His answer was rather surprising: "Google search for terms like "North Korean websites". It so appears, that the director of the IAWM has a personal interest in North Korea, and an accordingly dedicated web archiving collection to document it.

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.

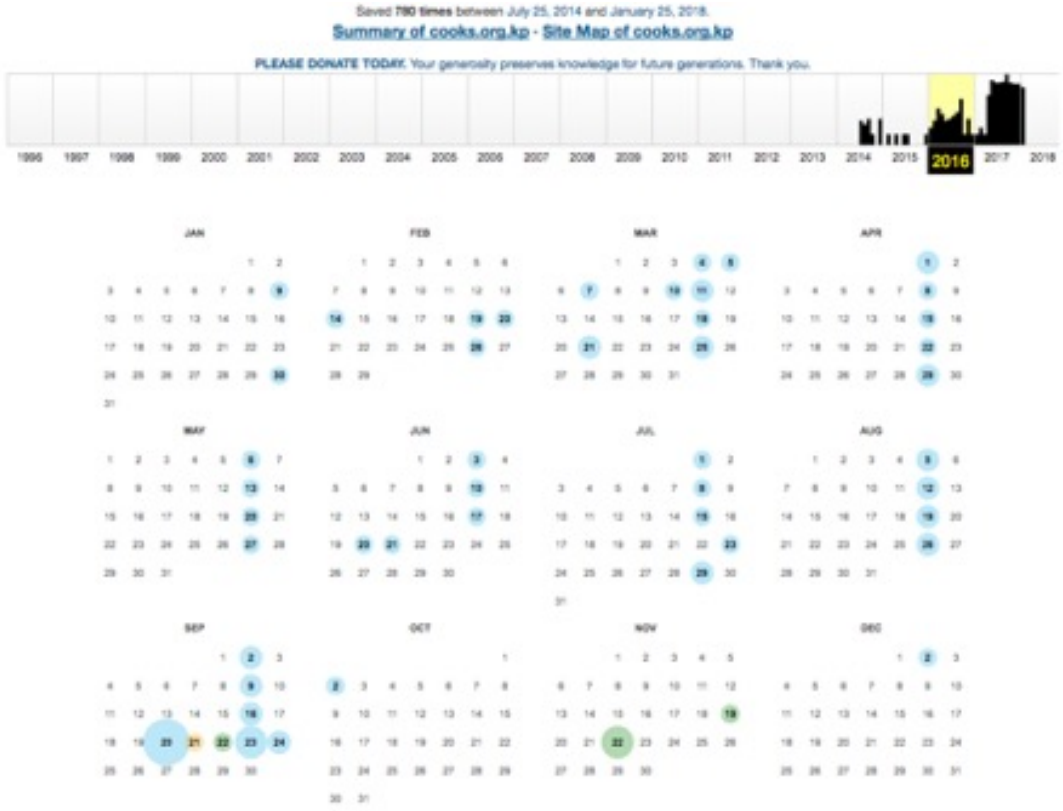


Figure 5. A snapshot from archive.org, displaying a calendar view of the number of captures of the North Korean website cooks.org.kp in 2016.

Is it possible that the North Korean websites were available on the live Web and retrievable through Google all along? Our analysis of the number of captures of the .kp websites on the IAWM in each year, summarized in Figure 1 and in Table 1, shows that for some websites, the archiving is inconsistent, displaying gaps in the years of the available captures. In the next section of this article we further problematize the stability and consistency of the IAWM captures of North Korea by tying it to the materialities and geographies of the IAWM. Before doing so, we would like to briefly summarize the IAWM'S epistemic processes we identified thus far in analyzing the case study of the North Korean web.

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.

Through analyzing the provenance data attached to each of the captures of the .kp websites in the IAWM, we got closer to understanding the complex epistemic culture behind the generation of historical evidence and facts on the IAWM. In doing so we have identified three epistemic processes, each involves different actors, technologies, and logic. The first epistemic process involves proactive human curation and intervention, either through Internet users saving pages through the Alexa Toolbar and the IAWM "Save Page Now" feature, or through individual curators such as Mark Graham and other Web archivists associated with the IAWM. The second process reifies the logic and content of external, authoritative web-based devices (The Archive Team and Wikipedia), which are built on crowd-sourced knowledge, and which are being automatically "fed" into the back-end of the IAWM. The third epistemic process is found in the work of the various automated and iterative crawls, which are built on top of a variety of changing seed-lists, and which have different (yet not publicly available) settings of depth, frequency, and rules of scope.

What Can be Learned from the North Korean Case Study on the Geopolitics of Web Archiving?

Earlier in this paper we declared our aim to reflect on our analytical attempts to understand knowledge-creation on the IAWM, through various "forensic" methods. By "reflection", we refer to reporting on our failures and unsuccessful attempts to get to the bottom of knowledge-processes, as well as to acknowledging our own biases and wrong assumption about our research materials. One of our biggest pitfalls was the assumption that the .kp websites were not available on the live web before and after the 2016 leak. Much to our surprise, Mark Graham's straightforward answer about using Google as a source for finding .kp URLs has made us realize that there might be differential access

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.
to .kp websites from different countries. Indeed, while our attempts to access the .kp websites from various Israeli ISPs always timed out, when we tried to access the same websites from proxies in the UK, most of them were suddenly accessible.

We previously noted in passing that the temporal representations of many of the snapshots of the .kp websites on the Internet Archive are inconsistent. Some snapshots may be available for a few years, followed by a gap of a year or two before the next capture (See Figure 1 and Table 1 in the Appendix). Could the same reasons that led us to wrongly assume that the .kp websites are inaccessible also affect their archival coverage on the Internet Archive? Put differently, do larger geopolitical processes contribute to or prevent knowledge creation on the IAWM?

To answer this question, we rely on recent studies on information infrastructures, which call for paying attention to the material, geographical and visible aspects of knowledge-creation in networked environments (Bowker, Baker, Millerand & Ribes, 2009; Starosielski, 2012). We therefore consider the physical location of the IAWM's servers in California, United States, as crucial for understanding the wider context affecting the IAWM as a knowledge device.

To understand whether or not the physical location of the IAWM's servers and crawlers in California is related to the level of archival coverage of the .kp websites over time, we used a method borrowed from Internet Censorship research. The method, developed by the Digital Methods Initiative at the University of Amsterdam (Rogers, 2013), uses a list of available proxies from various countries, and tests whether or not the same website can be accessed from a proxy of country known to have no restrictions on Internet access (such as the Netherlands), compared to proxies from countries that are suspected to apply various methods of Internet Censorship. The DMI method uses the error messages returned by the tested proxies, to assess the likelihood of restrictions

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.
on access to specific websites, from specific countries. While the error message most indicative of censorship is "403 Forbidden", the majority of error messages to inaccessible websites is "503 Time Out".

We 'reversed' the DMI method to analyse differential access rates to the .kp websites : instead of examining whether a list of websites is accessible from one country, we examined whether a list of websites from one country is accessible in 30 countries. Following our discovery that there is differential access to the North Korean websites from different parts of the world, our analysis does not examine internet censorship in North Korea, but rather a derivative of its Mosquito Network: instead of asking what is being censored inside North Korea, we are trying to identify whether or not North Korea does not let specific countries access its websites.

We therefore downloaded the same open proxy list used by the DMI Internet Censorship Tool (XROXY.com). Since the constantly-updated list of proxies from different countries contains many dysfunctional proxies, we first examined the availability of proxies for each of the 30 countries we studied, and only kept those that were available. Subsequently, we tried to access the list of the .kp website from each country's identified proxies (see Table 2 in the Appendix).

Our findings show significant differences in the accessibility of the North Korean websites from different countries. Countries that have relatively stable diplomatic relations with the North Korea (with reciprocal embassies) such as Russia and UK, display the highest access rates. Other countries known to have more conflictual ties with the North Korea, such the United States, display low access rates of about 7.6%. Note that our findings do not exclude the possibility that certain countries or ISPs are limiting access to .kp websites on their part. For example, our findings show

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.
strong indication of Internet censorship in Italy, where the majority of access attempts of .kp websites result in the definitive HTTP response "403 Forbidden".

While we are unable to confirm whether or not the IAWM's California-based servers were affected by possible IP filtering of the North-Korean websites, the data in Table 1 indicates that for some .kp websites, the number of inaccessible capture exceeds that of available ones, possibly indicating server timeouts. The website rodong.net.kp is a case in point. As indicated in Table 1, this news website has 204 inacessible captures, compared to 12 successful ones. The geographic distribution of the HTTP response codes to this websites shows that it is accessible in Europe, India, Afghanistan, Thailand and Mexico, but not from the United States, where the IAWM servers are based (see Figure 6). An overview of the 2016 captures on the IAWM's confirms that apart from the day after the DNS leak (blue circle), most captures are inaccessible due to server redirects (green circles). (See Figure 7.) Thus, even assuming that the URL was initially submitted to the IAWM from a country in which it was accessible, its future capturing depends on the geolocation of the IAWM's servers.

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.

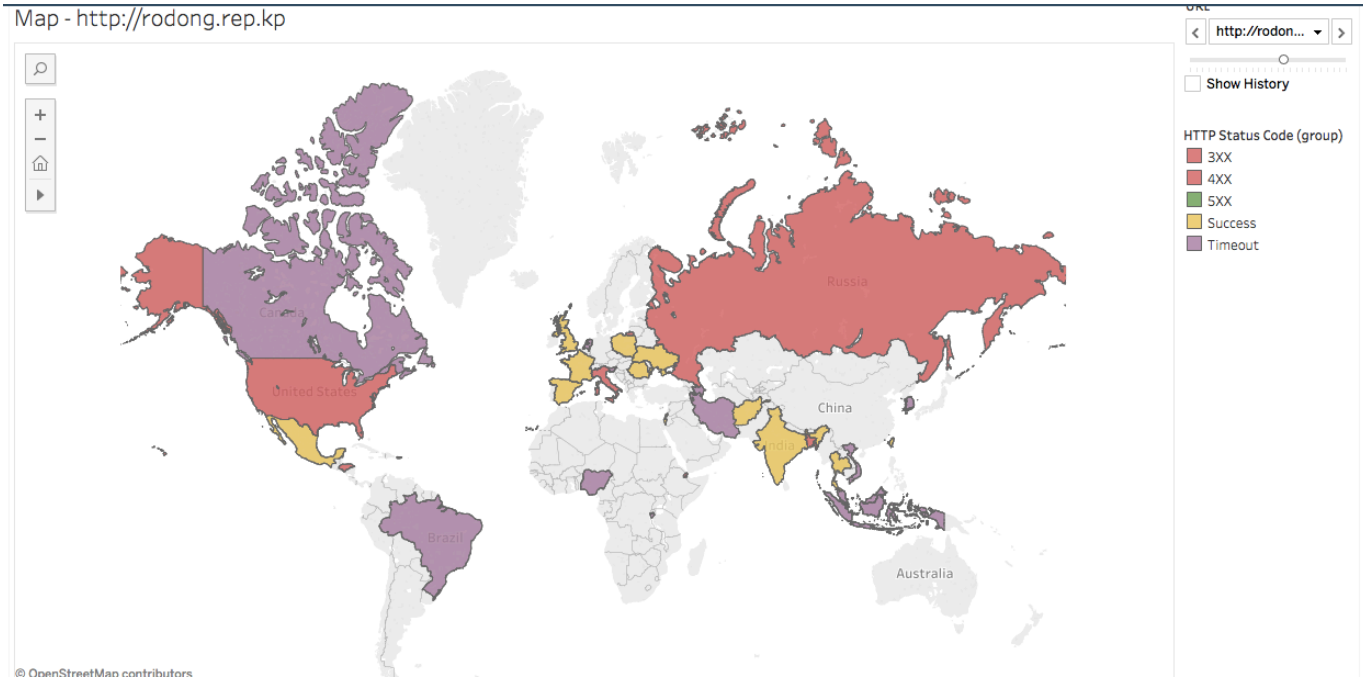


Figure 6. The Geographic distribution of HTTP response codes returned by attempts to access rodong.net.kp from 30 countries.

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.

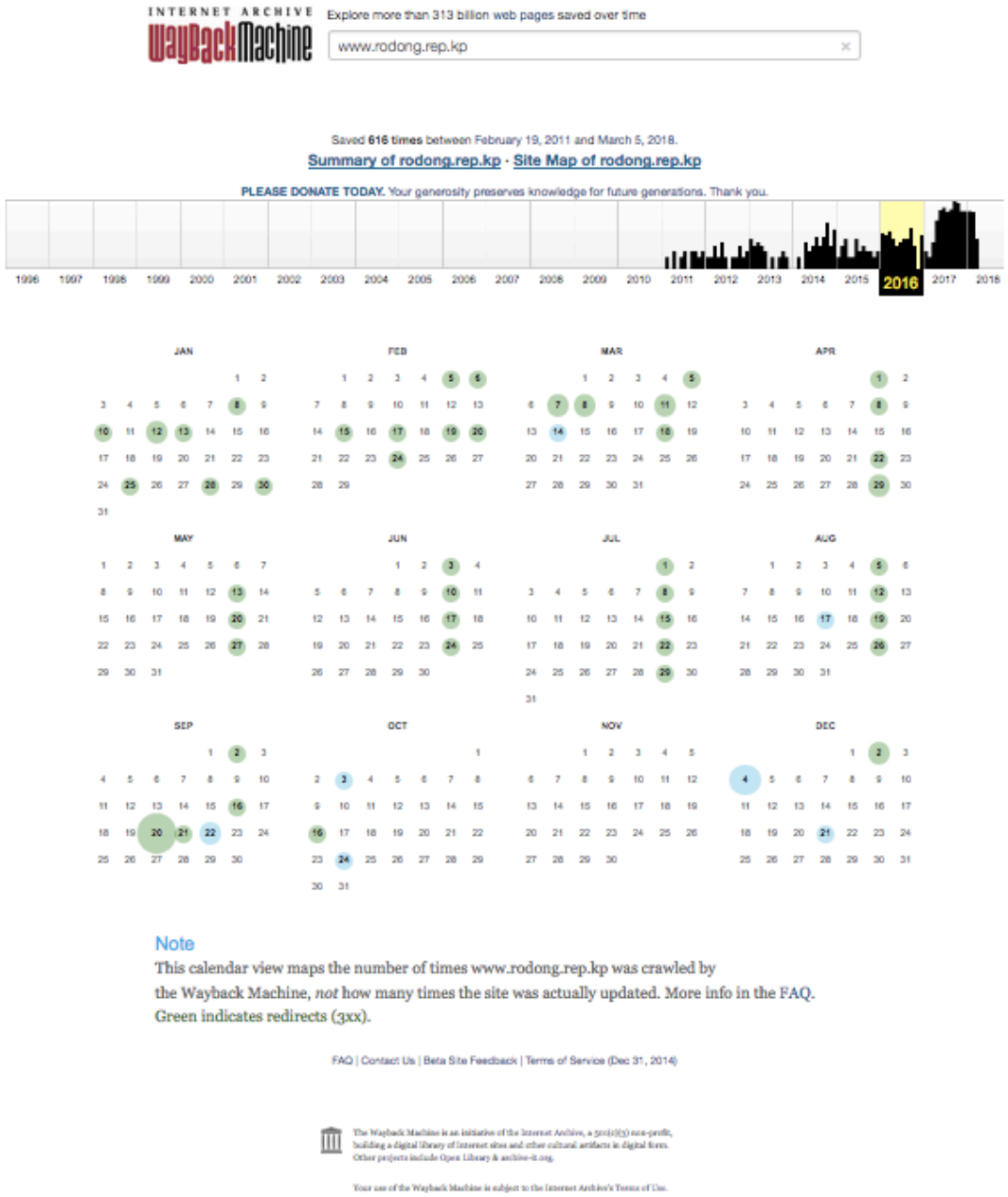


Figure 7. A snapshot from archive.org, displaying a calendar view of the number of accessible (blue) and inaccessible (green) captures of the North Korean website rodong.net.kp in 2016.

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.

Although location of the IAWM in the United States may potentially affect its crawlers' ability to capture content from other countries, the fact that the IAWM's content contribution processes are distributed and do no solely rely on automated crawlers physically based in one location, ensures that content may be contributed to the archive via individuals using the "Save Page Now" button, the "Alexa Toolbar", or adding references to Wikipedia from anywhere in the world. This makes the IAWM not only a robust knowledge device, but also one that may circumvent political attempts to prevent access to knowledge from different countries.

An Over-Knowing Device?

This article opened with a general problematizing of the construction of historical facts by the IAWM. Through the case study of the North Korean web, we attempted to perform a new kind of appraisal technique, tailored to the unique characteristics of the IAWM as a knowledge device. Theoretically, we asked whether or not the IAWM's output in the form of archived snapshots, along with their new provenance data, can be treated as other algorithmic "black boxes" of current digital platforms.

Our analysis of the North Korea case study has helped us answer this question to a certain extent. By mapping the organizations and collections attached to each snapshot of the .kp domain, we were able to identify specific epistemic processes running at the back end of the IAWM. While we are still unable to fully drill-down to the specific contribution of each capture (accounting, for example, for the circumstances around individual users' contribution of captures to the IAWM via the "Save Page Now" button, or whether or not the capture is part of the seed-list of a given Wide Crawl), the amalgamation of the multiple captures and their accumulation over time through different distributed socio-technical processes turns the IAWM into a repository that

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.
harbours knowledge that is far greater than each individual contribution or epistemic process.

Therefore, rather than pointing at the "black-boxed" elements of the IAWM, we conclude this article with a discussion about the robustness of the distributed socio-technical epistemic organization of the IAWM.

Unlike traditional archives and libraries, whose curation policies may have a top-down effect on knowledge-creation and canonization of cultural heritage, the multiple and parallel venues that inform the IAWM about relevant content is based on a combination of top-down and bottom-up processes. The top-down processes are embedded in the various crawlers of the IAWM, along with their specific "logic", "rules", and settings that are thus far relatively "black boxed". Yet these top-down processes are being continuously informed, and are continuously modified by human actors, who, as we have seen, intervene at strategic points in time and add specific captures when content is considered timely and relevant, and when access to this content is temporary or rare. The robustness of the IAWM's epistemic architecture therefore lies in its distributed and iterative process, which also includes embedded reification of "imported" knowledge. The iterative process is evident in the reciprocity between human and non-human contributions of content to the IAWM. Captures added individually, either through Archive-IT collections or through the Alexa Toolbar and the Save Now Feature, are immediately added to the seed lists of various crawls for further archiving. And distributed expert knowledge reifies the existing captures, either by contributing captures that might be inaccessible from other parts of the world, or, by adding in-depth content to the relatively shallow automated crawling process. Finally, the IAWM's staff intervenes at strategic points in time and ensures that timely events are properly covered.

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.

We therefore argue, resonating the theoretical discussion by Latour, Jensen, Venturini, Grauwin & Boullier (2012), which complicates the ties between mini and maxi digital knowledge structures, that the IAWM is a knowledge device whose sum is larger than its parts. As we noted earlier in the article, the IAWM has "known" about the existence and scope of the secretive .kp domain much earlier than the DNS leak revelation. It has aggregated distributed knowledge at different point in time and from various locations, while at each point in time the contributing organization / collection / actor may have not been aware to the entire contents of the archive. What could then be the implications of the IAWM's "over-knowing" for historical research? It may be argued that our meta-view of the "over-knowing" of the IAWM in the case of the .kp websites is an artefact gained at hindsight, but not in real time. Put differently, we would like to draw researchers' attention to the possible effect of "surplus knowledge" generated by studying the IAWM at hindsight, which was unknown on the live web during the studied period. Nevertheless, the multiple, parallel and distributed socio-technical epistemic channels for archiving web-content on the IAWM, make it indeed one of the last epistemic devices of the Web, which can be trusted for generating reliable, validated and trustworthy web-based historical facts.

References

AlNoamany, Y., AlSum, A., Weigle, M. C., & Nelson, M. L. (2014). Who and what links to the Internet Archive. *International Journal on Digital Libraries*, *14*(3-4), 101–115.

AskReddit (n.d). What is the rarest thing on the internet? North Korean websites, websites that end in .kp. Retrieved February 8, 2017, from https://www.reddit.com/r/AskReddit/comments/17dtor/what_is_the_rarest_thing_on_the_internet/

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.

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

Bowker, G. C., Baker, K., Millerand, F., & Ribes, D. (2009). Toward Information Infrastructure Studies: Ways of Knowing in a Networked Environment. In Hunsinger, J., Klastrup, L. & Allen, M. (Eds.), *International Handbook of Internet Research* (pp. 97–117). Dordrecht: Springer Netherlands.

Bruce, S. (2012). "The Information Age: N. Korean Style." T*he Diplomat, November 11*.

Bryant, M. (2016a, August 15). mandatoryprogrammer/TLDR. Retrieved December 8, 2016, from https://github.com/mandatoryprogrammer/TLDR

Bryant, M. (2016b, September 20). mandatoryprogrammer/NorthKoreaDNSLeak. Retrieved December 8, 2016, from https://github.com/mandatoryprogrammer/NorthKoreaDNSLeak

Brügger, N. (2009). Website history and the website as an object of stud. *New Media & Society*, 11(1-2), 115–132.

Brügger, N. (2016). Webraries and Web Archives–The Web Between Public and Private. In Evans, W., & Baker, D. (Eds.). *The End of Wisdom?* (pp. 185–190). Chandos Publishing.

Brügger, N., & Schroeder, R. (Eds.). (2017). *The Web as History: Using Web Archives to Understand the Past and the Present*. London: UCL Press.

Bucher, T. (2016). Neither black nor box: ways of knowing algorithms. In Kubitschko, S., & Kaun, A. (Eds.). *Innovative Methods in Media and Communication Research* (pp. 81–98). Springer International Publishing.

Chen, C., Ko, K., & Lee, J.-Y. (2010). North Korea's Internet strategy and its political implications. *The Pacific Review*, 23(5), 649–670.

Duff, W. M., & Johnson, C. A. (2002). Accidentally found on purpose: Information-seeking behavior of historians in archives. *The Library Quarterly*, *72*(4), 472-496.

Eltgrowth, D. R. (2009). Best evidence and the Wayback Machine: toward a workable authentication standard for archived Internet evidence. *Fordham L. Rev.*, *78*(1), 181.

Featherstone, M. (2006). Archive Problematizing Global Knowledge-Library/Archive/Museum. *Theory, Culture & Society*, *23*, 2591.

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.

Ford, H., Sen, S., Musicant, D. R. & Miller, N. (2013, August). Getting to the source: where does Wikipedia get its information from?. In *Proceedings of the 9th international symposium on open collaboration* (p. 9). ACM.

Ford, H., & Wajcman, J. (2017). 'Anyone can edit', not everyone does: Wikipedia's infrastructure and the gender gap. *Social Studies of Science, 47*(4), 511-527.

Geiger, R. S. (2014). Bots, bespoke, code and the materiality of software platforms. *Information, Communication & Society*, 17(3), 342–356.

Graham, M. (2016, October 26). More than 1 million formerly broken links in English Wikipedia updated to archived versions from the Wayback Machine. Retrieved April 20, 2017, from https://blog.archive.org/2016/10/26/more-than-1-million-formerly-broken-links-in-english-wikipedia-updated-to-archived-versions-from-the-wayback-machine/.

Greitens, S. C. (2013). Authoritarianism Online: What can we learn from internet data in nondemocracies?. *PS: Political Science & Politics*, 46(2), 262–-270.

Harsin, J. (2015). Regimes of posttruth, postpolitics, and attention economies. *Communication, Culture & Critique*, 8(2), 327–333.

Howell, B. A. (2006). Proving web history: How to use the Internet Archive. *Internet Journal of Law*, *9*(8), 3–9.

IANA — Report on the Delegation of .KP Top-Level Domain. (2007, September 11). Retrieved February 8, 2017, from https://www.iana.org/reports/2007/kp-report-11sep2007.html

Jones, S. (1999). Studying the Net: Intricacies and Issues. In S. Jones (ed.). *Doing internet research: Critical issues and methods for examining the net*. Thousand Oaks, CA: SAGE Publications Ltd. pp. 1–28.

Kahle, B. (2007). Universal access to all knowledge. *The American Archivist*, *70*(1), 23–31.

Karpf, D. (2012). Social science research methods in Internet time. *Information, Communication & Society, 15*(5), 639–661.

Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, Communication & Society, 20*(1), 14–29.

Knorr-Cetina,. K. (2009). *Epistemic cultures: How the sciences make knowledge*. Harvard University Press.

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.

Koop, P. (2014, January 17). New Detailed Analysis of How NSA and Its Foreign Partners Intercept Undersea Fiber Optic Cable Traffic. Retrieved February 8, 2017, from http://www.matthewaid.com/post/73733098307/new-detailed-analysis-of-how-nsa-and-its-foreign

Kucharski, A. (2016). Post-truth: Study epidemiology of fake news. *Nature, 540*(7634), 525.

Latour, B. (1999). *Pandora's hope: Essays on the reality of science studies*. Harvard university press.

Latour, B. (2005). Reassembling the Social. *Política y Sociedad, 43*(3): 127-130.

Latour, B., Jensen, P., Venturini, T., Grauwin, S. & Boullier, D. (2012). 'The whole is always smaller than its parts'–a digital test of Gabriel Tardes' monads. *The British journal of sociology, 63*(4), 590–615.

Marres, N. (2012). *Material Participation: Technology, the Environment and Everyday Publics*. London: Palgrave Macmillan.

McGoogan, C. (2016, September 21). North Korea's internet revealed to have just 28 websites. Retrieved April 22, 2017, from http://www.telegraph.co.uk/technology/2016/09/21/north-koreas-internet-revealed-to-have-just-28-websites/

McFarland, D. A., Lewis, K. & Goldberg, A. (2016). Sociology in the era of big data: The ascent of forensic social science. *The American Sociologist, 47*(1), 12–35.

McNeill, W. H. (1986). Mythistory, or truth, myth, history, and historians. *The American Historical Review*, *91*(1), 1–10.

Milligan, I. (2016). Lost in the infinite archive: The promise and pitfalls of web archives. *International Journal of Humanities and Arts Computing, 10*(1), 78–94.

Niederer, S. & Van Dijck, J. (2010). Wisdom of the crowd or technicity of content? Wikipedia as a sociotechnical system. *New Media & Society, 12*(8), 1368–1387.

Niu, J. (2012). Functionalities of web archives. *D-Lib Magazine*, *18*(3/4).

Paßmann, J. & Boersma, A. (2017). Unknowing Algorithms On Transparency of Unopenable Black Boxes. In chäfer, M. T., & van Es, K. (eds.) *The Datafied Society* (pp. 139–146). Chicago: University of Chicago Press.

Pinch, T. J. (1992). Opening black boxes: Science, technology and society. *Social studies of science, 22*(3), 487-510.

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.

Price, G. (2014, May 13). How To Save URLs To The Wayback Machine On Demand. Retrieved from https://searchengineland.com/save-urls-wayback-machine-demand-191150

Reddit.com. North Korea accidentally leaks DNS for .kp: only 28 domains. (n.d.). Retrieved December 8, 2016, from https://redd.it/53mr05

Rogers, R. (2002). Operating Issue Networks On the Web. *Science as Culture 11*(2): 191-213.

Rogers, R. (2013*). Digital Methods*. MIT press. pp. 61–82.

Rogers, R. (2017). Doing Web history with the Internet Archive: screencast documentaries. *Internet Histories*, *1*(1-2), 160–172.

Roush, W. (2005). The infinite library. *Technology Review*, *108*(5), 54–59.

Rieder, B. & Röhle, T. (2012). Digital Methods: Five Challenges. In Berry, D.M. (ed.) *Understanding Digital Humanities*. London: Palgrave Macmillan (pp. 67–84).

Rosenzweig, R. (2006). Can history be open source? Wikipedia and the future of the past. *The journal of American history*, *93*(1), 117–146.

Rossi, A. (2010, October 5). Worldwide Web Crawls. Retrieved from https://archive.org/details/widecrawl

Rossi, A. (2011, April 26). Live Web Proxy Crawls. Retrieved from https://archive.org/details/liveweb&tab=about

Rossi, A. (2012, December 17). Survey Crawls. Retrieved from https://archive.org/details/survey_crawl&tab=about

Rossi, A. (2013, March 25). Shallow Web Crawl 2013. Retrieved from https://archive.org/details/shallow_00003&tab=about

Ryfe, D., Mensing, D. & Kelley, R. (2015). What is the meaning of a news link? *Digital Journalism*, *4*(1), 41–54.

Schafer, V., Musiani, F. & Borelli, M. (2016). Web archiving, governance and STS. *French Journal of Media Research*, *6*. Retrieved from http://frenchjournalformediaresearch.com/docannexe/file/952/schafer_pdf.pdf

Star, S. L. (1999). The ethnography of infrastructure. *American Behavioral Scientist*, *43*(3), 377-391.

Starosielski, N. (2012). Warning: Do Not Dig': Negotiating the visibility of critical infrastructures. *Journal of Visual Culture*, *11*(1), 38–57.

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.

Summers, E. & Punzalan, R. (2017, February). Bots, seeds and people: Web archives as infrastructure. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 821-834). ACM.

Taylor, A. (2016, September 21). North Korea accidentally revealed it has just 28 websites - The Washington Post. Retrieved December 8, 2016, from https://www.washingtonpost.com/news/worldviews/wp/2016/09/21/north-korea-accidentally-revealed-it-has-just-28-websites/?utm_term=.f01d2d859275 .

Thelwall, M. & Vaughan, L. (2004). A fair history of the Web? Examining country balance in the Internet Archive. *Library & information science research, 26*(2), 162-176.

Thelwall, M., Vaughan, L. & Björneborn L. (2005). Webometrics. *Annual Review of Information Science and Technology. 39*(1): 81–135.

Warf, B. (2014). The Hermit Kingdom in cyberspace: unveiling the North Korean internet. *Information, Communication & Society*, *18*(1), 109–120.

Winner, L. (1993). Upon Opening the Black Box and Finding It Empty: Social Constructivism and the Philosophy of Technology. *Science, Technology, & Human Values*, *18*(3), 362–378.

Winters, J. (2017). Breaking in to the mainstream: demonstrating the value of internet (and web) histories. *Internet Histories*, *1*(1-2), 173-179.

Witt, S. W. (2015, August 17). World Sustainable Development Web Archive: Preserving and disseminating knowledge for sustainable growth. Retrieved from: http://hdl.handle.net/2142/90101.

Xroxy.com (http://www.xroxy.com/proxylist.php)

Appendix

Table 1. A list of the .kp websites discovered by the TLDR project during the DNS leak in September 2016, and their archival status on the IAWM. “Yes (error) indicates that a snapshot has been archived, but is unavailable for viewing.

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.

URL	**Description**	**Available on the IAWM?**	**Year of First Capture**	**Number of Available Captures**	**Number of inaccessible captures**
http://knic.com.kp	Korean National Insurance Corporation	yes	2015	1	111
http://www.airkoryo.com.kp/en	A flight ticket website	yes	2012	32	10
http://www.rodong.rep.kp	Official news website	yes	2011	22	204
http://www.vok.rep.kp	Voice of Korea	yes	2014	18	160
http://www.gnu.rep.kp	A religious group or well-being group.	yes	2013	118	13
http://koredufund.org.kp	A charity to increase the quality of education	yes	2010	165	6
http://korelcfund.org.kp	A charity for the elderly	yes	2011	127	2
http://korfilm.com.kp	Pyongyang International Film Festival	yes	2011	119	7

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.

http://nta.gov.kp	The Korean Tourism board	yes	2016	8	6
http://cooks.org.kp	Culinary website with recipes.	yes	2014	129	9
http://star.co.kp	ISP-related	yes	2011	6	-
http://kass.org.kp	Korean Association of Social Scientists	yes	2016	48	6
http://ma.gov.kp	Maritime Administration of Korea	yes	2014	57	5
http://friend.com.kp	The website of the Committee for Cultural Relations with Foreign Countries.	yes	2010	171	11
http://kiyctc.com.kp	Korean International Youth and Children's Travel Company	yes	2016	48	7
http://naenara.com.kp	"official" government site - information about the country	yes	2013	10	223
http://star-co.net.kp	ISP-related	yes (error)	-	-	10

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.

http://ryongnamsan.edu.kp	Kim Il Sung University	yes (error)	-	-	68
http://silibank.net.kp	ISP-related	yes (error)	-	-	3
http://star.edu.kp	ISP-related	yes (error)	-	-	1
http://sdprk.org.kp	Sports	yes (error)	-	-	47
http://star-di.net.kp	ISP-related	no	-	-	-
http://portal.net.kp	ISP-related	no	-	-	-
http://rcc.net.kp	ISP-related	no	-	-	-
http://star.net.kp	ISP-related	no	-	-	-

Table 2. A list of the number of http response codes to attempts to access 28 .kp websites from 30 countries. Response error codes are 302 Redirect, 401 Unauthorized, 403 Forbidden, 404 Not Found, 500 Internal Server Error, 503 Service Unavailable, 504 Gateway Timeout.

Country	**HTTP response Code**										
	Proxy Type	*Success*	*Request Timeout*	*302*	*401*	*403*	*404*	*500*	*502*	*503*	*504*
Afghanistan	Anonymous	13	12				1				

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.

Azerbaijan	Transparent	23					3				
Bangladesh	Transparent	4	18				4				
Brazil	Transparent	2	22				2				
Burundi	Transparent	3	19				4				
Canada	Transparent	8	7						1	10	
Djibouti	Transparent	9	10				7				
France	Transparent	11	7				1				7
Great Britain (UK)	Anonymous	14	3	1					1	7	
Honduras	Transparent	10	6		7						3
India	Transparent	8	7						2	9	
Indonesia	Transparent	5	14				7				
Iran	Anonymous	2	17							7	
Israel	Anonymous	13	4					1	8		

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.

Italy	Transparent	5				20		1			
Korea (South)	Transparent	7	16							1	2
Malaysia	Transparent	3	21							1	1
Mexico	Transparent	13	6	1							6
Netherlands	Transparent	7	13							1	5
Nigeria	Transparent	3	18				5				
Poland	Transparent	10	9							1	6
Romania	Transparent	12	6							1	7
Russian Federation	Transparent	11	3				2		10		
Slovenia	Transparent	5	18				3				
Spain	Anonymous	11	7						8		
Taiwan	Transparent	7	11							1	6
Thailand	Transparent	14	4							1	7

This is a preprint version of the published paper. Please refer to Ben-David, A., & 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.

Ukraine	Transparent	13	4							1	7
United States	Transparent	2	19				1				
Viet Nam	Transparent	3	14								