First Monday, Volume 20, Number 5 - 4 May 2015

PEER-REVIEWED JOURNAL ON THE INTERNET

Learning from failure: The case of the disappearing Web site

Francine Barone, David Zeitlyn, and Viktor Mayer-Schönberger

Abstract

This paper presents the findings of the *Gone Dark Project*, a joint study between the Institute of Social and Cultural Anthropology and the Oxford Internet Institute at Oxford University. The project has sought to give substance to frequent reports of Web sites "disappearing" (URLs that generate "404 not found" errors) by tracking and investigating cases of excellent and important Web sites which are no longer accessible online. We first address the rationale and research methods for the project before focusing on several key case studies illustrating some important challenges in Web preservation. Followed by a brief overview of the strengths and weaknesses of current Web archiving practice, the lessons learned from these case studies will inform practical recommendations that might be considered in order to improve the preservation of online content within and beyond existing approaches to Web preservation and archiving.

Contents

Introduction

Conducted during 2014, the *Gone Dark Project* has investigated instances of Web sites that are no longer online and which have not been captured by the Internet Archive or other archiving initiatives. We wanted to examine what has happened to Web sites, valuable archives and online resources that have disappeared, been shut down, or otherwise no longer exist publicly on the Internet. Web archiving services, including national libraries such as the British Library and U.S. Library of Congress as well as non-profit organizations like the Internet Archive, are dedicated to storing the contents of the Web and have had great success in preserving online content as part of recent human history (see, for example, BBC News, 2010; Lohr, 2010; Internet Archive, 2014). Despite these efforts, however, some important content has not been archived. Other research (cited below) shows that the lifespan of online content is pitifully short. The average lifespan of a Web page is difficult to determine, but estimates put it at a mere 100 days in 2003, up from just 44 days in 1997 (Taylor, 2011; Barksdale and Berman, 2007). As the Web evolves, different types of content become more susceptible to loss. In 2008, a survey by blog search engine Technorati found that a whopping 95 percent of its 133 million tracked blogs had been "abandoned" or not updated in 120 days (Quenqua, 2009). Ironically, in May 2014, Technorati unceremoniously shut down its famously extensive blog directory — once an indispensable tool to the blogosphere — with no prior announcement and little to no media coverage (Bhuiyan, 2014).

A study published in 2012 (SalahEldeen and Nelson, 2012) reveals that historically significant social media content decays at an alarming rate with 11 percent of timely media content lost within one year, rising to nearly 30 percent in two years (at a rate of .02 percent of shared resources lost per day). Compounding the problem of disappearing Web sites is the issue of link rot or hyperlink decay. In a study of academic references, Zittrain, *et al.* (2013) found that over 70 percent of URLs in academic journals and 50 percent found in U.S. Supreme Court opinions have broken or no longer link to the original citation information.

These losses and “gray areas” on the Web suggest that between what is automatically crawled and saved and what becomes lost without much impact in day-to-day activities on social media, lies a large swath of the Internet that we know very little about in terms of historical record. Continual reports of missing Web sites and 404 errors suggest that there are still Web pages that “go dark” on a regular basis. In fact, despite Google’s Cache, the Internet Archive’s Wayback Machine [1] and national digital preservation initiatives, it is still easy for a site to be completely lost from the public Web. Is there significant data loss in these situations? Even Web crawlers that have captured billions of pages cannot save all the content from sites before they vanish, especially if those sites are not widely known and/or indexed by major search engines or if the content is held in a database inaccessible to crawlers. There may be cases where the data is still held privately or off–line where Web crawlers cannot find it. Can it still be recovered?

In the light of this, the Gone Dark Project wanted to address the concern that there may be instances of culturally valuable Web sites which are no longer online and whose disappearance represents a major public or social loss. What, if anything, can be done to mitigate future losses of this kind?

As a collaborative project between Oxford Anthropology and the Oxford Internet Institute, this project benefited from both a technical and anthropological approach to the subject of digital content loss. We were able to investigate actual cases of content loss on the Web, including interviewing the original content owners or other involved parties, in order to better understand current practices and inform future innovations in pragmatic Web preservation.

Research methods and process

The Gone Dark Project was conducted over nine months from February to October 2014.

The questions guiding the research were:

1. How significant and/or widespread a problem is the disappearance of Web sites?
2. What common factors result in important Web content not being archived?
3. What practical steps or changes to Web preservation practices and/or policy can be identified to mitigate against reoccurrence in the future?

We should clarify that our principal concern was with sites which contain substantial or significant content, rather than either social media posts or collections of links to other sites. We explain this in more detail below.

The first task undertaken was to identify as many cases as possible of such sites known once to have existed, but which are no longer publicly available online (especially those that are not well-archived in some form or other). This allowed us to gauge the scope of the problem of Web sites ‘going dark’. It also brought up methodological challenges; notably, how to find digital artifacts that no longer exist by looking for clues around the Web.

Searching for ghosts of Web sites required pragmatic methods that evolved over the course of the research. For example, creative use of search engine filters and existing archive resources such as Google’s Cache and the Internet Archive’s Wayback Machine was essential. While reference was made in the initial scan to previously compiled lists of dead or endangered Web sites, apps and services (*e.g.*, ArchiveTeam’s DeathWatch [2]) these were largely of limited utility because most sites on popular lists were deemed to lie outside the scope of the Gone Dark Project (see below).

Potentially relevant sites or directories of pages were scanned for dead links using automated link checkers to find broken references to databases, repositories, or archives of original content. We also conducted some manual trawling through lists of links on older sites, including academic indexes. These links were then followed up to collect background information about the nature of the site, reasons for its disappearance and whether a public archived copy exists. For paradigm cases attempts were made to contact the relevant site owners and interviews were conducted. We wanted to learn about what happened to the site, how its loss might have been — or might still be — avoidable; and to trace the current whereabouts of the original content.

We also distributed links to the project via social media platforms, academic mailing lists and user forums. We especially encouraged academics and subject experts to send us information about content-rich sites and databases that might no longer exist.

On a day-to-day basis, social media accounts on Twitter and Facebook were used to foster dialogue with individual Web archivists and large organizations directly engaged in Web preservation, including the U.K. Web Archive [3], Austrian Web Archive [4], Internet Archive [5], NDIIP (U.S. Library of Congress) [6] and Internet Memory Foundation [7], among others. Throughout the project, active engagement with specialists in the fields of Web archiving, Web history and digital humanities helped to identify case studies of sites that have gone dark as well as to better understand the processes and professional standards for crawling and archiving the Web that are currently in place around the world. Informal surveys and interviews with social media followers were fruitful in pointing out strengths and weaknesses in current practices. The social media channels greatly informed the data analysis and final recommendations of the project.

Beyond link rot

Many of the reports of sites no longer available are attributable to 'linkrot'. In these cases, the original or referring URL no longer works (for many reasons) generating a 404 not found error. However, many sites whose published URLs no longer work do still exist, but at other URLs (site restructuring or redesign may break many links, as clearly needs to be pointed out to Web designers and the site managers who employ them). Even where the original site is no longer available, its content may have been preserved through one of the many initiatives to archive Web content. There are cases, however, as we shall see, in which the content in question has not been captured. These often involve sites where content is delivered by a user-searchable database such as a catalogue.

One such case in point is the Haddon catalogue developed by Marcus Banks in Oxford with support from the U.K.'s Economic and Social Research Council (Grant R000235891 [8]). The project sought to document early (pre-World War II) ethnographic films. In the course of the research, some 1,000 instances were identified and information about them was made available via a searchable catalogue which went live in 1996. As database and server technology advanced it became unavailable: the database engine was no longer compatible with current operating systems and it went down in 2005 [9]. Since the original project funding had finished, there were no longer resources available to reprocess the data (which had been securely archived) to make it available again using a different database engine. Now some support has been offered by colleagues in Manchester, so it is hoped that access to the Haddon catalogue will resume in 2015 or 2016, after 10 years of 'darkness'.

Typology of sites gone dark

After canvassing as many known defunct Web sites as possible across all fields of interest, the second task was to categorize our initial findings into manageable types. Potentially relevant case studies were organized by theme and the primary reason for the page's disappearance. This process helped to make more sense of the wider landscape of what can be described as the vanishing Web — sites, pages and genres of content that have gone as well as are in the process of going dark, and especially those which appear to be at greater risk of doing so, either for specific reasons or simply because they have a higher rate of unintentional decay.

Main types of sites:

1. Scientific and Academic: Databases, research tools and repositories ranging from the natural to social sciences and humanities. Losses of this type are commonly the result of the end of funding or institutional neglect, in which case the original data may still be held (*e.g.*, on university servers).
2. Political: Personal homepages of politicians, campaign pages, political speeches and/or repositories of once-public government files. Some journalistic sites also fall under this category (we discuss a case below).
3. Historical and Cultural: A range of sites with different origins fall within this category, including collated collections of historical documents, genealogies or research portals, as well as more professionally run film, video or music archives.
4. "Labours of love": Specialized project pages or information aggregation sites, typically self-hosted and curated by independent individuals with little to no institutional backing.
5. Social media: These include popular Web services on sites run by companies such as Google [10], Yahoo! [11] or Microsoft [12], including blogging platforms, social networking sites and other utilities that change hands or have been retired since social media platforms and startups evolve quickly and come and go easily, often leaving behind users with data they would prefer to keep (examples include several popular Web services from the late 1990s).

Main reasons for sites disappearing:

1. Neglect: Intentional or unintentional neglect is probably the most common reason that a site disappears, including allowing domain registrations to expire; losing or not updating files; and not keeping adequate backups.
2. Technical: Technical issues are usually bundled with some form of neglect or insufficient financial resources. Purely technological reasons for content loss include hardware malfunction, viruses, Web host errors and accidental file deletion.
3. Financial: A common factor among sites gone dark is the cost of site maintenance (hosting fees and/or server maintenance, plus staff costs where relevant.)

4. Natural disaster: Computer hardware is susceptible to fires, floods, rioting and neglect (just as are paper files). Although in principle "Lots of Copies Keeps Stuff Safe", for many reasons the many copies may not have been made or distributed.
5. High-risk situations: Tumultuous political climates are a nightmare for data loss. Sites can be shut down intentionally by hostile regimes or otherwise lost during human rights crises. Legal prosecution or the threat of this can lead to the removal of material: the international legal saga about the availability of material to do with the Church of Scientology is a case in point [13].
6. "Web wars": Competition between top Web companies such as Google, Yahoo!, MSN and AOL leads to aggressive acquisition of popular services that are subsequently abandoned, shut down or absorbed into a larger platform.

An anonymous *First Monday* reviewer points out that the interconnections between neglect, financial constraints, and technical issues are particularly insidious. A sort of fatal creeping obselescence can occur that is caused by a mix of under-funding, lack of investment in technical updating and neglect that is very different from a simple site crash or attack that exposes that the backups had not worked.

On the whole, it was clear that a) some sites are going dark across the Web without being archived and b) those sites vary widely in size, type and content. This confirmation is in itself a significant finding. However, the main aim for the Gone Dark Project was to focus on sites of particular socio-cultural value that constitute an irretrievable loss notably marked by large amounts of content not likely to be saved/crawled by automated software. Abandoned blogs, deleted user profiles and short-lived Web app startups are all digital losses, yet they have largely become accepted and even expected within the landscape of the Internet today. As the Web evolves, people move on and leave a patchy trail of online interactions in their wake. The dynamics and ethics of digital preservation of content like personal social media postings remain debatable and outside the scope of this paper (see Mayer-Schönberger, 2009).

While each and every site that goes dark arguably constitutes a lost piece of Internet history, the case studies chosen for deeper investigation were selected on the basis of being of cultural, heritage or social value whose loss represents a cautionary tale for Web preservation. There is certainly a considerable element of personal judgment about what constitutes an 'important' Web site containing 'valuable' material. Recognizing this, we would also observe that this is by no means a new problem: all archivists have always had to make decisions about what to include and what to reject from inclusion in the archive under their control. These judgments (albeit individually questionable) often include assessments of what future researchers, 'users', or lawyers might find helpful. The decision to archive may not be clear cut in any one case, but the intuition behind it remains clear. As one research discussing a film archive has it, an archive is a bet against the future — betting that these records will be found useful [14].

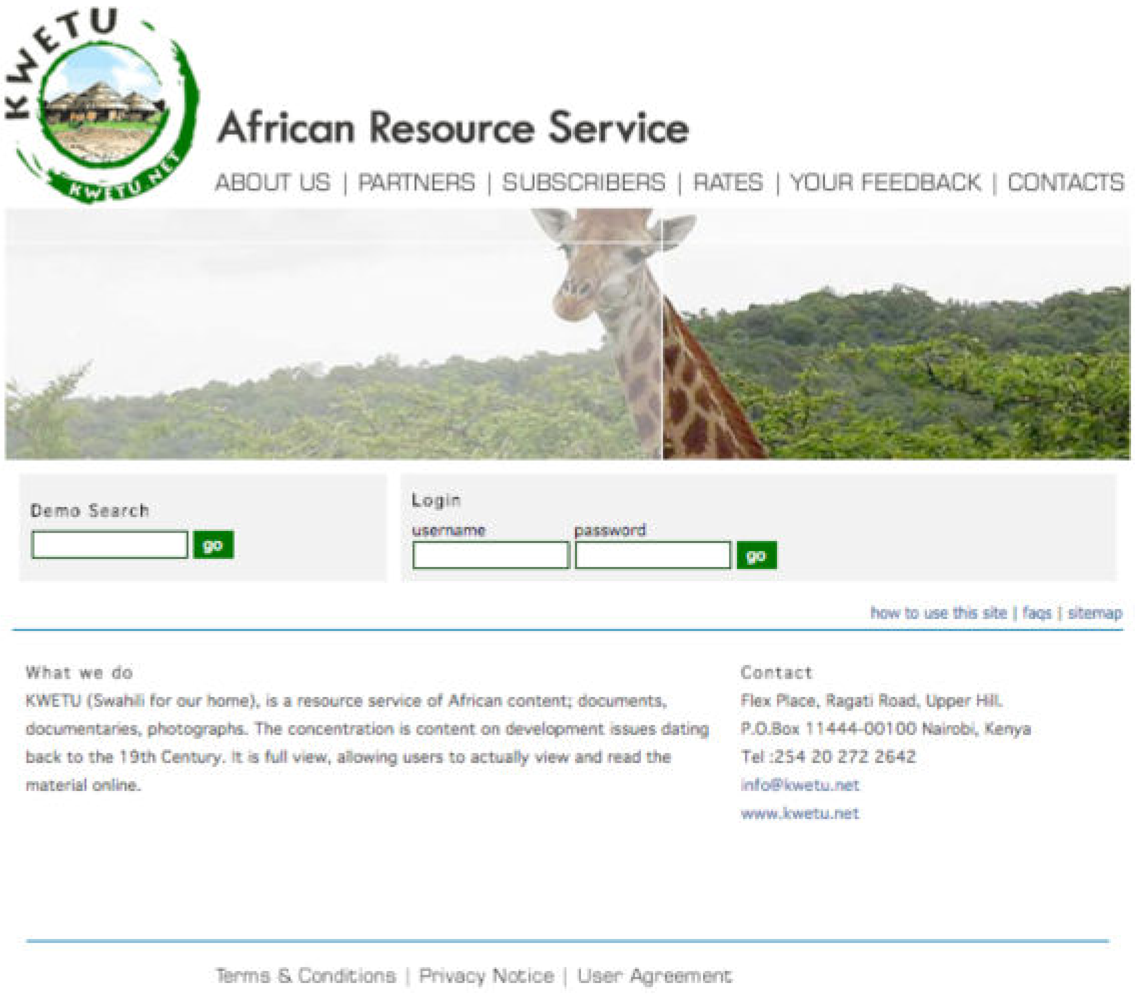
Once categorized, a more narrow focus was taken for the remainder of the project. The following section will focus on a selected number of illustrative cases of sites gone dark, including what happened to the data, if it still exists; and to interpret how each case can inform recommendations for future prevention. Rather than simply collect lists of defunct and unarchived pages, the Gone Dark Project sought out the original content owners in order to discover the individual stories behind the 404 error page, or, more simply, to find out what happens when a Web site dies.

Case studies

The selected case studies below illustrate various ways in which valuable Web sites can go dark. The first two examples represent instances where important digital resources that were once available online have gone dark for an extended period of time. In these cases, the original content still exists, but the challenge is making it available again. The final case takes a different tack, focusing on potentially more widespread but difficult to quantify circumstances with potential impact throughout extensive portions of the Web. Together, the cases illuminate where existing Web archiving practices are insufficient due to the fleeting and impermanent nature of some Web content as well as obstacles impeding content owners' ability to archive important data in a usable format before it is too late.

Kwetu.net

Kwetu.net [15] ("our home" in Kiswahili) was a privately owned grey literature site established in 2000 by Karani Nyamu and Luke Ouko as an online repository of photos and videos from Kenya, Uganda and Ethiopia [16], eventually expanding to include content from over 30 African countries. The same year, the *Economist* had declared Africa "The Hopeless Continent" (*Economist*, 2000). The rationale behind Kwetu.net, according to its founders, was to disprove that singular narrative and counteract the dominant, lopsided portrayal of Africa as a continent of war, poverty, disease and corruption. With Kwetu.net, they intended to make accessible to the rest of the world informative documents that would showcase Africa in a more positive light.


KWETU
KWETU NET
African Resource Service
ABOUT US | PARTNERS | SUBSCRIBERS | RATES | YOUR FEEDBACK | CONTACTS
Demo Search
go
Login
username
password
go
how to use this site | faqs | sitemap
Terms & Conditions | Privacy Notice | User Agreement

What we do

KWETU (Swahili for our home), is a resource service of African content; documents, documentaries, photographs. The concentration is content on development issues dating back to the 19th Century. It is full view, allowing users to actually view and read the material online.

Contact

Flex Place, Ragati Road, Upper Hill.
P.O.Box 11444-00100 Nairobi, Kenya
Tel :254 20 272 2642
info@kwetu.net
www.kwetu.net

Figure 1: A cached copy of Kwetu.Net's login page from 15 December 2005, accessed via Archive.org.

Its mission was therefore to offer the world a wide range of African content resources — in the form of "grey" literature focusing on health, social welfare and development issues in Africa — and to provide access to it irrespective of place and time [17] (see Figure 1). The type of documents sought by Kwetu.net included unpublished reports, baseline surveys, speeches, photos, videos, university theses and a mix of historical (dating as far back as the nineteenth century) and present-day information drawn from many fields or industries — agriculture, healthcare, conservation and politics to women's issues and urban development — that was otherwise difficult to find or largely inaccessible online [18]. As one founder put it, "if it was grey material, we sought it out". Much of this material was acquired from PDF documents produced by education and research institutions, government agencies or NGOs. In Kenya, the founders also partnered with national archives such as Kenya Railways and the African Medical Research Foundation to source content. At one point, Kwetu.Net had a list of 47 partners listed on their site [19].

Access was available on the following basis. The site offered a free demo/preview to all visitors, but access to the full site content was available by paid subscription only. Subsidized rates were available to African educational institutions and subscription fees for other institutions varied from US$800 to US$6,000 depending on the institution's size. Individuals could also subscribe for US$50 per annum. To secure content from producers and owners, there was a services-for-content system in place. For instance, a 30 percent subscription discount was applied to institutions that provided them with content [20]. Similarly, Kwetu.net had developed a network of correspondents in over 30 countries who helped to secure new content for the time they were in active operation. At their peak, the site had a subscription base of 15 African and U.S.-based universities, according to the founders. They also served a range of other institutions such as think tanks, embassies and foreign missions, civil society and donor agencies.

In terms of functionality, the founding team built the site from scratch, including a search engine [21] and an (A-Z) index-tagging system, with a diverse range of tags. This became extremely technically demanding. Extensive amounts of time would go into negotiating with the various content producers, uploading, curating, tagging and indexing the content to ensure ease of access and searchability. Over time, the demand for more content compounded this challenge. Even when the site reached upwards of one million manuscripts in its database, it became clear that it could not supply the demands for content put on it by its paying customers. Furthermore at this point, demand was not coming from local and regional universities — primarily because of low penetration rates and high Internet costs — which stalled the spread of a localized user base.

On top of the technical challenges, the primary motivation that led to Kwetu.net going dark was financial. The founders initially established the site as a "labour of love". Soon, the maintenance costs to keep it afloat — including paying the 30 correspondents connected to the site — exceeded the subscription revenues. The founders were therefore compelled to divert their attention to income-generating projects, and eventually away from Kwetu.net. A one-day delay in paying for renewal constituted in the loss of kwetu.net domain (www.kwetu.net is now hosted in Istanbul as a Turkish tourism site) and the team was unable to recover it.

The site officially went off–line in 2004, according to Nyamu, one of its founders. The original site including HTML, text and images is cached in the Wayback Machine, making it possible to view the skeleton of the old site. However, the search function does not work and no access to anything behind the search paywall is available, which is what made this a case of concern for the Gone Dark Project, since the paywall/database combination made the material inaccessible to the Web crawlers. Upon further investigation and interviews with the site's founders, it was possible to find out more about the large collection of data that was once available through Kwetu.net's search portal. Although it no longer exists on the Internet, the content that had been meticulously curated is still in the hands of the founders. The team, though now working on other ventures, is still passionate about what they had started and stated (in interviews in the course of this research) their interest in reviving the project.

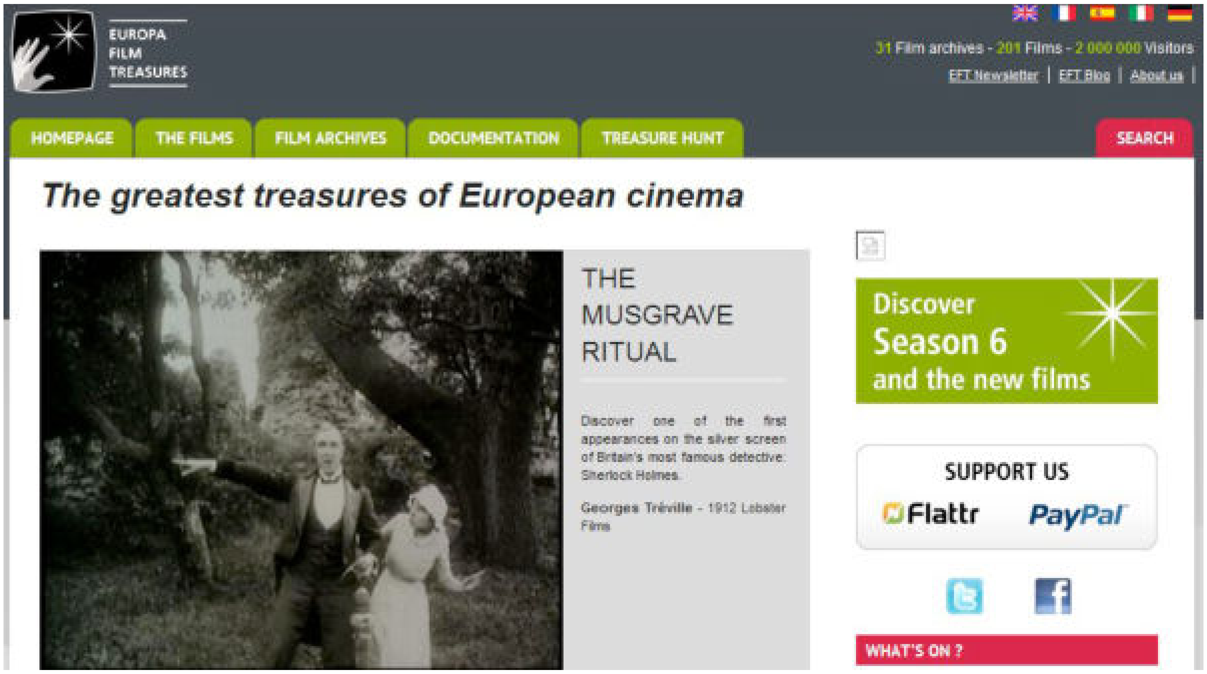
Despite the technical and financial challenges, they do not consider the site to be a failure. So what would it take to get Kwetu.net back online? The requirements to bring the site back would be primarily financial; that is, securing enough funding to keep the project sustainable. Whether a subscription model would still function in light of the Open Access movement in academia is not clear to these writers. While founders, Nyamu and Ouko, indicate that they still have access to the content and maintain relevant connections with their provider networks, at the moment they are focusing their attention on private sector clients.

In this case, a combination of technical, financial and human factors was involved. Referring to the list of common reasons for sites going dark presented above, at least three (neglect, technical and financial) apply here. One of the often overlooked aspects of Web preservation is the human time and energy it takes to keep Web sites alive, updated and functioning. All of the technical support fell to the original founders and immediate staff. Human oversight resulted in the domain name being lost, at which time, from the point of view of site visitors, it would have simply vanished. Finally, in cases such as this where the content is still in a state of preservation by its owners, but remains dormant due to a lack of resources, what can be done to restore it? We return to this question in our conclusion.

Europa Film Treasures

First launched in 2008 by Serge Bromberg, Europa Film Treasures (europafilmtreasures.eu) was "an online film museum", described by its founders as "an interactive tool for the promotion of film culture" [22]. Presented in English, French, Spanish, Italian and German, the Europa Film Treasures (EFT) homepage offered "free access to a scheduling of heritage films from the most prestigious European archives and film libraries". Online streaming of all full-length films was available without charge or geographic restrictions, making EFT an important repository and indeed a genuine public service:

> Faced with the vast choice offered on the Internet and the thousands of videos of uncertain quality and often-vague origins, we propose an entirely legal film offering on the Internet. Our principal commitment is to maintain this quality cultural offering, for it seems indispensable to us that all can access it without a tariff, geographic or linguistic barrier. [23]


EUROPA
FILM
TREASURES
31 Film archives - 201 Films - 2 000 000 Visitors
EFT Newsletter | EFT Blog | About us |
HOMEPAGE
THE FILMS
FILM ARCHIVES
DOCUMENTATION
TREASURE HUNT
SEARCH
Discover Season 6 and the new films
SUPPORT US
Flattr
PayPal
WHAT'S ON ?

The greatest treasures of European cinema

THE MUSGRAVE RITUAL

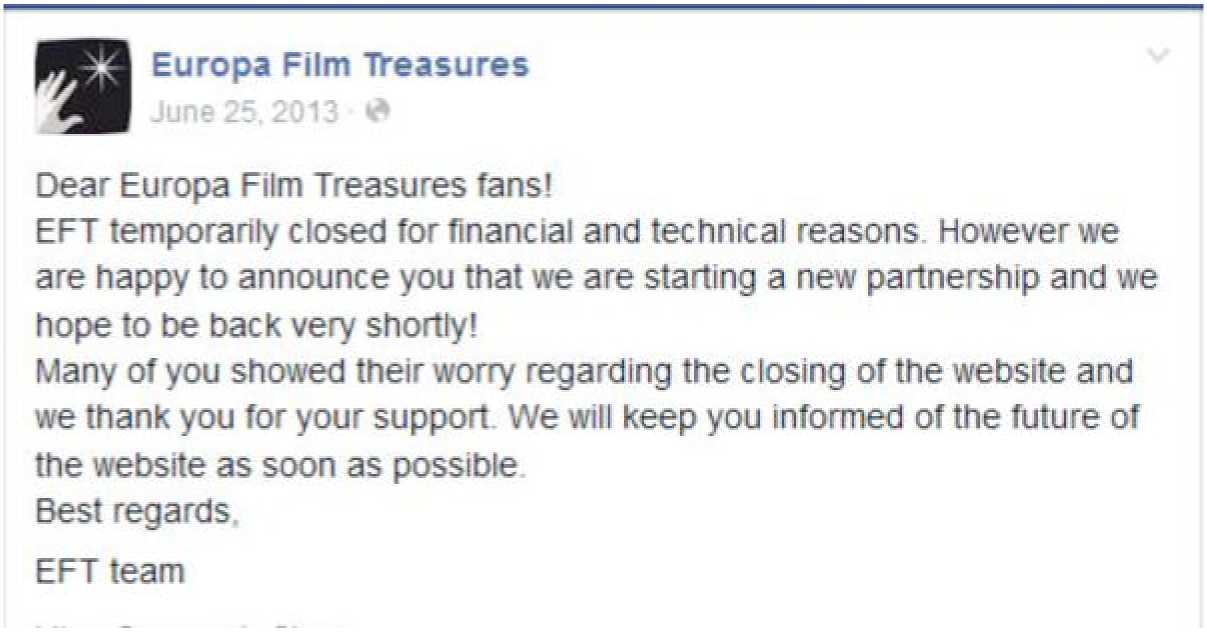
Discover one of the first appearances on the silver screen of Britain's most famous detective: Sherlock Holmes.

Georges Tréville - 1912 Lobster Films

Figure 2: Screenshot of the Europa Film Treasures Web site before it vanished (via Wayback Machine).

The Web site was made possible through partnership between Lobster Films, Sarl [24] — a film production and restoration company based in Paris — and 31 "prestigious" European archives including those of the British, Dutch, Danish, French, German, Irish, Italian, Finnish, Spanish and Swedish film institutes, among others. Internet service provider Enki Technologies handled the software, programs and file storage. It was also supported financially by the European Union's MEDIA program and other public and private partners. Copyright permissions to the original films for redistribution were secured from these partners, which effectively made EFT a heritage film aggregator that took on the hosting and maintenance responsibility for the films. According to the original "About" page, the collection contained 201 films dating from 1896 to 1999.

Another intention behind the site was education and instruction in the field of film preservation. All of the films, many of which were old, rare or from "relatively unknown film industries" were accompanied by an explanatory booklet of notes for better comprehension and a history of the film's "discovery and/or restoration". The original, full-length films together with additional film restoration resources, quizzes, puzzles, interviews and music composition notes comprised a valuable example of an interactive, public digital archive. When no music was available, Lobster would commission orchestral scores from music students (not more than one film per musician), and pay for them. Thus, the site as a whole became an important and much-loved resource with many valuable features.



Europa Film Treasures
June 25, 2013 ·

Dear Europa Film Treasures fans!
EFT temporarily closed for financial and technical reasons. However we are happy to announce you that we are starting a new partnership and we hope to be back very shortly!
Many of you showed their worry regarding the closing of the website and we thank you for your support. We will keep you informed of the future of the website as soon as possible.
Best regards,
EFT team

Figure 3: Announcement from the EFT Facebook page regarding the site's temporary closure in June 2013.

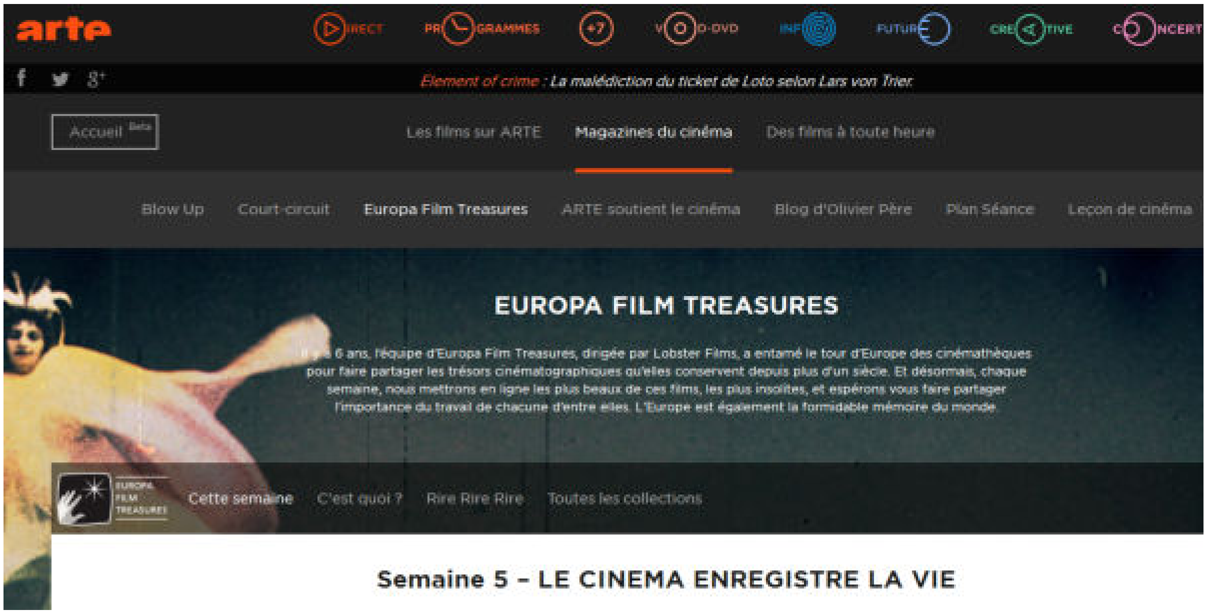
A message from 2013 on the official EFT Facebook informed users that the site has been temporarily closed for technical and financial reasons (see Figure 3) but no specifics were given. The post promised that a new partnership may result in the site re-opening very shortly. By September 2014, no new announcements had been made on the Facebook community or any other site regarding a re-launch despite the indication that users would be kept informed. In response to this post and in other places around the Web, many former users questioned what had happened to the site and expressed their dismay that they had lost access to the videos. The full details of the financial and technical reasons for a prolonged outage have remained mostly unexplained, leaving these former site visitors in the dark.

While the text and images making up the shell of the EFT Web site have been saved by the Internet Archive, the Wayback Machine has not saved copies of the actual films. When we contacted Lobster Films' CEO, Serge Bromberg, he provided the reason for the site's temporary disappearance:

> Enki Technologies went bankrupt, and before the last films went on line, we were told that the owner ... erased everything from his hard drives, and left without a trace. That was the end of the Web site as we knew it, but we of course still had the original masters for the films (on digi beta or Hdcam), the rights attached to them, and all the special contents created for the Web site. [25]

Similar to Kwetu.net and Haddon Online, the original files were luckily still in safe hands and awaiting an opportunity for restoration.

Thankfully, that time has arrived. Bromberg also informed us [26] that an as-yet unannounced new venture with Franco-German TV network ARTE will see the restoration of EFT films made available on a weekly basis via the ARTE Web site (see Figure 4). The timing of the launch of ARTE's cinema platform was serendipitous as it was an especially good fit for the EFT collection. In order to restore access, Lobster Films, backed by ARTE, covered the cost of reformatting the films for their new Web location. The Europa Film Treasures page at the ARTE Web site is currently live despite no official announcement being made at the time of writing.



Il y a 6 ans, l'équipe d'Europa Film Treasures, dirigée par Lobster Films, a entamé le tour d'Europe des cinémathèques pour faire partager les trésors cinématographiques qu'elles conservent depuis plus d'un siècle. Et désormais, chaque semaine, nous mettrons en ligne les plus beaux de ces films, les plus insolites, et espérons vous faire partager l'importance du travail de chacune d'entre elles. L'Europe est également la formidable mémoire du monde

Semaine 5 – LE CINEMA ENREGISTRE LA VIE

Figure 4: Europa Film Treasures' new home on the ARTE network as of November 2014 [27].

The return of EFT's film collection is still a work in progress. The films will be released weekly, so the full collection is not yet available. According to Bromberg, "When all the films are re-injected, we have already decided with ARTE to keep adding more films from the European Archive's vaults, with new explanatory texts attached." [28] However, the new site is only available in French and German, the languages of the Strasbourg-based ARTE network and, as yet, the guides, educational texts and interactive materials from the old site have not reappeared.

In its current form, EFT represents a successful case of bringing a once "dark" site back online, yet lessons can still be learned regarding the dangers of digital data loss. It took the sudden actions of just one person to wipe hard drives that would take down an entire Web site for nearly two years. Following that event, a great deal of dedicated effort, cost, negotiation and even luck went into the restoration process to bring EFT back online.

In addition, at the time of writing, the original URL (europafilmtreasure.eu) no longer resolves, so visitors to that URL or to the as-yet not updated Facebook page are none the wiser that the films are being re-released in a new location. Similarly, a search for "Europa Film Treasures" on Google (October 2014) does not yet bring up the new site. With some of the films online, but a continued lack of communication with the public both prior to and after the site was closed as well as in the lead up to its re-launch, the EFT case brings up some interesting gray zones that affect Web preservation. Even as former users made continued reports about the site going dark and that they desperately wanted restored access to the films, lack of transparency led many to assume that EFT was not coming back.

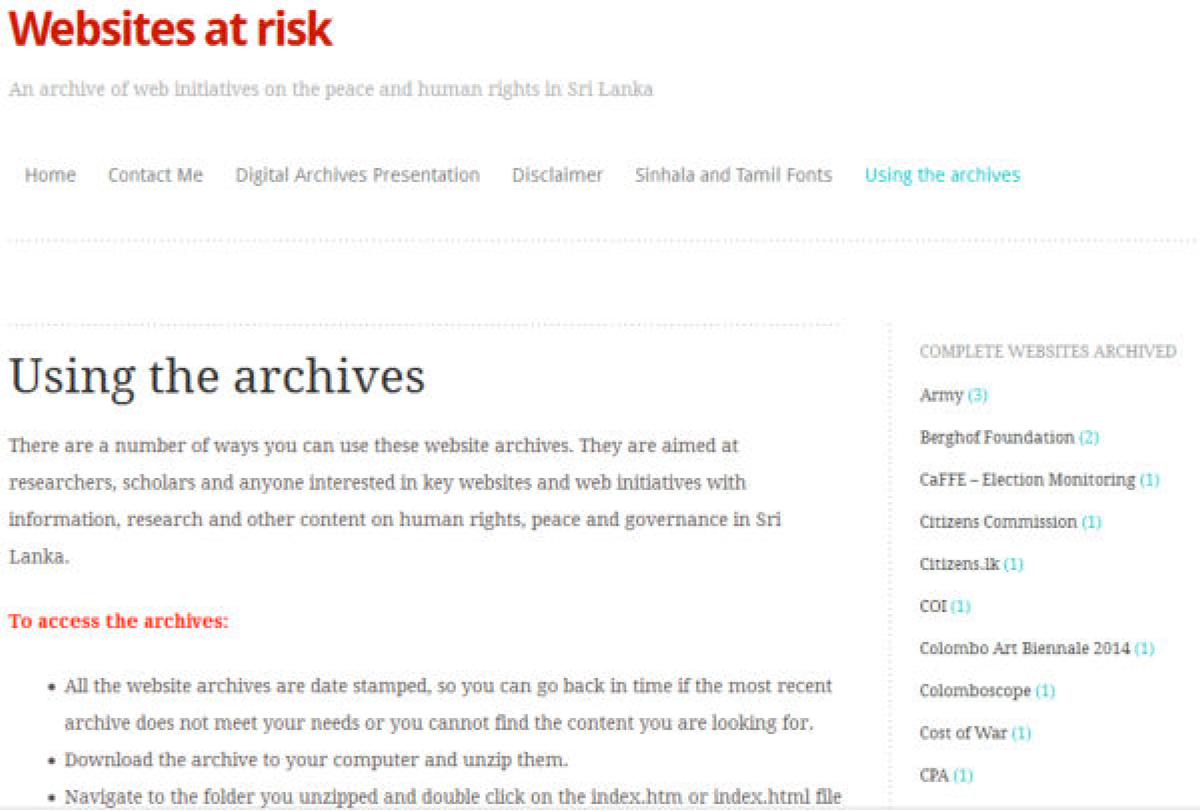
An anthropological case study approach proved effective in addressing both of the aforementioned cases. In each, there were complex organizational, financial and/or technical reasons that the content is no longer available to the public. Tracking down the relevant parties required prolonged investigation and multiple attempts at personal communication. While both Kwetu.net and EFT are the types of cases that Gone Dark researchers had expected to encounter in the course of the project, their circumstances would be difficult to pre-empt from a preservation standpoint. Each instance has localized peculiarities and complications. The benefit of both cases, however, is that the content is still in existence. The stories behind the loss of access and possible restoration can be used to evaluate what methods might be employed to restore sites like these in the future, or, preferably, to prevent such losses before they happen.

Sites at risk

Restoring a single site is a challenging enough task, but when a collection of related Web sites goes dark, prevention strategies are much more difficult to specify (and quantify) as losses can potentially include an entire digital ecosystem of information. A challenge for archivists is being able to tell the difference between isolated cases and more dispersed problems that may entirely wipe out a whole significant portion of the internet with serious social implications.

Hardware failure and technical neglect are not the only ways that online content can be lost. Most troubling in 2015 are Web resources that may be threatened by malicious parties (from hackers and militant groups to governments) who want to intentionally remove ‘conflicting’ information. The following case study focuses on conditions of political turmoil where external, and often non-digital, factors at play put the Web at risk every day. It shows that preventative backups are especially important when it is not always clear what information will become significant in the future.

Wherever there are volatile conditions on the ground, the Internet is susceptible to damage and loss. Human rights-related Web sites are therefore especially at-risk of going dark. In such cases, there are serious socio-political implications. When local news Web sites or cultural heritage organizations have their sites shut down during times of social upheaval, riots, or war, both daily communicative capabilities and the historical record can be irrevocably damaged. Unlike the two case studies above, the following study of at-risk sites in Sri Lanka from the perspective of a citizen archivist shows what can be learned from an expert who independently archives at-risk sites *before* they are lost forever.



Websites at risk

An archive of web initiatives on the peace and human rights in Sri Lanka

Home Contact Me Digital Archives Presentation Disclaimer Sinhala and Tamil Fonts Using the archives

Using the archives

There are a number of ways you can use these website archives. They are aimed at researchers, scholars and anyone interested in key websites and web initiatives with information, research and other content on human rights, peace and governance in Sri Lanka.

To access the archives:

- All the website archives are date stamped, so you can go back in time if the most recent archive does not meet your needs or you cannot find the content you are looking for.
- Download the archive to your computer and unzip them.
- Navigate to the folder you unzipped and double click on the index.htm or index.html file

COMPLETE WEBSITES ARCHIVED

Army (3)
Berghof Foundation (2)
CaFFE – Election Monitoring (1)
Citizens Commission (1)
Citizens.lk (1)
COI (1)
Colombo Art Biennale 2014 (1)
Colomboscope (1)
Cost of War (1)
CPA (1)

Figure 5: Screenshot of Sanjana Hattotuwa’s Sri Lankan archive Web site, Sites At Risk.

Sanjana Hattotuwa, human rights activist and creator/curator of *Groundviews* [29], Sri Lanka’s first citizen journalism Web site, is at the forefront of endangered Web site preservation in Sri Lanka. Hattotuwa’s personal blog, Sites at Risk Sri Lanka (see Figure 5), was created as an “archive of Web initiatives on peace and human rights in Sri Lanka” [30]. Inspired by the Internet Archive’s Wayback Machine, which Hattotuwa found does not adequately archive Sri Lankan civil society content “with any useful degree of comprehensiveness or frequency”, the purpose of this site is to keep downloadable .zip copies of entire “civil society and NGO Web sites and Web based initiatives on human rights, democratic governance and peacebuilding” for when they “suddenly go off–line or are rendered inaccessible in Sri Lanka” in order to preserve the content for scholars of peace and conflict [31].

Hattotuwa’s curated archive reveals the ease at which, one site at a time, an entire ecosystem of Web content can remain at continued risk due to conflict. He reflects on why it is important not to let this happen:

> The loss of digital resources for human rights activists is a significant one. The danger is two fold — one is of an enforced erasure and deletion of vital records, the other is deletion and erasure out negligence and technical failure. In both cases the failure to adequately and

> strategically adopt safeguards to backup information can exacerbate information loss. The issue with [human rights] documentation is that it is often irreplaceable — once lost, digitally, the same records cannot be regenerated from the field. Sometimes it is possible to go back to physical records, but most often the digital record is all that's there. [...] Digital information loss in this context can, as I have argued in the past, lead to the exacerbation of conflict. [32]

His expert knowledge of the political situation and key players in Sri Lankan human rights arena enable Hattotuwa to make pre-emptive and decisive steps towards archiving potentially vulnerable content with a higher success rate than relying on automated crawls. At the first hint of vulnerability, he saves a copy of the site in question before it can be lost. The content he is saving has a personal relevance and connection for him and he is well aware of the value of the archives of the information that he keeps.

Hattotuwa's memory of some of the greatest Web site losses he has witnessed reflects this:

> The most significant loss around Web site based data in Sri Lanka I have encountered [were] on two occasions. One [was] the Mayoral Campaign of a candidate I in fact stood publicly and vehemently opposed to. His campaign team created a Web site around their vision for the development of Colombo, engendering comments for each point in their manifesto in Sinhala, Tamil and English — the languages spoken in Sri Lanka. That Web site, soon after he lost, was taken down and yet was a treasure trove of ideas around governance and urban rejuvenation. The other site loss, arguably even more tragically, was the erstwhile site of the Lessons Learnt and Reconciliation Commission (LLRC), a process that looked into stories from citizens around the end of the war. There were citizen testimonies, records of the public hearings and associated documentation on the [government's] site that was for whatever reason just allowed to expire. ... My own archive of the LLRC plus another I helped set up are now the country's only archives of this content. [33]

In terms of technical solutions, the full archive for each saved Web site is stored as a .zip file. This allows them to be opened regardless of the user's operating system. The files themselves are hosted on a public Wordpress blog using Box.net storage. Anyone can download the copies and store them locally. The .zip files are self-contained copies of the entire Web site [34], with all pages, text and images, so that once downloaded, the site can be browsed off–line as if it were a live copy, even without the need for an Internet connection. For scholars, this type of repository system has advantages over simple screen grabs or surface crawls stored on a Web server. The format means that the files in the archive can be full-text searched quickly and efficiently using desktop search software.

The site itself is functional and can instruct and inspire others to create similar collections of important Web sites that are at risk of disappearing. For instance, he chose the site name to be "scaleable": "the idea was that each country or region would use sitesatrisk and at the end plug in their name — *e.g.*, sitesatriskuk, sitesatriskkosovo" (Hattotuwa, 2008). As yet, he is unaware of any other sites replicating his model, but understands why this is unsurprising: "I am constantly responding to some emergency or the other, constantly myself the subject of hate, hurt and harm. It's not easy, and so I understand why others haven't taken this up." Naturally, Sites at Risk Sri Lanka and all the files that appear there rely on Hattotuwa's personal dedication and upkeep. In such constantly perilous conditions, this is a difficult task to assume.

In the case of Sri Lankan human rights Web sites, Hattotuwa reveals that "a litany of issues" that are responsible for site loss, from "an incumbent regime viciously intolerant of critical perspectives on war and peace to a disturbing lack of awareness of, emphasis on and interest in safeguarding information and knowledge" by NGOs and civil society actors in Sri Lanka. What is worrying is that "most never learn, even when disaster strikes once" [35]. Thus, one solution to sites going dark is to improve general awareness of the fragility of online content.

Discussion

Who will save the Web?

Given the difficulty of tracing sites that have gone dark once they are off–line, we find that greater engagement with subject experts will be at the forefront of better Web preservation tools and practices. Since deep Web content like media or file repositories and research databases are absent from standard Web crawls (see below), *selective archiving* is best undertaken by those with firsthand knowledge of essential sites — such as career specialists, journalists, historians, hobbyists, activists, academics and private individuals. As experts typically engage in maintaining their own records of files and research repositories, they will be among the first to notice when a site goes dark and also, like Sanjana

Hattotuwa, able to prevent imminent losses.

As we note below, there is a problem that it is not clearly any one person or organization’s responsibility to ‘archive the Internet’ (the Internet Archive’s self-appointed, and limited role as discussed above, notwithstanding). Outside of dedicated university or national library archive programs [36], academics in particular may find themselves becoming inadvertent archivists, unaware that the copies of content they produce in their day-to-day work may be the only remaining copies of important Web archival materials. Certainly, many digital researchers do not begin their projects intending to keep permanent archives to make publicly available. Similarly, foresight and intuition for Web preservation is not always coupled with institutional or financial stability.

A good example of selective archiving by subject experts is the Internet Archive’s Archive-It program:

> Archive-It is a subscription Web archiving service from the Internet Archive that helps organizations to harvest, build, and preserve collections of digital content. Through our user-friendly Web application Archive-It partners can collect, catalog, and manage their collections of archived content with 24/7 access and full text search available for their use as well as their patrons. [37]

Current subscribers include college libraries, state archives, historical societies, NGOs, museums, public libraries, cities and counties. The success of such archives to maintain important timely content became evident when a curator from the Hoover Institution Library and Archives [38] had the foresight to include blog posts in Archive-It’s Ukraine Conflict Collection [39] in July 2014 (Hoover Institution, 2014). One such blog post became a piece of contentious evidence potentially tying separatist rebels in the Donetsk People's Republic in Eastern Ukraine to the Malaysian Airlines Flight 17 crash (Dewey, 2014) that killed 298 passengers and crew. When the live post was deleted, the evidence remained for international scrutiny in the Ukraine Conflict Collection, preserved because a proactive archivist recognized its importance before it was too late.

As Nicholas Taylor, Web Archiving Service Manager for Stanford University Libraries, explains:

> Internet Archive crawls the Web every few months, tends to seed those crawls from online directories or compiled lists of top Web sites that favor popular content, archives more broadly across Web sites than it does deeply on any given Web site, and embargoes archived content from public access for at least six months. These parameters make the Internet Archive Wayback Machine an incredible resource for the broadest possible swath of Web history in one place, but they don’t dispose it toward ensuring the archiving and immediate re-presentation of a blog post with a three-hour lifespan on a blog that was largely unknown until recently. [...] Though the key blog post was ultimately recorded through the Save Page Now feature, what’s clear is that subject area experts play a vital role in focusing Web archiving efforts and, in this case, facilitated the preservation of a vital document that would not otherwise have been archived. (Taylor, 2014)

Another example from Archive-It is the Occupy Movement Collection [40], which was started in December 2011 to capture ephemeral Web content to record the then rapidly developing global Occupy movement. In April 2014, researchers decided to look back at the 933 seed URLs amassed since 2011 to see how many of the pages were still live (Archive-It, 2014). They found that while 90 percent of archived news articles and 85 percent of social media content was still live on the Web, this number dropped to 41 percent for the 582 Web sites in the collection. Fifty-nine percent of all Web sites were no longer live and either returned 404 error messages or had been taken over by cybersquatters (Archive-It, 2014). This useful analysis shows that even with selective archiving, the work of Web preservation is ongoing. Sites are still going dark.

It would be interesting to see similar statistics for live or defunct Web sites for other Archive-It collections. Automating this process might be difficult, however, as in this case, “using a human to check the URL, rather an automated process, allowed for closer analysis of the live content to determine if it was on topic” (Archive-It, 2014). Interactions with archivists at various national libraries and organizations have shown throughout this project that there is a great deal of human intervention at work throughout the entire archiving and preservation process although it might look totally automated to outside observers. Most archives accept submissions and are grateful for notifications from the public about sites that need their attention.

In the following section, we address existing archiving practices and public perceptions before offering recommendations from this research that we hope will better facilitate more comprehensive solutions to sites going dark.

Current practices and perceptions

One popular view is that nothing on the Internet is ever truly "deleted"; that is, anything we put on the Web will be around forever to haunt us (Rosen, 2010; Whittaker, 2010) because in the digital age the default has shifted from "forgetting" to preservation (Mayer-Schönberger, 2009). This outlook makes it difficult to communicate to the general public the risk of content on the Web being lost to the world. Many Internet users have become familiar with the Wayback Machine, whose mission is to copy virtually every page on the public Web. This fantastic project has both alerted people to early Web sites with nostalgic value that might have been lost as well as become an essential tool for Web site creators or bloggers to find backups of their own pages that they might have accidentally deleted. As seen above, the Archive-It service and "Save Page Now" feature also enable others to submit links to supplement the Archive. As a result, it gives the appearance that every page on the Internet is being safeguarded and therefore needs no further intervention. Indeed, in the course of interviews for this project, we encountered cases where the former owner-operators of defunct URLs simply direct previous users of their site to the archived version of it in the Wayback Machine rather than attempt to save or restore the content otherwise.

Yet some fundamental misconceptions about the Wayback Machine are employed in this reasoning. For instance, evidence from this study suggest that — in the case of sites housing significant content of cultural or social value — it is not enough to simply leave a site to be captured by the Internet Archive with no other provision for saving its content for long-term preservation. While the Internet Archive's task to save a snapshot of the Web is useful, it is insufficient for sites that go beyond simple text and HTML.

Web crawlers like the Wayback Machine take a snapshot of surface content only. However, pages like EFT or Kwetu.net "may serve as the front end to a database, image repository, or a library management system, and Web crawlers capture none of the material contained in these so-called 'deep' Web resources" (Kenney, *et al.*, 2002). Contrary to popular belief, the searchable Web only represents a fraction of the pages on the Internet (Wright, 2009), omitting many types of sites and file repositories. Even where a standard crawl produces a sufficient facsimile of a largely text-based site to act as an archival record, it does not mean that it will be around forever.

Because the Internet Archive respects changes to the robots.txt file [41], site owners can decide to allow or disallow their sites to be crawled. Even more significant is that changes to the robots.txt file are retroactive. This means that changes made at any time will affect whether the historical record held by the Wayback Machine is erased. The official statement on this policy includes the following key section:

> While robots.txt has been adopted as the universal standard for robot exclusion, compliance with robots.txt is strictly voluntary. In fact most Web sites do not have a robots.txt file, and many Web crawlers are not programmed to obey the instructions anyway. However, Alexa Internet, the company that crawls the Web for the Internet Archive, does respect robots.txt instructions, and even does so retroactively. If a Web site owner decides he/she prefers not to have a Web crawler visiting his/her files and sets up robots.txt on the site, the Alexa crawlers will stop visiting those files and will make unavailable all files previously gathered from that site. [42]

As has been noted in several cases in discussions on archive.org [43] the retrospective application of changes to the robots.txt file threatens data permanence when it comes to domain squatting or simple changes of domain ownership and use. For example, recall that the founders of Kwetu.net lost their domain name by failing to renew it in time. If the current owner of Kwetu.net decided to make a small change to the current robots.txt file, all traces of Kwetu.net could vanish from the Wayback Machine forever. We address this later in the conclusions.

Crawlers are similarly bound by legal constraints, such as in 2002, when the Wayback Machine removed entire domains from its archive to comply with a request from the Church of Scientology's lawyers (Miller, 2002). There is also a six-month embargo on new sites, thus letting much timely and fleeting content — of the kind at risk in Sri Lanka — slip away before it can be crawled unless it is selectively archived by a human.

Furthermore, and extremely relevant to the Gone Dark Project, it can be argued that "automated approaches to collecting Web data tend to stop short of incorporating the means to manage the risks of content loss to valuable Web documents" (Kenney, *et al.*, 2002). That is, it does not address the root causes. As our case studies show, just having a record of sites that used to be live is not a sufficient preservation strategy, although it is definitely an indispensable service to maintain.

Problems arise when existing copies of pages come to be seen as a permanent backup solution and when insufficient attention is paid to the content or pages that are allowed to disappear. Relying on archives run by national libraries around the world to do all the work — especially those collections that are not made public (BBC News, 2013) — can cause content creators as well as average Web users to become complacent and therefore not take proactive steps to save large amounts of valuable content in a more functional format. As a result, automated crawls might inadvertently diminish long-term health of Web resources by encouraging a passive approach to backups coupled with the misguided impression that nothing on the Internet can be lost forever, when in fact it happens all the time.

Conclusions

National libraries and digital archive organizations continue to draw attention to the dangers of disappearing Web content. They are setting standards and taking action to save entire Web sites or ephemeral social media content from being lost forever. Despite the many laudable successes of current digital preservation efforts, however, some weak spots remain as we have demonstrated. For instance, automated Web preservation is restricted to the indexable or "surface" Web. We are limited by an inability to foresee and therefore prevent content loss that falls outside of this. Many sites, including those in the case studies above, may hold large repositories of culturally significant information behind a pay wall, a registration system or in a database. Similarly, large collections of Web sites dispersed throughout an entire ecosystem — such as human rights and activist Web sites — are fragile especially because of the difficulty in tracking ownership or preserving whole sections of the Web that may become vulnerable all at once.

What vanishing content reveals is that there is a problem in self-organisation for the network of bodies that keep the Internet running [44]. There is a lack of clarity about who is responsible for archiving material so in some cases it falls between cracks and vanishes. Again we recognise that this is not a new problem: it is at least as old as newspapers (no one was responsible for archiving a failing media business such as a local newspaper). So we may need more discussion of responsibility not only from Web archivists but also from content holders. Holders must be willing to have their content archived and made public (under certain conditions). And of course we need discussion of how all this is to be paid for. It is clear that archiving services are not without costs, especially as we think into the long term.

Sites that go dark do so for a variety of reasons from the financial to simple neglect and even malicious removal, but, as we have shown, that does not always mean that the original content has vanished. That we have been able, in the course of this project, to connect with relevant parties who report intentions to revive old pages given the right conditions means that better safeguards in place may be able to prevent such losses or hasten their return online. The following recommendations therefore consider what can be done to avoid future losses as well as suggesting ways to better preserve sites in imminent danger of vanishing.

A first step is to draw upon the excellent work already being done by Internet archivists to enhance the ease and regularity at which sites that fall within these grey areas can be saved. This means bridging the gap between professional librarians, academics, archivists and other dedicated individuals who can make worthwhile contributions. Paid, subscription services for Web site backup may be a good enterprise solution for the site owners with the means and access to make use of them (such as academic or corporate institutions), but the costs may be prohibitive to other users or they may simply lack of awareness of their existence. Encouraging dialogue between archive service providers and subject area experts will be the most effective way to save endangered or at-risk sites from going dark.

The background research we undertook revealed that many putative cases of disappearance were rather examples of link rot: the material was there but no longer at the same URL. This obscures a smaller number of actual instances of disappearance. Thus, arising from the case studies and interviews we have undertaken in this project, the broadest recommendation is to allow for more human intervention in the archival process such as appealing to subject experts who have first-hand knowledge of parts of the Web that are now at risk or may be in the future. Sanjana Hattotuwa exemplifies how specialist experience can inform better archiving practices based on actual needs and practices, while the Europa Film Treasures and Kwetu case studies show the importance of foresight and instilling good data management for long-term survival of Web content.

In addition, we have also encountered "inadvertent archivists": these are mainly researchers or academics who have found that they have unintentionally become the curators of the only surviving copy of old Web content that they captured in the course of their research. Among these are some original Web site owners who may have old pages stored on their hard drives, but no means to restore them to the Web.

What can we do to help those who find themselves with knowledge, or in possession, of Web content, but who do not know what to do with it? Because of the complex reality of sites going dark, we find that combinations of human and technical solutions are necessary.

There are practical considerations to remedy vanishing Web sites which vary on a case-by-case basis. Depending on the type of site and/or repository of media or information in question, making the data public — or indeed accessing the content without the aid of insiders — can be difficult. For this reason, collaborative solutions are needed which bring together those content owners or researchers aware of imminent site losses with archival professionals who can assist them. Ideally, this would include tailored services to better enable individuals in more perilous circumstances without the luxuries of institutional backing or secure funding sources to safeguard essential sites easily.

It is therefore important to continue developing tools to improve as well as open the archiving process to a wider audience. This will help to counteract the public perception that the entire Web is being backed up automatically when so much of it can remain at risk. Currently, the Internet Archive's Wayback Machine allows users to submit pages for archiving using the Save Page Now function, as do several other on-demand services. Yet none of these solutions reach out to the original site owner, make provisions for long-term preservation of original data, or endeavor to keep the site live. The backups they provide are also largely ineffective for recreating the site at a later date if the essential content is missing.

One technical solution to help bridge the gap may be an escrow-type backup system for the protection of endangered content. Such a solution would require archive professionals working closely with content owners or subject experts to produce preservation strategies that are easy to adopt, secure and flexible. The type of archive or backup system, its format and accessibility (open vs. restricted access) may vary depending on the needs of the organization or individuals wishing to secure their data and how sensitive that content may be. For instance, file format and integrity is a primary concern alongside legal requirements for preserving metadata to enable digital files to be used as a court record [45].

Working with experts would be of great value to help identify the type of sites we encountered in this project. At the same time, an interesting corollary is the need for improvements in the automated, technical side of Web preservation. Sanjana Hattotuwa adds this caveat: "machine and algorithmic curation can, with enough learning provided by analysing human curation, aid [the] archiving of content at risk esp. during violent conflict." [46] This can be invaluable in cases where the resources are simply not available to maintain fully staffed digital archives, such as in high-risk situations, with many non-profit companies, small organizations, poorer nations or NGOs. The same thing applies to small-scale sites whose owners are not available or otherwise up-to-date with good archiving practices.

Also worthy of consideration are open archive solutions to harness and analyze these aspects of human curation. Combining both technical and collaborative endeavors could result in a crowd-sourced solution that not only enabled users to submit sites at risk or already gone, but also then used submission data to predict other Web site candidates that may also be vulnerable. In the course of this project, we had expressions of interest from non-experts who wished to contribute more to efforts to save disappearing sites, but were unaware of any channels available to do so. Often, they could only offer an old URL for further investigation, which is where we were able to step in.

Lastly, developing solutions for safeguarding at-risk sites or reviving sites that have already gone dark requires improvements in how archives (and researchers) keep track of the disappearing Web over time. Inadvertently, this project has demonstrated the difficulties in identifying sites that may need help. It is certainly labor-intensive. One idea to remedy this lack of wider awareness about site losses is an early warning system for those parts of the Web that fall outside the scope of existing archival practices. An "endangered Web site alarm" could alert potential archivists of imminent content losses before or as they happen. For truly effective, proactive archiving solutions, this would go hand in hand with having clearer communication channels in place between archive service providers and others.

For example, while the Wayback Machine is an indispensable tool for Web research, as described above, several of its key restrictions limit its utility at present for pre-empting digital losses of sites that are not easily crawled by Alexa. That said, the Internet Archive expressed willingness to allow access to its collections by "researchers, historians and scholars" [47], and in 2012, even experimentally offered researchers access to a full 80 terabytes of archived Web crawl data by request (Rossi, 2012). We believe that the data that the Internet Archive, Alexa, Google, other search engines and even Wikipedia collect may offer valuable insight into the evolution of the Web if researchers had access to certain information.

Rather than search manually for broken links to find URLs returning 404 errors as was done in the course of the Gone Dark Project, it would be much more useful if there were a system to export data from automated crawls that indicated persistent 404 errors within a given period of time to give researchers a chance to investigate further before the data is completely lost from the public Web. Similarly, logs of changes to robots.txt files (as noted above, these changes are retroactive and permanent and can wipe archive records) could alert researchers or Web preservationists of unforeseen losses as they happen. It might be that a change of robots.txt file which would trigger retrospective deletion could only go back as far as the current ownership of a domain. This is automatable so we recommend it to the Internet Archive. Another possible way of using 404 errors to promote archiving might be if they could delay the wiping of cached copies by Internet search services such as Google and Bing.

In addition, the Wayback Machine once had a functional search engine called Recall, designed by Anna Patterson [48]. Looking back on our research, it was difficult to locate important sites that have gone dark because it is nearly impossible to search historical Web content. Live search engines like Google cannot search defunct pages, while sites cannot be retrieved from most internet archives without the original URL. Enabling full-text searching of old pages would be ideal.

In all three case studies, a key lesson learned has been that a priority for improving Web preservation needs to begin at source, educating site owners and content producers so they understand the value of Web archiving. This is perhaps most key for high-risk sites or large repositories. But the education process needs to go both ways: the best practices for archiving are those which meet the current and future needs of those whose content would benefit from long-term storage and also those who will be able to make use of the content in future, whether to restore it to the public Web or to safeguard in a restricted archive.

Solutions to the problems of sites going dark will require more awareness from all parties involved. Making archiving initiatives more accessible, collaborative and lowering boundaries to participation (at present, interested parties must have “reasonably advanced programming skills” [49] to work with the Internet Archive’s data crawls, which is prohibitive for many) is a good start. Beyond simply collecting snapshots from old URLs, the long-term health of essential Web resources will depend on working with content owners to find permanent homes for at-risk data.

Recommendations summary

1. major service providers should consider maintaining backups as dark archives/escrow services
2. Internet archive services should provide a mechanism for “inadvertent archivists” to upload material (possibly not their own)
3. Internet archive services should provide a mechanism for experts to flag material as being at risk for urgent archiving [50]
4. Patterns in 404 errors should be investigated — can they predict data loss?
5. Google and Bing (etc.) consider responding to persistent 404 errors by passing cached copies to archive services.
6. Internet Archive respect changes of robots.txt file which would trigger retrospective deletion only as far as the current ownership of a domain.

About the authors

Francine Barone is a social anthropologist and Internet researcher. Her ethnographic research focuses on the socio-cultural impacts of the digital age.
E-mail: fbarone [at] gmail [dot] com

David Zeitlyn is professor of social anthropology at the Institute of Social and Cultural Anthropology, University of Oxford. His field research is concentrated in Cameroon and he also works on archives and has been a pioneer of using the Internet to disseminate anthropology.
E-mail: david [dot] zeitlyn [at] anthro [dot] ox [dot] ac [dot] uk

Viktor Mayer-Schönberger is Professor of Internet Governance and Regulation at the Oxford Internet Institute, University of Oxford.
E-mail: viktor [dot] ms [at] oii [dot] ox [dot] ac [dot] uk

Notes

[1] http://www.archive.org.

[2] http://archiveteam.org/index.php?title=Deathwatch.

[3] http://www.webarchive.org.uk/.

[4] https://twitter.com/AT_Webarchive.

[5] http://www.archive.org.

[6] http://www.digitalpreservation.gov.

[7] http://www.internetmemory.org.

[8] Originally online at http://www.rsl.ox.ac.uk/isca/haddon/HADD_home.html.

[9] See http://web.archive.org/web/20050415000000*/http://www.isca.ox.ac.uk/haddon/HADD_home.html. The 4 April 2005 is last working snapshot before they become 404 not found.

[10] http://en.wikipedia.org/wiki/List_of_Google_products#Discontinued_products_and_services.

[11] http://en.wikipedia.org/wiki/List_of_Yahoo!-owned_sites_and_services#Closed.2Fdefunct_services.

[12] http://en.wikipedia.org/wiki/Windows_Live#Discontinued_services.

[13] There are many others. Some of the most prominent are mentioned at https://en.wikipedia.org/wiki/Wayback_Machine.

[14] Amad, 2010, p. 1; see also Zeitlyn (2012) and forthcoming.

[15] The research for this section was undertaken by Nanjira Sambuli, iHub Research, Kenya.

[16] See Stanford University's Library and Academic Information (Kenya) Resources listing: http://www-sul.stanford.edu/depts/ssrg/africa/kenya.html.

[17] According to the company profile page accessible through the Wayback Machine: http://web.archive.org/web/20030212235255/http://kwetu.net/about.asp.

[18] Source: http://www.library.upenn.edu/news/86.

[19] Available via Wayback Machine: http://web.archive.org/web/20060114003227/http://www.kwetu.net/partners.asp.

[20] http://web.archive.org/web/20060114024930/http://www.kwetu.net/subscribers.asp.

[21] A cached copy of the front-end of the Kwetu.net search engine from 2003 is available from: http://web.archive.org/web/20030812143045/http://kwetu.net/search.asp.

[22] http://www.openculture.com/2012/12/europa_film_treasures.html.

[23] According to the site's original "About" page: http://web.archive.org/web/20130327054908/http://www.europafilmtreasures.eu/about_us.htm.

[24] http://www.lobsterfilms.com/ANG/index.php.

[25] Personal communication, 21 January 2015.

[26] Personal communication, 18 September 2014.

[27] http://cinema.arte.tv/fr/magazine/europa-film-treasures.

[28] Personal communication, 21 January 2015.

[29] http://groundviews.org.

[30] http://sitesatrisksl.wordpress.com.

[31] *Ibid*.

[32] Personal communication, 3 May 2014.

[33] Personal communication, 3 May 2014.

[34] We note that this approach would not work for Web sites which access content via a database such as kwetu.net already discussed above.

[35] http://sitesatrisksl.wordpress.com/.

[36] The Human Rights Documentation Initiative at the University of Texas and Columbia University's Human Rights Web Archive, are both doing essential work for human rights Web preservation.

[37] https://www.archive-it.org/learn-more.

[38] http://hoover.org.

[39] https://archive-it.org/collections/4399.

[40] https://archive-it.org/collections/2950.

[41] A file that contains requests from site owners that can prevent Web crawling software from crawling certain pages. See: https://support.google.com/webmasters/answer/6062608?hl=en. Note that not all Web crawlers respect robots.txt files. The Internet Archive does.

[42] http://archive.org/about/faqs.php#14, (http://perma.cc/528A-QMPH, accessed 21 March 2015).

[43] See https://archive.org/post/406632/why-does-the-wayback-machine-pay-attention-to-robotstxt (http://perma.cc/NL3M-MNK9) and https://archive.org/post/188806/retroactive-robotstxt-and-domain-squatters (http://perma.cc/P6HL-VRWF).

[44] We are very grateful to *First Monday's* reviewers for suggesting that we acknowledge this point

explicitly — and for other points made in the review.

45. Sanjana Hattotuwa, personal communication, 11 November 2014.

46. Personal communication, 11 November 2014.

47. http://web.archive.org/web/20090924112618/ and http://www.archive.org/web/researcher/intended_users.php.

48. http://web.archive.org/web/20031204221423/ia00406.archive.org/about.html.

49. Explained here: http://web.archive.org/web/20090924112618/http://www.archive.org/web/researcher/intended_users.php.

50. Manual archiving is possible using services such as "Save Page Now" and "Archive-It". However, these share the same problem as the crawler-based automatic services of not having access to the content of Web-searchable databases.

References

Note: We have created perma.cc archive copies for our online sources. For completeness, we give both URLs although the permac.cc URL passes through to the original URL if it is still available, serving the archived copy only if the original URL generates a 404 error.

P. Amad, 2010. *Counter-archive: Film, the everyday, and Albert Kahn's Archives de la Planète*. New York: Columbia University Press.

Archive-It, 2014. "Only 41% of Occupy Movement URLs accessible on live Web," *Archive-It Blog*, ar https://archive-it.org/blog/only-41-of-occupy-movement-urls-accessible-on-live-web, accessed 13 November 2014; http://perma.cc/RJF6-CRXL, accessed 24 April 2015.

J. Barksdale and F. Berman, 2007. "Saving our digital heritage," *Washington Post* (16 May), at http://www.washingtonpost.com/wp-dyn/content/article/2007/05/15/AR2007051501873.html, accessed 14 February 2014; http://perma.cc/FNX5-Y97X, accessed 24 April 2015.

BBC News, 2013. "Web archive goes live but not online," *BBC News* (19 December), at http://www.bbc.com/news/technology-25446913, accessed 13 November 2014; http://perma.cc/5FGP-BCHQ, accessed 24 April 2015.

BBC News, 2010. "British Library warns UK's Web heritage 'could be lost'," " *BBC News* (25 February), at http://news.bbc.co.uk/2/hi/technology/8535384.stm, accessed 11 November 2014; http://perma.cc/U538-5F5M, accessed 24 April 2015.

O. Bhuiyan, 2014. "Technorati — the world's largest blog directory — is gone. *Business 2 Community* (16 June), at http://www.business2community.com/social-media/technorati-worlds-largest-blog-directory-gone-0915716, accessed 11 November 2014; http://perma.cc/4NZC-GQM4, accessed 24 April 2015.

C. Dewey, 2014. "How Web archivists and other digital sleuths are unraveling the mystery of MH17," *Washington Post* (21 July) (available on-line: http://www.washingtonpost.com/news/the-intersect/wp/2014/07/21/how-web-archivists-and-other-digital-sleuths-are-unraveling-the-mystery-of-mh17/, accessed 13 November 2014; http://perma.cc/7AMQ-K8ZP, accessed 24 April 2015.

Economist, 2000. "The hopeless continent," *Economist* (13 May), at http://www.economist.com/node/333429, accessed 12 November 2014; http://perma.cc/E6L8-Q3UT, accessed 24 April 2015.

S. Hattotuwa, 2008. "Websites at risk — Archiving information on human rights, governance and peace," *ICT for Peacebuilding (ICT4Peace)*, at http://ict4peace.wordpress.com/2008/04/02/websites-at-risk-archiving-information-on-human-rights-governance-and-peace/, accessed 13 November 2014; http://perma.cc/6EFV-X25D, accessed 24 April 2015.

Hoover Institution, 2014. "Archivists capture evidence in Malaysia Airlines Flight 17 crash" (25 July), at http://www.hoover.org/news/archivists-capture-evidence-malaysia-airlines-flight-17-crash, accessed 13 November 2014; http://perma.cc/54D2-RR6B, accessed 24 April 2015.

Internet Archive, 2014. "Wayback Machine hits 400,000,000,000!" *Internet Archive Blogs* (9 May), at http://blog.archive.org/2014/05/09/wayback-machine-hits-400000000000/, accessed 11 November 2014; http://perma.cc/RW5Y-2PSQ, accessed 24 April 2015.

A. Kenney, N. McGovern, P. Botticelli,, R. Entlich, C. Lagoze and S. Payette, 2002. "Preservation risk management for Web resources: Virtual remote control in Cornell's Project Prism," *D-Lib Magazine*, volume 8, number 1, at http://www.dlib.org/dlib/january02/kenney/01kenney.html, accessed 13 November 2014; http://perma.cc/BLQ8-D4Z2, accessed 24 April 2015.

S. Lohr, 2010. "Library of Congress will save tweets," *New York Times* (14 April), at http://www.nytimes.com/2010/04/15/technology/15twitter.html, accessed 11 November 2014; http://perma.cc/QA6Y-GU69, accessed 24 April 2015.

V. Mayer-Schönberger, 2009. *Delete: The virtue of forgetting in the digital age*. Princeton, N.J.: Princeton University Press.

E. Miller, 2014. "Sherman, set the Wayback Machine for scientology," *LawMeme* (24 September), at http://web.archive.org/web/20141025203224/http://lawmeme.research.yale.edu/modules.php?name=News&file=article&sid=350, accessed 13 November 2014; http://perma.cc/2JRV-5CX7, accessed 24 April 2015.

D. Quenqua, 2009. "Blogs falling in an empty forest," *New York Times* (5 June), at http://www.nytimes.com/2009/06/07/fashion/07blogs.html, accessed 11 November 2014; http://perma.cc/Z5F6-FRGJ, accessed 24 April 2015.

J. Rosen, 2010. "The Web means the end of forgetting," *New York Times* (21 July), at http://www.nytimes.com/2010/07/25/magazine/25privacy-t2.html, accessed 13 November 2014.

A. Rossi, 2012. "80 terabytes of archived Web crawl data available for research," *Internet Archive Blogs* (26 October), at http://blog.archive.org/2012/10/26/80-terabytes-of-archived-web-crawl-data-available-for-research/, accessed 22 December 2014; http://perma.cc/UPF4-3MQ4, accessed 24 April 2015.

H. SalahEldeen and M. Nelson, 2012. "Losing my revolution: How many resources shared on social media have been lost?" *arXiv* (13 September), at http://arxiv.org/abs/1209.3026, accessed 11 November 2014; http://perma.cc/U3Q2-R8YM, accessed 24 April 2015; also in: P. Zaphiris, G. Buchanan, E. Rasmussen and F. Loizides (editors). *Theory and practice of digital libraries: Second international conference, TPDL 2012, Paphos, Cyprus, September 23-27, 2012. Proceedings*. *Lecture Notes in Computer Science*, volume 7489. Berlin: Springer-Verlag, pp. 125–137.
doi: http://dx.doi.org/10.1007/978-3-642-33290-6_14, accessed 24 April 2015.

N. Taylor, 2014. "The MH17 crash and selective Web archiving," *The Signal: Digital Preservation* (28 July), at http://blogs.loc.gov/digitalpreservation/2014/07/21503/, accessed 13 November 2014; http://perma.cc/7TGU-BBWJ, accessed 24 April 2015.

N. Taylor, 2011. "The average lifespan of a Webpage," *The Signal: Digital Preservation* (8 November), at http://blogs.loc.gov/digitalpreservation/2011/11/the-average-lifespan-of-a-webpage/, accessed 11 November 2014; http://perma.cc/FL5X-R285, accessed 24 April 2015.

Z. Whittaker, 2010. "Facebook does not erase user-deleted content," *ZDNet* (28 April) at http://www.zdnet.com/blog/igeneration/facebook-does-not-erase-user-deleted-content/4808, accessed 13 November 2014; http://perma.cc/2ECY-HTRQ, accessed 24 April 2015.

A. Wright, 2009. "Exploring a 'deep Web' that Google can't grasp," *New York Times* (22 February), at http://www.nytimes.com/2009/02/23/technology/internet/23search.html, accessed 13 November 2014; http://perma.cc/MG5L-QRBW, accessed 24 April 2015.

D. Zeitlyn, forthcoming. "Looking forward, looking back," *History and Anthropology*.

D. Zeitlyn, 2012. "Anthropology in and of the archives: Possible futures and contingent pasts. Archives as anthropological surrogates," *Annual Review of Anthropology*, volume 41, pp. 461–480.
doi: http://dx.doi.org/10.1146/annurev-anthro-092611-145721, accessed 24 April 2015.

J. Zittrain, K. Albert and L. Lessig, 2013. "Perma: Scoping and addressing the problem of link and reference rot in legal citations," *Harvard Public Law Working Paper*, number 13–42, at http://papers.ssrn.com/abstract=2329161, accessed 24 February 2014; http://perma.cc/4DVN-DYS8, accessed 24 April 2015; also in *Harvard Law Review*, volume 127, number 4 (2014), at http://harvardlawreview.org/2014/03/perma-scoping-and-addressing-the-problem-of-link-and-reference-rot-in-legal-citations/, accessed 24 April 2015.

Editorial history

Received 28 January 2015; revised 30 March 2015; accepted 7 April 2015.

Learning from failure: The case of the disappearing Web site
by Francine Barone, David Zeitlyn, and Viktor Mayer-Schönberger.
First Monday, Volume 20, Number 5 - 4 May 2015

http://firstmonday.org/ojs/index.php/fm/rt/printerFriendly/5852/4456
doi: http://dx.doi.org/10.5210/fm.v20i5.5852