Introduction: The web as history

Ralph Schroeder and Niels Brügger

The web as a reflection of society

The web has been with us for more than a quarter of a century. It has become a daily and ubiquitous source of information in many peoples’ lives around the globe. But what does it tell us about historical and social change? For a researcher in the twenty-second century, it will seem unimaginable that someone studying the twenty-first century would do anything but draw heavily on the online world to tell them about peoples’ changing lives. Currently, however, the web remains an almost untapped source for research. This book aims to make a start in this direction.

If the importance of dusty – or digital – archived material seems like something that would be mainly of importance to academics, consider the following two examples: In late 2013, it was discovered that the UK Conservative Party had deleted political speeches that it might find inconvenient from the party’s websites and had also throttled access to these sites via Google and the Internet Archive. Cowl (2013) notes that, ironically, these speeches include one by the then Conservative leader David Cameron where he admonished politicians and others not to keep information secret. This discovery led, of course, to attempts to track down this material which had, as it turns out, been archived in a special collection by the British Library (Guardian, 2013). This incident highlights the importance of web archives as a matter of record, and in the end drew more negative attention to the websites than the Conservatives had hoped to avoid by deleting the information in the first place.

Another example is the 2014 shooting down of a passenger plane over the Ukraine during the war between Russians and Ukrainians. A Russian claimed to have shot down a Ukrainian military plane on social media, a post which was then deleted but found later via the Internet Archive, as the New York Times (2014) reported. There was
an extensive investigation which subsequently determined who was responsible for this incident. The point of both examples is that accurate records matter, and this applies to the digital realm just as much as it did for paper records and many other sources of evidence about the past.

As the following sections will show, while much has been written about the methodological and other challenges of using the web to understand the past; substantive studies which do just this are still thin on the ground. In this volume, we present a series of such studies which illustrate such early – but also rich and diverse – ways to use the web in this way. But before we summarize the chapters, it may be useful to discuss briefly what we know about how people use the web and how these uses shape and are shaped by the web. When using newspapers as a means to understand history, for example, we also want to know something about newspaper readers and journalists; though in the case of the web, the distinction between content consumers and producers (to put it differently) may be more difficult to define. This will serve as a background for the second section which reviews the history of the emergence of web archives and how the ground has been prepared for their use by researchers. The last section of this chapter will then give an overview of the volume’s contents.

The web in context

Before we discuss web archives and how they can be used to study social change, it is important to discuss a topic that is not covered in this volume (and indeed, about which little is known to date); namely, how the web is used. After all, if web pages are going to tell us about changes in society, we also need to know who reads – or watches, or listens to – the web. Part of the difficulty is that the web is a new medium, but like the internet, it has not yet been adequately theorized as such. To recognize this point, it can simply be noted that research about the web partly falls within media studies, which is concerned with communication, but also partly within information science, which deals with how people seek information. There are many difficulties here which cannot be resolved in a short space, but we will indicate briefly what we know about web ‘audiences’ or ‘consumers’ of online information. This is important because how the web is received in society will ultimately be a necessary backdrop for understanding the social significance of the patterns in the information that can be found online.
A good place to start is by considering the extent to which the web is a single entity – or if its use reflects offline political or cultural or linguistic borders. This is an interesting question because it has often been claimed that the web is a unique medium insofar as it can be accessed from anywhere – unlike traditional media that are confined, for example, by national broadcasting regulations or by the reach of transmitters and the like. In other cases, most notably in China, it has been argued conversely that the government and its censorship regime ringfence the web, making it into a cultural resource whose reach is circumscribed by the state. Both ideas are misleading, as Taneja and Wu (2014) have shown: first, in a certain sense, access to the web in China is no less densely bounded off from the global web than is the case for other non-English speaking large clusters on the web. The way that Taneja and Wu arrive at this finding is by examining traffic to the top 1000 websites (which together receive more than 99% share of attention globally), and then grouping these into sites that receive shared attention. Shared attention is defined as: if someone clicks on one site, they also visit another (after controlling for the statistical chance of co-visiting). One possible reason for this finding is that in the case of China, apart from language, an active policy by the party-state has promoted a Chinese-centric web, as in other cases of state-driven information technology policies such as Korea’s (see also chapter six by Schafer for the French case). But the Chinese web is not uniquely circumscribed by a wall of censorship, as some have argued; instead, it is mainly that Chinese citizens, like those of other nations, are primarily interested in content produced in China.

Wu and Taneja (2015) have extended this analysis to argue that the ‘thickening’ of the web has changed over time. Whereas in 2009 a Global/US cluster was predominant on the web and at the same time the largest, in 2011 it was overtaken by a Chinese cluster and there was no longer a Global/US cluster but instead in second place was a US/English cluster followed by a global cluster. The same two clusters occupied the top two spots by size in 2013, but the global cluster (of websites that are not language specific, such as Mozilla and Facebook) had slipped to 8th place (India was 9th and Germany 10th) followed by a number of other clusters including Japan and Russia but also Spain, Brazil and France. What we see here is the orientation of the web evolving towards the Global South (Spanish-speaking and Brazil, and also India). At the same time, it should be remembered that the shift towards the Global South is highly selective, as shown in a different study of the least connected continent, Africa (Boldi et al., 2002). In this case a web crawl of African websites revealed that the number of web pages
was approximately 2 million in 2002, a very small number, and that
almost 75% of these were in English, which is spoken as a first language
by far less than 1% of the African population.

In any event, returning to the study by Wu and Taneja (2015) we
see that, with time, the websites of ‘global’ status have become fewer in
number among the world’s top 1000 sites, and we see language playing
an increasing role over time. State policies promoting information and
communication technologies are one factor here, and shared language
another. Whatever the most important factors may turn out to be, the web
is not becoming a single whole, but rather a series of clusters – influenced
by linguistic factors and the policies of states and sites promoting shared
interests such as commerce or personal relations. In terms of the analyses
which are based on national and other large-scale domains in this volume,
or of the chapters which deal with cultural and social phenomena span-
nng multiple countries and languages, or of the several chapters which
use link analysis to identify clusters among issues, organizations and
transnational connections (especially chapter five by Cowls and Bright), it
is easy to see that where content is accessed will have major implications
for the changing shape of the web.

Online information in everyday life

In addition to a bird’s-eye perspective, we could also look at web uses
from the ground up, how people use the web in everyday life. Such
research on how people search for information, for example, is still
thin on the ground (Rieh, 2004; Savolainen, 2008; Aspray and Hayes,
2011; Schroeder, 2014). A major issue that has not yet been resolved in
media or communication studies is where to ‘put’ information seeking
in general. A simple way to grasp this point is to ask: where did people
seek information before the advent of the web, say, in the mid-1990s?
(The same point could be raised, of course, in relation to Wikipedia, and
search engine behaviour.) They might have consulted an offline ency-
clopaedia instead of Wikipedia, a travel agent instead of a travel website
(one of the chapters in this volume is about TripAdvisor), an offline pam-
phlet instead of a blog and so on. Yet these ‘media’ were also not much
studied. What makes the web different is that it contains all of this infor-
mation, but also that none of these uses of the web is easily categorized
within the study of offline behaviour or other digital media – or indeed
the study of mass and interpersonal communication. Where these uses
can be categorized is in the areas studied by information science, but
that is a discipline that barely overlaps with communication studies (and that deals mainly with educational, research and library searches). In any event, the web, in view of the fact that it is a large and accessible source of data and increasingly important in peoples’ lives, is bound to grow as a topic of research.

At this ‘micro’ end of the continuum, we could also examine the scholars who archive the web for a specific research project, the companies that keep web archives for legal reasons, or individuals and groups who simply want to preserve a portion of the web for whatever purpose. One study by Lindley et al. (2013) interviewed people – who were selected on the basis of being sophisticated users of digital technologies – about their personal digital archiving habits. One might expect such people would be starting to put their online materials together in a similar way to the manner in which they keep diaries, photo albums and other collections of mementos. What Lindley et al. found, however, was more complex. First, people archived their materials as part of a wider information management process, including the content on their social media sites, and their archiving was thus spread across a number of platforms. Second, the process of archiving was not an individual pursuit. Instead, people would, for example, rely on friends or family members to be able to keep a record of certain events. Third, much of the content is neither archived nor backed up since it is thought (often no doubt mistakenly) that it can be easily found again by searching through one’s file systems. Furthermore, much material, for example photos on a photo sharing site that is no longer used, are simply abandoned or discarded as not being worthwhile (again, there are many resonances, as the reader will find, with much web material that has been lost for one reason or another). Fourth, people regarded different sites or platforms as different facets of themselves, without any need for integration.

Hence, while one might expect people to be worried about keeping their personal material in an online storage system or controlled by organizations, in fact, they used diverse methods, abandoning certain sites and maintaining their records in collaboration with others in their networks. This indicates that the practices of curating one’s personal life online as a means of keeping a record has not yet settled down into a consistent and well-organized practice, and perhaps it never will. In this sense, it mirrors the early uncertainties of professional and academic archiving practices that will be mapped in the next section of this chapter. These individual-level archives also mirror the efforts of other entities – institutions such as firms, non-governmental organizations or
even governments – to keep records or institutional memories of themselves, which are also in a state of flux.

Finally, an obvious way to gauge the influence of the web is to measure the original audience for a given website, or collection thereof. Brügger (2012a: 318) has shown that one way to assess the influence of a given website is through analysis of the number of visitors overall, combined with the number of internet users in countries in which the website is most salient. Another approach is to use aggregate ranking sites such as Alexa (http://www.alexa.com/), but otherwise little is publicly known about who uses the web in general. Two exceptions are Waller (2011) who has examined information seeking by Australians, and Segev and Ahituv (2010) who provide a more global perspective. Wu and Taneja (2016) have more recently contributed to our understanding of attention paid to the world’s top websites by grouping them by format and genre and in terms of their popularity.

**Web archives and researchers**

Against this background of uses of the web, we can now turn to how the web can be used as a resource for scholarship. After the first web page was published in 1991 by Tim Berners-Lee, the inventor of the web, it took some five years before large-scale attempts to preserve the online web were initiated. From the mid-1990s the landscape of web archives started to evolve slowly with a number of web archives being established aimed at preserving the cultural heritage.

The landscape of the web of the past

Early attempts to archive material on the internet, including the web, were carried out in Canada in 1994–1995 (Brügger, 2011; Webster, 2017), but it was not until 1996 that the first major international initiative was launched, namely the Internet Archive. The Internet Archive was founded in 1996 by Brewster Kahle, who had made a considerable fortune as an internet entrepreneur. He established the Internet Archive as a non-profit organization, located in San Francisco and with the aim of preserving digital media, including the web. The Internet Archive began by creating a relatively small collection, namely the websites of the 1996 Presidential candidates (cf. Kimpton and Ubois, 2006: 202), but soon after initiated its broad web collections based on following hyperlinks. The Internet Archive collects that to which hyperlinks point, which is
why it is transnational by nature.\(^2\) As of today the Internet Archive holds the world’s largest collection of the preserved web from the past. It is also worth noting that the Internet Archive has established a priceless treasure trove, as well as being instrumental in promoting web archiving internationally. It has developed software that is widely used to collect web content (the web crawler software Heretrix), an archiving file format (ARC, and later WARC) and software to replay the archived web material (the Wayback Machine) (cf. Koerbin, 2017; Webster, 2017). Furthermore the Internet Archive has played an important role in the establishment, in 2003, of the International Internet Preservation Consortium (IIPC) that has since that time provided an important forum for debates, knowledge sharing and technical developments about web archiving.\(^3\)

In parallel with the establishment of the Internet Archive, a number of other national web archiving projects were initiated. These include ‘PANDORA Australia’s Web Archive’, and ‘The UK Government Web Archive’ in 1996, followed by the Swedish ‘Kulturarw3: Kungliga bibliotekets webbarkiv’ in 1997, the ‘New Zealand Web Archive’ in 1999 and the ‘Library of Congress Web Archive’ as well as the ‘Webarchiv – Czech Web Archive’ in 2000. National web archives really began to take off after the turn of the millennium: 2001 (Norway), 2002 (France, Japan), 2004 (Croatia, Iceland), 2005 (Denmark, Korea, Latvia, and the UK), just to mention a few. By and large, the establishment of national web archives has mirrored the general spread of the web. They were first established in North America, Northern Europe and in parts of Australasia. To the best of our knowledge there exist no national web archives in South America and Africa. Regarding South America, the University of Texas hosts the ‘Latin American Web Archiving Project’ (LAWAP) which, since 2005, has collected a variety of web material from throughout the Latin American continent (see http://lanic.utexas.edu/project/archives). As for Africa, there is a collaborative project entitled Current Events in Africa Web Archive (CEAWA) (led and funded by the Africana Librarians Council’s Title VI Librarians). Since 2014 this project has archived websites that document current events in African countries (https://archive.org/details/ArchiveIt-Collection-4426). Both of these initiatives are hosted by the Internet Archive’s subscription service Archive-It (see later in this section).

In many cases the national web archives have continuously developed their archiving remit as new legal frameworks were passed, allowing them to broaden their scope for collecting. For instance, the UK Web Archive started in 2005 as a collection of websites of leading UK institutions, based on selection criteria such as historical, social and cultural
significance. Since April 2013 the UK Web Archive has also been allowed to archive the whole of the UK web domain (as stated in The Legal Deposit Libraries (Non-Print Works) Regulations 2013, § 16). Hence, the UK Web Archive’s highly selective collection of a limited number of websites has been expanded with the Legal Deposit collection’s broad archiving of the entire national web domain.

It is also worth noting that the establishment of national web archiving initiatives is embedded in country-specific institutional settings which entail major differences in how each country approaches web archiving, ranging from countries with no national web archive (such as Belgium or the USA) via countries with only one national web archive (such as the Netherlands and Denmark) to countries with more than one national web archive, such as the UK which has the UK Web Archive plus the UK Government Web Archive (the latter preserves the UK government information published on the web) or France, where the Bibliothèque Nationale de France Web Archives focus on the French web in general while the web archive of the Institut National de l’Audiovisuel archives audiovisual media related to websites.

But web archives are not only to be found in the form of national archiving institutions. Many university libraries have also established web collections, in the main with a focus on specific topics of relevance for each university, be that the university’s own website, or research topics of importance for the faculty. Web archives at university libraries are particularly widespread in the USA, which is partly due to the absence of a national web archive, although the Library of Congress de facto to a large extent fills that function. For instance, the UCLA Library began web archiving in 1998 with a focus on election campaigns, in continuation of the library’s already established ‘UCLA Online Campaign Literature Archive’ that had a longstanding tradition of collecting campaign material related to Los Angeles and California elections. Some of the first to follow this lead were the Harvard University Library Web Archive Collection Service (2006), Stanford University Libraries (2007) and Columbia University Libraries (2008) (see Truman, 2016: 47–77 for an overview).

Other forms of institution such as museums and art communities have established web archives, an early example being the born-digital arts organization Rhizome’s ArtBase that since 1998 has collected more than 2000 pieces of internet art, including websites (http://rhizome.org/art/, and Truman, 2016: 67–8).

Five other types of web collections can be mentioned to complete this outline of where to find the web of the past. First, a number of
professional vendors offer web archiving services, such as the Internet Archive’s subscription service Archive-It, or the Internet Memory Research’s Archivethe.Net. In the main these services do not build their own collections, but rather function as operators for their subscribers, including national web archives, researchers, universities, museums, institutions and companies. These collections are often made accessible through the websites of the vendor alongside the website of the subscriber, as can be seen for instance with Archive-It (see https://archive-it.org). Second, there are web collections archived by researchers in relation to particular research projects. These collections can be very hard to find because no systematic overview exists, they may not be publicly available or they are not usable for other studies if they were created with a specific research project in mind. However, in some cases research collections have been established based on collaboration between researchers and university libraries, for example, The Human Rights Web Archive @ Columbia University (http://hrwa.cul.columbia.edu) (cf. also Webster, 2017). Third, there exist a number of publicly available collections, archived by individuals or groups with a strong interest in preserving specific parts of the web, but with no explicit obligation to cultural heritage. These collections include, among others, The Archive Team Geocities Snapshot (www.archiveteam.org), or Common Crawl’s open repository of web crawl data (commoncrawl.org). Fourth, specific parts of the web of the past that had actually disappeared may have been meticulously restored and put online. This is the case for the project ‘Restoring the first website’ which has restored material from the first web server info.cern.ch, including machine names and IP addresses (cf. http://first-website.web.cern.ch, see also Koerbin, 2017). Fifth, although it may not be considered a collection in the strict sense of the word, one should not forget the online web itself while looking for the web of the past. The web may still hold old web material, such as screen shots of Facebook pages or screen movies, or old material that is simply still available on the web (e.g. an early screenshot of browser windows on Tim Berners-Lee’s desktop, https://www.w3.org/MarkUp/tims_editor).

Making the web of the past useful for scholars

As with any other collection of documents or artefacts, so too for web archives: the ways in which things are collected, made accessible and documented have an impact on how they can later be used by researchers. Therefore a brief account will be useful for some of the fundamental
choices involved in the collecting of the web as well as in making the archived web accessible and documenting it.

Since it is impossible continuously to archive the web in its entirety, let alone a national web domain or even a smaller group of websites, an institution or person performing the archiving must have a strategy to decide what should be archived and what is deliberately omitted. Collection strategies can be placed on a continuum, ranging from selective collections of individual websites to broad collections with almost no limitations on what to include. An example of the first is the Australian PANDORA, while the Internet Archive is an example of the latter. In between, there are thematic collections related to events, to a topic, or other such demarcations (which are closer to the selective strategy), and strategies aiming to archive entire regional or national web domains, which are closer to the very broad collections. However, in most cases, web archives adopt a combination of several strategies, for instance the Danish Netarkivet uses three strategies (selective, thematic and broad national).

A collection without access does not make much sense, but for a variety of reasons (e.g. copyright, privacy, national legal frameworks), accessibility to web collections varies. It is important to distinguish access to the collection as such from access to the concrete material held in the collection. In terms of access to the collection, a scholar who wants to study the archived web is faced with a landscape where in some cases access may be online and open for all, and in other cases access may be so restricted that the web archive is literally closed. The Internet Archive, the Library of Congress, the Portuguese Web Archive, The Human Rights Web Archive @ Columbia University and The Archive Team all offer open access, whereas the Norwegian web archive offers only very restricted access. Between these two extremes, we find that different kinds of restrictions apply. Some web archives are open to a wider public but have to be accessed on site (such as the UK Web Archive’s Legal Deposit collection, or the Dutch web Archive), while others are only open to researchers, but once access is granted, they have access online (such as the Danish Netarkivet). And although access may be granted on site, severe restrictions on the use of the content may apply: with the UK Web Archive’s Legal Deposit Collection, for instance, users may only print a small portion of the archived content, no digital copies may be made, and a web page may not be accessed if it is being consulted at the same time by any other user in the library (cf. Webster, 2017).

Once a scholar has access to the collections, the next question is in what form he or she will get access to the concrete material held in the
collection. Since 2001, the main form of access to web collections has in most cases been through the interface of the Wayback Machine. The Wayback provides a browser-based interface where the user has to insert the web address (URL) of the web page he or she wants to retrieve, and once this is done, the Wayback presents the web page in a manner close to how it looked when online. From a researcher’s point of view, seeing the web page close to how it looked in the past is obviously beneficial, but this approach comes with a number of drawbacks, most notably that the scholar has to know the exact web address to find the material, and if s/he wants to find more than one web page, all the relevant web addresses have to be inserted and searched manually, one by one. Therefore, a number of web archives such as PANDORA, the Portuguese Web Archive and the Danish Netarkivet have established full text search, which means that the search interface allows for searching all types of content in the entire archive, including the body text itself. As an intermediate solution between URL and full text search, some collections have full text search of metadata (e.g. the Library of Congress). But full text search also comes with a number of challenges, including how to present and possibly rank thousands, or even millions, of hits in a user-friendly and relevant way (parcelling the search results by year and top-level domain name such as .com, .gov etc. may help, but the challenge is still significant).

In addition to URL and free text search, new ways of giving access to material in web archives have recently been launched. The Portuguese Web Archive, for instance, provides API (Application Program Interface) access to its collection, and the Internet Archive’s subscription service Archive-It has established the Archive-It Research Services (ARS) that provides access to data sets extracted from collections, such as metadata, link graphs and named entities.

Finally, the researcher who wants to study the web of the past is very likely to ask for documentation. In general, scholars would like to have access to as much information as possible about the provenance of what they study. For web archives, documentation, at different stages of the research process, can range from the collection level down to each individual web object, be that an image, a piece of graphics or a sound file. Whereas documentation about the collection is most likely to have been created manually, for instance by curators, the more fine grained types of documentation relating to individual web objects may be automatically generated. This is because although the relevant information is there, it has to be made available at the right moment in the research process and in a useful manner. However, for the time being, most web
archives only offer documentation either about their collections, or about individual websites if the collection is based on selective collecting and curation. But in many cases even this documentation is scarce.

In summary, a major challenge for the scholar wanting to study the archived web is to get an overview of where specific websites or clusters of websites may have been archived, if they are archived at all, since there is no overall registry of collections in all web archives. Once the relevant web material has been found, access has to be ensured, be that to the collection or to the archived materials, in such a form that supports the research project and that provides enough documentation on what is actually being studied.

Collaborations between web archives and scholars

Looking back on the history of web archives, and in particular large-scale transnational and national web archives, it is striking that in most cases they were not established to accommodate the needs or interests of researchers (cf. also Webster, 2017 on this point).6

The majority of web archiving projects were initiated either to preserve a variety of digital cultural products (e.g. the Internet Archive) or as a continuation of pre-existing national traditions of collecting and preserving the print or audio-visual cultural heritage. Hence, for a number of years web archives and researcher communities developed independently. Web archives were struggling to set up archiving procedures, hardware and software to keep pace with the seemingly endless flow of new web content and ever evolving software development, while little attention was paid to who might use the material in the archive, and how it might be used. And the research communities who could have benefited from accessing the archived web, including among others internet and media scholars, historians and social scientists, have shown little interest in a highly relevant source that could have added a novel dimension to their analyses.

However, within the last five years a shift has slowly emerged internationally – the first indication of which is the 2010 report Reseacher Engagement with Web Archives: State of the Art (Dougherty et al., 2010). Web archives are now more likely to involve researchers in discussions of collection policies and access forms, and increasingly scholars are starting to discover this new resource with all its pitfalls and challenges. As the contributions to the present volume highlight, web archives may hold a valuable potential for novel research projects as well as for approaching well-known research topics from a new
perspective. To fully realize this potential, sustainable collaborations must be created to ensure common standards, as well as researcher tools aimed at the skilled and novice web researcher, including sophisticated search tools, basic analytical software, tools for the creation of sub-collections and for exportation of data, and possibly also a wide range of API access-forms. And such initiatives must be combined with training courses with a view to disseminating knowledge to larger research communities. A number of collaborations between web archives and researchers have been initiated in recent years, and these projects can still serve as the inspiration for future joint ventures, whether in time-limited research projects or in long-term sustainable fora within already existing organizations, such as national research infrastructures or a transnational association such as the IIPC.7

Building on existing literature

This edited volume is the first book-length publication to focus on how the archived web of the past can be used as an entry point to analysing societal developments at large. But it builds upon several existing bodies of literature, including works on web archiving, the methodological challenges related to use of the archived web, internet and web history in general, and the broader field of digital history. The following brief account does not pretend to be comprehensive. Instead, by way of mentioning early examples, it will give an impression of how the literature originated and continues to feed into and inform the emerging nexus between the archived web and its use by researchers.

The first scholarly interest in the web of the past emerged within the web archiving communities: computer scientists, curators, software developers and others (e.g. Brown, 2006; Masanès, 2006; cf. the overview in Ayala, 2013). In general, this literature is not grounded in the traditions of scholarly users of web archives, but there is a very limited literature that highlights some of the impacts that the archiving process may have on researchers' use of the web archive (e.g. Brügger, 2005; Dougherty et al., 2010).

From the mid-2000s, publications started to reflect on some of the methodological challenges related to the scholarly use of the archived web (e.g. summarized in Brügger, 2011, 2012b), and in some cases were combined with empirical studies (e.g. Schneider and Foot, 2006). There are several books on general internet history (e.g. Naughton, 2012, 2015; Abbate, 2000; Poole, 2005; Goggin and McLelland, 2017) which
provide valuable insight into the history of the internet, though not as much about the history of the web. Empirical studies of the web exist (e.g. Gillies and Cailliau, 2000; Schneider and Foot, 2006; Banks, 2008; Brügger, 2010; Burns and Brügger, 2012; Salter and Murray, 2014), but this literature only partially examines the archived web.

Finally, there is an important body of literature about historiography and the digital (e.g. Cohen and Rosenzweig, 2006; Weller, 2013). However, this tradition is mainly concerned with the web as a medium for the distribution of sources and research results, and not as a historical source in its own right (exceptions being Rosenzweig, 2004, and more recently Graham et al., 2015).

As this brief account shows, the development of a literature relevant to someone wanting to use web archives to understand the past and the present mirrors the research process. Initially the literature concerned the sources to be studied and how they could be collected, preserved and made accessible; then came reflections on how these sources could be approached, and subsequently the first tentative empirical studies, in some cases inspired by internet history and digital history.

Thus the time is now ripe to take the next step and start considering the web as history, and to make the web of the past come alive, adding an important voice to our understanding of society in the last two decades. Recently, the field has taken a computational turn towards big data. Guldi and Armitage have argued that using big data allows for ‘realigning the archive to the intentions of history from below’ (2014: 93). This depends, however, on whether digital sources accurately represent the forces from ‘below’. And as we shall see, uses of web archives can take quantitative and qualitative approaches, and often both.

**Future research**

This book makes only a start in this nascent area of research. Before we summarize the chapters, it can be pointed out briefly that there are many possibilities for future research into using the web to shed light on the past and the present.

In outlining these, it can be reiterated that the web itself is changing. Hence one question that must be asked is where the boundaries of the archived web lie: is all content on social media included? Or app content? No doubt many further additions to the web will emerge, and capturing these, as they increasingly displace other media, will be a challenge. Second, there is the question of macro- versus micro-,
quantitative versus qualitative integration: how can we make sense of the relation between the global web, and individual sets of web pages pertaining to specific topics? Third, if the web is to be used as an indicator of historical and social or cultural change, there must be a way to understand how the web is used in everyday life: what information do we seek, and need, on a daily basis? Does the web shape, or is it shaped by, these needs? Finally, there are many ways to build on and extend the studies mentioned in this introduction: by examining the changing shape of the web as a whole (and the parts that have, and are still, disappearing or unarchived), or of national webs (and especially those parts of the world that have hitherto been neglected, like Africa), and the myriad subsets of pages, their coherence and disparateness, and the abundant materials that make up the web. These future topics also mean that there is much theoretical work to do: how can the findings from these studies be integrated into studies of other media? These areas of future research constitute wide and almost virgin territories for scholarship, and are bound to open many new directions, some of which are as yet difficult to foresee.

Overview of the chapters

Three contributions in this volume, grouped in Part one, take a quantitative approach to whole populations of web pages or to a whole national web sphere or domain. Chapter one by Meyer et al. examines the UK domain, or web pages ending in .uk, and in particular the academic part of this domain ending in the UK in ac.uk (in the USA, this would be .edu; other countries have different ways of marking university domains). What they find is that the .ac.uk domain was one of the initial driving forces of the web, which is indicative of the strong role that the universities played in the early days of the internet and web, but which then plateaued. The same applies to .gov.uk (the government’s websites), but not to .co.uk (in the USA, the equivalent is .com), which has continued to grow apace. Since the authors are examining all .uk domains, they can also do a link analysis, showing the interlinkages (numbers of hyperlinks) between these various sectors, which also include .org. This type of analysis of the changing shape and relations between subdomains may then shed light on changes in society, but to do this, it will probably be necessary to compare different national domains and their shapes or trajectories.

In chapter two, Hale et al. take a different approach, drawing on the same source of national level data of the whole .uk domain. They
examine the extent to which web pages within a commercial website – TripAdvisor, a popular travel site – are reflected in the UK data of the Internet Archive. What they find is that the two match each other only very unevenly: pages for the most popular or prominent tourist attractions are present, but pages for lesser known attractions are missing. The implication is that what gets archived is not a representative subset of the live web. One could go further: the fact that even a website as well-known as TripAdvisor is captured unevenly with a bias towards more prominent pages does not bode well for social science or humanities research requiring comprehensive or representative data.

A third approach by Brügger, Laursen and Nielsen in chapter three is to look at how domain names have changed over time, in this case comparing the Danish web archive and the Internet Archive, and focusing in particular not just on the growth, but also the disappearance of domain names. There are two contributions here: one is to compare the comprehensiveness or otherwise of two web archives. Since it is not yet established how solid various archives are as a matter of record, testing them will provide important indicators of their reliability. The second contribution is to show how much of the web is disappearing even as it is continuously experiencing growth. That too will be of interest to historians and others who are seeking to understand what gets lost in the record, which may also be important for how we view the past.

Part two, Media and Government, moves to chapters that combine quantitative and qualitative approaches. So, for example, in chapter four Weber charts the evolution of online newspapers in the USA, where there have been dramatic changes. But apart from the larger American picture, Weber also analyses local online newspapers in New Jersey, using, like others, the Internet Archive to do so. He shows, via an analysis of the links between these local papers, that larger national transformations – where there has been a shakeout with only a few players surviving – are also replicated at the local level. Obviously a link analysis is only one way to chart these changes; others would include measuring the changing revenues of newspapers. But since links are tied to visibility, this kind of analysis can provide an important starting point.

Using a different quantitative methodology, Cowl and Bright, in chapter five, also analyse the evolution of news, in this case the international links to and from the website of BBC News. This chapter ties to larger debates about whether, with the increasing globalization of news, news content is nevertheless biased to richer countries or countries which have other characteristics such as military conflict or economic
ties. Their findings are that less peaceful countries receive fewer out-
links from the BBC site, even when they receive more coverage, and that
countries using the English language also receive more links. Again,
there are implications for visibility of certain places or languages.
Perhaps more importantly, the BBC news site has a very wide readership
and is well-known around the world. Thus it will be necessary to study
the spread of other online news sites with an international reach to com-
plement this analysis of one media organization. Once this is done, we
will have a powerful understanding of whether the shift online is lead-
ing to greater global interconnectedness – or rendering certain parts of
the world even less visible.

Government uses of the web are also revealing. Schafer has
unearthed the early and difficult attempts of the French government
to reach out and engage with its citizens. As she shows in chapter six,
this effort was driven partly by French politicians who rode the wave of
enthusiasm for digital solutions during the 1990s, on the one hand, and
by the distinctive culture of the French internet, with its national Minitel
system, on the other. Yet grand ideas about making the administration
more efficient and interacting with citizens mostly petered out into spo-
radic informational web pages led by a few local administrations. These
were innovative at the time, but nowadays strike us as rather dated.
Here we can see how history ‘on the ground’ looks different from his-
tory as it is written by reference to French politicians who were largely
responding to the rhetoric of vice-president Al Gore’s idea of an ‘infor-
mation superhighway’.

Part three delves deeper into particular cultural phenomena.
Milligan’s chapter seven is about a virtual ghost town: GeoCities was
among the first and largest online community spaces on the web, a
thriving place where people put up web pages in the manner of cre-
ating a home in an online neighbourhood. Milligan describes how
people presented themselves on their virtual homesteads and how
they interacted with each other. This is a fascinating story, although
GeoCities was abandoned – partly for commercial reasons but also
partly because having an online webspace became much more com-
monplace and because the geographical metaphor became increas-
ingly outdated. Another part of the story is how GeoCities was only
preserved due to the efforts of some of its dedicated former inhabitant
web archivists, reminiscent of offline organizations for the preserva-
tion of historical monuments. Again, we see that what almost disap-
peared from view is just as important as what remains, as with other
historical artefacts.
Tracking how a particular issue has been represented on the web is yet another approach to using web archives. Ackland and Evans do this in chapter eight for the abortion debate in Australia by means of hyperlink analysis and text analysis. Among other outcomes, they find that the commercial sites offering abortion drugs became more prominent, partly because they had not been approved at the start of the decade examined by Ackland and Evans. Another finding is that while the pro-choice and pro-life sides were roughly equally prominent at the start of the analysis period, over the decade the pro-choice side became more visible (though also more diffuse), which may give a clue about the direction of public sentiment over this period. Finally, the authors discuss how they used Google to gauge visibility, which can be justified inasmuch as Google is by far the most widely used search engine in Australia (and a link can be made here to the study by Waller, mentioned in the section above on the web in context, which examined Australians’ search behaviour using Google). However, as in other big data studies, the reliability of this source and its bias towards commercial websites may hamper its usefulness as a source for researchers.

A different contentious online issue is discussed by Webster in chapter nine which considers sharia law in the UK. Webster looks at how the Church of England, via its figurehead the archbishop of Canterbury, became embroiled in a controversy over Muslim–Christian relations. Webster also details the involvement of the British National Party, an extreme right-wing organization which responded vehemently to the archbishop’s position. Webster uses link analysis, as do several other chapters, but he also engages in a close reading of some of the relevant web pages and situates the debate in the larger context of the role of the Church of England in Britain. Inasmuch as controversies such as these increasingly take place online, and where there are multiple sources through which to document them (although as Webster notes, some, such as the official papers of the archbishop, will not become available for a long time), the web can be a powerful resource to chart the relations between contending factions on major issues. And the relation between Islam and other faiths in Britain and beyond is certainly bound to be an issue of continuing interest.

Islam is refracted through a quite different lens in Dougherty’s chapter ten, which follows the development of Islamic Punk in North America. This subcultural movement has already faded, so the web provides a major record of its rise and fall. Documenting the movement will be a valuable source for understanding the cultural norms of young Muslims in North America who, for a brief period, took to a particular
form of punk music to express their allegiance both to Islam and to a popular genre of music. The discussions within this subculture, over what it means to be Muslim and at the same time identify with music that saw itself as non-conformist, shed interesting light on how young Muslims saw themselves during a time when there was considerable controversy about the place of Islam outside of its homeland, and in the USA in particular. Dougherty also raises questions about where the most revealing documentation can be found, since many of the relevant discussions can be found not in the Internet Archives, but often on moderated lists such as Reddit, or on Wikipedia and blogs, as well as via various news sources.

Finally, chapter 11 by Cowls, summarizes the outputs from a project with which the co-editors were involved, the Big UK domain data for the Humanities (BUDDAH) project (see footnote 7). This project gave bursaries to a number of pilot studies which used the UK web archive to examine particular topics. The topics covered a wide range, including the web presences of the UK military, of Beat poets, of disabilities charities and of right-wing political groups. The studies illustrate some of the challenges of using web archives to research specific topics, since searches for these topics in the archive will yield a vast number of web pages. But they also show that non-specialists – most of the studies were written by non-academics – can produce valuable documentation and insights to chart the evolution of organizations, movements, and cultural and historical trends by using different approaches. For readers who are interested in pursuing research using the web themselves, this chapter will be a good place to start.

The Coda to this volume is written by Jane Winters, who led the BUDDAH project and who is Professor of Digital Humanities at the School of Advanced Study at the University of London. Winters reflects on the many debates that have surrounded the use of the web for research, especially in history where there have been vigorous discussions about whether big data and quantitative approaches are useful. Such debates are to be expected in a nascent field, especially in the humanities where there is little task certainty and mutual dependence (Meyer and Schroeder, 2015) – unlike in other areas of research – but where many new directions in scholarship have rapidly taken off, proliferated and gained much attention.