NEGOTIATING THE WEB OF THE PAST

Valérie Schafer, Francesca Musiani, Marguerite Borelli

French Journal For Media Research – n° 6/2016 – ISSN 2264-4733

NEGOTIATING THE WEB OF THE PAST

Web archiving, governance and STS

Valérie Schafer
Chargée de recherche / Researcher
Head of the ANR Web90 Project
ISCC, CNRS/Paris-Sorbonne/UPMC
valerie.schafer@cnrs.fr

Francesca Musiani
Chargée de recherche / Researcher
ISCC, CNRS/Paris-Sorbonne/UPMC
francesca.musiani@cnrs.fr

Marguerite Borelli
Scholar
UCL, Londres
margueriteborelli@gmail.com

Résumé

Les dimensions matérielles, pratiques et théoriques de l'archivage du Web sont étroitement liées : les processus et infrastructures, bien que souvent discrets ou invisibles, tiennent une place centrale dans la constitution de ce patrimoine nativement numérique. Cet article entend montrer dans quelle mesure les approches développées par les *Science and Technology Studies* (STS) sont particulièrement appropriées pour étudier la fabrique des archives et de l'archivage du Web.

Mots-clés :

Archives du Web, Patrimoine nativement numérique, Archivage du Web, STS, gouvernance

Abstract

The material, practical, theoretical elements of Web archiving as an ensemble of practices and a terrain of inquiry are inextricably entwined. Thus, its processes and infrastructures – often discreet and invisible – are increasingly relevant. Approaches inspired by Science and Technology Studies (STS) can contribute to shed light on the shaping of Web archives.

Keywords

Web archives, Born Digital Heritage, Web archiving, STS, governance

INTRODUCTION

Forbes Magazine published "The Internet Archive Turns 20: A Behind the Scenes Look at Archiving the Web" on January 18th, 2016, opening the long list of articles and celebrations honouring, on its twentieth birthday, the American foundation Internet Archive, founded by Brewster Kahle only a few years after the creation of the Web.

A pioneer in Web archiving, the Internet Archive has been joined in its effort over the years by a variety of actors, amongst which are national institutions, such as the Library of Congress (LoC) in the USA, or the *Bibliothèque Nationale de France* (BnF) and the *Institut national d'audiovisuel* (INA), the two institutions responsible for the legal deposit of websites since 2006 in France, but also research institutes, such as George Mason University's Roy Rosenzweig Center.

As born-digital heritage is progressively being institutionalised – its importance, alongside that of digitalised heritage, was recognized by the UNESCO in its 2003 *Charter on the Preservation of Digital Heritage* – its collections are growing exponentially. Indeed, in February 2016, 466 billion web pages were accessible through the Internet Archive's Wayback Machine, and the volume of Internet Archive's Web archives was estimated to be fifty times bigger than that of the LoC's print collections – with all due reservations that this comparison to a library, that Claude Shannon already held as reference in 1949, can raise (Milligan, 2012).

Nicolas Delalande and Julien Vincent have observed that "the conservation of websites is [...] a well-identified challenge since several years, and one whose relevance is still growing" (2011)[1]. Despite this, their appropriation and study by scholars is slower to develop, and Web archives still lack visibility as a source; so much so, that Richard Rogers stated that "we even speak of a crisis in certain fields of Digital Humanities, because scholars haven't received/acknowledged these new documents in a proportional manner to the rush towards digitalisation. For instance, Web archives remain only seldom used for scientific purposes [...]" (2015)[2].

However, Web archives hold great potential, both as a source (Schafer and Thierry, 2015) and as the object of analyses, as observed by Hélène Bourdeloie: "the real innovation is the fact that digital media and technologies are no longer simply tools at the service of research, but instead, they are also objects of research in themselves. They are, all at once, instrument, method, field of research and object of study" (2013)[3].

This article provides an analysis of the shaping of Web archives as informed by Science and Technology Studies and sociology of innovation; it is based on a synthesis of theoretical and pragmatic approaches, interviews with Web archiving professionals and observations conducted within Web archiving institutions. We show how a grounded understanding of the collection of Web archives sheds light on their "heritagization" (Davallon's notion of *patrimonialisation*, 2006). Similarly, we address in this light the multiple socio-technical

[1] Our translation.
[2] Idem.
[3] Idem.

mediations, arrangements, and agencies mobilised throughout the archiving process – be they technical or human – or the choices that decide their selection, presentation, and accessibility. A number of dynamics and negotiations contribute to qualify and define born-digital documents as heritage: all these dynamics are a testament to the plurality of actors engaged in this process, their incentives and their objectives, at times either complementary or divergent. The close observation of such negotiations, formal and informal, allows to move towards a relational conception of authority, one based on networking and "facilitation of linkages between social worlds" as a form of power (Flyverbom, 2011, p. 96). Moreover, the web archive can be understood as a boundary object (Star and Griesemer, 1989; Bowker and al., 2015), claimed by multiple communities whose aspirations and logics are just as diverse as those expressed on the "live" Web – and more generally on the Internet.

For all these reasons, Web archives can be better used as sources once they have been understood as an object of study.

OPENING THE BLACK BOXES …

"Where I end up is seriously jealous of the possibilities; and seriously wondering what the 'object of study' might be. In the nature of an archive, the UK Web Archive imagines itself as an 'object of study'; created in the service of an imaginary scholar. The question it raises is how do we turn something we really can't understand, cannot really capture as an object of study, to serious purpose? How do we think at one and the same time of the web as alive and dead, as code, text, and image – all in dynamic conversation one with the other. And even if we can hold all that at once, what is it we are asking?",

recently reflected Tim Hitchcock on the topic of Web archives (2015). This modernist historian, today a professor of Digital History at the Sussex Humanities Lab, highlights Web archives' inherent novel, fluid and evolving nature that makes them initially difficult to grasp. As Claude Mussou notes as well:

"[...] whereas a traditional archive is constituted of documents whose use value is no longer relevant, selected and organised in precisely defined funds, according to predetermined selection criteria, the Web archive is continuously constructed through the automated collection of contents which have not, at least for some, lost their value" (2012)[4].

Thus, 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.

[4] Our translation.

Reconstructions

"In general it is impossible to archive online web content on a scale of 1:1. [...] What is in the web archive therefore may prove to be inconsistent with what was once online: obviously, something has been lost, but something may also have been archived which was never online at the same time. In this sense, the web archive could turn out to have too little or too much web material, and it can be very hard to determine with certainty what the online web actually looked like at a specific point in time. What we are witnessing here is that even in the age of digital reproduction, we are not only making copies. On the contrary: the Web archiving process creates unique versions, each with their individual 'aura',"

summarized with clarity Niels Brügger in 2012. He and Megan S. Ankerson (2015) have convincingly demonstrated that each web archive is a reconstruction that can sensibly differ from the original web page or website at the time of its capture (2012). There are several reasons for this, the first of which is the depth of the harvest and of the capture. Indeed, very often websites are only partly archived, because the crawler-robot is programmed to capture them only to the depth of a few clicks – which explains why users regularly find themselves facing missing or unfound web pages. In their effort to capture vast, representative screenshots of the contemporary Web in all of its diversity, many institutions have opted for this "superficial" method of collection. For instance, the BnF favours quantity for its so-called broad[5] crawls, which resulted in the capture of over 4.4 million websites at about two clicks of depth in 2015. Moreover, if these websites are rarely preserved in their entirety, rarely are their web pages integrally preserved, and as such, they can also be incomplete: advertisements, pop-ups and banners are amongst the elements that can be voluntarily blocked out of crawls – and indeed, they often are, leading to the omission of a considerable part of born-digital heritage, a frequent source of discomfort for Web users that nonetheless remains an important illustration of business models, communication strategies, and the attention economy (Kessous, 2012; Citton, 2014) of the Web of our times. Furthermore, fonts can differ from their originals in Web archives: if at the time of archiving a web page's font was not embedded in its original source code, but rather used by default, then it will be the present-day browser's default font settings that are displayed on the archived web page.

Another feature of Web archives is the capture and preservation of the images that are on the Web; several archived Web pages from the 1990s contain empty frames where their images used to be. Probably, the reason for this is less so the technical difficulty of the capture, than the "impatience" of crawler robots and harvesting objectives of the time. Indeed, the Internet Archive was linked to Brewster Kahle's Alexa, a company whose specialty was the ranking and indexing of websites, rather than the preservation of images. In the present day, to avoid duplicates, those are not systematically recollected. If their URL has not changed from one crawl to another, they may be recovered from the most recent crawl, instead of being recaptured. This explains some of the inconsistencies that can arise when surfing the archived Web, such as when a calendar widget shows a different date than the date of capture of the web page. A concrete example is the archive of the French National Centre for Scientific Research (CNRS)' website: a web capture by the BnF on August 1st, 2015[6], shows the institution' logo in black (signifying mourning), instead of its usual blue colour. In fact, the

[5] See BnF website: http://www.bnf.fr/en/collections_and_services/book_press_media/a.internet_archives.html
[6] French National Library (BnF), Web Archives, the CNRS website on August 1st, 2015, 9.07 GMT http://archivesinternetbnf.fr/20150801090709/http://www.cnrs.fr/

black logo was captured *ex post facto*[7], when the BnF crawled the website again on November 23, 2015[8], following the Paris attacks that took place a few days earlier. Because the website's logo was not recollected on the August 1st crawl, the access software reconstructed the web page with its most recent capture of the logo (on November 23) and inserted an anachronistic element onto the August 1st Web archive – creating a document that had never existed in this form on the live Web.

Other such peculiarities are also manifest to regular observers of the archived Web: unfound pages due to the crawler not following links, temporal jumps, and hybrid archived pages. This raises the question of the authenticity of these sources, and further complicates their external, as well as internal, critique (Schafer and Thierry, 2015).

At the 2014 General Assembly of the *International Internet Preservation Consortium* (IIPC), Louise Merzeau argued that although the history of Web archiving is short, it has already seen paradigm shifts whose consequences are mirrored by the archives. Indeed, in the 1990s, when the Internet Archive's founding project was born, Web archiving followed the "documentary model," aimed at a universal archiving of the Web anchored in traditional models, most notably that of the library. Then, at the turn of the century, the "documentary model" was briefly replaced by the logic of the archive as memory, based on the model of the scholarly copy or *exemplar*, with "tinkering" used as a method for lack of a better alternative. During this second phase, the main emphasis of Web archiving was preserving, and maybe even freezing the Web, by saving each corpus element piece by piece. And finally, since the end of the 2000s, Web archives are constructed under the logic of the "temporal archive," which seeks to fully capture the instability of the Web, through the development of dynamic archiving methods at the image of the Web itself. This instability, initially considered a contingent dysfunction, is more and more seen as an essential dynamic. Louise Merzeau also notes that paradoxically, by increasingly sticking to the variations of the live Web, the archived Web is getting further and further away from the idea of restitution, thus requiring a proficient understanding of flows from academics (2014).

The questions of temporality and intelligibility are also analysed by Niels Ole Finneman (2015), who emphasizes that all Web archive corpuses are devised around a threefold temporal dimension: original content, accumulation and transformation, and exploration of the archive by a scholar. The latter is grounded in his own era, and thus introduces its biases, possibly inducing a presentist or nostalgic reading (Schafer, 2015).

Re-mediations

Along those lines, one might add a fourth dimension to the three temporal dimensions suggested by Niels Ole Finneman. Among the many mediations incurred by Web archives, their encryption onto a user interface is especially important. As shown by Megan S. Ankerson, the interface of the Wayback Machine (such as its 2003 "Take Me Back" button,

[7] As shown on the CNRS' web page of August 3rd 2015, as archived by the Internet Archive https://web.archive.org/web/20150803194822/http://www.cnrs.fr/

[8] French National Library (BnF), Web Archives, the CNRS website on November 23, 2015, 15.02 GMT http://archivesinternetbnf.fr/20151123150246/http://www.cnrs.fr/

which subsequently became "Browse History") and its temporal browsing interface in its present-day version, are all expressions of its creators' vision of accompanying users in their journey through the archives of the Internet Archive (2015). In the same way, the new user interface launched by the Ina at the beginning of 2016 articulates the institution's representation and experience of the scholarly uses of their archive, and goes so far as anticipating their users' expectations and the searches they may want to perform. This influence of host institutions on user interfaces is established through subtle negotiation between these elements: the archives' specificities, ergonomics and technical constraints, and foreseeable developments, as observed by the Ina's Web legal deposit team:

"Our choice for the vertical menu is not only based on ergonomic concerns but also on the requirements of future developments. Our mid-term goal is to provide a uniform search platform without distinguishing the different types of archived objects, by using the left menu like a dashboard. For example, if a user searches for a specific url, we would like to provide not only the navigation for this url, but also corresponding metadata, the list of videos found on the page and the list of tweets that point to this url" (Pehlivan, 2016).

Furthermore, the launch of oldweb.today[9], a service that allows for the navigation of Web archives on browsers of their times, can help with their contextualisation. Beyond the importance of considering the visual and material specificities which framed the context of production and accessibility of the pages – terminal, network and bandwidth – reading a Web archive implies an understanding of what the web page, the website and even its hyperlink meant in their time. Anne Helmond shows how much the meaning of "hyperlink" has changed over time; she studies

"the history of the hyperlink from a medium-specific perspective by analyzing the technical reconfiguration of the hyperlink by engines and platforms over time. Hyperlinks may be seen as having different roles belonging to specific periods, including the role of the hyperlink as a unit of navigation, a relationship marker, a reputation indicator and a currency of the web" (Helmond, 2013).

Technical and human negotiations at both levels of collection and consultation of the Web archive include many operations: the choices of particular crawl frequencies, depths, domains to be collected, programming of robots, data deduplication processes; the recreation of links and filling of URLs by the access software; the exclusion of specific elements such as advertisements; the creation of platforms and consultation environments offering different designs and functionalities. All these operations bear witness to the ongoing choices that reflect the scope and ambitions established by and for the actors of Web archiving.

Frames and environments

Web archives are part of complex environments that go well beyond the consultation interface seen by the user. For instance, the conditions of access vary from one fund to another, as a function of the institutional and legal frameworks applicable to the archive. Indeed, the very open online access model of the Wayback Machine or the Portuguese Web archives[10]

[9] http://oldweb.today
[10] http://arquivo.pt

contrasts starkly with the strictly-restricted in-house access models of the BnF, the Ina (and some regional French libraries), or the Swiss Web archive, whose website reads: "For copyright reasons, conditions of access to the archives are limited. Any reproduction of the materials – such as downloading, printing, etc., is therefore blocked"[11]. In contrast, the access policies of the Internet Archive and the Portuguese Web archives are rooted in the open culture of Web and informatics that is very different to the legal logic of other national institutions. A European exception, the Portuguese Web archives have been developed by the Portuguese NREN (National Research and Education Network), whereas in many other states web archiving has been made the prerogative of libraries and is framed within the traditions and legislation initially conceived for print materials. However, even before Web archives, videos and composite multimedia documents (1975), and finally multimedia, software and databases (1992) were already included in the French legal deposit, alongside prints and etchings, maps and plans (Oury in Cohen and Verlaine, 2013).

The archive's target audience also influences its collections in an important way. The Ina, whose archives are linked to the audiovisual sector, is the obvious example, but other actors have also chosen to refine their archive's perimeter through more or less selective collection policies. Indeed, Switzerland only captures the part of the .ch domain deemed "of heritage value"[12], whereas Great Britain[13] and France aim at collecting large representative samples of the existing Web[14]. But these broad policies also have their limits; the BnF bases its selection on the lists of the Afnic (French Network Information Center) and the ISP OVH, allowing the library to collect about 4.5 million domains, but leaving out those 3 million that are hosted by other companies such as Gandi. This propensity for representativeness does however result in successful screenshots of the Web in all of its diversity, both institutional and "vernacular," as studied by Olia Lialina (2005). The preservation of this vernacular Web is also the self-attributed mission of the Archive Team, a (Web-)hacktivist group affiliated with the Internet Archive and rallied around the slogan "We are going to rescue your shit,"[15] who have notably saved GeoCities and Mobileme – both important platforms of the "ordinary" expression and creativity of Web users – from falling into oblivion.

Legal frameworks can pose constraints on the accessibility and reproducibility of archives, and largely determine the perimeter of selection. But in some cases – such as those of French and Swiss legal deposit institutions – they can also legally allow for the bypassing of robots.txt restrictions, which the Internet Archive respects. The distinctions between the seemingly opposed universalistic, open logic of the Internet Archive, and the restrictive, territorial logic of European institutions, seem thus more complex than what a binary reading could suggest.

---

[11] https://www.nb.admin.ch/nb_professionnel/01693/01695/01705/03333/index.html?lang=fr#sprungmarke1_50

[12] The Swiss National Library has opted for a selective strategy. It proceeds to archive the digital heritage Web sites that have a strong link with Switzerland and are freely accessible (e.g. Web sites on cantons or municipalities). Canton-level libraries, and other specialized ones, are mostly those that select Web sites. Special collections are put together on the occasion of particular events (e.g. federal elections of 2011). https://www.nb.admin.ch/nb_professionnel/01693/01695/01705/03333/index.html?lang=fr#sprungmarke1_50

[13] http://www.webarchive.org.uk/ukwa/

[14] The collection perimeters also raise the question of representativeness and sampling. The quest for both varies depending on institutions, who are in any case aware that fully inclusive archiving is not possible. See Huuderman and al. (2015) for a stimulating approach to uncover unarchived web pages and websites and to reconstruct different types of descriptions for these pages and sites, based on links and anchor text in the set of crawled pages.

[15] http://archiveteam.org/index.php?title=Main_Page

Understanding the complex stakes and processes at work in web archiving calls for a detailed analysis of its multiple levels, layers and stakeholders, for which the tools and approaches of science and technology studies (STS) are especially relevant. STS provides conceptual and methodological tools to describe how human and non-human actors exercise joint agency in mediated environments, and to study the way in which infrastructure and culture have progressively merged. Janet Abbate (2012) states in this regard:

"The Internet, as a communications platform, challenges the historian to reconsider her hypotheses on the boundaries between human and non-human, infrastructure and culture, technology and society. STS concepts can help make sense out of these blurred categories. Rather than a passive medium of communications, the Internet emerges as a driving force, a complex network of actors both automated and spontaneous. By focusing our attention on the internal mechanisms of technology, we can observe how programmers' cultural backgrounds are implicated in the infrastructures they create. Hybrid agency provides a theoretical framework for the study of the links between programmers, software and computational media users, whereby people and machines are co-authors of online contents. The exploration of this human/machine collaboration, its origins and socio-political implications, is only just beginning."[16]

From the Internet and the Web that are the focus of Abbate's research, this reasoning may be applied to new yet related terrains such as Web archives, digital and born digital heritage, to highlight the socio-technical negotiations that take place in Web archiving, from its earliest stages to their exploitation. Research in media studies has certainly not neglected these aspects; yet, it has rarely adopted negotiations as the primary subject of inquiry or focused explicitly its attention on their mundane, procedural aspects – something that STS approaches, inextricable blend of methods and theory, allow to do.

…THROUGH STS TOOLS AND APPROACHES

The Internet's upper and lower layers are very closely linked – both feed into a history of infrastructure. Web archives are no exception, as infrastructures that bring the knowledge of the past into the present. Thus, an exploration of the "negotiations of the Web of the past" that contribute to its current governance needs to include a socio-technical analysis of Web archives' backstage as shown before. This entails looking into both the "plumbing" (Musiani, 2012), the devices composing the system, and into what Susan Leigh Star has effectively labelled as the "invisible work" (1999: 385), the design processes that have led to the creation and evolutions of such devices, and eventually entailed controversies about the different paths they could take. The different and conflicting perimeters of archiving, institutions' missions, the frequencies of webpage captures and the software subtending them, the variety of actors and budgets involved in Web archiving, the plethora of tools and protocols used by those different actors, the different access modalities proposed or imposed upon the user, all of these issues and artefacts, most of the time left (intentionally) invisible to whoever accesses a webpage of the past, take centre stage in STS-informed approaches to the study of Web archives.

[16] Our translation.

Pipes, bricks, mortar

It is of little surprise that the "pervasive enabling resources in network form" (Bowker et al., 2010: 98) that constitute infrastructure are increasingly being studied by scholars of information and communication technologies, the Internet first and foremost. Indeed, the "invisible" layers of the Internet -- underlying practices, uses and exchanges in this networked system of systems – informs its adoption and (re)appropriation by users, its regulation, and its organizational forms.

As Geoffrey Bowker and colleagues note, the term "infrastructure" first evokes large sets of material, collective equipment necessary to human organization and activity - to name but a few examples, buildings, roads, bridges and communications networks. However, "beyond bricks, mortar, pipes or wires, infrastructure also encompasses more abstract entities, such as protocols (human and computer), standards, and memory," and in the case of the Internet, "digital facilities and services [...such as] computational services, help desks, and data repositories" to name a few (Bowker et al., 2010: 97). The field of STS has explored the social and organizational dimensions of infrastructure, paying particular attention to a number of its characteristics that make it an extremely interesting, albeit "discreet" subject of study for scholars of complex socio-technical systems. Namely, the fact that infrastructure typically exists in the background, it is invisible, and it is frequently taken for granted (Star & Ruhleder, 1994). This invisibility often extends to the workers ensuring its operation and maintenance.

We can easily see how these features are consubstantial or may be applied to the subject at hand, and the interest, for scholars of Web archives, of addressing them through an STS lens. Following and accounting for the different operations needed to ensure the maintenance and preservation of Web archives as well as their circulation, or the set of technical features that constrain and enable their appropriation by users and institutions themselves, researchers may benefit from STS concepts and tools such as technical democracy, co-production, and boundary objects (Callon and al., 2001; Star & Griesemer, 1989).

Agencies

"...from your terms of use:
'...Further, you agree not to recirculate your password to other people.'
This is a hardship.
I had previously done this because I didn't realize you had the provision there.
Sometimes, I want to contribute a large file to the archive, but my internet connection is slow or limited by a data plan. In those instances, I have to give my credentials to another worker so he can do it for me.
Thus, I'm asking an exemption."

Andrew Bontrager, an Internet Archive user,
commenting a change in the Web archiving site's Terms of Use, January 2015.[17]

[17] http://blog.archive.org/2014/12/30/update-to-terms-of-use/

What does user Andrew's commentary (which we will come back to later) tell us about the usefulness of STS approaches to the study of Web archives? Possibly that, if socio-technical (and techno-legal) aspects of Web infrastructure need to be unveiled in a thorough analysis of Web archives, so does their design. Indeed, design processes -- both of the original web page and of the tools destined to archive and retrieve it -- in several instances become prescription thanks to passwords, bottlenecks, clashes of formats, lack of interoperability, or the "silos" created by national Web archives in knowledge infrastructures (Edwards, 2013): in short, the several socio-technical mediations undertaken by Web archives. The commentary also shows hybrid human and non-human agencies at work, bearing witness to both the dimension of collective action, and the "power of actors to formulate constructive criticism, thus *changing* the course of the reproduction of the social world" [18] (Proulx, 2009).

Knowledge infrastructures and national silos

The first mediation concerns the new "architectures" of networks and knowledge created by Web archives: by creating silos between national archives at the European level – based in particular on domain names – they raise the issue of interoperability, of bridges (and their absence), of Web fragmentation, of the links between national sovereignty and networks, of digital, national and institutional boundaries. They also raise the question of the articulation between the universal issues linked to born-digital heritage (as defined in the UNESCO charter of 2003), such as the "knowledge society," the Internet as a common good, and the shaping of "knowledge infrastructures," the delegation of legal deposit to institutions, its perimeters and the exception to copyright. And finally, the issue of the status of the metadata produced by Web archiving institutions, and the conditions of their circulation, is also an important one.

The possibility of gathering European collections, and of sharing metadata, is part of a reflexion led by the European research network RESAW (A research infrastructure for the Study of Archived Web materials),[19] initiated by the Netlab, Aarhus University. Our research project "From #JeSuisCharlie to #Offenturen: the archiving of born digital heritage and terror attacks" [20] offers a good example of what is at stake in the sharing of metadata, in interoperability and documentation of collections and funds, as well as in their fragmentation, and what they all mean for the scholarly exploitation of Web archives. In addition to the crawls operated by the BnF and the Ina at the time of the attacks (Twitter was archived on the night of the attacks by the Ina, and "emergency" crawls were conducted by both institutions in the following days and months), Internet Archive's Archive-It service [21] also conducted crawls, based on URL lists signalled by the BnF and other institutions of the IIPC, such as the National Library of Spain, the UCLA Library, or the Denmark State and University Library. Although the BnF and Archive-It collections about the attacks are based on the same domain

[18] Our translation.
[19] http://resaw.eu
[20] https://asap.hypotheses.org. Based on Web archives, and with the help of the BnF and Ina's Digital Legal Deposit teams, this interdisciplinary research project, funded by the CNRS in 2016, seeks to document the archiving of the Web and Twitter during the events, to question the conditions and possibilities of elaborating corpora, and to extract, from this considerable quantity of data, some elements of analysis of the events' online "making".
[21] https://archive-it.org/explore?q=Charlie+Hebdo&page=1&show=Sites

selection, the temporalities of their captures differ, meaning that both collections hold different contents, and are complementary. However, as of now it remains impossible to exploit these complementary funds as the coherent whole that they are, because they have not been merged onto a single platform. The consequence of this for their academic exploitation is an important one, as it means that it is impossible to use analysis tools on the entire corpus (e.g. metadata analysis or format and image analysis through Facetredux software[22]). Scholars also need to get acquainted with the selection policies that shaped these funds, so as to understand what bias the capture introduced in the corpus, which fails to render a faithful image of the way the entire Web "vibrated" in January and November 2015 (Boullier, 2015).

Blazing news, slow(er) archives?

At the moment of the early-2015 terrorist attacks in France, the "emergency" collection related to the events by the BnF leads also us to reflect on the temporalities of archiving, and the articulation between events, history and memory. In this regard, Camille Paloque-Berges highlights the ongoing paradox between the "trusted third party" mission delegated to institutions, and the responsiveness to ad-hoc reactions. Do heritage institutions have to follow the speedy pace of current news?

The same question can also be raised in the so-called "Ferguson case." This label refers to the shooting of African-American teenager Michael Brown by (Caucasian) police officer Darren Wilson, which occurred on August 9, 2014, in the city of Ferguson, Missouri in the United States. The shooting of Brown, who was unarmed, sparked unrest in Ferguson and was the subject of widespread attention in the U.S. and worldwide - unrest and attention that were mirrored in intensive Twitter practices. Later that month, the Society of American Archivists discussed the potential role of Web archivists in documenting the event, with a reported wide agreement that "Ferguson was a painful reminder of the type of event that archivists working to 'interrogate the role of power, ethics, and regulation in information systems' should be documenting."[23] Shortly after, the Archive-It service run by the Internet Archive announced their collection of seed URLs for a Web archive related to Ferguson, and parallel archiving attempts took place, destined however to remain incomplete due to the crushing volume of tweets[24]. Interestingly, the discourse about "instant archiving" of the Ferguson-related online uproar mixed with other infrastructure-related issues such as net neutrality and algorithmic filtering, and how these combined issues may have affected present and future visibility of the issue: "What happens to #Ferguson affects what happens to Ferguson" (Tufekci, 2014).

As in the wake of the Charlie Hebdo, Hyper Cacher or Saint-Denis and Bataclan killings in France, the response of actors in the archiving ecosystem to the online (and offline) unrest that followed the shooting of Michael Brown leads us to re-think the weight of actors and their choices in the creation of memory and digital heritage – "human"-originated restrictions adding up to technology-embedded ones, such as terms of use, robot.txt exclusions, technical locks.

---

[22] https://github.com/INA-DLWeb/FacetRedux/tree/master/src/main/java/fr/ina/dlweb/proprioception/facetRedux

[23] http://inkdroid.org/2014/08/30/a-ferguson-twitter-archive/

[24] Ibid. "There were some gaps because of [...] the data just moving too fast for me to recover from them: most of August 13th is missing, as well as part of August 22nd. I'll know better next time how to manage this higher volume collection."

Restrictions and « locks »

Controversies such as the Suzanne Shell vs. Internet Archive legal case of 2006[25] reveal the importance of legal-digital regimes and licenses. In this case, the defendant, Suzanne Shell, an American activist, alleged that the copying of her site by the Internet Archive constituted an acceptance of the site's terms of use, which require the payment of high fees for the copying activity. Shell's site stated that whoever copied or distributed any content from it was "entering into a contract." The argument at the core of her lawsuit was that the Internet Archive's periodical, implicit, "by-program" visitation of her site constituted an acceptance of her terms; she argued this against two factors, the inability of an automatic Web crawler to actually understand the terms, and the absence, on Shell's website, of a robots.txt file to deter crawlers. Even if the lawsuit was eventually settled, this case posed very interesting questions for Web archiving and more broadly for sites based on automated content gathering, first and foremost the liability of automated software programs for their "actions." Will the Internet Archive need to "teach their Web spiders how to read contracts" (Claburn, 2007)?

Further along those lines is the issue of the capture of the increasingly numerous password-protected pages or intranets. Indeed, the constitution of heritage is often contingent upon the accessibility of pages, rather than their content – the device determining the (im-)possibility of inclusion, the design becoming prescription.

Finally, let us go back to Andrew Bontrager's very revealing commentary left on the Internet Archive blog, cited earlier on and on which this whole section of the article sheds light: contributions to the elaboration of born-digital heritage, such as voluntary user Andrew's, take shape (or are prevented from doing so) by the velocity of an Internet connection and the possibility of accessing it in a constant manner; by the "locks" that make it impossible to archive on the Web password-protected pages; by a shared task conducted via different protocols and tools, and the lack of interoperability that ensues. As in other large socio-technical systems, be they networked or not, technical devices that are constitutive of Web archives both inform and are informed by practices, which allows us to speak of "boundary objects" – entities that serve as an interface between the perspectives of different actors and social worlds (Star and Griesemer, 1989).

Boundary objects

Several features of born-digital heritage allow for its qualification as a boundary object; first and foremost, the discussion lists, newsgroups or websites dating back to the first half of the 1990s that Camille Paloque-Berges studied in an effort to reconstruct the trajectories of innovation within pioneering user groups (2016). Through an analysis of the collective and a-synchronous forms of computer-mediated communication or CMC (online lists and discussion groups), she highlights the uses of these technologies in the mid-1990s, and observes that they served a logic of confrontation to social (rules of sociability in online public speaking), political (equipment and techno-scientific development, governance and regulation of networks) and economic (transition from non-commercial networks to the digital economy) norms.

[25] http://blog.ericgoldman.org/archives/waybackshell.pdf

"Online mail is the first form of this new media genre which is the CMC. Deployed on computational networks since the 1970s, its historical value as a socio-technical accompaniment to the development of the Internet is well documented in the social history of technologies (Abbate, 2000) and in the sociology of media, sciences and technology (Flichy, 2001; Paravel, 2007). At first a channel of communication between scientists and engineers involved in the construction of these networks, CMC is, from the beginning, a recursive driving force of their development: it is to better communicate with peers that networks were developed, and it is to better develop the networks that it was used to communicate with peers through its channels. I specifically focus on collective and asynchronous forms of the CMC: lists and online discussion groups. As the socio-technical media of distributed communication, they constitute one of the areas of the construction, communication and debating of contemporary knowledge in digital networks. They sketch the contours of the 'computer scientists' republic' (Flichy, 2001)[26] *at the core origin of networks, whose ordinary enunciative rituals (between technical and ordinary language) (Mourlhon-Dallies & Colin, 2004, Hert, 1998) can be analysed, as well as the expressions of technical democracy experimented online (Paravel, 2007)." (Paloque-Berges, 2016)*[27].

Not only the discussions themselves, but also their archiving, facilitated the negotiations and cooperation taking place in debates between programmers and pioneer users, and even those situated within the programming community, by helping to shed light on the reconfiguration and re-appropriation of innovation (Star and Griesemer, 1989). Moreover, they also contributed to the emergence of the spatial and temporal practical dimension of innovation, as noted by Guillaume Latzko-Toth :

"If the concept of boundary object sheds light on the translation process at work in scientific and technological activity, it also invites us to rethink the relationship between the ends and means of innovation: the artefacts do not (always) constitute the end of technological activity, but they are also (and maybe more often) the foundation for it to unfold as a practice" (2010).[28]

Sufficiently "malleable to adapt itself to the local needs and constraints of its different user-types,", and capable of existing in different social worlds, all the while satisfying the "informational needs" of each, and being sufficiently robust to maintain a common identity throughout these adaptations (Star and Griesemer, 1989), born digital heritage carries within itself the challenges of its maintenance, memory, but also of its governance, as the aforementioned newsgroups are nowadays maintained by… Google.

WEB ARCHIVING GOVERNANCE

In 2005, the Working Group on Internet Governance defined Internet governance as "the development and application by Governments, the private sector and civil society, in their respective roles, of shared principles, norms, rules, decision-making procedures, and programmes that shape the evolution and use of the Internet," adding that "it also includes other significant public policy issues, such as critical Internet resources, the security and safety of the Internet, and developmental aspects and issues pertaining to the use of the

[26] Our translation.
[27] Idem.
[28] Idem.

Internet".[29] A definition that fits Web archiving very well, if it replaces systematically the word "Internet." Indeed, Web archiving practices involve a variety of actors, moved by different motivations; they encompass different and evolving definitions, values, imaginaries of the very notion of Web archiving; they suggest different ways in which control and responsibility can be exerted, be it through social norms, technical standards or policies. Which raises the question: is Web archiving a microcosm of Internet governance? There is no doubt they share several similarities.

A microcosm of Internet governance

Web archiving relies upon a multi-stakeholder model. It is the prerogative of foundations such as the Internet Archive; national institutions; transnational organizations such as IIPC (International Internet Preservation Consortium);[30] civil society (the Archive Team militants, other initiatives by researcher communities); the private sector (e.g. Google, which becomes an actor of born-digital heritage by making Usenet newsgroups available, as previously mentioned).

Observing stakeholders "coming together" entails looking into the scripts (Akrich, 1992) that, embedded in technology, perform relationships between actors, role-sharing and the distribution of competencies; the plasticity of users' technical choices beyond the initial control of inventors and power users; the delegation of rule enforcement to algorithms and automated devices. Some places and spaces, off and on line, are of particular interest to the study of those confrontations between the different stakeholders, and of the models somewhat resembling hybrid forums (Callon *et al.*, 2001). It is the case for the IIPC, which we will come back to, but also for the Wayback Machine.

"As Mark Graham, Director of the Wayback Machine put in an email, the Internet Archive's web materials are comprised of 'many different collections driven by many organizations that have different approaches to crawling.' At the time of this writing, the primary web of the Archive total more than 4.1 million items across 7,357 distinct collections, while its Archive-It program has over 440 partner organizations overseeing specific targeted collections. Contributors range from middle school students in Battle Ground, WA to the National Library of France" (Leetaru, 2016).

The notion of co-construction has made its way into Web archiving, where the main categories of Internet governance actors may be found – as well as their tensions. Collaboration experiences between archiving institutions and researchers are regularly undertaken: e.g., the French National Library recently associated our research team, Web90[31], to a reflection on the implementation of plain text in the Web archives of the Nineties; the INA organized ateliers on the legal deposit of the Web to foster cross-community dialogue;[32] the RESAW network mixes researchers and archiving professionals. The Internet Archive

[30] See netpreserve.org

[31] http://web90.hypotheses.org. This work is supported by the French National Agency (ANR-14-CE29-0012-01).

[32] http://atelier-dlweb.fr/blog/

goes even further, by explicitly promoting bottom-up initiatives: "We thought the machines were going to save us — crawling the web, digitizing the books, organizing the information — but we were wrong," Brewster Kahle says: "communities of people are at the heart of curation" (Kahle in Streitfeld, 2014).

However, Web archiving illustrates the tension between the common good and proprietary formats, and between different imaginaries of the Internet and the Web. In this regard, the missions established by, or delegated to, Web archiving organizations are interesting to observe. Since August 1, 2006, the BnF has the mission of collecting, preserving and communicating Internet sites pertaining to the "French domain" according to the legal deposit. This mission is carried out within the frame of intellectual property law and personal data protection. Out of respect of the former, collections are not accessible online, except in the dedicated libraries rooms. On its end, the legal deposit, instituted in the 16th century, has the objective of preserving the memory of the French editorial production as a whole, whatever the intended audience (scientific results, artistic production, entertainment). This framework can be compared to that of the Archive Team, where the availability of computational resources and the willingness to share them are prominently featured:

"Since 2009 this variant force of nature has caught wind of shutdowns, shutoffs, mergers, and plain old deletions - and done our best to save the history before it's lost forever. Along the way, we've gotten attention, resistance, press and discussion, but most importantly, we've gotten the message out: IT DOESN'T HAVE TO BE THIS WAY[33]."

"This project is composed of volunteers, currently coordinated by Jason Scott.
If you're wondering where to stick your nose in, we could use:

- *Warriors, You will run the Archive Team Warrior on any PC's you have with spare bandwidth. [...]*
- *Writers, who can create clear essays and instructions for archivists and concerned parties.*
- *People with Lots of Hosted Disk Space who have a proper hosted webserver and fat pipe, who are willing (when asked) to consider hosting mirrored dead sites or archives. [...] ".[34]*

In the first case, we see the weight of historical heritage and the sovereignty issues historically carried by the legal deposit – and, in the second case, the link between the individual's technical capacity and his or her ability to contribute. The contributor is understood in its technical and computational capacity, as well as human[35].

Web archiving illustrates the presence of geopolitical tensions too, as Brewster Kahle's September 2014 appeal shows:

"China started blocking the Internet Archive again a couple of months ago, we believe, because they do not like our open access policies. In this way, we have started to understand

[33] http://archiveteam.org/index.php?title=Main_Page
[34] http://archiveteam.org/index.php?title=Who_We_Are
[35] http://archiveteam.org/index.php?title=Dev

the power in the hands of the Internet service providers. Let's keep our access to Internet sites 'Neutral' and not at the discretion of companies and governments" (Kahle, 2014).

Finally, one of the first Internet governance typologies emphasized the plurality of governance systems, from technology to the market, from international and/or transnational concertation to non-legal standards and law (Bygrave and Bing, 2009); this dialectic between different practices and sources of normativity, concurring or complementary, may as well be found in Web archiving. The alleged "independence of cyberspace" reflects in the Archive Team's flamboyant motto ("We are going to rescue your shit!"), and its rescue of Geocities from Yahoo!'s shutdown is its best example. The "governance by markets" finds its equivalent in the collection/capture of private data and archives by Twitter and Facebook. The role of national as well as international ("traditionally political") institutions is exemplified by the legal deposit and by entities such as the UNESCO chart, the Internet Archive, the International Internet Preservation Consortium (IIPC) – with nuances ranging from the "international" as the sum of national initiatives, to the "transnational" approaches. The IIPC case also hints at the issue of technical governance, leading to a possible reflection on the place of experts and of standards.

To sum up, Web archiving reactivates the same polarizations, negotiations and dynamics between actors which had emerged at the time of Internet governance's birth, notably during the World Summit on the Information Society (WSIS) held in Geneva in 2003, and Tunis in 2005. It is therefore unsurprising to discover, in Web archiving, another "classical" issue of Internet governance such as the digital divide: Web archives mirror the fact that the present-day digital world is still developing unevenly.

DIGITAL DIVIDE

A striking feature of the Web archiving community is that it comprises almost exclusively institutions from the "Global North" (Gomes *et al.*, 2011). Indeed, developing countries' presence on the archived Web is far from proportional to their growing presence on the live one.

This, however, is beginning to evolve, as initiatives to preserve the Web are currently developing in the "Global South": Chile, South Africa, China and Malaysia are in the process of developing Web archives; and a 2010 study of the Diet National Library of Japan for the CDNLAO[36] showed a strong interest of South-East Asian national libraries in setting up Web archives; as expressed by Indonesia, Fiji, Mongolia, Papua New Guinea, Nepal, the Maldives, Vietnam and Sri Lanka. When asked about the obstacles they faced in their respective countries, respondents repeatedly cited technical infrastructure and knowledge, as well as a general lack of awareness on the part of governments about Web archiving that led to legal obstacles (legal deposit legislation not up-to-date). The kind of help they were most interested in from the more developed Web archives of the CDNLAO (South Korea, Japan, China and Singapore) was training for librarians, and the sharing of technical knowledge and project planning techniques (National Diet Library of Japan, 2010). Although this study was

[36] The Conference of Directors of National Libraries in Asia and Oceania.

conducted in Asia and Oceania, one could also expect national libraries in other countries and continents to share those obstacles and needs.

Other noteworthy points highlighted by this survey were that the libraries with operational Web archives encouraged the others to join the IIPC, and expressed interest in the CDNLAO as a regional forum for Web archiving cooperation amongst member states to focus on region-specific issues, such as the development of software specifically tailored to contents published in Asian alphabets (IIPC software is developed for the Western alphabet). This suggests that in addition to the IIPC – which operates at the global level, regional associations could also be important actors of Web archiving in the future, serving as sub-forums for states to exchange on issues specific to their regions, and to coordinate practical skill-transfers from the more developed archives to the least, using geographical and cultural proximity to their advantage.

However, there are still regions of the world – and thus of the Web, which remain largely un-archived. Notably, India, the Middle East, Latin America and Africa are populous regions where the Web is rapidly changing and expanding, but is vulnerable to oblivion for lack of indigenous archiving initiatives. The fact that most of the states in these regions do not archive their national Web spheres, nor plan to do so in the near future constitutes a considerable risk of losing valuable cultural and scholarly sources; and poses the moral questions of whether their Web spheres should be archived, and if so, by whom. Indeed, for now the 'Northern' research community has taken upon itself to preserve political websites in some developing countries: in 2001, the DACHS archive at the University of Heidelberg was launched, to preserve the Chinese socio-political Web; in 2004, the feasibility study "Political Communications Web Archiving" was published under the leadership of four major American universities and followed by the launches, in 2005 of the Latin American Web Archiving Project based at the University of Texas at Austin, and in 2008 of the Web Archiving Project for the Pacific Islands based at the University of Hawaii.

Who should select what is to be archived, so that the collection is not biased? Should consent of website owners be obtained? Who should be able to access the archives, and under what conditions? In their 2004 article, Lor & Briz review those moral challenges specific to the archival of websites of the "Global South" by countries of the "Global North," and put forward a comprehensive framework for such practices based on human rights, social justice and the African Studies' Association (ASA) guidelines for ethical research (Lor and Britz, 2004: 457).

CONCLUSION

In 2006, historian Paul E. Ceruzzi said: "The Internet is a technological construction with a magnitude and scope comparable to the hydroelectric dams, railroads, aircraft, and electric power systems of an earlier era." Indeed, Web archives and Web archiving need, as other sets of practices and devices building upon and embedded in the Internet, a socio-technical de-construction that takes the mundane, the material, the invisible, the automated and semi-automated agency of technical artefacts, fully into account. The processes subtending Web archiving, from the dialogues conducted in international, multi-stakeholder fora to the user-to-user sharing of a password and identifier in order to add material to the Internet Archive, all contribute to the present-day negotiations on the perimeter, the definition, the shaping of the Web of the past – and as such, shed light on formal and informal mechanisms of Internet governance.

This paper has explored how an approach that gives conceptual and analytical prominence to the "unpacking" of socio-technical black-boxes, and links them to governance issues, has two main implications for the analysis of Web archiving. On one hand, it provides yet another illustration – less prominent perhaps, but no less interesting than Internet governance itself – of the distributed, diffused and technology-embedded nature of power in the age of digital networks (DeNardis, 2014). On the other, it also provides novel arenas to observe the extent to which data, the oft-alleged "fundamental stuff of truth itself" are a cultural resource that, as such, is – and needs to be – generated, protected, interpreted (Gitelman, 2013), via a set of conceptual, political and design choices that are anything but a given.

This paper has provided myriad examples of how the study of infrastructure – what Geoffrey Bowker has described as "pervasive enabling resources in network form" and Susan Leigh Star as the "invisible work" underlying practices, uses and exchanges in a networked system – helps us to understand the impact of perimeters of archiving, institutions' missions, frequencies of captures, actors and budgets, tools and protocols used by such actors, access modalities, on Web archiving and the processes of "heritagization" of the digital. Archiving practices are also affected, in parallel, by geopolitical issues concerning multi-stakeholderism in Web archiving, by "classic" political questions such as the digital divide, the new silos and knowledge infrastructures created by national Web archives, the techno-legal aspects of Web infrastructure. In parallel, the design becomes prescription due to passwords, bottlenecks, clashes of formats – the several socio-technical mediations undertaken by Web archives. As shown by Niels Brugger (2012), archives are rarely the same thing as the original website – this paper has explored the different levels at which there is agency of both human and machines that lead to these transformations, and the intersections of these levels.

Just as Internet governance is in-the-making and not yet stabilized, after at least 2001, Web archives "naturally" join some of the questions related to it, while drawing from previous experiences and lessons learned[37]. Will they clash with controversies and tensions as well, as the broader Internet governance ecosystem does (Musiani, 2015)? How will Web archives further illustrate the distribution of power on the Internet, as well as its bottlenecks and choke points (DeNardis and Musiani, 2016)? Should we consider, as we do Internet governance, the governance of Web archives as a political and geopolitical issue as well as a socio-technical one? As an ecosystem constrained and enabled by markets, technology and practices as well as institutions (Massit-Folléa et al., 2013)? This paper has suggested several reasons why this may be the case, and has hopefully shed light on the value of analytical instruments borrowed from Science and Technology Studies in order to explore them.

[37] As highlighted by the mission of IIPC: "The IIPC is a membership organization dedicated to improving the tools, standards and best practices of web archiving while promoting international collaboration and the broad access and use of Web archives for research and cultural heritage." http://netpreserve.org/general-assembly/general-assembly-2014-schedule

REFERENCES

Abbate, J. (2000) *Inventing the Internet*. Cambrige, MA: The MIT Press.

Abbate, J. (2012). L'histoire de l'internet au prisme des STS. *Le Temps des medias*, 18, 170-180.

Akrich, M. (1992). The De-scription of Technical Objects. In W. Bijker, J. Law (eds.), *Shaping Technology/Building Society. Studies in Sociotechnical Change*, 205-224. Cambridge, MA: The MIT Press.

Ankerson, M.S. (2015). Take me back! Web history as chronotourism of the digital archive. *Times and Temporalities of the Web International Symposium*, Paris.

Ankerson, M.S. (2015). Read/Write the Digital Archive: Strategies for Historical Web Research. In E. Hargittai, C. Sandvig (eds.), *Digital Research Confidential. The Secrets of Studying Behavior Online.* Cambridge, MA: The MIT Press.

Boullier, D. (2015). Les sciences sociales face aux traces du big data : société, opinion ou vibrations ? *Revue Française de Science Politique*, 65, 805-828. Retrieved from https://halshs.archives-ouvertes.fr/halshs-01141120/

Bourdeloie, H. (2013). Ce que le numérique fait aux sciences humaines et sociales. *tic&société*, 7(2). Retrieved from: https://ticetsociete.revues.org/1500#text

Bowker, G.C., Baker, K., Millerand, F., Ribes, D. (2010). Toward Information infrastructure Studies: Ways of Knowing in a Networked Environment. In J. Hunsinger, L. Klastrup, M. Allen (eds.), *International Handbook of Internet Research*, 97-117. Dordrecht, Netherlands: Springer.

Bowker, G.C., Timmermans, S., Clarke, A.E., Balka, E. (2016). *Boundary Objects and Beyond*. Cambridge, MA: The MIT Press.

Brügger, N. (2012). Web History and the Web as a Historical Source. *Zeithistorische Forschungen/Studies in Contemporary History*, 9, 316-325. Retrieved from: http://www.zeithistorische-forschungen.de/2-2012/id%3D4426

Bygrave L., Bing J. (eds.). (2009). *Internet Governance. Infrastructure and Institutions*. Oxford: Oxford University Press.

Callon M., Lascoumes P., Barthe Y. (2001). *Agir dans un monde incertain. Essai sur la démocratie technique*. Paris: Seuil.

Ceruzzi, P. (2006). The Materiality of the Internet. *IEEE Annals of the History of Computing*, vol. 28, no. 3, 96, c3.

Citton, Y. (ed.) (2014). *L'économie de l'attention : Nouvel horizon du capitalisme ?*. Paris: Éditions La découverte.

Claburn, T. (2007). Colorado Woman Sues To Hold Web Crawlers To Contracts. Retrieved on February 18, 2016 from the *InformationWeek* website: http://www.informationweek.com/colorado-woman-sues-to-hold-web-crawlers-to-contracts/d/d-id/1053075

Cohen, E., Verlaine, J. (2013). Le dépôt légal de l'internet français à la Bibliothèque nationale de France. *Sociétés & Représentations*, 1(35), 209-218.

Davallon, J. (2006). *Le don du patrimoine : Une approche communicationnelle de la patrimonialisation*. Paris, Hermes sciences publications.

Delalande, N., Vincent, J. (2011). Portrait de l'historien-ne en cyborg. *Revue d'histoire moderne et contemporaine*, 5(58-4bis), 5-29.

DeNardis, L. (2014). *The Global War for Internet Governance*. New Haven: Yale University Press.

DeNardis, L. and Musiani, F. Governance by Infrastructure. In F. Musiani, D. Cogburn, L. DeNardis, and N. Levinson, *The Turn to Infrastructure In Internet Governance*, 3-21. New York: Palgrave-Macmillan.

Edwards, P.N. and al. (2012). Knowledge Infrastructures: Intellectual Frameworks and Research Challenges. *Report of a workshop sponsored by the National Science Foundation and the Sloan Foundation University of Michigan School of Information, 25-28 May*. http://pne.people.si.umich.edu/PDF/Edwards_etal_2013_Knowledge_Infrastructures.pdf

Finnemann, N.O. (2015). Hypertextual relations in digital born materials Hypertext and time: Towards a genre analysis of heterogeneous digital materials. *RESAW Conference*, Aarhus University, Denmark, 8-10 June.

Flichy, P. (2001). *L'imaginaire d'Internet*. Paris: La Découverte.

Flyverbom, M. (2011). *The Power of Networks: Organizing the Global Politics of the Internet*. Cheltenham, UK: Edward Elgar Publishing.

Gitelman, L. (ed., 2013). *"Raw Data" Is An Oxymoron*. Cambridge, MA: The MIT Press.

Gomes, D., Miranda, J., Costa, M. (2011). A Survey on Web Archiving Initiatives. *TPDL'11 Proceedings of the 15th international conference on Theory and practice of digital libraries:*

research and advanced technology for digital libraries. Berlin/Heidelberg/New York: Springer.

http://www.researchgate.net/profile/Miguel_Costa4/publication/221176373_A_Survey_on_Web_Archiving_Initiatives/links/004635276bb1ddf6e7000000.pdf

Helmond, A. (2013). The Algorithmization of the Hyperlink. *Computational Culture*. Retrived from: http://computationalculture.net/article/the-algorithmization-of-the-hyperlink

Hert, P. (1998). *Jeux, écritures, espaces d'énonciations. Contribution à une étude anthropologique de l'usage d'Internet en milieu scientifique* (doctoral thesis). Université Louis Pasteur – Strasbourg I, Strasbourg, France.

Hitchcock, T. (2015). The UK Web Archive, born-digital sources, and rethinking the future of research. Retrieved on February 18, 2016 from the *Web Archives for Historians* blog: http://webarchivehistorians.org/tag/tim-hitchcock/

Huuderman, H., Kamps, J., Samar, T., de Vries, A., Ben-David, A., Rogers, R. (2015). Lost but not forgotten: finding pages on the unarchived web. *International Journal on Digital Libraries*, 16, 3, 247-265. DOI: 10.1007/s00799-015-0153-3

Kahle, B., Streitfeld, D. (2014). The Internet Archive, Trying to Encompass All Creation. Retrieved on February 18, 2016 from *the New York Times* website: http://bits.blogs.nytimes.com/2014/10/31/the-internet-archive-trying-to-encompass-all-creation/

Kahle, B. (2014). Please Help Protect Net Neutrality. Retrieved on February 18, 2016 from the *Internet Archive Blogs* blog: https://blog.archive.org/2014/09/10/please-help-protect-net-neutrality/

Kessous, E. (2012). *L'attention au monde : Sociologie des données personnelles à l'ère numérique*. Paris: Armand Colin, coll. « Recherches ».

Latzko-Toth, G. (2010). *La co-construction d'un dispositif socio-technique de communication : le cas de l'Internet Relay Chat* (doctoral thesis). Université du Québec à Montréal, Montréal, Canada.

Leetaru, K. (2016). The Internet Archive Turn 20: A Behind the Scenes Look at Archiving The Web. Retrieved on February 18, 2016 from the *Forbes Magazine* website: http://www.forbes.com/sites/kalevleetaru/2016/01/18/the-internet-archive-turns-20-a-behind-the-scenes-look-at-archiving-the-web/#5e85d7887800

Lialina, O. (2005). A Vernacular Web. *Extended version of a talk at the Decade of Web Design Conference in Amsterdam, January 2005*. Retrieved from: http://art.teleportacia.org/observation/vernacular/

Lor, P., Britz, J. (2004). A moral perspective on South-North web archiving. *Journal of Information Science*, 30(6), 540-549

Massit-Folléa, F., Méadel, C., Monnoyer-Smith, L. (eds., 2012). *Normative Experience in Internet Politics*. Paris: Presses des Mines.

Merzeau, L. (2014). Vers un Web temporel. Talk at the IIPC General Assembly. Retrieved from : http://merzeau.net/vers-un-web-temporel/

Milligan, I. (2012). Mining the Internet Graveyard: Exploring Canada's Digital Collections Projects. *Presentation at the Canadian Historical Association Annual Meeting*. Retrieved from: https://ianmilligan.ca/2012/05/28/mining-the-internet-graveyard-exploring-canadas-digital-collections-projects-presentation-at-the-canadian-historical-association/

Mourlhon-Dallies, F., Colin J-Y. (1995). Les rituels énonciatifs des réseaux informatiques entre scientifiques. *Les Carnets du Cediscor*, 3, 161–172.

Musiani, F. (2015). Practice, Plurality, Performativity and Plumbing: Internet Governance Research Meets Science and Technology Studies. *Science, Technology and Human Values*, 40(2), 272-286.

Mussou, C. (2012). Et le Web devint archive: enjeux et défis. *Le Temps des médias*, 2(19), 259-266.

National Diet Library of Japan. (2010). CDNLAO Questionnaire Survey on Web Archiving (Q2-Q7), Document 2. In *Report on the Questionnaire Survey on Web Archiving, at the 18th CDNLAO Annual Meeting 2010*. Retrieved from: http://www.ndl.go.jp/en/cdnlao/meetings/2010.html

Paloque-Berges, C. (in press, 2016). Patrimoine des sciences et des techniques du numérique : vers des lieux de mémoire réticulaires ?. *RESET*.

Paravel, V. (2007). De la plume d'oie à la souris : la recherche en réseaux. In C. Jacob (ed.), *Lieux de Savoir (vol. 1). Espaces et Communautés*, 1095–1118. Paris: Albin Michel,

Pehlivan, Z. (member of the R&D team of the Ina legal deposit service). (2016, 11 February). Email to Valérie Schafer.

Proulx, S. (2009). L'intelligence du grand nombre : la puissance d'agir des contributeurs sur Internet – limites et possibilités. *7ème colloque du chapitre français de l'ISKO, Intelligence*

collective et organisation des connaissances, Lyon, 24-26 juin 2009. http://pro.ovh.net/~iskofran/pdf/isko2009/PROULX.pdf

Rogers, R. (2015). Au-delà de la critique big data. La recherche sociale et politique à l'ère numérique. In M. Severo, A. Romele (eds.) *Traces numériques et territoires*, 18. Paris: Presses des Mines.

Schafer, V. (2015). *En construction. Une histoire française du Web des années 1990*. HDR, volume 2. Université Paris-Sorbonne, Paris, France.

Schafer, V., Thierry, B. (2015). L'ogre et la Toile. Le rendez-vous de l'histoire et des archives du Web. *Socio*, 4, 75-96.

Star S. L., Griesemer J. (1989). Institutional Ecology, "Translations" and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science*, 19(3), 387- 420.

Tufekci, Z. (2014). What Happens to #Ferguson Affects Ferguson: Net Neutrality, Algorithmic Filtering and Ferguson. Retrieved on February 18, 2016 from *The Message* blog: https://medium.com/message/ferguson-is-also-a-net-neutrality-issue-6d2f3db51eb0#.gvv8qfuq8

Unesco. (2003). *Charte sur la conservation du patrimoine numérique*. Retrieved from: http://portal.unesco.org/fr/ev.php-URL_ID=17721&URL_DO=DO_TOPIC&URL_SECTION=201.html