



The

Web



as

History

Edited by **Niels Brügger** and **Ralph Schroeder**

 **UCLPRESS**

The Web as History

*Using Web Archives to Understand
the Past and the Present*

Edited by

Niels Brügger and Ralph Schroeder

 **UCL**PRESS

First published in 2017 by
UCL Press
University College London
Gower Street
London WC1E 6BT

Available to download free: www.ucl.ac.uk/ucl-press

Text © Contributors, 2017

Images © Contributors and copyright holders named in captions, 2017

A CIP catalogue record for this book is available
from The British Library.

This book is published under a Creative Commons 4.0 International license (CC BY 4.0). This license allows you to share, copy, distribute and transmit the work; to adapt the work and to make commercial use of the work providing attribution is made to the authors (but not in any way that suggests that they endorse you or your use of the work). Attribution should include the following information:

Niels Brügger and Ralph Schroeder (eds.), *The Web as History*. London, UCL Press, 2017. <https://doi.org/10.14324/111.9781911307563>

Further details about CC BY licenses are available at <http://creativecommons.org/licenses/>

This book was published with support from the School of Advanced Study, University of London, Aarhus University Research Foundation, and Webster Research and Consulting.

ISBN: 978-1-911307-42-6 (Hbk.)

ISBN: 978-1-911307-55-6 (Pbk.)

ISBN: 978-1-911307-56-3 (PDF)

ISBN: 978-1-911307-58-7 (epub)

ISBN: 978-1-911307-57-0 (mobi)

ISBN: 978-1-911307-59-4 (html)

DOI: <https://doi.org/10.14324/111.9781911307563>

Analysing the UK web domain and exploring 15 years of UK universities on the web

Eric T. Meyer, Taha Yasseri, Scott A. Hale, Josh Cows, Ralph Schroeder and Helen Margetts

Introduction

The World Wide Web is enormous and in constant flux, with more web content lost to time than is currently accessible via the live web. The growing body of archived web material available to researchers is potentially immensely valuable as a record of important aspects of modern society, but there have previously been few tools available to facilitate research using archived web materials (Dougherty and Meyer, 2014). Furthermore, based on the many talks we have given over the years to a variety of audiences, some researchers are not even aware of the existence of web archives or their possible uses. However, with the development of new tools and techniques such as those used in this chapter and others in this volume, the use of web archives to understand the history of the web itself and shed light on broader changes in society is emerging as a promising research area (Dougherty et al., 2010). The web is likely to provide insight into social changes just as other historical artefacts, such as newspapers and books, have done for scholars interested in the pre-digital world. As the web becomes increasingly embedded in all spheres of everyday life and the number of web pages continues to grow, there is a compelling case to be made for examining changes in both the structure and content of the web. However, while interfaces such as the Wayback Machine¹ allow access to individual web pages one at a time, there have been relatively few attempts to work with large collections of web archive data using computational approaches across the corpus.

The research presented in this chapter used hyperlink data extracted from the Jisc UK Web Domain Dataset (Jisc, n.d.-a) covering the period from 1996 to 2010 to undertake a longitudinal analysis of the United Kingdom (UK) national web domain, .uk, focusing on the four largest second level domains: .co.uk, .org.uk, .gov.uk, and .ac.uk. We explore the growth of these domains, and examine the link density within and between them. Next we look in more detail at the academic second-level domain, .ac.uk, to understand the relationship between link density among UK academic institutions and measures of affiliation, status, performance and geographic distance. Overall, these results are used both to understand the growth and structure of the .uk domain, but also to demonstrate the benefits and challenges of this type of analysis more generally.

Background

Archiving national web domains

National web domains represent one approach to web archive analysis for researchers seeking an overview of a single country's web presence (Brügger, 2011). Any particular national web domain offers the potential of both diversity and completeness in its coverage (Baeza-Yates et al., 2007), although there are limitations in terms of generalizability beyond the country in question and frequently in terms of the completeness of the analysis based on technical factors (see section on the UK web domain below). At the same time, limiting the focus to a single country reduces the number of contextual differences (such as multiple dominant languages, different internet and broadband penetration rates, different degrees of political openness and so forth), and thus is a sound strategy for demonstrating the potential of this new type of analysis.

Research in this area is at an early stage, and there are conceptual challenges associated with analysing national web domains. The content and structure of country-code top-level domains (ccTLDs), such as .uk for the UK and .fr for France, are governed more by tradition than rules (Masanès, 2006), complicating efforts to reach a comprehensive definition of what they represent. Brügger (2014) discusses the difficulty, for example, of deciding how national presences should be delimited. In the case presented here, the domain name .uk is used, but this does not cover all the web pages originating in the UK as it is possible for UK companies, organizations and individuals to

use generic top-level domains (.com, .org, etc.) or those assigned elsewhere. Moreover web pages ending with .uk are also used for websites which arguably belong to a different country, as when multinational companies headquartered outside the UK have affiliates within the UK with a .uk address. Finally, it might be contended that not only web pages with a .uk address be examined, but also those that link to and from these web pages. However, for the purposes of this research, these limitations can mostly be noted for future research and do not seriously limit the ability to understand the broad patterns within the UK national web presence. Furthermore, when we focus on UK universities, as we do in the later part of this chapter, we avoid both false positives and false negatives as the academic domain (.ac.uk) is stable and predictable in a way that the commercial domains are not. Essentially, all universities in the United Kingdom have a main address in the .ac.uk domain, and almost all addresses in the .ac.uk domain are universities (with a few exceptions for academic-affiliated organizations that are not themselves universities).²

Another issue that must be decided when undertaking analysis of web domains is the appropriate level of detail. This includes the temporal resolution to use for analysis (since while the web is constantly changing, the number of snapshots available in Internet Archive data vary over time based on the crawl settings in place when the data were gathered). In addition, the level of detail to be extracted from web pages must be determined (i.e. the appropriate level of resolution of page content, link information, page metadata, and so forth). Previous research on the .uk ccTLD has examined monthly snapshots over a one year period, finding that page-level hyperlinks change frequently month to month (Bordino et al., 2008). As Brügger (2013) notes, there are several reasons why archived websites are different from other archived material in respect to these details: choices must be made not just about what to capture but there are also technical issues about what can be archived and how the archiving process itself shapes the later availability of the archived materials.

Previous research using national web archives

While there have been a number of papers describing the practices of constructing national web archives (see for instance Masanès, 2005; Gomes et al., 2006; Baeza-Yates et al., 2007; Žabička and Matjka, 2007; Aubry, 2010; Hockx-Yu, 2011; Rogers et al., 2013), there are few that report using national web archives using large-scale (or even medium-scale) computational methods.

Thelwall and Vaughan (2004) used data from the Internet Archive to assess international bias in the coverage of the archive's collection. At the time of their study, however, it was not possible to access the data in the archive via automated means, so they were limited to relatively small samples of between 94 and 143 websites for each of four countries (total $N = 382$), accessed via the public Wayback Machine interface. They determined with these methods that there was an unbalanced representation of different countries in the archive, partially explained by technical factors rather than by biased policies.

The *Analytical Access to the Domain Dark Archive* (AADDA) project³ and then later the *Big Data: Demonstrating the Value of the UK Web Domain Dataset for Social Science Research* project⁴ and the *Big UK Domain Data for the Arts and Humanities* project⁵ enabled researchers to use UK Web Archive data for analytical study. These projects also demonstrate one of the legal issues of working with web archive data: the UK web archive data held by the British Library can be made available to researchers for use, but full-text content is only available via systems at the British Library. The raw data in the ARC/WARC files cannot be moved outside the Library's computer systems. As a result, many of the demonstrator projects that came out of these bigger projects focused on more qualitative, close analysis (see for instance Gorsky, 2015; Huc-Hepher, 2015) that was *enabled* by computational methods involving search, indexing and ontologies created by the project developers, the actual researchers largely used the extracted results in non-computational ways (see Chapter 11). It is important to note, however, that derivative datasets such as the list of web pages in the archive and the list of hyperlinks can be distributed more widely, which enables some large-scale approaches as we do in this chapter.

Another European project on *Longitudinal Analytics of Web Archive Data*⁶ published a number of technical reports and papers that demonstrate computational approaches to working with web archive data but, as far as we are able to determine, there have not been the same sort of domain investigations as those done using the tools we report here.

The lack of studies using web archives in general, and using large-scale computational approaches in particular, has been documented in earlier work by members of this team (Dougherty et al., 2010; Thomas et al., 2010; Meyer et al., 2011; Dougherty and Meyer, 2014). In those papers and reports, we found that there remains a disconnect between the relatively active community engaged in archiving the web, and the relative lack of any community forming around large-scale analysis of web archives. This study is in part an attempt to fill that very clear gap.

The UK web domain

The .uk country-code top-level domain is managed by the internet registrar Nominet.⁷ Below the .uk top-level domain are several second-level domains (SLDs), the largest of which are .co.uk (commercial enterprises), .org.uk (non-commercial organizations), .gov.uk (government bodies), and .ac.uk (academic establishments).⁸ This chapter examines third-level domain data such as *nominet.org.uk* (Nominet), *fco.gov.uk* (the Foreign and Commonwealth Office of the UK government), or *ox.ac.uk* (the University of Oxford).

In the case of web archives (or indeed of other archived material which takes the approach of archiving all that can be archived, without a particular topic in mind), it is not scholarly interest in any particular topic that has set the data collection agenda. Instead it has been the goal of the archiving institution to accumulate material for the sake of preservation, leaving the question of the eventual uses of the archive data to later researchers. This means that the scope of the archived material and the level of detail available, as with other historical materials, is a function of the archiving processes used to gather and store the data. Thus, unlike web archive research done on the live web using researcher-implemented data collection mechanisms (e.g. Escher et al., 2006; Foot and Schneider, 2006), for the purpose of this study the dataset itself should be seen as a given. However, it can be mentioned that the Internet Archive's data comprise the most comprehensive archive of the web available (Ainsworth et al., 2011).

It is important to note that while the Internet Archive (IA) is the *most* comprehensive archive of the web available, that should not be confused with thinking that the IA crawls represent a *fully* comprehensive record of the web. The data collected over the 15-year period we are examining used a variety of methodologies and were done at varying levels of granularity. Data from the earliest years came from Alexa with 'no visibility into how this data is crawled', and the IA obeys robots.txt restrictions set by site owners (Jisc, n.d.-b), which can result in some websites missing pages or even being excluded completely from the archive (see chapter two by Hale et al.). The time between crawls is variable for any given page, resulting in some pages having more captures over time than others. Furthermore, the Internet Archive does not use the zone file from Nominet, which forms a complete list of all domains within .uk. Instead the Internet Archive relies on discovering websites through hyperlinks and other methods.

Data

Data preparation

The data for this study originally come from the Internet Archive, which began archiving pages from all domains in 1996 (Kahle, 1997). For the .uk domain that will be examined here, the data are sourced from copies of the approximately 30 terabytes of compressed archive data relating to the UK domain (the .uk ccTLD). Archive files were provided to the British Library by the Internet Archive with the specific purpose of creating the basis of a national archive of the web in the UK. These data form the 'Jisc UK Web Domain Dataset' (Jisc, n.d.-a).⁹ The data provided to the research team by the British Library do not include the full text of all the pages crawled due to legal restrictions on use outside the British Library, but do include the link data and other metadata extracted from the full archive.¹⁰

The data were cleaned by removing error pages (e.g. 404 Not Found pages) as well as pages not within the .uk ccTLD. This resulted in a plain-text list of all page Uniform Resource Locators (URLs) remaining in the collection and the date and times they were crawled, and an additional plain-text list of all outgoing hyperlinks starting from pages within the dataset.

For this study, we started with this list of hyperlinks and filtered it to only include links between different third-level domains. We further grouped pages crawled at similar times (within 1,000 seconds) together and assigned the hyperlink pair a weight based on the number of hyperlinks between the two third-level domains in that time period. For each year, if there are multiple crawls within the dataset we take the crawl with the largest number of captured hyperlinks between any two domains. We also formed one list of all third-level domains present in the dataset each year and the number of pages crawled within each third-level domain. These data were loaded into Apache Hive for the analysis that we present here.

Data analysis

In what follows, we undertake a longitudinal network analysis, charting the .uk domain and its core second-level domains over time. As Brügger (2013) points out, this type of analysis is not concerned with who produced what, nor with how the web content was used, but rather with what was created and thus 'the web which is' – or was – 'actually available to users'.

First, we present an overall longitudinal view of the second-level domains within the .uk domain. We investigate the growth of the entire domain between 1996 and 2010, broken down into its four largest constituent parts, .co.uk, .org.uk, .gov.uk, and .ac.uk. Analysis of these SLDs allows us to investigate the role of different sectors of UK society in the growth of the UK web presence.

The second section looks at the link density within and between second-level domains. We examine the internal link density of each SLD, and analyse how they interact with each other: whether, for example, there are more links between certain subdomains, and whether linking is reciprocal between domains or whether it is unbalanced.

The third and final section of the findings takes a closer look at the academic second-level domain .ac.uk. This research builds on earlier longitudinal analyses of academic web pages, which have investigated, for example, the stability of outlinks (Thelwall et al., 2003; Payne and Thelwall, 2007). Our findings update earlier studies by extending the period of analysis to the end of 2010 and assessing the effect of new variables, including institutional affiliation, league table ranking and geographic location on link practices between different universities.

Results

Overview of growth in the .uk web domain

Figure 1.1 displays the overall growth of the .uk ccTLD, showing the total number of nodes (on a logarithmic scale) within each of the four main SLDs we analysed over the period from 1996 to 2010. The insert in the figure shows the size of the entire .uk domain (on a linear scale). There is a clear change in the trend of the growth around 2001 for .co.uk and .org.uk as both domains continue to increase in size, but at a lower speed. Furthermore, .ac.uk and .gov.uk seem to almost stabilize in size at around the same time.

Figure 1.2 shows the relative size of the second-level domains .co.uk, .org.uk, .ac.uk, and .gov.uk across the 15-year period, standardized as each SLD's proportion of the total nodes (i.e. domains/websites, not web pages) in the collection in each year. While these are not the only second-level domains in use within the .uk domain, they are the four largest in terms of number of nodes across the whole period.

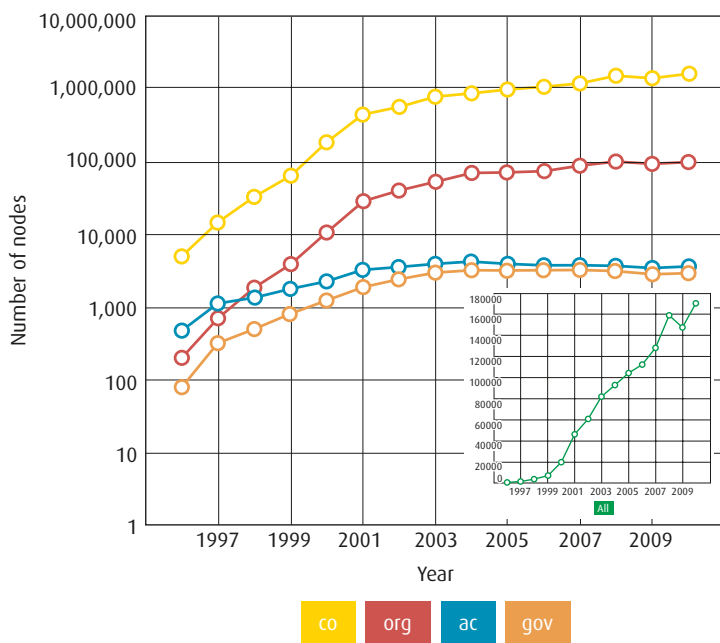


Figure 1.1 Number of nodes (third-level domains) within each second-level domain over time. The inset shows the sum over all second-level domains

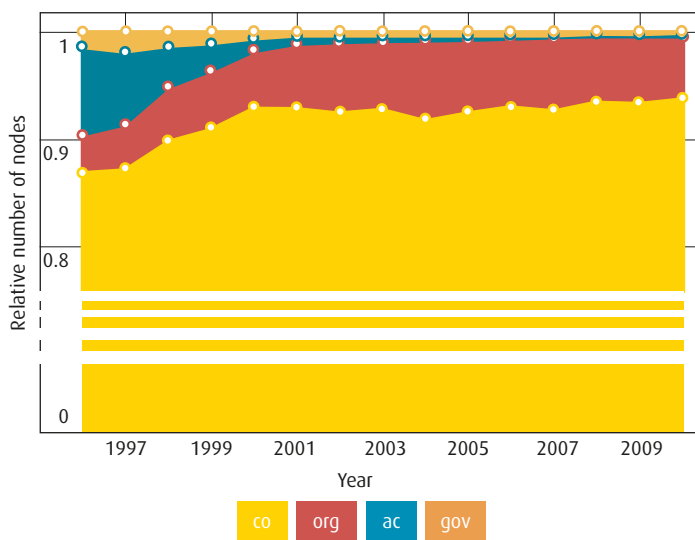


Figure 1.2 Relative size of second-level domains in the .uk top-level domain over time

As Figure 1.2 shows, .co.uk is the predominant second-level domain throughout the entire period, with .co.uk sites never accounting for less than 85% of the total. However, also apparent is the large proportion of governmental and, especially, academic sites in the early recorded history of the UK web. This is consistent with the role that universities played in the early establishment, adoption and development of the web (Leiner et al., 2009). Over time, however, this early presence was greatly overshadowed in terms of absolute numbers of nodes when compared to the continued growth of the .co.uk and .org.uk domains.

Link density within and between second-level domains

Up to this point the analysis has drawn only on node data; that is, the number of websites making up each domain. However, link analysis can offer insight into how well connected each SLD is with itself and with other domains. A link from one site to another has been used as an indicator of awareness between blogs (Hale, 2012) and recognition between academic sites (Thelwall et al., 2003). Figure 1.3 shows, for each subdomain, how many total links there are for every node over time, where a fluctuating relationship between the number of nodes and links to other nodes for each second-level domain is visible. Over the whole period, the .ac.uk academic SLD and, from 1997 onwards, the .gov.uk governmental SLD are the most internally dense SLDs. This observation may reflect the fact that registration for the .ac.uk and .gov.uk subdomains is restricted, whereas .org.uk and .co.uk sites can be registered easily by any party. In addition, the .ac.uk and .gov.uk subdomains are likely constituted by a narrower and more cohesive set of institutions, creating, on average, a stronger basis for linking within the SLDs. Furthermore, there is likely more competition and thus less reason to link within the .co.uk commercial subdomain compared to .ac.uk or .gov.uk. Higher link density within the .org and .gov domains in comparison to the .com domain has previously been observed during a smaller scale, topical study about climate change (Rogers and Marres, 2000).

Also of note is the general rise of links in the middle of the period, particularly in the substantial .co.uk subdomain. This peaks sharply in 2004 before falling sharply back to around pre-2001 levels by 2009. This trend has no easy explanation, suggesting that further research is required to explain this pattern. Possible explanations include that the norm of including lists of links on web pages such as blogs fell out of favour in the middle of this period or that more websites increasingly linked outside of the .uk ccTLD.

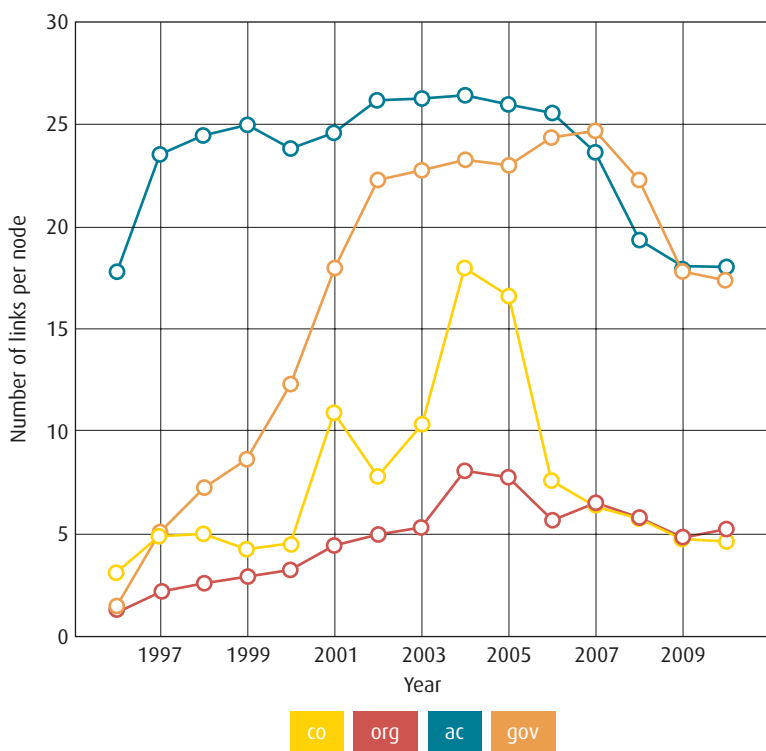


Figure 1.3 Number of within-SLD links per node in four .uk SLDs, 1996–2010

Not only can web domain data tell us how well integrated an SLD is internally, but we can also investigate how well SLDs are connected to each other. Figures 1.4a and 1.4b show the quantity of links between SLDs for 2010, the last year in the dataset, where the size of an arc relates to the volume of links from one SLD to another. The colour of each arc relates to links sent in one direction, from the host SLD outwards. For example, green arcs show links from the .co.uk domain to others. Figure 1.4a shows the absolute volume of links, while the size of the arcs in Figure 1.4b are normalized in relation to the number of nodes in the target subdomain. (Note that Figure 1.4a does not display links within a single SLD, as the volume of links between .co.uk sites dwarfs all other relationships. As Figure 1.4b controls for the number of nodes in each SLD, the adjusted .co.uk arc is much smaller and links within a single SLD are therefore included.)

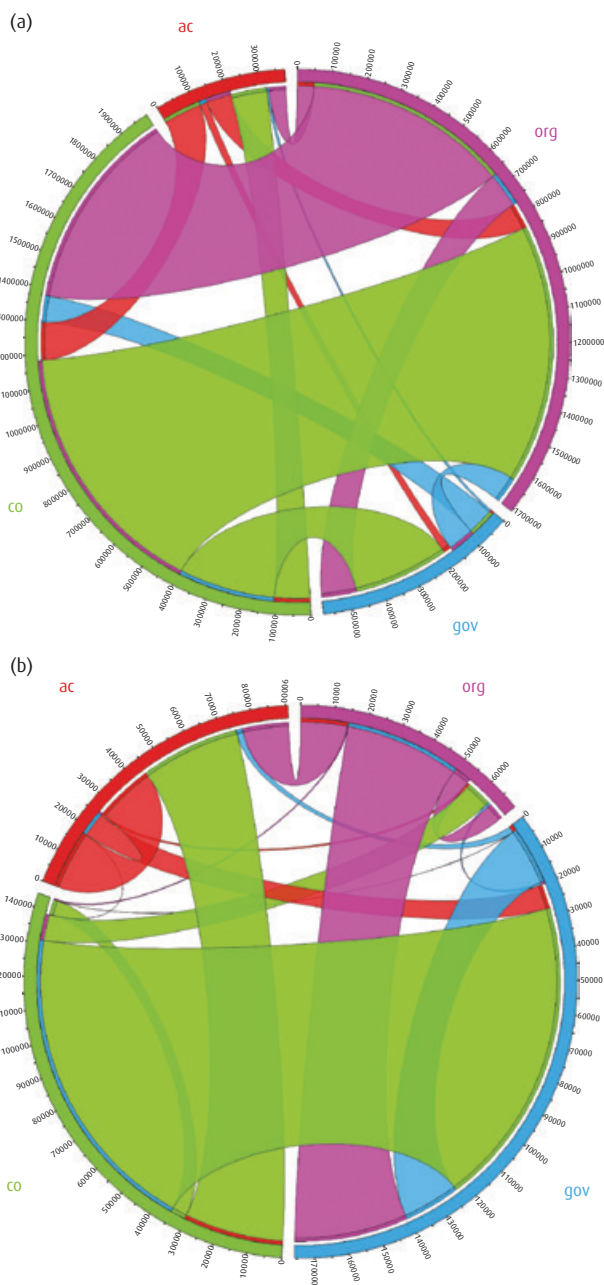


Figure 1.4 Links between four second-level domains. Panel *a* shows the absolute number of links between different SLDs (self-loops are excluded), and panel *b* shows the relative number of links normalized by the size of target subdomain

Figure 1.4a shows that the largest volume of links between SLDs in 2010 flowed from .co.uk sites to .org.uk sites, and this relationship is fairly reciprocal, with .org.uk sites sending almost as many links back. Links between other domains are much lower in terms of absolute volume. When controlling for the size of the target subdomain, however, the picture changes somewhat. As Figure 1.2 showed, by 2010 the number of nodes in the .org.uk subdomain far outweighed those in the .ac.uk and .gov.uk subdomains. Figure 1.4b, adjusting for this, shows that the .gov.uk and, to a lesser extent, the .ac.uk subdomains punch above their weight, receiving proportionally more links from .co.uk and .org.uk sites. Once again, the more restrictive registration policies for these SLDs may be a factor here, driving up the average quality and ‘link-worthiness’ of sites in these subdomains compared to .co.uk and .org.uk sites. However, this discrepancy may also be related to other factors such as the comparative homogeneity of these SLDs, the perception of objectivity or balance on academic or government websites as opposed to sites oriented towards sales or persuasion, or even the international standing of many UK universities, although understanding these factors would require further investigation.

For the .gov.uk subdomain, the finding that sites link out less than they are linked to suggests a lack of ‘outward-lookingness’, compared to the other sectors. In contrast, Escher et al. (2006) found the UK Foreign and Commonwealth Office to be relatively more outward-looking than its equivalents in Australia and the USA. However, foreign offices, given their outward facing role, could easily be an exception to a more general government-wide propensity not to link out.

In addition, it is worth noting the relatively heavy proportion of links within the .ac.uk SLD shown in Figure 1.4b in the red arc that curves from ‘ac’ back into ‘ac’. This propensity of academic institutions to link heavily to other academic institutions (more so than the other domains) reflects (taking a positive view) a strong network among academic institutions, but also potentially (taking a negative view) a tendency towards inward-looking, within-domain links. We examine these links in more depth in the next section.

The UK academic subdomain

At this stage we turn our attention to one particular subdomain, the .ac.uk academic subdomain of the UK web. To be eligible for a third-level domain within .ac.uk, an organization must have a permanent physical presence in the UK and either have the majority of its activities publicly funded by

UK government funding bodies or be a Learned Society. In addition, the organization must satisfy at least one of the following criteria: the organization must provide tertiary-level education with central government funding, conduct publicly funded academic research, have a primary purpose of supporting tertiary-level educational establishments, or have the status of a Learned Society ('a society that exists to promote an academic discipline or group of disciplines').¹¹

The academy was at the forefront of the development of the web, and, as Figure 1.2 shows, .ac.uk sites constituted a sizeable minority of .uk sites in 1996. Over time, this proportion waned, even as more UK universities established a substantial web presence. In this subsection we use the longitudinal data collected to examine the relationship between universities' linking practices and three variables: institutional affiliation, league table ranking and geographic location. Our hypothesis in doing so was that higher status academic institutions would be more strongly linked to than lower status institutions and would also be more strongly interconnected with their peer institutions.

For the analysis, we built a list of the 121 universities listed in the 2014 *Sunday Times* University Guide.¹² Each of these universities has a website, all of which use the .ac.uk suffix. We obtained the third-level domain (e.g. ox.ac.uk) for each. Further data collection as necessary is described in the respective subsections that follow.

Group affiliation

Many UK universities belong to associations, formed to represent their interests and facilitate collaboration. The groups are neither mutually exclusive nor exhaustive, meaning that universities can belong to none, one or more than one group, but for practical and political reasons most universities belong to only one. We collected data on the memberships of five groups, the Russell Group,¹³ the 1994 Group,¹⁴ the University Alliance,¹⁵ the Million+ Group,¹⁶ and the Cathedrals Group.¹⁷

The best known of these is perhaps the Russell Group of research-intensive, highly ranked universities, formed in 1994 and now constituted of 24 members. The 1994 Group, which represented smaller research institutions, was formed in response to the Russell Group, but disbanded in 2013. Given the time frame of the dataset we include the 11 final members of the group in our analysis. Of the remaining three groups, the University Alliance is formed of 22 business-oriented UK universities, the Million+ Group is made up of 17 mostly 'new' (post-1992) institutions, and the Cathedrals Group is made up

of 16 universities originally instituted as church-led teacher training colleges. The stated purposes of these groups differ somewhat, but each are constituted broadly to serve the research and educational interests of their members.

In comparing group membership to the density of links between different universities, we sought to discover whether academic affiliation was associated with the density of links between institutions. To do this, we performed a network analysis, investigating whether the universities clustered on the basis of group affiliation. Figure 1.5 shows a network diagram, with different affiliations marked by different colours.

To the naked eye, Figure 1.5 shows no discernible clustering on the basis of group affiliation, and network analysis bears this out. The division of the network by affiliations has a modularity score (Newman, 2006) of -0.003 , indicating that the division of the network into clusters based on university affiliation is no better than dividing the network into five random clusters. On an individual basis, only one group, the Russell Group, has many internal links and comparatively fewer links to institutions outside the group. It is the most strongly connected group with an internal hyperlink density of 0.71. The Russell Group, which includes 24 of the leading international UK universities with some of the highest levels of research funding, arguably represents most if not all of the elite universities in the UK. It contains nine of the ten top-ranked UK universities, including both Oxford and Cambridge. That these universities are more strongly linked to each other is likely related at least in part to their active research cultures, with many collaborations existing between researchers at these top institutions. The lack of strong web connections in the other associations, however, suggests that while these institutions may or may not have strong connections among their members by other measures, there is no evidence that universities strongly link to the websites of institutions with which they share group affiliation over institutions outside of the group.

League table ranking

University league tables are an important if imperfect indicator of a university's prominence. Modern league tables incorporate a whole range of measures, including factors related to teaching, research and student satisfaction. As such, we investigated whether a university's league table ranking is associated with its web presence, and whether the relationship has changed over time, in terms of both increasing adoption and development of an institution's web presence and its changes in league

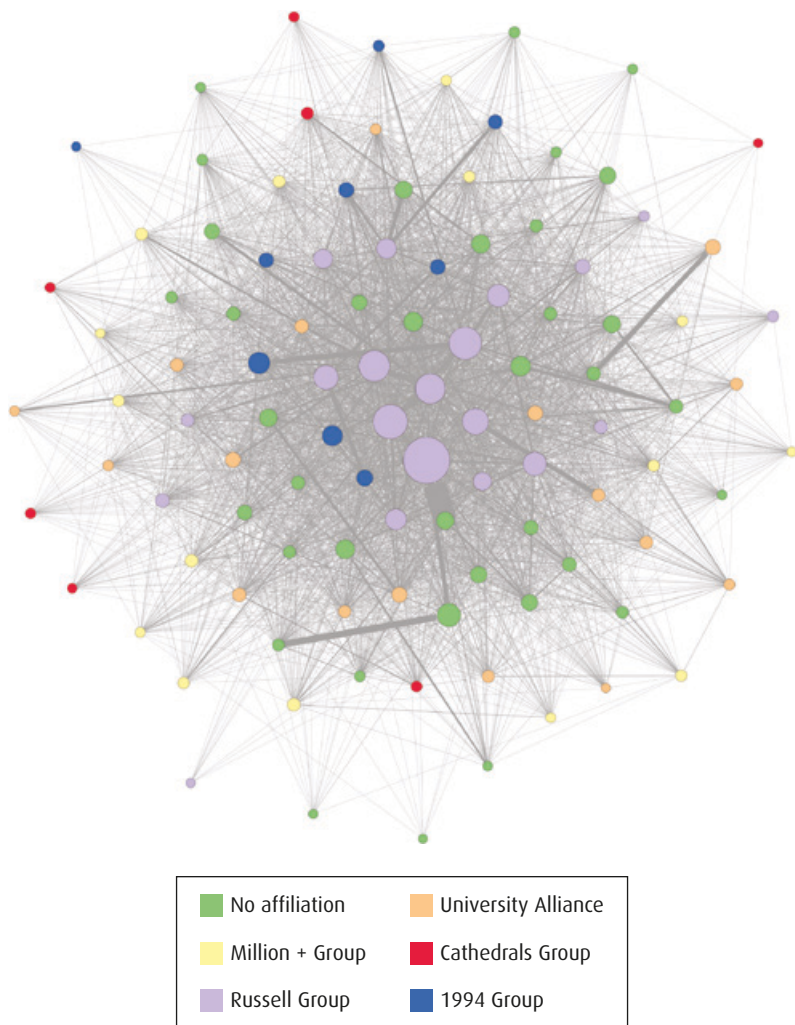


Figure 1.5 Network diagram of hyperlinks between universities. Different colours indicate different university affiliations

table ranking over time. For this analysis, we collected the rankings of UK universities published in *The Times* Good University Guide for three years, 2000, 2005 and 2010, and compared these rankings with data from crawls conducted in the same three years.

In conducting the analysis, we used ten common measures of network centrality for each of the three different years to gauge the relationship between each university's league ranking and its position in the

network of hyperlinks flowing between university third-level domains. We then produced lists ranking the universities for each year by each centrality measure and computed Spearman's rank correlation coefficient for each centrality ranking and league table ranking combination. These correlation coefficients are shown in Figure 1.6.

For most measures of centrality used, a pattern emerges: the data for 2010 show the strongest correlation between league table ranking and centrality, while the relationship is less evident for 2000 and 2005. The most strongly correlated measure is in-strength, a sum of all the hyperlinks linking to a given web domain. This measure uses the weight of each edge, which corresponds to the number of hyperlinks between any two third-level domains. This differs from in-degree which measures the number of other domains that link to a given web domain. Figure 1.7 shows the fairly strong correlation between universities' league table rankings and their network positions as measured by in-strength. What Figure 1.6 and Figure 1.7 suggest is two-fold: first, that a university's prominence, as measured by its league table position, is an increasingly stronger predictor of the number of links to that institution over the 2000–2010 period. Whether this is an example of the Matthew Effect ('the rich get richer') (Merton, 1968) whereby highly prominent institutions become well-linked institutions largely as a result of their prominence (and conversely, marginal institutions become more marginalized as a result of their lack of prominence), or whether there is another independent factor at play here cannot be determined from these data. However, the second conclusion is clear: the hyperlink patterns within the UK academic subdomain support the notion that the web does not inherently challenge existing power structures. Instead, the saturation of the .ac.uk subdomain, in terms of the presence of essentially all possible academic institutions by 2003 (as shown in Figure 1.1), resulted in a subdomain in which network centrality closely mirrors prominence as measured by league tables by 2010.

Role of geography

Finally, we investigated whether any association exists between the geographic proximity of UK universities and the density of hyperlinks between them. This analysis builds upon work by Pan et al. (2012) who found, at a global scale, that rates of academic citations and collaborations between two cities diminish as the distance between them increases, following gravity laws. We conduct a similar analysis,

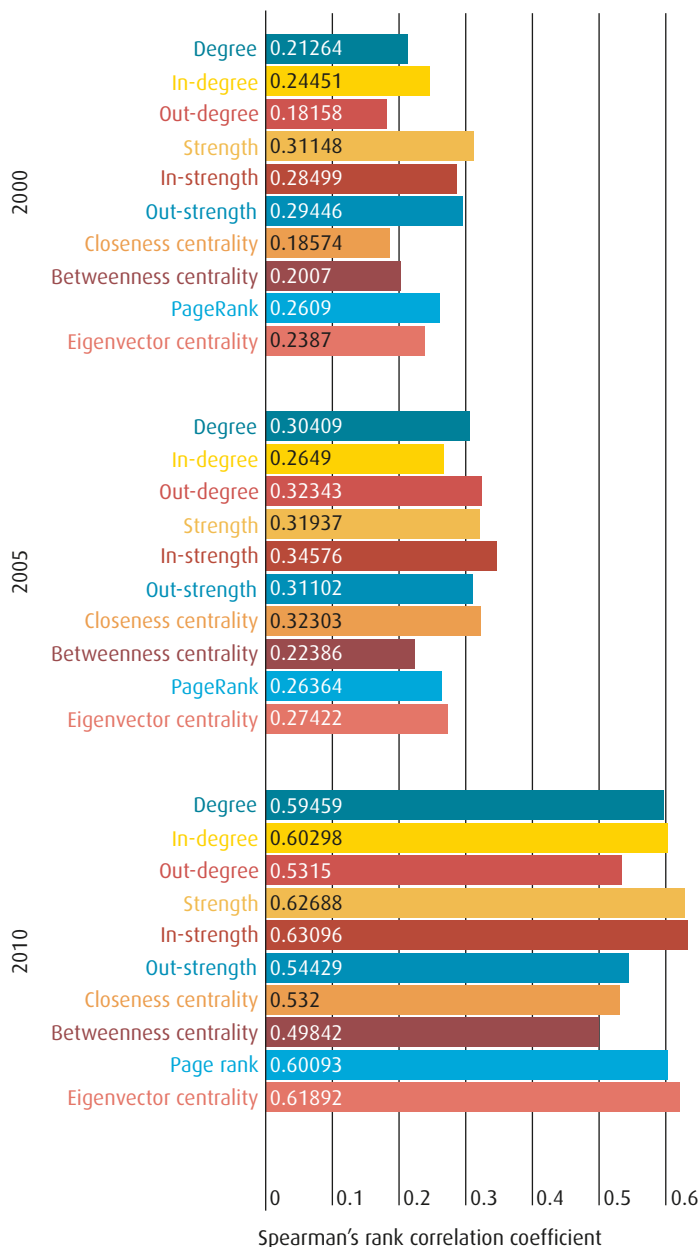


Figure 1.6 Spearman's rank correlation coefficients between university league table rankings and ten different network centrality measures for three years

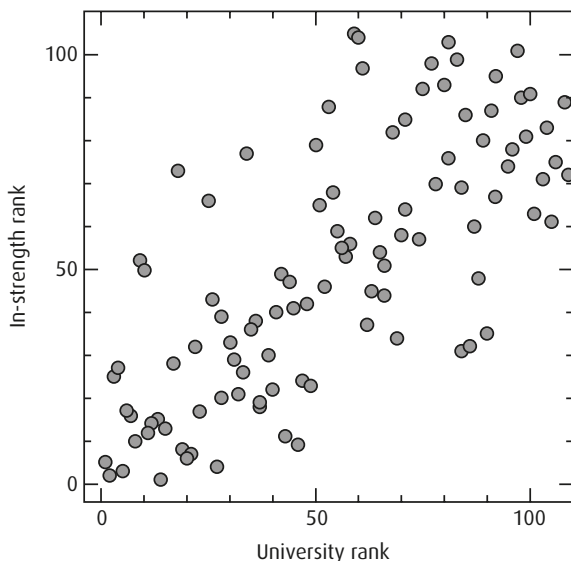


Figure 1.7 University in-strength rankings compared to university league table rankings for 2010. Spearman’s rank correlation is 0.63

replacing citations and collaborations with hyperlinks collected in the web domain data.

We collected geographic coordinates for the UK universities in the list using simple Google Maps searches. Universities can be spatially complex, sometimes having multiple campuses and satellite sites; so, some discretion was occasionally required in identifying the centre of each university.

The standard, naïve gravity law approach would suggest that the number of hyperlinks, or the strength of the connection, between two given universities is inversely proportional to the square of the distance between the two universities. We let S_{ij} denote the strength from university i to university j . Focusing on the data from 2010, the left frame of Figure 1.8 shows that the relationship between this measure and the geographical distance between the two universities is very noisy. To correct for the different sizes of universities and their different linking practices (some universities may just link more than others), we normalize these strengths. We divide S_{ij} by the sum of the weights of all edges coming from university i (S_i^{out}) multiplied by the sum of the weights of all edges linking to university j (S_j^{in}). We denote this normalized measure σ_{ij} and plot it against physical distance in the right frame of Figure 1.8. With this normalization, the relationship between distance and the

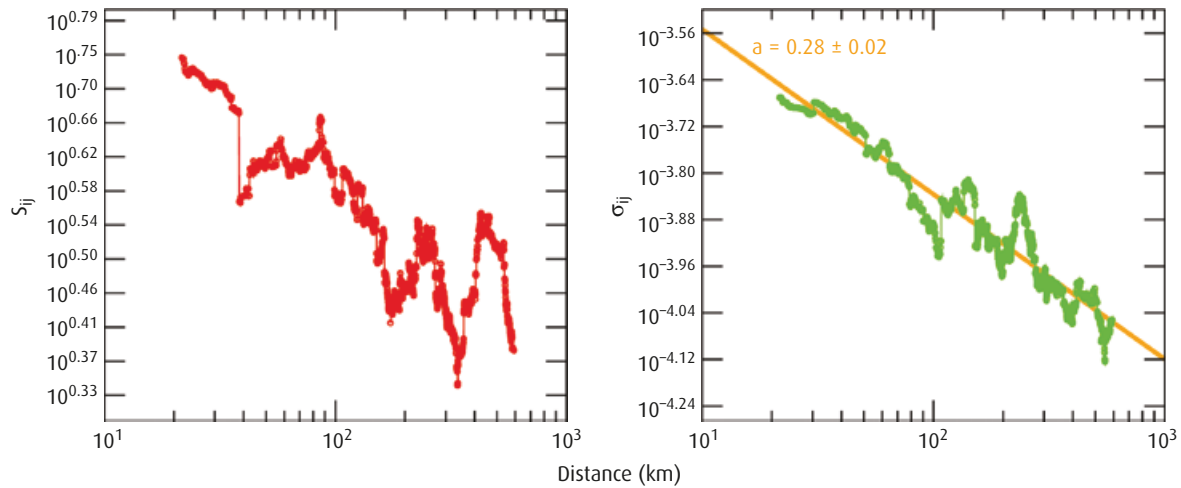


Figure 1.8 Left: Raw hyperlink strength (S_{ij}) between universities versus geographical distance. Right: Normalized hyperlink strength (σ_{ij}) between universities *versus* geographical distance. The normalized measure follows a gravity-law model with an exponent of $a=0.28\pm0.02$

number of hyperlinks (strength) between universities is very clear. In both frames, we use a moving average window with a length of 500 data points and therefore a lower bound of 20km is introduced. An upper bound is induced by considering only the universities within the UK in this study. However, the gravity law holds significantly within a large distance range of 30–600km.

Letting d_{ij} denote the geographical distance between two universities, we then seek the exponent a , which best fits the observed data following $\sigma_{ij} \propto d_{ij}^{-a}$. Using the least squares method, we fit a linear function to the logarithmically transformed data and find $a = 0.28 \pm 0.02$, which closely matches the findings of Pan et al. (2012) for citation and collaboration networks. In that study, Pan et al. found an exponent of $a = 0.30$ for the citation network before any normalization, while finding an even stronger role for geographical distance ($a = 0.77$) after applying a similar normalization to the one we apply here.

Figure 1.9 maps the universities in the sample along with the connections between them coloured according to σ . It is evident, especially in the map of 2010, that the longer connections generally have weaker strength. It is worth noting that the size limit of the dataset and the geographical constraints—such as the dense region of London extended to Oxford and Cambridge, which includes a large number of universities in our dataset – could partially drive the strong geographical dependency we observed. This dense region is particularly visible in the map of 2005 in Figure 1.9.

Conclusion

In this chapter we have reported findings based on longitudinal analysis of the recorded history of the UK web domain from 1996 to 2010. While this analysis is by necessity at a macro-level in terms of detail, it nevertheless demonstrates the potential of these data for detecting changes in patterns in web linking behaviour over time. Such evidence is related to the growth and expansion of the web and uneven patterns of linking within subdomains, such as the academic .ac.uk subdomain discussed in this chapter. We have shown that even though the growth of the commercial side of the web has resulted in increasing commercial dominance of the UK ccTLD in terms of absolute number of nodes, the academic and government subdomains receive proportionally more inlinks per domain. In examining the academic subdomain in particular, we have shown that while there is no generalized

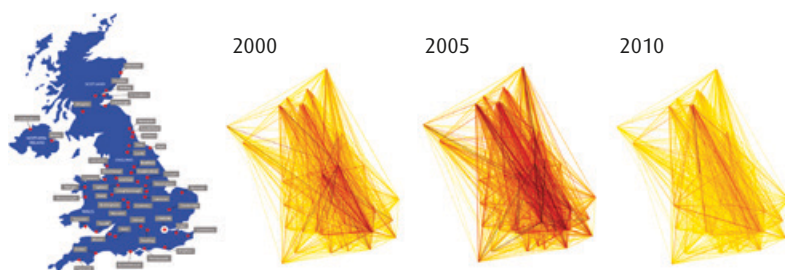


Figure 1.9 Maps of the UK universities under study for three years: 2000, 2005 and 2010. The connections are the hyperlinks and colour corresponds to the normalized strength of each link (σ_{ij}). The reddest links correspond to the strongest connections

clustering based on the affiliation of academic institutions, there are clear patterns in terms of a higher number of inlinks to academic institutions with higher statuses and stronger connections between geographically-closer institutions.

This research has also demonstrated some of the benefits and challenges of this type of analysis. The methods and results described here have allowed us to paint a reliable portrait of the .uk web domain over a period of growth spanning 15 years, which would otherwise be impossible without using web archives (unless a researcher had started collecting similar data themselves over the same time period, which could work going forward, but not retrospectively). We have also shown that it is possible, within the limits of an admittedly incomplete national web archive, to understand certain domains in greater detail, as we have done with the academic portion of the UK web domain.

Challenges, however, remain. Working with these data was neither simple nor quick, and the link data required significant cleaning before they were usable. Also, while the file structure for the link data was very simple, the sheer size of the data necessitated the use of larger processing infrastructure (Apache Hive) that not all researchers have access to or the skills to use. Further, because of legal limitations on the distribution of actual page content, questions that arose over inconsistencies in the link data that might have been easier to understand by looking at the context of the link were more difficult to resolve.

The biggest challenge, however, to using web archives in computational ways remains finding the right questions that are both interesting and capable of being answered within the limits of the web archive data and the extent to which any given web archive contains appropriate coverage over the time period of interest.

This analysis suggests many future possibilities for research with these web archive data, including more detailed micro-level analysis of linking behaviour within various subdomains over time, discovery of networks of collaboration between subunits of institutions, comparison between link measures and other measures of prominence such as citation networks and analysis of other subdomains besides .ac.uk. In addition, there are ongoing efforts to prepare the full-text corpus extracted from the web archive for research (rather than the link corpus used here), which it will be possible to combine with these data to answer more detailed questions about the content of the web, the context for links and discourses on the web.

Acknowledgements

The authors would like to thank Ning Wang for his advice and support on data cleaning for the original project and Andreas Kaltenbrunner for his help with creating the original geographic visualizations. The authors are also grateful for funding from UK Jisc for the ‘Big Data: Demonstrating the Value of the UKWeb Domain Dataset for Social Science Research’ grant (16/11 Enhancing the Sustainability of Digital Collections) that supported the data extraction and early analysis, and further funding for analysis from the UK Arts and Humanities Research Council for the ‘Big UK Domain Data for the Arts and Humanities’ grant (AH/L009854/1). Finally, the authors would like to thank our anonymous reviewers for their helpful comments on both this chapter and the earlier version of this research, which was published in the Proceedings of the 2014 ACM Conference on Web Science (Hale et al., 2014) and is updated here with permission from ACM.

Notes

Introduction

- 1 For a detailed overview of web archives and links to existing web archives, refer to Member Archives, List of Web archiving initiatives, and Truman, 2016. On web archiving, see Brügger, 2005, 2011; Masanès, 2006; Brown, 2006, and the comprehensive bibliography in Ayala, 2013. See also Webster, 2017 for a first attempt to write the cultural history of web archiving initiatives.
- 2 For a more detailed history of the Internet Archive, see Kimpton and Ubois, 2006; Webster, 2017.
- 3 As of 2016, the IIPC has 33 member institutions, see <http://netpreserve.org/resources/member-archives>. Accessed 20 June 2016.
- 4 However, the web page in the Wayback Machine consists of bits and pieces that may have entered the archive at different points in time, thus rendering a web page that may not have looked exactly the same when it was online (cf. the reflections on Memento in Jane Winters' Coda in this volume).
- 5 It is also worth noting that the archived web that is not found in a dedicated web archive often has to be downloaded to the users' own computer as a number of individual files and with no built-in interface, which is, for example, the case with The Archive Team GeoCities Snapshot.
- 6 A rare exception being the Danish case, as described by Webster, 2017; as stated in Danish law, the Netarkivet shall have a standing editorial committee, including researchers, and appointed by the Ministry of Culture. In addition, two early examples exist of researchers being involved in the creation of special collections, namely the Dutch project Archipol (2000), and webarchivist.org (2001), see Brügger, 2011: 32, 40.
- 7 Previous collaborations include: (1) Research projects such as the UK based 'Big Data: Demonstrating the Value of the UK Web Domain Dataset for Social Science Research' (2012–14), Analytical Access to the Domain Dark Archive (AADDA), Big UK Domain Data for the Arts and Humanities (BUDDAH, 2013–14), and 'Born digital big data and methods for history and the humanities' (2016–17), the Danish 'Probing a Nation's Web Domain' (2016), the Dutch 'WebART: Web Archive Retrieval Tools' (2012–14), the French 'Web90, Patrimoine, Mémoires et Histoire du Web des années 1990' (2014–16) and 'De #jesu-ischarlie à #offenturen: archives et archivage du patrimoine nativement numérique face aux attentats' (2016), and the Canadian 'A Longitudinal Analysis of the Canadian World Wide Web as a Historical Resource' (2015–16); (2) Training and networks such as the research infrastructure project NetLab in Denmark, the French workshops 'Atelier DL web Ina', the Canadian/American 'Archives Unleashed: Web Archives Hackathon' (2016), and the network 'Working with Internet Archives for Research' (WIRE, US, Rutgers University, 2014–15); (3) Conferences such as the IIPC's annual General Assembly, which for a number of years has provided an invaluable venue for collaborations, 'Web Archives as Scholarly Sources: Issues, Practices and Perspectives' (Aarhus, 2015), and 'Time(s) and temporalities of the Web' (Paris, 2015).

Chapter 1

- 1 <http://web.archive.org/>
- 2 Examples include jobs.ac.uk, which is an academic job listing service operated by University of Warwick; bl.ac.uk, which is the British Library; and funders such as Jisc (jisc.ac.uk), the Wellcome Trust (wellcome.ac.uk), and the Economic and Social Research Council (ESRC, esrc.ac.uk), among others.
- 3 <http://domaindarkarchive.blogspot.co.uk/>
- 4 <http://www.oii.ox.ac.uk/research/projects/?id=88>
- 5 <http://buddah.projects.history.ac.uk/>
- 6 <http://www.lawa-project.eu/>
- 7 <http://www.nominet.org.uk/>
- 8 <http://www.nominet.org.uk/uk-domain-names/about-domain-names/uk-domain-subdomains/second-level-domains>
- 9 <http://data.webarchive.org.uk/opendata/ukwa.ds.2/>
- 10 The specific data we begin with in this project are the Web Archive Transform (WAT) files generated from the full dataset ([https://webarchive.jira.com/wiki/display/Iresearch/Web+Archive+Transformation+\(WAT\)+Specification,+Utilities,+and+Usage+Overview](https://webarchive.jira.com/wiki/display/Iresearch/Web+Archive+Transformation+(WAT)+Specification,+Utilities,+and+Usage+Overview)).
- 11 <https://community.ja.net/library/janet-services-documentation/eligibility-guidelines>
- 12 http://www.thesundaytimes.co.uk/sto/University_Guide/
- 13 <http://www.russellgroup.ac.uk/our-universities/>
- 14 <http://www.timeshighereducation.co.uk/news/was-1994-groups-demise-triggered-by-relaunch-delays/2008999.article>
- 15 <http://www.unialliance.ac.uk/member/>
- 16 <http://www.millionplus.ac.uk/who-we-are/members>
- 17 <http://cathedralsgroup.org.uk/Members.aspx>

Chapter 2

- 1 This paper is an updated version of Ainsworth, AlSum, SalahEldeen, Weigle and Nelson (2011).
- 2 TripAdvisor operates a number of domain names (e.g. tripadvisor.com, tripadvisor.es, etc.) in over 30 countries; however, most of the content about specific attractions on these sites is the same.
- 3 In addition to the JISC UK Domain Dataset comprised entirely of Internet Archive data, the British Library has also independently collected web content related to the UK. Prior to 2014, the British Library manually selected important UK websites and crawled the websites whose owners could be contacted and gave permission to be included in the BL Web Archive. In 2014, the British Library started running more complete crawls of the .uk domain, completely separate from the Internet Archive. We do not use any data that the British Library crawled itself as the selective crawls did not include TripAdvisor and the 2014 crawl was not available at the time we extracted our data.
- 4 We use a technique called kernel density estimation with a Gaussian kernel to estimate the distributions of the two datasets. We also use a standard hypothesis-testing technique, a one-sample *t*-test, to compare the mean of a sample to a known population mean in order to assess the probability that the sample (the archived data) was drawn from the population (the live data).

Chapter 3

- 1 Cf. also the overview of more technical studies in Brügger (2016).
- 2 The project was initiated in 2014 by NetLab, a unit within the national Danish research infrastructure project Digital Humanities Lab Denmark (DIGHUMLAB), and conducted in close collaboration with the national Danish web archive Netarkivet, and funded by the Danish Ministry of Culture (grant recipient: the State and University Library) and by the Danish e-Infrastructure Cooperation (DeIC).

- 3 When delimiting a corpus in a web archive, a number of issues have to be taken into consideration (cf. Brügger, 2016). For comments on web corpus building on the online web in corpus linguistics, see Hundt et al. (2007).
- 4 In some web archives this type of material has been already identified when archiving, which can be a great help. However, this identification always mirrors the archiving policy, the available resources, etc. at any given time in the web archive's history (cf. Zierau, 2015).
- 5 For more information, see the newsletters published by Netarkivet (n.d.).
- 6 For this reason, we are very thankful for the assistance of one of Netarkivet's IT-developers, Per Møldrup-Dalum.
- 7 R is a programming language and a software environment that can be used for statistical computing and for graphics (<https://www.r-project.org/about.html>).
- 8 E.g. the question of how many domains from 2005 had disappeared by 2009 can be asked like this: `domains %>% filter(y2005 == TRUE, y2009 == FALSE) %>% count()`
- 9 As the lists are protected by national privacy acts, we cannot provide names, distinguishing features or the like.
- 10 <https://web.archive.org/web/20050308155332/http://www.dk-hostmaster.dk/dkhostcms/bs?pageid=101&action=cmsview&language=da> (last accessed 20 October 2016).
- 11 We are very grateful for help from Vinay Goel, Jefferson Bailey and John Lekashman from the Internet Archive.
- 12 Many individuals, of course, also have their own website.

Chapter 4

- 1 Web 2.0 is commonly defined as the 'network of interconnected devices and applications that enable the production, consumption and remixing of technologies at both the individual and group level, ultimately leading to an architecture of participation' (O'Reilly, 2005).
- 2 The paywall model refers to the decision by a website to place all or a portion of its content behind a login page; users are then required to pay for an account in order to access the content (Chiou and Tucker, 2013).
- 3 See <http://web.archive.org/web/20030603182856/http://www.whitehouse.gov/news/releases/2003/05/iraq/20030501-15.html>. Accessed 12 October 2016.
- 4 See <http://web.archive.org/web/20031001200908/http://www.whitehouse.gov/news/releases/2003/05/iraq/20030501-15.html>. Accessed 12 October 2016.

Chapter 5

- 1 <http://www.bbc.co.uk/mediacentre/worldnews/bbc-world-news-web-figures.html> (Accessed 16 September 2016).
- 2 <http://www.alexa.com/topsites/countries/GB> (Accessed 16 September 2016).
- 3 <https://web.archive.org/web/20051231123944/http://news.bbc.co.uk/1/hi/help/3676692.stm> (Accessed 16 September 2016).
- 4 <http://data.webarchive.org.uk/opendata/ukwa.ds.2/> (Accessed 16 September 2016).
- 5 We do not count outlinks from the same source and destination page more than once per day.
- 6 We removed the domains 'tv', 'us', 'fm', 'io' and 'me' for this reason.
- 7 Our data covers a broad timespan and yet the variables collected are largely measured at the year level. All of the variables collected were the values for 2005, a year near the midpoint of our timespan, unless otherwise specified.
- 8 <http://data.worldbank.org/indicator/SP.POP.TOTL?page=1> (Accessed 19 November 2014).
- 9 <https://www.uktradeinfo.com/Statistics/BuildYourOwnTables/Pages/Table.aspx>. Total combined trade represents total value of imports plus exports. (Referred to as 'dispatches' and arrivals' for European Community countries.) (Accessed 19 November 2014).

- 10 <http://data.worldbank.org/indicator/NY.GDP.PCAP.CD/countries?page=1>. GDP per capita is gross domestic product divided by midyear population. (Accessed 19 November 2014).
- 11 This used Google Map data, and, for countries for which this was unobtainable, information from <http://www.timeanddate.com/worldclock/distances.html?n=136>. (Accessed 19 November 2014).
- 12 http://worldriskreport.entwicklung-hilft.de/uploads/media/WorldRiskReport_2013_online_01.pdf This data was collected for the latest available year. (Accessed 19 November 2014).
- 13 http://static.visionofhumanity.org/sites/default/files/Global%20Peace%20Index%20Report%202015_0.pdf This data was collected for the latest available year. (Accessed 19 November 2014).
- 14 http://www.unodc.org/documents/gsh/pdfs/2014_GLOBAL_HOMICIDE_BOOK_web.pdf This data was collected for the latest available year. (Accessed 19 November 2014).
- 15 The variable is incremented by 1 before transformation to preserve the small number of observations which have 0 mentions. The fit of the model was investigated with residual plots and lack-of-fit tests for all variables and the model itself. Following several other transformations, which are noted in Table 5.2, these tests gave no cause for concern. Due to some missing data in the independent variables, the N for this regression is 148.
- 16 This model was again investigated with residual plots and lack-of-fit tests. These tests showed some evidence that in fact the model was on the borderline of acceptability in terms of fitting well. Further investigation revealed that this was due to several outliers in the dataset, hence a robust regression was also fitted. This regression, however, provided identical results to simple OLS regression (in terms of statistical significance and direction of effect), hence we preserve this simple regression here for the sake of consistency.

Chapter 6

- 1 In 1999 he was the co-founder, with Valentin Lacambre, of Gandi, a small company selling domain names at much cheaper rates than Network Solution Inc.
- 2 On web archives, useful information can be found in Brügger, 2009, 2012a, 2012b, 2012c; Dougherty et al., 2010; Mussou, 2012; Ben-David and Huurdeman, 2014.
- 3 Adding to the confusion is the evolution of some domain names: france.diplomatie.fr, appearing in 1997, later became diplomatie.gouv.fr. Cross-referencing sources is necessary to avoid the mistake of thinking that the website diplomatie.gouv.fr, which the Wayback Machine references only since 15 July 2005, exists only since that date. The *Guide du Routard de l'Internet* (collective, 1998) mentions france.diplomatie.fr, and the Wayback Machine confirms the change of name when offering a similar answer to a reply using both URLs: <https://web.archive.org/web/20051015220307/http://www.france.diplomatie.fr/fr/> and <https://web.archive.org/web/20051015183638/http://www.diplomatie.gouv.fr/fr/>
- 4 The hôtel Matignon is the official residence of the French Prime Minister.
- 5 The address indicates hosting by the *École nationale supérieure des télécommunications* (ENST).
- 6 Launched in France in 1993, Compuserve allowed internet access while relying on a proprietary system, with forums and services dedicated to its users.
- 7 Isabelle Falque-Pierrotin is the author of the report *Internet. Enjeux juridiques (Internet, Legal issues)* in 1997.
- 8 See Versailles Court of Appeals, 12th chamber, Judgement of 13 September 2007, Semi-public company Issy Média et al. v. Mohamed E., Issy on Line.
 Mohamed E. registered the domain names Issy.net, Issytv.com, Issytv.org and the trademark Issy TV on 13 January 2004. He was the founder and president of the association Issy on Line. The city of Issy-les-Moulineaux and the company Issy Média deemed that the use by Mohamed E. and the association Issy on Line under their trademark and domain names of Internet sites was likely to create confusion with their name. http://www.legalis.net/spip.php?page=jurisprudence-decision&id_article=2049, last accessed on 25 July 2015.
- 9 Ranging from a forum on the Strasbourg Board of Education website allowing parents to communicate with schoolchildren on a trip to the seaside, to the Calvados *Direction départementale* offering real time information on the processing of building permits, to the variety of events presented by the Ministry of Culture.

- 10 Quantity is not sufficient: one will encounter websites of a prefecture or a decentralized service boasting 1,500 or even 2,500 pages, or the site of a ministry claiming 21,000 pages, but this does not allow the citizen or the user to be satisfied: contents accessible to the initiated only, ill-adapted navigation, stale information, all shatter the effects of such ambitious endeavors (DIRE, 2001).

Chapter 7

- 1 While there is a global story to be told of GeoCities, for reasons of feasibility I am largely constraining myself to North American conclusions: drawing on North American media reactions, for example, and the literature that emerged around it there.
- 2 The focus of this chapter rests on the substantive findings from the GeoCities archive, rather than method. Our analysis was generated in part through the warbase platform, a web archiving analytics platform led by Jimmy Lin (University of Waterloo) available at <http://warbase.org>.
- 3 A later option would allow people to purchase ‘vanity’ addresses, such as <http://geocities.com/~janesmith>.
- 4 The basic HTML editor is discussed extensively in Sawyer and Greely, 1999. We know less about the GeoCities experience of 1996 than we do about its subsequent 1998 evolution, as the Internet Archive could not preserve the dynamic content of the web form. We have snapshots of individual pages, as well as user reflections on how easy the basic editor was. In any case, it is clear that a user without technical expertise could create a simple template-driven website with personalized textual content quite easily.

Chapter 8

- 1 RU486 is the common name for the abortion drugs Mifepristone and Misoprostol.
- 2 The Pharmaceutical Benefits Schemes makes pharmaceutical products available at subsidised prices.
- 3 A typical Web 1.0 website provides content (often reflecting organizational goals, background, services, etc.) that does not change regularly and does not allow a lot of interactivity.
- 4 A web crawler is software that automatically traverses a web site, in a manner similar to the way a human user enters the homepage of a website, and then clicks internal links to visit other parts of the website. The crawler can be designed to collect and store text content and hyperlinks (both internal and external) from each page it visits.
- 5 According to Experian Hitwise, Google Australia was the top-ranked website in January 2016 with a 11.2% share of traffic, and no other search engine is in the top-10 (source: <http://www.experian.com.au/hitwise/online-trends.html>, accessed 27 January 2016).
- 6 These search results would most likely have been affected by Google search customizations associated with the location (based on IP address) of the computer which was used for the search (there are national, and potentially even sub-national differences in search results). There is also a chance the browse and search history of the computer used for conducting the search could have impacted on the search results. We note these potential biases in the search results, but it was beyond the scope of this chapter to investigate their magnitude and significance.
- 7 We did not take this step here because our initial Google search was fairly extensive and we expect that including additional sites into the analysis is unlikely to qualitatively impact on the research findings.
- 8 We acknowledge that the use of hostnames is a somewhat rudimentary way of representing websites (and indeed groups or organizations). For example, it could be that a single organization has more than one subdomain (e.g. subdomain1.website.com and subdomain2.website.com) and both of these hostnames would be present in the dataset. Another problem is that different organizations could share a hostname (e.g. that of a commercial web hosting company), and these different organizations would then effectively be merged into a single data point. Casual inspection of our data lead us to conclude that this is not a major problem, in that it would not impact qualitatively on our results.

- 9 For more on social network analysis see, for example, Wasserman and Faust (2004) and Hanneman and Riddle (2005).
- 10 The reader may wonder why, in Table 8.6, facebook.com and youtube.com are classified as 'neutral' while twitter.com is classified as 'unknown'. The reason is that pages from Facebook and YouTube appeared in the 2015 Google searches, and these websites were classified as 'neutral' since the companies hosting the sites are not participants in the abortion debate. In contrast, Twitter did not feature in the Google search results (but it was picked up from the web crawl), and hence it was not classified.
- 11 We also used a word 'stop list' to ensure that commonly used words (e.g. 'and', 'but', 'the') were not included in the analysis.
- 12 The visualizations were created using the tm and wordcloud packages in the R statistical software.

Chapter 9

- 1 The data also includes a limited amount of data from non .uk hosts, being only those resources necessary to render the main series.
- 2 The results of a crawl of the open web are influenced by its starting point(s): that is, the list of URLs with which the crawl begins (seed URLs).
- 3 The interface itself is available at <http://webarchive.org.uk/shine>; the codebase may be found at <https://github.com/ukwa/shine/>
- 4 The Host Link Graph has the DOI <http://dx.doi.org/10.5259/ukwa.ds.2/host.linkage/1>
- 5 2005: 59.5 million; 2006: 53.1 million; 2007: 92.0 million; 2008: 32.4 million. (UK Web Archive, 2015a).
- 6 The Host Link Graph shows 2008 as the first year in which *bnp.org.uk* linked to the archbishop's domain. It is likely that this resource was the one containing that first link.
- 7 The Internet Archive's capture of the page does not include the video content, which is however still available on the live web at <http://bnptv.org.uk/2008/07/christian-doctrine-is-offensive-to-muslims/> (retrieved 15 September 2015).

Chapter 10

- 1 <http://www.mideastyouth.com/2009/09/22/althawra/> Accessed 14 September 2015.

Chapter 12

- 1 Papers dealing with web archives have been accepted for the first time at the ADHO's DH2016 conference, Kraków.
- 2 The still necessary focus on the developed world has been highlighted in the introduction to this volume. No doubt this emphasis will change in the coming decades.
- 3 Referring to 'the archived web' rather than 'web archives' has proven to be useful in distinguishing, for historians at least, between digitized historical material online and the archive(s) of the web itself. Some flexibility about terminology for different audiences is perhaps inevitable in an emerging field.
- 4 The closure in May 2016, after only two months, of a new print-only newspaper in the UK, the *New Day*, is illustrative of this general trend (although, of course, there were other factors at work too) (Sweney, 2016).
- 5 Big UK Domain Data for the Arts and Humanities was funded by the Arts and Humanities Research Council (AHRC) as part of its Digital Transformations in the Arts and Humanities theme (grant reference AH/L009854/1).
- 6 The formulation 'n.d. [no date]' is, however, not uncommon when dealing with early printed books and in some ways analogous to the difficulty of establishing a date of publication for an archived web page. I am grateful to Jonathan Blaney for suggesting this comparison.

- 7 The editor of a critical or scholarly edition aims to produce a 'best' text by comparing various extant versions (witnesses), usually choosing what they deem to be the most authoritative variant as the copy, or base, text. In the case of web archives, an algorithm rather than a human editor is producing the 'best' version of the web page. The role of the researcher is then to recognize the temporal incoherence and assess its significance, and this can only happen if archiving institutions make their processes transparent.
- 8 The *Oxford English Dictionary* defines diplomatic as 'The science of diplomas, or of ancient writings, literary and public documents, letters, decrees, charters, codicils, etc., which has for its object to decipher old writings, to ascertain their authenticity, their date, signatures, etc.' Perhaps more helpful is the definition given in Wikipedia: 'a scholarly discipline centred on the critical analysis of documents [...] It focuses on the conventions, protocols and formulae that have been used by document creators, and uses these to increase understanding of the processes of document creation, of information transmission, and of the relationships between the facts which the documents purport to record and reality'.
- 9 The Digging into Data Challenge, which has run periodically since 2009, perfectly captures in its title this aspect of humanities research.
- 10 *OED*: 'In the United Kingdom (originally the south of England): a young person of a type characterized by brash and loutish behaviour and the wearing of designer-style clothes (esp. sportswear); usually with connotations of a low social status'.
- 11 *OED*: 'a severe reduction in lending by banks and other financial institutions, typically as a result of widespread (or anticipated) defaulting on loans, mortgages, etc.; (also) a period characterized by this'.
- 12 The potential of web archives for linguistic research is clear from a resource such as the Corpus of Global Web-based English (GloWbE).
- 13 The ongoing importance of the Text Encoding Initiative consortium is one example of this.
- 14 I am very grateful to Jonathan Blaney for this suggestion. Interestingly, in Denmark steps have been taken to at least partially overcome this problem. For example, as pointed out by the editors of this volume, Netarkivet is allowed to archive password-protected data if the option to acquire a password is publicly available (either freely or in exchange for payment). Agreements are also in place with some of the larger commercial websites to allow access based on IP address.
- 15 Apps have not replaced websites to the extent that might have been expected when smartphones began to become ubiquitous (see, for example, Newton, n.d.). I owe this reference to Jonathan Blaney.
- 16 In the UK, online petitions opened at the official parliament website which gain 10,000 signatures will receive a formal government response, while 100,000 signatures are sufficient to trigger a debate in parliament.
- 17 Andy Jackson at the British Library has conducted some fascinating research on this in relation to sites added to the open UK Web Archive between 2004 and 2014. The finding that stands out is that '50% of resources [are] unrecognisable or gone after 1 year' (Jackson, 2015).
- 18 The British Library, the National Library of Wales, the National Library of Scotland, the Library of Trinity College, Dublin, the Bodleian Libraries at the University of Oxford and Cambridge University Library.
- 19 This is the case in Denmark, for example, where Netarkivet 'cannot be accessed by the general public. The archive is only accessible to researchers who have requested and been granted special permission to use the collection for specific research purposes'. However, in contrast to provision at the British Library, once researchers have been granted access they may conduct their research remotely and there are no limits placed on the number of concurrent users.
- 20 For example, the Research Infrastructure for the Study of Archived Web materials, led by Niels Brügger; and the Born Digital Big Data and Approaches for History and the Humanities network, of which I am the Principal Investigator (grant reference AH/N006178/1).

References

Introduction

- Abbate, J. (2000). *Inventing the Internet*. Cambridge, MA: MIT Press.
- Aspray, W. and Hayes, B. (eds) (2011). *Everyday Information: the Evolution of Information Seeking in America*. Cambridge MA: MIT Press.
- Ayala, B. R. (2013). Web Archiving Bibliography 2013. UNT Digital Library. <http://digital.library.unt.edu/ark:/67531/metadc172362/>. Accessed 5 June 2016.
- Banks, M. A. (2008). *On the Way to the Web: The Secret History of the Internet and its Founders*. New York: Apress.
- Boldi, P., Codenotti, B., Santini, M., and Vigna, S. (2002). Structural properties of the African web. The Eleventh International WWW Conference, <http://vigna.di.unimi.it/ftp/papers/www2002b/poster.pdf> Accessed 20 June 2016.
- Brown, A. (2006). *Archiving Websites. A Practical Guide for Information Management Professionals*. London: Facet Publishing.
- Brügger, N. (2005). *Archiving Websites: General Considerations and Strategies*. Aarhus: Center for Internet Studies.
- Brügger, N. (ed.) (2010). *Web History*. New York: Peter Lang.
- Brügger, N. (2011). Web archiving – between past, present, and future. In M. Consalvo and C. Ess (eds), *The Handbook of Internet Studies*. Oxford: Wiley-Blackwell, 24–42.
- Brügger, Niels (2012a). Web history and the web as a historical source. *Zeithistorische Forschungen* 9(2): 316–25.
- Brügger, N. (2012b). When the present web is later the past: Web historiography, digital history and internet studies. *Historical Social Research* 37(4): 102–17.
- Brügger, Niels (2014). Web Archives and Big Data. Paper accepted for the 2nd Workshop on Big Humanities Data, Washington, DC.
- Burns, M. and Brügger, N. (2012). *Histories of Public Service Broadcasters on the Web*. New York: Peter Lang.
- Cohen, D. J. and Rosenzweig, R. (2006). *Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web*. Philadelphia: University of Pennsylvania Press.
- Cowls, J. (2013). Digital Deletion is incompatible with democracy <https://joshcowls.com/2013/11/15/fahrenheit-401-digital-deletion-is-incompatible-with-democracy/>. Accessed 20 June 2016.
- Dougherty, M., Meyer, E. T., Madsen, C., van den Heuvel, C., Thomas, A., and Wyatt, S. (2010). *Researcher Engagement with Web Archives: State of the Art*. London: JISC.
- Gillies, J. and R. Cailliau (2000). *How the Web was Born: The Story of the World Wide Web*. Oxford: Oxford University Press.
- Goggin, G. and McLelland, M. (eds) (2017). *The Routledge Companion to Global Internet Histories*. New York: Routledge.
- Graham, S., Milligan, I., and Weingart, S. (2015). *Exploring Big Historical Data: The Historian's Macroscope*. London: Imperial College Press.

- Guardian (2013). Conservative party deletes archive of speeches from internet. <http://www.theguardian.com/politics/2013/nov/13/conservative-party-archive-speeches-internet>. Accessed 20 June 2016.
- Guldi, J. and Armitage, D. (2014). *The History Manifesto*. Cambridge: Cambridge University Press.
- Kimpton, M. and Ubois, J. (2006). Year-by-year: From an archive of the internet to an archive on the Internet. In J. Masanès (ed.), *Web Archiving*. Berlin: Springer, 201–12.
- Koerbin, P. (2017). Revisiting the world wide web as artefact: Case studies in archiving small data for the National Library of Australia's PANDORA Archive. In N. Brügger (ed.), *Web 25: Histories from the first 25 Years of the World Wide Web*. New York: Peter Lang.
- Lindley, S. E., Marshall, C. C., Banks, R., Sellen, A., and Regan, T. (2013). Rethinking the web as a personal archive. In Proceedings of the 22nd International Conference on World Wide Web, pp. 749–760.
- List of Web archiving initiatives (n.d.), https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives. Accessed 20 June 2016.
- Masanès, J. (ed.) (2006). *Web Archiving*. Berlin: Springer.
- Member archives (n.d.), <http://netpreserve.org/resources/member-archives>. Accessed 20 June 2016.
- Meyer, E. T. and Schroeder, R. (2015). *Knowledge Machines: Digital Transformations of the sciences and humanities*. Cambridge MA: MIT Press.
- Naughton, J. (2015). *A Brief History of the Future: The Origins of the Internet*. London: Weidenfeld and Nicolson.
- Naughton, J. (2012). *From Gutenberg to Zuckerberg: What You Really Need to Know About the Internet*. London: Quercus.
- New York Times (2014). What Happened to Malaysia Airlines Flight 17. <http://www.nytimes.com/interactive/2014/07/18/world/europe/malaysia-airlines-flight-mh17-q-a.html>. Accessed 20 June 2016.
- Poole, H. W. (ed.) (2005). *The Internet. A Historical Encyclopedia*. Santa Barbara, CA: ABC/Clio.
- Rieh, S. Y. (2004). On the Web at home: Information seeking and web searching in the home environment. *Journal of the American Society for Information Science and Technology* 55(8): 743–53.
- Rosenzweig, R. (2004). How will the net's history be written? Historians and the internet. In P. Nissenbaum, H. and M. E. Price (eds), *Academy & the Internet*. New York: Peter Lang, 1–34.
- Salter, A. and Murray, J. (2014). *Flash: Building the Interactive Web*. Cambridge, MA: MIT Press.
- Savolainen, R. (2008). *Everyday Information Practices: A Social Phenomenological Perspective*. Lanham, MD: Scarecrow Press.
- Schneider, S. M. and Foot, K. A. (2006). *Web Campaigning*. Cambridge, MA: MIT Press.
- Schroeder, R. (2014). Does Google shape what we know? *Prometheus: Critical Studies in Innovation* 32(2): 145–60.
- Segev, E. and Ahituv, N. (2010). Popular searches in Google and Yahoo! A 'digital divide' in information uses? *The Information Society* 26(1): 17–37.
- Taneja, H. and Wu, A. X. (2014). Does the Great Firewall really isolate the Chinese? Integrating access blockage with cultural factors to explain web user behavior. *The Information Society* 30(5): 297–309.
- Truman, G. (2016). *WebArchiving Environmental Scan*. Harvard Library Report. Cambridge, MA: Harvard Library. <http://nrs.harvard.edu/urn-3:HUL.InstRepos:25658314>. Accessed January 2016.
- Waller, V. (2011). Not just information: who searches for what on the search engine Google? *Journal of the American Society for Information Science and Technology* 62(4): 761–75.
- Webster, P. (2017). Users, technologies, organisations: Towards a cultural history of world web archiving. In N. Brügger (ed.), *Web 25: Histories from the first 25 Years of the World Wide Web*. New York: Peter Lang.
- Weller, T. (ed.) (2013). *History in the Digital Age*. London: Routledge.
- Wu, A. X. and Taneja, H. (2015). Reimagining internet geographies: A user-centric ethnological mapping of the world wide web. arXiv preprint arXiv:1510.04411.
- Wu, A. X. and Taneja, H. (2016). Reimagining internet geographies: A user-centric ethnological mapping of the world wide web. *Journal of Computer-Mediated Communication*, DOI: 10.1111/jcc4.12157.

Chapter 1

- Ainsworth, S. G., Alsum, A., SalahEldeen, H., Weigle, M. C., and Nelson, M. L. (2011). How much of the web is archived? In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, New York: ACM, 133–6.
- Aubry, S. (2010). Introducing web archives as a new library service: the experience of the national library of France. *Liber Quarterly* 20(2).
- Baeza-Yates, R., Castillo, C., and Efthimiadis, E. N. (2007). Characterization of national web domains. *ACM Transactions on Internet Technology (TOIT)* 7(2): 1–32.
- Bordino, I., Boldi, P., Donato, D., Santini, M., and Vigna, Sebastiano. (2008). Temporal Evolution of the UK Web. In *Proceedings of the ICDMW '08: IEEE International Conference on Data Mining Workshops, 2008*, New York: IEEE, 909–18.
- Brügger, N. (2011). Web archiving—Between past, present, and future. In M. Consalvo and C. Ess (eds), *The Handbook of Internet Studies*. Oxford: Wiley-Blackwell, 24–42.
- Brügger, N. (2013). Historical Network Analysis of the Web. *Social Science Computer Review* 31(3): 306–21.
- Brügger, N. (2014). Probing a nation's web sphere: A new approach to web history and a new kind of historical source. Paper presented at the 64th Annual Conference of the International Communication Association, Seattle.
- Dougherty, M. and Meyer, E. T. (2014). Community, tools, and practices in web archiving: The state-of-the-art in relation to social science and humanities research needs. *Journal of the Association for Information Science and Technology* 65(11): 2195–209.
- Dougherty, M., Meyer, E. T., Madsen, C., Van den Heuvel, C., Thomas, A., and Wyatt, S. (2010). *Researcher Engagement with Web Archives: State of the Art*. London: JISC. Accessed 27 June 2016 from <http://ssrn.com/abstract=1714997> and <http://ie-repository.jisc.ac.uk/544/>.
- Escher, T., Margetts, H., Petricek, V., and Cox, I. (2006). Governing from the centre: comparing the nodality of digital governments. Paper presented at the American Political Science Association Annual Conference.
- Foot, K. A. and Schneider, S. M. (2006). *Web Campaigning*. Cambridge, MA: MIT Press.
- Gomes, D., Freitas, S., and Silva, M. J. (2006). Design and Selection Criteria for a National Web Archive. In J. Gonzalo, C. Thanos, M. Felisa Verdejo, and R. C. Carrasco (eds), *Research and Advanced Technology for Digital Libraries: 10th European Conference, ECDL 2006, Alicante, Spain, September 17–22, 2006*. Berlin, Heidelberg: Springer Berlin Heidelberg, 196–207.
- Gorsky, M. (2015). Into the Dark Domain: The UK Web Archive as a Source for the Contemporary History of Public Health. *Social History of Medicine* 28(3): 596–616.
- Hale, S. A. (2012). Net increase? Cross-lingual linking in the blogosphere. *Journal of Computer-Mediated Communication* 17(2): 135–51.
- Hale, S. A., Yasseri, T., Cows, J., Meyer, E. T., Schroeder, R., and Margetts, H. (2014). Mapping the UK webspace: Fifteen years of British universities on the web. In *Proceedings of the 2014 ACM Conference on Web Science (WebSci '14)*, 62–70. New York: ACM.
- Hockx-Yu, H. (2011). The past issue of the web. In *Proceedings of the 3rd International Web Science Conference*, 1–8. New York: ACM.
- Huc-Hepher, S. (2015). Big Web data, small focus: An ethnosemiotic approach to culturally themed selective Web archiving. *Big Data & Society* 2(2): 2053951715595823.
- Jisc. (n.d.-a). Jisc UK Web Domain Dataset, Accessed 24 October 2016 from <http://data.webarchive.org.uk/opendata/ukwa.ds.2/host-linkage/>
- Jisc. (n.d.-b). Jisc UK Web Domain Dataset Description, Accessed 27 June 2016 from <http://dx.doi.org/10.5259/ukwa.ds.2/1>
- Kahle, B. (1997). Preserving the Internet. *Scientific American* 276(3): 82–3.
- Leiner, B. M., Cerf, V. G., Clark, D. D., Kahn, R. E., Kleinrock, L., Lynch, D. C., et al. (2009). A brief history of the Internet. *ACM SIGCOMM Computer Communication Review* 39(5): 22–31.
- Masanès, J. (2005). Web archiving methods and approaches: A comparative study. *Library Trends* 54(1): 72–90.
- Masanès, J. (2006). Web archiving: Issues and methods. In J. Masanès (ed.), *Web Archiving*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 1–54.
- Merton, R. K. (1968). The Matthew Effect in Science. *Science*, 159(3810), 56–63.
- Meyer, E. T., Thomas, A., and Schroeder, R. (2011). *Web Archives: The Future(s)*. London: IIPC. Accessed 27 June 2016 from <http://ssrn.com/paper=1830025>.

- Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences* 103(23): 8577–82.
- Pan, R. K., Kaski, K., and Fortunato, S. (2012). World citation and collaboration networks: uncovering the role of geography in science. [Article]. *Scientific Reports* 2: 902.
- Payne, N. and Thelwall, M. (2007). A longitudinal study of academic webs: Growth and stabilisation. *Scientometrics* 71(3): 523–39.
- Rogers, R. and Marres, N. (2000). Landscaping climate change: A mapping technique for understanding science and technology debates on the world wide web. *Public Understanding of Science* 9(2): 141–63.
- Rogers, R., Weltevrede, E., Borra, E., and Niederer, S. (2013). National Web Studies. In J. Hartley, J. Burgess, and A. Bruns (eds), *A Companion to New Media Dynamics*. Oxford: Wiley-Blackwell, 142–66.
- Thelwall, M., Tang, R., and Price, L. (2003). Linguistic patterns of academic web use in Western Europe. *Scientometrics* 56(3): 417–32.
- Thelwall, M. and Vaughan, L. (2004). A fair history of the Web? Examining country balance in the Internet Archive. *Library & Information Science Research* 26(2): 162–76.
- Thomas, A., Meyer, E. T., Dougherty, M., Van den Heuvel, C., Madsen, C., and Wyatt, S. (2010). *Researcher Engagement with Web Archives: Challenges and Opportunities for Investment*. London: JISC. Accessed 27 June 2016 from <http://ssrn.com/abstract=1715000> and <http://ie-repository.jisc.ac.uk/543/>.
- Žabička, P. and Matjka, L. (2007). Czech web archive analysis. *New Review of Hypermedia and Multimedia* 13(1): 27–37.

Chapter 2

- Ainsworth, S., AlSum, A., SalahEldeen, H., Weigle, M. C. and Nelson, M. L., (2011). How Much of the Web is Archived? *JCDL 2011*, ACM Press, Ottawa, Canada, 133–36.
- Ainsworth, S., AlSum, A., SalahEldeen, H., Weigle, M. C. and Nelson, M. L. (2013). How Much of the Web is Archived? Technical Report arXiv:1212.6177v2.
- Alexander, V. D., Blank, G. and Hale, S. A. (in preparation). How People Think about Distinction: Using Digital Trace Data to Examine User-Generated Cultural Hierarchies.
- Arms, W., Huttenlocher, D., Kleinberg, J., Macy, M. and Strang, D. (2006). From Wayback Machine to Yesternet: New opportunities for social science. Paper presented at *The 2nd International Conference on e-Social Science*, Manchester, UK, 29–30 June. Retrieved from <http://ent.cs.nccu.edu.tw/drupal/files/ArmsWaybackMachineToYesternet.pdf>, accessed 18 October 2016.
- Ayeh, J. K., Au, N. and Law, R. (2013). Do we believe in TripAdvisor? Examining credibility perceptions and online travelers' attitude toward using user-generated content, *Journal of Travel Research* 52(4): 437–52.
- Brügger, N. (2017). Probing a nation's web domain: A new approach to web history and a new kind of historical source. In G. Goggin and M. McLelland (eds), *The Routledge Companion to Global Internet Histories*. London: Routledge.
- Chu, S. C., Leung, L. C., Hui, Y. V. and Cheung, W. (2007). Evolution of e-commerce web sites: A conceptual framework and a longitudinal study. *Information and Management* 44(2): 154–64.
- Cunningham, P., Smyth, B., Wu, G. and Greene, D. (2010). Does TripAdvisor make hotels better? Technical Report, UCD-CSI-2010-06, School of Computer Science & Informatics, University College Dublin.
- Fetterly, D., Manasse, M., Najork, M. and Wiener, J. (2004). A large-scale study of the evolution of web pages. *Software-Practice and Experience* 34: 213–37.
- Hackett, S. and Parmanto, B. (2005). A longitudinal evaluation of accessibility: Higher education web sites. *Internet Research* 15(3): 281–94.
- Hale, S. A., Yasseri, T., Cows, J., Meyer, E. T., Schroeder, R. and Margetts, H. (2014). Mapping the UK webspace: Fifteen years of British universities on the Web. In *Proceedings of the 2014 ACM Conference on Web Science (WebSci '14)*. ACM, New York, 62–70. <http://dx.doi.org/10.1145/2615569.2615691>

- Internet Archive. (2014). Internet Archive: Petabox. Retrieved from <https://archive.org/web/petabox.php>, accessed 18 October 2016.
- Kimpton, M. and Ubois, J. (2006). Year-by-year: From an archive of the Internet to an archive on the Internet. In J. Masanès (ed.), *Web Archiving*. Berlin: Springer, 201–12.
- O'Connor, P. (2008). User-generated content and travel: A case study on tripadvisor.com. *Information and Communication Technologies in Tourism*, 47–58.
- Payne, N. and Thelwall, M. (2007). A longitudinal study of academic webs: Growth and stabilization. *Scientometrics*, 71(3), 523–539.
- Russell, E. and Kane, J. (2008). The missing link: Assessing the reliability of Internet citations in history journals. *Technology and Culture* 49(2): 420–9.
- Scott, S.V. and Orlikowski, W. J. (2012). Reconfiguring relations of accountability: Materialization of social media in the travel sector. *Accounting, Organizations and Society* 37: 26–40.
- Sparks, B. A. and Browning, V. (2010). Complaining in cyberspace: The motives and forms of hotel guests' complaints online. *Journal of Hospitality Marketing & Management* 19(7): 797–818.
- Stringam, B. B., and Gerdes, J. Jr (2010). An analysis of word-of-mouth ratings and guest comments of online hotel distribution sites. *Journal of Hospitality Marketing & Management* 19(7): 773–96.
- Thelwall, M. and Vaughan, L. (2004). A fair history of the Web? Examining country balance in the Internet Archive. *Library & Information Science Research* 26(2): 162–76.
- Thelwall, M. and Wilkinson, D. (2003). Three target document range metrics for university web-sites. *Journal of the American Society for Information Science and Technology* 54(1): 29–38.
- TripAdvisor (2014). About TripAdvisor. Retrieved from http://www.tripadvisor.co.uk/PressCenter-c6-About_Us.html (https://web.archive.org/web/20150505015934/http://www.tripadvisor.co.uk/PressCenter-c6-About_Us.html), accessed 5 May 2015.
- TripAdvisor (2015). Top Things to Do in London. Retrieved from http://www.tripadvisor.co.uk/Attractions-g186338-Activities-London_England.html, accessed 1 August 2015.
- UK Web Archive Open Data (n.d.). JISC UK Web Domain Dataset (1996–2013). Retrieved from <http://data.webarchive.org.uk/opendata/ukwa.ds.2/>, accessed 18 October 2016.
- Weinreich, H., Obendorf, H., Herder, E. and Mayer, M. (2008). Not quite the average: An empirical study of Web use. *ACM Transactions on the Web* 2(1): 1–31.
- Van de Sompel, H., Nelson, M. L., Sanderson, R., Balakireva, L. L., Ainsworth, S. and Shankar, H. (2009). Memento: Time travel for the Web. Technical Report, arXiv:0911.1112.
- Van de Sompel, H., Sanderson, R., Nelson, M., Balakireva, L., Shankar, H. and Ainsworth, S. (2010). An HTTP-based versioning mechanism for linked data. In *Proceedings of Linked Data on the Web Workshop (LDOW2010)*. Retrieved from http://events.linkedata.org/ldow2010/papers/ldow2010_paper13.pdf, accessed 18 October 2016.
- Vaughn, L. and Thelwall, M. (2003). Scholarly use of the web: What are the key inducers of links to journal web sites? *Journal of the American Society for Information Science and Technology* 54(1): 29–38.
- Weber, M. (2014). Observing the web by understanding the past: Archival Internet research. In *Proceedings of the 14th International World Wide Web Conference (WWW'14 Companion)*. Seoul, Korea, 1031–36. <http://dx.doi.org/10.1145/2567948.2579213>.

Chapter 3

- Agata, T., Miyata, Y., Ishita, E., Ikeuchi, A. and Ueda, S. (2014). Life span of web pages: A survey of 10 million pages collected in 2001. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*. New York: IEEE Press, 463–64.
- Andersen, B. (2006) The DK-domain: in words and figures. <http://netarkivet.dk/wp-content/uploads/DK-domaenet-i-ord-og-tal.pdf>, accessed 21 October 2016.
- Ben-David, A. (2014). Mapping minority webspaces: The case of the Arabic webspace in Israel. In D. Caspi and N. Elias (eds), *Ethnic Minorities and Media in the Holy Land*. London: Vallentine-Mitchell Academic, 137–57.
- Ben-David, A. (2016). What does the web remember of its deleted past? An archival reconstruction of the former Yugoslav top-level domain. *New Media & Society*, 18(7): 1103–19.
- Berlingske Business (2009, 19 June). Domæne-direktør vil skærpe haj-jagt, <http://www.business.dk/digital/domaene-direktoer-vil-skaerpe-haj-jagt>, accessed 27 May 2016.

- Brügger, N. (2009). Website history and the website as an object of study. *New Media & Society* 11(1–2): 115–32.
- Brügger, N. (2016, in press). Probing a nation's web domain: A new approach to web history and a new kind of historical source. In G. Goggin and M. McLelland (eds), *The Routledge Companion to Global Internet Histories*. New York: Routledge.
- Hale, S. A., Yasseri, T., Cows, J., Meyer, E. T., Schroeder, R. and Margetts, H. (2014). Mapping the UK webspace: fifteen years of British universities on the web. *WebSci '14 Proceedings of the 2014 ACM conference on Web science*, Bloomington, Indiana, June. DOI 10.1145/2615569.2615691.
- Hundt, M., Nesselhauf, N. and Biewer, C. (eds) (2007). *Corpus Linguistics and the Web*. Amsterdam & New York: Rodopi.
- Jackson, A. (2015) Ten years of the UK web archive: what have we saved? http://netpreserve.org/sites/default/files/attachments/2015_IIPC-GA_Slides_03_Jackson.pptx, accessed 21 October 2016.
- Laursen, D. and Møldrup-Dalum, P. (2017). Looking back, looking forward: 10 years of development to collect, preserve, and access the Danish web. In N. Brügger (ed.), *Web 25: Histories from the first 25 Years of the World Wide Web*. New York: Peter Lang.
- Millennium Development Goals Indicators. The official United Nations site for the MDG Indicators, <http://mdgs.un.org/unsd/mdg/SeriesDetail.aspx?srid=605>, accessed 27 May 2016.
- Moretti, F. (2000). Conjectures on world literature. *New left review*, 1, Jan.–Feb.: 56–8.
- Netarkivet (n.d.). Newsletters, 2006–2011, <http://netarkivet.dk/om-netarkivet/nyhedsbreve/#newsletters>, accessed 31 May 2016.
- Netarkivet (2015). Statistik, <http://netarkivet.dk/om-netarkivet/statistik/>, accessed 31 May 2016.
- Nominet (2013). Annual report and Accounts 2013. http://www.nominet.uk/wp-content/uploads/2015/08/nominet_report_and_accounts_2013.pdf, accessed 21 October 2016.
- Rogers, R., Weltevrede, E., Borra, E. and Niederer, S. (2013). National web studies: the case of Iran Online. In J. Hartley, J. Burgess and A. Bruns (eds), *A Companion to New Media Dynamics*. Oxford: Blackwell, 142–66.
- Schostag, S. and Fønss-Jørgensen, E. (2012). Webarchiving: Legal deposit of internet in Denmark. A curatorial perspective. *MDR* 41: 110–20.
- Storm, K. F. (1988). DKnet. *DKUUG-nyt*, 18, 13–19. <http://www.dkuug.dk/wp-content/themes/dkuug/arkiv/dkuug-nyt-018.pdf>, accessed 21 October 2016.
- Zierau, E. (2015). Identifying national parts of the internet. http://netpreserve.org/sites/default/files/attachments/2015_IIPC-GA_Slides_13b_Zierau.pptx, accessed 21 October 2016.

Chapter 4

- Ammann, R. (2011). Reciprocity, Social Curation and the Emergence of Blogging: A Study in Community Formation. *Procedia – Social and Behavioral Sciences* 22: 26–36.
- Anderson, M. and Caumont, A. (2014). How social media is reshaping news. Retrieved from <http://www.pewresearch.org/fact-tank/2014/09/24/how-social-media-is-reshaping-news/>. Accessed 12 October 2016.
- Benton, J. (2015, 24 March). A wave of distributed content is coming – will publishers sink or swim? Retrieved from <http://www.niemanlab.org/2015/03/a-wave-of-distributed-content-is-coming-will-publishers-sink-or-swim/>. Accessed 12 October 2016.
- Berners-Lee, T. (1991). The World Wide Web – past, present and future. *Journal of Digital Information* 1(1). Retrieved from <https://journals.tdl.org/>. Accessed 12 October 2016.
- Boczkowski, P. J. (1999). Understanding the development of online newspapers: Using computer-mediated communication theorizing to study Internet publishing. *New Media & Society* 1(1):, 101–26.
- Boczkowski, P. J. (2004a). *Digitizing the News: Innovation in online newspapers*. Cambridge, MA: MIT Press.
- Boczkowski, P. J. (2004b). The processes of adopting multimedia and interactivity in three online newsrooms. *Journal of Communication* 54(2): 16.x

- Boczkowski, P. J. (2010). *News at Work: Imitation in an age of information abundance*. Chicago: The University of Chicago Press.
- boyd, d. and Ellison, N. (2008). Social network sites: Definition, history and scholarship. *Journal of Computer-Mediated Communication* 13(1).
- Chiou, L. and Tucker, C. (2013). Paywalls and the demand for news. *Information Economics and Policy* 25(2): 61–9.
- Chung, D. S. (2007). Profits and perils: Online news producers' perceptions of interactivity and uses of interactive features. *Convergence: The International Journal of Research into New Media Technologies* 13(1): 43–61.
- Cruz-Cunha, M. M., Gonzales, P., Lopes, N., Miranda, E. M. and Putnik, G. D. (2011). Preface. In M. M. Cruz-Cunha, P. Gonzales, N. Lopes, E. M. Miranda and G. D. Putnik (eds), *Handbook of Research on Business Social Networking: Organizational, Managerial, and Technological Dimensions*. Hershey, PA: IGI Global.
- Deuze, M. (2003). The Web and its journalism: Considering the consequences of different types of newsmedia online. *New Media & Society* 5(2): 203–30.
- Digital: Top 50 online news entities. (2015). Retrieved from <http://www.journalism.org/media-indicators/digital-top-50-online-news-entities-2015/>. Accessed 11 October 2016.
- Downie, L. and Schudson, M. (2009). The reconstruction of American journalism. *Columbia Journalism Review* 19. Retrieved from http://www.cjr.org/reconstruction/the_reconstruction_of_american.php. Accessed 11 October 2016.
- Falkenberg, V. (2010). (R)evolution under construction: The dual history of online newspapers and newspapers online. In N. Bruggen (ed.), *Web History*. New York: Peter Lang, 233–56.
- Gao, Y. and Vaughn, L. (2006). Web hyperlink profiles of news sites: A comparison of newspapers of USA, Canada and China. *ASLIB Proceedings* 57(5): 398–411.
- Greer, J. D. and Mensing, D. (2004). U.S. news web sites better, but small papers still lag. *Newspaper Research Journal* 25(2): 98–112.
- Hayes, D. and Lawless, J. L. (2015). As local news goes, so goes citizen engagement: Media, knowledge, and participation in US House elections. *The Journal of Politics* 77(2): 447–62.
- Karimi, J. and Walter, Z. (2015). The role of dynamic capabilities in responding to digital disruption: A factor-based study of the newspaper industry. *Journal of Management Information Systems* 32(1): 39–81.
- Kawamoto, K. (2003). *Digital Journalism: Emerging media and the changing horizons of journalism*. London: Rowman & Littlefield Publishers.
- Kohut, A., Doherty, C., Dimock, M. and Keeter, S. (2012). *Trends in News Consumption: 1991–2012*. Washington, DC: Pew Research Center.
- LaFrance, A. and Meyer, R. (2015). The eternal return of Buzzfeed. *The Atlantic*.
- Lewis, S. C. (2011). Journalism innovation and participation: An analysis of the Knight News Challenge. *International Journal of Communication*, 5: 1623–48.
- Liu, J. and Birnbaum, L. (2008). Localsavvy: Aggregating Local Points of View about News Issues. Paper presented at the Proceedings of the first international workshop on Location and the web, Beijing, China.
- Matheson, D. (2004). Weblogs and the epistemology of the news: some trends in online journalism. *New Media and Society* 6(4): 443–68.
- Mitchell, A. and Rosenstiel, T. (2010). *State of the News Media 2010*. Washington, DC: Pew Research Center.
- Moy, P., McCluskey, M. R., McCoy, K. and Spratt, M. A. (2004). Political correlates of local news media use. *Journal of Communication* 54(3): 532–46.
- Mysiani, F. (2013). Governance by algorithms. *Internet Policy Review* 2(3).
- Napoli, P. M., Stonbely, S., McCollough, K. and Renninger, B. (2015). *Assessing the Health of Local Journalism Ecosystems*. New Brunswick, NJ: Rutgers University.
- Nielsen, R. K. (2015). *Local Journalism: The decline of newspapers and the rise of digital media*. Oxford: Reuters Institute for the Study of Journalism.
- O'Reilly, T. (2005). What is Web 2.0: Design patterns and business models for the next generation of software. Retrieved from <http://oreilly.com/web2/archive/what-is-web-20.html>
- Paek, H.-J., Yoon, S.-H. and Shah, D. V. (2005). Local news, social integration, and community participation: Hierarchical linear modeling of contextual and cross-level effects. *Journalism & Mass Communication Quarterly* 82(3): 587–606.

- Patterson, T. (2007). *Creative Destruction: An exploratory look at news on the Internet*. Boston, MA: Joan Shorenstein Center on the Press, Politics and Public Policy.
- Perelman, J. (2014). Content and distribution are the keys to brand building on the social web. *Journal of Digital & Social Media Marketing* 2(1): 12–18.
- Perren, A. (2010). Business as unusual: Conglomerate-sized challenges for film and television in the digital arena. *Journal of Popular Film & Television* 38(2): 72–8.
- Pickard, V. and Williams, A. T. (2013). Salvation or folly? *Digital Journalism* 2(2): 195–213.
- Rettberg, J. W. (2008). *Blogging*. Malden, MA: Polity Press.
- Saltzis, K. (2012). Breaking news online. *Journalism Practice* 6(5–6): 702–10.
- Schlesinger, P. and Doyle, G. (2015). From organizational crisis to multi-platform salvation? Creative destruction and the recomposition of news media. *Journalism* 16(3): 305–23.
- Shumate, M. and Lipp, J. (2008). Connective collective action online: An examination of the hyperlink network structure of an NGO issue network. *Journal of Computer-Mediated Communication* 14: 178–201.
- Sood, S., Owsley, S., Hammond, K. J. and Birnbaum, L. (2007). TagAssist: Automatic Tag Suggestion for Blog Posts. Paper presented at the ICWSM.
- Stovall, J. G. (2004). *Web journalism: Practice and promise of a new medium*. Boston, MA: Pearson Education.
- Suler, J. (2004). The online disinhibition effect. *CyberPsychology & Behavior* 7(3): 321–26.
- Sylvie, G. and Witherspoon, P. D. (2002). *Time, Change and the American Newspaper*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Tsui, L. (2008). The hyperlink in newspapers and blogs. In J. Turow and L. Tsui (eds), *The Hyperlinked Society: Questioning Connections in the Digital Age*. Ann Arbor: University of Michigan Press, 70–84.
- Tuchman, G. (1978). *Making News*. New York City: Free Press.
- Turow, J. and Tsui, L. (2008). *The Hyperlinked Society: Questioning Connections in the Digital Age*. Ann Arbor: University of Michigan Press.
- Usher, N. (2014). Making news at *The New York Times*. Ann Arbor: University of Michigan Press.
- Wadbring, I. and Bergström, A. (2015). A print crisis or a local crisis? *Journalism Studies* 1–16.
- Weber, M. S. (2012). Newspapers and the long-term implications of hyperlinking. *Journal of Computer-Mediated Communication* 17(2): 187–201.
- Weber, M. S. and Monge, P. (2011). The flow of digital news in a network of sources, authorities, and hubs. *Journal of Communication* 61(6): 1062–81.
- Weber, M. S. and Monge, P. (2014). Industries in turmoil: Driving transformation during periods of disruption. *Communication Research* 1–30.
- Winter, S., Brückner, C. and Krämer, N. C. (2015). They came, they liked, they commented: social influence on Facebook news channels. *Cyberpsychology, Behavior, and Social Networking* 18(8): 431–6.

Chapter 5

- Barnett, G. A., Chung, C. J. and Park, H. W. (2011). Uncovering transnational hyperlink patterns and web-mediated contents: A new approach based on cracking .com domain. *Social Science Computer Review* 29(3): 369–84.
- Van Belle, D. A. (2000). New York Times and network TV news coverage of foreign disasters: The significance of the insignificant variables. *Journalism & Mass Communication Quarterly* 77(1): 50–70.
- Born, G. (2003). From Reithian ethic to managerial discourse: Accountability and audit at the BBC. *The Public* 10(2): 63–80.
- Bright, J. (2015). The Social Gap: How social media shares socially important news, and why it matters. Mimeo.
- Bright, J. and Nicholls, T. (2014). The life and death of political news: Measuring the impact of the audience agenda using online data. *Social Science Computer Review* 32(2): 170–81.
- Chang, K. K. and Lee, T. T. (2009). International news determinants in U.S. news media in the post-cold war era. In G. Golan, T. Johnson and W. Wanta (eds), *International Media Communication in a Global Age* (pp. 71–88). London: Routledge.

- Chang, T. K., Shoemaker, P. J. and Brendlinger, N. (1987). Determinants of international news coverage in the US media. *Communication Research* 14(4): 396–414.
- Charles, J., Shore, L. and Todd, R. (1979). The New York Times coverage of equatorial and lower Africa. *Journal of Communication* 29(2): 148–55.
- Coddington, M. (2012). Building frames link by link: The linking practices of blogs and news sites. *International Journal of Communication* 6: 20.
- Connor, A. (2007). Revolution Not Evolution, BBC Online. Accessible at http://www.bbc.co.uk/blogs/bbcinternet/2007/12/revolution_not_evolution.html. (Accessed 27 August 2015).
- Dupree, J. D. (1971). International communication: View from 'A window on the world'. *Gazette* 17:, 224–35.
- Golan, G. J. (2008). Where in the world is Africa? Predicting coverage of Africa by US television networks. *International Communication Gazette*, 70(1), 41–57.
- Golan, G. and Wanta, W. (2003). International elections on US network news: an examination of factors affecting newsworthiness. *International Communication Gazette* 65(1): 25–39.
- Graf, P. (2004). Report of the Independent Review of BBC Online. Accessible at http://news.bbc.co.uk/nol/shared/bsp/hi/pdfs/05_07_04_graf.pdf. Accessed 16 September 2016.
- Hale, S. A., Yasseri, T., Cows, J., Meyer, E. T., Schroeder, R. and Margetts, H. (2014). Mapping the UK Webspace: Fifteen Years of British Universities on the Web. In *Proceedings of the 2014 ACM Conference on Web Science*. Bloomington, IN, 62–70.
- Huggers, E. (2011). Reshaping BBC Online. Accessible at <http://www.bbc.co.uk/blogs/about-thebbc/2011/01/delivering-quality-first.shtml>. Accessed 16 September 2016.
- Ishii, K. (1996). Is the US over-reported in the Japanese Press? *Gazette* 57: 135–44.
- Kahle, B. (1997). Preserving the internet. *Scientific American* 276(3): 82–3.
- Kim, K. and Barnett, G. A. (1996). The determinants of international news flow. A network analysis. *Communication Research* 23(3): 323–52.
- Lippmann, W. (1922). The world outside and the pictures in our heads. *Public Opinion* 4: 1–22.
- McCombs, M. E. and Shaw, D. L. (1993). The evolution of agenda-setting research: Twenty-five years in the marketplace of ideas. *Journal of Communication* 43:, 58–67.
- McCombs, M. E. and Shaw, D. L. (1972). The agenda-setting function of mass media. *Public Opinion Quarterly*: 176–87.
- Martin, F. (2005). Net worth: The unlikely rise of ABC Online. In G. Goggin (ed.), *Virtual Nation: The Internet in Australia* (pp. 193–208). Sydney: UNSW Press.
- Meyer, W. H. (1989). Global news flows dependency and neoimperialism. *Comparative Political Studies* 22(3): 243–64.
- Moe, H. (2003). Digital television and the state of public service broadcasting (Report 54). Bergen, Norway: University of Bergen, Department of Media Studies.
- Nnaemeka, T. and Richstad, J. (1980). Structured relations and foreign news flow in the Pacific region. *International Communication Gazette* 26(4): 235–57.
- Norris, P. and Inglehart, R. (2009). *Cosmopolitan Communications: Cultural Diversity in a Globalized World*. Cambridge: Cambridge University Press.
- OXIS. (2013). Cultures of the Internet: The Internet in Britain. Oxford Internet Survey 2013 Report. Available from: <http://oxis.oii.ox.ac.uk/wp-content/uploads/2014/11/OxIS-2013.pdf> (Accessed 11 September 2015).
- Park, H. W., Barnett, G. A. and Chung, C. J. (2011). Structural changes in the 2003–2009 global hyperlink network. *Global Networks* 11(4): 522–42.
- Shoemaker, P. J. (1991). *Gatekeeping*. London: Sage Publications.
- Shoemaker, P. J., Chang, T. K. and Brendlinger, N. (1986). Deviance as a predictor of newsworthiness: Coverage of international events in the U.S. media. In M. L. McLaughlin (ed.), *Communication Yearbook*, 10. Newbury Park, CA: Sage.
- Skurnik, W. A. E. (1981). Foreign news coverage in six African newspapers: The potency of national interests. *International Communication Gazette* 28(2): 117–30.
- Thorsen, E. (2010). BBC News Online: A brief history of past and present. In N. Brügger (ed.), *Web History* (pp. 213–32). New York: Peter Lang.
- Wilke, J. (1987). Foreign news coverage and international news flow over three centuries. *Gazette* 39(3): 147–80.
- Wu, H. D. (2007). A brave new world for international news? Exploring the determinants of the coverage of foreign news on US websites. *International Communication Gazette* 69(6): 539–51.
- Wu, H. D. (2000). Systemic determinants of international news coverage: A comparison of 38 countries. *Journal of Communication* 50(2): 110–30.

Chapter 6

- Adminet (n.d.). Un bouquet de rapports. <http://admi.net/literacy/bouquet.html>. Accessed 22 August 2015.
- Ankerson, M. S. (2009). Historicizing web design: Software, style and the look of the web. In J. Staiger, J. and S. Hake (eds), *Convergence Media History* (pp. 192–203). New York, London: Routledge.
- Bangemann, M. et al. (1994). L'Europe et la société de l'information planétaire – Recommandations au Conseil des ministres de l'Union européenne. Bruxelles. <http://urlz.fr/2dj7>. Accessed 22 August 2015.
- Baquiast, J.-P. (1998). *Propositions sur les apports d'Internet à la modernisation du fonctionnement de l'État: rapport d'orientation*. Paris: La documentation française.
- Ben-David, A. and Huurdeman, H. (2014). Web archive search as research: Methodological and theoretical implications. *Alexandria* 25(1): 93–111.
- Bortzmeyer, S. (1996). Refere UEJF: mon compte-rendu. 18 March. [https://groups.google.com/forum/#!searchin/fr.network.internet/bortzmeyer\\$20AUI\\$20procès\\$20Renater/fr.network.internet/bNkar8_8gE4/PgrvMHUIDZMJ](https://groups.google.com/forum/#!searchin/fr.network.internet/bortzmeyer$20AUI$20procès$20Renater/fr.network.internet/bNkar8_8gE4/PgrvMHUIDZMJ). Accessed 22 August 2015.
- Bouquillion, P. and Pailliant, I. (2006). *Le déploiement des TIC dans les territoires*. Grenoble: PUG.
- Brügger, N. (2009). Website history and the website as an object of study. *New Media & Society* 11(1–2): 115–32.
- Brügger, N. (2012a). L'historiographie de sites Web: Quelques enjeux fondamentaux. *Le Temps des Médias* 18(1): 159–69.
- Brügger, N. (2012b). When the present web is later the past: Web historiography, digital history, and internet studies. *Historical Social Research* 37(4): 102–17.
- Brügger, N. (2012c). Web history and the web as a historical source. *Zeithistorische Forschungen* 9: 316–25.
- Cern (1994). WWW94 Awards. Cern website. 28 May. <http://www94.web.cern.ch/WWW94/Awards0529.html>. Accessed 22 August 2015.
- Chemla, L. (2002). *Confessions d'un voleur. Internet: la liberté confisquée*. Paris: Denoël.
- Cohen, D. and Debonneuil, M. (2000). *Nouvelle économie*. Paris: La documentation française.
- Collective (1998). *Le Guide du Routard Internet*. Paris: Hachette.
- Curtill, C. (1996). *La carte française des inforoutes*. Paris: Hermes Science Publications.
- d'Attilio, H. (1998). *Le développement des Nouvelles Technologies d'Information et de Communication dans les Collectivités Locales: de l'expérimentation à la généralisation, Rapport au Premier Ministre*. Paris: La documentation française.
- Délégation interministérielle à la réforme de l'État (DIRE). 2001. Le développement des sites Internet des services de l'État. Évaluation au printemps 2000. <http://www.ladocumentationfrancaise.fr/rapports-publics/014000796/index.shtml>. Accessed 22 August 2015.
- Desautz, L. (2000). L'État planche sur la mise au Net des services publics. *La Tribune*. 27 April.
- Dougherty, M. et al. (2010). *Researcher Engagement with Web Archives: State of the Art*. London: JISC.
- Eko, L. S. (2013). *American Exceptionalism, the French Exception, and Digital Media Law*. New York: Lexington Books.
- Eveno, E. (1998). Parthenay, modèle français et européen de ville numérisée. In A. Lefebvre and Tremblay, G. (eds), *Autoroutes de l'information et Dynamiques territoriales*. Paris: PUF, 129–48.
- Falque-Pierrotin, I. (1997). *Internet. Enjeux juridiques. Rapport au ministre délégué à la Poste, aux Télécommunications et à l'Espace et au ministre de la Culture*. Paris: La documentation française. <http://www.ladocumentationfrancaise.fr/rapports-publics/974057500/index.shtml>. Accessed 22 August 2015.
- Flichy, P. (1996). Présentation. *Réseaux* 14(77): 5–6.
- Hallier, A. and Rassat, B. (2007). Documentary *Quand l'Internet fait des bulles*, part 1. 13ème Rue. <https://www.youtube.com/watch?v=Hj7KoLITX0k>. Accessed 22 August 2015.
- INA Archives (1995a). Recette Bombe Internet. Midi 2, 3 Augus3. <http://www.ina.fr/video/CAB95042655> Last accessed 22 August 2015.
- INA Archives (1996a). Pédophilie sur Internet. 19/20, France 3, 7 May. <http://www.ina.fr/video/CAC96019270/pedophilie-sur-internet-video.html>. Accessed 22 August 2015.
- INA Archives (1996b). Médicaments/Internet. JA2 20H, 4 October. <http://www.ina.fr/video/CAB96050811>. Accessed 22 August 2015.

- Internet Archive (1997a). Homepage from the Strasbourg Board of Education website. Archived 12 January 1997. <http://web.archive.org/web/19970112024736/> <http://www.ac-strasbourg.fr/>. Accessed 24 July 2015.
- Internet Archive (1997b). Homepage from the Strasbourg Board of Education website. Archived 10 December. <http://web.archive.org/web/19971210212812/> <http://www.ac-strasbourg.fr/>. Accessed 24 July 2015.
- Internet Archive (1998). Websites in.gouv.fr listed by NIC France website. Archived 4 February 1998. <http://web.archive.org/web/19980204192838/> <http://www.nic.fr/Annuaire/france/gouv/gouv.html>. Accessed 22 August 2015.
- Internet Archive (1999). Cyberi Homepage. Issy-les-Moulineaux. Archived 29 January 1999. <http://web.archive.org/web/19990129025023/> <http://www.issy.com/club-int/cyberi.html>. Accessed 2 December 2015.
- Internet Archive (2000). Page from the Strasbourg Board of Education website. Archived 17 August 2000. <http://web.archive.org/web/20000817041856/> <http://www.ac-strasbourg.fr/>. Accessed 24 July 2015.
- Jospin, L. (1997). Préparer l'entrée de la France dans la société de l'information. Hourtin, Université de la communication. <http://www.admiroutes.asso.fr/action/theme/politic/lionel.htm>. Accessed 15 October 2016.
- Lacambre, V. (2012). Interview by Valérie Schafer. 4 January. Paris: France.
- L'Atelier (1999). L'expérience d'Intranet local 'In-Town-Net' menée par Parthenay demeure relativement unique. Paris. <http://www.atelier.net/trends/articles/lexperience-dintranet-local-lin-town-net-menee-parthenay-demeure-relativement-unique> Accessed 22 August 2015.
- Marchandise, J.-F., Dupuis, C. and Kaplan, D. (1999). Commissariat général du plan, étude de l'usage pratique des NTIC au sein de l'administration. Final Report. Terra Nova Studio. <http://www.ladocumentationfrancaise.fr/rapports-publics/014000796/index.shtml>. Accessed 22 August 2015.
- Mussou, C. (2012). Et le Web devint archive: enjeux et défis. *Le Temps des Médias* 19: 259–66.
- Pioch, N. (1995). Art sur W3. FLTEACH ARCHIVES. <https://listserv.buffalo.edu/cgi-bin/wa?A2=ind9501&L=fteach&D=1&F=P&P=546612>, Accessed 22 August 2015.
- Ponterio, R. (1995). France shuts down the Weblouvre. ListServ Homepage, FLTEACH Archives. <https://listserv.buffalo.edu/cgi-bin/wa?A2=ind9501&L=fteach&D=1&F=P&P=546612>. Accessed 22 August 2015.
- Prot, M. (2003). Naissance du projet CIM@ISE de refonte du site Internet du musée du Louvre. Paris: École du Louvre. <http://www.archimuse.com/publishing/ichim03/119C.pdf>. Accessed 22 August 2015.
- Russell, A. L. and Schafer, V. (2014). In the shadow of ARPANET and internet: Louis Pouzin and the Cyclades Network in the 1970s. *Technology and Culture* 55(4): 880–907.
- Tronc, J.-N. (2011). Interview by Valérie Schafer. 6 September. Paris: France.
- Théry, G. (1994). *Les autoroutes de l'information*. Paris: La documentation française.
- Vidal, P. (2007). La permanence d'une politique publique TIC: de Parthenay, Ville 'numérisée' à Parthenay 'Ville numérique'. *Netcom: Networks and Communication Studies* 21(1–2): 137–64.

Chapter 7

- Alaitha (1997, 1 March). Introduction – The Elements of Web Page Style – Shady Oaks. Retrieved 25 July 2013, from <http://web.archive.org/web/19970301083309/> <http://www1.geocities.com/Heartland/5419/elements.htm>
- Anderson, B. (1991). *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. London: Verso.
- Archive Team (2009) The Archive Team GeoCities Snapshot. Retrieved 19 November 2015, from <https://archive.org/details/2009-archiveteam-geocities-part1>.
- Augusta Golf Neighborhood (Unknown). Augusta Award Application. Retrieved 13 July 2015, from <http://www.oocities.org/augusta/1020/birdform.htm>
- Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research* 3: 993–1022.

- Business Wire (1995). Beverly Hills Internet, builder of interactive cyber cities, launches 4 more virtual communities linked to real places. *Business Wire*, 5 July. Retrieved from <http://web.archive.org/web/19961221091944/http://www.thefreelibrary.com/Beverly+Hills+Internet,+builder+of+interactive+cyber+cities,+launches...-a017190114>
- Doheny-Farina, S. (1996) *The Wired Neighbourhood*. New Haven, CT: Yale University Press.
- GeoCities. (1996a, 21 December). Athens' Community Leaders. Retrieved 25 July 2013, from <http://web.archive.org/web/19961221091944/http://www.geocities.com/Athens/9999/>
- GeoCities. (1996b, 21 December). GeoCities Heartland Community Leaders. Retrieved 12 July 2015, from <http://web.archive.org/web/19961221015607/http://www.geocities.com/Heartland/leader.html>
- GeoCities. (1996c, 21 December). GeoCities FAQ Page 3. Retrieved 20 October 2016, from <http://web.archive.org/web/19961221005732/http://www.geocities.com/homestead/FAQ/faqpage3.html>
- GeoCities. (1996d, 21 December). GeoCities Community Leaders. Retrieved 12 October 2016, from <http://web.archive.org/web/19961221010319/http://www.geocities.com/homestead/homeleader.html>
- GeoCities. (1996e, 1996). GeoCities Homesteading on the World Wide Web – Q & A. Retrieved 20 October 2016, from http://www.ewebtribe.com/remember/GC_FAQ_old.html
- GeoCities. (1997a, February 22). GeoCities Homesteading Program. Retrieved 12 October 2016, from <http://web.archive.org/web/19970222174816/http://www1.geocities.com/homestead/>
- GeoCities. (1997b, April 13). GeoCities Neighborhood Watch. Retrieved 12 October 2016, from http://web.archive.org/web/19970413014952/http://www7.geocities.com/homestead/neighbor_watch.html
- GeoCities. (1997c, March 1). About the Heartland Community Leaders. Retrieved 12 July 2015, from <http://web.archive.org/web/19970301082611/http://www1.geocities.com/Heartland/7546/hclabout.html>
- GeoCities. (1998, 5 July). Homestead Add-Ons. Retrieved 13 July 2015, from <http://web.archive.org/web/19980705020058/http://www6.geocities.com/members/addons>
- Graham, G. (1999). *The Internet: A Philosophical Inquiry*. Abingdon: Routledge.
- Hansell, S. (1998) The Neighbourhood Business; GeoCities' Cyberworld is Vibrant, but Can it Make Money? *New York Times*, 13 July.
- Hill, B. (2000). *Yahoo For Dummies* (2nd edition). Foster City, CA: For Dummies.
- Htm_help. (1996). The "Home Page" Home Page. Retrieved 13 July 2015, from <http://web.archive.org/web/19961221005656/http://www.geocities.com/Athens/2090/>
- Jockers, M. L. (2011, 29 September). The LDA Buffet is Now Open; or, Latent Dirichlet Allocation for English Majors. Retrieved 18 October 2016, from <http://www.matthewjockers.net/2011/09/29/the-lda-buffet-is-now-open-or-latent-dirichlet-allocation-for-english-majors/>
- Kamvar, S., Haveliwala, T., Manning, C. and Golub, G. (2003). Exploiting the Block Structure of the Web for Computing PageRank. Stanford. Retrieved 18 October 2016, from <http://ilpubs.stanford.edu:8090/579/1/2003-17.pdf>
- Karlins, D. (2003). *Build Your Own Web Site* (1st edition). New York: McGraw-Hill Osborne Media.
- Kendall, L. (2011) Community and the internet. In M. Consalvo and C. Ess, *The Handbook of Internet Studies*. Wiley-Blackwell, 309–25.
- Lialina, O. (2013) Some remarks on #neocities @kyledrake. *One Terabyte of Kilobyte Age*. Retrieved 18 October 2016 from <http://contemporary-home-computing.org/1tb/archives/4012>.
- Licklider, J. C. R. and Taylor, R. W. (1968). The computer as a communication device. *Science and Technology*, 20–41.
- Logie, J. (2002). Homestead acts: Rhetoric and property in the American West, and on the World Wide Web. *Rhetoric Society Quarterly* 32(3): 33–59.
- Manovitch, L. (2012). Guide to Visualizing Video and Image Sequences. Retrieved 5 June 2014 from <https://docs.google.com/document/d/1PqSZmKwQwSIFrbmVievbStbt7PrtsxNgC3W1oY5C4/edit>
- Montello, D. R., Fabrikant, S. I., Ruocco, M. and Middleton, R. S. (2003). Testing the first law of cognitive geography on point-display spatializations. In W. Kuhn, M. F. Worboys and S. Timpf (eds), *Spatial Information Theory. Foundations of Geographic Information Science*. Springer Berlin Heidelberg, 316–31. Retrieved 18 October 2016 from http://link.springer.com/chapter/10.1007/978-3-540-39923-0_21

- Moschovitis, C. J. P. (1999) *History of the Internet: A Chronology, 1843 to the Present*. Santa Barbara, CA: ABC-Clío.
- Motavalli, J. (2004). *Bamboozled at the Revolution: How Big Media Lost Billions in the Battle for the Internet*. New York: Penguin Putnam.
- Ocamb, K. (2012). David Bohnett: Social change through community commitment. *Frontiers*, 16 October. 18.
- Porter, C. (2004) A typology of virtual communities: A multi-disciplinary foundation for future research. *Journal of Computer-Mediated Communication*, 10(1).
- Putnam, R. D. (2000). *Bowling Alone: The Collapse and Revival of American Community*. New York: Simon & Schuster.
- RainForest Community Leaders (Unknown). OuttaSite Awards. Retrieved 13 July 2015, from http://www.oocities.org/rainforest/9900/cl_only/clouttasite.html
- Rheingold, H. (2000) *The Virtual Community: Homesteading on the Electronic Frontier*. Cambridge, MA: MIT Press.
- Ridey, R. (1996) Roger Ridey travels under the volcano, and also discovers a Web full of creepy-crawlies. *The Independent* (UK), 12 February.
- Sawyer, B. and Greely, D. (1999). *Creating Geocities Websites*. Cincinnati, OH: Music Sales Corporation.
- Scott, J. (Unknown) Please be patient – This Page is Under Construction! Accessed 18 October 2016 from <http://www.textfiles.com/underconstruction/>.
- Turner, F. (2008). *From Counterculture to Cyberculture: Stewart Brand, the Whole Earth Network, and the Rise of Digital Utopianism*. Chicago: University of Chicago Press.
- Walker, K. (2000) 'It's Difficult to Hide It': The presentation of self on internet home pages. *Qualitative Sociology* 23(1): 99–120.
- Zacharek, S. (1999). Addicted to eBay. *Salon*. Retrieved 18 October 2016 from http://www.salon.com/1999/12/30/feature_237/.

Chapter 8

- Ackland, R. (2009). Social network services as data sources and platforms for e-researching social networks. *Social Science Computer Review Special Issue on e-Social Science* 27(4): 481–92.
- Ackland, R. (2013). *Web Social Science: Concepts, Data and Tools for Social Scientists in the Digital Age*. London: Sage Publications.
- Ackland, R. and Evans, A. (2005). The visibility of abortion-related information on the World Wide Web. Presented to the Australian Sociological Association Conference, University of Tasmania, Sandy Bay Campus, 5–8 December. http://voson.anu.edu.au/papers/TASA2005_Ackland_Evans_for_web.pdf. Accessed 22 September 2015.
- Ackland, R. and O'Neil, M. (2011). Online collective identity: The case of the environmental movement. *Social Networks* 33: 177–90.
- Ackland, R. and Shorish, J. (2014). Political homophily on the web. In M. Cantijoch, R. Gibson and S. Ward (eds), *Analysing Social Media Data and Web Networks*, Basingstoke: Palgrave Macmillan.
- Ackland, R. and Zhu, J. (2015). Social network analysis. In P. Halfpenny and R. Procter (eds), *Innovations in Digital Research Methods*. London: Sage Publications.
- Ackland, R., Gibson, R., Lusoli, W. and Ward, S. (2010). Engaging with the public? Assessing the online presence and communication practices of the nanotechnology industry. *Social Science Computer Review* 28(4): 443–65.
- Adamic, L. and Glance, N. (2005). The political blogosphere and the 2004 U.S. election: Divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery (LINKDD 2005)*. New York: Association for Computing Machinery, 6–43.
- Albury, R. (1999). *The Politics of Reproduction: Beyond the Slogans*. St Leonards: Allen & Unwin.
- Andrew, M. and Maddison, S. (2010). Damaged but determined: The Australian Women's Movement, 1996–2007. *Social Movement Studies* 9(2): 171–85.

- Bounegru, L. (2011). Mapping the Abortion Debate on the Romanian Web: Top Google Rankings as measure of popularity or marginality? Paper presented at the Digital Methods Initiative mini-conference, January 2011. <http://web.mit.edu/comm-forum/mit7/papers/Bounegru.pdf>. Accessed 1 February 2016.
- Brügger, N. (2012). Historical network analysis of the web. *Social Science Computer Review* 31(3): 306–21.
- Davenport, E. and Cronin, B. (2000). The citation network as a prototype for representing trust in virtual environments. In B. Cronin and Atkins, H. (eds), *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield*. Metford, NJ: Information Today.
- Ferree, M., Gamson, W., Gerhards, J. and Rucht, D. (2002). *Shaping Abortion Discourse: Democracy and the Public Sphere in Germany and the United States*. Cambridge: Cambridge University Press.
- Foot, K. A., Schneider, S. M., Dougherty, M., Xenos, M. and Larsen, E. (2003). Analyzing linking practices: Candidate sites in the 2002 U.S. electoral web sphere. *Journal of Computer-Mediated Communication* 8(4).
- Fulk, J., Flanagin, A., Kalman, M., Monge, P. and Ryan, T. (1996). Connective and communal public goods in interactive communication systems. *Communication Theory* 6: 60–87.
- Hanneman, R. A. and Riddle, M. (2005). Introduction to social network methods. University of California Riverside, Published in digital form at <http://faculty.ucr.edu/~hanneman/net-text>. Accessed 25 October 2016.
- Hargittai, E., Gallo, J. and Kane, M. (2008). Cross-ideological discussions among conservative and liberal bloggers. *Public Choice* 134: 67–86.
- Hindman, M. (2008). *The Myth of Digital Democracy*. Princeton, NJ: Princeton University Press.
- Jackson, M. H. (1997). Assessing the structure of communication on the World Wide Web. *Journal of Computer-Mediated Communication* 3(1): 273–99.
- Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM* 46(5): 604–32.
- Lusher, D. and Ackland, R. (2011). A relational hyperlink analysis of an online social movement. *Journal of Social Structure* 12(5). <http://www.cmu.edu/joss/content/articles/volume12/Lusher/>. Accessed 25 October 2016.
- McLaren, K. (2013). The emotional imperative of the visual: Images of the fetus in contemporary Australian pro-life politics. *Advances in the Visual Analysis of Social Movements* 35: 81–103.
- Park, H. W. and Thelwall, M. (2003). Hyperlink analyses of the world wide web: A review. *Journal of Computer-Mediated Communication* 8(4).
- Park, H. W., Kim, C. S. and Barnett, G. A. (2004). Socio-communicational structure among political actors on the web in South Korea: The dynamics of digital presence in cyberspace. *New Media & Society* 6(3): 403–23.
- Parliamentary Library (2005). How many abortions are there in Australia? A discussion of abortion statistics, their limitations, and options for improved statistical collection. Department of Parliamentary Services, Canberra.
- Putnam, R. (2000). *Bowling Alone: The Collapse and Revival of American Community*. New York: Simon & Schuster.
- Rogers, R. and Marres, N. (2000). Landscaping climate change: A mapping technique for understanding science and technology debates on the world wide web. *Public Understanding of Science* 9(2): 141–63.
- Rogers, R. and Zelman, A. (2002). Surfing for knowledge in the information society. In G. Elmer (ed.), *Critical Perspectives on the Internet*. Lanham, MD: Rowman & Littlefield, 63–86.
- Shumate, M. and Dewitt, L. (2008). The North/South divide in NGO hyperlink networks. *Journal of Computer Mediated Communication* 13: 405–28.
- Siedlecky, S. (2005). The abortion issue all over again. *New Doctor* 83: 16–18, 21.
- Sunstein, C. (2001). *Republic.com*. Princeton, NJ: Princeton University Press.
- Wasserman, S. and Faust, K. (2004). *Social Network Analysis*. Cambridge: Cambridge University Press.
- Wyatt, D. and Hughes, K. (2009). When discourse defies belief: anti-abortionists in contemporary Australia. *Journal of Sociology* 45(3): 235–53.

Chapter 9

- Amarasingam, A. (2010). Introduction: what is the New Atheism? In A. Amarasingam (ed.), *Religion and the New Atheism. A critical appraisal*. Leiden: Brill, 1–8.
- Burns, M. and Brügger, N. (eds) (2012). *Histories of Public Service Broadcasters on the Web*. New York: Peter Lang.
- Campbell, H. (2011). Internet and Religion. In C. Ess and M. Consalvo (eds), *The Handbook of Internet Studies*. Oxford: Blackwell.
- Copsey, N. (2003). Extremism on the net. The extreme right and the value of the Internet. In R. Gibson, P. Nixon and S. Ward (eds), *Political Parties and the Internet. Net Gain?* London: Routledge, 218–33.
- Copsey, N. and Macklin, G. (2011). The BNP and the media in contemporary Britain. In N. Copsey and G. Macklin (eds), *The British National Party. Contemporary perspectives*. London: Routledge, 81–102.
- De-la-Noy, M. (1990). *Michael Ramsey. A portrait*. London: Collins.
- Dorey, P. and Kelso, A. (2011). *House of Lords Reform since 1911. Must the Lords go?* Basingstoke: Palgrave.
- Goddard, A. (2013). *Rowan Williams. His legacy*. Oxford: Lion.
- Guardian (2008a). Giles Fraser, 'In an age of red-top fury, here is a hero' (12 February 2008). Retrieved 14 September 2010 from <http://www.theguardian.com/commentisfree/2008/feb/12/anglicanism.islam1>
- Guardian (2008b). Madeleine Bunting, 'A noble, reckless rebellion' (9 February 2008). Retrieved 14 September 2010 from <http://www.theguardian.com/commentisfree/2008/feb/09/religion.politics>
- Hastings, A. (1991). *Robert Runcie*. London: Mowbray.
- Internet Archive (2001). British National Party. Islam: the bloody track record! <http://web.archive.org/web/20011229150036/http://www.bnp.org.uk:80/article78.html>.
- Internet Archive (2003). Traditional image was the strength behind the church (Voice of Freedom. The Monthly Newspaper of the British National Party, no date) <http://web.archive.org/web/20031219060308/http://www.bnp.org.uk:80/freedom/church.html>
- Internet Archive (2006a). Diocese of York <https://web.archive.org/web/20061007041241/http://www.dioceseofyork.org.uk/archbishop.shtml>
- Internet Archive (2006b). British National Party: The cowardice of the Church, http://web.archive.org/web/20060427023333/http://www.bnp.org.uk:80/reg_showarticle.php?contentID=666
- Internet Archive (2007). BBC News: Archbishop makes Zimbabwe protest (9 December 2007), at https://web.archive.org/web/20071209183436/http://news.bbc.co.uk/2/hi/uk_news/7135087.stm
- Internet Archive (2008a). BBC News: In full: Rowan Williams interview, <https://web.archive.org/web/20080217194245/http://news.bbc.co.uk/1/hi/uk/7239283.stm>
- Internet Archive (2008b). Army Rumour Service: 'Williams is dangerous, he must be resisted' <https://web.archive.org/web/20080211102043/http://www.arrse.co.uk/cpgn2/index.php?name=Forums&file=viewtopic&t=88654#1779435>
- Internet Archive (2008c). On the Wrong Planet: 'Poor Rowan – he is doing his best' <https://web.archive.org/web/20080311081621/http://blog.atrevorsmith.co.uk/>
- Internet Archive (2008d). Tigra Networks: 'God bless Rowan Williams' <https://web.archive.org/web/20080214231017/http://community.tigranetworks.co.uk/>
- Internet Archive (2008e). British National Party: 'Archbishop of Canterbury: "Sharia law in Britain is unavoidable"' <http://web.archive.org/web/20080209140740/http://www.bnp.org.uk:80/2008/02/07/archbishop-of-canterbury-sharia-law-in-britain-is-unavoidable/>
- Internet Archive (2008f). British National Party: 'New Labour to disestablish the Church of England' <http://web.archive.org/web/20080118034524/http://www.bnp.org.uk:80/2008/01/10/new-labour-to-disestablish-the-church-of-england/>
- Internet Archive (2008g). British National Party: 'The betrayal of Charles Martel' <http://web.archive.org/web/20081217014333/http://bnp.org.uk:80/2008/12/the-betrayal-of-charles-martel-mosque-cornerstone-laid-in-tours/>
- Internet Archive (2008h). British National Party: 'Christian doctrine is offensive to Muslims' <http://web.archive.org/web/20080820110919/http://www.bnp.org.uk/2008/07/christian-doctrine-is-offensive-to-muslims-nick-griffin-video-response/>

- Jackson, P. (2010). Extremes of faith and nation: British Fascism and Christianity. *Religion Compass* 4: 507–27.
- Kearns, P. (2008). The end of blasphemy law. *Amicus Curiae* 76: 25–7.
- Marshall, R. (2004). *Hope the Archbishop: A Portrait*. London: Continuum.
- Rogers, R. (2013). *Digital Methods*. Cambridge, MA: MIT Press.
- Shortt, R. (2008). *Rowan's Rule. The biography of the archbishop*. London: Hodder.
- Thurlow, R. (1998). *Fascism in Britain. From Oswald Mosley's Blackshirts to the National Front*. Revised edition, London: I.B. Tauris.
- UK Web Archive (2015a). JISC UK Web Domain Dataset 1996–2013 Introduction. Retrieved 3 September 2015, from <http://data.webarchive.org.uk/opendata/ukwa.ds.2/>
- UK Web Archive (2015b). Geo-location in the 2014 UK Domain Crawl (24 July 2015), retrieved 3 September 2015 from <http://britishlibrary.typepad.co.uk/webarchive/2015/07/geo-location-in-the-2014-uk-domain-crawl.html>
- Webster, P. (2008). Rowan Williams and sharia, retrieved 7 September 2015 from <http://peterwebster.me/2008/03/01/rowan-williams-and-sharia/>
- Webster, P. (2015). *Archbishop Ramsey. The shape of the church*. Farnham: Ashgate.

Chapter 10

- Abdou, M. (2009). Anarca-Islam. (Unpublished Master's thesis). Queen's University, Kingston, Ontario, Canada.
- Andersen, A., Lingner, B., Ernst, N., Tadini, N. and Coelli, T. (2011). Looking for Cultural Space – Discourses of Identity Formation on the Case of Taqwacore. (Unpublished Masters' Project). Roskilde University, Denmark.
- Attolino, P. (2010). U-communities and the Taqwacores: towards the construction of a (neither) American (nor) Muslim identity. In M. Palander-Collin, P. S. M. Vesalainen, M. Nevala and H. Lenk (eds), *Constructing Identity in Interpersonal Communication*. Helsinki: Societe Neophilologique de Helsinki, 215–26.
- Bohlman, P. V. (2002). *World Music: A Very Short Introduction*. Oxford: Oxford University Press.
- Darrell, I. (1999). Straight edge subculture: Examining the youths' drug-free way. *Journal of Drug Issues* 29(2): 365–80.
- Davidson, A. J. (2011). *Punk Islam? Muslim Punk?: Taqwacore as a Multivalent Means Through which to Counteract a Monolithic Image of Islam*. Portland, OR: Reed College.
- Davies, M. (2005). Do it yourself: Punk and the disalienation of international relations. In M. I. Franklin (ed.), *Resounding International Relations: On Music, Culture, and Politics*. Palgrave Macmillan, 113–40.
- Duncombe, S. and Tremblay, M. (eds) (2011). *White Riot: Punk Rock and the Politics of Race*. New York: Verso Books.
- Feixa, C. (2006). Tribus Urbanas and Chavos Banda: Being punk in Catalonia and Mexico. In P. Nilan and C. Feixa (eds), *Global Youth? Hybrid Identities, Plural Worlds*. New York: Routledge, 149–66.
- Foot, K. (2006). Web sphere analysis and cybercultural studies. In D. Silver and A. Massanari (eds), *Critical Cyberculture Studies: Current Terrains, Future Directions*, New York: New York University.
- Furness, Z. (ed.) (2012). *Punkademics: The Basement Show in the Ivory Tower*. London: Minor Compositions.
- Gansauge, B. (2009). The punk and hardcore youth subcultures in the USA since the 1980s. (Unpublished Seminar Paper). Institute for English and American Studies-Technical Institute, Dresden, Germany.
- Hall, S. (2003). Cultural identity and diaspora. In J. E. Braziel and A. Mannur (eds), *Theorizing Diaspora: A Reader*. Malden, MA: Blackwell, 233–46.
- Hebdige, D. (1979). *Subculture: The Meaning of Style*. New York: Routledge.
- Hosman, S. S. (2009). Muslim punk rock in the United States: A social history of *The Taqwacores*. (Unpublished Master's thesis). The University of North Carolina at Greensboro, Greensboro, NC.
- Knight, M. M. (2004). *The Taqwacores*. Brooklyn, NY: Soft Skull Press.

- Knight, M. M. (2006). *Blue-Eyed Devil: A Road Odyssey Through Islamic America*. Brooklyn, NY: Soft Skull Press.
- Levine, N. (2008). *Dharma Punx: A Memoir Against the Stream*. San Francisco: HarperOne.
- Luhr, E. (2010). Punk, metal and American religions. *Religion Compass* 4(7): 443–51.
- Marcus, G. (1990). *Lipstick Traces: A Secret History of the Twentieth Century*. Cambridge, MA: Harvard University Press.
- Murthy, D. (2010). Muslim punks online: A diasporic Pakistani music subculture on the Internet. *South Asian Popular Culture* 8(2): 181–194.
- Nguyen, M. T. (2012). Afterword. In Z. Furness (ed.), *Punkademics: The Basement Show in the Ivory Tower*. New York: Autonomedia, 217–23.
- Nikpour, G. (2012). White riot: Another failure ... *Maximum Rocknroll* 345.
- Ortiz-Torres, R. (2012). Mexipunks. In Z. Furness (ed.), *Punkademics: The Basement Show in the Ivory Tower*. New York: Autonomedia, 187–202.
- Pollock, D. C. and van Reken, R. (2001). *Third Culture Kids: The Experience of Growing up among Worlds*. Boston, MA: Nicholas Brealey Publishing.
- Richards, C. (2008). *Forever Young: Essays on Young Adult Fictions*. New York: Peter Lang.
- Stewart, F. E. (2011). 'Punk Rock Is My Religion': An Exploration of Straight Edge punk as a Surrogate of Religion. (Unpublished doctoral dissertation). University of Stirling, Stirling, UK.
- Yulianto, W. (2011). Desacralization and critiques to Islamic Orthodoxy in Michael Muhammad Knight's *Taqwacores*. (Unpublished Master's thesis). University of Arkansas, Fayetteville, AR.

Chapter 11

- Aust, R. (2015). Online reactions to institutional crises: BBC Online and the aftermath of Jimmy Savile. Retrieved 19 April 2016, from <http://sas-space.sas.ac.uk/6100/>
- Big UK Domain Data for the Arts and Humanities (n.d.). Retrieved 19 April 2016, from <http://buddah.projects.history.ac.uk/about/aims-and-objectives/>
- Brügger, N. (2012). Web history and the web as a historical source. *Zeithistorische Forschungen* 9(2): 316–25.
- Cran, R. (2015). 'all writing is in fact cut ups': The UK Web Archive and Beat literature. Retrieved 19 April 2016, from <http://sas-space.sas.ac.uk/6101/>
- Deswarte, R. (2015). Revealing British Euroscepticism in the UK Web Domain and Archive Case Study. Retrieved 19 April 2016, from <http://sas-space.sas.ac.uk/6103/>
- Huc-Hepher, S. (2015). Searching for Home in the Historic Web: An Ethnosemiotic Study of London-French Habitus as Displayed in Blogs. Retrieved 19 April 2016, from <http://sas-space.sas.ac.uk/6252/>
- Jackson, A. (2016). Introducing SHINE 2.0 – A Historical Search Engine. UK Web Archive Blog. Retrieved 7 March 2016, from <http://britishlibrary.typepad.co.uk/webarchive/2016/02/updating-our-historical-search-service.html>
- Kahle, B. (1997). Preserving the internet. *Scientific American* 276(3): 82–3.
- Kay, A. (2015). Capture, commemoration and the citizen historian: digital shoebox archives relating to PoWs in the Second World War. Retrieved 19 April 2016 from <http://sas-space.sas.ac.uk/6248/>
- Millward, G. (2015). Digital barriers and the accessible web: disabled people, information and the internet. Retrieved 19 April 2016, from <http://sas-space.sas.ac.uk/6104/>
- Musso, M. (2015). A history of UK companies on the web. Retrieved 19 April 2016, from <http://sas-space.sas.ac.uk/6251/>
- Raffal, H. (2015). The Online Development of the Ministry of Defence (MoD) and Armed Forces. Retrieved 19 April 2016 from <http://sas-space.sas.ac.uk/6250/>
- Richardson, L. (2015). Looking for public archaeology in the web archives. Retrieved 19 April 2016 from <http://sas-space.sas.ac.uk/6249/>
- Schneider, S. M. and Foot, K. A. (2004). The web as an object of study. *New Media and Society* 6(1): 114–22.
- Taylor, H. (2015). Do online networks exist for the poetry community? Retrieved 19 April 2016, from <http://sas-space.sas.ac.uk/6105/>

Chapter 12

- Big UK Domain Data for the Arts and Humanities (n.d.). Retrieved 9 May 2016, from <http://buddah.projects.history.ac.uk/>
- Born Digital Big Data and Approaches for History and the Humanities (n.d.). Retrieved 25 May 2016, from <https://borndigitaldata.blogs.sas.ac.uk/>
- British Library home page (20 July 2009). Retrieved 10 July 2015, from <http://timetravel.mementoweb.org/reconstruct/20090720093000/http://www.bl.uk>
- Common Crawl (n.d.). Retrieved 13 May 2016, from <http://commoncrawl.org/>
- Davies, M. (2013). Corpus of Global Web-based English (GloWbE). Retrieved 13 May 2016, from <http://corpus.byu.edu/glowbe/>
- Depositing Websites and Web Pages (n.d.). Retrieved 25 May 2016, from <http://www.bl.uk/aboutus/legaldeposit/websites/websites/>
- Digging into Data (n.d.). Retrieved 13 May 2016, from <http://diggingintodata.org/>
- Digital Humanities 2016, Kraków 11–16 July (n.d.). Retrieved 9 May 2016, from <http://dh2016.adho.org/program/>
- Diplomatics (n.d.). Retrieved 9 May 2016, from <https://en.wikipedia.org/wiki/Diplomatics>
- Graham, S., Milligan, I. and Weingart, S. (2015). *Exploring Big Historical Data: The Historian's Macroscope*. London: Imperial College Press. Retrieved 13 May 2016, from <http://www.the-macroscopic.org/2.0/>
- Guldi, J. and Armitage, D. (2014). *The History Manifesto*. Cambridge: Cambridge University Press. Retrieved 9 May 2016, from <http://historymanifesto.cambridge.org/read/conclusion-public-future-past/>
- Hitchcock, T. (9 November 2014). Big data, small data and meaning. Retrieved 13 May 2016, from http://historyonics.blogspot.co.uk/2014/11/big-data-small-data-and-meaning_9.html
- Host Link Graph: Jisc UK Web Domain Dataset (1996–2010) (n.d.). DOI: 10.5259/ukwa.ds.2/host.linkage/1
- IHR-Info. Hypertext Internet Server (n.d.). Retrieved 13 May 2016, from <http://web.archive.org/web/19961227133909/http://ihr.sas.ac.uk/>
- Internet Archive – About (n.d.). Retrieved 9 May 2016, from <https://archive.org/about/>
- Jackson, A. (27 April 2015). Ten years of the UK web archive: what have we saved? Retrieved 13 May 2016, from <http://www.slideshare.net/andrewnjackson/ten-years-of-the-uk-web-archive-what-have-we-saved>
- John Johnson Collection of Printed Ephemera (n.d.). Retrieved 13 May 2016, from <http://www.bodleian.ox.ac.uk/johnson>
- Ketelaar, E. (2007). Archives in the digital age: new uses for an old science. *Archives & Social Studies: A Journal of Interdisciplinary Research* 1: 167–91.
- The National Archives (n.d.). *20-year rule*. Retrieved 9 May 2016, from <http://www.nationalarchives.gov.uk/about/our-role/plans-policies-performance-and-projects/our-projects/20-year-rule/>
- The National Archives (2016). *The Digital Landscape in Government 2014–15*. Kew: The National Archives. Retrieved 25 May 2016, from <http://www.nationalarchives.gov.uk/documents/digital-landscape-in-government-2014-15.pdf>
- Netarkivet (n.d.). Retrieved 13 May 2016, from <http://netarkivet.dk/in-english/>
- Newton, C. (n.d.). Life and death in the app store. *The Verge*. Retrieved 9 May 2016, from <http://www.theverge.com/2016/3/2/11140928/app-store-economy-apple-android-pixite-bankruptcy>
- Petitions: UK Government and Parliament (n.d.). Retrieved 13 May 2016, from <https://petition.parliament.uk/>
- Research Infrastructure for the Study of Archived Web materials (n.d.). Retrieved 25 May 2016, from <http://resaw.eu/>
- Segell, G. (1993). *Guide to IHR-Info: Hypertext Internet Server*. London: Institute of Historical Research.
- 'Shine' image search for 'cat' (n.d.). Retrieved 13 May 2016, from https://www.webarchive.org.uk/shine/search?query=cat&tab=results&action=search&facet.in.content_type_norm=%22image%22
- 'Shine' trend graph (n.d.). Retrieved 13 May 2016, from <https://www.webarchive.org.uk/shine/graph>

- Sweney, M. (5 May 2016). *The New Day newspaper to shut just two months after launch*. Retrieved 9 May 2016, from <http://www.theguardian.com/media/2016/may/04/new-day-newspaper-shut-two-months-launch-trinity-mirror>
- Text Encoding Initiative (n.d.). Retrieved 13 May 2016, from <http://www.tei-c.org/index.xml>
- UK Web Archive (n.d.). Retrieved 13 May 2016, from <http://www.webarchive.org.uk/ukwa/>
- UK Web Format Profile 1996–2010 (n.d.). Retrieved 13 May 2016, from <http://www.webarchive.org.uk/ukwa/visualisation/ukwa.ds.2/fmt>
- Voyant Tools (n.d.). Retrieved 13 May 2016, from <http://voyant-tools.org/>
- Warburg Institute Iconographic Database (n.d.). Retrieved 13 May 2016, from http://warburg.sas.ac.uk/vpc/VPC_search/main_page.php
- Webster, P. (29 June 2015). Why hoping private companies will just do the Right Thing doesn't work. Retrieved 25 May 2016, from <https://peterwebster.me/2015/06/29/why-hoping-private-companies-will-just-do-the-right-thing-doesnt-work/>

Index

- ABC (Australian Broadcasting Corporation), 106
- abortion drug, 160–1, 186
- abortion service, 161, 181, 183, 186, 188
- .ac.uk, 24, 29–34, 34–42, 43
- access, barriers to, 246
- ACM Web Science conference, 238
- ADHO (Alliance of Digital Humanities Organisations), 238
- Adobe 177
- advocacy, digital, 239
- Africa, 3, 7, 15
- Al-Thawra 215
- Alexa, 6
- algorithms, 99
- America Online, 175
- Apache Hive, 28, 43
- API (Application Program Interface), 189, 246
- apps, 245
- Arab Spring, 215–6
- ARC, 7
- Archbishop of Canterbury, 191, 193–202
- Archbishop of York, 198–9, 202
- archive, 170, 189
- Archive Team, 137, 140
 - Geocities Snapshot, 9, 10
- Archive-It, 9, 11
- Archivethe.Net, 9
- Armitage, David, 241
- Aust, Rowan, 221–2, 233
- Australia, 7
- authenticity, 205, 210, 212
- awards, role in community, 153, 154

- Bangemann, 118
- BBC (British Broadcasting Corporation)
 - Jimmy Savile scandal, 221–2, 233
 - news website, 102, 105, 107–15
 - online, 106–7
 - radio, 194
 - television, 19
- Beat literature, 222–3, 231
- Belgium, 8
- Berners-Lee, Tim, 6, 9, 86, 140
- Beverly Hills Internet, 138
- bias, 45, 56, 59–61
- Bibliothèque Nationale de France Web Archives, 8
- Birt, John, 106
- Bitly, 59
- blogosphere, 164, 188
- blog (weblog), 88, 89, 162–4, 188, 197, 205, 212–5, 218
- born digital big data, 247–8
- Brazil, 3
- British Library
 - homepage of, 240, 245
 - Internet Archive data, 28, 47, 52, 220
 - own crawls of UK websites, 59
 - SHINE, 192, 200
 - UK Web Archive, 1, 8, 10, 26, 191, 192, 200, 246
 - visualization, 242–3
- British National Party, 197, 199–202
- britishblogs.co.uk, 197
- broad crawl, 66–7, 73–8
- BUDDAH (Big UK Domain Data for the Arts and Humanities) project, 239, 247
- Burroughs, William, 223

- Canada, 6
- Carey, George (*See also* Archbishop of Canterbury), 194
- cat, image of, 243
- Caterpillar boots, image of, 243
- ccTLD (Country code Top-Level Domain), 63–7, 76–9, 192
- chav, 242
- China, 3
- Church of England, 190, 191, 193, 196, 201
- collective action, 163
- collective identity, 163
- Columbia University Libraries, 8
- comment threads, 201
- commerce, 186
- Common Crawl, 9, 246
- community leaders, 151, 152
- community, 137–8, 140–1, 143–9, 151–8
- Comparison cloud, 180, 182–7
- completeness, 47–8, 52–4, 58–61
- connected component, 173
- consumerism, digital, 241
- content analysis, 161, 164–5, 170, 188
- .co.uk, 29–34, 42
- coverage, of an individual website, 47–8, 52–4, 58–61
- coverage, web archives, 26
- Cran, Rona, 222–3, 231, 235
- crawl profile, 192

- credit crunch, 242
- Croatia, 7
- cyberbalkanization, 164
- data loss, 244–5
- data protection, 247–8
- data, open, 245, 246
- data, portability of, 246
- deduplication, policies on, 192, 195
- deleted content, 53, 55, 60
- Delicious, 58
- Denmark, 7, 8, 10, 63–80
- density, network, 173
- Deswarte, Richard, 223–4, 231, 232, 234–6
- development of the web, 62–80
- diaspora, 205
- digital dark age, 244
- digital humanities, 238, 239, 242, 246
- diplomatic, 240
- distribution, 55
- .dk registry, 67, 72–6
- domain name registry, 67, 72–6
- ephemerality, of data, 245–6
- European Union, 223
- event crawl, 66
- everyday life, 4
- evidence, survival of, 244
- Facebook
 - commercial service provider, 243
 - links to, 94
 - news, 98
 - screen shots of, 9
 - size, 3
 - social media, 160, 162–3, 174, 188, 245
 - Taqwacore 205, 210, 215, 218
- Farage, Nigel, 223
- Flickr, 162
- France, 3, 7, 118–23, 126–33
- France Télécom, 126
- gatekeeping, 101
- GeoCities, 137–58, 175
- geographic distance, 38–42
- Germany, 3
- Google, 1, 58, 163, 165–6, 168–71, 186–9, 235
- .gov.uk, 29–34
- gTLD (generic Top-Level Domain), 65
- guestbooks, role in community, 155, 156
- Guldi, Jo, 241
- Harvard University Library Web Archive
 - Collection Service, 8
- health 165, 180–2
- Heretrix, 7
- History Manifesto, the, 241
- historical method, 195–7, 202–3
- history of the web, 23
- Hitchcock, Tim, 241
- homophily, 164, 173
- Hope, David, 198 (*See also* Archbishop of York)
- HTML (HyperText Markup Language), 86, 141, 142, 152, 153
- Huc-Hepher, Saskia, 224–5, 231
- Human Rights Web Archive @ Columbia University, 9, 10
- hyperlink, 24, 25, 28, 34
- hyperlink network, 159, 161–5, 170–1, 174–5, 187–9
- Iceland, 7
- IIPC (International Internet Preservation Consortium), 7
- image analysis, 149, 150, 151
- inclusiveness, network, 173
- indegree, network 173–7
- India, 3
- information highway, 123–4
- information public good, 163
- information studies, 2
- Instagram, 243
- Institut National de l'Audiovisuel, 8
- Institute of Historical Research, University of London, 220, 244, 245
- institutional history, digital, 241
- Internet Archive, 1, 10, 26, 84, 107, 140, 239, 244–5
 - Archive-It, 9
 - biases of, 45–6
 - Danish web, 62–4, 66, 72–5, 78–80
 - establishing of, 6, 51
 - Geocities, 140, 158
 - UK web domain, 27–8, 191, 195, 197–8, 201, 220, 231
- Internet Memory Research, 9
- Issuercrawler, 187
- Issy-Les-Moulineaux, 125–6
- Japan, 3, 7
- JavaScript, 51, 60
- JISC (Joint Information Systems Committee), 52, 220
- JISC UK Web Domain Dataset, 191
 - Host Link Graph, 192, 193, 195, 196
- John Johnson Collection of Political Ephemera, 243
- Kay, Alison, 225–6, 231, 233, 236
- kernel density, 55 (Figure 2.5), 56 (Figure 2.6), 57 (Figure 2.7)
- Ketelaar, Eric, 240
- Knight, Michael Muhammed, 204, 205, 207, 209, 211–2, 215–6
- Kominas, the, 215–6, 218
- Korea, 7
- language development, 242
- latent content, 165, 177
- Latvia, 7
- law, religious, sharia law, 194, 196, 200, 202
- legal deposit legislation, UK, 246
- legal frameworks, 246
- legislation, health 160
- Library of Congress, 7, 8, 10, 11
- link density, 29, 36
- links, interpretation of, 193, 195–6
- Lippman, Walter, 101
- local news media, 93, 97
- London, 46, 49
- longitudinal analysis, 45–6, 58, 60

- macro-historical research, 241
- macroscope, historical, 241
- manifest content, 165, 177
- marriage equality, 188
- media studies, 2
- Memento API, 47
- Memento protocol, 240
- meta words, 170, 177, 180–5
- Meyer, Eric T., 247
- micro-historical research, 241
- migration online, 106
- Millward, Gareth, 226–7, 232, 236
- Mind (UK charity), 226
- Ministry of Defence, 228–9, 233
- Minitel, 118–121, 123–4, 126, 133
- mobile web, 98
- modelling, 60
- Mosaic, 86
- music catalogue, image of, 243
- Musso, Martha, 227–8
- MySpace, 205, 215

- n-grams, 242
- National Archives of the UK, the, 239
- national web archives, 24, 25
- national web, 62–80, 192
- neologisms, 242
- Ness of Brodgar, 229
- Netarkivet, 10, 11, 62–4, 66–7, 72–5, 78–80
- Netherlands, 8, 10
- network analysis, 242
- network centrality, 38
- New Zealand, 7
- newsgroup, 162, 163
- newspapers, 83
- Non-Print Legal Deposit, 192
- Norway, 7, 10

- one-sample t-test, 55, 56
- online religion, object of study, 191
- Open Director Project (DMOZ), 47, 58
- .org.uk, 29–34
- outdegree, network, 174, 177, 179
- Oxford English Dictionary, 242
- Oxford Internet Institute, University of Oxford, 220

- PageRank, 187
- PANDORA, 7, 10, 11
- Parliament (United Kingdom), House of Lords, 190, 193, 198
- Parthenay, 125
- personalization, 60
- petitioning, online, 246
- political party, 168, 188
- politician, 168, 188
- politics, 159, 160
- Portuguese Web Archive, the, 10, 11
- power law, 187
- probability sample, 47, 58–9
- provenance, 240
- publication, date of, 240
- Python, 51

- Raffal, Harry, 228–9
- Reddit, 216–7

- regular expressions, 51
- religion, 168, 180, 182–3
- religious leaders, relationship with news media, 193, 194, 195, 198, 202
- reviews, 48
- Rhizome's ArtBase, 8
- Richardson, Lorna, 229, 231, 233
- Robot Wisdom, 87
- Rogers, Richard, 193
- Royal National Institute of the Blind, 226–7
- Russia, 3

- sample, 47, 58–9, 165
- Savile, Jimmy, 221–2, 233
- scholarly editing, 240
- Scope (UK charity), 226
- search engine 165–6, 177, 181, 184, 187, 189
- search methodologies, 243–4
- Second World War, 225
- seed URLs, 192
- selective crawl, 66
- Sentamu, John, 198–9 (*See also* Archbishop of York)
- SHINE (web archive search interface), 192, 200, 242, 243
- sitemap, 51
- SixDegrees.com, 88
- social issue, 163, 185, 187–8
- social media, 160, 162, 174, 188–9, 241–2
- social network analysis, 159, 161–2
- social networking sites, 88, 98
- South America, 7
- spam, 168, 186
- Spanish, 3
- Stanford University Libraries, 8
- Stonehenge, 229
- Sweden, 7

- t-test, one-sample, 55, 56
- Taylor, Helen, 229–30, 233
- text analysis, 148, 163, 177, 242–3
- text content, 159, 161, 164–6, 168, 170–1, 182–3, 187–8
- Thirty Year Rule, 239
- top-level domains, 24–5
- topic drift, 171
- topic modelling, 148, 149
- tourism, 48
- travel, 48
- TripAdvisor, 45–61
- Twitter, 160, 162–3, 174, 188

- UCLA Library, 8
- UK Conservative Party, 1
- UK Government Web Archive, 8
- UK Web Archive, 1, 8, 10, 26, 191, 192, 200, 246
- UK web domain, 24–5, 27, 28, 29–33
- UK, 7–8
- Ukraine, 1–2
- United Kingdom Independence Party, 223
- United Kingdom, 7–8
- universities, 34–42, 58
- online, 247
- Russell Group, 35, 36, 38
- URL (Uniform resource locator), 85

- USA, 3
- user-generated content, 48–49
- Vallely, Paul, 195
- visualization, 164, 173, 180
- VOSON crawler, 170–1
- Voyant, 246
- Warburg Institute Iconographic Database, 243
- WARC, 7
- Warcbase, 137
- Wayback Machine, 11, 58, 84, 140, 157
- Web 1.0, 161–2, 165, 175, 177, 188–9
- Web 2.0, 162
- web archives, computational analysis of, 26
 - history of, 6–9
- web archiving strategies, 10, 66
- web archiving, 26, 27
- web crawler, 51, 165–6, 170–1, 187
- web crawls, 25
- web, Danish, 62–80
- web, development of the, 62–80
- web, history of the, 23
- web, use of, 2–6
- Wikipedia, 4
- Williams, Rowan, 194–7, 200–2 (*See also* Archbishop of Canterbury)
- Word cloud, 180–3
- word of the year, 242
- World Wide Web Consortium, 226
- Yahoo!, 137, 139, 140, 142, 145
- YouTube, 174, 175, 243
- zines, 204

'No other work as cohesively, clearly, forcefully and successfully argues for the Web's centrality in contemporary society and social science. While scholars of new media tend to turn their attention to the newest and latest new media phenomena, the Web is and will continue to be crucial to understanding online phenomena generally and, just as critically, providing a record of online discourse and events.'

– **Steve Jones**, *UIC Distinguished Professor of Communication, University of Illinois at Chicago*

The World Wide Web has now been in use for more than 20 years. From early browsers to today's principal source of information, entertainment and much else, the Web is an integral part of our daily lives, to the extent that some people believe 'if it's not online, it doesn't exist'. While this statement is not entirely true, it is becoming increasingly accurate, and reflects the Web's role as an indispensable treasure trove. It is curious, therefore, that historians and social scientists have thus far made little use of the Web to investigate historical patterns of culture and society, despite making good use of letters, novels, newspapers, radio and television programmes, and other pre-digital artefacts. This volume argues that now is the time to ask what we have learnt from the Web so far. The 12 chapters explore this topic from a number of interdisciplinary angles – through histories of national web spaces and case studies of different government and media domains – as well as an Introduction that provides an overview of this exciting new area of research.

Niels Brügger is Professor and Head of the Centre for Internet Studies and of the internet research infrastructure NetLab, Aarhus University.

Ralph Schroeder is Professor and Director of the Master's course in Social Science of the Internet at the Oxford Internet Institute, University of Oxford.

 **UCLPRESS**

Free open access versions available from
www.ucl.ac.uk/ucl-press

Cover design:
Liron Gilenberg

£40.00

