The Web as History

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

UCL PRESS

The Web as History

The Web as History

Using Web Archives to Understand the Past and the Present

Edited by

Niels Brügger and Ralph Schroeder

UCLPRESS

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

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

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

Acknowledgements

We would like to thank especially Lara Speicher at UCL Press for being a great help, and of course the authors of the volume. The Arts and Humanities Research Council funded project The Big UK domain data for the Humanities (BUDDAH) with which both editors were involved and which provided the initial impetus for the book. This project is also the basis of several chapters. We would also like to thank the School of Advanced Study, University of London, Aarhus University Research Foundation, and Webster Research and Consulting for contributing to open access publication.

Contents

List of figures ix
List of tables xii
List of contributors xiii

Introduction: The Web as History 1
Ralph Schroeder and Niels Brügger

PART ONE THE SIZE AND SHAPE OF WEB DOMAINS

1. Analysing the UK web domain and exploring 15 years of UK universities on the web 23
   Eric T. Meyer, Taha Yasseri, Scott A. Hale, Josh Cowls, Ralph Schroeder and Helen Margetts
2. Live *versus* archive: Comparing a web archive to a population of web pages 45
   Scott A. Hale, Grant Blank and Victoria D. Alexander
3. Exploring the domain names of the Danish web 62
   Niels Brügger, Ditte Laursen and Janne Nielsen

PART TWO MEDIA AND GOVERNMENT

4. The tumultuous history of news on the web 83
   Matthew S. Weber
5. International hyperlinks in online news media 101
   Josh Cowls and Jonathan Bright
6. From *far away* to *a click away*: The French state and public services in the 1990s 117
   Valérie Schafer

PART THREE CULTURAL AND POLITICAL HISTORIES

7. Welcome to the web: The online community of GeoCities during the early years of the World Wide Web 137
   Ian Milligan

8. Using the web to examine the evolution of the abortion debate in Australia, 2005–2015 159
   Robert Ackland and Ann Evans

9. Religious discourse in the archived web: Rowan Williams, Archbishop of Canterbury, and the sharia law controversy of 2008 190
   Peter Webster

10. 'Taqwacore is Dead. Long Live Taqwacore' or punk's not dead?: Studying the online evolution of the Islamic punk scene 204
    Meghan Dougherty

11. Cultures of the UK web 220
    Josh Cowls

12. Coda: Web archives for humanities research – some reflections 238
    Jane Winters

Notes 249
References 256
Index 275

List of figures

Figure 1.1	Number of nodes (third-level domains) within each second-level domain over time	30
Figure 1.2	Relative size of second-level domains in the .uk top-level domain over time	30
Figure 1.3	Number of within-SLD links per node in four .uk SLDs, 1996–2010	32
Figure 1.4	Links between four second-level domains	33
Figure 1.5	Network diagram of hyperlinks between universities	37
Figure 1.6	Spearman's rank correlation coefficients between university league table rankings and ten different network centrality measures for three years	39
Figure 1.7	University in-strength rankings compared to university league table rankings for 2010	40
Figure 1.8	Left: Raw hyperlink strength (S_{ij}) between universities versus geographical distance, and Right: Normalized hyperlink strength (σ_{ij}) between universities *versus* geographical distance	41
Figure 1.9	Maps of the UK universities under study for three years: 2000, 2005 and 2010	43
Figure 2.1	Cumulative number of reviews in the live dataset	53
Figure 2.2	Cumulative number of attractions in the live dataset by first appearance	53
Figure 2.3	The number of new London attractions added each month to the TripAdvisor website based on archived data and live data	54
Figure 2.4	The proportion of attractions stored in the archived dataset increased irregularly to around 24% of all attractions on the TripAdvisor website from 2007 to 2013 even as the overall number of attractions on TripAdvisor continued to grow	54

Figure 2.5	Distribution of reviews per attraction in the live dataset and the archived data	55
Figure 2.6	Distribution of star ratings in live dataset and the archived data	56
Figure 2.7	Distribution of attraction rankings in the live dataset and the archived data	57
Figure 3.1	Extract from the .dk domain name list	68
Figure 3.2	Number of .dk domains over time	69
Figure 3.3	Registered and disappearing .dk domain names over time	69
Figure 3.4	Relationship in 2012 between ownership and domains (anonymous registrants removed)	71
Figure 3.5	Number of .dk domains over time	72
Figure 3.6	Number of domains in the .dk registry list and in Netarkivet	73
Figure 3.7	Number of .dk domains in the .dk registry, Netarkivet, and the Internet Archive	74
Figure 3.8	Domain names in the Internet Archive not found in the .dk registry	75
Figure 4.1	Connections between newspapers and other websites on the web in 1999	90
Figure 4.2	Connections between newspapers and other websites on the web in 2005	91
Figure 4.3	New Jersey local news ecosystem, 2008	97
Figure 4.4	New Jersey local news ecosystem, 2012	97
Figure 5.1	Evolution of outlinks to top five country domains over time	110
Figure 5.2	Correlation between outlinks and mentions of a country in BBC News Online	112
Figure 6.1	Cyberi Homepage. Issy-les-Moulineaux	126
Figure 6.2	Homepage from the Strasbourg Board of Education website	130
Figure 6.3	Homepage from the Strasbourg Board of Education website	131
Figure 6.4	Homepage for the Strasbourg Board of Education, displaying links to one access page for each category of visitor	131
Figure 6.5	Page from the Strasbourg Board of Education website	132

Figure 7.1	The exploding size of GeoCities, 1995–1997	139
Figure 7.2	Relative frequency of keywords 'Community' and 'Neighborhood' in Lexis|Nexis database, 1995–2013	146
Figure 7.3	Montage of 5,690 images extracted from the EnchantedForest	150
Figure 7.4	Image borrowing in the EnchantedForest	150
Figure 7.5	Word cloud of all community leader pages, 1996–1997 over six crawls	153
Figure 7.6	Awards taken from a random assortment of websites	154
Figure 8.1	Hyperlink network of participants in abortion debate in Australia, 2005	174
Figure 8.2	Hyperlink network of participants in abortion debate in Australia, 2015	175
Figure 8.3	Word cloud (meta words) – pro-choice, 2005	180
Figure 8.4	Word cloud (meta words) – pro-life, 2005	181
Figure 8.5	Word cloud (meta words) – pro-choice, 2015	182
Figure 8.6	Word cloud (meta words) – pro-life, 2015	183
Figure 8.7	Comparison cloud (meta words) – 2005	184
Figure 8.8	Comparison cloud (meta words) – 2015	185
Figure 8.9	Comparison cloud (page words) – 2005	186
Figure 8.10	Comparison cloud (page words) – 2015	187

List of tables

Table 2.1	Categories of attractions on TripAdvisor in 2015	50
Table 2.2	Percentages in each attraction category in the live data and archived data	57
Table 3.1	Selection of broad crawls	67
Table 3.2	Number of .dk domains and .dk owners	70
Table 4.1	Network analysis of local New Jersey news websites, 2008–2012	95
Table 5.1	Descriptive statistics	111
Table 5.2	Linear regression model explaining amount of country news mentions on BBC online	113
Table 5.3	Linear regression model explaining amount of country outlinks on BBC online	115
Table 6.1	Evaluation of the navigation and user interface of state websites	128
Table 7.1	Topics in three selected GeoCities neighbourhoods	149
Table 8.1	Direction and manifestation of ties in online networks	163
Table 8.2	Composition of sites (abortion stance)	167
Table 8.3	Composition of sites (site type)	167
Table 8.4	Top-20 sites ranked by Google, 2005 and 2015	169
Table 8.5	Network statistics	172
Table 8.6	Top-20 sites by indegree (full network)	176
Table 8.7	Top-20 sites by indegree (participant subnetwork)	178
Table 8.8	Top-20 sites by outdegree (full network)	179
Table 11.1	Comparing strategies for web archive research	234

List of contributors

Robert Ackland is a Senior Fellow in the Research School of Social Sciences at the Australian National University (ANU). He gained his PhD in economics at the ANU, focusing on index number theory in the context of cross-country comparisons of income and inequality. Robert has been studying online social and organizational networks since the early 2000s and in 2005, he established the Virtual Observatory for the Study of Online Networks lab (http://vosonlab.net). He teaches in the ANU's Master of Social Research (Social Science of the Internet specialisation), and his book *Web Social Science: Concepts, Data and Tools for Social Scientists in the Digital Age* (SAGE) was published in July 2013.

Victoria D. Alexander (AB, Princeton; AM, PhD, Stanford) is Senior Lecturer of Arts Management at Goldsmiths, University of London. Her research falls in the intersection of sociology of the arts, visual culture, sociology of organizations and sociology of culture. She has studied the funding of art museums, the use of information technology in museums, cultural policy in comparative perspective, sociology of the arts, neighbourhoods and visual sociology. Her books include *Sociology of the Arts*; *Museums and Money*; *Art and the State* (co-authored) and *Art and the Challenge of Markets* (forthcoming, co-edited).

Grant Blank is the Survey Research Fellow at the Oxford Internet Institute, University of Oxford. He is a sociologist specializing in the political and social impact of computers and the internet, the digital divide, statistical and qualitative methods, and cultural sociology. He is currently working on a project asking how cultural hierarchies are constructed in online reviews of cultural attractions. His other project links sample survey data with census data to generate small area estimates of Internet use in Great Britain. He holds a PhD from the University of Chicago.

Jonathan Bright is a Research Fellow at the Oxford Internet Institute, University of Oxford. He is a political scientist specialising in political communication and computational social science (especially 'big data' approaches to the social sciences). His research concerns how people get information about politics, and how this process is changing in the internet era. He finished a PhD in political science at the European University Institute in 2012, and also holds a BSc in Computer Science from the University of Bristol.

Niels Brügger is Professor and head of the Centre for Internet Studies as well as of the internet research infrastructure NetLab, Aarhus University, Denmark. His research interests are web historiography, web archiving and media theory. Within these fields he has published monographs and a number of edited books as well as articles and book chapters. He is co-founder and Managing Editor of the newly founded international journal *Internet Histories: Digital Technology, Culture and Society* (Taylor & Francis/Routledge). Recent books and guest edited journals include *Web History* (ed., Peter Lang 2010), *Histories of Public Service Broadcasters on the Web* (co-edited with M. Burns, Peter Lang 2012) and *Web25*, themed issue of *New Media & Society*.

Josh Cowls is a graduate student and researcher in Comparative Media Studies at the Massachusetts Institute of Technology. Prior to joining MIT, Josh completed his MSc in Social Science of the Internet, and served as a research assistant at the Oxford Internet Institute. His work covers the impact of new technology and data on areas including political campaigns, academia and the media.

Meghan Dougherty (PhD, Communication, University of Washington, Seattle) is an Associate Professor of Digital Communication at Loyola University Chicago's School of Communication. She studies the preservation of web cultural heritage, research methods for web history, and web archiving as an emerging cyberinfrastructure for e-research. Before joining the faculty at Loyola, Dougherty was a researcher for Webarchivist.org. As a member of the Webarchivist team, Dougherty participated in a number of web archiving projects including the September 11 Web Archive, and the Web Campaigning Digital Supplement. She built Wayfinder, a personalizable research interface for web archives, as an addition to the Webarchivist suite of research tools. Her forthcoming book, *Virtual Digs*, on web archival research methodology is supported by University of Toronto Press.

Ann Evans gained her PhD in Demography at the Australian National University (ANU). She is currently a Fellow in the School of Demography and Associate Dean (Research) in the ANU College of Arts and Social Sciences. Ann's primary research interest lies in the area of family demography, and she undertakes research in the following areas: cohabitation, relationship formation and dissolution, fertility and contraception, young motherhood and transition to adulthood.

Scott A. Hale is a Senior Data Scientist at the Oxford Internet Institute, University of Oxford, and a Faculty Fellow at the Alan Turing Institute. His research spans the social and computational sciences and focuses on knowledge discovery, data mining and the visualization of human behaviour in three substantive areas: multilingualism and user experience, mobilization/collective action and human mobility.

Ditte Laursen, PhD, is Head of department, The Royal Library Denmark. Experienced in collection management, it governance and research and development. Her special interests include digital cultural heritage, digital humanities and digital research infrastructures. She is author or co-author of numerous publications on digital archives, social interaction in, around and across digital media, and users' engagement with archives, museums and libraries, all published in international peer-reviewed journals and anthologies.

Helen Margetts is Director of the Oxford Internet Institute, University of Oxford, where she is Professor of Society and the Internet, and a Fellow of Mansfield College. She is a political scientist specializing in digital government and internet-mediated collective action. She is co-author (with Patrick Dunleavy) of *Digital Era Governance: IT Corporations, the State and e-Government* (Oxford University Press, 2006, 2008) and (with Peter John, Scott Hale and Taha Yasseri) *Political Turbulence: How Social Media Shape Collective Action* (Princeton University Press, 2015).

Eric T. Meyer is Professor of Social Informatics and Director of Graduate Studies at the Oxford Internet Institute, where he has been on the faculty since 2007. Meyer's research focuses on the transition from analogue to digital technologies in research and knowledge creation across disciplines in the sciences, social sciences, arts and humanities. His research has included both qualitative and quantitative work with marine biologists, genetics researchers, physicists, digital humanities scholars, social scientists using big data, theatre artists, librarians and

organizations involved in computational approaches to research. He has authored many articles and, with Ralph Schroeder, the book *Knowledge Machines: Digital Transformations of the Sciences and Humanities* (MIT Press, 2015).

Ian Milligan is an Assistant Professor of digital and Canadian history at the University of Waterloo. He studies how historians can engage with web archives, by exploring the large files that underlie the Internet Archive's Wayback Machine. His Social Sciences and Humanities Research Council of Canada-funded work on web archives has appeared in the *International Journal of Humanities and Arts Computing*, the *Journal of the Canadian Historical Association and Social History/Histoire Sociale*, as well as several peer-reviewed conference papers. He is also a proponent of historians learning to develop computational skills, and to that end is a co-editor of the website ProgrammingHistorian.org.

Janne Nielsen is an Assistant Professor in Media Studies, and a board member of the Centre for Internet Studies, Aarhus University. She is part of the Danish research infrastructure project Digital Humanities Lab where she participates in both the research infrastructure for the study of internet materials, NetLab, and the research infrastructure for the study of audio and visual materials. She holds a PhD in Media Studies for her work on the historical use of cross media in the educational activities of the Danish Broadcasting Corporation (DR). Her research interests include media history, cross media, web historiography, and web archiving.

Valérie Schafer is a researcher at the French National Center for Scientific Research (Institute for Communication Sciences, CNRS/Paris-Sorbonne/UPMC). She specializes in history of computing and telecommunications. Her current research deals with the internet and web history and she leads the Web90 project funded by the French National Research Agency (ANR) and dedicated to the French Heritage, Memories and History of the Web in the 90s. She is the author of *La France en réseaux (années 1960–1980)* [France in Networks (1960–1980)] (2012) and co-authored with Benjamin Thierry, *Le Minitel, l'enfance numérique de la France* [The Minitel, the French Digital Childhood] (2012) and with Bernard Tuy *Dans les coulisses de l'Internet. RENATER, 20 ans de technologie, d'enseignement et de recherche* [On the Internet's Sidelines: RENATER, 20 Years of Technology, Teaching and Research] (2013).

Ralph Schroeder is Professor at the Oxford Internet Institute at the University of Oxford. He is director of its Master's degree in 'Social Science of the Internet'. Before coming to Oxford, he was Professor at Chalmers

University in Gothenburg, Sweden. His books include *Rethinking Science, Technology and Social Change* (Stanford University Press 2007), *Being there Together: Social Interaction in Virtual Environments* (Oxford University Press, 2010), and (with Eric Meyer) *Knowledge Machines: Digital Transformations of the Sciences and Humanities* (MIT Press, 2015).

Matthew Weber is an Assistant Professor in the School of Communication and Information, and Co-Director of Rutgers' NetSCI Network Science research lab. Matthew's research examines organizational change and adaptation, both internal and external, in response to new information communication technology. His recent work focuses on the transformation of the news media industry in the United States in reaction to new forms of media production. This includes a large-scale longitudinal study examining strategies employed by media organizations for disseminating news and information in online networks. He is also leading an initiative to provide researchers with access to the Internet Archive in order to study digital traces of organizational networks. Matthew utilizes mixed methods in his work, including social network analysis, archival research and interviews. Matthew received his PhD in 2010 from the Annenberg School of Journalism and Communication at the University of Southern California.

Peter Webster is an historian of contemporary Britain, with interests in the history of Christianity in late twentieth century Britain, particularly the relation of church, law and state. He has published widely on the place of religious debate in Parliament, inter-faith encounter and permissive law reform in the period since 1945. His study of Michael Ramsey, archbishop of Canterbury (1961–1974), was published by Ashgate in 2015. Much of his professional life has been spent at the interface between historical scholarship and digital technologies, with particular interests in digital history, web archiving and digital curation. Before founding Webster Research and Consulting, he was Web Archiving Engagement and Liaison Manager at the British Library.

Jane Winters is a Professor of Digital Humanities at the School of Advanced Study, University of London. Among her current and past research projects are British History Online, Connected Histories, Digging into Linked Parliamentary Data, Big UK Domain Data for the Arts and Humanities, and Traces through Time: Prosopography in Practice across Big Data. Her research interests include digital history, big (and born digital) data for humanities research, new models of peer review, digital scholarly editing, the use of social media in an academic context and open access publishing.

Taha Yasseri is a Research Fellow in Computational Social Science at the Oxford Internet Institute, a Faculty Fellow at the Alan Turing Institute for Data Science, and Research Fellow in Humanities and Social Sciences at Wolfson College, University of Oxford. He completed his PhD in Complex Systems Physics in 2010. Prior to coming to Oxford, he spent two years as a Postdoctoral Researcher at the Budapest University of Technology and Economics, working on the socio-physical aspects of the community of Wikipedia editors, focusing on conflict and editorial wars, along with Big Data analysis to understand human dynamics, language complexity, and popularity spread. Yasseri's main research interests are in human dynamics, social networks and collective behaviour.

Introduction: The web as history

Ralph Schroeder and Niels Brügger

The web as a reflection of society

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

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

Another example is the 2014 shooting down of a passenger plane over the Ukraine during the war between Russians and Ukrainians. A Russian claimed to have shot down a Ukrainian military plane on social media, a post which was then deleted but found later via the Internet Archive, as the *New York Times* (2014) reported. There was

an extensive investigation which subsequently determined who was responsible for this incident. The point of both examples is that accurate records matter, and this applies to the digital realm just as much as it did for paper records and many other sources of evidence about the past.

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

The web in context

Before we discuss web archives and how they can be used to study social change, it is important to discuss a topic that is not covered in this volume (and indeed, about which little is known to date); namely, how the web is used. After all, if web pages are going to tell us about changes in society, we also need to know who reads – or watches, or listens to – the web. Part of the difficulty is that the web is a new medium, but like the internet, it has not yet been adequately theorized as such. To recognize this point, it can simply be noted that research about the web partly falls within media studies, which is concerned with communication, but also partly within information science, which deals with how people seek information. There are many difficulties here which cannot be resolved in a short space, but we will indicate briefly what we know about web 'audiences' or 'consumers' of online information. This is important because how the web is received in society will ultimately be a necessary backdrop for understanding the social significance of the patterns in the information that can be found online.

A good place to start is by considering the extent to which the web is a single entity – or if its use reflects offline political or cultural or linguistic borders. This is an interesting question because it has often been claimed that the web is a unique medium insofar as it can be accessed from anywhere – unlike traditional media that are confined, for example, by national broadcasting regulations or by the reach of transmitters and the like. In other cases, most notably in China, it has been argued conversely that the government and its censorship regime ringfence the web, making it into a cultural resource whose reach is circumscribed by the state. Both ideas are misleading, as Taneja and Wu (2014) have shown: first, in a certain sense, access to the web in China is no less densely bounded off from the global web than is the case for other non-English speaking large clusters on the web. The way that Taneja and Wu arrive at this finding is by examining traffic to the top 1000 websites (which together receive more than 99% share of attention globally), and then grouping these into sites that receive shared attention. Shared attention is defined as: if someone clicks on one site, they also visit another (after controlling for the statistical chance of co-visiting). One possible reason for this finding is that in the case of China, apart from language, an active policy by the party-state has promoted a Chinese-centric web, as in other cases of state-driven information technology policies such as Korea's (see also chapter six by Schafer for the French case). But the Chinese web is not uniquely circumscribed by a wall of censorship, as some have argued; instead, it is mainly that Chinese citizens, like those of other nations, are primarily interested in content produced in China.

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

was approximately 2 million in 2002, a very small number, and that almost 75% of these were in English, which is spoken as a first language by far less than 1% of the African population.

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

Online information in everyday life

In addition to a bird's-eye perspective, we could also look at web uses from the ground up, how people use the web in everyday life. Such research on how people search for information, for example, is still thin on the ground (Rieh, 2004; Savolainen, 2008; Aspray and Hayes, 2011; Schroeder, 2014). A major issue that has not yet been resolved in media or communication studies is where to 'put' information seeking in general. A simple way to grasp this point is to ask: where did people seek information before the advent of the web, say, in the mid-1990s? (The same point could be raised, of course, in relation to Wikipedia, and search engine behaviour.) They might have consulted an offline encyclopaedia instead of Wikipedia, a travel agent instead of a travel website (one of the chapters in this volume is about TripAdvisor), an offline pamphlet instead of a blog and so on. Yet these 'media' were also not much studied. What makes the web different is that it contains all of this information, but also that none of these uses of the web is easily categorized within the study of offline behaviour or other digital media – or indeed the study of mass and interpersonal communication. Where these uses can be categorized is in the areas studied by information science, but

that is a discipline that barely overlaps with communication studies (and that deals mainly with educational, research and library searches). In any event, the web, in view of the fact that it is a large and accessible source of data and increasingly important in peoples' lives, is bound to grow as a topic of research.

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

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

even governments – to keep records or institutional memories of themselves, which are also in a state of flux.

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

Web archives and researchers

Against this background of uses of the web, we can now turn to how the web can be used as a resource for scholarship. After the first web page was published in 1991 by Tim Berners-Lee, the inventor of the web, it took some five years before large-scale attempts to preserve the online web were initiated. From the mid-1990s the landscape of web archives started to evolve slowly with a number of web archives being established aimed at preserving the cultural heritage.[1]

The landscape of the web of the past

Early attempts to archive material on the internet, including the web, were carried out in Canada in 1994–1995 (Brügger, 2011; Webster, 2017), but it was not until 1996 that the first major international initiative was launched, namely the Internet Archive. The Internet Archive was founded in 1996 by Brewster Kahle, who had made a considerable fortune as an internet entrepreneur. He established the Internet Archive as a non-profit organization, located in San Francisco and with the aim of preserving digital media, including the web. The Internet Archive began by creating a relatively small collection, namely the websites of the 1996 Presidential candidates (cf. Kimpton and Ubois, 2006: 202), but soon after initiated its broad web collections based on following hyperlinks. The Internet Archive collects that to which hyperlinks point, which is

why it is transnational by nature.[2] As of today the Internet Archive holds the world's largest collection of the preserved web from the past. It is also worth noting that the Internet Archive has established a priceless treasure trove, as well as being instrumental in promoting web archiving internationally. It has developed software that is widely used to collect web content (the web crawler software Heretrix), an archiving file format (ARC, and later WARC) and software to replay the archived web material (the Wayback Machine) (cf. Koerbin, 2017; Webster, 2017). Furthermore the Internet Archive has played an important role in the establishment, in 2003, of the International Internet Preservation Consortium (IIPC) that has since that time provided an important forum for debates, knowledge sharing and technical developments about web archiving.[3]

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

In many cases the national web archives have continuously developed their archiving remit as new legal frameworks were passed, allowing them to broaden their scope for collecting. For instance, the UK Web Archive started in 2005 as a collection of websites of leading UK institutions, based on selection criteria such as historical, social and cultural

significance. Since April 2013 the UK Web Archive has also been allowed to archive the whole of the UK web domain (as stated in The Legal Deposit Libraries (Non-Print Works) Regulations 2013, § 16). Hence, the UK Web Archive's highly selective collection of a limited number of websites has been expanded with the Legal Deposit collection's broad archiving of the entire national web domain.

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

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

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

Five other types of web collections can be mentioned to complete this outline of where to find the web of the past. First, a number of

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

Making the web of the past useful for scholars

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

choices involved in the collecting of the web as well as in making the archived web accessible and documenting it.

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

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

Once a scholar has access to the collections, the next question is in what form he or she will get access to the concrete material held in the

collection. Since 2001, the main form of access to web collections has in most cases been through the interface of the Wayback Machine. The Wayback provides a browser-based interface where the user has to insert the web address (URL) of the web page he or she wants to retrieve, and once this is done, the Wayback presents the web page in a manner close to how it looked when online.[4] From a researcher's point of view, seeing the web page close to how it looked in the past is obviously beneficial, but this approach comes with a number of drawbacks, most notably that the scholar has to know the exact web address to find the material, and if s/he wants to find more than one web page, all the relevant web addresses have to be inserted and searched manually, one by one. Therefore, a number of web archives such as PANDORA, the Portuguese Web Archive and the Danish Netarkivet have established full text search, which means that the search interface allows for searching all types of content in the entire archive, including the body text itself. As an intermediate solution between URL and full text search, some collections have full text search of metadata (e.g. the Library of Congress). But full text search also comes with a number of challenges, including how to present and possibly rank thousands, or even millions, of hits in a user-friendly and relevant way (parcelling the search results by year and top-level domain name such as .com, .gov etc. may help, but the challenge is still significant).[5]

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

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

archives only offer documentation either about their collections, or about individual websites if the collection is based on selective collecting and curation. But in many cases even this documentation is scarce.

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

Collaborations between web archives and scholars

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

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

However, within the last five years a shift has slowly emerged internationally – the first indication of which is the 2010 report *Researcher Engagement with Web Archives: State of the Art* (Dougherty et al., 2010). Web archives are now more likely to involve researchers in discussions of collection policies and access forms, and increasingly scholars are starting to discover this new resource with all its pitfalls and challenges. As the contributions to the present volume highlight, web archives may hold a valuable potential for novel research projects as well as for approaching well-known research topics from a new

perspective. To fully realize this potential, sustainable collaborations must be created to ensure common standards, as well as researcher tools aimed at the skilled and novice web researcher, including sophisticated search tools, basic analytical software, tools for the creation of sub-collections and for exportation of data, and possibly also a wide range of API access-forms. And such initiatives must be combined with training courses with a view to disseminating knowledge to larger research communities. A number of collaborations between web archives and researchers have been initiated in recent years, and these projects can still serve as the inspiration for future joint ventures, whether in time-limited research projects or in long-term sustainable fora within already existing organizations, such as national research infrastructures or a transnational association such as the IIPC.[7]

Building on existing literature

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

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

From the mid-2000s, publications started to reflect on some of the methodological challenges related to the scholarly use of the archived web (e.g. summarized in Brügger, 2011, 2012b), and in some cases were combined with empirical studies (e.g. Schneider and Foot, 2006). There are several books on general internet history (e.g. Naughton, 2012, 2015; Abbate, 2000; Poole, 2005; Goggin and McLelland, 2017) which

provide valuable insight into the history of the internet, though not as much about the history of the web. Empirical studies of the web exist (e.g. Gillies and Cailliau, 2000; Schneider and Foot, 2006; Banks, 2008; Brügger, 2010; Burns and Brügger, 2012; Salter and Murray, 2014), but this literature only partially examines the archived web.

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

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

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

Future research

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

In outlining these, it can be reiterated that the web itself is changing. Hence one question that must be asked is where the boundaries of the archived web lie: is all content on social media included? Or app content? No doubt many further additions to the web will emerge, and capturing these, as they increasingly displace other media, will be a challenge. Second, there is the question of macro- *versus* micro-, or

quantitative *versus* qualitative integration: how can we make sense of the relation between the global web, and individual sets of web pages pertaining to specific topics? Third, if the web is to be used as an indicator of historical and social or cultural change, there must be a way to understand how the web is used in everyday life: what information do we seek, and need, on a daily basis? Does the web shape, or is it shaped by, these needs? Finally, there are many ways to build on and extend the studies mentioned in this introduction: by examining the changing shape of the web as a whole (and the parts that have, and are still, disappearing or unarchived), or of national webs (and especially those parts of the world that have hitherto been neglected, like Africa), and the myriad subsets of pages, their coherence and disparateness, and the abundant materials that make up the web. These future topics also mean that there is much theoretical work to do: how can the findings from these studies be integrated into studies of other media? These areas of future research constitute wide and almost virgin territories for scholarship, and are bound to open many new directions, some of which are as yet difficult to foresee.

Overview of the chapters

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

In chapter two, Hale et al. take a different approach, drawing on the same source of national level data of the whole .uk domain. They

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

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

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

Using a different quantitative methodology, Cowls and Bright, in chapter five, also analyse the evolution of news, in this case the international links to and from the website of BBC News. This chapter ties to larger debates about whether, with the increasing globalization of news, news content is nevertheless biased to richer countries or countries which have other characteristics such as military conflict or economic

ties. Their findings are that less peaceful countries receive fewer outlinks from the BBC site, even when they receive more coverage, and that countries using the English language also receive more links. Again, there are implications for visibility of certain places or languages. Perhaps more importantly, the BBC news site has a very wide readership and is well-known around the world. Thus it will be necessary to study the spread of other online news sites with an international reach to complement this analysis of one media organization. Once this is done, we will have a powerful understanding of whether the shift online is leading to greater global interconnectedness – or rendering certain parts of the world even less visible.

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

Part three delves deeper into particular cultural phenomena. Milligan's chapter seven is about a virtual ghost town: GeoCities was among the first and largest online community spaces on the web, a thriving place where people put up web pages in the manner of creating a home in an online neighbourhood. Milligan describes how people presented themselves on their virtual homesteads and how they interacted with each other. This is a fascinating story, although GeoCities was abandoned – partly for commercial reasons but also partly because having an online webspace became much more commonplace and because the geographical metaphor became increasingly outdated. Another part of the story is how GeoCities was only preserved due to the efforts of some of its dedicated former inhabitant web archivists, reminiscent of offline organizations for the preservation of historical monuments. Again, we see that what almost disappeared from view is just as important as what remains, as with other historical artefacts.

Tracking how a particular issue has been represented on the web is yet another approach to using web archives. Ackland and Evans do this in chapter eight for the abortion debate in Australia by means of hyperlink analysis and text analysis. Among other outcomes, they find that the commercial sites offering abortion drugs became more prominent, partly because they had not been approved at the start of the decade examined by Ackland and Evans. Another finding is that while the pro-choice and pro-life sides were roughly equally prominent at the start of the analysis period, over the decade the pro-choice side became more visible (though also more diffuse), which may give a clue about the direction of public sentiment over this period. Finally, the authors discuss how they used Google to gauge visibility, which can be justified inasmuch as Google is by far the most widely used search engine in Australia (and a link can be made here to the study by Waller, mentioned in the section above on the web in context, which examined Australians' search behaviour using Google). However, as in other big data studies, the reliability of this source and its bias towards commercial websites may hamper its usefulness as a source for researchers.

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

Islam is refracted through a quite different lens in Dougherty's chapter ten, which follows the development of Islamic Punk in North America. This subcultural movement has already faded, so the web provides a major record of its rise and fall. Documenting the movement will be a valuable source for understanding the cultural norms of young Muslims in North America who, for a brief period, took to a particular

form of punk music to express their allegiance both to Islam and to a popular genre of music. The discussions within this subculture, over what it means to be Muslim and at the same time identify with music that saw itself as non-conformist, shed interesting light on how young Muslims saw themselves during a time when there was considerable controversy about the place of Islam outside of its homeland, and in the USA in particular. Dougherty also raises questions about where the most revealing documentation can be found, since many of the relevant discussions can be found not in the Internet Archives, but often on moderated lists such as Reddit, or on Wikipedia and blogs, as well as via various news sources.

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

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

PART ONE
THE SIZE AND SHAPE OF WEB DOMAINS

1

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

Eric T. Meyer, Taha Yasseri, Scott A. Hale, Josh Cowls,
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 Machine[1] 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) project[3] and then later the *Big Data: Demonstrating the Value of the UK Web Domain Dataset for Social Science Research* project[4] and the *Big UK Domain Data for the Arts and Humanities* project[5] 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.

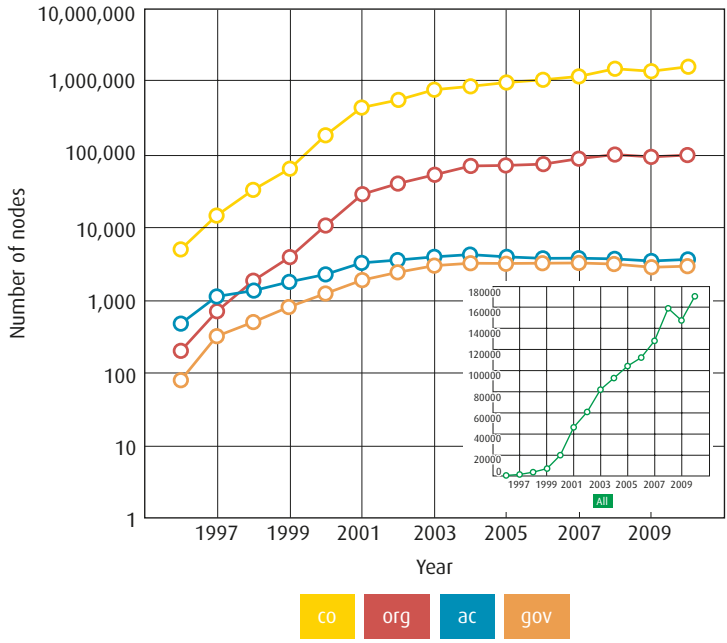

Figure 1.1 Number of nodes (third-level domains) within each second-level domain over time. The inset shows the sum over all second-level domains

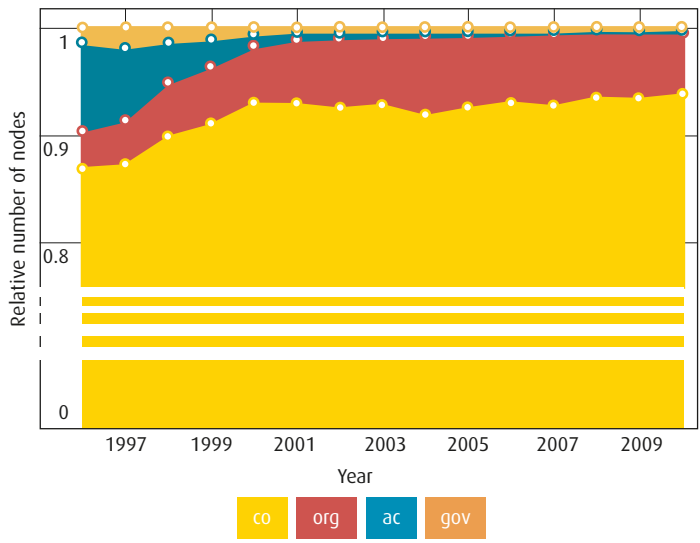

Figure 1.2 Relative size of second-level domains in the .uk top-level domain over time

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.

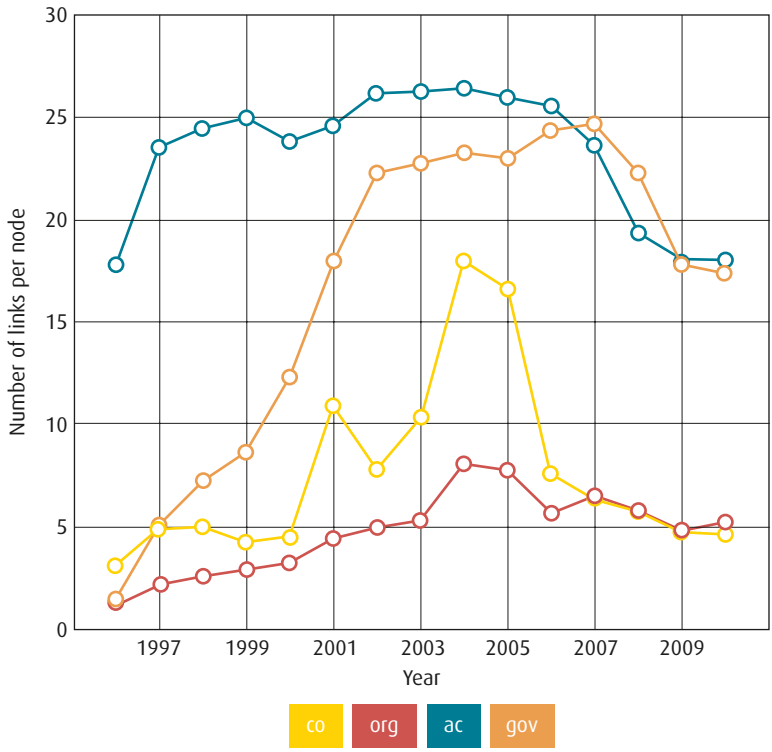

Figure 1.3 Number of within-SLD links per node in four .uk SLDs, 1996–2010

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.)

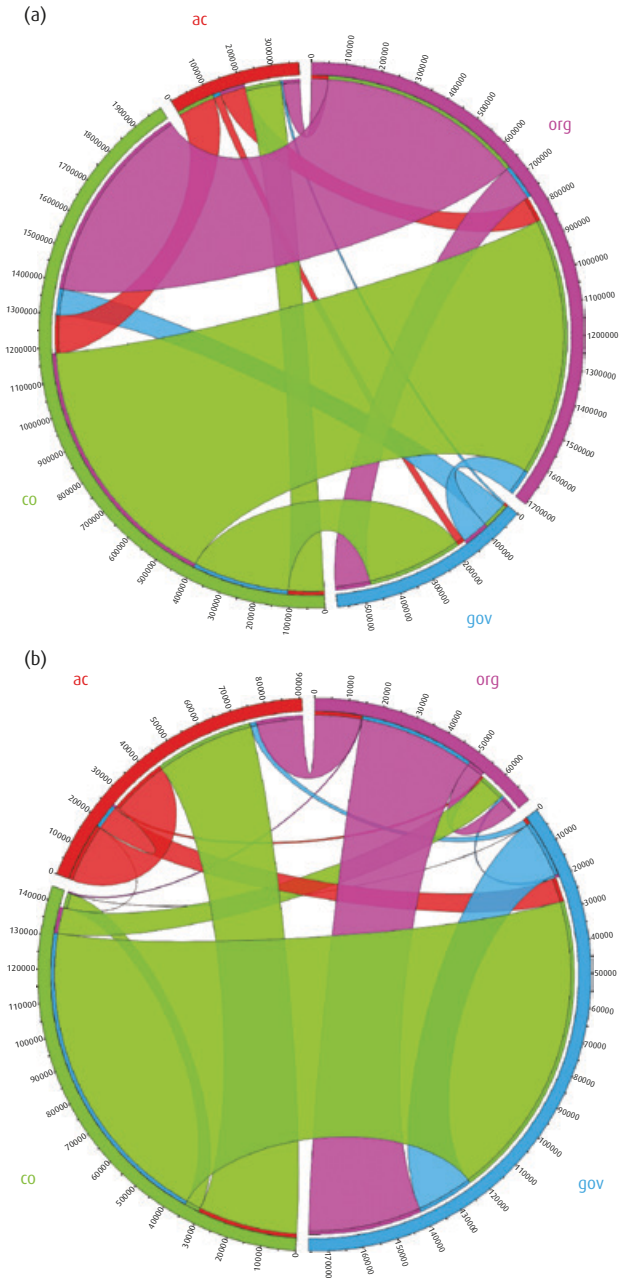

Figure 1.4 Links between four second-level domains. Panel *a* shows the absolute number of links between different SLDs (self-loops are excluded), and panel *b* shows the relative number of links normalized by the size of target subdomain

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 'linkworthiness' 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

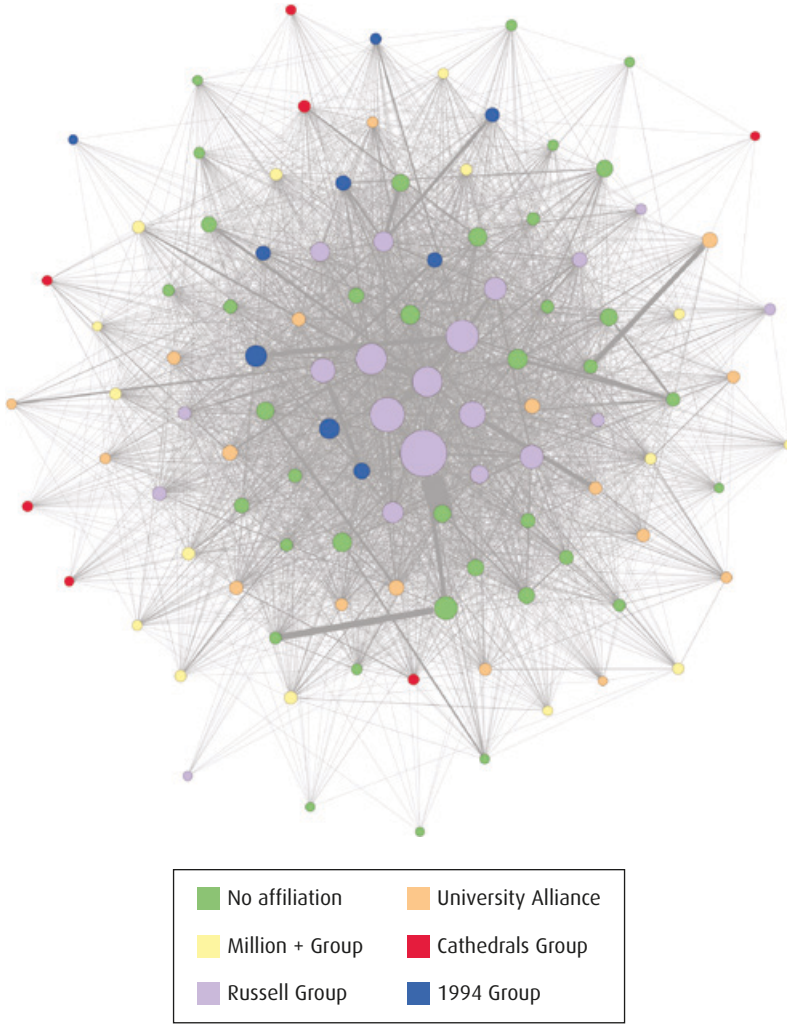

Figure 1.5 Network diagram of hyperlinks between universities. Different colours indicate different university affiliations

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,

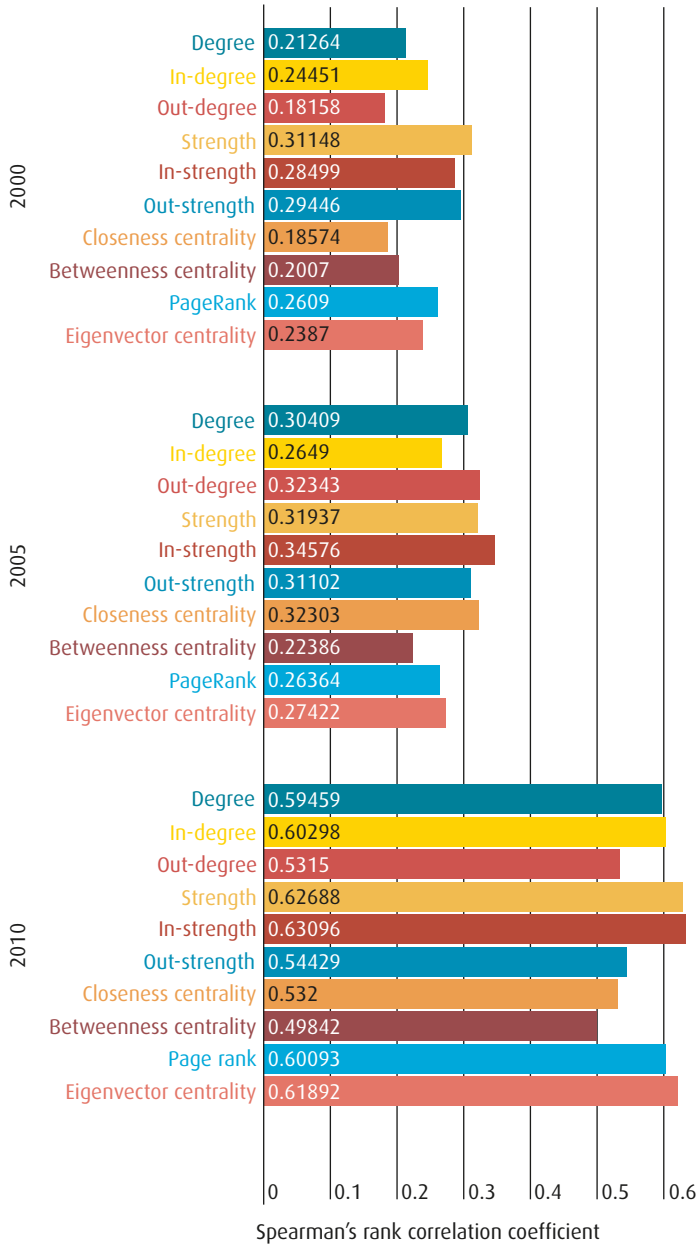

Figure 1.6 Spearman's rank correlation coefficients between university league table rankings and ten different network centrality measures for three years

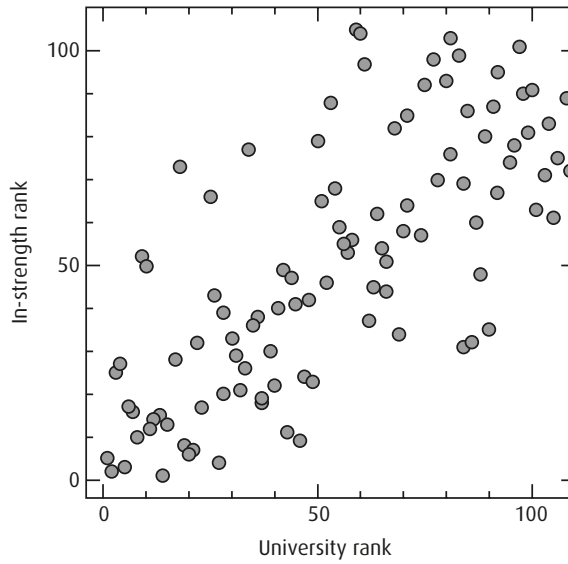

Figure 1.7 University in-strength rankings compared to university league table rankings for 2010. Spearman's rank correlation is 0.63

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

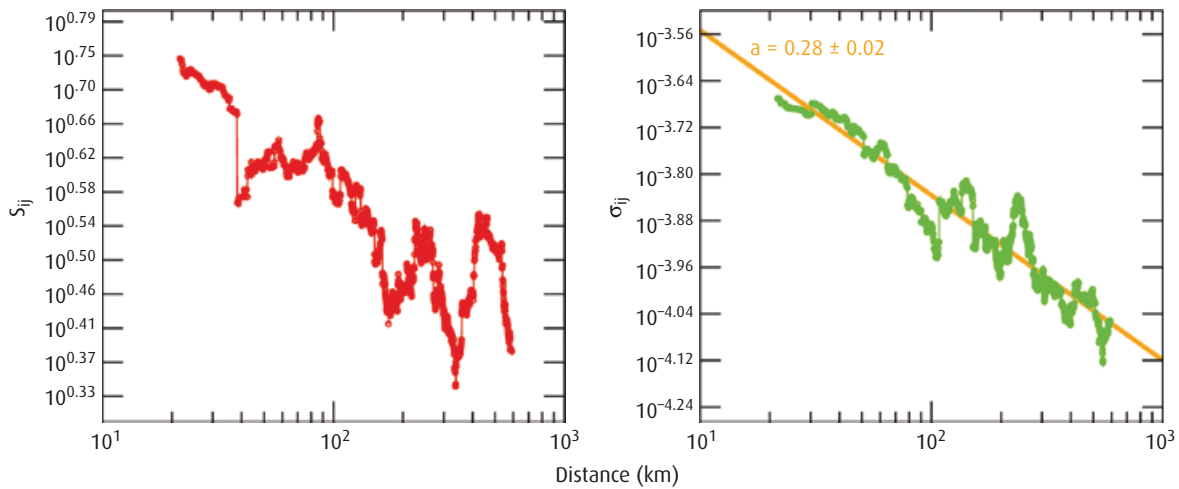

Figure 1.8 Left: Raw hyperlink strength (S_{ij}) between universities versus geographical distance. Right: Normalized hyperlink strength (σ_{ij}) between universities *versus* geographical distance. The normalized measure follows a gravity-law model with an exponent of a=0.28±0.02

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 nothing 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

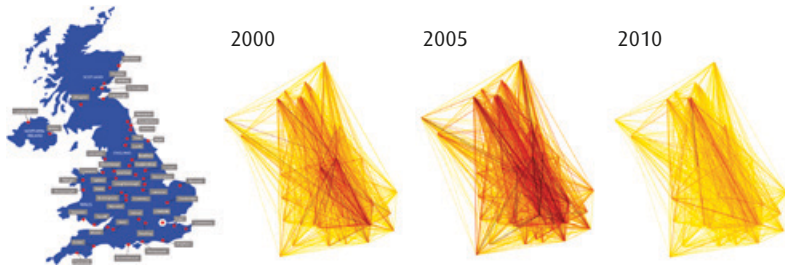

Figure 1.9 Maps of the UK universities under study for three years: 2000, 2005 and 2010. The connections are the hyperlinks and colour corresponds to the normalized strength of each link (σ_{ij}). The reddest links correspond to the strongest connections

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.

2

Live *versus* archive: Comparing a web archive to a population of web pages

Scott A. Hale, Grant Blank and Victoria D. Alexander

Introduction

With its seemingly limitless scope, the World Wide Web promises enormous advantages, along with enormous problems, to researchers who seek to use it as a source of data. Websites change continually and a high level of flux makes it challenging to capture a snapshot of the web, or even a cross-section of a small subset of the web. Web archives, such as those at the Internet Archive, promise to store and deliver repeated cross-sections of the web, offering the potential for longitudinal analysis. Whether this potential is realized depends on the extent to which the archive has fully captured the web. Therefore, a crucial question for Internet researchers is: 'How good are the archival data?'

We ask if there are systematic biases in the Internet Archive, using a case study to address this question. Specifically, we are interested in whether biases exist in the British websites stored in the Internet Archive data. We find that the Internet Archive contains a surprisingly small subset, about 24%, of the web pages of the website used for our case study (the travel site, TripAdvisor). Furthermore, the subset of data we found in the Internet Archive appears to be biased and is not a random sample of the web pages on the site. The archived data we examine has a bias toward prominent web pages. This bias could create serious problems for research using archived websites, and we discuss this issue at the end of the chapter.

The web has always been an extremely dynamic object. One widely quoted study found that 35–40% of web pages changed content in any

given week (Fetterly et al., 2004). Another study found that 26% of all web pages visited by users twice within an hour had changed content, and 69% of web pages revisited within a day had changed (Weinreich et al., 2008). For researchers interested in the evolution of the web or any part of the web (such as the diffusion of certain web technologies), this is a serious challenge. They need historical data, and almost all of this history is lost.

This problem was recognized early in the development of the web, and the Internet Archive was incorporated in 1996 by Bruce Gilliat and Brewster Kahle (Kimpton and Ubios, 2006). The goal of the Internet Archive is to collect digital data in danger of disappearing. There has never been any way to completely enumerate all web pages; so, all attempts to archive the web are to some extent incomplete. The general approach is to use a web crawler, a software program that starts with a list of Uniform Resource Locators (URLs) to visit (a seed list) and downloads a copy of the content at each of these URLs. Each downloaded web page is examined to find all the hyperlinks, which are then added to the list of URLs to be downloaded (subject to certain policies about how much content and what types of content to download). In this way, the software 'crawls' from page to page following hyperlinks somewhat like snowball sampling. Despite its best efforts the Internet Archive cannot collect everything. This leads to the question: How much of the web is archived?

In order to answer this question, we looked at two different collections of web pages, one that was collected and archived by the Internet Archive, and one that we collected ourselves. In this way, we are able to examine the completeness of the data that are held in the Internet Archive, at least with respect to our case study. To achieve this, we needed a case where we could reasonably find and download the full population of historical web pages. It is extremely difficult to find such a population since the Internet is constantly changing, and purposely collected archives are often the only source of historical web pages. We chose TripAdvisor as our case study as the website stores all reviews, including those written years ago, and thus allows us to reconstruct a historical population of web pages.

Our case study compares a full population of web pages from TripAdvisor with the subset stored by the Internet Archive. We defined our population as all tourist attractions in London listed on the TripAdvisor website. We downloaded these attractions from the current TripAdvisor site and found the earliest review of each

attraction. We call this data the 'live data', and compare it to Internet Archive data. The specific data we use for comparison are a copy of all the Internet Archive data for all web pages in the .uk country-code top-level domain from 1996 to 2013 that were copied to the British Library, which is where we obtained them. We refer to these data as the 'archived data' and note that they form a 'subset' rather than a 'sample' of the web because the Internet Archive does not claim to select a probability sample.

While others have looked at archive coverage in terms of web pages (URLs) generally, notably Ainsworth et al. (2013), this chapter is the first attempt to look at the extent of coverage of an individual website in depth. The remainder of this chapter is organized as follows. We review the existing literature comparing archived coverage to the web. We describe the Internet Archive and the source of our data before discussing TripAdvisor. We report our methodology and results and then turn to the implications of these results for research using web archival data.

Literature

Prior research on the success of web archiving is surprisingly sparse. Two studies, based on small subsets, address this issue. Thelwall and Vaughan (2004) studied differences in website coverage. They used randomly constructed names up to four letters long to find a total of 521 commercial websites related to four countries: the USA, Taiwan, China and Singapore and found large differences across the countries. They found that the Internet Archive in 2004 had at least one page stored for 92% of the US commercial websites, but had at least one page stored for only 58% of the Chinese commercial websites. Russell and Kane (2008) looked at web citations in history journals. They attempted to retrieve, from the Internet Archive, those citations that were no longer available on live websites. Only 57% of the citations not available online were retrievable from the Internet Archive.

Both of these studies examined only a small number of websites, and Russell and Kane's selection was not a random sample. The most complete study on the extent to which the web is archived is Ainsworth et al. (2013).[1] They sampled 1,000 URLs each from the Open Directory Project (DMOZ), the recent URLs bookmarked on the social bookmarking site Delicious, randomly created hash values from Bitly, and the Google search engine index. They used the Memento API (Van de

Sompel et al., 2009; Van de Sompel et al., 2010) to search 12 archives (including the Internet Archive) for each of the samples of 1,000 URLs and found that between 35% and 90% of the web was archived.

This is not a very satisfactory answer because it is such a wide range, but it broadly confirms the results from the smaller projects of Thelwall and Vaughan (2004) and Russell and Kane (2008). Large parts of the web are not included in any archive. A major weakness of these studies is a lack of detail about how much of each website has been archived. Thelwall and Vaughan (2004) counted a website as present in the archive as long as at least one page was archived. Ainsworth et al. (2013) and Russell and Kane (2008) looked at web pages (URLs) from many websites but did not examine how much of each site was in the archive. We address this gap by analysing how much of a website has been archived and whether the archived pages in the website differ in a systematic way from the population of all pages on the website.

There is a large literature on the use of Internet Archive data. However, this literature is less helpful to scholars than it could be, as it largely discusses what authors think should be possible without reference to the reality of what actually is possible (e.g. Arms et al., 2006; Weber, 2014). Our study uses a computational approach to assess what can actually be learnt from Internet Archive data.

Case selection

We study London attractions found on the travel website TripAdvisor (TripAdvisor.co.uk). TripAdvisor, according to its own strapline, is the 'world's largest travel website'. TripAdvisor (2014) cites Google Analytics as showing that it received an average of 315 million unique visitors each month in the third quarter of 2014. This figure shows the extraordinary importance of TripAdvisor in the travel business. It is therefore not surprising that most academic research on TripAdvisor is found in the tourism literature and focuses on hotel reviews. Previous studies tend to focus on practical issues such as how users decide how to trust reviews, the response of hotels to reviews, or the content of negative reviews and complaints (O'Connor, 2008; Cunningham et al., 2010; Sparks and Browning, 2010; Stringam and Gerdes, 2010; Ayeh et al., 2013). In contrast, our substantive interest, discussed elsewhere, is in how TripAdvisor works to convey cultural meanings. By studying reviews of cultural organizations, we examine the blurring of distinctions between

high and popular culture and between commercial and non-profit venues (Alexander et al., in preparation).

TripAdvisor displays user-generated reviews across categories such as hotels, restaurants and attractions. (Attractions encompass all elements of a city that are not restaurants or hotels.) Each review comprises a star rating, a title and a textual description. When starting a review, users enter the name of the hotel, restaurant or attraction, and if the target has been reviewed already, TripAdvisor suggests matches. Users can choose to review an item that already exists in TripAdvisor, or they can create an entry for a new, previously unreviewed establishment. For each review, users must choose a star rating, ranging from one star (negative) to five stars (positive). It is not possible for users to post reviews without choosing a star rating. Users then enter a short title or description in a free-form text box, and this serves as the title of their review. They then write the review itself, which can be as short or as long as they wish. TripAdvisor ranks hotels and attractions within categories based on their reviews using a proprietary method and these rankings may have a profound effect on the livelihood of hoteliers (Scott and Orlikowski, 2012). From our perspective, however, a crucial benefit of the reviews is that they provide a simple star rating combined with a more nuanced textual description. The star ratings allow an explicit comparison across different types of data, in this case, the archived data and our own live data.

We limited our live data to TripAdvisor's user-generated reviews of London attractions on TripAdvisor's UK site (tripadvisor.co.uk). This offers major advantages. London is a world-class metropolis with an enormous variety of attractions, providing us with a large range of reviews. Despite its size, however, London is still a bounded space so that our dataset can include the entire population of attractions and the entire population of reviews. Using TripAdvisor's UK site for London attractions makes it an appropriate vehicle for comparison to the archived data.[2]

At the time of data collection, the British Museum was the top attraction in London, and was described as '#1 of 1,277 things to do in London' (TripAdvisor, 2015). We have compiled a dataset of these attractions, as detailed in Table 2.1. This allows us to compare across datasets (live data versus archived data) on easily measured variables, such as number of attractions and reviews, the average star rating for each attraction, and the dates of reviews. Table 2.1 lists example attractions in each of TripAdvisor's top-level categories.

Table 2.1 Categories of attractions on TripAdvisor in 2015

Category	Number of attractions in category[a]	Example attractions
Amusement parks	3	The London Dungeon; Shrek's Adventure!
Boat tours & watersports	45	Canal and River Cruises Day Tours; Capital Pleasure Boats
Casinos & gambling	17	Hippodrome Casino; Kempton Park Racecourse
Classes & workshops	90	Hairy Goat Photography Tours; Bread Angels; East London Wine School
Food & drink[b]	120	Eating London Food Tours; Spice Monkey Cookery School
Fun & games	232	ClueQuest – The Live Escape Game; HintHunt; Secret Studio
Museums	280	Victoria and Albert Museum; National Gallery
Nature & parks	129	St James's Park; Thames River;
Nightlife	1231	City of London Distillery; Comedy Store London; The Cavern Freehouse
Outdoor activities	139	London Duck Tours; Moo Canoes Ltd.; Fishing London Coaching and Guide Service
Shopping	571	Covent Garden; Harrods
Sites & landmarks	519	Houses of Parliament; Big Ben
Spas & wellness	210	Pure Massage; The Body Retreat
Theatre & concerts	292	Les Miserables; Brick Lane Music Hall
Tours & activities	521	Alternative London Tours; BrakeAway Bike Tours; Shoreditch Street Art Tours
Transportation	67	London Tube; King's Cross Station
Traveller resources	30	Barbican Centre; City of London Information Centre
Zoos & aquariums	6	London Zoo

[a] Attractions often appear in more than one category; so, the total adds to more than the number of attractions in the dataset.

[b] The Food and Drink category does not include restaurants, but does include food and drink available in other attractions, such as a museum café, cookery school, or food-related tour.

Source: Data on categories and number of subtopics is from the live data on TripAdvisor. The number of attractions per category and examples are drawn from TripAdvisor (2015).

Data and methods

There are many technical issues to resolve in order to study web pages. We found all the London attraction pages on TripAdvisor had the form of 'Attraction_Review-.*-London_England.html' where '.*' indicates any (or no) characters. We used the sitemap files published by tripadvisor.co.uk that list all web pages on the site to create a complete list of all the attractions in London available on TripAdvisor for the current, live site and wrote a custom web crawler in Python3 to fetch the HTML of all the pages. Each attraction page had up to ten user reviews on it. For attractions with more than ten reviews, we downloaded all the additional pages of reviews.

We crafted regular expressions to extract the elements of the attractions and user reviews in which we were interested. For attractions, we extracted the following elements:

- the name of the attraction;
- the number of reviews for the attraction;
- the average star rating of the attraction;
- the category of the attraction as determined by TripAdvisor/its users;
- the ranking of the attraction among other attractions in London;
- the total number of 5-star, 4-star, 3-star, 2-star and 1-star reviews.

We also extracted the date that each review was added to each attraction. We performed all data collection in July 2015. Our final live dataset therefore contains all London attractions listed on TripAdvisor at that time and all available reviews to these attractions.

TripAdvisor, like many websites, does not include all content in the HTML of each web page, but loads some content separately using JavaScript. For TripAdvisor, the text of all user reviews is truncated in the HTML page and foreign-language reviews are not included at all. As the website still exists, we were able to emulate the JavaScript requests needed to collect the full text of reviews as well as foreign-language reviews for the live site but not for the archived data. Even so, within the live dataset, we were unable to collect 123 foreign-language reviews and hence our dataset contains 516,641 (99.98%) of the 516,764 reviews available in July 2015.

The Internet Archive is the oldest and biggest web archive, founded in 1996. A non-profit organization headquartered in San Francisco, it

was created to preserve a historical copy of the World Wide Web. The UK Joint Information Systems Committee (JISC, now 'Jisc', a third-sector, charitable body) commissioned the Internet Archive to extract all stored web pages within the .uk domain from its archives. These data were stored in a new data centre at the British Library and form the JISC UK Domain Dataset (UK Web Archive Open Data, n.d.). These Internet Archive data are the data we use within this chapter, and note that these data form the broadest dataset of UK domains available for the time period we study (1996–2013).[3] In partnership with the British Library, we extracted all TripAdvisor web pages stored in the archive with URLs matching 'Attraction_Review-.*-London_England.html'. The data include the HTML of the web pages as well as information about when the pages were added to the archive. We refer to these data simply as the archived data.

Results

Data overview

The earliest review in the live dataset was written on 26 August 2001, and the number of reviews on the site has been growing exponentially since that time (Figure 2.1, note that the vertical axis is a logarithmic scale).

TripAdvisor does not indicate when an attraction was first added to the website; we therefore take the date of the earliest review as a proxy for this measure. Measuring growth in this way, we found that the number of attractions on the website has also been growing each year (Figure 2.2, again note the logarithmic scale on the vertical axis).

The archived data contains 1,169 TripAdvisor web pages containing 340 unique attractions. The web pages of most attractions (57%) were only archived once, but some attractions were archived multiple times. The median number of copies was 1, the mean 3.4, and the maximum 31 (the most-archived attraction was 'Alternative London Tours').

The most recent data in the archived dataset are from 1 May 2013. Using the live dataset and the date of the first review for each attraction as a proxy for when that attraction was added to TripAdvisor, we estimate there were at least 1,406 attractions listed on the TripAdvisor website at that time. Thus, the 340 attractions covered in the archived dataset represent at most 24% of all the attractions available on the site at that time. This is the first indication of what proportion of the website

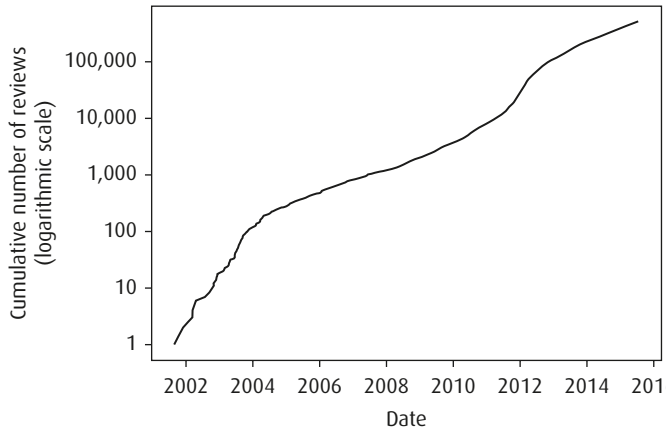

Figure 2.1 Cumulative number of reviews in the live dataset

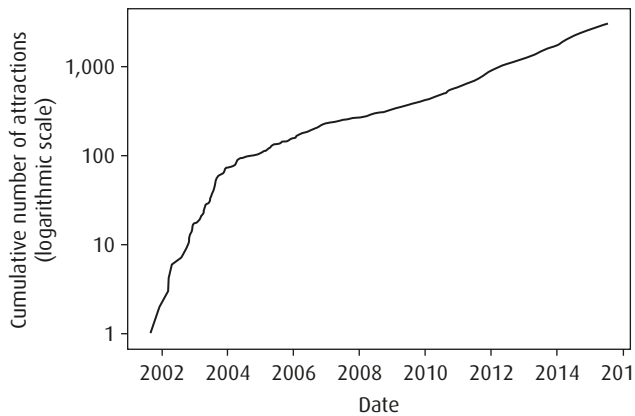

Figure 2.2 Cumulative number of attractions in the live dataset by first appearance. The date of the earliest review is used as the date the attraction first appeared on the site

is contained within the archived dataset. The top panel of Figure 2.3 shows the number of new attractions added to the archived dataset each month based on the date that the web page was crawled. The bottom panel of Figure 2.3 shows the number of new attractions added to the live website each month based on the date of the earliest review. Figure 2.4 shows the estimated proportion of attractions in the archived data compared to the live dataset.

The actual percentage of attractions stored in the archived dataset is probably lower as the live dataset does not include attractions

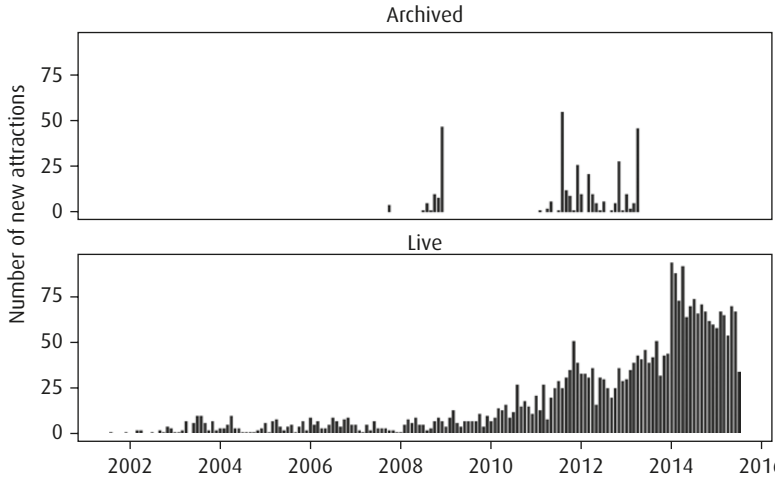

Figure 2.3 The number of new London attractions added each month to the TripAdvisor website based on the archived data and live data. For the archived data the date of a new attraction is the date that the webpage of the attraction was first crawled, while for the live data the date of a new attraction is the date of the oldest review for that attraction

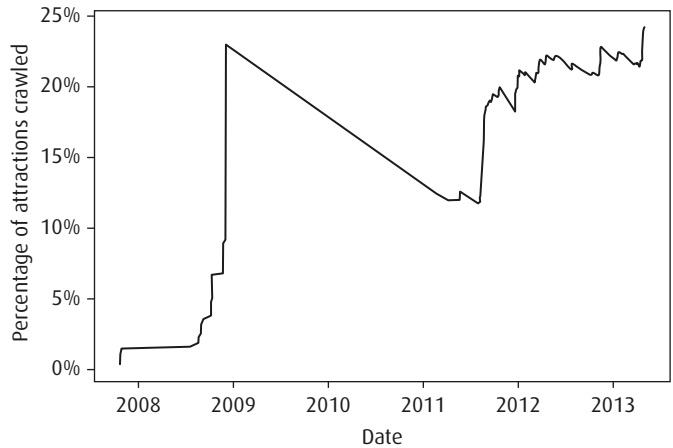

Figure 2.4 The proportion of attractions stored in the archived dataset increased irregularly to around 24% of all attractions on the TripAdvisor website from 2007 to 2013 even as the overall number of attractions on TripAdvisor continued to grow

that were on TripAdvisor but later removed. This appears to apply to 37 attractions in the archived dataset that do not appear in the live dataset. This means that there are actually 303 attractions in both the archived data and the live data. In addition, our numbers do not include the 734 attractions in the live data (8 of these are in the archived data) with no reviews and hence no proxy for when they were added.

Comparing the two datasets

We proceed by comparing the 303 attractions in both the archived dataset and the live site with the 1,409 attractions known to be on the live site at the last date of a new page being added to the archived data. We find that the attractions in the archived dataset differ significantly and are not representative of those on the live site.

Attractions within the archived dataset have a considerably different distribution of reviews per attraction than attractions in the live dataset. We demonstrate these differences using two statistical techniques.[4] Figure 2.5 shows the distribution of the number of reviews per attraction using a kernel density (note that the horizontal axis uses a logarithmic scale). Since the live data represents the actual population, we use a one-sample *t*-test, which shows that the mean number of reviews per attraction in the archived data differs significantly from the population mean ($t = 5.7, p < 0.001, N = 303$). The distribution of the archived

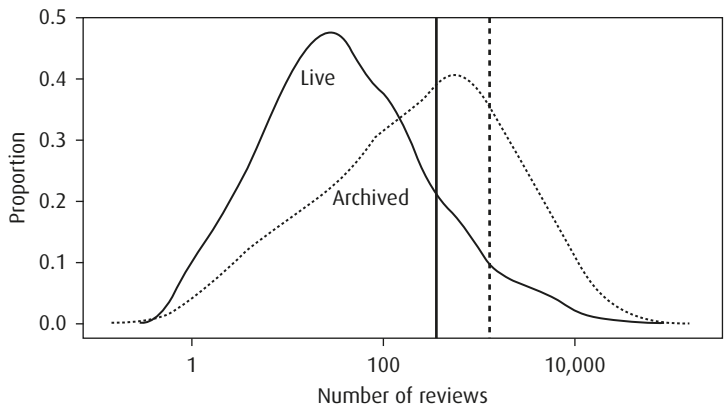

Figure 2.5 Distribution of reviews per attraction in the live dataset and the archived data. Vertical lines are means. Note that the horizontal axis uses a logarithmic scale

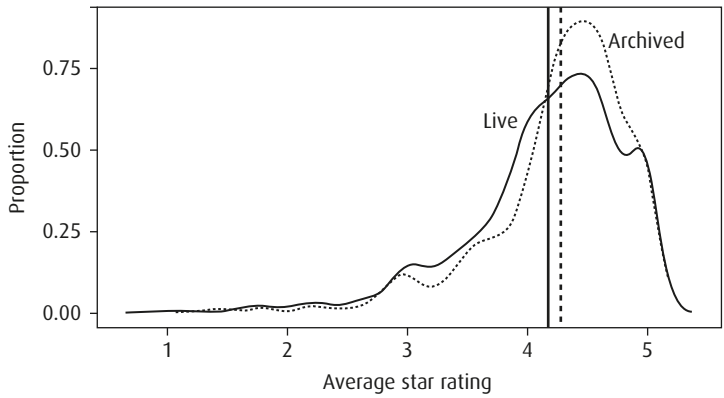

Figure 2.6 Distribution of star ratings in live dataset and the archived data. Vertical lines are means

data is skewed to the right; it contains attractions with 928 more reviews on average, probably an indication that the archived data have a bias towards more visible and prominent web pages. Figure 2.6 (also a kernel density, but with linear scales) shows that attractions in the archived dataset have higher average star ratings compared to attractions in the live dataset: an indication that the archived data tend to be biased toward more popular attractions. This difference is confirmed by a one-sample *t*-test ($t = 3.2$, $p = 0.002$, $N = 303$). Finally, Figure 2.7 (also a kernel density with linear scales) shows that attractions in the archived dataset tend to have a similar distribution of ranks. A one-sample *t*-test shows that the mean rank of attractions in the archived data does not differ significantly from the mean of the population, the live data ($t = –1.2$, $p = 0.22$, $N = 303$). The fact that one of the three measures of bias does not show a statistically significant difference is noteworthy; however, rankings are probably the least useful indicator because TripAdvisor reports attraction rankings within a number of different subcategories and the particular ranking criteria are not public.

Finally, in Table 2.2 we examine the percentage of attractions in each dataset in each of the 18 top-level categories on the current TripAdvisor website. Museums are most overrepresented in the archived dataset, 9 percentage points higher than in the live data. The archived data also include an excessive number of Tours and Activities (6.6 percentage points higher). Nightlife is the most underrepresented, 6.9 percentage points less in the archived data compared to the live data. If a researcher were interested in using the archived data as a proxy for attractions, these deviations could certainly bias results.

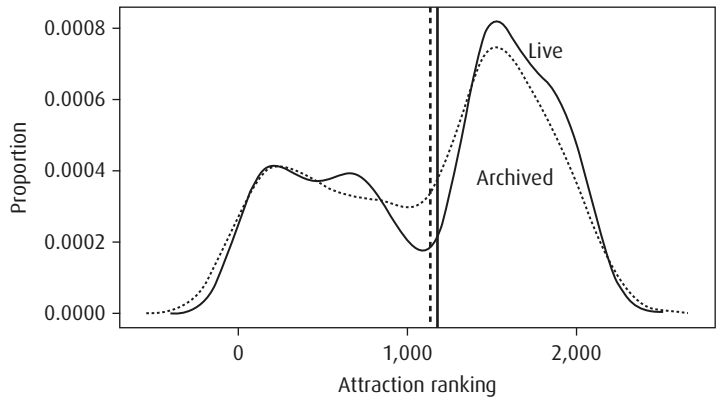

Figure 2.7 Distribution of attraction rankings in the live dataset and the archived data. Vertical lines are means

Table 2.2 Percentages in each attraction category in the live data and archived data

Category	Live data	Archived data	Difference
Amusement parks	0.1	0.4	0.3
Boat tours & water sports	1.5	2.3	0.8
Casinos & gambling	0.5	0.8	0.3
Classes & workshops	1.9	1.9	0.0
Food & drink	1.4	1.2	–0.3
Fun & games	5.8	5.0	–0.8
Museums	11.8	20.8	9.0
Nature & parks	5.6	5.8	0.2
Nightlife	18.1	11.2	–6.9
Outdoor activities	3.6	5.8	2.1
Shopping	15.3	12.3	–3.0
Sights & landmarks	22.0	24.2	2.2
Spas & wellness	4.0	0.8	–3.2
Theatre & concerts	11.2	12.7	1.5
Tours & activities	15.7	22.3	6.6
Transportation	0.7	1.9	1.2
Traveller resources	1.3	1.2	–0.1
Zoos & aquariums	0.3	1.2	0.9

Note: The percentages in the live data and the archived data add to more than 100% because some attractions are categorized in more than one category.

Discussion

Much has been promised for the use of web archives, and there have been a number of studies. For example, Chu et al. (2007) tracked the longitudinal development of site content on e-commerce websites. Mike Thelwall with various colleagues (Thelwall and Wilkinson, 2003; Vaughn and Thelwall, 2003; Payne and Thelwall, 2007) used web data to demonstrate the interdependence of academic institutions on the web. Hackett and Parmanto (2005) used the Internet Archive's Wayback Machine to analyse how technological advances were manifest in changes in website design over time. Hale et al. (2014) studied the evolution of the presence of British universities on the web using the same .uk web archive dataset that we used here.

The work with web archives has not been as extensive as the original founders anticipated, because, at least in part, there remain major challenges to using web archives. Scholars using the biggest archive, the Internet Archive, are mining data from a 9-petabyte dataset as of August 2014 (Internet Archive, 2014). Confronted with this enormous amount of data, few tools exist to help scholars find information. Furthermore, web pages are not well-structured or consistently structured, and they can be extremely difficult to transform into a format that can be used for large-scale quantitative research. In addition, changes in web page format and changes in content often occur simultaneously. This complicates longitudinal research because just getting the data into a consistent format may be difficult and slow. It may not be something that many scholars will want to invest in, given pressures to publish.

Once the data have been put into a consistent format what, exactly, do researchers have? This is the question we have addressed. First, researchers using web archive data have a subset of the full web. Using Ainsworth et al.'s (2013) estimates of web pages they might have between 35% and 90% of the web. By constructing their sample of URLs from DMOZ, Delicious, Bitly, and Google, Ainsworth et al. (2013) almost certainly examined the inclusion of more popular and prominent URLs (i.e. the URLs included in DMOZ or added to Delicious are by definition more popular and prominent than the URLs that no one adds to these platforms). We have avoided this bias by comparing archived data to the entire population of London attraction web pages on TripAdvisor. Although TripAdvisor is a prominent website, we still found that only 24% of the web pages about London attractions were archived.

This suggests that previous results are dramatic overestimates of the amount of the web that has been stored in archives. Our findings also complement the results from previous studies that have examined the percentage of websites included in web archives (e.g. Thelwall and Vaughan, 2004). Whereas these studies looked at the inclusion of at least one page of a website in the archive, we looked deeper into the site itself at whether web pages within the site are stored. Even though the TripAdvisor site itself is included in our archived data, only at most 24% of the pages about London attractions have been stored. This may also suggest that there are enormous variations in the archival coverage, and the simple presence of one web page from a website in the archive does not provide an indication of how much of that website is actually within the archive.

We also found that the archived pages do not resemble a random probability sample. There is a clear bias toward prominent, well-known and highly-rated web pages. Smaller, less well-known and lower-rated web pages are less likely to be archived. It is worth noting that all the archived data we used came from the Internet Archive; so, the archived data are probably the best, most complete source possible for this time period but it is clearly not complete, and it contains significant biases. In 2014, the British Library began conducting its own crawls of UK websites, but the representativeness and completeness of these data are yet to be determined.

What are the implications of these results for research using web archives? Much of the appeal of the Internet is that it seems to provide broader data than conventional sources. Advocates talk about it being unrestricted in scale or geographic scope. One reason web archives were seen as valuable was because they promised to provide full historical data on things such as diffusion of innovations, community formation, emergence of issues and the formation and dynamics of networks (Arms et al., 2006). The Internet is certainly broader than most conventional data sources, but the web archive we examined is broader in a certain way. It focuses on the big and the prominent. Due to the limits on the number of pages found and crawled from any one website, web archives are necessarily incomplete even when they start with a seed list of all domain names (as is now the case for the British Library crawls of the .uk country-code top-level domain). In some instances the limit on the number of pages for each website is relatively high – as is the case of the national web archive in Denmark (see Brügger, 2017) – but it remains difficult to assess what content is not archived (as archiving strategies change over time and technical issues in capturing

dynamic/JavaScript content arise). Therefore, a web archive-based study of diffusion of innovation on the Internet would actually be a study of diffusion among prominent, highly-rated web pages, not among all web pages. A study of network formation or network dynamics would be a study of networks of well-known, highly-rated web pages. It would not be a study of diffusion among all web pages. Hale et al.'s (2014) study of British university websites, for instance, is a study biased toward hyperlinks on more prominent web pages.

The incomplete nature of web archives limits the type of analyses available to researchers. We were only able to conduct our analysis, for instance, at the level of attractions in London and not about the content of reviews: the archived data are so incomplete with reference to review text that it did not make sense to even attempt such a comparison. These problems are only getting worse as content moves off the web to other channels (e.g. mobile apps), personalization means there is no definitive version and dynamic sites use JavaScript or other technologies to fetch content separately from the HTML pages.

The promise raised by Arms et al. (2006) was that web archives would eliminate the need to proactively collect data for longitudinal studies of networks, innovations, community formation, etc., and instead allow for fine-grained, retrospective analyses over longer periods of time. Web archive data can certainly provide insights that would otherwise be unavailable (e.g. we were able to find attractions that had been deleted from TripAdvisor in the archive that were unavailable on the live site). With suitable modelling, networks of hyperlinks from web archive data may be compared to null model controls. However, our study highlights that web archive data does not replace the need to collect specific data proactively over set periods of time for many types of longitudinal analysis. The level of incompleteness of web archive data also raises questions about the extent to which archived web data can be used to conduct longitudinal research at all. An approach that would yield much higher quality data is the same as we might have used for pre-Internet longitudinal data. That is, collect repeated cross-sectional datasets proactively in real time and then do retrospective, time-series analyses of the data only at the end of the study period. The irony is striking, but the point is that web archives do not provide a free lunch to good research.

These are serious problems. Web archives are an extensive and permanent record, but they are also an incomplete and biased record. While it is certainly possible to analyse larger numbers of many things,

are large, biased numbers a good idea? The answer is that a biased set of data remains biased no matter how many cases it contains and biased datasets provide biased answers regardless of their sizes. So researchers have to confront the bias problem. Web archives do not contain a complete population, except perhaps in certain limited areas, and what is missing from the archives is often unknown.

3

Exploring the domain names of the Danish web

Niels Brügger, Ditte Laursen and Janne Nielsen

Introduction

What does an entire national web domain look like? And how can its development over time be understood? Using the Danish web as our case study, this chapter explores these questions by studying the historical development of the .dk domain names and the .dk domains archived in the Danish national web archive, Netarkivet, as well as in the international US-based web archive Internet Archive. The analysis is a first step in a larger study of the development of the Danish web. This chapter will also address the broad questions above by combining different sources and developing methods that access and analyse materials from the archives in different ways.

An entire national web domain is something that we never experience as such when browsing the web, but nevertheless it is always there as a horizon, as the national context of our browsing. Therefore we need to understand national web domains not only to grasp the national web in its entirety, but also to allow in-depth analyses of web activities within the boundaries of the nation. Large scale analyses of the development of a national web may also be used to shed light on the nation's life outside the web, by comparing outgoing links from the national web domain with migration, immigration, travelling and trade.

Studies of a national web domain inevitably move from the close and detailed reading of individual web elements such as images, web pages or websites to what the literary scholar Franco Moretti calls 'distant reading'. This refers to a reading that zooms out from the individual document to encompass a vast amount of texts (Moretti, 2000). The aim

of a distant reading is to identify systems, structures, patterns and tendencies that transcend the individual texts, at the expense of complete knowledge about each entity in the mass of texts.

The historical study of an entire national web is a rather new field, and only few articles about national web studies exist. Some of the studies focus on national webs at a given point in time and use material archived by the scholars themselves, in contrast to material in web archives (Rogers et al., 2013; Ben-David, 2014, 2016). Clearly historical studies exist, some of which are based on the archived web, but are limited to studying hyperlink networks (e.g. Hale et al., 2014). One study, based on Yugoslavia (.yu, deleted from the internet in 2010) has investigated how the history of an entire country code top-level domain (ccTLD) can be reconstructed, based on URL-lists and material in the Internet Archive (Ben-David, 2016).[1] Hence, best practice is only slowly emerging. In most cases theories as well as methods and the selected source material have to be developed as the research progresses.

This is also the case with the study described in this chapter. It is part of a larger research project 'Probing a nation's web sphere – the historical development of the Danish web'.[2] The aim of the project is to analyse the development of the Danish national web from 2005 to 2015 as it has been archived in the Danish national web archive, Netarkivet. As part of the project we are developing methods and tools to delimit what constitutes 'the national web' at a given point in time. This is necessary because Netarkivet holds several versions of the same online web entity, even within a limited time span. It is therefore imperative to create a smaller collection from the entire web archive, in other words: a corpus.[3] Once the corpus is in place we will perform detailed analyses of the following five focal points: (1) size (size of the entire web domain, of file types and of websites), (2) space (geographical distribution of websites), (3) structure (networks of hyperlinks), (4) aliveness (new/disappeared domain names and frequency of updating), and (5) content (file and software types, language, and semantics, e.g. word frequencies, sentiment analysis/topic modelling). Due to technical limitations it has not yet been possible to perform these planned analyses.

However, the technical challenges have highlighted another way to approach the development of the national Danish web, namely to study the development of the domain names which constitute the Danish web. Lists of all the registered Danish domain names, year by year, can be found in Netarkivet as they were used as the so-called seed-list which was loaded into the web crawler to tell it what to archive. Later in the project we will be studying the archived content itself. First,

we will analyse the archive's meta-content via the list of registered domain names.

Therefore, the aim of the chapter is threefold: first, we investigate how the list of domain names can be studied as a historical source in its own right. Second, we present the results of what the domain names can tell us about the development of the Danish web, and compare the domain name lists to the number of domains that have been archived in Netarkivet and the Internet Archive. Third, we discuss how the results of this study may be used as part of a broader analysis of the development of the Danish web as it was archived by Netarkivet.

Studying the development of a national web domain

When setting out to study the historical development of an entire national web domain, a number of sources may be relevant, from user statistics and texts in news media to oral history accounts, as well as preserved copies of the web of the past. An example of a research project based on a great variety of sources is the French 'Web90 – Patrimoine, Mémoires et Histoire du Web dans les années 1990' (http://web90.hypotheses.org) which studies the development of the web in France in the 1990s. In contrast to a wide range of sources, Hale et al.'s (2014) longitudinal analysis of the UK national web domain, .uk, was based exclusively on the archived web, that is, the web as archived by the Internet Archive. However, one highly relevant source is often overlooked, namely the domain names allocated to a given nation. In order to get an impression of the size and development of a national web domain, access to comprehensive lists of all existing domain names from different points in time would be extremely valuable.

Domain names as a historical source

An inherent question in any study of a nation's web domain is where the national web starts and ends on the global web. The simple answer is that the national web is any web activity related to the nation state in question. However, operationalizing this answer is both easy and a challenge. It is easy since the web comes with its own institutionalized national delimitations, namely the system of ccTLD domain names such as .uk, .dk, .fr. It is fair to say that whatever activity takes place on a ccTLD is related to the nation state in question, thus forming a national domain name space that we can define as 'the national web'. But this

approach can also present a challenge. On the one hand, there may very well exist web material related to a given nation state outside of the ccTLD on other ccTLDs or on generic top-level domains (gTLD) such as .com, .org or .net. Identifying this material can be very time consuming, if it is possible at all.[4] On the other hand, not all nation states can be identified exhaustively by a ccTLD, most notably the USA. There is a national ccTLD, .us, but the vast majority of US related material is found on gTLDs.

Nevertheless, the institutionalized national delimitation mirrored in the ccTLD constitutes an appropriate first step in identifying a national web, or as Ben-David (2016) puts it, the domain name system is 'the Internet's most strict authenticator of nation-states'. The official national lists of domain names are managed by a national organization. The management of a ccTLD is delegated by the global domain name registrar ICANN (Internet Corporation for Assigned Names and Numbers), such as Nominet in the UK, DK Hostmaster in Denmark, and AFNIC in France. These registrars handle the internet's address system within each of the two-letter suffixes for countries and territories such as .uk, .dk, or .fr. Since the lists of ccTLD domain names provided by these organizations constitute a comprehensive inventory of all the web domains within the national domain, it is relevant to include them in any study of the development of a national web. On the one hand, because they delineate the outer limits of the national domain name space and, on the other, because they attest to the development of the national web domain. The domain name list itself can help to answer research questions regarding, for instance, the number of domain names per year, the number (and names) of domain names that have disappeared or been added since last year, and the number of domain names per domain name owner.

Inventories of the physical space and its inhabitants have been known and used as historical sources for centuries: maps, registers of land and real estate, and population registers. However, the historical use of registrars of digital real estate is still uncharted territory. To the best of our knowledge, only one study exists which aims to map a national web domain based on a study of domain names, namely the above mentioned study of the history of former Yugoslavia's web domain .yu (Ben-David, 2016).

This chapter will investigate how the domain names of the Danish ccTLD .dk can be used as a source, and what they can tell us about the development of the Danish web. The principal focus is on 2005–2015, but the study will also look back to the period after 1987 when the Danish

ccTLD was initially registered. The main source is the complete list of domain names from one date each year, supplemented with information about the domain names from other sources, particularly yearly statistical overviews as well as information from Netarkivet and the Internet Archive. In general, domain name lists are not publicly available, but the national registrar DK Hostmaster provides the Danish list to Netarkivet, where it is the basis for the web archive's broad crawls of the entire .dk domain (cf. below). We have had access to the domain name lists for the present study, but they are protected by national privacy acts and must therefore be processed accordingly. This study is therefore in contrast to Ben-David's (2016) study, which deliberately analysed a disappeared ccTLD, .yu, with a view to demonstrating the challenges of reconstructing a domain name list of a disappeared web domain. The present analysis has access to a complete list of domain names for the Danish ccTLD (at least for the period 2005–2015), and it can rely on a national web archive where the web domains to which the domain names refer can be found.

The national Danish web archive Netarkivet and the Danish ccTLD list

The Danish web is preserved in Netarkivet. Netarkivet was established in 2005 by collaboration between the two national libraries – the State and University Library, and the Royal Library. Since then it has collected and preserved the Danish web based on a legal deposit law (Andersen, 2006; Schostag and Fønss-Jørgensen, 2012). Netarkivet is not delimited to material on the ccTLD .dk. The archive also collects material on any other domain name if it is aimed at a Danish audience or treats themes of relevance for a Danish readership (this material is called 'Danica').

Netarkivet uses three archiving strategies: (1) broad crawls where the entire .dk domain and Danica are archived (four times per year from 2012, fewer in 2005–2011); (2) selective crawls where up till 100 frequently updated websites are archived (e.g. news sites on a daily/weekly basis); and (3) event harvests where websites in relation to events are collected (e.g. elections, disasters, sports events, 3–4 events per year). In November 2015 Netarkivet's collection was approximately 654 TB, according to Netarkivet's website (Netarkivet, 2015). A broad crawl in Denmark is a snapshot of all .dk domains as well as Danish websites published under other extensions, such as .com, .org, etc. The broad crawl is performed by harvesting software, which downloads as much web content as possible from the websites on the domain list, including links and

the websites that the domains link to (for more details, see Andersen, 2006). A broad crawl takes two to four months to perform. In the following we will analyse the development of the Danish web based on the lists from 2006, 2009, 2012 and 2015. From 2012, the lists also contain the names of domain name owners. Table 3.1 shows the broad crawls that are studied in the project.

Table 3.1 Selection of broad crawls

Name of harvest definition	Start date	End date
2005–4–10MB (step 1)	16/12/05	10/02/06
2005–4–500MB (step 2)	20/02/06	30/05/06
2009–1–10MB (step 1)	26/02/09	06/03/09
2009–1–4GB (step 2)	10/04/09	06/07/09
2012–1–10MB (step 1)	23/02/12	13/03/12
2012–1–8GB (step 2)	16/03/12	18/04/12
2015–1–10MB (step 1)	22/01/15	28/01/15
2015–1–10GB (step 2)	04/02/15	24/03/15

As can be seen in Table 3.1, the broad crawls are done in two steps. First, all domains are harvested up to a limit of 10 MB (cf. the names of harvest definitions). Most Danish websites contain less than 10 MB, so this step will harvest approximately 85% of the websites (Schostag and Fønss-Jørgensen, 2012). The second step harvests the larger websites, and as Table 3.1 shows, the limit per domain in the second step has been raised over time as the size of the largest websites has increased. The start and end date of the broad crawl and the time spans vary due to different technical issues (Schostag and Fønss-Jørgensen, 2012).[5]

The development of the domain names of the Danish web

The registry of .dk domains is simply a long list of domain names. The list of domain names constitutes a complete inventory of all the domain names on the national ccTLD at a given point in time. Therefore, it can be used to describe the development of the Danish web without looking in the web archive. Since its beginning, Netarkivet has received lists on a recurring schedule from the national domain name registrar. The data are in fixed width format with domain name, registrant name and email information.

Solmark.dk
Solmarken.dk

DOMAIN

Solmarksvej.dk
Solmaster.dk

Figure 3.1 Extract from the .dk domain name list

When handling data spanning ten years, it becomes apparent that no processing and analysing can be performed without standardizing and cleaning up the data.[6] For instance, the data were standardized into UTF–8 because years ago other character encodings were used. Also, the data were cleaned to remove traces from earlier attempts at handling the data. Dirty data were erased; for instance in one year the list had the remains of some sort of pagination headers, so three rows were deleted in 97 instances (one empty, two purple ones, in Figure 3.1). In other years invisible tab characters were detected that could hinder the data load process.

After cleaning, the data were put into the R system for analysis and charting/visualizations.[7] In R, the individual lists were joined into one data frame that became the base for the analysis.

Number of Danish domain names and ownership 2005–2015

In the analysis of the lists, the following questions were asked: (1) What are the total number of domain names over time? (2) How many domain names have disappeared and have been registered compared to previous years? (3) How many domain names have changed hands compared to previous years? (4) What is the relationship between ownership and domains over time?[8]

The number of domain names is, of course, the simplest question.

Figure 3.2 shows that over ten years, the number of domains has been increasing – but also that the increase is decelerating. This could indicate that the number of domain names is stabilizing. This result is in fact in line with a similar result in a study of the .uk domain (Nominet, 2013). Of course, the result says nothing about the number of active and non-active domains.

For the second question – how many domain names have disappeared and have been registered compared to previous years? – the chart illustrated in Figure 3.3 was created.

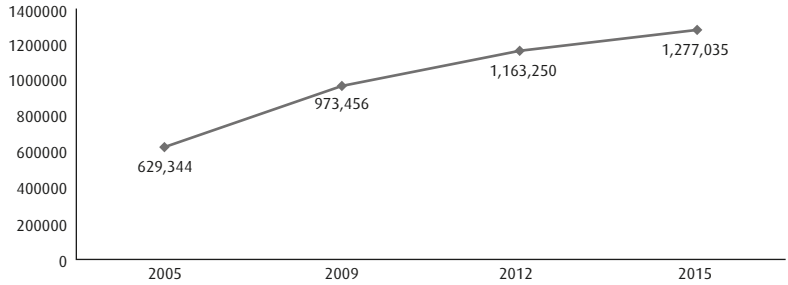

Figure 3.2 Number of .dk domains over time

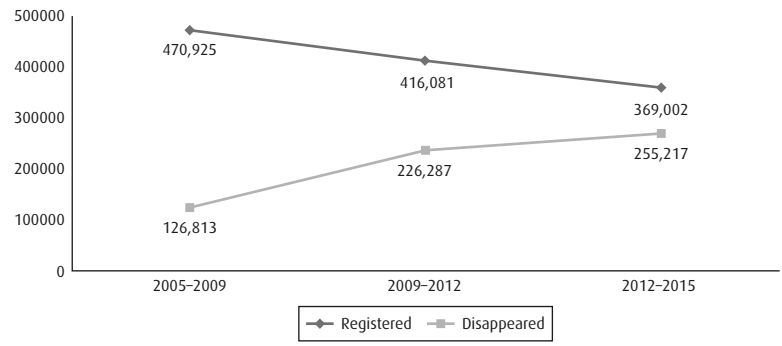

Figure 3.3 Registered and disappearing .dk domain names over time

Figure 3.3 shows that from 2005 to 2009, 126,818 domain names disappeared, and 470,925 domain names were registered. From 2009 to 2012, 226,287 domain names disappeared, and 416,081 domain names were registered. And from 2012 to 2015, 255,217 domain names disappeared while 369,002 domain names were registered. Thus, there has been an increase in the number of disappeared domain names over a three-year interval and a decrease in the number of registered domain names in the same three-year interval. That the two lines approach each other correlates with the gradually slower increase in the total number of domain names indicated in Figure 3.2. As the two lines get closer, the line indicating the total number of domain names will approach the horizontal. Interestingly, Figure 3.3 also says something about the Danish web domain's dynamics or 'aliveness'. At first glance, it looks very dynamic, with many domains being registered and many domains disappearing. However, if we look more closely at the total number of domains that change (that are either registered or disappear), we find

that the numbers add up to approximately 600,000 in all three intervals (2005–2009: 597,738; 2009–2012: 642,368; 2012–2015: 624,219). The dynamics or aliveness of the domain names can therefore be said to be stable over the ten years. This stability can also be seen in the way that the two lines are almost symmetrical around an invisible horizontal line around the number 300,000. In other words, the relationship over time between the increase in disappeared domain names and the decrease in registered domain names is stable.

For the third question concerning the number of domain names that have changed hands over time, we can only compare data from 2012 and 2015, as shown in Table 3.2.

The ratio of domains to owners is approximately the same in 2012 and 2015, with an average of around 2.3 websites per owner. When studying this relationship, however, we find that looking at the average might not be the most relevant way to approach the numbers, as in reality, the domains are not evenly dispersed. In both 2012 and in 2015, just short of 10% of the total number of owners owned 50% of the Danish domains. In addition, in both 2012 and in 2015, if an owner owned more than three domains, s/he belonged to the top 10% of domain owners.[9] When analysing the changes in domain name ownership to answer our third question, we find that in 2015, 14% of the domains from 2012 had changed owner.

In relation to the fourth question – what is the relationship between ownership and domains over time? – the chart in Figure 3.4 shows the results for 2012.

There is no visual difference between 2012 and 2015, and hence no change over the three years. Notably, however, there are two owners who own more than 3,000 domain names, while most owners own one or two domain names.

All four questions are simple questions which reveal something about the development of the Danish web over ten years. The results can be investigated further by means of qualitative analysis. For instance, a closer look at the (types of) domains that have disappeared could uncover interesting patterns. Aspects like these will be studied at a later point in the project.

Table 3.2 Number of .dk domains and .dk owners

Year	Domains	Owners	Anonymous
2012	1,163,250	513,326	46,727
2015	1,277,035	549,978	58,710

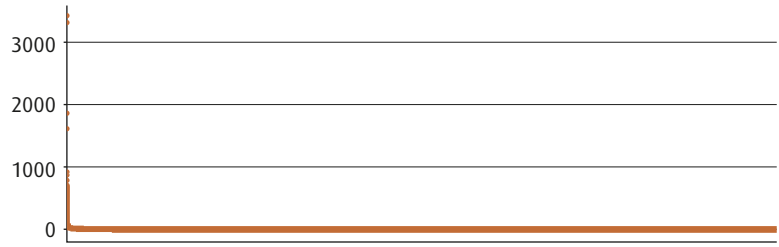

Figure 3.4 Relationship in 2012 between ownership and domains (anonymous registrants removed)

The above results can be further elaborated and put into perspective when combined with data from other sources containing information about national Danish domain names. We have done this in two ways. First, we expanded our analysis of the number of domains over 2005–2015 with data from other sources for the period 1987–2005. Second, we compared our results with data on the number of Danish domains for the period 2005–2015 in Netarkivet and in the Internet Archive, respectively, to see how many of the available domains have actually been archived.

Danish domain names before 2005

In 1987, the internet domain .dk was created. According to an early issue of the magazine of the Danish UNIX User Group *DKUUG-nyt* no 18 (Storm, 1988), the number of registered domain names grew from 49 in 1987 to 70 in 1988. For the years 1989–1995, it has not been possible to locate information on the number of registered .dk domains. But for 1996–2004, a statistics web page from the Danish ccTLD registrar DK Hostmaster's website was found at the Internet Archive.[10] By interpolating from 1988–1996, the chart from 2005–2015 can be expanded as shown in Figure 3.5 (Laursen and Møldrup-Dalum, 2017).

Figure 3.5 shows a slow increase in the years 1987–1997, a steady increase from 1997, a steep increase taking off in the late 1990s, and a slower increase from 2010. There may be various reasons for this development, among which the following three are plausible and could be borne in mind. First, since domain name owners probably prefer as short a domain name as possible, the number of potential names will gradually diminish over the years. Second, the increase in registered domain names correlates with the spread of internet use in Denmark during the

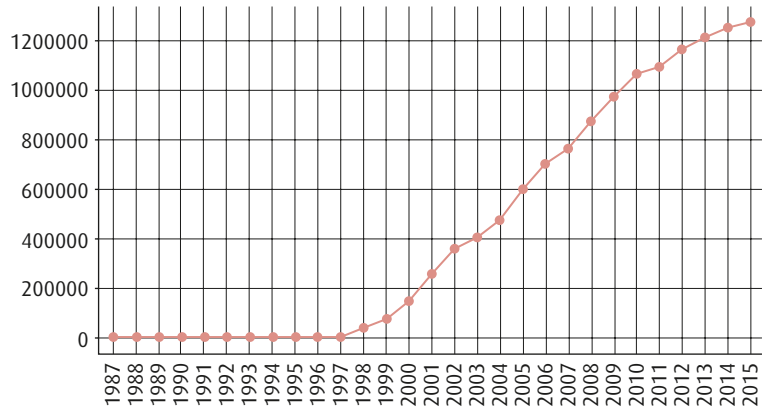

Figure 3.5 Number of .dk domains over time

same period, but with a delay of 2–3 years. The number of internet users slowly increased until 1996 (5%), followed by a steep increase which ended in approximately 2006 (87%) when the curve flattens out until internet access reached 96% in 2014 (Millennium Development Goals Indicators). Not surprisingly, once people have access to the internet, more content is needed, and thus more web domains for content are registered. Third, in 2009 the Danish web domain registrar DK Hostmaster ran a campaign against so-called 'domain name sharks' who bought domain names for 'typosquatting', that is domain names that were misspellings of frequently used domain names (Berlingske Business, 2009).

The Danish domain names in Netarkivet and in the Internet Archive

Our analysis of the domain name lists was compared with data from the archives showing which Danish domains have actually been crawled and archived in the period 2005–2015 to see whether the domain name lists match what is found in the archive. A comparison between the .dk registry list and the domains archived in Netarkivet is shown in Figure 3.6.

As Figure 3.6 shows, more .dk domains are found in the crawled data than on the domain name registry list. This can be explained by differences in time: the registry list is a moment in time, while the crawled data covers a period of time. As time passes, new domains are born. Thus, the two datasets offer two fundamentally different views on the Danish web, where one is no more correct than the other. In addition, Figure 3.6 indicates that the difference in numbers between the registry list and

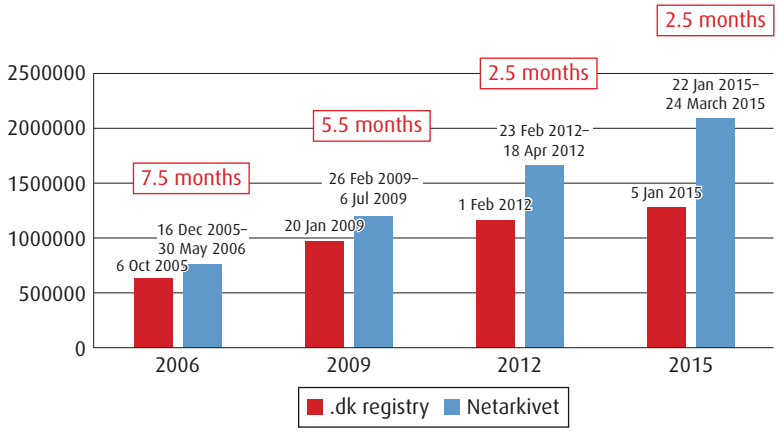

Figure 3.6 Number of domains in the .dk registry list and in Netarkivet

the crawled data increases over time. This could be a sign of aliveness, that there is an increase in the speed at which domains are registered. However, the data are skewed because the crawled data are cumulative – so all the known domain names in the archive are included, even though some may not be active anymore. What we could have done was to exclude domain names with 0 bytes harvested. However, even if we had done that, the data would still not be directly comparable: not only are we trying to compare one moment in time with a period of time, but we are also working with different time spans of the crawls. This means that a comparison between the two kinds of data (and even between the different crawls) has to be done carefully, and taken into consideration when analysing the results.

If we then compare our results with the data from the Internet Archive, the outcome is as shown in Figure 3.7.[11]

Figure 3.7 shows a lot fewer .dk domains in the crawled Internet Archive data than on the domain name registry list. However, again, the data are not directly comparable since the Internet Archive's data, like the numbers from Netarkivet, are based on crawl logs and the .dk registry is not. In addition, data are not from the exact same periods of time. The dates of the .dk registry precede the dates from the Internet Archive, and also the time spans differ: 7.5 months (2006), 5.5 months (2009), 2.5 months (2012) and 2.5 months (2015). Finally, and most importantly, the Internet Archive time spans may or may not cover the archive's broad crawls. Because the intent was to compare crawl log data from the two archives from the same time span, the Internet Archive's

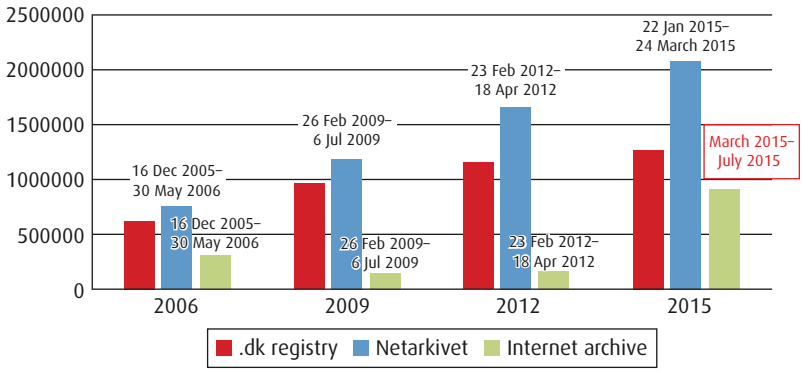

Figure 3.7 Number of .dk domains in the .dk registry, Netarkivet, and the Internet Archive

time periods from 2005, 2009 and 2012 correspond with the start and end date for broad crawls in the Danish web archive. In retrospect, it might have given a more accurate picture had we used the number of .dk domains from broad crawls in the Internet Archive, while still choosing crawls that were as close as possible to the date of the .dk registry list and the dates of Netarkivet's broad crawls. A comparison of broad crawls from both archives would have enabled a less biased result. In 2015, the Internet Archive data do in fact cover an Internet Archive broad crawl, according to information from the archive. Noticeably, this is the year when the number of .dk domains in the Internet Archive is closest to the number of .dk domains on the registry list, that is, 28% less than the .dk registry list.

Comparing numbers of domains, however, does not take into account that domain names may not be the same in the two data sets. For this reason, domain names in the Internet Archive were compared with domain names on the .dk registry list.

Figure 3.8 shows that the Internet Archive contains .dk domain names not found in the .dk registry list, even though the Internet Archive in total contains fewer domain names than the .dk registry list. The difference between domain names in the .dk registry and in the Internet Archive can be explained by the same fact as mentioned above in relation to comparison between the .dk registry list and the domain names in Netarkivet. The new domain names appear in the time span of the crawl (cf. Figure 3.6). For instance, the .dk registry list for 2006 is from 6 October 2005, while the domain names from the Internet Archive are from 16 December 2005 to 30 May 2006. This makes it likely that the

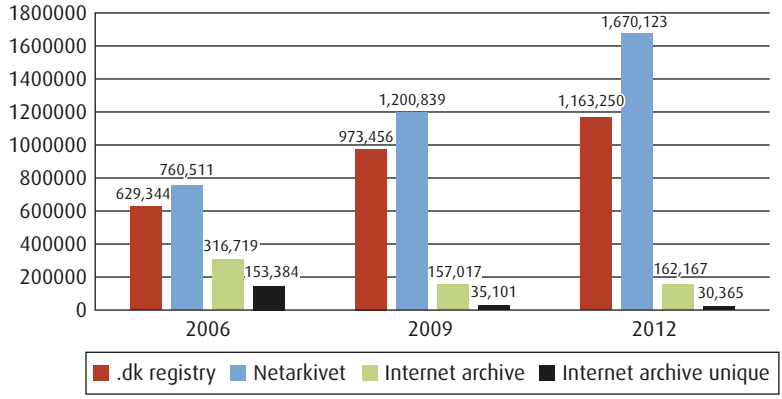

Figure 3.8 Domain names in the Internet Archive not found in the .dk registry

Internet Archive will contain some domain names that do not appear on the .dk registry list.

However, the data offer several possible explanations. One possibility is that the Internet Archive is bad at capturing Danish domains: In 2012, for instance, the Internet Archive collected only about 14% (162,167) of the number of domain names found on the .dk registry list (1,163,250) (cf. Figures 3.2 and 3.7). But from another perspective, the Internet Archive collected about 100% of the statistical increase in newly registered domains from 2009 to 2012 (cf. Figure 3.2). This could be a sign that the Internet Archive is actually very good at capturing new domain names (and that only the new ones are captured). A more likely explanation, however, is changed harvesting settings, which gives bad data or bad calculations. Again, this makes comparing the data a complex matter.

In summary, the number of domains in the Internet Archive does not correspond to the number of domains on the .dk list. The Internet Archive has the aim of capturing all domains and following the links of domains to do so. Consequently, recently registered .dk domains or .dk domains with no or very few ingoing links will have a hard time getting captured. Further studies can provide more insight into the extent to which the difference can be ascribed to the number of .dk domains recently registered as against the number of .dk domains with no ingoing links. Moreover, the Internet Archive captures domains not found in the .dk registry list. This makes it likely that the Internet Archive complements the Danish web archive with regard to some domains. Further

studies can specify the relation between domains on the .dk list and the .dk domains in the Danish web archive. In theory, the domains should be the same, but since a broad crawl takes more than two months, domains may have disappeared from the web before they were crawled. Moreover, .dk domains not on the registry list, i.e. domains that have appeared since the list was made, may have been captured if other .dk domains linked to them.

Finally, a comparison of crawl log data from a broad crawl from both archives could provide a more accurate picture of the capturing of the .dk domains in the two archives and the development of this capture over time. However, a complete comparison will probably not be possible if we take into account that periods of crawling differ and that domains are appearing and disappearing. Even an experiment that started the crawls at the same point in time would make the periods of crawling differ, since different settings and different scopes in the two archives would make crawling end at different times. For this reason, different archives will always complement each other to some extent.

Domain names and archived web

The study of domain names is significant in itself, but it also constitutes an important element in a more comprehensive analysis of the entire ecosystem that constitutes a national web domain. In this section, we will recapitulate what the analysis of the domain names tells us about the Danish web domain, before briefly outlining some of the ways in which a domain name analysis and an analysis of what can be found in a web archive can supplement each other.

What domain names can tell us about the Danish web domain

Digging into the development of domain names can tell us something about three things at least. First, the size of the Danish web can be described through the number of domain names on the Danish ccTLD. However, size understood as the number of domain names does not say anything about how big the Danish web domain is in terms of bytes or number of digital objects. The analysis of the development of the number of domain names indicates that the Danish web name space has grown steadily, until it reached a level where the pace of growth slowed down. To put this another way: the number of street names has increased which in itself is significant, but this does not tell us anything

about how broad or long the streets are, how many houses are on them and how big those houses are.

Second, analysis of the domain names tells us something about the 'aliveness' of the Danish web domain. In the early years, many new domains were registered; the Danish name space was much like a 'wild west' with new streets and houses appearing rapidly. During that time, fewer domains were disappearing compared to later in the study period when the number of disappearing domains began to close in on the number of registered domains. Whether the increasing number of disappearing domains in the later periods signals that more of these early 'streets and houses' are being removed at a more rapid pace, we can only speculate at this point. This could be a subject for further study. Some registered domains might disappear within the same intervals as they were registered, so studies could also be done in order to determine how often this is the case. In addition, it would be interesting to study the average lifespan of a Danish website throughout the period studied.

Third, the domain name lists are very indicative when it comes to ownership. They chart a domain name space where, on the one hand, only a limited number of domain names change hands and, on the other, that this is a space that can be charted as a long tail. Many people each own a very limited number of domain names, and few people own a relatively large proportion of the domain names. This can be likened to a physical space with many small home owners and a few big land owners, or maybe even property speculators.

All three types of result could be correlated with the nation's life outside the web by looking into information found in other sources. For example, by comparing the number of domain names with the number of households with internet access, or by finding more information about the owners who have many domain names.

Domain name studies and archived web content

It would not have been possible to arrive at the results reported here regarding the Danish web domain simply by looking into the web content in the web archive. This is in itself indicative of the value of studying domain name spaces. However, this should not overshadow the fact that combining domain name analysis with analysis of actual archived web content opens up new avenues of inquiry. On the one hand, the archived web can shed light on the development of domain names, and, on the other hand, domain name analysis can help us understand the archived web. Let us have a closer look at these two avenues.

Although the analysis of domain names provides valuable insights, for instance, about which domain names have been established, it does not tell us anything about these web domains: were they actually used by the owner, or were they just an almost empty web page 'under construction'? And what could one actually find on these websites? Following our earlier analogy: what did the streets and houses actually look like? An analysis of archived web content can enhance the analysis of the domain names.

However, the opposite is also the case, since the results of the domain name analysis can supplement analyses of the archived web content and function as a stepping stone to the archived content in at least two ways. The first concerns the completeness of the web archive, the other concerns the generation of new research questions and hypotheses.

The domain name list itself is an important key to evaluating the completeness of how much of a national web domain is in fact a web archive. As an inventory of all the domain names on the national ccTLD, the list can measure the completeness of archived web content in studies where a nation's web domain is delimited by the ccTLD. Access to the historical ccTLD domain name list is particularly important if the corpus of the national web domain is extracted from a web archive which has not been based on a ccTLD domain name list, as is the case with the Internet Archive. If the established corpus is not compared to the domain name list from the same point in the past, however, it is difficult to evaluate the completeness of the corpus at the domain name level. In short, a complete domain name list can help us establish to what extent the archived content actually mirrors 'the nation's web' of the past. However, things may prove to be more complicated than this. A complete domain name list does not in itself provide a solid baseline for what the Danish web actually looked like at a given point in the past, simply because archiving takes time. Each of Netarkivet's broad crawls starts with a comprehensive seed list in the form of the authoritative ccTLD domain name list. However, a comparison of the domain names archived in Netarkivet and those archived in the Internet Archive reveals that there may have been Danish web domains that are not included in the broad crawl based on the ccTLD list. This is simply because web domains are likely to have been registered or to have disappeared during the two to four months it takes to archive the entire Danish web domain. An analysis of domain names can draw our attention to the tension between the static list of ccTLDs that is used at the launch of each broad crawl and the dynamic evolution of the domain name list during the time it takes to archive the ccTLD. There is no easy way to solve this problem. The longer it takes to archive the ccTLD, the greater the chance for an inconsistency between

the initial list of ccTLDs and the evolving list. With a shorter archiving time, the possible inconsistency is smaller, but at the expense of fewer web domains being archived. However, the insights revealed in this study may indicate a need to combine different collections of the archived web.

The second way in which domain name analysis can supplement the archived web content is that it can help generate new research questions and hypotheses. The following two examples can illustrate this. First, one could investigate 'the disappeared web'; that is, all the web domains that have disappeared year by year: What did the disappeared web domains look like? Have specific types or genres of websites disappeared? And are there any patterns or trends in these types compared over time? Second, one could dig deeper into ownership and investigate the many web domains that are with the same owner: Do they belong to a specific content genre, or are they diverse? And how are they interlinked? For instance, a hyperlink analysis of the web domains having the same owner could identify link patterns and maybe tell us something about the extent to which these websites do or do not cluster. In addition, one could correlate the postal address of the owner with actual geo-information on the websites (postal codes, city names, etc.) with a view to investigating whether the web domains 'live' at the same place as the owner. One could investigate the real estate domain of the name landscape. Are there any patterns with regard to content type for the web domains that are often passed to another owner? A final subject for study are the web domains that have 'disappeared' because they were never archived in Netarkivet, having fallen prey to the temporal lapse between the initial ccTLD list and its evolution during the archiving process. Once these web domains have been identified in another web archive such as, for instance, the Internet Archive, one could look for patterns in terms of content or genre.

Conclusion

Many levels of analysis are necessary to derive a comprehensive analysis of the entire ecosystem that constitutes a national web domain. In this chapter, we have taken the first steps in answering the big question(s) of what an entire national web domain could look like and how the concept has developed over time. Our study was based on the historical development of the backbone of the Danish web: the national domain names (ccTLD). This approach has focused on the analysis of historical changes in the .dk domains as they appear in three different sources: lists from the

Danish national registrar, the Danish national web archive, Netarkivet, and the international web archive, Internet Archive. The analysis of the domain name list shows that the number of domain names has increased over the years but that the pace has changed. From a slow start at the end of the 1980s, there was a lot of activity from the late 1990s more or less corresponding with the spread of internet use in Denmark. Since 2010 there has been a tendency for the curve to level off, but it still shows a steady upward slope. This is not surprising. We would expect the Danish web to become gradually larger (here understood as the number of domains) following the general spread and growth of the internet. It has become the norm for companies, institutions and organizations to have their own website, often on their own domain.[12] While the number of domains is increasing, we see that the relationship between registered and disappearing names is relatively stable, highlighting that the dynamics of the Danish web are more complex than just the appearance of more domains. Another aspect relates to the ownership of domains and studies of the top 10% of domain name owners might shed light on parts of the dynamic in the Danish domain name scape.

An important lesson from this study is that the three datasets – the .dk registry list and the list of archived domains from each of the two web archives – offer different insights into the development of the Danish web. Combining them contributes not just to furthering the understanding of each data set, but also to understanding the complete picture of the ecosystem. It is important to supplement the results of the domain name analysis with more analyses of the archived web content, and we propose to do this by creating a corpus for each year and analysing these corpora focusing on the size, space, structure, aliveness and content (as described in the introduction to this chapter). Both approaches constitute valuable methods to the understanding of the evolution of a nation's web domain, and they could both be included as best practice in the toolbox of similar studies in the future. Another way to enhance the results is to combine a quantitative approach with qualitative analyses, studying selected websites in more detail. Hence, a multitude of approaches and sources could possibly be included in further research, and they will all be useful in gaining a comprehensive understanding of the historical development of a national web.

PART TWO
MEDIA AND GOVERNMENT

4

The tumultuous history of news on the web

Matthew S. Weber

News and the web

Tuchman (1978) noted in her ethnographic research that newspapers were responsible for creating a constant flow of information to consumers, continually moulding our comprehension of society. In the 1970s, more than 62 million newspapers were sold in the USA each day. Readership had been growing or stable for more than 50 years. Subsequently, what was a constant flow has become a torrent of news today, and the news industry has entered a period of remarkable tumult. News and information today flows to consumers via many traditional media, but it is increasingly complemented, and in some cases preceded, by computers, tablets, mobile phones and other emerging devices. The internet today is the pipeline that feeds information to consumers, and for the time being the web is a primary window through which that information is distributed, accessed and retrieved.

In the past 20 years, the relationship between news media companies and internet technology has been a tenuous one at best. Traditional print newspaper circulation has declined sharply since the early 1990s; hedging their bets, many newspaper companies began experimenting with internet technology as early as the late 1980s. Much of that experimentation, however, was often done simply as an extension of existing printed products (Boczkowski, 2004a). Most newspaper executives were sceptical that the web represented a credible threat, and as a result most innovation occurred at arm's length (Chung, 2007). In recent years, however, the need for change has been widespread and notable; even television news has faced challenges from the rise of online video

and the looming threat of online-only programming (Perren, 2010). Thus, although there were early innovators in the newspaper industry, it is only in recent years that revolutionary change has occurred on a large scale (Karimi and Walter, 2015; Schlesinger and Doyle, 2015).

History of news on the web, from the web

The following discussion traces the tumultuous history of news media on the web. More specifically, this chapter focuses on the history of newspapers in the USA as they have grappled with adapting to new digital technology by tracing their development through their websites. A multitude of notable volumes exist that trace change in the newspaper industry in response to digital technology (see, for example, Boczkowski, 1999, 2004a; Kawamoto, 2003; Usher, 2014). Contrasting prior work, this chapter is not intended as a comprehensive history of news online; rather, this chapter aims to illustrate the adaptation of the newspaper industry to the web through an examination of the content of the web itself, including snapshots of the broad ecosystem.

The web as history: news online

This story is told through an examination of archived news media content maintained within the Internet Archive. There are few cohesive sources of archived news content available for large-scale research; the largest source of archived internet content is available via the Internet Archive (accessible via archive.org). Founded in 1996, the San Francisco-based Internet Archive is a non-profit organization that was established with a mission to preserve the history of the web, and to build an internet library containing that history. The Internet Archive is best known for the Wayback Machine. The Wayback Machine is the graphic interface for the Internet Archive databases, and allows users to freely access the information stored within the archive. Beyond the Wayback Machine, the Internet Archive maintains a rich repository of web pages, photographs, videos and other types of digital content. As of 2015, the Internet Archive had recorded, parsed and archived 438 billion web pages occupying 23 petabytes (PB) of storage space. While the Internet Archive does not contain a complete record of the internet, it is the single most extensive archive of the internet.

Most of the statistical analysis and web pages referenced in following sections were enabled through the ArchiveHub project.

ArchiveHub is a National Science Foundation-funded project intended to enable researcher access to archival internet content. The ArchiveHub project includes a substantial collection of media content stored in a researcher accessible database format, with extensive metadata to aid scholarly work. For example, a dataset of media websites from 2008 to 2013 includes 1.3 billion captures of web pages representing 540 million unique URLs. A second smaller dataset contains a total of 25,628 websites from 1996 through 2007; the dataset contains roughly 300 million total captures. The aggregate datasets provide a robust source for better understanding the history of news and newspapers on the web. The focus of the ArchiveHub project is on the development of tools that extract hyperlinks from Internet Archive records. Hyperlinks are useful for understanding the degree of connectivity between websites, and for mapping the flow of information (Weber and Monge, 2011). In addition to extracting hyperlinks, the tools provide additional information including keywords relevant to a given web page, and information such as the size of the archived website content measured in megabytes (MB).

Thus, in the context of this chapter, the web is both the source of change, and the means by which the story is told. The dramatic change that has occurred in the news media industry was predicated by the introduction of new technological standards in the 1990s; and yet today I am able to tell that story thanks to that very technology. Recent scholarship examining online news has often relied on data from the web to help tell the story of changes in the modern news media landscape. A 2012 study examined newspaper websites in the UK. This illustrated the fluidity of online news, with sources and content being added and deleted over time as a particular story or news event evolved; the traditional notion of a news article as a fixed unit is less prominent in the digital space (Saltzis, 2012). Similarly, web content has been used to show how competitive news outlets often mimic the coverage of competitors in order to ensure relevance (Boczkowski, 2010).

I draw on a number of different examples and analyses to illustrate the role of the web in telling the history of news on the web. Large-scale data are paired with a series of case studies to demonstrate how individual news media responded to competitive threats, first from hyperlinks and the free flow of content, and subsequently from social media. The growth of internet technology flattened the competitive landscape for newspapers; while much has been made of innovation within industries, the interaction between print newspapers and new media brought about rapid transformation across the news landscape. The following

vignettes focus first on the early period of the web, from 1990 through 2005; the latter sections fast-forward and highlight more recent developments between 2010 and 2015.

The early days of news on the web: 1990–2005

In 1991, Tim Berners-Lee published the first website on the first web server at the CERN laboratory in Switzerland (Berners-Lee, 1991). Newspapers were quick to join the web, building on their early experimentation with internet technology. The introduction of new technology delivered a shock to the newspaper population, but the industry responded with measured innovation rather than any significant restructuring (Sylvie and Witherspoon, 2002; Boczkowski, 2004a, 2004b; Patterson, 2007). In this way, newspapers have taken advantage of internet technology since its infancy, although many early experiments were not successful. The Columbus Dispatch was the first daily newspaper in the USA to provide an online version for its customers. In 1980 the newspaper provided access to an online version via the internet service provider CompuServe (Kawamoto, 2003). A number of newspapers also experimented with Videotex, an early digital information transmission system. Knight-Ridder, a former American news company, even developed a proprietary Videotex system.

The newspaper industry, in general, viewed online technology as a new medium for distributing an existing product, and for nearly a decade newspaper organizations focused on products that simply delivered the print product digitally (Boczkowski, 2004a). To this end, Falkenberg (2010) delineates between 'online newspapers' and 'newspapers on the web', with the early period of the web primarily occupied by newspapers on the web as replications of their printed products. In 1993, the first commercial graphic web browser, Mosaic, was launched, and by 1999 more than 4,900 newspapers globally had launched web versions of their newspapers.

Web technology gave rise to the newspaper websites that many consumers are familiar with today; 1991 to 1993 represented a juncture in the history of online news, because World Wide Web protocols including hypertext markup language (HTML) enabled a new visual interface for accessing news via the internet (Stovall, 2004). In 1994 *Raleigh News & Observer* launched Nando.net as a web-based version of their newspaper; this is one of the first examples of a web-based newspaper living on the web outside of an internet service provider's intranets. The first

available record of Nando.net is available in the Internet Archive. The visual nature of the interface is clear, as is the differentiation from the traditional print product.

The archival pages of Nando.net include articles, images and hyperlinks to other early websites. This type of record allows for an examination of the type of content available on the early web, as well as the type of communication enabled by early web protocols. For instance, the earliest records of Nando.net include a rich repository of photographs drawn from news wire services, contrary to the perception that early web-based newspapers lacked graphics or visuals. The *Raleigh News & Observer* generally published photographs with sports content, but photographs were for the most part not included with general news articles. Based on statistics on archived web pages of Nando.net, the website received about 14 million visitors per week in 1999. Despite the arguable success of early web ventures such as Nando.net, newspapers were quickly failing to replace lost print advertising revenue with the equivalent in online advertising revenue (Weber and Monge, 2014).

The rise of blogs

Newspaper advertising revenue has generally been tied directly to the number of readers purchasing a newspaper. In the 1990s, new digital sources were drawing readers away from printed products in large numbers, and in time, rapidly drawing away revenue.

In 1997 and 1998, the first variations of a new type of web content emerged online in the form of weblogs, more broadly referred to as blogs. The World Wide Web was not a tool born of the newspaper industry. Rather, it is a technological disruption that originated as a government and research communication tool. Consequently, innovation on the web came from many sources. By 1994, early innovators were using web pages as tools for online diaries and personal commentaries. In 1997, Jorn Barger launched Robot Wisdom, which featured a listing of links that Barger liked to visit, as well as updates from Barger's daily life (Rettberg, 2008). Similar types of sites began to pop up en masse, but generally failed to attract large audiences. The term 'blog' was first used to describe these sites by Peter Merholz in 1999; the term was a shortening of 'weblog' (web log), which Merholz thought ought to be pronounced 'wee-blog' and later shortened to 'blog'. The large-scale emergence of blogs served as an exchange network of sorts, whereby users were able to share hyperlinks with one another to identify information sources of interest (Ammann, 2011). But a high barrier to entry

plagued these early sites; a user seeking to build a blog was required to have a sufficient amount of technological expertise in order to build and maintain the site. In October 1998, however, Open Diary was founded to offer users space on the web with free hosting and easy-to-use online publishing options. Within four months, the site had 25,000 hosted online diaries. Pitas launched in 1999 offering free blogging tools, followed by the launch of Blogger. Traditional newspapers continued to adhere to the strict routines of printed newspapers, but blogs allow writers and reporters to share opinions and publish relatively raw content outside the bounds of journalistic hierarchy. Early blogs were relatively simple hypertext documents updated on a relatively frequent basis, with content ranging from a few roughly assembled sentences to complete magazine-length features (Matheson, 2004).

Weblogs were one of the first forms to dramatically reinvent the form of daily news. The Drudge Report, an early variant of weblogs, first appears in the Internet Archive on 10 December 1997 (the Drudge Report was founded in 1996, but in the early archive there is an occasional delay between actual founding and appearance in the database). The Drudge Report and other 'news' blogs were simply aggregations of links to other websites and news articles. The Drudge Report continues as a popular and successful news source today. Early blogs were a harbinger of future change, but were rudimentary in nature.

And then came social media

In 1997, SixDegrees.com launched, allowing users to create profiles and connect with other friends on the site. SixDegrees is generally credited with being the first social networking site (SNS) (boyd and Ellison, 2008). Subsequently, numerous imitators emerged, and many were successful in improving social networking as a platform. In 2002, Friendster.com launched and quickly gathered a following of more than 300,000 users. From 2003 onward, SNSs established themselves as mainstream media platforms, due largely to the development of Web 2.0 technology. Web 2.0[1] technologies are a class of platforms that enable consumer participation and interaction in online environments, including discussion and creation of the news. Today, an SNS is viewed as a website that 'connects and presents people based on information gathered about them, as stored in their user profile' (Cruz-Cunha et al., 2011: xviiii). boyd and Ellison (2008) distinguish SNSs as websites that allow users to (1) create a public-facing profile, (2) construct a list of users to whom they are connected, and (3) navigate lists of connections

for individuals and their connections. More broadly, SNSs are online resources that allow users to create 'maps' of their social networks, and to share information through these networks. By 2000, SNSs gained further traction as a means of sharing information between users; eventually this included pointing others to news articles, and providing links to news on websites (Suler, 2004).

The challenge of adapting to the web

The introduction of World Wide Web protocols was a first tipping point; the rise of Web 2.0 technology created a second tipping point in the history of news on the web. By the turn of the century, social networking sites were gaining in number and popularity. With the popularization of Web 2.0 technology, blogs became increasingly widespread and interactive. During this period, successful news blogs such as Huffington Post and Gawker were launched. Thus, during this period the notion of online newspapers began to reach maturity, as content was being produced exclusively for the web (Falkenberg, 2010). For instance, when the Huffington Post officially launched on 9 May 2005, its interface was driven by a strong visual design, and the site included features that allowed users to comment on the news and to engage with the website.

The rise of bloggers and blogs presented a clear challenge to traditional newspapers, as did the rise of Web 2.0 content and social media. In part, newspapers were challenged by their own structural inertia; given the storied history of many of the large newspapers in the USA, it is not surprising that many organizations were hesitant to transform completely to a digital platform (Weber, 2012; Weber and Monge, 2014).

For example, Figures 4.1 and 4.2 illustrate the changing dynamics of interaction between traditional newspapers on the web, and blogs on the web. The red circles represent established newspapers with a presence on the web; the blue circles represent blogs, online communities, and online-only news sources. The data shown in the illustrations focuses on a subset of the larger dataset, with the subset containing 269 blogs, 192 online communities and social networking sites and 487 newspapers. Web archiving technology changed significantly during this early period, and thus the subset was selected by identifying the websites for which data was consistently available for the period of interest. Hyperlinks to advertising websites were removed. A connection between two websites exists if a hyperlink existed between two websites and was present at least three times in a given year. Hyperlinks are useful for analysing the relationship that existed between media

organizations; in cases where a hyperlink persists over time, prior research has established a connection between the presence of hyperlinks, and the presence of a relationship between organizations (Gao and Vaughn, 2006; Shumate and Lipp, 2008; Tsui, 2008; Turow and Tsui, 2008).

As Figures 4.1 and 4.2 demonstrate, there was little interaction between bloggers and newspapers during this early period. There is a clear shift over time, however, as the two disparate groups become more intertwined. The visualizations were generated by conducting an analysis of the hyperlinks between newspaper websites and bloggers via the Internet Archive, and illustrate the connections between all websites in the network.

A subsequent statistical analysis provides further insight. For instance, the density of the networks visualized in Figures 4.1 and 4.2 measures the percentage of hyperlinks that exist as a percentage of the total possible hyperlinks. Controlling for growth of the web as a whole,

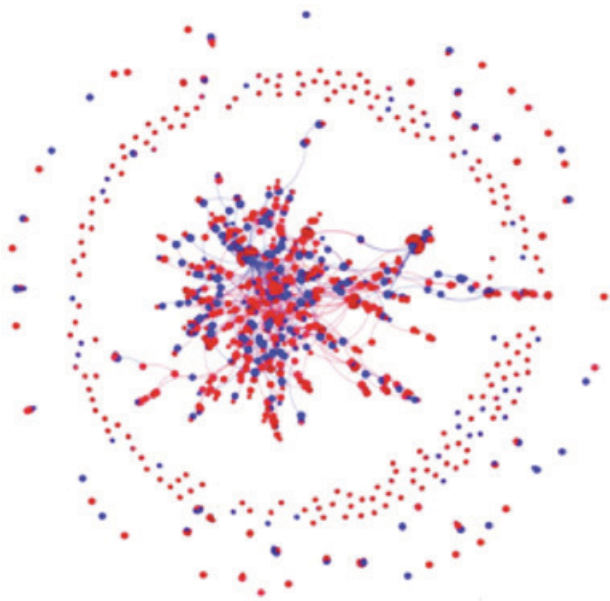

Figure 4.1 Connections between newspapers and other websites on the web in 1999 (red indicates traditional newspapers; blue indicates online native entities)

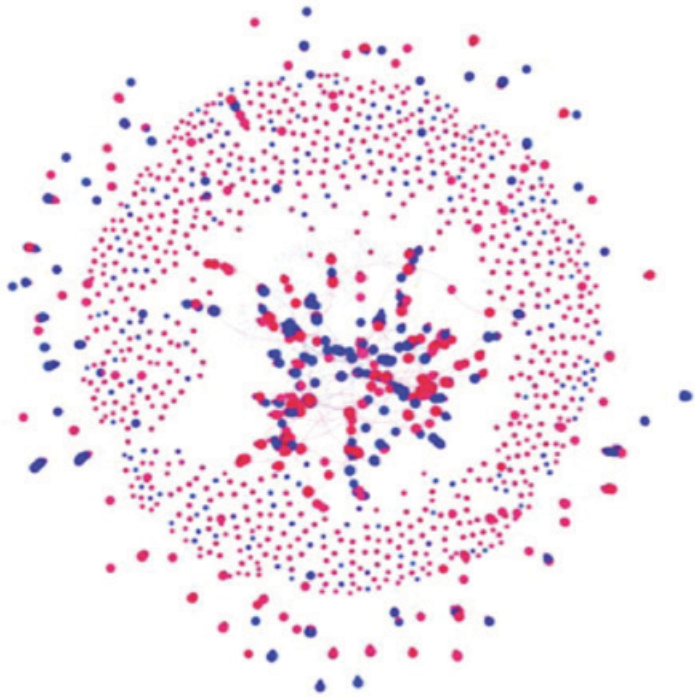

Figure 4.2 Connections between newspapers and other websites on the web in 2005 (red indicates traditional newspapers; blue indicates online native entities)

the density of the network decreased; in 1999 the graph density was 0.34%, and in 2005 it had decreased to 0.22%. The web has always been a vast space, and the increase in news outlets and websites over time helped to create 'pockets' of news early on. Websites clustered together in relatively disparate groups, as evidenced by the interconnectivity of early websites. At the same time, connections between different types of websites increased; in 1999, only 12% of hyperlinks in the above sample existed between website types; that number increased to 32% in 2005. This helps to explain why the visualization is more cohesive in the later time period, and demonstrates the slow erosion of the barriers between the disparate pockets of news.

By the end of this critical period, the online news ecosystem no longer existed as an ecosystem of disparate entities. The new news

ecosystem that existed in 2005 developed as an agglomeration of traditional newspapers, blogs and early social networking sites; many of the virtual barriers were eroded by this point in time. The early period of transformation saw the online news ecosystem move from one of isolated hubs of news websites – dispersed with relatively little interconnectivity – to an integrated network of news and information websites.

The web, grown up: 2010–2015

What a difference a decade makes. According to the Pew Research Center's 2012 News Consumption Survey (Kohut et al., 2012), in 2000, 23% of Americans reported that they went online for news at least three days a week. In 2010, that number had increased to 46%. From 2010 onwards, the disruption of the news media landscape accelerated. This is, in part, due to broad societal shifts. Based on data from the US Census, the percentage of households with internet access increased from 41.5% in 2010 to 71.1% in 2015. The number of web-based sources increased with the growth in online traffic, and yet, users were increasingly concentrated among a handful of websites. In 2010, an analysis of 4,600 news and information websites tracked by Nielsen Media found that the top 7% of news media websites collected 80% of the overall traffic (Mitchell and Rosenstiel, 2010). According to ComScore ('Digital: Top 50 online news entities', 2015), by 2015, the top online sources for news in the USA were Yahoo!/ABC News Network, CNN, NBC News Digital, HuffingtonPost.com and CBS News; the top newspaper, *USA Today*, came in at sixth in terms of unique visitors.

During this period, newspaper print circulation continued to slip, although the rate of decline slowed to less than a 5% decline per year. Yet significant shifts in audience preferences had already occurred. Data from the American Press Institute found in 2014 that 69% of Americans accessed news via their computers, 56% accessed news via their cell phones and 29% accessed via their tablets (based on sources used for news in the past week). Audience preferences had clearly taken a dramatic swing towards a wide array of digital devices.

Revenue challenges persisted hand-in-hand with readership challenges; digital revenue continues to be a small fraction of overall revenue for legacy newspapers, as well as other legacy news sources including magazine and television. Despite the slow decline of print newspapers, many continued to attract a substantial audience. In print, *USA Today* had an average daily print circulation of 3.3 million, based on 2014 data.

Comparatively, however, *USA Today* is estimated as the largest online newspaper in the USA, reaching 54.5 million users online in January 2015. Elsewhere, *The New York Times* maintained a daily circulation of 2.1 million in 2014. *The New York Times* has also had success driving revenue online; the newspaper is an example of a traditional news organization that has succeeded in developing a revenue model for its online content (Pickard and Williams, 2013); despite this success, a relatively small proportion of print newspapers have succeeded in replicating a 'paywall' model.[2]

At the same time, native online news services began to emerge as key providers of news and information; SNS are increasingly driving consumers to news content, and serving as a key portal through which consumers discover news (Perelman, 2014). Native online news services refer to organizations that create and distribute news media solely through the web. Today, there are many different iterations of native online news services; despite the popularity of online news platforms, there is still significant experimentation. This is consistent with a growing industry (Weber and Monge, 2014), and is a trend that will likely continue for the near-term future. For example, Buzzfeed first gained prominence when it launched in 2006 as an incubator of digital content, but as it has evolved into a news provider it has gone on to develop its own newsroom (LaFrance and Meyer, 2015). In another example, ProPublica was launched in 2007 as a non-profit news organization. The news service's primary goal is to publish in-depth investigative journalism; in 2010, ProPublica became the first online news service to win a Pulitzer Prize. But not every venture has been successful; for instance, GigaOm, a popular technology news blog, launched in 2006 but shut down in 2015 due to declining revenue.

The changing news landscape as a local story

As the preceding discussion illustrates, the period of time from 2010 to 2015 has seen continued changes in the news media landscape. Despite ominous predictions about the death of the printed newspaper, and despite the contraction of the newspaper industry as a whole (Deuze, 2003), newspapers continue to publish both in print and online, albeit in a diminished capacity. More recent data from the Internet Archive, covering 2010 to 2015, provides a snapshot of recent changes to the national news landscape. For instance, an examination of outbound hyperlinking of the top 25 national newspapers in the USA reveals that

from 2012 onwards, 98% of those newspapers' websites contained outbound links to Twitter and Facebook, illustrating the growing role of SNS in the news ecosystem.

As the web has grown up, one area where there has been a profound change is in the provision of local news. Much of the early innovation by newspapers occurred at larger newspapers; small community newspapers lagged behind in terms of the development of web content (Greer and Mensing, 2004). And yet, community engagement with local news is well established as a key predictor of community health, and helps to foster social interaction within communities (Paek et al., 2005). Readership of local news is also directly related to the likelihood of voting in elections (Moy et al., 2004).

As web-based news moved towards maturity, many saw an opportunity for web-platforms to improve local news coverage. For instance, Downie and Schudson (2009) predicted that the launch of local news websites by entrepreneurial journalists would help to improve local democracy. Similarly, Lewis (2011) observed that foundation-funded hyperlocal websites managed to successfully pair new technology with high quality journalism, creating an opportunity for growth.

Despite early optimism, by most accounts the digitization of news has had a negative impact on local news. As of 2010, many local newspapers were in a state of crisis (Nielsen, 2015; Wadbring & Bergström, 2015), and in markets where coverage had decreased there was already an indication of a decline in political participation (Hayes and Lawless, 2015). A 2015 report from the Democracy Fund highlights the plight of local news; the report examined the health of local news in New Jersey, and found that there are stark gaps in local news coverage (Napoli et al., 2015). For instance, the report observed that there are only 0.58 sources of news for every 10,000 people in Newark, NJ, with a population of 277,000 and a per capita income of $13,009. Comparatively, Morristown, NJ, with a population of 18,000 and per capita income of $37,573, has 6.11 sources per 10,000 people.

A new perspective on change in local news coverage is enabled through an examination of the Internet Archive's records of local news websites. In order to better assess changes in local news ecosystems, a longitudinal examination of the New Jersey news ecosystem was conducted using a subset of data from the Internet Archive. The subset of websites was selected by hand coding local news websites in New Jersey and extracting those websites from the Internet Archive's repositories. In this case, an analysis of the hyperlinks between local news websites allows for an examination of both the scale of the local news ecosystem,

as well as the cohesiveness and coherence of the ecosystem. The analysis is based on a subset of local news websites extracted from the Internet Archive. A list of 390 local New Jersey news websites was created based on websites that operated between 2008 and 2012; because of the structure of the market, some key websites from Philadelphia and New York were included as key information sources. The resulting dataset includes approximately 1.6 million captured websites across the five-year period, which includes both the focal local news websites as well as websites that are connected to those organizations. Moreover, the number of captures does not directly reflect the number of websites. It is not unusual for a website to have hundreds, or even thousands, of web pages with a given domain, and those domains are crawled many times in a given year.

Again, using the ArchiveHub system, hyperlinks between websites were extracted from the Internet Archive data. This allowed for an examination of the flow of information between news websites (see Weber, 2012, for a further discussion of the role of hyperlinks as a tool for guiding information flow). In addition, it was possible to summarize the amount of information within each domain (measured in megabytes). The amount of information in megabytes can be considered a proxy for the amount of text and images on a web page.

Table 4.1 provides descriptive information regarding the state of local news in New Jersey from 2008 through 2012. In order to examine changes in the local news landscape, this analysis focuses on changes in the core sample of websites, based on average degree, average path, density, connected strong components and clustering. Average degree measures the average number of connections per website. Average path measures the average of the shortest path that exists between all

Table 4.1 Network analysis of local New Jersey news websites, 2008–2012

	2008	2009	2010	2011	2012
Websites	90	193	216	338	203
Connections	105	246	261	388	315
Average degree	1.17	1.28	1.24	1.14	1.56
Average path	1.48	2.25	2.05	3.11	2.45
Density	0.013	0.007	0.006	0.003	0.008
Connected strong component	90	106	207	335	194
Clustering	0.014	0.029	0.0016	0.007	0.095

websites; it gives an indicator of how connected all websites are to one another. Density, as previously mentioned, accounts for the number of connections between websites as a percentage of the possible connections. The number of connected strong components gives a measure of the number of clusters of websites that exist. Clustering gives a measure of the degree of clustering on a scale of 0 to 1.

The descriptive analysis in Table 4.1 reveals some critical changes in the local New Jersey news landscape. Between 2008 and 2011, the number of local news websites increased significantly, but the data begin to reflect the broader trend of decline in 2012. Although the number of websites present declines in 2012, the connectedness of websites remains relatively stable, as seen in the number of connections, as well as changes in the average degree and path length. With regards to clustering, the number of components increased at first, but declined by 2012. On the other hand, the degree of clustering decreased.

In aggregate, these data illustrate a story of a decreasing number of local news websites that are increasingly clustered together. Echoing prior research, this type of story would be consistent with a declining number of websites increasingly sharing information with one another.

Figures 4.3 and 4.4 provide further context, illustrating the connections that existed between key websites in the New Jersey local news ecosystem based on hyperlinking. The two visualizations illustrate the top 30% of websites in each year, based on the degree of connectivity to other websites.

Comparing the two visuals, it is clear that the ecosystem in 2012 had become more tightly clustered. Moreover, there are fewer organizations engaged in this central cluster; for instance, Philly.com is no longer central in this network, and the dailyrecord.com is less prominent. Looking to 2012, the connections within this smaller cluster are also stronger, as illustrated by the thickness of the connections between websites. Orange connections represent relationships with less prominent websites, whereas blue connections are relationships between equally popular websites.

This analysis provides an overview of the changing local news landscape at the beginning of this critical period. Clearly this is a single analysis of a single state, but the trends are consistent with previous research on the topic. Large-scale web data provides a unique vantage point for assessing the health of local news environments; furthermore, the nature of web archives provides a tool with which research can code and analyse the actual content, creating a fertile resource for future research.

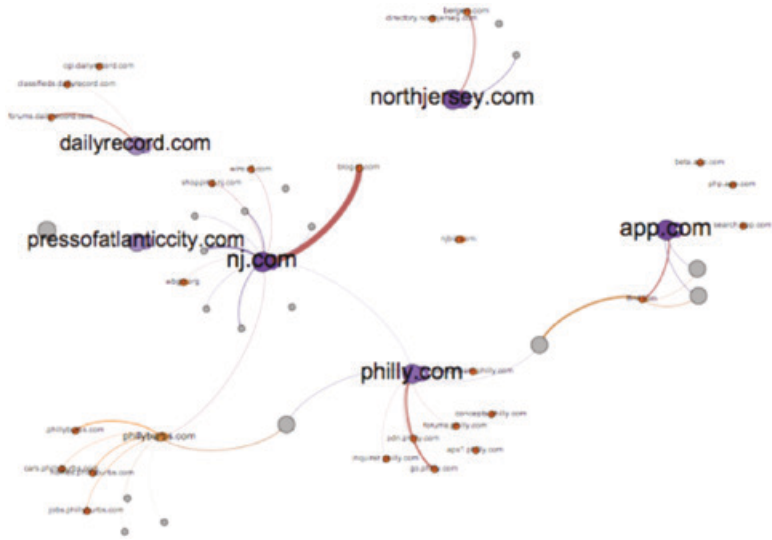

Figure 4.3 New Jersey local news ecosystem, 2008

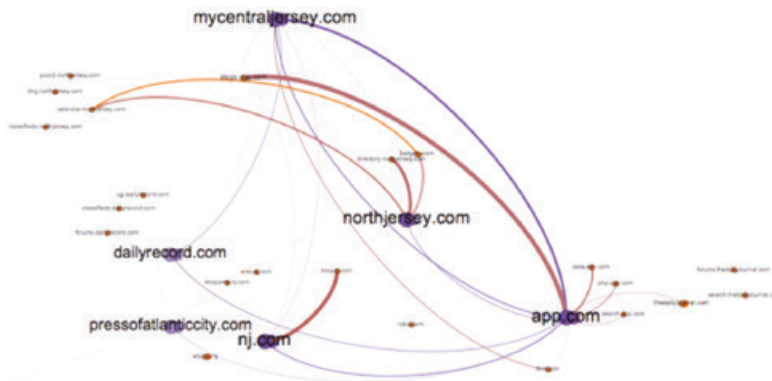

Figure 4.4 New Jersey local news ecosystem, 2012

The next generation of news on the web, and beyond

In sum, despite the promise of digital news platforms, traditional newspaper models have struggled in a web-based environment. National newspapers have transitioned to digital platforms, but success is largely isolated among the few largest newspapers. Local newspapers, on the other hand, are continuing to decline, and local communities are suffering as a consequence.

In aggregate, prior research, and the illustrations provided in this chapter, underscore the ongoing turmoil within the news media industry. The continuing growth of new platforms shows that change is expected to continue, and in many cases, news is moving away from the web towards mobile and application based platforms.

Echoes of social networking

The role of the web is perhaps as unclear today as it was in 1995; this is particularly true in light of recent trends. For instance, today SNSs are mainstream sources of information. In 2015 Facebook reported that it had 1 billion active users in a single day. SNSs are shifting the way that they provide consumers with news. In recent years, SNSs have been a key driver of consumers to newspaper websites. For instance, a 2014 Pew Report found that 30% of the general population cited Facebook as the place they turned to for news; in turn, those users are then connected via hyperlinks to specific newspaper websites (Anderson and Caumont, 2014).

In 2015, moreover, a number of SNSs have developed new platforms that provide consumers with news from content partners within the SNSs' own application. SNSs have quickly developed into platforms where users can discover and engage with news (Winter et al., 2015), and it is increasingly unnecessary to go to other websites. Thus, distributed content networks that draw content from partners onto SNSs keeps users on sites such as Facebook, Twitter or Snapchat, and drives revenue for content providers through in-application advertising (Benton, 2015). Facebook introduced its Instant Articles platform, and Apple introduced its Apple News application; both provide content from third-party providers directly in the application. Through partnership agreements, some of the applications also provide consumers with exclusive content. For example, Apple formed a partnership with Wired Magazine that included exclusive access to certain feature news articles. SNSs are quickly shifting to control the distribution of news media, and moving to a system whereby news media is distributed by a given platform's own application.

Mobility, automation and everything after

The growth of SNSs as a platform for news, and the increasing prominence of mobile applications, further underscores a move away from the distribution of content solely via the web. For instance, of the

54.5 million users who accessed USA Today online in January 2015, 34 million of those users accessed the website via mobile devices. In 2012, 39% of survey respondents indicated they received their news in the past 24 hours from a mobile device (Kohut et al., 2012). Mobile platforms often access information via the web, but are increasingly reliant on applications. The growth of media applications has created yet another new avenue for news distribution. In another vein, the move to digital news on the web has increased the speed with which consumers receive information, but future iterations are likely to see the provision of news automated to a certain degree. Indeed, the move to mobility and application-based news is also bringing forth a new focus on algorithms and the automation of information flow. Algorithms are increasingly being used to automatically decide what news consumers see based on their reading habits (Mysiani, 2013). Others are working to develop algorithms that produce summaries of content and routine news stories such as financial summaries and sports recaps, automating the actual production of the news (Sood et al., 2007; Liu and Birnbaum, 2008).

Why the web's history of news matters

In the midst of previous and forthcoming change, the web provides a critical resource for examining the nature of change for newspapers. Not only does the history of the web allow for sensemaking of previous changes, but it also provides context for understanding future change. As a history of news, the web also provides a critical record for understanding what actually was. For example, archived internet records provide one of the best records of the *Rocky Mountain News*, a daily Denver newspaper that ceased publication in 2009. The history of the web can even provide a critical perspective on news events.

For example, on 1 May 2003, then-President George Bush stood on the deck of the USS Abraham Lincoln and declared the end of combat operations in Iraq following Operation Iraqi Freedom. The Internet Archive captured the subsequent press release on 6 May 2003.[3]

When it became clear that combat would continue in Iraq, the press release was modified on WhiteHouse.gov. A capture from 1 October 2003 shows the change.[4] Despite apparent attempts to modify the narrative of historical events, the archived web provides a critical record of actual history; the headlines in the two screen captures show the attempt to change the narrative, as the headline was modified to

read ‘major combat operations’, as opposed to ‘combat operations’ in the original press release.

The web is a living history; changes to the web are a story in and of themselves. As this chapter has illustrated, the web is also a tool for exploring the nature of change over time. Whether the focus is on a single story, a single newspaper or an entire ecosystem of information, there is validity and power in being able to trace history through the web.

5

International hyperlinks in online news media

Josh Cowls and Jonathan Bright

Introduction: international news coverage on- and off-line

The disjuncture between 'the world outside and the picture in our heads' remains as relevant today as it was when Walter Lippman described it in 1922. At that time, a dramatically more efficient means of mass communication – radio – was making the spreading of news much faster and easier compared with what had come before; news of the outbreak of the First World War spread in hours and days, rather than weeks. Yet, Lippman argued, this quantitative shift in speed of transmission was not matched by a qualitative shift in the nature of mass communication. At the outbreak of war there was still an interval in which the picture of Europe held in people's heads did not correspond to the new reality of Europe at war (Lippmann, 1922).

Lippman's notion has been refined in the decades since, in parallel with the emergence of still more efficient communication platforms. As research has continued, however, attention has also started to shift from the *chronological* interval between the occurrence of an event and the transmission of news of it to the *contextual* interval between the myriad potentially newsworthy events occurring every day, and the far smaller number which are actually reported to the public at large. Media networks play a powerful 'gatekeeping' role by deciding which stories are given valuable airtime and column inches (Shoemaker, 1991); in turn, these decisions are shown to have a substantial effect on the formation of public opinion (McCombs and

Shaw, 1972, 1993) in a variety of areas, including their perceptions of international affairs.

The most dramatic contemporary shift in patterns of news consumption has been the emergence of online media as important locations for where people receive their news. Traditional media organizations have, by and large, persisted in the internet era: major print, radio and television outlets from the pre-internet age have been largely successful in transferring their success online. And this transfer has brought readers with them: a recent survey found that 68% of those who read newspapers do so online at least some of the time (OXIS, 2013: 45).

One important yet understudied consequence of this transfer is the embedding of hyperlinks in news content, which have the potential to direct readers to other sources of which they were not necessarily aware. Whilst most news organizations naturally want readers to remain on their websites, they also frequently link out to other websites, often to provide factual support or background to a story, 'placing news events in a thematic frame' (Coddington, 2012). The choices these news organizations make about which sites to link out to are, potentially, of huge significance, as the volume of web readership they receive means that they can boost the profile of another website considerably if they choose to link to it. For example, the BBC News website, the focus of this chapter, was viewed by just under 100 million people a month in 2014[1]: a link outwards from the BBC website therefore has considerable value in terms of directing traffic. Such choices are particularly important in terms of the debate about Lippman's notion of the world outside, as they can lead individuals straight to websites about the countries referred to in a given article. In other words, they provide the potential to enhance what Norris and Inglehart (2009) called 'cosmopolitan communications'.

This chapter seeks to explain patterns in news website outlinking practices, placing a particular focus on the country of origin of the website being linked to. It is structured in the following way. In the next section, we develop elements of a theory of outlinks, which necessarily relates to international news coverage patterns in general. We then describe the data used in our study: a large collection of news articles collected from a web archive of the BBC News website. We then test two analytical models which attempt to explain outlinking practices. Our results are discussed in the context of an internet which helps to globalize patterns of information consumption and news reading.

Theorizing international outlinking patterns

As Coddington has highlighted, the major reason for news websites to link outwards to other websites, especially in different countries, is as a way of providing context or support to a story they are publishing themselves. Hence any theory of international outlinking also has to take into account explanations for patterns in international coverage.

Many studies have set out to investigate international news coverage, seeking to explain what factors drive coverage of different stories in different countries around the world. Chang et al. (1987) draw the useful distinction between 'context-oriented' factors – that is, factors relating to the national and international context within which particular events occur – and 'event-oriented' factors, which relate to the nature of the news events themselves. We will review each of these factors in turn here.

Within the remit of *contextual* factors, previous studies have investigated the effect of an array of variables – economic, demographic, geographic and political – on the distribution of coverage. Multiple studies show that economic factors are especially significant in predicting coverage. A large-scale analysis of news coverage on multiple media platforms in 38 countries (Wu, 2000) found that the volume of trade between two countries was (alongside the presence or absence of news agency bureaus) the most significant variable driving coverage. This finding was replicated in a later study which also encompassed news websites (Wu, 2007). A country's economic development can also be an influential factor: Golan's (2008) analysis of US television found that, as well as trade, national GDP can predict coverage of African countries, while Kim and Barnett's (1996) analysis of international newspaper trade data found economic development to be the most significant of a host of contextual factors.

Other studies have investigated the role of demographic and geographic factors. When population is used as a variable it is usually found to have some predictive power (Dupree, 1971; Ishii, 1996; Charles et al., 1979) though this typically accounts for less influence than other factors. The distance between two countries has also been shown to partially drive coverage; a 1987 literature review of research in this area suggested 'a relatively stable pattern of foreign news coverage [which] above all is characterized by "regionalism", i.e. a preference for news from nearby countries' (Wilke, 1987: 150). Analysis of the *New York Times*' coverage of foreign disasters similarly found that distance from the USA was the only pertinent contextual factor (van Belle, 2000), and

the regionalism effect was also detected in Oceania (Nnaemeka and Richstad, 1980). This finding was moderated in Chang and Lee (2009), which found that geographic proximity was significant only for television coverage, not newspapers.

Finally, political factors have sometimes been found to play a role. The 'relevance' of a country to the USA was found to be significant in Chang et al. (1987) (albeit this was measured in a somewhat limited fashion as a dichotomous variable). More systematic analysis of six African newspapers in 1981 found that former colonial ties still exert some influence over coverage, although this is related more to coverage of fellow former colonies in Africa than ex-colonial powers in Western Europe (Skurnik, 1981). Meyer (1989) also found 'neoimperial' effects in the flow of news, in relation to former French and British colonies in Africa and the sphere of US influence in Latin America.

The volume of previous research that has tested for the effects of *event-oriented* variables is slight compared to contextual factors. Nonetheless, some studies have found these types of variables to be significant. Chang et al. (1987) built on the concept of event 'deviance', developed in Shoemaker et al. (1986), as a factor that could explain news coverage. Of the seven contextual and event-oriented variables they tested for in a content analysis of American newspapers and television broadcasts, they found that both the normative deviance of an event in relation to the norms of the USA, and the potential for social change in the country in which the event took place, were among the most statistically significant factors explaining coverage. Thus events which *would* threaten domestic norms were they to occur at home, and events which *did* pose the possibility of social change within the foreign country, were both more likely to be covered. However, a later study which sought to replicate this earlier work with data from 1994 found that only the latter factor was significant in newspaper and television coverage ten years on (Chang and Lee, 2009). ('Loss of lives or property', a variable added to the 1994 data analysis, was found to be statistically significant for newspaper but not television coverage.)

Other studies have assessed the effect of other event-oriented factors. Van Belle (2000) found that the number of people killed in natural disasters is a statistically significant factor in the volume of coverage of that event. Golan and Wanta (2003) found that, in election coverage, elections were significantly more likely to be covered in regions where conflict was taking place. Golan (2008: 53) showed that the majority of American coverage of African stories 'focused on negative and highly deviant issues such as conflict and disasters both natural and human

caused', although overall rates of death among a country's population was not significant in relative levels of coverage between countries.

In this study, we test for the effects of many of the contextual and event-oriented factors outlined here. However, we also expect that these factors do not explain everything there is to know about online hyperlinking. Rather, we also expect these linking practices to obey certain logics of their own, within the overall structures conditioning international news reporting. Several aspects are worth considering here.

First and most obvious is the language of the website being linked to, with news websites likely to favour external sites that have the same language as they do. While obvious and mechanical, this hypothesis nevertheless has significant implications, as it means news readers are much more likely to learn more about countries with which they share a language through this mechanism. The second factor is the number of websites available relevant to the country of interest: countries with a larger digital presence are more likely to attract web links. This is again significant as larger and more developed countries inevitably have more of a web presence.

Third, a variety of other more subtle factors about the perceived trustworthiness of the content being linked to may come into play. This may relate to background knowledge the journalists themselves have about the country in question, or perceptions generated by reading websites related to any given country. Finally, it is important to note that we restrict our study here to the coverage of only one news organization, BBC News. The outlinking decisions taken by BBC journalists are undoubtedly also shaped to some degree by the characteristics of the organization: as an established, esteemed, publicly funded broadcaster, the priorities of and pressures on BBC reportage are likely to diverge from, say, up-and-coming and/or commercially funded news outlets. One strength of our single-organization approach is that these factors are controlled for across different national domains, but one trade-off is that we are not able to generalize fully to a wider array of broadcasting organizations.

Data, methods and descriptive statistics

In this chapter, we test these propositions by focusing on the case of BBC News Online. Before describing the dataset collected in more detail, it is worth reflecting a little on this organization. The migration of traditional news media organizations on to the internet has typically been

uneven and often inchoate. This is due in part to confusion – on the part of governments as well as news organizations themselves – over the increasingly sophisticated affordances of the web, in relation to existing broadcasting technology. Traditional broadcasting organizations are licensed by governments, a policy which has its roots in the original 'scarcity' of broadcast frequencies (Moe, 2003). In this context, the web's increasing support for audio and video playback led to tortured definitions of what constitutes 'broadcasting' on the internet, as in the case of Australia's state broadcaster ABC (Martin, 2005).

The experience of bringing the BBC online was also somewhat uneven. The Conservative government's original aim in the mid-1990s was for the BBC's web presence to be commercial (Born, 2003). The last minute decision of the then-BBC director general John Birt to pull out of a commercial deal in 1996 was described by a BBC executive as 'the most important thing [Birt] ever did' (Connor, 2007). In 1999, this shift towards a public service provision was solidified with the BBC's submission to the licence fee review panel; significantly, the first core element of the online provision was 'the provision of news and information' (Graf, 2004: 69). In practice, too, the technical development of the BBC's online public service offering was driven largely by real-world news events: the 1996 budget, the 1997 and 2001 general elections, the terrorist attacks of September 2001 in New York and July 2005 in London and the Indian Ocean tsunami of 2004 all yielded new capacities and approaches for the BBC website (Thorsen, 2010).

To date, the BBC has continued to innovate and iterate its online services, now firmly under the rubric of public service delivery. The BBC's digital services were grouped under the Future Media division in 2011 after a restructure, and innovation efforts continue in the BBC News Labs project. As part of measures aimed at cutting the online budget by 25% by 2013, many subsections of the BBC's website were taken down, yet 'high quality news' remained the top of the list of the corporation's revised online strategy in 2011 (Huggers, 2011). The BBC's continued investment and innovation has been vindicated by its consistent popularity among UK web users: at the time of writing, it was the seventh most visited site in the UK, and the only British organization represented among the top ten most visited sites in the UK.[2] As a large and prominent media organization, the BBC has navigated initial confusion over the status of public broadcasters online – as well as recent budget cutbacks – to sustain a popular, resourceful web presence over the course of 20 years, with the reporting of domestic and international news as its flagship function.

The size and prominence of the BBC makes it an excellent case study with which to test our hypotheses. However, it should also be noted that this case does come with certain compromises. First, as the BBC's newsgathering activity must meet with stringent editorial standards, its hyperlinking should as well – suggesting that material which the BBC links to should not be objectionable. (Although it is noteworthy that under the list of 'Related Internet Links' common to BBC news stories in our period of investigation, the phrase 'The BBC is not responsible for the content of external internet sites' appears as a disclaimer, suggesting that these standards are perhaps not as complete as for the content actually published by the BBC.) These standards will naturally differ in different organizations. Second, the BBC operates an automatic external link generation system which contributes some of the external links found on its web pages, especially those relating to foreign news organizations.[3] Again, this system is rather unique to the BBC. Both of these factors decrease the generalizability of our findings.

In order to test our hypotheses, particularly those relating to sporadic and infrequent events across multiple countries, a dataset which covers as wide a time period as possible is required. For this reason, we chose to collect our data from the Internet Archive (IA), an organization which has been capturing and archiving web pages since 1996 (Kahle, 1997). The IA made available a large set of data on web pages specifically emerging from the .uk country level domain, which constitutes the 'JISC UK Web Domain Dataset'.[4] From this dataset, a set of hyperlinks was extracted during the course of a separate project (see Hale et al., 2014), together with the text to which the hyperlink was attached. These data were then filtered out to include only links emerging from the BBC itself. The web archive dataset is considerable, containing data from almost 17.5 million BBC news pages. It has excellent coverage for the period 2002–2010, when the BBC was visited and archived on average 354 days per year, and reasonable coverage for the period 1999–2001, when on average 205 days per year were captured (not much was archived before 1999). It is difficult to estimate, however, the absolute coverage of the hyperlink dataset, as we do not know to what extent archival visits to the BBC were complete (i.e. the IA may have saved some of the pages but not all of them). However, we have no reason to suspect that the IA's visits were biased to including coverage of one country more than another.

These data are used to create the two major variables used in the dataset. First, we count the number of links made from the BBC website to other country specific 'top-level domains' (TLDs) (across the entire time period of the archive).[5] Such links frequently appear to provide

extra background and context to ongoing news stories. For example, the BBC often links to the government page of a particular country if it is reporting on a news story from that country; or it might link to the website of a particular organization, if the story is about an organization. A top-level domain, in a general sense, is the last part of a hyperlink which indicates the top level of the website in question. For example, the '.fr' in www.lemonde.fr indicates that the website has a French top-level domain. In our analysis, a TLD is taken to include all second-level content (such as .edu.au, where .au indicates the TLD for Australia and .edu indicates Australia's academic SLD). We focus our analysis solely on 'country code top-level domains' (ccTLDs) – that is, TLDs which are reserved for countries and other recognized territories. As such, for practical reasons we exclude generic TLDs such as .com, which are typically country-neutral (although other research has suggested ways to incorporate the .com domain into studies of the international hyperlink network: cf Barnett et al., 2011). Moreover, we restrict our analysis to ccTLDs which can be unequivocally linked to one country, removing ccTLDs which have come to be used for non-country specific purposes. For example, the '.tv' domain is partially owned by the island nation of Tuvalu, but since the government's leasing of the TLD in 1999, it has frequently been used for websites which aim to broadcast television and video content.[6] In total that left us with 222 ccTLDs which had at least one outlink from the BBC website in our dataset.

Second, we counted the number of times each country was mentioned in the text of links to news articles found on the BBC News pages. News articles themselves were identified on the basis of a previously developed schema used in other research (Bright and Nicholls, 2014; Bright, 2015). The text of the link is most frequently the title of the news article, and hence can be used as a means of identifying what the article is about. Based on a list of country names, and common abbreviations for those countries, we checked each title to see how many times a country had been mentioned. This provides an indication of the level of coverage that country receives.

Of course, this method is not a perfect proxy: in particular, it is likely to understate the total amount of coverage each country receives, because not every article about a country will have the name of the country within it (for example, it might refer instead to the capital of the country, or that country's prime minister). However, we do not expect this understatement to be uneven across different countries, hence as a measure of the relative distribution of coverage between countries we still expect this to be valid.

One issue to highlight with the dataset is that the IA's method of archiving pages is quite ad-hoc, based on a web 'spider' which crawls over the internet following hyperlinks from one page to another. There is therefore no guarantee that the same page will be archived consistently over time. Furthermore, as we highlighted above, the volume of pages captured is also not constant over time. However, we do not expect these sampling issues to affect one country disproportionately, since our analysis relies solely on the presence of BBC News pages in the IA over time. We need not assume that the BBC News website was captured in its entirety throughout the period, because we see no reason for the IA to have 'over-captured' BBC News pages covering a particular country compared to any other. Therefore, we still believe that these measures can be used as effective proxies. This is a contention supported by Figure 5.1 below, which shows how absolute counts of outlinks to selected country domains fluctuate over time, but the relative order of countries remains largely unaffected (in Figure 5.1, each point represents the total number of outlinks observed during a one month period in the archive).

We created the following independent variables for this study.[7] Beginning with contextual factors: first, to investigate whether the sheer size of a country influences coverage, we collected data on the total population for each country from World Bank statistics.[8] Second, to assess whether trade flows with the UK affect news coverage, we collected the total combined trade between the UK and other countries from official UK trade data.[9] Third, as a gauge to measure the importance of a country's overall wealth we collected GDP per capita (in current US dollars).[10] Fourth, we collected data on the geographic distance, in kilometres, from London to every other capital city.[11] Fifth, we created the dichotomous variable of whether a country was a member of the Commonwealth of Nations – an intergovernmental organization of member states, most of which were formerly territories of the British Empire – to assess whether the historical legacy of colonization affects modern news coverage or outlinking.

As we highlight above, alongside these general, context-oriented indicators, we also expect factors that are related more specifically to the 'newsworthiness' of a country to have an effect on both the amount of coverage it receives and the amount of links it receives. It is worth noting here that – in contrast to earlier research in this area, which typically gauged the impact of events on a qualitative case-by-case basis – since the dataset we have is so large, we operationalize the 'eventful-ness' of countries by using summary statistics. As such, we introduce three variables which measure a country's newsworthiness, as a proxy for event-oriented

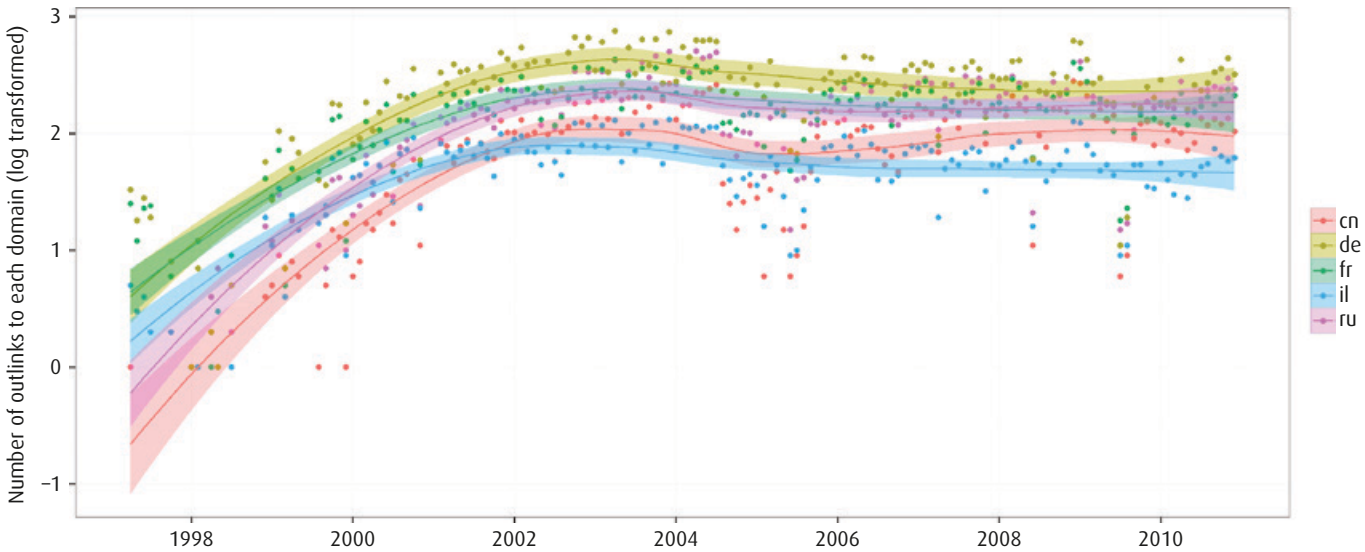

Figure 5.1 Evolution of outlinks to top five country domains over time

hypotheses. First, we look at a country's 'disaster risk', based on data from the World Risk Report.[12] This report measures both the potential for natural disasters such as earthquakes to occur in a country, and the extent to which the country in question is prepared to deal with such an event (see World Risk Report: p. 46). Higher scores on this scale mean a country is at greater risk of disasters. Second, we look at the extent to which a country is 'peaceful', using data from the Global Peace Index.[13] This index measures internal safety and security within a society (taking into account factors such as violent demonstrations), the extent to which it is involved in domestic and international conflicts, and the extent of its militarization (Global Peace Index: p. 2). Higher scores on this scale mean a country is less peaceful. Finally, we measure the crime rate of a country, looking in particular at murder rate statistics provided by the UNODC.[14] This statistic, it should be noted, is also taken into account in the peace index; but it is more specific, focusing solely on internal crime rather than also taking war into account. Higher scores on this scale mean more homicides per 100,000 people within a given country.

Initial descriptive statistics are provided in Table 5.1. As will be apparent from comparing the mean and median values, many of the variables in the dataset (including our key independent variables) are highly skewed. This means both news coverage and outlinking patterns are distributed unevenly, with a small amount of countries receiving a large proportion of the attention. It also suggests that transformations of these variables is appropriate to improve the fit of our statistical models; these transformations are discussed below in the analytical section.

Table 5.1 Descriptive statistics

	Mean	**Median**
Context-oriented variables		
Outlinks (whole period)	17,147	4,096
Mentions (whole period)	1,213	418
Population (2005)	31,410,000	5,904,000
GDP per capita ($) (2005)	11,375	3,172
Distance from London	6,600	6,600
Trade with UK (2005)	2,053,000,000	78,550,000
Event-oriented variables		
Disaster risk (2015)	0.07	0.07
Peace Index (2015)	2.02	1.98
Homicide rate (per 100,000, 2015)	8.80	4.8

Analysis

The main aim of this chapter is to explain outlinking patterns from the BBC to different country top-level domains. In this section, we will explore this question using a series of regression models. As highlighted above, the nature of BBC outlinks means that we expect country coverage to have a significant impact on outlinks themselves, as outlinks are prevalent on news articles, and are themed to the article in question. In fact, a major aim of the chapter is to explain how outlinks vary when taking these differential levels of coverage into account.

Figure 5.2 is a scatter plot of the relationship between country mentions and observed outlinks for the entire time period. It provides strong support for the idea that coverage is a major underlying driver of outlinks, as we might expect. A strong positive correlation between mentions and outlinks can be observed when points are plotted on a log 10 scale (R = 0.72). In other words, as the amount of times a country is mentioned by the BBC goes up, so does the number of outlinks to domains linked to that country. With this in mind, an initial analytical task is to explain news mentions themselves. This is something we tackle

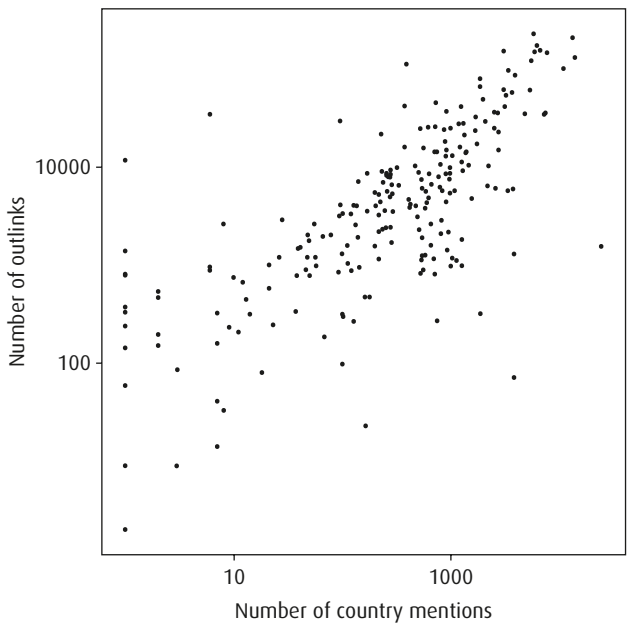

Figure 5.2 Correlation between outlinks and mentions of a country in BBC News Online

Table 5.2 Linear regression model explaining amount of country news mentions on BBC online

Variable	Coefficient	Standard error
Population (log transformed)	0.63***	–0.11
Trade with UK (log transformed)	–0.09	–0.1
GDP per capita (log transformed)	0.32	–0.17
Distance from UK (log transformed)	0.08	–0.14
Homicide rate	–0.02*	–0.01
Peace Index	0.53*	–0.23
Disaster risk	–3.13	–2.59
Commonwealth ember	–0.34	–0.36
Internet penetration	0	–0.01
English as an official or primary language	0.89*	–0.34
adj. R-squared	0.43	
N	148	

in Table 5.2, with a linear regression model, which enables us to assess how multiple explanatory variables might relate to the BBC increasing or decreasing its coverage of a certain country. The dependent variable in this model is the log transformed mentions variable.[15]

Three main drivers of international news coverage can be observed. First, population size is strongly and positively associated with news coverage, with more populous countries receiving more mentions in the news; replicating some previous findings (Dupree, 1971; Charles et al., 1979; Ishii 1996). GDP is found to have a positive correlation, however the p-value is above the conventional cut-off for statistical significance, thus providing no real support for previous findings (Kim and Barnett, 1996; Golan, 2008). Volume of trade with the UK and distance from the UK were also not found to have any effect, despite strong findings in previous research.

In terms of event-driven factors, the Peace Index also shows a strong positive correlation, with less peaceful countries receiving more coverage. This supports earlier research suggesting greater coverage of less peaceful regions (Golan and Wanta, 2003). The homicide rate has a negative correlation which is also statistically significant, although we have no theory to explain why a higher level of homicides might lead to

less coverage. Finally, despite good theoretical cause to believe disasters should increase coverage, we find no evidence for this in our data. Overall, the relative scarcity of evidence for the effects of event-specific coverage here chimes with the fairly limited findings of significance in earlier research.

We also include three variables which we considered of importance in explaining outlinks: membership of the Commonwealth, internet penetration and use of English as a primary or official language. Using these factors in the mentions model is useful in order to provide a basis for comparison in the next model. Commonwealth membership and internet penetration are shown to have no impact on news coverage. However, use of English is shown to be significant and positively associated with increasing news coverage. Of final importance in this model is the adjusted R^2 of 0.41. This indicates that the model explains a reasonable amount of the variance in coverage, but also that a significant portion of it goes unexplained.

We will now move on to look at factors driving outlinking itself. We fit the same model as in Table 5.2, but with news coverage now included as an independent variable. This model, which will allow us to assess factors which seem to drive outlinking whilst controlling for news mentions, is presented in Table 5.3. The dependent variable, outlinks, is again log transformed.[16] It is worth highlighting that the adjusted R^2 of this model is 0.69, meaning that it explains a considerable amount of the total variance in observed outlinks.

As we would expect given Figure 5.2, news mentions are strongly correlated with news outlinks. Population and GDP also continue to be important factors. Interestingly, however, the Peace Index is now negatively correlated with outlinks. Internet penetration in the country of destination also appears as an important factor, as does the use of English as either a primary or official language of the country. Being a member of the Commonwealth has, however, no effect. This analysis underlines our hypothesis that outlinks on news articles are understood to serve a different function for the user than the coverage itself. Where the country being covered has a robust internet infrastructure, with content likely to be in English, journalists and editors may see greater utility in linking to content local to those domains. Where this is not the case – and where non- and inter-governmental organization websites might provide more authoritative sources of additional information, in the case of conflict-ridden countries – we may have detected less of an urge to link to 'native' content.

Table 5.3 Linear regression model explaining amount of country outlinks on BBC online

Variable	Coefficient	Standard error
Population (log transformed)	0.52***	(0.10)
Trade with UK (log transformed)	0.02	(0.08)
GDP per capita (log transformed)	0.11	(0.14)
Distance from UK (log transformed)	0.04	(0.11)
News mentions (log transformed)	0.16*	(0.07)
Homicide rate	0.01	(0.01)
Peace Index	–0.44*	(0.19)
Disaster risk	–3.34	(2.05)
Commonwealth member	–0.41	(0.29)
Internet penetration	0.02***	(0.01)
English as an official or primary language	0.66*	(0.27)
adj. R-squared	0.69	
N	148	

Discussion

This chapter has sought to explain variation in the rate that the news media link outwards to websites in different countries, even when taking into account variations in the news coverage of those countries. Based on a long-term temporal dataset of articles from the BBC News Online website, it has shown that more populous countries and less peaceful countries receive greater levels of news coverage. However, it has also shown that, with the level of news coverage taken into account, less peaceful countries receive fewer outlinks. Countries with greater internet penetration and which use English as a primary or official language are also more likely to receive links.

In closing, we offer a few remarks on the significance of these findings. As we argued in the introduction, the disjuncture between 'the world outside and the picture in our heads' is a key topic in research on international news coverage, and it is also a topic which the emergence of the internet as a crucial venue for news consumption has the potential to revolutionize. Even if major news outlets continue to dominate in the online environment, selective outlinking from these sites has the

potential to inform audiences like never before, as they can be taken to media portals directly within the country they are reading about. Given that the international hyperlink network was perfectly interconnected by 2009, the potential of the hyperlinked web to take users 'closer' to the geographic source of a given news story is clearly possible in principle (Park et al., 2011). However, as we have shown (and as the aforementioned study would suggest), this outlinking is not evenly distributed, or simply a function of how much a given country is in the news. Rather, certain countries do appear to receive more links than others. This means that the impact of these links will be skewed, and the 'picture in our heads' only partially corrected.

Further, it is important to situate this study in the far longer-running stream of analysis of international news coverage. For example, the fact that we have observed here significant differences between the factors explaining coverage, on one hand, and those which explain outlinks, on the other, highlights the value of analysing phenomena specific to the internet as a location of news coverage. Journalists and editors have – perhaps unconsciously – adopted forms of practice relating to the affordances of web-specific phenomena such as hyperlinks. This sort of finding is nothing new to the study of communications in general, which has long demonstrated the importance of the form and affordances of a given medium on how it is used. But it is nonetheless important to note the extension of this phenomenon here, in a study of news coverage on the internet.

It is appropriate to conclude, however, by highlighting again the limitations of this study. Chief among these is the focus only on the BBC, which is an organization with a specific set of values and practices. Furthermore, the identification of 'country relevant' websites is also potentially problematic: not every country uses its top-level domain equally, and many nationally specific websites nevertheless use the generic .com TLD. The study could also be improved by studying temporal variation in outlinks, and rather than just the level observed over the entire period. Finally, we have done little to differentiate between different types of websites within these TLDs, something which may obscure important secondary patterns (for example, how many governments around the world are linked to from the BBC). Further work could usefully pursue these questions, and thus offer us a fuller explanation of the determinants of international hyperlinking practices in online news media.

6
From *far away* to *a click away*: The French state and public services in the 1990s

Valérie Schafer

> […] The first scandal to spark controversy […] came about right after the first TV show mentioning the Net. *Le Grand Secret*, the notorious book by François Mitterrand's personal physician, had just been published; in it, the doctor revealed how he had lied for years about the president's disease. (Chemla,[1] 2002: 172)

All the ingredients for a political and media scandal were present in this Gubler affair: a well-known public figure (the deceased former French President), a secret (finally revealed by the doctor Gubler), a court ruling (stopping book sales) and a private entrepreneur operating an internet café who decided to provide online access to the book in 1996. Immediately relayed through a British website, the book's contents soon found their way on to dozens of servers, in what became a textbook case for the debates on rights and freedoms on the internet (Eko, 2013: 100–1).

Indeed, by the mid-1990s, the legal and political aspects of the internet became apparent with the attempted censoring of newsgroups and of the web. Attempts to establish a legal framework did little to invalidate the notion that politicians fail to understand information and communication technologies – and since the very early days of the web, the involvement of the French state has been highly controversial. The debates hinted at the inability of politicians to grasp the ins and outs of information and communication technologies, being rapidly left behind by the fast pace of innovation, left outside of the global scale of regulations and unaware of the ingenuity of entrepreneurs and users.

In 1994, as the European Bangemann report was promoting the topic of the information society (Bangemann et al., 1994), the French government was nevertheless aware of the promises of the 'New Economy' (Cohen and Debonneuil, 1998) and the issues linked to an 'information society'. In the second half of the decade, the government took a proactive stance, as evident in the 1998 Government Action Plan for the Information Society (or PAGSI for *Plan d'Action Gouvernemental pour la Société de l'Information*). In his address at the summer conference on communication held in Hourtin the previous year, Prime Minister Lionel Jospin claimed that the state should present itself as a driving force and a role model for its citizens. Although the word 'web' was never pronounced (there is only one mention of 'Internet sites', at a time when the vocabulary was not entirely stabilized), the state took part in the development of the web, promoting and encouraging access to state services.

By 2000, the plan was fully implemented, with 600 websites accumulating 5 million hits per month. While some of them provided state-of-the art interactivity and interaction with users, others experienced difficulty in identifying their audience and, as noted in a 2001 report, 'appear(ed) to target everyone – and in the worst cases, no one at all' (DIRE, 2001). For today's historians, this reality is not easily retraced in web archives.[2] Some websites are readily identifiable through their domain names (gouv.fr) or the visual identity of the government, but others prove harder to unearth. A search limited to the domain name is tempting, but yields poor results, as the 34 websites ending in .gouv.fr indexed by the Network Information Center (NIC France) for February 1998 (Internet Archive, 1998) hardly match the figure of 600 websites given by the DIRE report. A number of addresses do not include the .gouv subdomain name (legislative and educational websites, for example).[3]

However, making use of resources such as web archives recovered through the Wayback Machine, newsgroups, oral interviews, state reports, press and audio-visual archives, this chapter describes the relationship between the French state and the web in the second half of the 1990s at different levels, highlighting cultural impediments and state impetus towards the web, legal issues and the heritage of the Minitel. It finally analyses government involvement, at the end of the decade, in the development of a 'French webosphere', how its designers perceived the role of a website, and how users and their needs were understood.

The web loathes a vacuum

> When I first arrived to Matignon,[4] one of my first decisions has been to distribute the *Official Journal* on the Internet for free, which was a great surprise at the time, as the dedicated services were about to launch a paying service. Once the decision was taken, it only took a few months. Why? Because all databases were digitized and we just needed to put them on the web. (Tronc, 2011)

Jean-Noël Tronc, a key actor of the digital-based policy initiated by Prime Minister Lionel Jospin starting in 1997, here provides an explanation to an apparent paradox. With Minitel in the 1980s and the early 1990s, France has been an exception in terms of the wide distribution and appropriation of a culture of online services by the general public. However, France has been slower in its appropriation of the web's potential; early and pioneer content providers, such as administrative services, were reluctant to engage fully in the process. The main reason is that Minitel, and its business model based on the 'Kiosque' system, that relied on the duration of connections rather than on distance, was at the time clearly profitable, while the web struggled to find an economic model. Moreover, faced with the immobility of public and state services, a number of external and peripheral initiatives emerged.

Exogenous and peripheral initiatives

In 1995, Christian Scherer, senior civil servant at the Ministry of Industry, launched Adminet, the first French website focusing on public administration. The reaction of the government was very negative:

> 'In 1995, a number of French embassies had decided to create Internet sites to promote France: tourism, culture, administrative procedures, lyrics for *La Marseillaise* …,' he remembers. 'The Ministry of Interior plainly had the sites shut down. Their motivation: the United States have the Internet, France has the Minitel.' (Desautez, 2000)

Christian Scherer had to shut down a number of pages, even as he was sharing information that was already in the public domain. In particular, he was blamed for publishing samples of the *Official Journal* as a private company, OR Télématique, had had a concession since 1992

from the French state to reproduce the *Official Journal* on CD-ROMs and telematics services and was about to initiate a fee-based Minitel service for retrieving the same material.

> After this very predictable episode, the only legal source on the Internet for a year was the website of Jérôme Rabenou, a Master's degree law student, who had himself taken care of uploading the content of the main legal instruments of our good republic, so as not to infringe any copyright. (Chemla, 2002)

The second case is that of Nicolas Pioch and Weblouvre. What makes the history of this website so strikingly unique is that after its creation by a student in 1994, it gained international fame with a Best of the Web Award (Cern, 1994) in the Best Use of Multiple Media category, alongside Xerox, MIT and the National Center for Supercomputing Applications. The reaction of the Louvre was strongly negative:

> The domain name is owned by us again, recovered from a 'cyber-squatting' engineering student who had taken hold of it for a personal website. Recovering the domain name naturally meant creating a website. (Prot, 2003)

The reaction of the Louvre, forced to hasten its arrival on the web, stirred an outcry in the community of internet users, inside and outside of France (Ponterio, 1995). As for Nicolas Pioch, he had elected to transfer his entire document base during the previous month, from his original server (mistral.enst.fr[5]) to the University of North Carolina and the Tokyo University of Science (Pioch, 1995).

The third case involved the *Société Nationale des Chemins de Fer Français* (the French National Railways), which was confronted with an external initiative from a CNRS (National Center for Scientific Research) researcher who noticed that the SNCF did not offer train schedules online.

> This researcher, acting for the greater good, endeavoured to write the few lines of codes allowing SNCF schedules to be posted online. This he achieved all the more easily as the software used for scheduling […] was perfectly adapted to web development and, by a happy coincidence, available to him.
>
> Unfortunately, the SNCF pressured the CNRS into shutting down the website hosted on its server. Profits from the very

> expensive 3615 SNCF Minitel service all but trumped the satisfaction of travellers, and the competition of a free website giving away information that the Minitel was offering for a price, was entirely unacceptable. (Chemla, 2002: 61–3)

These initiatives, appearing in the period 1995–1996, were blocked in the first and third cases by a particular culture of telematics, but were also more broadly blocked by a political and administrative culture that failed to embrace the internet – yet in fact indistinguishable from the web.

From newsgroups to websites: political and legal issues

The Gubler affair, in early 1996, was only the first of a series of trials marking the entrance of the French internet and web into years of legal and political wrangles. In March of the same year, the Union of Jewish Students of France (UEJF for *Union des Étudiants Juifs de France*) engaged in a legal action against nine internet service providers (ISPs). All ISPs claimed their neutrality and lack of responsibility, while they argued for specificities – Compuserve, for example, clarified that it was not an internet provider,[6] but rather a 'competitor to the Internet' (Bortzmeyer, 1996). ISPs formed a united front against the prospect of a filtering system: 'In terms of filtering Internet content, it's all or nothing. It is impossible to filter selectively (Axone/IBM).' 'It as well considers that a service provider is only a conduit, neutral to the information conveyed (Oléane)' (Bortzmeyer, 1996).

That same year, the managers of Francenet and Worldnet were indicted for circulation of child pornography through their newsgroups and servers (INA, 1996a).

The government reacted to these affairs with a bill proposed by the Minister of Post and Telecommunication, François Fillon, protecting intermediaries from legal action in cases of acts and content that do not fall within their responsibility. However, the Minister also proposed the creation of a public law entity with the power to censor content deemed illegal.

The law never came to fruition, much to the relief of the Association of Internet Users (AUI), formed in 1996 and opposed to the creation of an administration tasked with deciding, in lieu of the legal system, which websites should be censored. However, not even a year later, it was the turn of hosting service providers to be put under scrutiny with the Costes affair.

Valentin Lacambre, one of the first free hosting providers for personal websites, including controversial and provocative performance

artist Jean-Louis Costes, stated that within just a few years, he and his company AlternB had been the target of over fifteen lawsuits (resulting in only two convictions) (Lacambre, 2012).

In the context of a standoff between innovative regulations and the implementation of older measures, and faced with new online expressions of illegal and criminal activities, such as the glorification of terrorism (INA, 1995a), unchecked sales of prescription drugs (INA, 1996b), fraud and scams, child pornography and the like, the state was tempted to search for a stricter legal framework – especially as the issues highlighted by the legal cases of the end of the decade (sales of Nazi memorabilia on Yahoo!, incrimination of the website Front14 which hosted over 300 websites advocating Nazism) were a matter of ethics as much as they were political affairs. Although the legal issues were the main highlights of the reports that the state commissioned at the time, the government simultaneously tried to address other issues such as the impact of networks on the French economy, on small- and medium-sized enterprises or on public administration.

A reluctant administrative culture

When tasked in 1998 to report on the impact of the internet on the modernization of state administration (Baquiast, 1998), Jean-Paul Baquiast had an opportunity to assess how obstacles and constraints could be overcome, and he was quickly faced with scepticism:

> Your report will join the pile of reports on the Internet in France drafted over the past three years, barely read and forgotten as soon as published. […] To start with, the necessary funds will never be made available – and in any case the mindset of the civil servant within the administration, and that of citizens themselves, are at the polar opposite of the Internet mindset. (Baquiast, 1998)

Baquiast was aware that there was some truth to these arguments, as the number and quality of the personal computers used by the French administration were clearly insufficient, and the 'administrative culture' was not yet ready to adapt to networks:

> In many domains, people avoid initiative when it comes to public authorities, deemed too distant or too stiff. […] These tools are

> not designed for the application of orders and instructions, like a computer charging a taxpayer. These are tools of questioning and invention. (Baquiast, 1998)

How could this new mindset be translated into administrative practices, within the administration and in its relation with citizens? For Jean-Noël Tronc, the answer lay in the engagement of the French state:

> The first thing that strikes me when I arrive in the Prime Minister's offices in Matignon is that there is barely any computer equipment. There is no network. Back then secretaries would show up in hallways with 3.5″ floppy disks with the contents of the files.
>
> I ask for a computer, which I'm given without too much complication. I ask for a printer, and they tell me I have a secretary and don't need a printer.
>
> For me the first role of the state is to send a message. And especially in a country like France, where everyone is a critic when it comes to political power, a lot is expected from the state in terms of showing the way. (Tronc in Hallier and Rassat, 2007)

1997–1998: The impetus

An entire generation of TV viewers remembers the episode of the *Guignols de l'Info* (the French spitting image, a satirical TV programme) in 1997 on Canal+ lampooning the disarray of President Jacques Chirac when faced with a computer mouse trying to surf the web (Les guignols de l'info, 1997). The satire helped solidify the notion, still prevalent today, that politicians are incompetent in technological matters.

Entering the information society

However, at the highest level, 1997 was the year when the intention to move toward the internet and the web was first explicitly stated. Even before that, the state had engaged in an analysis of what was then called the 'information highway', following the language coined in the USA. Yet, in accordance with French habits, the analysis was to be undertaken by the former Director of Telecommunications, Gérard Théry, who had supported the development of the Minitel system. While it established a basis for an understanding of the issues to come, the Théry report (1994)

remained highly critical of the internet, and lacked any insight into how quickly it was developing (in addition to failing to mention the web):

> It does not include any security system. […] The delivery of messages is not guaranteed. High traffic may jam the system for minutes or even hours, and lead to the loss of messages. Lastly, there exists no directory of users or services. Word of mouth appears to be the most common mode of operation of this network.
>
> Additionally, no billing systems exist on the Internet, outside of subscription to services, which are then accessed through a password. This makes the network poorly adapted to commercial services. The global revenue for its services amounts to only a twelfth of Minitel's.
>
> The limits of the Internet show that it may not, in the long term, constitute in and by itself the global network of highways. (Théry, 1994: 17)

The arguments made here are fairly common in the rhetoric of French telecoms since the very beginnings of the internet, especially concerning the quality of services and the poor reliability of data transfer (Russell and Schafer, 2014). Nevertheless, the Théry report was followed in 1994 by a call for proposals for experiments relating to new services on information highways (Curtil, 1996: 41). However, 1996 and 1997 were also the years when a 'bouquet of reports', as Adminet dubbed their abundance (Adminet, n.d.), would fully blossom. Although Jean-Noël Tronc confirms that elected officials struggled in their approach to the internet and the web, he describes the indifference of politicians toward digital affairs as follows:

> The state is composed of three tiers: the major players, the decision-makers, where no one really sees the issue. The second tier is the Minister's offices, where people like Sorbier, Baquiast, Scherrer, myself, Isabelle,[7] strongly feel that something must be done. And there is a third tier made up of lesser known individuals keeping to themselves, who are moving forward. […] There are folks who created dre.org without the knowledge of their central administration, and who exchange information by email while diplomats keep using the diplomatic cable, a clunky thing where everything is typed in upper case, there are no accents, everything passes through a cipher […]. In large administrations, you could find people who started to move forward. Similarly, within local

> administrations, there are a number of pioneering elected officials […]. (Tronc, 2011)

Even before the 1998 mission report from Henri d'Attilio and its emphasis on how 'local administrations have a decisive role to play in accelerating the advent of the information society' (d'Attilio, 1998); before these parts of the administration started to take advantage of the opportunities afforded by the PAGSI (the Government Action Plan for the Information Society, see below), they benefited from the support of the state, for example within the 'projects of national interest' in 1995 and 1996.

Some of them developed pioneering website experiences, for instance the city of Issy-Les-Moulineaux next to Paris, or the rural township of Parthenay, that wished to offer the image of a 'digitized city' (Eveno, 1998), aiming to build an identity that was damaged in the 1970s (Vidal, 2007: 139). In 1996, the city of Parthenay stands out as particularly innovative in the domain of 'digital citizenship'. After opening one of the first French digital spaces within the town hall, with 20 internet-connected computers freely accessible to all residents, the town became its own ISP in 1996 and took part in the '*1000 micro*' (1000 PCs) operation, giving access, for 300 francs (45 euros per month), to a computer and 200 hours of free access to the local server. Its local intranet, 'In-Town-Net', offered free website hosting. In 1998, over 200 individuals shared content online, leading some to note that 'residents spend more time on In-Town-Net than they do on the Internet. In-Town-Net is a sharing community' (d'Atillio, 1998). The people of Parthenay 'went on In-Town-Net' before they even 'went on the Internet and the web'.

Issy-Les-Moulineaux, mindful of its image as a city invested in digital media (which had brought it a number of awards and labels), launched into battle to protect the 'trademark' Issy registered on 28 February 1996 and its own domain name,[8] and placed itself at the vanguard of personal page hosting with Cyberi, while innovating with an interactive city council (Internet Archive, 1999) (Figure 6.1). However, like Parthenay it was a 'social laboratory for the experimentation of new information technologies' (L'Atelier, 1999) and the two cities were not representative of the general situation.

The Hourtin address and the PAGSI

While some governmental reports showed a growing awareness of what was at stake with the internet and the web, a decisive signal from the

Figure 6.1 Internet Archive. (1999). Cyberi Homepage. Issy-les-Moulineaux. Archived on 29 January 1999 http://web.archive.org/web/19990129025023/ http://www.issy.com/club-int/cyberi.html Last accessed on 2 December 2015

government was still needed: Lionel Jospin's speech in Hourtin offered that cue.

Delivered in August 1997 during the annual 'Université d'été de la communication' [Summer University about Communication], the address argued that entry into the information society would be made through the internet and the web (while the latter was not named, it was present in the speech through use of the term 'sites'). Jospin mentioned the internet twice in the first seconds of his speech. The role of telecommunications, of the Minitel and the motive behind French lateness were explicitly underlined, as the Prime Minister wished

> that France Télécom offer incentives for the progressive migration of the very large number of Minitel services toward the Internet, a migration where the government shall be leading by example. (Jospin, 1997)

The main traits that would define the PAGSI (the Government Action Plan for the Information Society) the following year were hinted at in the address (administrative services on the web, development of ICT training in schools and the like), which was a founding act with immediate political effect as well as an undeniable legacy.

Just one (but most likely more) clicks away

Two years after the PAGSI, the first assessment from the Interministerial Delegation for the Reform of the State (or DIRE for *Délégation interministérielle à la réforme de l'État*), tasked with a yearly review of the state internet services, offered an interesting perspective on the presence of the state on the web.

As Michel Sapin (then Minister of Public Service and State Reform) noted in his foreword, while online services offered a heterogeneous rather than a unified front,[9] 'the public Internet is a reality today, with 600 websites of state services' (DIRE, 2001).

These websites referenced by the DIRE in 2001 boasted an accumulated 5 million visits per month, and were ranked first in Europe for their range and the quality of their information by Andersen Consulting in 2000 and the Maastricht-Amsterdam summer summit of 1999, thus qualifying the idea of a 'French delay'. But, individually, they offered highly contrasting profiles: some were rudimentary, others undecipherable due to the wealth of information; some were regularly updated and maintained, others left unattended.

The web: learning years and childhood illnesses

The methodological approach chosen by the DIRE for its 2001 evaluation deserves some attention: the reviewers assumed the viewpoint of citizens – 'will the user find on the website the information, the service, the resources they are looking for? A website can be a perfectly clear window into an administration, or a satisfying technological effort, and yet fail to meet the needs of users' (DIRE, 2001: 4). Technical and social interactivity was clearly a more important criterion than any quantitative measure, although the synthesis for the study of 142 sites (about a quarter of the existing websites) stressed that they

> remain 'institutional' in the sense that their primary function is the presentation of the administration responsible for their creation. Very few of them (10%) are portals offering first level information and user orientation'. (DIRE, 2001: 4)

Far from a negative assessment, the report highlighted the steep rise in the numbers of views of these websites – from 6 million hits in 1998 to 27 million in 1999. This progress may be linked to the growing number of websites, their improving quality and access to services, as much as

it may be related to the general growth of the number of internet users in France (the latter still slower than the fourfold acceleration of site views).

The report was resolutely optimistic despite the nuanced data, as shown for example in the evaluation of user interfaces and navigation (Table 6.1).

However, the report noted a number of 'childhood diseases' (which were not specific to state-managed websites), such as the lack of user orientation or a poor distribution of information among separate websites. The profusion of administrative desks had its online counterpart, and the lack of administrative continuity was visible in the state's completed online projects.

Another shortcoming noted by the review was the lack of basic information, such as a website summary or the opening hours of a service. The report also warned against over-informative pages and counter-productive information, such as hit counters:

> The presence of view counters on the first page is in general a rather bad idea (one prefecture proudly states upon loading the page that 'You are our 167th visitor,' which is not a lot, and not very significant). (DIRE, 2001)

Some sites were left in an abandoned state. This issue may be correlated with the small size of web teams – usually one to three people – where the scope of the work includes development, content writing and site administration. In addition to their employees, many administrations turned to subcontractors for website management. About 40% of the

Table 6.1 Evaluation of the navigation and user interface of state websites

Navigation and User-friendliness (qualitative analysis)	**Poor**	**Insufficient**	**Satisfying**	**Good/Very good**
	*	**	***	****
Visuals	12%	47%	32%	9%
User interface	16%	37%	42%	6%
Ease of use	14%	38%	43%	4%
Speed	13%	42%	42%	3%

Source: DIRE (2001: 18).

websites for the central administration were hosted internally, but outside contractors were often used for technical, design and development aspects.

> At times they reveal a significant discrepancy between decisions and their execution, primarily on account of the absolutely strict deadlines of government contracts, with time frames seldom under six months. (DIRE, 2001)

A last finger-wagging went out to the bad taste of third-person praises, 'the narcissism of iconography (focused only, for example, on "the superb building" of the service or the promotion of the office director)' and of 'pretentious home pages or irritating Flash animations' (DIRE, 2001). Flash animations had already lost their appeal too; Megan S. Ankerson shows well how they belong to a bustling, pre-internet-bubble age only to be considered, in the early 2000s, ostentatious (Ankerson, 2009). Through its critical reviews, the report also hinted at what a good website should be, singling out a number of noteworthy sites.

Exemplary websites

The authors of the report, unfazed by the impressive quantity alone,[10] placed value on the targeting and positioning of the websites, as well as on the credibility of the data presented and the ease of access to information. The DIRE valued theme-based information across administrations rather than an institutional approach:

> The 'online pamphlet' aspect is often necessary, so far as it provides information about the identity and mission of a service. Still, this is not the priority for users, and can be cumbersome. (DIRE, 2001)

While it may still appear relevant today, the vision expressed in the report was far from obvious for members of the various administrations. The collected statements of 40 agents from all administrative categories in 1999 showed that the use of ICTs was still perceived as an 'image' factor making administrations seem advanced:

> The administration is seen as old-fashioned, outdated and closed to the outside world, it's time for a more modern image, and that's what ICTs are for […]. Still, everyone assumes that Internet users, now a small minority, will never be the majority of users of

> administrative services, and that the necessary work of uploading content and services is an additional workload, as current forms of services should persist and improve. (Marchandise et al., 1999)

In that context, the educational goals that the DIRE set for itself seemed far from unnecessary. The ten 'exemplary websites' presented in the report were selected for their qualities: proper understanding of their target audience and of users' profiles, clever segmentation of services, clear organization of information, easy follow-up on ongoing requests as well as the efficiency of the search engine on the website. Not all were novice websites, and they did seem to benefit from a solid amount of experience: throughout the 1990s, the Strasbourg Board of Education committed itself to videotext and later online services. Its website underwent at least two overhauls before presenting the design applauded by the 2000 DIRE report: in 1997, the homepage displayed a 'Cyber School' theme (Figure 6.2), before turning to a richer content page showing real attention to user orientation, as exemplified by its top menu where teachers and staff, students, school parents and visitors each had their own access.

In December of 1997, another version appeared. Only partially archived (part of the images are lost), it showed a new format, still simple and uncluttered, but where the homepage had obtained a menu (Figure 6.3). In the interval, the Board had put aside the Cyber School theme.

Figure 6.2 Homepage from the Strasbourg Board of Education website, archived by Internet Archive on 12 January 1997 http://web.archive.org/web/19970112024736/ http://www.ac-strasbourg.fr/ Last accessed on 24 July 2015

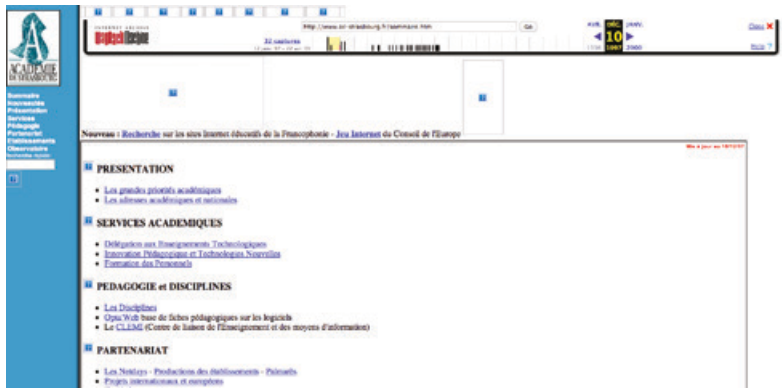

Figure 6.3 Homepage from the Strasbourg Board of Education website, archived by Internet Archive on 10 December 1997 http://web.archive.org/web/19971210212812/ http://www.ac-strasbourg.fr/ Last accessed on 24 July 2015

Figure 6.4 Homepage for the Strasbourg Board of Education, displaying links to one access page for each category of visitor (DIRE, 2001)

Access to the 2000 website, successor to the two previous versions, is not possible any more through the Wayback Machine; fortunately the DIRE report gives us an idea of its design (Figures 6.4 and 6.5) – while giving additional confirmation that it is necessary for historians to cross-reference web archives with other sources.

Figure 6.5 Page from the Strasbourg Board of Education website, archived by Internet Archive on 17 August 2000 http://web.archive.org/web/20000817041856/ http://www.ac-strasbourg.fr/ Last accessed on 24 July 2015

Conclusion

A few years later, Bouquillion and Pailliart still remarked that

> Online democracy firmly remains conventional since, for all the talk on the interactivity of Internet sites, the medium is predominantly used to reproduce information available on other media, primarily the municipal journal. The diffusion of information remains one essential aspect of democratic activities and, in this case, the difference between print media and new technologies is small. […] It allows for the development of the political in its most institutional dimension. (Bouquillion and Pailliart, 2006: 24).

However, within a few years, during the second part of the 1990s, the state was able to take full measure of the challenge, which was still understood as a matter of information more than communication, but stopped being perceived as an outside constraint.

The French approach was clearly one of adaptation and appropriation – one might say of creolization – more than a transposition of US methods and influences. In order to seduce the general public, some ISPs providing web content were indeed betting on the 'French spirit';

this was the case of Club Internet and of Infonie, aware of users' need to have content in their own language at their disposal. The 8pm France 2 televized newscast dedicated a report to these two services, titled 'The Internet, the French way'. In it, Fabrice Sergent underlined that 'Club Internet was in the first place Internet in French, made by the French, for the French', highlighting the role of curation and selection of online content that his service was proposing.

The development of the web in France was the creation of a digital culture inventing itself within national spaces, dealing with the Minitel heritage and the administrative culture, social initiatives and political agendas, in a manner very much related to that Patrice Flichy noted in 1996:

> Unlike Christian Huitema, we do not think that *God created the Internet*, nor that the development of the network of networks is determined by its technical essence. As a matter of fact, the Internet finds itself in the same situation the radio was in the 1910s, or personal computing in the 1970s. It is not a medium yet, but more of a portmanteau-object: the juxtaposition of a number of technical devices and social projects. (Flichy, 1996: 5–6).

PART THREE
CULTURAL AND POLITICAL HISTORIES

7

Welcome to the web: The online community of GeoCities during the early years of the World Wide Web

Ian Milligan

Introduction

As the World Wide Web entered mainstream North American society in the mid- to late 1990s, GeoCities was there to welcome users with open arms.[1] GeoCities helped to facilitate their first steps into publishing, so they could reach previously unimaginable audiences. For the first time, users could create their own web pages without having to worry about the intimidating acronym soup of FTP, HTML, and the like. It was in places like GeoCities where users would become parts of virtual communities held together by volunteers, neighbourhood watches, web rings and guest books. These methods, grounded in the rhetoric of both place and community, helped make the web accessible to tens of millions of users.

GeoCities is dead today, leaving behind little more than its web archive. While in 1999 it was by some counts the web's *third* most popular website, today it is a holding place for Yahoo! advertisements. Saved by the concerted efforts of the Internet Archive, which has a few scrapes going back to the late 1990s, and the Herculean end-of-life efforts of the Archive Team, the digital ruins of this once mighty community today offer rich terrain for historians to explore.

Through a combination of distant, computational reading using web archival analytics platforms such as warcbase (http://warcbase.org) (studying websites as a collective whole, rather than as individual documents) and more focused, targeted reading, this chapter will

address the charge, put forward by several scholars (discussed later in this chapter), that GeoCities was nothing more than an unconnected assemblage of places. I explore what we can learn as we virtually stroll GeoCities' now ghostly 'streets' and 'avenues', from the child-focused EnchantedForest to the festive BourbonStreet. Here, many early web users teased out their relationship with the web, building a foundation for the blogging and social networking explosion that would take place in the new millennium. Together they built a vibrant, interconnected virtual city.

What was GeoCities? A brief history of its rise and fall

What would eventually grow to be millions of websites had simple beginnings. In November 1994, in Beverly Hills, California, a web server flickered to life. David Bohnett, fresh from the software industry and heartbroken by the recent death of his companion, launched a new venture – Beverly Hills Internet – that would let users create their own free web pages. The geographically specific name spoke to the desire for community that lay at the heart of the undertaking. As Bohnett later recalled (as quoted in Ocamb, 2012), 'We all have something to share with each other, which enriches both their lives and ours as well.' Some of the impetus came from Bohnett's own background; he told the *New York Times* (Hansell, 1998) that a lot of what he did had 'to do with being gay and part of a minority that had not had an equal voice in society.' While Beverly Hills Internet was not alone in providing free web hosting, part of a broader trend that included competitors such as Tripod.com (1994) and Angelfire.com (1996), its unique focus on community gave it a distinctive presence on the early web.

In the heady days of the early web, there was a marked desire among users to situate themselves on the web: it was the new 'frontier' sermonized by *Wired* magazine and exalted by technological utopians across the political spectrum, from Newt Gingrich to anarchical socialists (Turner, 2008). The geographical community metaphor meshed well with a public that was conditioned to think of GeoCities, the renamed Beverly Hills Internet, as an ever-expanding geographical space. Five weeks after GeoCities opened, it had received over 600,000 hits and by summer 1995, it was hosting 1,400 websites (Business Wire, 1995). Numbers subsequently skyrocketed (see Figure 7.1).

By mid-1998, the site was one of the top ten draws on the web and was growing by 18,000 new users a day (Motavalli, 2004).

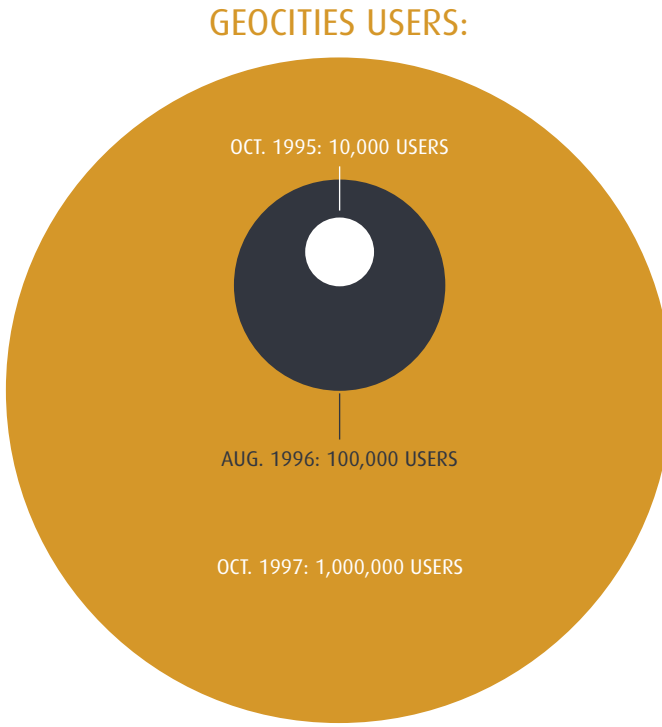

Figure 7.1 The exploding size of GeoCities, 1995–1997

The media began to take notice. Echoing marketing rhetoric, commentators relied on metaphors of space and place. 'What if you want to do more than just look at live images from Hollywood?' asked Roger Ridey (1996) in the English newspaper *The Independent*, 'What if you want to live there? Now you can.' The web was no longer something understood by the public as being a passive area of consumption; it was presented as something that you could live in. Most importantly, it was easy to move in.

If Bohnett and the early web explosion represent the chronological beginning of this chapter, it is bookended by Yahoo!'s purchase of GeoCities. GeoCities went public in August 1998, its share value skyrocketing to around $40 from its initial offer at $17. Yahoo!, a web behemoth then best known for its directory service, began inquiring and in January 1999 purchased GeoCities for $4.6 billion, or $117 a share. This price helps show just how significant GeoCities was seen by many at the time. As John Motavalli (2004: 194) notes,

> At the time of the Yahoo! deal, GeoCities was getting 55 million page views a day, and it was the number-three site, according to Media Metrix. Yahoo! was number one, and AOL was number two. GeoCities called the final sale price a 'kingmaker premium.'

The purchase, however, would also herald significant changes for the site. Yahoo! scrapped the neighbourhood structure that had made GeoCities distinct – rather than having an address, users quickly moved over to URLs based on their usernames – and the site began to decline in attention and user numbers. For these reasons, this study ends then.

If the study ends in 1999, however, the story of GeoCities itself did not. It muddled along under Yahoo's ownership, although media coverage substantially declined almost immediately after its purchase. In 1998 and 1999, respectively, *Lexis|Nexis* has 208 and 247 news items about GeoCities, by 2000 it had dwindled to only 20 and by 2003 only 7. As Yahoo! shifted their business emphasis, they decided in 2009 to shutter GeoCities and delete all user content. While they gave a few months' notice, many of these e-mails would have gone to the e-mail addresses that users signed up to create their websites over ten years ago; there was also no export tool, and to save a website users were encouraged to manually save each page on their website. If it had not been for the efforts of the Internet Archive and Archive Team, an ad-hoc collective of guerrilla archivists, today we would have no record of GeoCities. It would have meant a large gap in our collective understanding of the early web.

As Archive Team declared, 'Yahoo! succeeded in destroying the most amount of history in the shortest amount of time, certainly on purpose, in known memory. Millions of files, user accounts, all gone' (Archive Team, 2009). Their torrent of what they could download *en masse* from GeoCities in 2009 forms the main source base of this chapter, alongside the regular web scrapes that the Internet Archive carried out between 1996 and 2009. It thus forms a relatively unique web archival dataset, available at https://archive.org/details/2009-archiveteam-geocities-part1, that lets us explore a web archive without having to use the Internet Archive's relatively circumscribed Wayback Machine. We also received the final GeoCities scrape from the Internet Archive itself, allowing us to explore and access their web archive files directly. This chapter thus also demonstrates what we can learn from these old web archives, and that they are worth preserving.[2]

Moving into GeoCities: reconstructing first web steps

GeoCities was an experiment in accessible, user-generated content. Users could fill out a straightforward template or a series of forms, making a few clicks here and there, without having to worry about credit card payments or maintenance settings. A GeoCities site was not a work of art, especially by our standards: they were clunky, text heavy, with repetitive backgrounds and garish clipart. But a site offered a powerful publishing platform, the ability to reach a large audience, and in many ways helped realize Berners-Lee's original vision of a read–write web.

For no cost, anybody with an email address could create a GeoCities page with an initial size limit of one megabyte. Accessibility helped GeoCities break a potentially vicious cycle that might have militated against widespread web usage: if people were going to visit the web, they needed meaningful content to view; but for creators to want to generate meaningful content, they needed visitors.

The real key, however, was the neighbourhood system that lay at the heart of GeoCities and to which each free website belonged.[3] I will discuss the neighbourhood concept in depth shortly, but in brief, the first step in establishing a site in GeoCities was to sift through the neighbourhoods one by one, reading up on the sorts of sites each welcomed. For example, the Area 51 neighbourhood welcomed 'Fanzines for Star Trek, The X-Files, The Twilight Zone', among other things.

The explicit attempt to form community through familiar space- and place-based metaphors and rhetoric was GeoCities' hallmark. This did not just take place through the neighbourhood system, although that was critical. GeoCities also attempted to link cyberspace with the 'real' world through the innovative use of web cameras placed in locations such as the intersection of Hollywood and Vine in Beverley Hills, or in Tokyo or Paris. The intent was to amplify 'the sense of place' (Business Wire, 1995). The neighbourhood approach and physical space came together at times. During the 1996 holiday season, for example, a special NorthPole neighbourhood was established for users to launch Christmas-related websites. A webcam simultaneously broadcast a Christmas tree at GeoCities headquarters adorned with comments mailed into the office by users.

The process of doing web history on the 'moving in' process is illuminating. To reconstruct what it was like for future GeoCitizens to take their first steps, we need to use a combination of technological (various text analysis mechanisms, as well as link extraction and image analysis)

and traditional research methods (from closely reading individual web pages to researching media coverage and print resources). For example, web page builders are dynamic and thus eluded the period's web crawlers, so I relied upon traditional print resources.[4]

To create their pages, users had two options back in 1996: they could use a simple template-driven creator, or if they knew HTML they were welcome to use the advanced editor to create a more sophisticated site. The former was akin to the 'Wizard' feature of a Microsoft product (for example, in Word, you might fill out a series of questions to generate a letter template, such as 'who is this letter being addressed to?' and 'what is your address?'). Users entered filenames such as index.html for the home page and anything else for subsequent ones, selected their background and text colours, and then entered the text they wished to see in their body, header and footers. The format accepted HTML input if a user wanted to make something bold or italicized, but also encouraged simple text.

The network effects inherent in GeoCities quickly manifested itself. Users who wanted to learn *how* to use HTML were sent to other users to learn the basics, specifically to http://GeoCities.com/Athens/2090 (hereafter, I will refer to sites by their neighbourhood and address alone). Athens/2090, 'The "Home Page" Home Page' (html_help, 1996), provided straightforward instructions on how to code basic HTML, as well as helpful comparisons to the then-dominant WordPerfect word processing program, which also used markup.

By fall 1998, there were five new ways for users to create their web pages: from the form-based and sponsored 'Intel.com Web Page Wizard' to the GeoBuilder. GeoBuilder was the most significant, helping to democratize free website design and setting the stage for what GeoCities would become. It was a what you see is what you get (WYSIWYG) editor, which let users drag and drop elements such as a text box or a graphic onto the page or template. Occupying similar market space to that of products such as Microsoft's FrontPage, GeoBuilder mixed artistic expression with ease of use. There were many templates to choose from (incidentally similar to today's Wordpress themes): technology focused, academic, social, professional resume/CV, travel diary, personal advertisement, a food website, or a wedding theme. GeoBuilder continued to develop, adding new templates and other options, into 1999, when Yahoo!'s acquisition of GeoCities saw it converted into a downloadable program called PageBuilder (Hill, 2000; Karlins, 2003).

From all of this, we can see the degree to which GeoCities presented itself as an accessible alternative to other web development

options at the time. What can we learn from this massive collection of public speech about online life in the mid- to late 1990s? In the web archives, we can see the broad contours of a community emerge.

Using web archives to explore community

Exploring a dead collection of websites can be eerie, reminiscent of an abandoned cityscape in the films *The Andromeda Strain* or *28 Days Later*. Websites are frozen in time: old guest books, dead links, stopped hit counters, animated GIFs long since pulled from the live web. Yet in these frozen artefacts are the former building blocks of virtual communities, something Internet pioneers saw as early as 1968 as leading to greater happiness because 'the people with whom one interacts most strongly will be selected more by commonality of interests and goals than by accidents of proximity' (Licklider and Taylor, 1968: 30–1).

Community, both offline and online, is difficult to define; communities come in different shapes, from the 'imagined' communities that draw people together by shared media practices (Anderson, 1991), to physical and virtual ones. Constance Porter (2004) defines virtual communities as follows:

> an aggregate of individuals or business partners who interact around a shared interest, where the interaction is at least partially supported and/or mediated by technology and guided by some protocols or norms.

Other scholars contest this emphasis on virtual communities as marketing tools; Lori Kendall (2011) argues that virtual communities are a means to facilitate deeper human connections. In *The Virtual Community*, Howard Rheingold (2000) advanced the following definition of virtual communities:

> social aggregations that emerge from the Net when enough people carry on those public discussions long enough, with sufficient human feeling, to form webs of personal relationships in cyberspace.

He noted in particular the emergence of a gift-based economy, where people give their time without direct reward – although, perhaps, down the road somebody will help them out. It is not enough to simply declare

that community exists, in a website splash page or a press release; it must be enacted, received and perceived as such by members.

In short, community requires effort. As Stephen Doheny-Farina (1996: 37) notes,

> A community is bound by place, which always include complex social and environmental necessities. It is not something you can easily join. You can't subscribe to a community as you subscribe to a discussion group of the Net. It must be lived. It is entwined, contradictory, and involves all our senses.

The sheer ease of joining GeoCities, of firing up PageBuilder and moving into the neighbourhood, has led some scholars to dismiss out of hand the notion that it was a community. Christos Moschovitis (1999) is frank: simply offering web space and email was insufficient, he argues, noting that most joined for the free storage, rather than community elements. True. Certainly many, probably the majority, of GeoCities users were just like that: they signed up, created websites, and did not interact with fellow users *any differently* than they would have with users from other parts of the web. In this they may have been more reminiscent of the suburbanites of Robert Putnam's *Bowling Alone* (2000) – people isolated without sharing civic associations.

Some evidence bears this out. A writer for the online newsmagazine *Salon*, Stephanie Zacharek, discovered this the hard way when she arrived at her new online home in 1999:

> Welcome to my home at GeoCities. I live at 9258 Fashion Avenue, in a neighborhood appropriately called Salon. I moved in here earlier last week because I was told that 'Design, Beauty and Glamour are the toast of Fashion Avenue,' but so far there's not a whiff of glamour to be seen – my neighborhood is a ghost town of hundreds of empty pages, half-started websites and vacant lots; only a handful of the members seem to be at all interested in fashion. (Zacharek, 1999)

While Zacharek was a bit late for the heyday of community, as my explorations reveal here, her point is an important one and captures what may have been a not-uncommon experience. Many users never did get past the 'Under Construction' stage of a brand-new site, as Jason Scott's (Scott, Unknown) collection of construction images aptly reveals.

Yet for a non-trivial minority, we can see traces of virtual community in this web archive. This community structure largely endured between 1995 and 1999; when Yahoo! acquired GeoCities and rearranged the community structure, users moved toward 'vanity' websites (such as http://geocities.com/~janesmith) rather than neighbourhood addresses. But during that earlier time, GeoCities sought to be a new kind of web place for its new arrivals: a place where you learned how to make a first website, with the possibility of friendly neighbours and helpful advice, and might even win a few blinking awards to help bolster your confidence. The web might have seemed infinitely big, but that did not mean you could not have a home there.

Homesteading on the electronic frontier

The central metaphor that governed new GeoCities users was homesteading. It was a consciously chosen metaphor, in keeping with the spirit of the frontier and the heady expansionary rhetoric so common during the web's early days. Think of the Electronic Frontier Foundation, for example, or the many other instances recounted in Fred Turner's *From Counterculture to Cyberculture* (2008). GeoCities' (1997a) central administration defined a homestead in four ways:

> 1. a dwelling with its land and buildings occupied by the owner as a home. 2. any dwelling with its land and buildings where a family makes its home. – v.t. 3. to acquire or settle on (land) as a homestead. – v.i. 4. to acquire or settle on a homestead. – home'stead'er, n.

Each homestead was located in a neighbourhood. This meshed well with the visions of founders Bohnett and John Rezner (the latter joined the team in August 1995 as the technical builder), who saw in 'neighbourhoods, and the people that live in them, the foundation of community' (Sawyer and Greely, 1999: 57–9).

The neighbourhoods and the concept of community were indelibly linked. Surveying a corpus of 1,000 such entries in the Lexis|Nexis database reveals the rise and fall of these two concepts (see Figure 7.2).

The marked decline after 1999 is not surprising; when Yahoo! purchased GeoCities that year, they phased out the neighbourhoods for new entrants. As Olia Lialina (2013), a professor of new media and co-author of the blog *One Terabyte of Kilobyte Age*, has noted: 'Users became isolated'. By 2003, users were asked what topic they were interested in when

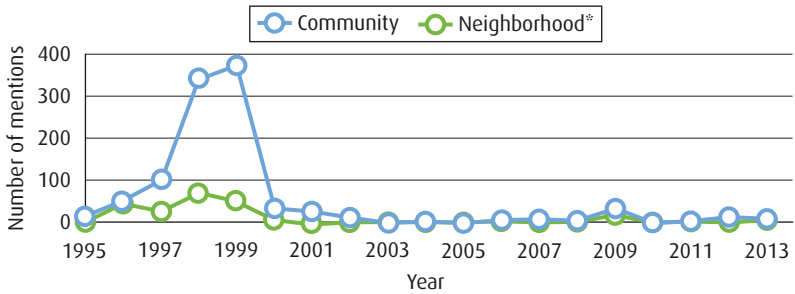

Figure 7.2 Relative frequency of keywords 'Community' and 'Neighborhood' in Lexis|Nexis database, 1995–2013

they created their websites – from alternative lifestyles, computers, the military, pets, romance, science, women and so forth – not to build community, but for the purpose of targeted advertisements (Karlins, 2003). The new GeoCities was very different from what had come before.

Let us return to the late 1990s, when the system was in full swing. When users arrived to create their sites, they were presented with a list of the neighbourhoods they could move into. We have already encountered a few of these places. Those writing about 'education, literature, poetry, philosophy' were encouraged to settle in Athens; political wonks in CapitolHill; small businesspeople or those working from home in Eureka; and so on. Some neighbourhoods came with restrictions and explicit guidance, such as the very protective and regulated EnchantedForest, for young children who wanted their own websites. Others were much wider, such as the largest neighbourhood, Heartland, which focused on 'families, pets, hometown values'. Each enjoined users to settle in, and gave lists of sample topics and websites (in Heartland, for example, in addition to the above three topics, pages about genealogical research and local events were also encouraged).

Popular neighbourhoods filled up quickly, necessitating a sprawl into the 'suburbs': Heartland/Plains or Heartland/Hills were two such destinations. Each neighbourhood or suburb was limited to 9,000 sites (addresses ranged between 1,000–9,999). By 1999, Heartland had 41 suburbs, from the Acres to the Country, the Grove to the Woods. Each had its own support apparatus: community leaders, coding guidelines, web rings, property standards and so forth. Content standards were maintained by the 'Neighbourhood Watch', which was centrally managed by GeoCities (1997b): 'If you notice any of your neighbors not following our policies, please let us know', volunteer watchpeople were directed.

After finding a neighbourhood, users selected their actual address – akin to a street number. If the user wanted '6084', for example, they had to choose the neighbourhood and then see if that particular number was free. If it wasn't, they could either choose a new number or move to one of the emerging suburbs – such as the 'Plains' of Heartland. While the dynamic website that allowed users to pick addresses was not preserved by the Internet Archive, Gordon Graham's *The Internet: A Philosophical Inquiry* (1999) provides a contemporary description:

> Within these townships, each user has a 'homesteading site'; there are users who 'live' next door and others who 'live' further off. All these features can be represented visually. Typically the icons supplied reflect something of the spirit of the township. So, for instance, in *Pentagon* the homesteads are military-style tents, while in *Enchanted Forest* (a site for and by children) the homestead icons are 'cute' cottages. (Graham 1999: 148)

Neighbourhoods, addresses and representations as cottages and tents all comprised the spatial dimension of GeoCities. It was founded on finite land: only one person could hold Heartland/8132, for example, and if addresses ran out suburbs were necessary. The single megabyte of storage came with only one major proviso: 'In order to keep the neighbourhoods a lively and enjoyable place, we would like you to move in within a week after you have received your password and confirmation Email', GeoCities' management advised in a FAQ archived by a user (GeoCities, 1996e). 'Your neighbors would prefer to live next door to someone who has moved in rather than a vacant lot.'

These instructions had significant conceptual overlap with the idea of homesteading. There was only one way to gain more property: continual improvement. *Money* could buy you more storage – you could upgrade to 10 megabytes with the GeoPlus program – but it would not buy you a second address. For that, you had to be a good citizen. 'Part of your responsibility as a resident of GeoCities is to keep your home page fresh and exciting', GeoCities (1996c) explained to those seeking a second site. 'If your original page is kept current, and is consistent with the theme of the neighborhood, you may apply for a second GeoCities address.' John Logie (2002) explored this point in an article in *Rhetoric Society Quarterly*, noting that metaphors within GeoCities aped the central points of the 1862 Homestead Act (US).

The neighbourhoods held GeoCities together. As of late 1996, there were 29 of them. They were an attempt to cluster users based on

pre-existing interests, to facilitate greater traffic within and throughout the community, and to encourage members to use the advertisement-supported infrastructure pages.

Neighbourhood cohesiveness

Exploring the digital ruins of GeoCities today presents unique challenges for historians who use web archives. How can we extract meaningful historical information from such a large set of information? We cannot read every single page, or even a reasonable sample of them. Even if it were possible to view every single picture or read each line of text, by the end of the journey we would have forgotten most things. Computational methods are necessary.

These can range from counting words, which can be useful for the relative frequency of a given word but obscure the context in which a word appears, to more sophisticated approaches such as topic modelling. The latter finds clusters of words that appear frequently together, or topics (Blei et al., 2003). For example, when we write about our families we use words like *husband*, *wife*, *kids*, *pets*, and *home*. Or when we write about work we use words such as *productivity*, *office*, *commute*, *pain*, and *boss* (Jockers, 2011). Latent Dirichlet allocation, or topic modelling, uses a sophisticated mathematical algorithm to go through documents and put the words back into the baskets from which they came. A researcher reading emails in the future might then see two bags of words: *husband*, *wife*, *kids*, and *office*, *commute*, *pain* and call them *home* and *work*, respectively. Without reading individual emails, researchers can gain a sense of what the user wrote about.

We can use a similar method with the neighbourhoods of GeoCities. In Table 7.1, I list the top two topics for a specific subset of neighbourhoods. Neighbourhood place descriptions are from the GeoCities page that invited users to choose which neighbourhood would suit them best. Table 7.1 offers three representative selections.

The data demonstrates that such correlation was not universal, however. The EnchantedForest remained child focused, due in part to the efforts of engaged community leaders in a context of fears around online child exploitation. Pentagon expanded beyond its initial aim of connecting widely deployed and constantly moving military members: it became a forum for military history and for activism and political discussion. Heartland, a significant GeoCities hub, advanced a

Table 7.1 Topics in three selected GeoCities neighbhourhoods

Neighbourhoods	Top Two Topics in each Neighbourhood
Athens *'… based on education, teaching, reading, writing and philosophy'.*	people things time person sense life man work world human good mind soul make nature body case made point part parts goddess witch healing incense witchcraft love energy pagan shaman witches sun spirit protection light circle earth religion
EnchantedForest *'A place for and about kids. Games, stories, educational sites, and homepages created by kids themselves.'*	blue page school home day kids clues fun-time year room birthday family mom jordan play great party friends jq battalion show st jonny horse batteryar-mored lt artillery camp sailor army field col pingu war area quest
Heartland *'A family oriented neighbor-hood that represents Main Street in cyberspace. This is the place to find parenting, pets, and home town values.'*	people time children book years child infor-mation year work make life school person system state world books government good family county church home years informa-tion st city born state war school mrs history birth records great cemetery death

*Topics appear in the neighbourhoods that they should appear in.

particular vision of 'family': focused on the Christian faith, domestic issues, and – significantly – genealogy.

Other metrics also establish significant degrees of cohesiveness. Images extracted from GeoCities give us a sense of how the neighbourhoods worked, as Figure 7.3 demonstrates.

Drawing on the methodologies of Lev Manovitch (2012), I extracted every image from each neighbourhood and arranged them as montages. They need to be used with caution, of course: presented with a randomly arranged montage, we tend to privilege up–down relationships over left–right relationships, even if they are identical (Montello et al., 2003). Yet, there is clear evidence of borrowing and cohesiveness across these communities: the children's community really did have children's pictures, and so forth.

Indeed, if we examine image borrowing – how images travelled around the network – we get results such as those in Figure 7.4.

The animated GIF of Tigger hopping up and down is the 11th most popular image in the EnchantedForest, appearing 48 times. The graph to the right shows that the image is evenly distributed across

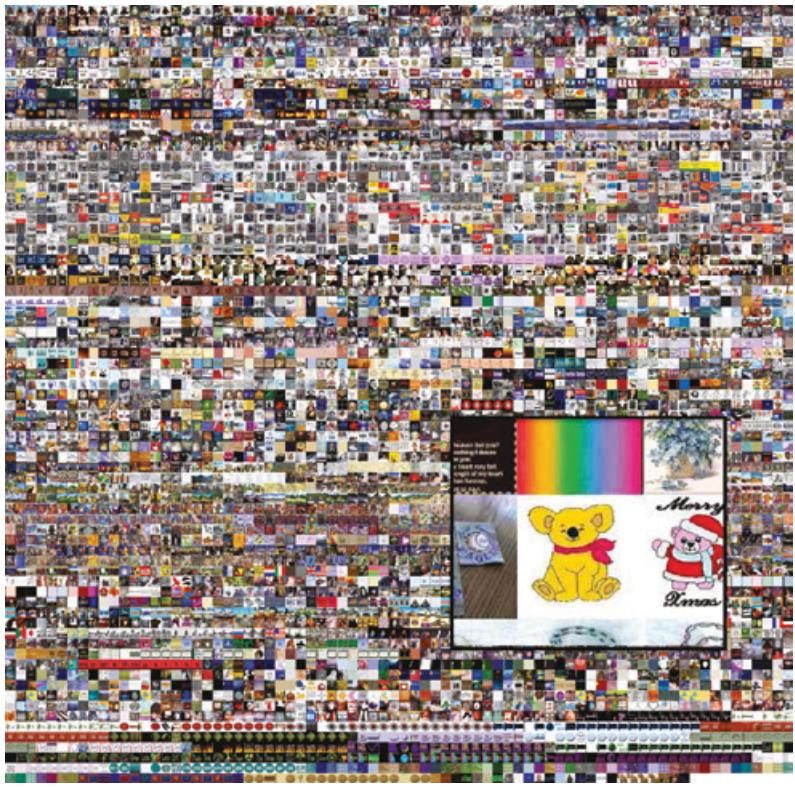

Figure 7.3 Montage of 5,690 images extracted from the EnchantedForest

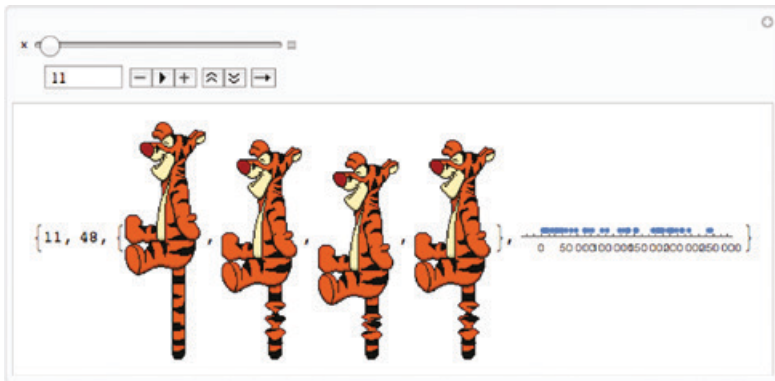

Figure 7.4 Image borrowing in the EnchantedForest

the tens of thousands of individual files that make up the community. People borrowed from each other. This holds true for many GeoCities neighbourhoods. Popular culture communities contain grabs from popular television programmes and movies. Athens, for example, contains a disproportionate number of black-and-white images of historical figures, pointing to the community's educational and philosophical underpinnings.

Finding what we expect to find according to GeoCities' classification of these neighbourhoods is meaningful. Despite the massive array of websites, each zone is relatively homogenous. Heartland was for families; SiliconValley was for computer nerds; and Hollywood dealt with movies, television shows and the like. How did this happen, though? How did these remarkably homogenous communities form online? The neighbourhoods were held together primarily through three methods: community leaders, guest books and community awards. In them, we see the tendrils of community that ran through these websites.

Beyond imposed community: the peer-driven glue

The first method by which GeoCities built communities was 'community leaders'. They helped new users settle into their homesteads, edited newspapers, reviewed websites and provided an accessible human face for people figuring out the World Wide Web. While they provided different services in different communities, in general at the very minimum they were frequent participants in chat rooms, newsgroups and made their emails accessible to users (GeoCities, 1996b). GeoCities (1996d) presented these leaders as a response to user demand – 'many homesteaders have asked us how they can contribute to the development of the GeoCities communit[y]' – but it is unclear whether their role evolved organically or whether the GeoCities leadership team created it. These leaders were selected volunteers who were delegated responsibilities ranging from responding to user emails, to identifying particularly promising sites, policing content guidelines, and acting as the primary intermediary layer between GeoCities management proper and users.

It is testament to the power of community that so many leaders took to the program with such aplomb. Volunteers received few perks: a bit more disk space and a few GeoPoints that could be redeemed for consumer products such as GeoCities clothing. Yet as the program itself admitted, these were miniscule compared to the work asked of

the volunteers: 'If that's the only reason you want to be a leader, think again. It's hard work. Many of our leaders spend several hours each day answering questions and helping their neighbors set up their sites' (GeoCities, 1996d). Applicants were selected based on the quality of their own GeoCities pages, past leadership experience, and an essay on why they would be a good candidate.

After making it through the selection process, the volunteers were assigned a given block of addresses to steward. Some neighbourhoods assigned leaders based on their addresses: for example, if in March 1997 you resided in the 2650–2999 block of the Heartland neighbourhood, your leader would be 'Alison (AKA Alaithea)', who was an expert in a host of things ranging from HTML to Microsoft's Internet Explorer (GeoCities, 1996b). Alison's own website provided information on 'color, layout, navigation, graphics & more', and sensible advice on how to create an attractive website (with still valid advice on the ideal size of text blocks and limiting length of pages). She also provided galleries of attractive backgrounds, even allowing users dynamic previews for their own home pages (Alaithea, 1997). She was the model of a community leader: helpful, generous, accessible and welcoming. Alison also shows how GeoCities provided community leadership roles to women users: in Heartland, 15 of the 25 community leaders were female, drawing on their use of pronouns in their third-person descriptive biographies.

Other neighbourhoods operated on an 'at large' model: each street did not have a dedicated leader but was served instead by a general pool of leaders. Much of Athens, for example, operated on this model (GeoCities, 1996a). Universally, however, these leaders offered help with basic HTML and design and offered themselves as the first contact when users had complaints.

As GeoCities bridged the gap between the earlier model of bulletin board systems – where users could 'yell for SysOp' and actually make the administrators' computers beep to grab their attention – and the more open, impersonal world of the web, these community leaders formed a critical connective tissue. If we download all the descriptions of these 1,040 community leaders and look at keywords, we get a sense of what they offered (see Figure 7.5).

Word clouds – where the more often a word appears in the examined text, the bigger it is in the cloud – are not perfect. For one thing, they obscure context. But they do convey the overall dimensions of the program without bogging us down in a word frequency chart.

Beyond offering help, community leaders facilitated connection by playing an integral part in conferring GeoCities' website awards.

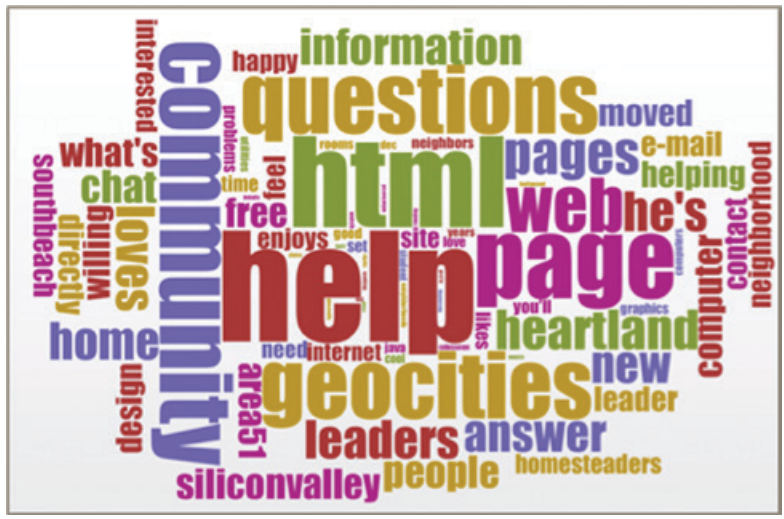

Figure 7.5 Word cloud of all community leader pages, 1996–1997 over six crawls. Generated by http://voyant-tools.org/

A traipse through GeoCities reveals a surprising number of awards, in various shapes and sizes. Official committees of community leaders awarded some, such as the 'Heartland Award of Excellence', voted upon by the volunteer leaders. To get these, new users would submit their web pages for review, a vetting based on whether they adhered to community standards (from having multimedia to having clearly written text), and they would win an award if their pages met a certain threshold. In assigning these awards, community leaders had the ulterior motive of ensuring that sites fit into the prevailing community, that they used efficient and well-written HTML, and that they merged meaningful content with JavaScript and multimedia pop-ups (see, for example, Augusta Golf Neighborhood, n.d.; RainForest Community Leaders, n.d.). Community leaders had explicit instructions to find the 'best sites' in the neighbourhoods to showcase. Other awards were unofficial: users exchanged them to help cement community. Through these exchanges, an internal awards system emerged.

Users could usually click on an award to learn more about it and easily find opportunities to submit or give awards. In any case the community leaders made it clear that potential awards were only a review away. Recipients would often, but not always, receive a badge to adorn their page, as seen in Figure 7.6.

Figure 7.6 Awards taken from a random assortment of websites. From top-left, clockwise, 'Annika's Award' is from Heartland/Hills/9073; 'Chris's Award' from Petsburgh/1098; 'Heartland Heartbeat Award', from Heartland/Lane/8195; 'Best of the "Web '98"', MotorCity/Downs/3148; 'Tropics Choice Award' from TheTropics/5555; 'Heartland Award of Excellence', from Heartland/Bluffs/8336

These awards helped to make community tangible; they were a constant reminder of the webs that tied sites together, woven directly into GeoCities' fabric.

If awards celebrated the 'best' sites and provided a way to exchange favours between users, guest books served as another, less bombastic but equally important, connective tissue between community members.

Seemingly omnipresent throughout websites of the late 1990s and early 2000s, guest books were an important community-building tool for users on the GeoCities platform. They were more than just a way to thank or complement a particularly useful or enjoyable website: for that, there was email. If that mode of communication occupied the 'private' side of the communication spectrum, guest books came in somewhere just short of 'public'. Guest books were not discussion forums: they did not support threaded discussion, replies to authors and so forth.

Coupled with the ubiquitous web page counter (a small set of digits on GeoCities sites that increased by one every time a visitor arrived), guest books were a prime means of evaluating a site's reception. They took various shapes and sizes. At a minimum, they were user-generated snippets: visitors could click on the guest book to fill out a short form with their name, website, email, physical location and a few comments. Users savvy with HTML could incorporate an image into their comment, which led to quite a few advertisements spamming these books.

Why were guest books ubiquitous across GeoCities? A major reason was the decision to include them in the default list of simple add-ons to your website. They were an easy way to facilitate user engagement: designing forms yourself required a level of technical know-how. To install a guest book, members merely had to navigate to the add-ons page, click on 'guest book', provide their site details and then make a few customizations: colour, greetings and questions (GeoCities, 1998). By default, visitors were asked for their name, URL and email address, and guest book owners could add up to nine custom fields.

Guest books played a critical role in community. In her study of personal home pages, carried out in 1998 and published in 2000, sociologist Katherine Walker placed them within the broader genre of web self-presentation. Seeing guest books as akin to the web page counter, Walker argued that they functioned 'as a testament to popularity and a confirmation that others regard the created page and the identity it represents as worthy' (2000: 106). She held that they also played a significant role for the person leaving a comment:

> Leaving a message with an address might lead to response not only from the guest book's owner, but also from others reading the guest book. As such, the audience may potentially receive a greater reward from filling in a guest book than from just sending a private email message. Guest books are a form of role support. (Walker 2000: 106)

Guests often left invitations to visit their own web pages, discussed mutual interests, and provided public email addresses to help them build up a network of contacts and engage the GeoCities public.

Comments were almost universally positive and personalized. When we run textual analysis on these corpuses, overwhelmingly the most common words to emerge are *my*, *you*, *I*, *your*, and other such informal pronouns. *Great*, *love*, *enjoyed*, *thanks*, *wonderful*, and other hyperbole were common instances of gratitude and expression. People liked to thank each other for their content. In more developed form, some of these guest books resembled elaborate questionnaires. Drawing on selective keyword-in-context explorations of the guest books, my research found that questions included, in order of popularity: favourite music, favourite animal, favourite book, favourite website, favourite food, favourite singer, favourite TV show and so forth. Within communities focused on a particular animal, singer, actor or band, the questions became more focused: favourite Shania Twain song, Keanu Reeves movie or dog breed. Through these questionnaires, community was reinforced on a continuous basis.

The URLs that users entered in their guest books are also useful for the web archiving explorer – they represent a sort of calling card indicating where the visitor was from. Were the users coming from all over the web? Or were they GeoCities users commenting and discussing on neighbours' sites? To explore these questions, I extracted all the URLs mentioned in a large sample of guest books. These were mostly the entries provided for the URL or 'my URL' part of a guest book, as well as additional websites that people mentioned in their comments. In total, I extracted 8,147 URLs. In general, GeoCities link structures do not indicate that the community was more cohesive than any other major part of the web – one study compared it to Stanford University sites, which of course have more links to each other than to external sites (Kamvar et al., 2003). Yet when it comes to guest books, we certainly see strong community among users: 43% of links in the guest books came from other GeoCities domains. Given the large numbers of users who would not have their own web pages, or have hosting elsewhere, this is suggestive at least that among a subset of active GeoCities users – those who commented on and provided guest books – there was significant engagement with each other's websites. Unfortunately, as we do not have longitudinal data, it is difficult to see how this might have waxed and waned over time, but it is another factor that helped to contribute to a sense of community.

Conclusions: web archives and the story of community

Between 1994 and 1999, GeoCities users carved out an active online community, preserved as remnants among web archives. This community did not include every user by any means, but rather a sizeable minority of users. Those who sought it out could find meaningful connections within GeoCities: from the community leaders who welcomed them, to the awards they might receive and proudly display on their sites, to the guest books they signed and the invitations they issued.

Through these web archives, limited as they are and circumscribed by a single scrape, we can learn a lot about these digital places. They are the ruins of a robust web community that mattered to the lives of many people. Community leaders volunteered their time, awards were given, web rings connected sites both from necessity and from a desire for connectivity, and neighbours dropped by geographically situated websites to leave friendly messages for other users. GeoCities provided a sense of belonging to a significant minority of users.

There are limits, of course, to this kind of scholarship. Much of the evidentiary basis for this chapter relied upon media coverage of GeoCities, which could have been confused and more importantly susceptible to the dot.com hype cycle. Better contextualization could come from seeing GeoCities within the broader sweep of the 1996–1999 web archive, as well as seeing what connections GeoCities had with the rest of the early web. As the Internet Archive prepares to re-launch their Wayback Machine in 2017 with some form of full-text search, this kind of research will become more accessible. However, access to the underlying WebARChive (ARC and WARC) files that comprise these holdings would be essential to facilitate the sort of research done on GeoCities in this chapter at scale.

Even within GeoCities, however, this chapter also presents the study of these early web archives as a legitimate window onto the lives of the early web and of community more broadly. As a youth and childhood historian by training, I am currently beginning to explore the EnchantedForest more closely, reflecting on what it means to have thousands of historical sources left by children and youth – who, throughout the sweep of historiography, rarely leave sources and need to be understood by adults. Or, a more serious look at the gender dynamics of GeoCities would help inform contemporary discussions around contemporary technical and gaming communities. In short, a serious book is waiting to be written here.

It also sheds light on the broader questions of online communities, of which GeoCities was just a part. GeoCities was and is unique in two respects: first, in its ease of use for everyday web users in the mid-to-late 1990s; secondly, in the ability to download the entire torrent from the Internet Archive to explore as a cohesive whole. It is an unparalleled resource of downloadable content.

Ultimately, the pre-Web 2.0-era is a fascinating one, showing us how user engagement and contributions took shape before the rise of social media. The same desire for connectivity was there, expressed through content, hyperlinks and guest books. Instead of showing appreciation through a 'heart' on Twitter, or a 'like' on Facebook, a handmade Microsoft Paint award was there: more meaningful, perhaps, given the level of detail needed to successfully spread this sort of community. In any case, among the ruins of GeoCities we can see how new web users teased out their relationship to the web. They were not alone but were part of a larger community. Web archives present an interesting opportunity to look back to the days between 1994 and 1999 and to how – spread out across time and space – users figured out what the web would mean to them. GeoCities, a massive assemblage of non-commercialized public speech, presents an interesting introduction to the history of the early World Wide Web – and to the potential found within web archives.

8

Using the web to examine the evolution of the abortion debate in Australia, 2005–2015

Robert Ackland and Ann Evans

Introduction

What can we learn about the evolution of the abortion debate in Australia over the past ten years using data from the World Wide Web? In this chapter, we analyse hyperlink network and website text content data collected in 2005 and 2015 from websites related to the abortion issue in Australia. We use social network analysis (SNA) and quantitative text analysis in an attempt to answer the following questions: Has the relative prominence or visibility of pro-choice and pro-life websites changed? Have other significant sites joined or left the network? Has language used by each side of the debate changed over time? And to what extent do these changes (if any) in the hyperlink network and text content reflect what has happened in the 'real world'?

The politics of abortion has received much attention in the USA and some research attention has been directed at describing the nature of the public debate surrounding pro- and anti-choice campaigning on the web. In order to quantify the abortion debate on the web we examine 'who' speaks and 'what' they communicate. These are two elements of Ferree et al.'s (2002) measures of the quality of discourse. By 'who' we measure the type of organization posting information on the web. 'What' they communicate is determined using a quantitative analysis of the words that are posted.

Our analysis of the hyperlink networks and text content of participants in the abortion debate allows us to provide some interesting

insights into the evolution of abortion as an issue in Australia over the past ten years. In addition, focusing on a particular topic or issue such as abortion allows us to provide evidence on marked changes in web use over the past ten years in Australia and more generally, such as the increasing commercialization of the web and also the shift of activity to social media such as Twitter and Facebook. Our research also demonstrates the difficulty of distinguishing behavioural change on the web relating to social phenomena (such as changing attitudes or policies relating to abortion) from technology-induced behavioural change (resulting from the emergence of social media, for example).

Abortion in Australia

Abortion has been widely available in Australia since the early 1970s. Although available, it was still legislated through various criminal codes rather than through health legislation until the 1990s. In 2015, abortion is legal in most states but is still a highly contested and controversial issue, although the debate does not have the same heat as that in the USA (Albury, 1999). In 2004 the then Federal Health Minister, Tony Abbott, declared an 'abortion epidemic' in Australia. While disputed by some, little data were available at the time to respond to this claim. This led to a parliamentary library report on abortion data collection (Parliamentary Library, 2005). Wyatt and Hughes (2009) argue that conservative politics enabled and encouraged a resurgence of the abortion debate in 2004. However, Siedlecky (2005), a long time pro-choice commentator, insists that the debate never really goes away. McLaren (2013) suggests that the debate in Australia continues because the position of each side of the debate does not change. She argues that this is because the debate is grounded in symbolism and emotion which leads to the language around the arguments also remaining unchanged over time.

Abortion is legislated by individual states and territories and there is a lot of inconsistency across the different jurisdictions. In the ten-year period being examined in this chapter, there has been legislative change in Victoria, Queensland (twice) and in Tasmania. One high profile court case also occurred in this period. In April 2009, a 19-year-old Cairns woman was charged for procuring her own miscarriage. Her partner was charged for assisting her. The case was heard in the Cairns District Court in October 2010, where the jury brought down not guilty verdicts in both charges. The charges related to the use of the drug known as RU486.[1]

At the federal level, there has been a senate debate on transparency in advertising (a bill designed to ensure consumers are aware of the abortion stance of pregnancy counselling services). There has also been a senate debate about removing funding subsidies for second trimester abortions. In 2006 the ban on using RU486 was lifted, paving the way for the provision of a medical alternative to surgical abortion. In 2012 RU486 was registered by the Therapeutic Goods Administration and made available under the Pharmaceutical Benefits Scheme[2] in 2013. Since 2005 there has been continuous debate and legislative change surrounding the provision of abortion services in Australia. Federal changes to the availability and provision of RU486 are evident in the analysis in this chapter.

Hyperlink network and text content analysis – some background

In this section, we provide some context for the two analytical techniques used in this chapter.

Hyperlink network analysis

Hyperlink network analysis involves the construction of a network where the nodes are websites or web pages and the connections between nodes are hyperlinks. In the present research, we focus on networks of websites, since we are more interested in mapping the inferred connections between organizations or groups who are involved in the abortion debate, rather than the connections between individual resources (web pages).

Social scientists recognized the potential of hyperlink networks to provide insights into society in the first phase of the development of the web; what is now known as the Web 1.0 era.[3] However, while Jackson (1997) argued that the idea of using 'a methodology based on the metaphor of a network to examine a communication medium based on the metaphor of a web seems to be so obvious that it threatens to be trivial', the author was not convinced that concepts and methods from social network analysis (SNA) could successfully be implemented with hyperlink networks. In particular, Jackson had concerns whether nodes in a hyperlink network (pages or sites) could reasonably be described as social actors and also whether a hyperlink network could satisfy one of the core assumptions of SNA – the interdependence of actors. Other

authors have similarly expressed caution about the potential use of hyperlink networks for social science research, with Park and Thelwall (2003) noting that hyperlink data can be used to 'potentially discern fingerprints of social relations', and Brügger (2012) has suggested that hyperlink data may need to be supplemented by other data and methods (for example, interviews) in order for a website to be equated with a node and a hyperlink to be equated with a tie, and social network analysis techniques applied.

The evolution of the web from Web 1.0 to Web 2.0, where there is a blurring of the distinction between webmasters and users with social media services such as Facebook and Twitter enabling non-technical people to both produce and consume content, has led to broader interest from social scientists in the web as a source of data for social network research. Ackland (2009) noted that while both theoretical and methodological concerns can make it challenging to regard an unobtrusively-collected hyperlink network as a social network, many of these concerns are not present in the case of networks derived from social network services such as Facebook.

It is now possible to describe a typology of online networks, and the place of hyperlink networks within this typology. For example, Ackland and Zhu (2015) identify two dimensions of ties in online networks (Table 8.1): *directionality* refers to whether a tie between any pair of nodes is directed *versus* undirected, while *manifestation* refers to the substantiality of the relations between nodes, with active acts (e.g. invitation, acceptance) leading to explicit ties, while implicit ties are more inferred (e.g. co-occurrence or interactions). The typology leads to four categories or types of online networks. Networks which are the closest to the classic notion of social networks result from explicitly undirected ties, that is, friendships that require mutual consent to be established (Facebook is an example). Explicitly directed ties involve a one-way, public (or broadcast) mode of relations among users (Twitter is an example). Implicitly undirected ties are inferred by social network analysts post hoc, based on semantic similarity (e.g. co-usage or co-occurrence of keywords or tags) between pairs of nodes (the Flickr photo tagging site is an example). Finally, implicitly directed ties can be extracted from the interactions of people in newsgroups or blogs; these ties are implicit because while a person might reply or respond to another person in a newsgroup, such 'opinion exchanges' are really only inferred connections between the people. Hyperlinks between websites are also examples of implicitly directed ties, since their existence implies a connection between the sites (or the organizations running the sites) but the exact

nature of the connection is generally unknown to the researcher (in the context of large-scale unobtrusive data collection).

Hyperlinks have been described as the 'essence of the Web' (Jackson, 1997; Foot et al., 2003) and their implicit nature means that various interpretations have been ascribed to the existence of a hyperlink between two websites. At one level, a hyperlink can be thought of purely in terms of information provision and hence a sign of authority (Kleinberg, 1999) or trust (Davenport and Cronin, 2000) regarding the information on the page that is being hyperlinked to and the author of the information (the website owner). However, in the context of debate or contention over a social issue (such as abortion), it is also relevant to think of hyperlinks as reflecting communicative or strategic choices (Rogers and Marres, 2000), organizational alliance building and message amplification (Park et al., 2004), and tools for the construction of information public goods in the context of collective action (Fulk et al., 1996; Shumate and Dewitt, 2008) and online collective identity (Ackland and O'Neil, 2011).

We note that all these potential interpretations of the meaning of a hyperlink imply that the tie has a positive effect, that is, the sender of the hyperlink is attempting to confer some positive benefit on the receiver of the hyperlink. As noted by authors such as Brügger (2012), in the absence of some form of analysis of the text surrounding the hyperlink (for example, the 'anchor text'), it is problematic to assume that a hyperlink has positive effect. However, it is technically challenging to conduct such text analysis in the context of large-scale unobtrusive data collection. In their study of the hyperlink networks of refugee and asylum

Table 8.1 Direction and manifestation of ties in online networks

		Direction of ties	
		Undirected	**Directed**
Manifestation of ties	Explicit	Friendship networks	Microblog networks
		(e.g. Facebook, Google+)	(e.g. Twitter, Sina Weibo)
	Implicit	Semantic networks	Threaded conversation & hyperlink networks
		(e.g. recommendation systems, social tagging systems)	(e.g. newsgroups, blogs, WWW hyperlink networks)

Source: from Ackland and Zhu (2015).

seeker advocacy groups in Australia, Lusher and Ackland (2011) purposely did not include government websites such as the Commonwealth Department of Immigration in the analysis (even though this site was linked to by advocacy group websites) precisely for the reason that they wanted to remove negative effect relations from the network (e.g. advocacy groups linking to government policies they do not agree with), which would have complicated the interpretation of hyperlinks in that study.

In studying hyperlink networks of actors who are participants in debates over social or political issues, the focus has been directed to assessing the extent of connections between opposing sides in the debate. Adamic and Glance's (2005) 'Divided they blog' study found evidence of marked polarization in the US political blogosphere. Hargittai et al. (2008) also found substantial polarization among political blogs but no evidence that this was increasing over time, leading the authors to refute the existence of cyberbalkanization – a fragmenting of the online population into narrowly-focused groups of individuals who share similar opinions (Putnam, 2000; Sunstein, 2001).

Another interesting aspect of this research has been whether conservatives and liberals display similar levels of political homophily, or the tendency to connect with actors of similar political persuasion. Adamic and Glance (2005) found some evidence that conservative weblogs tended to cite other conservative weblogs more frequently than liberal weblogs cited other liberal weblogs: 'Through […] visualizations, we see that right-leaning blogs have a denser structure of strong connections than the left, although liberal blogs do have a few exceptionally strong reciprocated connections' (Adamic and Glance, 2005: 40). Ackland and Shorish (2014) did not find evidence of a marked differential political homophily using the 2004 blog data collected by Adamic and Glance (2005), but with a replication of the Adamic and Glance dataset collected in 2011, Ackland and Shorish found that a conservative weblog was around eight times more likely to hyperlink to another conservative weblog, while a liberal blogger was only about four times more likely to link to another liberal blogger.

Text content analysis

The second analytical approach used in this chapter is quantitative analysis of web content, and it is useful to first briefly summarize the key features of this approach.

First, it is necessary to identify the population that is being studied and the sampling approach (this point is also relevant to the construction of the hyperlink network). If the objective of the study is to only collect content from websites of organizations that have an offline presence, then it may be possible to sample from, for example, official registers of such organizations (e.g. a register of non-profit organizations). However, in many examples of Web 1.0 research it is not possible to identify the population from which a sample is being drawn. Authors such as Lusher and Ackland (2011) and Ackland and O'Neil (2011) have identified samples of websites using techniques similar to those proposed by Rogers and Zelman (2002) for researching 'issue networks': entering key words or terms into search engines to identify relevant websites and then using a web crawler to iteratively discover other relevant websites (this is an example of what Rogers and Zelman refer to as 'public trust logics' – finding groups commonly linked to by players trusted to be important in the debate).[4] This technique of using a search engine and web crawler to construct a sample of websites is a form of snowball sampling and it must be emphasized that it does not lead to a representative sample. Caution is therefore required when making inferences about the underlying population, based on analysis of the sample.

Second, one needs to decide whether the focus is on the manifest content (content that exists objectively and unambiguously in the text, i.e. what the author actually wrote) or the latent content (content that is more conceptual and not directly observed in the text, i.e. what the author meant). Social scientific quantitative web content analysis will often involve latent content; for example Ackland et al. (2010) conducted a principal components analysis of content from websites of organizations involved in nanotechnology (manufacturing, research, commercialization) and found three main discourses or orientations: an industrial or proactive discourse (focusing on business, investment and opportunity), a science or education discourse and a social or critical discourse (stressing health risks and the need for political discussion).

Constructing the sample of websites via Google search results

We attempted to use exactly the same approach for data collection in both 2005 and 2015. Our first step was to search Google using the query 'abortion Australia', and collect the top 500 pages returned. The 2005

data were collected in October 2005 (Ackland and Evans, 2005) and the 2015 data were collected in June of that year.

When we collected the 2005 data, Google was the dominant search engine and while there are now significant competitors in the search space (e.g. Bing), Google is still the dominant search engine today.[5] We chose to use Google as our starting point because in 2005 search engines were, and still are today, a first step for many people who are seeking information. Hence, we contend that by using Google we are constructing a sample of websites that people searching for information on abortion are most likely to encounter.[6]

In our present example, it is not possible to identify the population of websites run by organizations engaged in the abortion debate in Australia, and instead we are using the Google search engine to identify web pages which Google ranks as being relevant to the topic, and then we identify our sample of websites (the 'seed sites') from the list of returned web pages. A second step would be to use the web crawler to identify further websites relevant to the study – a website that sends a hyperlink to or receives a hyperlink from one of our seed sites might also be run by an organization or group engaged in the abortion debate in Australia (even if it was not in the list of sites returned by Google).[7] As noted above, it is important to mention that our sample is not representative of the underlying population (which in this case, cannot be identified).

It should also be noted that while our search query was designed to locate web pages focused on the issue of abortion in Australia, we placed no restriction on the actual geographic location of the website or the organization running the website. That is, we did not restrict Google to return pages only from websites in Australia (based on either IP address or country code top-level domain) and similarly, we did not attempt to identify (using the who-is service, for example) the geographic location of the organization. Thus, as will be clear in the following discussion, some of the sites in our sample are not Australian, but they are still participating in the abortion debate in Australia via the nature of the content hosted on their websites.

As discussed above, while our unit of data collected is the web page, our analysis is conducted at the level of the website, and our search resulted in 343 unique websites in 2005 and 376 websites in 2015. We identified websites using the hostname part of the URL. For example, Family Planning New South Wales (NSW) has two web pages that were collected in the 2015 Google search: http://www.fpnsw.org.au/374118_8.html, and http://www.fpnsw.org.au/144423_8.html; in

the analysis, these pages were collapsed to a single website, based on the hostname: www.fpnsw.org.au.[8]

We classified the websites according to abortion stance (Table 8.2) and the type of site or the organization/group running the site (Table 8.3). The classification of abortion stance was done manually. Each site was viewed and a judgement was made about the stance of the organization. This was relatively easy in most cases however some required discussion, and further investigation of the website or the host

Table 8.2 Composition of sites (abortion stance)

	2005		2015	
Stance	**N**	**Prop**	**N**	**Prop**
Neutral	155	0.45	174	0.46
Pro-choice	83	0.24	63	0.17
Pro-life	96	0.28	57	0.15
Unrelated	9	0.03	82	0.22
All	343	1.00	376	1.00

Table 8.3 Composition of sites (site type)

	2005		2015	
Type	**N**	**Prop**	**N**	**Prop**
Abortion provider	8	0.02	8	0.02
Academic	57	0.17	23	0.06
Blogsite	12	0.03	13	0.03
Commercial	12	0.03	74	0.2
Directory/portal	38	0.11	46	0.12
Government	14	0.04	18	0.05
Individual	11	0.03	7	0.02
Info-discussion	18	0.05	15	0.04
Media	41	0.12	76	0.20
NGO	70	0.20	60	0.16
Political party	8	0.02	3	0.01
Politician homepage	4	0.01	0	0.00
Religious organization	32	0.09	15	0.04
Religious-media	18	0.05	10	0.03
Unknown	0	0.00	8	0.02
All	343	1.00	376	1.00

organization. Generally, websites hosting academic articles were classified as neutral unless the article listed in the Google search results was clearly pushing one side of the debate.

Between 2005 and 2015 there was a marked change in the composition of sites returned by the Google searches. The proportion of sites that are 'participants' in the abortion debate (pro-choice or pro-life) decreased from 52% in 2005 to 37% in 2015, with the decrease more pronounced for pro-life sites (Table 8.2). The proportion of sites that are neutral in the abortion debate was roughly constant between the two years, but the proportion of unrelated sites increased markedly from 3% in 2005 to 22% in 2015. This was due to a large increase in the number of spam pages (containing unrelated content) and 'attack' pages (pages or sites identified by Google as potentially hosting malicious code designed to steal private information or otherwise damage computer systems) that appeared in the Google search results.

Table 8.3 shows that the sites ranked by Google as being related to abortion in Australia became less academic (falling from 17 to 6% of the returned sites) and more commercial (growing from 3 to 20% of sites) over the last 10 years. The proportion of media sites increased from 12 to 20%. It is also notable that the presence of political parties and politicians declined; in 2005 there were eight political party sites in the Google search results, but only three sites in 2015. The change for politician websites was even more marked, falling from four in 2005 to none in 2015. Finally, there was a significant decline in religious presence, with the proportion of sites belonging to religious organizations halving from 9 to 4% and the proportion of religious media sites also declining (from 5 to 3%).

The above analysis was of all sites returned by the search query (which collected the first 500 search results), but the reality of search behaviour is that most people do not search beyond the first couple of pages of search results. In order to better assess the visibility of different participants in the abortion debate in Australia (and changes thereof in the past 10 years), Table 8.4 shows the top 20 sites returned for the query, for the two years. The decline in the prominence of pro-life sites is apparent: in 2005 there were six pro-life sites in the top-20, but by 2015 this had halved to three sites, while over the period the number of pro-choice sites in the top-20 remained constant at seven. Perhaps even more tellingly, while there were two pro-life sites in the top-10 in 2005, there were none in 2015 (while over the period the number of pro-choice sites increased from four to five).

Table 8.4 Top-20 sites ranked by Google, 2005 and 2015

	2005			2015		
Rank	**URL**	**Stance**	**Type**	**URL**	**Stance**	**Type**
1	abortion-facts.com	PL	NGO	childrenby-choice.org.au	PC	NGO
2	better-health.vic.gov.au	N	Government	en.wikipedia.org	N	Academic
3	healthinsite.gov.au	N	Directory/portal	betterhealth.vic.gov.au	N	Government
4	wel.org.au	PC	NGO	au.reachout.com	N	NGO
5	mariestopes.com.au	PC	Abortion provider	aph.gov.au	N	Government
6	gynpages.com	PC	Abortion provider	drmarie.org.au	PC	Abortion provider
7	survivorsofabortion.org.au	PL	Religious organization	australia.angloinfo.com	N	Commercial
8	bibpurl.oclc.org	N	Academic	fpnsw.org.au	PC	NGO
9	mhcs.health.nsw.gov.au	N	Government	mja.com.au	PC	Academic
10	atheistfoundation.org.au	PC	NGO	abortion.org.au	PC	Blogsite
11	rtlaust.com	PL	NGO	abc.net.au	N	Media
12	mja.com.au	PC	Academic	health.wa.gov.au	N	Government
13	endeavourforum.org.au	PL	NGO	emilysvoice.com	PL	NGO
14	childrenbychoice.org.au	PC	NGO	theconversation.com	N	Media
15	theage.com.au	N	Media	abortiongrief.asn.au	PL	NGO

(*Continued*)

Table 8.4 (*Contd*.)

	2005			2015		
Rank	**URL**	**Stance**	**Type**	**URL**	**Stance**	**Type**
16	abortion-clinicgold-coast.com	PC	Abortion provider	mariestopes.org.au	PC	NGO
17	cathnews.com	PL	Religious-media	pregnancy-counselling.com.au	N	NGO
18	utas.edu.au	N	Academic	thewomens.org.au	PC	NGO
19	aph.gov.au	N	Government	pregnancysupport.com.au	PL	NGO
20	nswrtl.org.au	PL	NGO	smh.com.au	N	Media

Note: PC – pro-choice, PL – pro-life, N – neutral, U – unknown.

Hyperlink network and website text content analysis

The VOSON software (see Chapter four in Ackland, 2013) incorporates a web crawler, which was used to collect hyperlink and website text content data (meta keywords, body text) in both years. The 2005 hyperlink and text data were collected in October 2005, while the 2015 data were collected in June 2015. So, while this research involves analysis of historical web data (from 2005), the data were collected and archived by the authors using the VOSON software in 2005, rather than via access to institutional repositories of archived web data (we return to this in the discussion section below).

The first step in the data collection involved setting the crawler parameters such that the crawler would visit each of the 'seed' pages returned by the Google searches, collect text content from each page, and then leave the page. That is, in this first step, the crawler was set so it would not iteratively crawl throughout the entire website, but only collect text content from the seed page. This was done for practical reasons (the version of VOSON in 2005 was more limited in the amount of text content it could store) but also for methodological reasons: the Google search engine has returned these pages because they contain text content relevant to the topic of abortion in Australia, and by allowing the

crawler to collect text content from other pages in the website, this is likely to introduce irrelevant text content into the analysis (this is known as topic drift in information retrieval).

The second step in the data collection was the collection of hyperlinks. While text content was collected from all the seed pages, hyperlinks were only collected from the seed pages identified as belonging to websites that are participants in the abortion debate (i.e. either pro-life or pro-choice websites). Again, this was done in order to prevent 'topic drift' – by crawling sites deemed irrelevant to the research topic, we would simply be collecting hyperlink data that would not be used in the research – and also as a means of preserving bandwidth resources. The VOSON crawler only collected outbound hyperlinks, and the crawler stopped when it had collected either 1,000 links to external pages or else had crawled 100 internal pages.

Network-level analysis

As discussed in the previous section, our unit of analysis is the website rather than the web page, and this affects the construction of the hyperlink networks. Specifically, the crawling process results in a network of web pages, but a data processing step reduces this to a network of websites where, as was the case with the Google search data discussed above, nodes in this research are websites (identified by hostname) rather than web pages. Thus, in the case of Family Planning NSW, this organization had 248 web pages in the hyperlink network of web pages (the two seed pages discussed above, and 246 pages that the VOSON crawler identified as being hyperlinked to by various seed pages), however in the network of websites this organization is represented by a single node: www.fpnsw.org.au which reflects all the connections to and from pages in this website.

This process of 'collapsing' from a network of pages to a network of websites results in a significant reduction in the scale of the data. While the 2005 (2015) full network of pages (by 'full', we mean it contains all the seed pages identified by the Google searches and all the new pages identified by crawling these pages) contains 40,776 (71,644) nodes, as shown in Table 8.5, the corresponding full network of websites contains only 13,240 (6,192) nodes.

Table 8.5 shows key network statistics for four networks for each of the two years: the full network, the participant network (pro-life and pro-choice sites), and separate networks for each of the pro-life and pro-choice groups.[9] The first thing to note is that the size of the

Table 8.5 Network statistics

	2005				2015			
Metric	**Full**	**Participants**	**Pro-choice**	**Pro-life**	**Full**	**Participants**	**Pro-choice**	**Pro-life**
Network size	13240	179	83	96	6192	120	63	57
Number components	3	1	1	1	27	3	2	1
Number isolates	83	18	8	13	129	25	14	17
Inclusiveness	0.9937	0.8994	0.9036	0.8646	0.9792	0.7917	0.7778	0.7018
Density	0.0001	0.018	0.0306	0.0319	0.0002	0.0137	0.0256	0.0226
Density*			0.0375	0.0428			0.0425	0.0462
Average indegree			2.506	3.031			1.587	1.263

Note: * – density calculated for subnetwork with isolate nodes removed.

full network halved between 2005 and 2015 (from 13,240 to 6,192 nodes) and it also became more disconnected, with the number of connected components (sets of nodes that are connected) increasing from three to 27 and inclusiveness (the proportion of non-isolated nodes as a proportion of total network size) falling from 99.4 to 97.9%. The conclusion is that over the past ten years, pro-life and pro-choice sites collectively significantly reduced the number of hyperlinks they make to other sites.

The decline in hyperlinking activity is even more apparent when we consider the subnetworks for participants (pro-life and pro-choice), and for these networks we can also see a marked decline in network density, which is the number of ties as a proportion of the total possible number of ties that could exist. Researchers such as Adamic and Glance (2005) have found some evidence that conservative actors create denser online networks, compared with their liberal counterparts. As shown in Table 8.5, the network densities for 2005 for the pro-life and pro-choice subnetworks were very similar (0.0306 for the pro-choice subnetwork, compared with 0.0319 for the pro-life subnetwork). However, once isolates have been removed, there is some evidence that the pro-life network is more densely connected, with pro-life sites in 2005 creating 4.28% of the hyperlinks that potentially could be created and pro-choice sites only creating 3.75% of the potential hyperlinks. This difference remained in 2015 (at least as calculated for the networks with isolates removed).

Table 8.5 also reports average indegree for the pro-choice and pro-life subnetworks, in both years. In 2005, the average pro-choice site received 2.5 inbound hyperlinks from other pro-choice sites, while the average pro-life site received three inbound hyperlinks from other pro-life sites. Thus, in 2005 pro-life sites were on average more active in sending hyperlinks to other pro-life sites, compared with their pro-choice counterparts. By 2015 there had been a drop in hyperlinking activity, most markedly for pro-life sites, with pro-choice (pro-life) sites receiving an average of 1.6 (1.3) inlinks.

The changes in the participant subnetwork are visually apparent in Figures 8.1 and 8.2. In these visualizations, node size is proportional to indegree and node colour reflects abortion stance (pro-life is red, pro-choice is blue). The force-directed graphing algorithm has produced clusters that are very clearly demarcated according to abortion stance, a visual representation of the existence of homophily in hyperlinking behaviour.

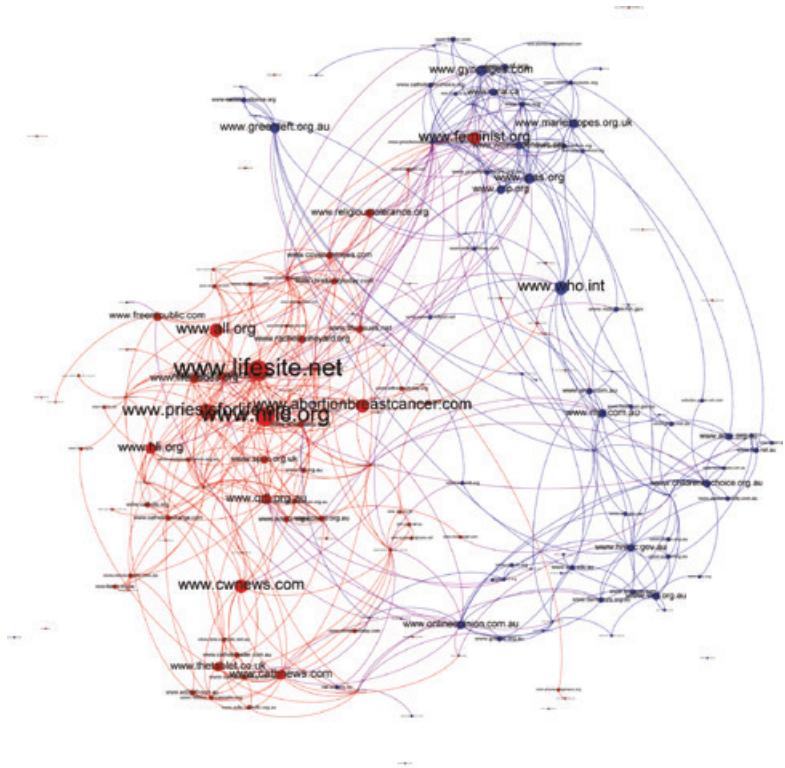

Figure 8.1 Hyperlink network of participants in abortion debate in Australia, 2005. Note: pro-life – red, pro-choice – blue. Node size is proportional to indegree

Prominent sites

There are many different node-level metrics that can be used to identify nodes that are taking significant or prominent roles within a network. In this chapter we focus on the simplest of these measures: indegree (number of inbound hyperlinks) as a measure of visibility and outdegree (number of outbound hyperlinks) as a measure of activity. Table 8.6 shows the top-20 sites by indegree in the full hyperlink networks for the two years. The most striking (but not unexpected) finding is the rise of social media; in 2005 Twitter, Facebook and YouTube either did not exist or had been barely launched, while in 2015 these were the top-three sites in terms of indegree.[10] These sites are prominent because abortion-related sites are providing links to their accounts on social media (e.g. 'follow us on Twitter') but these sites are also providing links

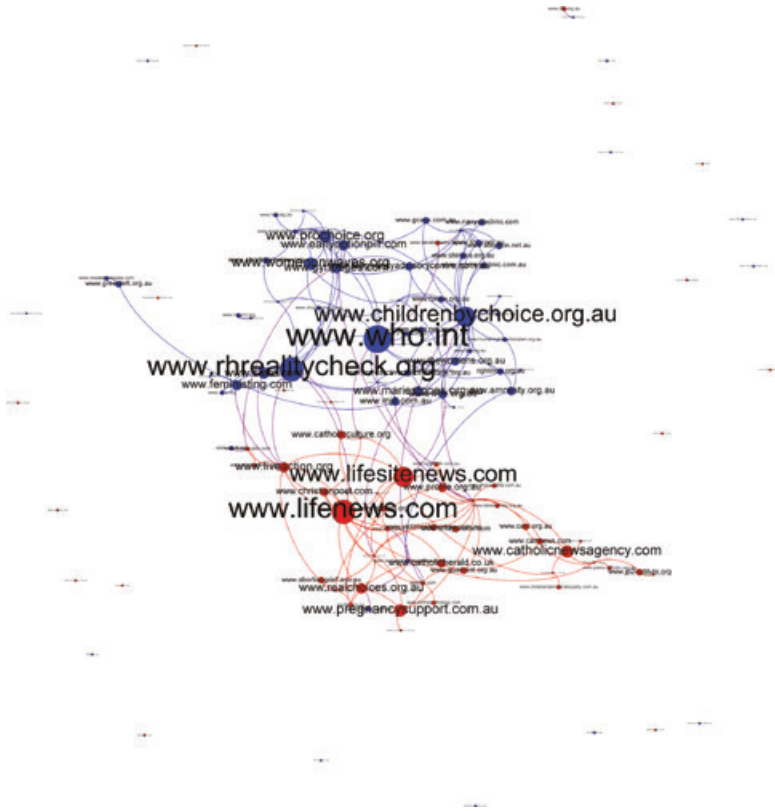

Figure 8.2 Hyperlink network of participants in abortion debate in Australia, 2015. Note: pro-life – red, pro-choice – blue. Node size is proportional to indegree

to resources such as videos on YouTube. Media sites became prominent over the last ten years, with the number of media sites in the top-20 increasing from five to seven, and Australian media sites are relatively more highly ranked in 2015, compared with ten years ago.

The apparent decline of the Web 1.0 presence of pro-life groups identified above is reinforced by Table 8.6; while there were two pro-life sites in the top-20 in 2005, there were none in 2015 (in contrast, there were no top-20 pro-choice sites in 2005, but one in 2015). There are some other interesting findings in Table 8.6 that point to general changes in the web that have occurred over the past decade. For example, two sites that were popular for hosting small websites run by individuals and groups (geocities.com, aol.com) were in the top-20 in 2005 but are no longer providing this service in 2015 (for more on GeoCities,

Table 8.6 Top-20 sites by indegree (full network)

2005				2015			
Indeg	**URL**	**Stance**	**Type**	**Indeg**	**URL**	**Stance**	**Type**
41	abc.net.au	N	Media	103	facebook.com	N	Unknown
37	adobe.com	U	Unknown	89	twitter.com	U	Unknown
35	geocities.com	N	Political party	50	youtube.com	N	Unknown
34	news.bbc.co.uk	N	Media	33	abc.net.au	N	Media
33	amazon.com	N	Academic	30	smh.com.au	N	Media
31	washingtonpost.com	U	Unknown	29	en.wikipedia.org	N	Academic
28	smh.com.au	N	Media	27	linkedin.com	N	Unknown
26	nytimes.com	U	Unknown	26	theage.com.au	N	Media
25	theage.com.au	N	Media	23	theaustralian.com.au	N	Media
24	cnn.com	U	Unknown	20	theguardian.com	N	Media
24	guardian.co.uk	U	Unknown	20	pinterest.com	U	Unknown
23	lifesite.net	PL	NGO	19	ncbi.nlm.nih.gov	N	Government
21	google.com	N	Directory/portal	18	washingtonpost.com	U	Unknown
21	theaustralian.news.com.au	N	Media	18	news.com.au	N	Media
21	un.org	N	Academic	18	amazon.com	N	Commercial
21	abcnews.go.com	U	Unknown	18	nytimes.com	N	Media
21	nrlc.org	PL	NGO	17	childrenbychoice.org.au	PC	NGO
20	aph.gov.au	N	Government	17	instagram.com	U	Unknown
19	msnbc.msn.com	U	Unknown	16	google.com	U	Unknown
19	members.aol.com	U	Unknown	16	heraldsun.com.au	U	Unknown

Note: PC – pro-choice, PL – pro-life, N – neutral, U – unknown.

see Milligan, chapter seven in this volume). It is also notable that in 2005 the second ranked site was adobe.com but in 2015 this site does not make the top-20 as PDFs are ubiquitous and website owners no longer feel the need to provide a link to the Adobe PDF reader.

Table 8.7 shows the top-20 sites ranked by indegree in the abortion debate participant subnetwork, and this table reinforces evidence of the decline of the position and activity of pro-life organizations on Web 1.0. While in 2005 eight of the top-10 sites based on indegree (in the participants' network) were pro-life sites, by 2015 this had declined to just three sites.

Finally, Table 8.8 shows the top-20 sites on the basis of outdegree in the full network and it is apparent that while pro-life sites have declined, relatively, in terms of numbers of sites, they are still active in terms of their linking behaviour, with half of the sites in the top-10 being pro-life (in 2015 six of the top-10 sites were pro-life). From this we can surmise that the relative decline in the visibility of pro-life sites on the web is more due to the decline in numbers of sites, rather than a decline in the number of hyperlinks being created.

Text analysis

Text analysis further deepens our understanding of the patterns described above. The text analysis presented here only involves manifest content (we do not attempt to discern latent content). We focus on what text content is prevalent on abortion-related websites (frequency analysis) and whether these keywords or terms are related to the type of organization behind the website (pro-choice or pro-life). The text analysis involves two types of text extracted from the web pages: 'meta words' are words extracted from the page meta data (keywords, title, description), and 'page words' are words extracted from the body of the web page. In the case of meta words, if a website owner used a pair of words in the meta keyword section of the web page (for example, 'abortion clinic') then the pair of words is treated as a single term (i.e. it will appear as 'abortion_clinic' in the text analysis). However with the page words, only single words are used in the analysis, that is, 'abortion clinic' would be split into two words 'abortion' and 'clinic'. The other thing to note is that the words 'abortion' and 'australia' were excluded since they were likely to be appearing on all of the sites, given the search query, and hence do not add to the analysis.[11]

Table 8.7 Top-20 sites by indegree (participant subnetwork)

2005				2015			
Indeg	**URL**	**Stance**	**Type**	**Indeg**	**URL**	**Stance**	**Type**
23	lifesite.net	PL	NGO	17	childrenbychoice.org.au	PC	NGO
21	nrlc.org	PL	NGO	15	who.int	PC	NGO
16	priestsforlife.org	PL	Religious organization	15	rhrealitycheck.org	PC	Media
14	cwnews.com	PL	Religious-media	13	lifenews.com	PL	Media
14	abortionbreastcancer.com	PL	NGO	10	lifesitenews.com	PL	Media
14	all.org	PL	Religious organization	7	fpnsw.org.au	PC	NGO
14	who.int	PC	NGO	7	mja.com.au	PC	Academic
13	feminist.org	PL	NGO	6	pregnancyadvisorycentre.com.au	PC	Abortion provider
11	hli.org	PL	Religious organization	6	pregnancysupport.com.au	PL	NGO
10	gynpages.com	PC	Abortion provider	6	womenonwaves.org	PC	NGO
10	cathnews.com	PL	Religious-media	5	prochoice.org	PC	NGO
10	ipas.org	PC	NGO	5	catholicnewsagency.com	PL	Religious-media
10	greenleft.org.au	PC	NGO	5	feministing.com	PC	Info-discussion
10	qrtl.org.au	PL	NGO	4	slate.com	PC	Media
9	thetablet.co.uk	PL	Religious-media	4	realchoices.org.au	PL	NGO
9	lifeissues.org	PL	NGO	4	liveaction.org	PL	Blogsite
9	mariestopes.org.uk	PC	NGO	4	mariestopes.org.au	PC	NGO
8	freerepublic.com	PL	NGO	4	earlyoptionpill.com	PC	Commercial
8	mja.com.au	PC	Academic	4	nanyaraclinic.com	PC	Abortion provider
8	onlineopinion.com.au	PC	Info-discussion	4	gynpages.com	PC	Directory/portal

Note: PC – pro-choice, PL – pro-life, N – neutral, U – unknown.

Table 8.8 Top-20 sites by outdegree (full network)

2005				2015			
Outdeg	**URL**	**Stance**	**Type**	**Outdeg**	**URL**	**Stance**	**Type**
877	blogicus.com	PL	Blogsite	510	conservapedia.com	PL	Info-discussion
826	womensenews.org	PC	Media	433	freerepublic.com	PL	Info-discussion
695	trevorcook.typepad.com	PC	Blogsite	425	saltshakers.org.au	PL	Religious organization
656	multiline.com.au	PL	Individual	371	prochoice.org	PC	NGO
572	jonjayray.tripod.com	PL	Blogsite	314	feministing.com	PC	Info-discussion
534	fwhc.org	PC	NGO	297	rhrealitycheck.org	PC	Media
504	covenantnews.com	PL	Religious-media	283	slate.com	PC	Media
486	mwilliams.info	PL	Blogsite	279	gynpages.com	PC	Directory/portal
463	ourcommunity.com.au	PC	Directory/portal	257	liveaction.org	PL	Blogsite
460	prolifeblogs.com	PL	Blogsite	195	cathnews.acu.edu.au	PL	Religious-media
415	johnstonsarchive.net	PL	Individual	175	bioedge.org	PL	Media
366	christianitytoday.com	PL	Religious-media	166	christianpost.com	PL	Religious-media
350	seattlecatholic.com	PL	Religious-media	164	mediaisland.org	PC	NGO
345	gynpages.com	PC	Abortion provider	141	medicalabortionconsortium.org	PC	NGO
345	religioustolerance.org	PL	Religious organization	139	childrenbychoice.org.au	PC	NGO
329	tennesseerighttolife.org	PL	Directory/portal	138	lifenews.com	PL	Media
321	isteve.com	PL	Individual	134	rightnow.org.au	PC	Media
310	prwatch.org	PC	Info-discussion	125	bladesplace.id.au	PC	Blogsite
256	hreoc.gov.au	PC	Government	125	bernardgaynor.com.au	PL	Individual
252	media.anglican.com.au	PL	Religious-media	121	acl.org.au	PL	Religious organization

Note: PC – pro-choice, PL – pro-life, N – neutral, U – unknown.

Figure 8.3 Word cloud (meta words) – pro-choice, 2005

Two types of visualizations are used.[12] Word clouds are a random placement of the words, with size reflecting the number of times the word appeared across all of the sites in the group (pro-choice or pro-life). Comparison clouds provide a means of comparing across groups, by placing the word clouds for both groups on the same page and, importantly, they display the words that are predominantly associated with each group.

In 2005 there was a noticeable difference in the meta words used by pro-life and pro-choice websites (Figures 8.3 and 8.4). The word cloud for pro-choice meta words is dominated by the words health, women, pregnancy, clinic, rights, information, whereas the word cloud for the pro-life meta words is dominated by Catholic, life, prolife, Christian, human, news, family. This shows the obvious association with religion and religious pages linked with the pro-life movement.

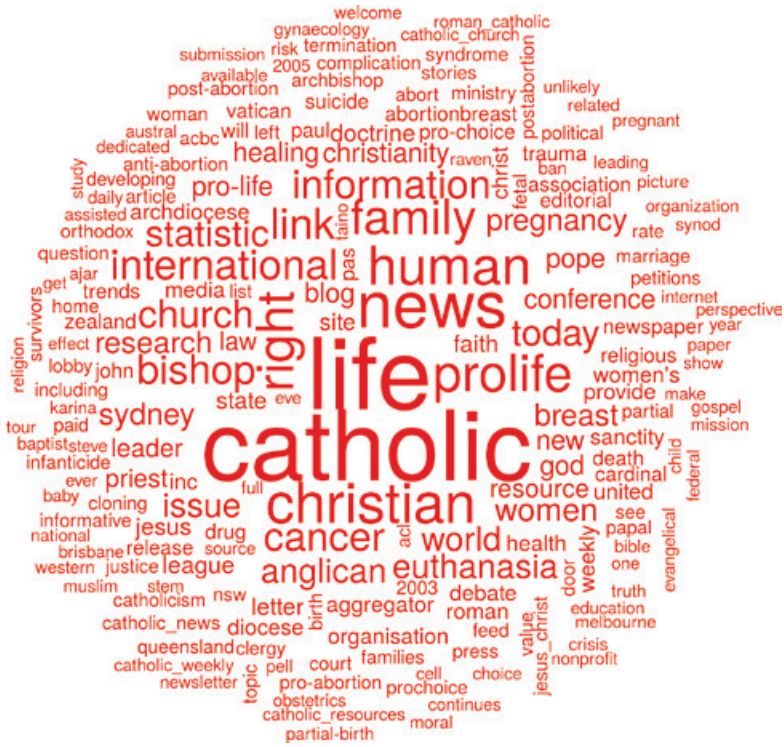

Figure 8.4 Word cloud (meta words) – pro-life, 2005

In 2015 the difference in the type of words still exists (Figures 8.5 and 8.6). However, the websites of both the pro-life and pro-choice sides are using fewer meta words. This likely reflects a change in behaviour of webmasters in response to the fact that meta keywords are no longer as important as they used to be for ensuring appropriate search engine ranking, since search engines now make use of page text (and indeed, other information such as click through behaviour in search results), in addition to meta words.

For reasons of space, the word clouds for the page words are not displayed, but they follow a similar pattern to what was found with meta keywords, in terms of the comparison between pro-choice and pro-life sites. The pro-choice page words emphasize the service and health nature of pregnancy termination (services, access, public, safe, women, right, health). On the other hand, the pro-life page words are more focused on the individual (will, women, children, life, human, child, time). The overall number of page words in the word clouds does not

Figure 8.5 Word cloud (meta words) – pro-choice, 2015

decrease between 2005 and 2015, unlike the results for meta keywords, and this supports our contention that the reduction of meta keywords was a response of webmasters who no longer saw them as being necessary for good search results.

As noted above, the comparison clouds highlight the differences between the language of the two sides of the abortion debate by displaying the words that are associated with each side of the debate. The comparison cloud for meta words in 2005 shows a clear difference with 'Catholic' and 'life' dominating the pro-life side, and 'health' and 'women' dominating the pro-choice language (Figure 8.7). By 2015 the meta words comparison cloud shows a change in the words used by the pro-choice pages, with the following words now dominating: clinical, medical, health (Figure 8.8). The pro-life meta words in 2015 are more dispersed with no clearly dominating language, although religious words are still visible as is the word 'unborn'.

Figure 8.6 Word cloud (meta words) – pro-life, 2015

The unique page words collected from the pro-life sites in 2005 are already related to service provision, but also include criminal and law, referencing the push for legislative change that in 2005 was still to occur (Figure 8.9). The unique pro-life page words in 2005 include religious references, death, babies, human and cancer. The comparison cloud in 2015 shows an even greater focus of pro-choice sites on services relating to abortion while, as was found for the meta keywords, the pro-life sites present a more diffuse set of words with no apparent major themes (Figure 8.10).

The harvesting of meta- and page-words provides the opportunity to add a depth of understanding of the differences between types of organization on the web that cannot be gained with hyperlink analysis on its own. The analysis here shows that pro-life and pro-choice groups use different words and have a different focus on the content of their websites. Pro-choice sites are dominated by information about services, whereas pro-life sites focus on religious beliefs about abortion. A qualitative analysis of these websites has not been conducted and would be a

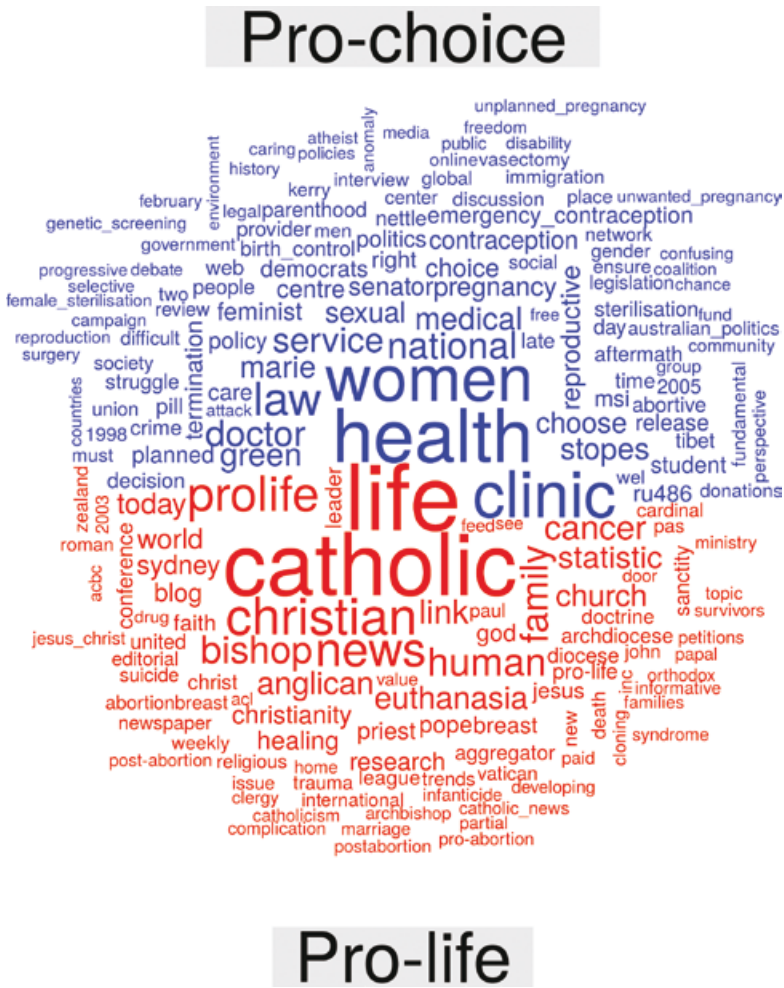

Figure 8.7 Comparison cloud (meta words) – 2005

fruitful endeavour, but one that is beyond the scope of this chapter. We do note, however, that our results are similar in nature to those found in the USA and Germany in an analysis of newspaper text (Ferree et al., 2002) and in Australia (McLaren, 2013) through an analysis of the use of foetal images.

Additionally, the analysis shows the decreasing use of meta words over time as organizations change their web behaviour in light of changing search engine technology. In general, we discerned that the pro-life 'message' became relatively more diffuse over the past ten years.

Figure 8.8 Comparison cloud (meta words) – 2015

Discussion and conclusions

Ten years is a long time, especially on the web. Over the past ten years there have been changes in the social issue in Australia that we have focused on (abortion), but there have been even greater changes in the technological space which is the source of data for our analysis. It is a challenge for us to be able to distinguish changes that originate in the behaviour of the actors we are studying (participants in the abortion

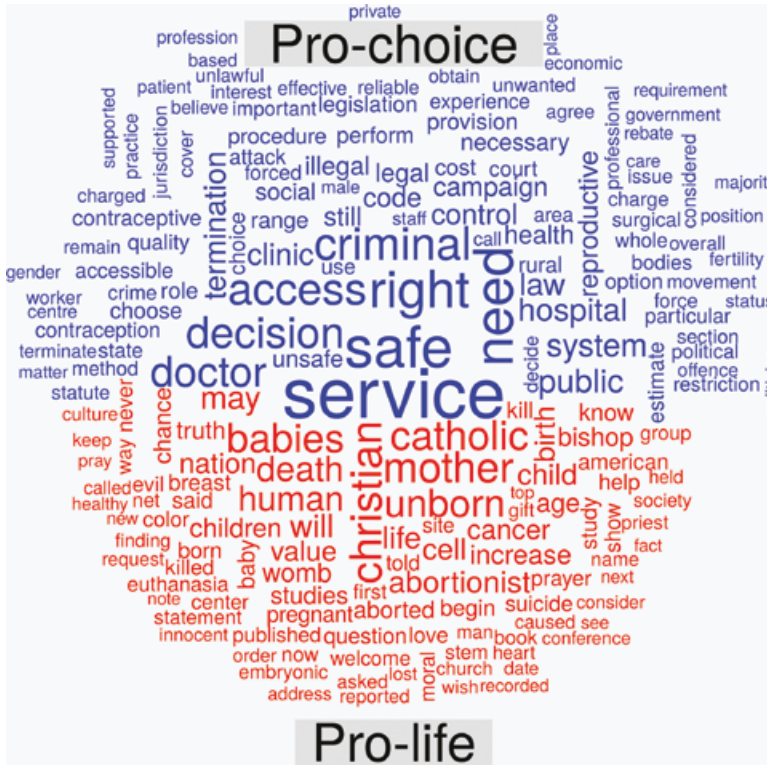

Figure 8.9 Comparison cloud (page words) – 2005

debate) and changes that relate to the technological space in which they operate (the web).

One of the marked changes in the web over the past ten years is that it has become even more commercially oriented, and this is reflected in the number of commerce-related sites appearing in the Google search results. In 2005, abortion drugs had not been approved for use in Australia and so the Google search in that year tended to return pages and sites that were focusing on abortion as a social and policy issue. In contrast, after ten years of legal access to abortion drugs and services, the 2015 search results returned many more sites that were providing access to these services and drugs (and during this period, there has been a marked commercialization of the web). We also noticed a marked increase in the number of spam and attack pages in the Google results.

The other major finding relates to the relative presence of pro-choice and pro-life sites on the Australian web, and how this has changed over the past ten years. Analysis of the Google search results and the

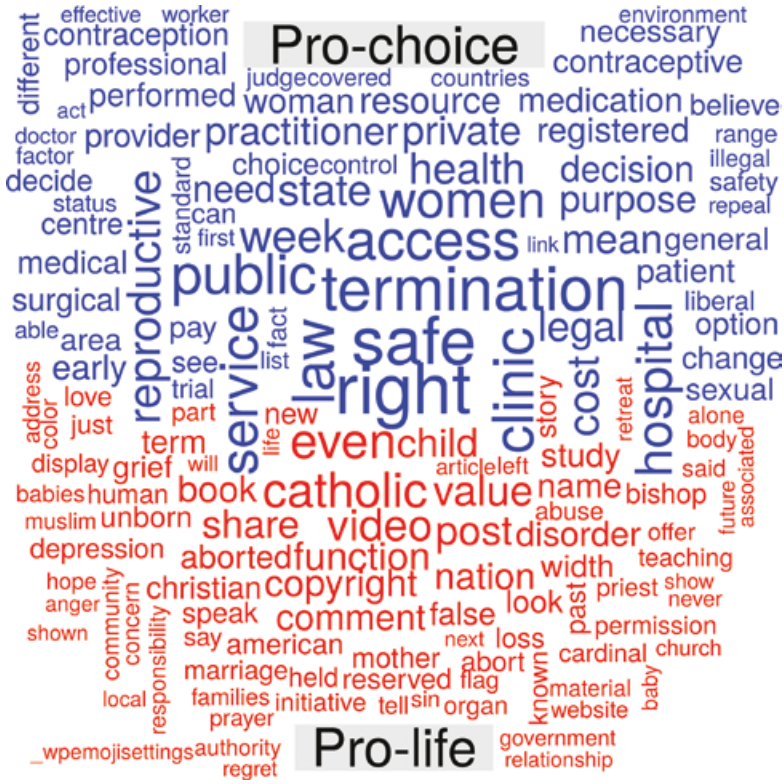

Figure 8.10 Comparison cloud (page words) – 2015

hyperlink networks suggests that both sides of the issue were active and visible on the web in 2005; while the top-ranked site was pro-life, there were roughly equal numbers of pro-choice and pro-life sites in the top-20. So, one cannot say that a particular side of the debate was dominating the web in 2005. This contrasts with the findings of Bounegru (2011) who studied the abortion issue on the Romanian web and found that pro-life sites dominated both the Google search results and the associated hyperlink network constructed using the Issuecrawler software. It is also at odds with Hindman's (2008) contention that the existence of power laws on the web (a marked inequality in the distribution of inlinks, with a few sites receiving the lion's share of inlinks, and hence attention), combined with the fact that Google rankings are (or at least were, in the original formulation of the PageRank algorithm that underlies the Google search engine) largely influenced by inlinks from relevant sites, which meant that social issues such as abortion could become dominated by particular voices on the web.

We found that there has been a marked decline of the presence and visibility of pro-life sites over the past ten years, both in terms of absolute numbers of sites and also their visibility in terms of Google ranking and centrality in the hyperlink network. The content analysis also suggested that the pro-life message became more diffuse over the past ten years.

We attribute the decline in the web presence of pro-life sites to a number of factors. First, the web of 2005 was largely a web without social media, as we know it today. While bloggers were active by 2005 (although it was really only the US presidential election of 2004 where the potential power of the blogosphere to influence the news media cycle became widely recognized), Facebook had not yet emerged from the US college system (Facebook was launched in February 2004), and Twitter was not launched until July 2006. So in 2005, an organization or group wanting to have a web presence needed to run a website. However, even today, websites are costly and technically challenging to run, compared with setting up and using a Facebook or Twitter profile, and so we surmise that one of the reasons for the decline in pro-life sites in the past ten years is the fact that since pro-life groups are likely to be smaller and less well resourced, they have left Web 1.0 in order to focus their activities on social media, which is cheaper and arguably, more effective. This argument is therefore about the technical and economic change in the web from 2005 to 2015, which we believe has differentially impacted on the presence/visibility of pro-life versus pro-choice sites.

The second potential reason why pro-life presence has declined on Web 1.0 is purely related to the social issue of abortion. In Australia the abortion debate has largely been won by the pro-choice side, with abortion services legally and widely available. We contend that for this reason, many pro-life groups have largely left the abortion battleground and are instead focusing their attention on current social issues that are still in policy contention, such as marriage equality. Meanwhile, many pro-choice groups are still active on Web 1.0 because they are involved with service provision, for example, and require a (Web 1.0) web presence for those activities.

Another notable finding is the decline in political parties and individual politicians as a presence in the abortion debate on Web 1.0. As with the above, it is difficult to ascertain the reason why political party and politician web pages did not appear in the Google search in 2015. Is it because in 2005 abortion was a policy issue and hence parties and politicians were making statements while in 2015, abortion is no longer a political issue? Or is it because of technological change, with parties (and in particular politicians) in 2015 focusing on cheaper (and

potentially more effective) social media channels rather than Web 1.0 websites?

Our final comments are about methodology. Historical analysis of Web 1.0 hyperlink networks is challenging. This is because in order to construct large-scale hyperlink networks from web archives, it is necessary that these archives allow crawlers or else provide publicly available application programming interfaces (APIs) so that the hyperlink network data can be programmatically extracted at scale. There does not exist an Australian web archive with such capabilities and hence, we could not have conducted the research presented in this chapter without having crawled the live web at both time points (2005 and 2015), that is, effectively creating a purpose-built archive of hyperlink and website text data. Thus, historical hyperlink network analysis typically requires researchers to collect snapshots from the live web over time.

Conducting comparable web research ten years apart is not straightforward, even when (as in this case) one of the authors is the lead developer of the web crawl software that was used. One technical challenge we faced was that while in 2005 it was possible to use the Google API to find all 500 pages that mentioned our search criteria, the 2015 version of the Google API only allows one to return the first 100 search results (even if one is prepared to pay for API access). So that meant we needed to manually copy and paste the Google search results.

Finally, the fact that we use Google also needs noting. Google was the dominant search engine in 2005 and it is still the dominant search engine in 2015. However, Google is not synonymous with the web (even Web 1.0) and so our finding that, for example, the composition of the sites related to abortion in Australia became markedly more commercial over the past ten years could simply reflect a change in Google's ranking algorithm, that is, that Google is promoting more commercial websites, and not that the web itself has become more commercial. However, we regard the latter as being likely to be true.

Acknowledgements

We would like to thank Eleanor Bettini and Francisca Borquez for their assistance with this research.

9

Religious discourse in the archived web: Rowan Williams, Archbishop of Canterbury, and the sharia law controversy of 2008

Peter Webster

Introduction

To anticipate the judgment of later historians on the very recent past has its risks. However, it may be that the decade following the turn of the millennium will come to be seen as a period marked by a shift in the nature of public discussion of the place of religion in public life in the UK. World events, indeed, made it likely that this should be so. The 9/11 terrorist attacks on the USA in 2001, the British involvement in the US-led war in Iraq in 2003 and the terrorist bombings in London in 2005 prompted an outpouring of anxiety in the media and in public life concerning the place of Islam in British politics and society.

These were not, however, irruptions in an otherwise stable field of discourse, for the period saw other significant changes in the religious landscape. The 2008 Criminal Justice and Immigration Act finally ended the statutory protection of Christianity from blasphemy (Kearns, 2008). The year before, the Labour government had signalled its willingness to relinquish its control over senior appointments in the established Church of England, with the report on *The Governance of England*. Legislation in 1999 removed hereditary peers from the House of Lords which for many left a job half-finished, since the position in the House of the bishops of the Church of England was still in question (Dorey and Kelso, 2011: 171–216). Taken together, these trends suggested that the uniquely privileged position of Christianity, and the Church of England

in particular, was under discussion to a greater extent than had been the case for decades.

The period also saw the coming to prominence of what has sometimes been termed the 'New Atheism': a polemically vigorous form of radical atheism that goes beyond a mere argument for a secular settlement in public life to a frontal assault on all forms of theistic faith. Prominent in this was *The God Delusion* by Richard Dawkins, which appeared in 2006, followed in 2007 by *God is not Great: How Religion Poisons Everything* by Christopher Hitchens (Amarasingam, 2010: 1–4).

To point out the coincidence of these three trends is not to assert any causal link between them. Such a determination must wait until the passage of time gives a longer perspective, and access to sources that now remain closed. But coincide they certainly did, and this chapter sets out to examine the potential of the archived web as a class of source material in which to observe the coincidence. Although there is a well-developed literature in the field of online religion, scholars have been slow to begin to exploit the archived web, as opposed to the live web (Campbell, 2011: 232–50). Whilst important groundwork has been done on the nature of the news media in an online environment (e.g. Burns and Brügger, 2012), there is also still room for studies of particular events and themes as they play out between mainstream media channels and the rest of the web.

Fully to examine the changing shape of religious discourse on the open UK web in this period would of course require a much larger study than the space available here would allow. This chapter confines itself to the exploration of a set of interrelated details of the larger picture, and by doing so proposes an approach to just such a larger study. All of the aspects treated relate to the unique position in British religious life of the archbishop of Canterbury, the leader of both the established Church of England and the global Anglican Communion.

Methods and sources

The primary data used for the research on which this chapter is based is the JISC UK Web Domain Dataset, held by the British Library. Acquired with funding from the UK agency the Joint Information Systems Committee, it is an extraction of all the resources in the Internet Archive from the country code top-level domain (ccTLD) for the UK (.uk) for the period 1996–2010.[1] Two notes as to its contents are necessary. First, users of web archives must always deal with the fact that content can and does appear, change and subsequently disappear on the live web

without having been visited by a crawler. In the case of this dataset in particular, comprehensive documentation of the crawl profiling – of matters such as how lists of seed URLs were compiled, how frequently and how deeply sites were crawled, and polices on deduplication of identical resources – is not available.[2] The fact that a resource does not appear in the data cannot be safely read as indicating that such a resource was not in fact on the live web at the time in question, and the lack of understanding of the crawl profile means that it is difficult to hypothesize as to which content is more or less likely to be missing. As such, research questions must be framed in such ways as to avoid needing to equate an absence in the data with an absence in fact and then draw conclusions from the latter: in other words, to avoid the so-called 'argument from silence' (UK Web Archive, 2015a).

There are also significant limits on the scope of the data as a source from which to generalize about the whole experience of the UK. The criteria by which content should be included or excluded from a national domain crawl are expressed in the various implementations of Non-Print Legal Deposit in different nations, with varying treatments of ownership, geographical location and language. Be that as it may, there would be general agreement among web archivists that the ccTLD alone cannot encompass the whole of a national web sphere. In the UK, many organizations including political parties, banks, train companies and churches have non .uk domain names. Efforts to understand the scale of national web content that lies outside ccTLDs are in their infancy. However, a recent investigation by the British Library found more than 2.5 million hosts that were physically hosted in the UK without having .uk domain names (UK Web Archive, 2015b). As a result, the study presented here is cautious in making generalizations about the national web sphere for the UK from the more limited .uk data available.

The full JISC UK Web Domain Dataset is not available for use by individual researchers as a dataset, since it is some 32 TB in size and thus is unmanageable for the majority of users. However, this study makes use of a prototype user interface to a full-text index of the data – known as SHINE – made publicly available by the British Library.[3] The UK Web Archive team have also placed in the public domain the Host Link Graph, derived from the larger dataset, which summarizes links between individual hosts in each year. The data appears in the following format:

2001 | host1.co.uk | host2.co.uk | 27

This states that the data contains 27 individual resources from *host1.co.uk* that were crawled in 2001 and which contain one or more links to a resource at *host2.co.uk*.[4]

This chapter makes no use of the total numbers of linking resources per host given in the Host Link Graph, since to interpret them properly would necessitate an understanding of the total number of resources present on a host. A single link on a very small host might be thought to have a different significance than ten linking resources on a very large host. The chapter also does not analyse the significance of individual links, which are not recorded in the data at hand in a usable way. Instead, it focuses solely on host-to-host relations as a proxy measure of attention paid by the individual or organization by whom the linking host is controlled. Richard Rogers among others has explored the meaning of the link in a web context (Rogers, 2013: 39–59). The quality of that attention may of course be positive, negative or neutral; links may equally well be intended to draw the reader's attention to content which the author deplores as much as to content (s)he endorses. In thinking about the nature of 'hyperlink diplomacy' between organizations, Rogers has characterized links as 'cordial, critical or aspirational' (Rogers, 2013: 45). This kind of close qualitative analysis of the sentiment of linkage is a matter for a larger study. Here, the concern is simply with attention.

The archbishops of Canterbury

The archbishop of Canterbury occupies a place in British public life for which it is difficult to find precise parallels elsewhere. As well as being the figurehead of the worldwide Anglican Communion, he is also leader of the Church of England, which is formally established as a state church whilst the Anglican churches in Scotland, Wales and Northern Ireland are not. Despite this, his position as the man who places the crown on the head of each new monarch of the United Kingdom, and as leader of the bishops in the UK parliament, has historically led many to regard him as in some poorly defined way the representative of all Britain's Christians. As such, successive archbishops have understood it to be part of the role to intervene in matters of public controversy, even if some were less disposed to do so than others (Hastings, 1991: 84–98; Webster, 2015: 115–31). In their turn, for decades the mainstream media have tended to treat the interventions of the archbishop in different terms to those of the leaders of the other Christian churches and of the other faiths.

Rowan Williams and the 2008 sharia law controversy

Rowan Williams was named as archbishop of Canterbury in 2002 as successor to George Carey, and enthroned in 2003. Translated to Canterbury from the position of archbishop of Wales, Williams arrived with a reputation as an intervener in national affairs from the left wing of the political spectrum, having been shaped by the socialism of his native Wales (Shortt, 2008: 82–5). The essays collected in the 2012 volume *Faith in the Public Square* are directly concerned with the interplay of faith and politics. There was, however, one particular episode for which Williams may well be remembered for longest, and which prompted a significant change in the way in which his role was reflected in the UK web.

On 7 February 2008, Williams visited the Royal Courts of Justice in London to deliver a lecture to an audience of legal professionals, although admission was also open to the public. Among the members of the public in the audience was the present author, who remembers only a complex but clear and scrupulously balanced argument concerning the interaction of the secular civil law and religious law (and Islamic sharia law in particular), particularly in the area of the law of marriage. Put simply, Williams argued that many people who looked to religious principles to settle certain kinds of disputes did not find any reflection of that fact in the law, which contributed to their perception of marginalization. If this desire was unavoidable, it would be better to accommodate it within the law, and thus to a degree to control it, than to have it operate at a local level without any kind of restraint. There were other circumstances in which the law allowed parties in a dispute to go to arbitration without troubling the courts. Such an arrangement ought to be possible in these cases (Shortt, 2008: 390–402).

Unfortunately for Williams, the majority of the public were not at the lecture itself, and heard instead an interview given in advance to BBC radio, in which the archbishop suggested that some kind of accommodation of sharia was unavoidable (Internet Archive, 2008a). Although Williams' time as archbishop up to this point had been dominated in the minds of church insiders by controversy over the ordination of gay clergy, the sharia law dispute brought him to public attention in a new way. All sections of the news media engaged with the story, some with outrage, and others with calls for the archbishop to resign, or for the Church of England to be disestablished. Williams' predecessor George Carey described the suggestion as 'disastrous' in the tabloid *News of the World* (Webster, 2008). For many, the very limited accommodation that Williams proposed was lost in lurid

visions of stoning for adulterous women and the amputation of the hands of shoplifters.

Criticism also centred on Williams' alleged naivety about the media and the likely reaction to the story (Goddard, 2013: 234–8). As an episode in media history, it was part of a recurring theme: the portrayal of senior religious leaders as well-meaning men who were either unaware of, or careless of the reaction which their interventions would provoke (De-la-Noy, 1990: 184–5; Webster, 2015: 125–7). In Williams' case, a media narrative had already been established of his supposed intellectualism and inability to express ideas in a concise and clear way. Some of the staff at Lambeth Palace privately regretted that they had not enough time beforehand to digest the speech and its likely implications (Shortt, 2008: 401). Even sympathetic commentators were caught in the contradictory position of both praising Williams' courage in raising a complex and emotive issue, whilst regretting that its expression had not been more easily digested by the media (Guardian, 2008a, 2008b).

There was also at the time some awareness that that supposed naivety in media handling extended particularly to the web. The experienced religion journalist Paul Vallely observed that 'diligent website watchers' had noticed the rapid online reaction:

> 'as this crudeness of response was transmitted, and magnified, with increasing volatility by this new communications technology. [However] it seems there were no diligent website watchers at Lambeth. Or if there were, and they pointed out […] how seriously awry things were going, [Williams] failed to hear the electronic alarm bells. He would be a fool if he made the same mistake next time. And there will be a next time, make no mistake. Welcome to the world of the new media.' (as quoted by Shortt, 2008: 401)

Reading a press storm in the web archive

Using the available data, it is possible to observe just this online reaction through the traces it has left. Extracted from the Host Link Graph dataset were all the occurrences of *archbishopofcanterbury.org* (the archbishop's own site). Results which were outward links from captures of the archbishop's site itself were removed, as were duplicates, where the Internet Archive had captured content from the same host more than once in a single year. In cases where there were multiple hosts that were part of a larger domain, these were not deduplicated. Although it would

have been straightforward to do so in the case of the larger media organizations such as the *Guardian*, which has multiple hosts (*society.guardian.co.uk*, *education.guardian.co.uk*, etc.) it was difficult to do so reliably for all such cases without examining individual archived pages, which was not possible at this scale. In any case, these accounted for less than 5% of the total. In the analysis that follows, it is assumed that a host *abc.co.uk* held the same content as *www.abc.co.uk*. It is also assumed that the Internet Archive was no more likely to miss hosts that linked to the Canterbury site than ones that did not. That is to say, if there are gaps in what the Internet Archive found (and there certainly will have been such gaps), there is no reason to suppose that they systematically skew this particular analysis.

In addressing the question of the sharia law controversy, it is convenient that it occurred very near the beginning of a calendar year (7 February). Although the absolute numbers are relatively small (347 for 2008), it is possible to see a significant rise in the total number of unique hosts found linking to the Canterbury site in 2008. The total for 2008 represents an increase of 49% on the previous year; it is 42% higher than the mean average of the three previous years; and it is 24% higher than the previous peak in 2004. An examination of the size of the dataset itself suggests that this is not to be accounted for by trends within the whole Host Link Graph, since the total number of linked pairs for 2008 is in fact considerably *lower* than any of the previous three years.[5] A distant reading of the link graph therefore suggests that more attention was being paid to the Canterbury domain in 2008 than previously.

Of the hosts found linking to the Canterbury domain in 2008, some 153 (44%) were appearing in the data for the first time. Whilst always bearing in mind the fact that those hosts may actually have linked to the Canterbury domain before 2008 but were not captured by the Internet Archive doing so, what does the patterning in this group of hosts suggest about the degree and kind of attention being paid to the archbishop that year?

Mostly absent from this subset of the data were those hosts which formed part of the infrastructure of the religious (and indeed secularist) web. Very few national organizations within the Church of England or the other churches (and the ecumenical apparatus that links them) are to be found linking first in 2008, since most had begun to link before this date. The same can be said for the main secularist campaigning organizations such as the National Secular Society, and also for the mainstream media organizations.

Instead, the data show signs of a widening in the kinds of sites involved. Among those hosts linking for the first time, there was an academic journal dealing with the nature of the mass media, as well as mainstream social affairs thinktanks such as Demos. Also in the list are sites from within the fields of public relations consultancy, family law and political organizations such as the British National Party (of which more below). The Army Rumour Service, an unofficial and widely used chat forum for the British Army, contained an (overwhelmingly negative) thread about the matter from 9 February, two days after the lecture (Internet Archive, 2008b). However, the most significant group of hosts referring for the first time in 2008 are from the blogosphere.

An inspection of the blogs shows that while some of those recorded were written by Christians, or had established secularist or anti-religious themes, the majority had no particular religious agenda but were rather outlets for the miscellaneous opinions of their contributors. This would suggest that many bloggers who had previously not been particularly engaged with religious affairs in 2008 became interested enough to link to the archbishop's domain. Some of the posts concerning Williams occur in blogs ostensibly dedicated to unrelated matters, such as that of an IT firm serving small businesses, which broke off from discussing instant messenger viruses to declare 'God bless Rowan Williams' (Internet Archive, 2008d). Some bloggers were positive; one, writing on 9 February, thought that the reaction was born of 'deep prejudice and bigotry' and that over time Williams might well be viewed as a 'precipitator of a turning point in cultural reconciliation' (Internet Archive, 2008c). However, the majority of the sample voiced similar sentiments to the more hostile voices in the mainstream media.

Of course, many of the largest and most commonly used blog platforms are not part of the UK ccTLD. That the apparent increase of attention to Williams in blogs hosted within the UK ccTLD is matched in those outside is indicated by one particular blog aggregation site, *britishblogs.co.uk*. The site was first captured by the Internet Archive in February 2006, at which time it referred to 16 posts tagged with the term 'religion'. Before 2008, none of the content from the site captured by the Internet Archive contained links to the Canterbury domain. In 2008, by contrast, there were some 1,597 resources that did so. Even allowing for the very considerable probable levels of duplication involved in the way that sites such as this are crawled (with the same content captured multiple times in views by different subject tags), this suggests a step-change in the way in which Williams was being represented in the UK blogosphere.

York and Canterbury

England is unusual among those countries with an Anglican history in having not one but two archbishops, ostensibly of equal rank. When viewed in domestic terms, the archbishop of York has precisely the same authority within the northern province as does Canterbury in the south. However, the common media habit has for decades been to attend to public statements from Canterbury rather more closely than to those emanating from York.

Whether cause or consequence of this public and media perception, archbishops of York in modern times have tended to be less politically active, although there are exceptions to the pattern, most notably William Temple in the 1930s and 1940s. However, the most recent holder of the office, John Sentamu, has also not fitted the model, after succeeding David Hope in 2005. To take one measure of political activism, Hope had rarely intervened in the House of Lords and was indeed somewhat uncomfortable in principle with the role of the bishops in the House (Marshall, 2004: 67–8). Sentamu, by contrast, intervened regularly in the House and elsewhere on controversial matters from the outset. The most dramatic of Sentamu's gestures was to cut up his clerical collar on BBC television in protest at the regime of Robert Mugabe in Zimbabwe (Sentamu is Ugandan by birth) (Internet Archive, 2007).

Sentamu, like Williams, also had a web domain dedicated to his work as archbishop. This first appears in the Internet Archive in October 2006 as a directory on the domain of the diocese of York before moving to its own domain (*archbishopofyork.org*) in early 2008 (Internet Archive, 2006a). Since Sentamu was arguably at least as politically active as Rowan Williams in the period under discussion, what does an analysis of inbound links to the York domain between 2008 and 2010 reveal about the relative attention paid to the pronouncements coming from the two men?

In order to address this, the method already documented in relation to Canterbury was repeated, with an extraction of all link pairs from the JISC Host Link Graph involving *archbishopofyork.org*, and a visual inspection and classification of all inbound referring hosts, using the live web or the Internet Archive. A total of 78 individual hosts were found, all but one of which it was possible to identify.

When compared with the same data for Canterbury, this absolute number of unique referring hosts was considerably smaller than the several hundred referring to Canterbury. Among them there are seven hosts

from the mainstream media, representing five organizations, including the BBC and several of the broadsheet newspapers. Predictably, slightly fewer than half were Christian organizations: parts of the national structure such as other dioceses, Anglican organizations within the diocese of York, and local congregations from around England.

The contrast between this data and that for Canterbury is in the relatively low number of inbound hosts from outside the churches and the media. None of the main campaigning secularist and humanist organizations are to be found linking to the York domain, and there are also few personal blogs. An examination of those individual bloggers shows that the references to Sentamu are often incidental, and do not demonstrate any sustained attention to the archbishop as a public figure. Despite Sentamu's interventions in controversial national issues, then, the link structure of the UK web confirms the older pattern: that those outside the churches were still more likely to pay attention to the archbishop of Canterbury than to his northern colleague.

Changing patterns of religious discourse: a case study – the British National Party

British political parties of the far right have for long had a relationship of polemical tension with the archbishops of Canterbury. This has been the case since at least the late 1960s when Michael Ramsey was heckled by members of the National Front on account of his work on behalf of recent immigrants from the Commonwealth (Webster, 2015: 127–9). A consistent component of neo-fascist rhetoric has been a claim that the Christianity of the native English was under threat from uncontrolled immigration, and that the churches, and the established Church of England in particular, had colluded in allowing the crisis to arise (Jackson, 2010: *passim*). Whilst the British National Party has never won a seat in a UK general election to Parliament, it has at times achieved some success in elections to local government. After a period of apparent decline, the party enjoyed a significant increase in popular support in the years following the terrorist attacks on the USA in 2001 (Thurlow, 1998: 268–72). This rise in popularity coincided with a shift in the polemical strategy of the party, away from a traditional preoccupation with British Jewry and towards the 'Menace of Islam', in order to align the party more with public opinion (Copsey and Macklin, 2011: 85–6). This section considers the evolution of the website of the British National Party and its engagement with the archbishop of Canterbury.

The BNP site first appears at its current domain (*bnp.org.uk*) in the Internet Archive in 2001. This revamped version of the site, launched in July of that year, was evidence of a marked professionalization of the party's mode of operation, with improved graphic design, use of PDF to deliver documents born in print, and the use of audio and video content (Copsey, 2003: 227–8). Using full-text search results derived from the SHINE interface provided by the British Library, this section examines resources from the BNP domain containing the search term 'archbishop'. Whilst this result set could include references to other Anglican primates in Scotland or Wales, or indeed to their Roman Catholic counterparts, a qualitative examination of the archived content itself suggests that the prime concern of the BNP was with the Anglican archbishops in England, and Rowan Williams in particular.

Between 2001 and 2007 there were examples of an older complaint, familiar in conservative religious rhetoric generally, about a modernizing church leaving its traditional adherents behind. One article, first captured in 2003, connected the trend with the controversial appointment of Jeffrey John, an openly gay man, to be bishop of Reading, only for the decision to be reversed by Williams after sharp dissent from within the Church of England (Internet Archive, 2003). There were also early examples of BNP rhetoric shifting towards Islam. Already in 2001 the party was reporting the implementation of sharia law in Africa as evidence of the danger of Islamism (Internet Archive, 2001). The site also published an attack by a party activist on the Christian–Muslim Forum, a consultative group set up early in 2006, the launch of which Williams had hosted in the company of Prime Minister Tony Blair (Internet Archive, 2006b; Shortt, 2008: 336–7). That Williams himself was already a marked man in BNP circles was evident in an article published just weeks before the sharia controversy on proposals from the Labour government to repeal the historic laws protecting Christianity from blasphemy. BNP members should not expect Williams to fight this 'constitutional vandalism', it argued, since 'Archbishop Rowan Williams is simply a Marxist in a dog collar, a man more interested in gay rights, immigrants and inter faith dialogue than in standing up for the Church of England' (Internet Archive, 2008f).

So although the volume of content referring to the archbishops was limited before 2008, the themes that were to break into public consciousness in that year were nonetheless already evident in the BNP domain. However, 2008 saw a very significant increase in the amount of content containing the string 'archbishop' appearing in the archive. Between 2001 and 2007 there were fewer than 20 unique resources

found in the BNP domain containing the string 'archbishop'; in 2008, the cumulative total (including those first published before that year) was in excess of 100. A proportion of the increased number of occurrences of the search string are accounted for by links to the initial post fed to other pages on the site. However, the bulk of the increase can be accounted for as new primary content, or new user comments on existing content. Some others contain references to other archbishops, such as to the alleged collusion of the Roman Catholic archbishop of Tours in the 'betrayal of Charles Martel', as the foundations of a mosque were laid in Tours, the scene of Martel's battle against Islamic invaders in 732 (Internet Archive, 2008g). However, the bulk of the increase can be attributed to the controversy.

The Internet Archive happened to crawl the BNP domain on 9 February, two days after Williams' lecture at the Royal Courts of Justice. The site had carried a news item about the speech, summarizing it without particular comment, save for posting a link to the archbishop's site for users 'to let him know how you feel about his comments!'.[6] Within 48 hours the post had received 167 comments, almost without exception hostile to Williams' argument. Alongside the merely vituperative, many of the commenters latched onto the episode as yet more evidence of the desertion of the Church of England of its historic duty to defend Christian England, and painted lurid pictures of beheadings and the amputation of hands on the streets of London. A number had evidently emailed the archbishop, and posted their email text as a comment (Internet Archive, 2008e).

From this point onwards, reportage on the BNP site in relation either to social cohesion at home or the consequences of alleged Islamic domination abroad was repeatedly connected to Williams and the 2008 controversy. If the authors of the content did not make the connection explicit, those adding comments very often did. Several of those commenting on the initial report thought that Williams' comments were a publicity coup for the party (Internet Archive, 2008e), and it would seem that the level of engagement with the story from users of the site persuaded the party leadership that the issue should be pursued. Seizing on a particular phrase in an open letter from Williams to Islamic scholars, made public in July, party leader Nick Griffin was filmed in front of Lambeth Palace, issuing a call to British Christians to resist their own leaders' collusion with the 'Islamification' of Britain (Internet Archive, 2008h).[7] Even though the party had been publicly denouncing both sharia law and the archbishop before 2008, the evidence of the archived web shows that the controversy over Williams' lecture led to a significant

upswing in content relating to both matters, both editorial and from users. Such a qualitative study of the evolution of specific hosts in the web archive affords the historian a new way of observing the development of religious discourse over time.

Conclusion

It is a brave historian who attempts to interpret the very recent past, as opposed to merely documenting it. As with most aspects of very recent history, the full significance of Rowan Williams' lecture about sharia law will only become clear as the passage of time grants the historian a sufficiently long perspective from which to view it. An exhaustive qualitative examination of both the published record, and memoirs and private papers that are as yet inaccessible (not least the papers of the archbishop himself, not due to be released until 2038) will be needed to place the episode in its fullest context. Without these, we cannot yet know how changes in patterns of communication that are observable in the archived web were motivated, or how opinions expressed online related to broader patterns of social and intellectual change.

However, even if it is difficult to explain changing patterns of religious discourse on the web, we may nonetheless document those changes. First, the sharia law episode prompted a step-change in the levels of attention paid to the domain of the archbishop of Canterbury, as evidenced by the incidence of inbound links, and also a broadening of the types of hosts that contained those links. Second, a comparison of the inbound links to the Canterbury domain to that of the archbishop of York suggests that the historic privilege given to the views of Canterbury over those of York was extended onto the web. Regardless of their actual status in relation to each other within the Church of England, the media and the public at large seemed only to pay attention to Canterbury. Finally, a qualitative examination of the site of the British National Party shows that at least one organization, with a very particular concern with the place of Islam in British life, certainly took new account of the person of the archbishop as a result of the 2008 controversy.

This chapter has also sought to use the episode as a means of demonstrating both the potential for historians to utilize the archived web to address older questions in a new way, and some of the particular issues of method that web archives present. At one level, the methodological complications presented here – understanding the

meaning of a link from one resource to another, say – are peculiar to the archived web and must be understood anew. As with all other born-digital sources, there is work to be done among historians in understanding these issues of method, and in acquiring the skills needed to handle data at scale. At the same time, it is part of the historian's stock-in-trade to assess the provenance of a body of sources, its completeness and the contexts in which those sources were transmitted and received. The task at hand is in fact the application of older critical methods to a new kind of source: a challenge which historians have confronted and overcome before.

This chapter has also tried to show some of the potential available to historians, should they accept the challenge. In the study of public controversy, the archived web allows the detection of changing communication patterns at scale that would be impossible using a traditional qualitative method. It also enables the detection of attention being paid online in places where a scholar would not think to look. More generally, the chapter has attempted to outline an approach that combines quantitative readings of the links in web archives with qualitative examination of particular subsets of resources. When dealing with a new superabundance of historical sources, a combination of distant and close reading will be required to understand the archived web.

Acknowledgements

My thanks are due to Ian Milligan, the editors and the anonymous peer reviewer for their comments on this chapter.

10

‘Taqwacore is Dead. Long Live Taqwacore’ or punk’s not dead?: Studying the online evolution of the Islamic punk scene

Meghan Dougherty

Introduction

In 2003 a much-photocopied but unpublished novel by Michael Muhammad Knight was passed around young Muslims across the USA. The novel, called *The Taqwacores*, told the story of a group of young people living in a shared house in Buffalo, New York. Each character embodies some different combination of religious and political subcultures including a burqa-wearing riot grrl, a straight-edge and tattooed Sunni Muslim, a Sufi punk, and the main character – a straight-laced Islamic engineering student – who questions his own identity as he is introduced to the alternative views of his housemates. The novel spoke to young Muslims who saw the stories of their own lives echoed in its pages. The fiction gave shape and coherence to a growing movement that ties punk, straight-edge hardcore, feminism, rebellion, and Islamic faith in a vibrant subculture called Taqwacore – a combination of the Arabic word *taqwa* for piety, and *core* for varying versions of hardcore punk.

Since the novel was published by Soft Skull Press in 2004, inspiring a documentary in 2009, and a full-length feature film in 2010, the subculture took form. Taqwacore, a subcultural collection of young Muslims who are politically active, rebellious, devout and who identify with a punk ethos, gained popularity. New bands, zines, a cross-country music tour and other cultural markers began to solidify into a legitimate subcultural scene. It drew much media attention in its early days, and media

attention persisted, but many blogs, online forums, band websites and other online spaces where the subculture had begun to take form slowed or were abandoned shortly after they were started. Traces of the scene can be found woven into the larger cultural landscape. Conversations around Taqwacore take place on Twitter, Facebook and other social media platforms; Taqwacore bands maintain MySpace pages dedicated to their music and blogs discussing their views on events in the Middle East and global politics; fans post videos of basement concerts and missives on their punk stance towards Islam and politics in the Middle East.

Taqwacore is a small but complex subculture of loosely knit Muslim anti-ideals and political fervour stemming from a punk ethos.

> Punk rock means deliberately bad music, deliberately bad clothing, deliberately bad language and deliberately bad behavior. [It] means shooting yourself in the foot when it comes to every expectation society will ever have for you but still standing tall about it, loving who you are and somehow forging a shared community with all the other fuck-ups. Taqwacore is the application of this virtue to Islam. I was surrounded by deliberately bad Muslims but they loved Allah with a gonzo kind of passion that escaped sleepy brainless ritualism and the dumb fantasy-camp Islam claiming that our deen had some inherent moral superiority making the world rightfully ours. I think its a good thing. […] Be Muslim on your own terms. (Knight, 2004: 212)

There are numerous areas for possible exploration with regard to Taqwacore using punk as a lynchpin in youth subculture to structure politics, mobilization and resistance, cyberpolitics, diaspora and globalization, DIY as applied to cultural hybridity, anomie and marginalization of hybrid identities, authenticity and various themes in race, class and gender. The social history of Taqwacore is complex and includes controversies stirred up within the community when alternate readings challenge commonly held norms. As yet, it has not been presented as a social history that can lead to insights about larger youth subcultural social structures for action.

This is a youth subculture with an idolized, reluctant leader, amorphously defined borders around which much debate has been sparked within the community, and among outsiders looking in. Taqwacore had its heyday in the early 2000s into the 2010s, but has since faded. The subculture came together online, leveraging technology to draw together dislocated plural worlds into a robust web sphere, or 'a unit of analysis

[that] is boundable by time and object-orientation, and is sensitive to developmental changes, within which social, political and cultural relations can be analyzed in a variety of ways' (Foot, 2006). This web sphere reflects the quickly changing digital landscape of the 2000s and 2010s.

As popular discourse repeatedly rehashed the surprise of this seemingly contradictory hybrid subculture, academics turned their attention to exploring Taqwacore in more depth. Scholarly attention was paid specifically to the analysis of Taqwacore as a subset of punk subculture and gained some momentum with a few dissertations (Davidson, 2011; Stewart, 2011), theses and other student projects (Abdou, 2009; Hosman, 2009; Andersen et al., 2011; Yulianto, 2011), and a small handful of peer-reviewed articles (Luhr, 2010; Murthy, 2010) published between 2009 and 2012 – the high point of the scene's momentum. This chapter offers a synthesis of that scholarly literature on Taqwacore in the context of literature on punk and youth subculture movements constructing identity within religious and political movements. It adds to that literature an analysis of traces the scene left behind in web archives, as much of the scene no longer maintains an active presence on the live web. This chapter covers definitions and a framing of Taqwacore that lead to dominant narratives. Much of the extant narratives of Taqwacore do not consider the scene's online presence. This chapter offers an analysis of Taqwacore's presence on the web, and commentary on methodological obstacles in studying small, short-lived cultural phenomena using web archives.

Framing Taqwacore, youth subculture and punk

Scholarly attention on Tawqacore focuses on defining the subculture to resolve the perceived conflict between its two primary driving forces of punk and Islam. Much attention is paid to determining the approach to Taqwacore as a subculture, post-subculture, imagined community or youth movement, and locating it with immigrant populations in the USA or in other diasporic Muslim communities. Conceptual frames focus on identity construction, and findings discuss a somewhat reluctant leader, self-representation and media representations, and their role in forming a definition for the subculture. Research methods used in these few studies are limited to variations on discourse analysis and ethnography. Studies focused specifically on Taqwacore frame the movement as developing in and as a result of a post-9/11, postcolonial, diasporic, international or postmodern worldview.

Defining Taqwacore and punk

What happens when seemingly incompatible subcultures are appropriated and merged? This question drives much of the discourse around music cultures and Islam. There is an overwhelming attempt to define Taqwacore as an unlikely mashup of subcultures. This perspective defines Taqwacore as an instance of the Other borrowing or appropriating Western cultural markers to integrate with his/her own (e.g. a picture of a Muslim girl wearing an Iron Maiden t-shirt with a hijab or shayla). Many of these studies begin with an autoethnographic account of the author first learning about punk Islam or heavy metal Islam and being surprised that such a self-contradictory thing existed (Levine, 2004). It is as if Islamic youth didn't rebel, feel frustration in the face of rules and social forces, have political opinions or want to explore alternate definitions of social norms. With the implication that they necessarily had to borrow models from the West in the form of punk or metal applied to Islam in order to express what could be seen as predictable rites of passage.

This analysis is where Michael Muhammed Knight, the author of *The Taqwacores*, loses interest in Taqwacore as a cultural phenomenon. In his book *Blue-Eyed Devil: A Road Odyssey Through Islamic America* (2006) he explains that, with *The Taqwacores* and his other writing, he aims to describe a kind of indigenous American Islam, and explains that curiosity inspired by a seemingly unlikely cultural combination is more patronizing than anything else. This patronizing view from outside is restrictive. The notion that punk ideology is appropriated and applied to Islam is an oversimplification that does not allow for what Taqwacores went looking for in the first place: a frame to understand disillusion, a filter for anger, a creative outlet for discovering alternative shared meaning, etc. (Knight, 2004).

Few approach the idea from a perspective of diasporic Islam merging in different localities with other subcultures. In this oversimplified perspective, studies begin with a review of studies on punk in order to set a framework for understanding it so we may then understand how punk diffused elsewhere. This is a problematic notion of punk. It is reductionist, and limiting in terms of global possibilities for punk. It is important to revisit this foundational concept in the Taqwacore literature because the stories we tell about punk matter; they are not only about punk.

Punk, as a cultural phenomenon and not simply a music genre, emerges in different localities interwoven with global dimensions.

Rubin Ortiz-Torres's (2012) autoethnographic narrative essay, *Mexipunx*, approaches the topic from this global perspective aiming to tell a more indigenous story of punk rather than tracing the diffusion of punk to Mexican culture. Using a similar critique that seeks to unsettle the notion that punk necessarily moves from the West to the rest of the world, Mimi Thi Nguyen points out in her Afterword in *Punkademics*, 'Too often, punk studies replicate a historical consciousness, through which punk unfolds from an imperial center alongside modernity and capitalism – such that anthropological accounts, or news reportage, describe punks in the so-called Third World through a sense of their belated arrival, their distance from "our" here and now' (Nguyen, 2012: 221).

Because punk is often narrated as a white and frequently male-dominated phenomenon (Duncome and Tremblay, 2011), it can be surprising to see it evolve in other multisubcultural arenas. However, Nguyen explains that punk, no matter what multisubcultural version of it you consider is 'a product of a particular historical moment in the global city, a moment that is rife with tensions not only between colony and metropole, but also town and country' (Nguyen, 2012: 222). Nguyen explains that it is important to take from this multisubcultural view on punk that 'punks themselves are already theorizing these questions – more empirical and nuanced inquiries about multiple racial, global projects that crisscross each other in webs of connectivity and exchange' (Nguyen, 2012: 222). And so she explains, for that reason, the stories we tell about punk matter in many different ways. The multisubcultural view that rejects the West to the rest perspective is a valuable avenue for studies in Taqwacore. And if it is indeed a product of a particular cultural moment, then Taqwacore's online presence must be considered in any robust analysis as a primary influence on the shape of the scene.

Framing subcultures

Much of the scholarly writing about Taqwacore is found in dissertations and theses written between 2009 and 2011 in sociology, religion and philosophy departments. These explorations begin by describing what Taqwacore is, attempting to draw boundaries by identifying key actors, events and themes for debating what Taqwacore means to those who identify with it (Abdou, 2009; Hosman, 2009; Murthy, 2010; Andersen et al., 2011). Indeed this academic treatment of Taqwacore echoes the discussion happening among those who claim the label, those who deny it and those who live on the periphery struggling with the notion

of 'belonging' on a more fundamental level. These writings are largely social histories of Taqwacore between 2004 at the publication of the novel and some time shortly before or after the release of the Taqwacore documentary in 2009.

This line of inquiry – determining the ontological nature of Taqwacore: is it a youth culture, youth subculture, postmodern tribe, philosophy, ideology or post-subculture among numerous other labels – follows directly from academic thought on punk. Choosing one of these labels certainly gives the researcher a conceptual framework to deconstruct the thing, but ultimately clouds the literature and forces the scholarly discussion to be about how to categorize the thing rather than discussing the thing itself. In his edited volume, *Punkademics*, Zach Furness (2012) collects chapters that use lived experience of punk as a more grounded examination of the topic. Furness explains 'what gets missed [in the scholarly literature on punk], for instance in the habitual focus on punk's origins, its shining stars, its hottest locations, and its most obvious but nonetheless vital contributions, such as punk's amplification (with all that term implies) of independent music and art – are the everyday practices, processes, struggles, ruptures and people that make it so interesting in the first place' (2012: 18). And Michael Muhammed Knight explained the implications of this kind of approach in an interview with MTV:

> I'm going to have to say this quick and then retreat, so here goes: Taqwacore is my friends, a growing circle of friends. That's it. Some of us happen to be artists: writers, musicians, photographers, filmmakers. You could say that we've built our own culture, but every circle of friends builds its own culture. Because the idea of brown kids having mohawks remains provocative to the media, Taqwacore has received a significant amount of coverage—Rolling Stone, Newsweek, The Guardian, NBC, BBC, The New York Times, Globe & Mail, Mother Jones, and so on. Reporting that we've seduced all the confused American Muslim kids, journalists have made us into the movement that they wanted to see, and Taqwacore is apparently real now. I just hope that it never gets real enough for Mountain Dew to throw money at us. (Andersen et al., 2011: 41–2)

The focus in academic study of Taqwacore is on the individual and ultimately group or community formation. The literature has started with conceptual framing as an attempt to determine what the subculture is

as evidenced by the cultural artefacts, discourse, interactions and social behaviours of those who name themselves as part of the group. This subculture was given a name and a few fictional examples as a foundation. Discussion among those who feel affinity for this foundational definition, and possibly more importantly those who do not feel affinity for the movement, but have been labelled as part of it, test the boundaries set by these foundational definitions.

Dominant narratives about Taqwacore

Self-representations (interviews with scholars, documentaries, fictions, music, Facebook posts, Twitter activity, etc.), media representations (articles reporting on Taqwacore) and a search for clear definitions of Taqwacore are common threads of discussion. Other dominant narratives explore the search for community and belonging, the role of computer mediated communication in enabling the diasporic character of Taqwacore (Murthy, 2010), and debates about authenticity, leadership and origin stories (Hosman, 2009; Andersen et al., 2011).

On representations, community, authenticity and leadership

One dominant narrative explains that self-representations of Taqwacore are made in reaction to American stereotypes of Arab and Muslim cultures after 9/11. Self-representations include definitions offered by many who identify as Taqwacore but may or may not be Arab or Muslim, and may or may not practise Islam to different degrees. Self-representations describe an effort to create a space for an alternative view of Arabs and Muslims in the USA – one that is defiant, individually motivated, and politically and culturally engaged. Media representations tell a tale of angry youth rebelling against authority, rebelling against Muslim stereotypes, and yet inexplicably, according to the tone of most academic study of Taqwacore, also devout. The media highlight the seemingly contradictory nature of blending a punk rock identity with a devout Muslim identity. It seems that this contradiction is felt within the scene, but it is far more complicated to those who feel that Taqwacore is a cultural space for belonging.

> While not all of the characters in the [novel] are punks and not all of them are devout Muslims, they all question Islam, what it means to be Muslim, what it means to be 'punk', and how these

> two identities and lifestyles are compatible. The novel's main character, Yusef, is not a punk rocker but, in living in the house [that he shares with other characters], he discovers an authentic interpretation of Islam in those around him. (Hosman, 2009)

Youth movements that centre on music/lifestyle genres can bring individuals together and bring relief from the estrangement felt by global youth living hybrid identities in plural worlds. Sometimes, such movements can create common cause to resist hegemonic forces. Taqwacores carve out a bit of autonomy within which they can authentically imagine, share and validate their alternatives to the status quo. This is a sincere attitude that is often recast and dismissed as anger, aggression and even violence by turning readers' attention to the most shocking examples of cultural product – the frequent use of this quote from The Taqwacores, 'In this so-called clash of civilizations, Taqwacore is about sticking the middle finger in both directions' (Knight, 2004) or citing Rage Against the Machine's lyrics, 'Fuck you I won't do what you tell me!' (Levine, 2004).

Rather than simple, boorish resistance, part of the drive for these subcultures is a reaction against the thoughtlessness of peers (Darrell, 1999). A set of rules would enable thoughtlessness on a different level. These subcultures strike a fine balance between codifying cultural markers and conformity for the sake of inclusion. Members of these subcultures want to belong, but they want belonging to be meaningful, thoughtful, and individually owned, not simply a set of rules. They look to a kind of leader for examples they may choose to build their own identity.

The evidence of online action within this youth movement tells a more complex story of leveraging technology that moves beyond the stories of self-representation that other studies explore. This exploration of inclusion and exclusion, of finding belonging in online spaces that bridge the multiple localities that play into the multisubcultural space of Taqwacore, is an under-represented theme in the Taqwacore literature (Murthy, 2010).

On finding definitions

Definitions run the gamut from identifying punk as a surrogate for religion (Stewart, 2011) to lending it political purchase by locating it at the centre of a political resistance movement (Abdou, 2009). Murthy

(2010) locates it as an emerging Muslim youth subculture in South East Asia. Others, including Michael Muhammed Knight himself, describe Taqwacore as inherently American, claiming that the dislocation and diasporic nature of Taqwacore make it what it is – unlocatable except as a tenuous common thread that holds together multiple alternate interpretations of multiple subcultures. Few studies discuss anything about social inclusion as a measure of success (Murthy, 2010). The concept of shared meaning as a marker for cultural identity is well-covered in the literature, but is quickly followed by a discussion of exclusion and authenticity (Hosman, 2009; Attolino, 2010; Luhr, 2010; Andersen et al., 2011).

Taqwacore is often defined using a frame of rejection (Attolino, 2010). Definitions of Taqwacore tend toward the negative. Short definitions tend to highlight the pairing of a rejection of Islamic hegemony with rejections of American hegemony to form some sort of curious super-rejection model that is Taqwacore. This is an oversimplification, and directs exploration toward highlighting the more negative aspects of the subculture. Attolino (2010) introduced Taqwacore as 'an emerging subculture of young American Muslims which, in the post 9/11 climate, rejects parts of both American and Islamic culture under the flag of punk music'. This reduces the subculture to its more base elements without appreciating the complexity of third culture kids (Pollock and Van Reken, 2009), hybrid or plural identities, and the rich evolution of rebellion in punk sensibilities. Descriptions of Taqwacore frame it as rebellious, angry and rejecting, whilst deriving this reading from evidence that claims the opposite. The immediate expression of Taqwacore is anger, but the deeper meaning is hopeful. Punk is described in Knight's novel as somehow similar to Islam, in that it is a 'flag, an open symbol representing not things, but ideas. You cannot hold Punk or Islam in your hands. So what could they mean besides what you want them to?' (2004: 7).

The complications of studying Taqwacore

Common approaches in studies of Taqwacore include some form of discourse analysis to examine structure, meaning, interaction, and social behaviour. Sites of mediated subcultural activity are identified in and across different online media including blogs and twitter feeds, band sites and blogs, and other offline cultural texts. More often than not, studies in the literature provide a contrast between outsiders' definitions

of the subculture represented by news and insiders' definitions represented by fictional and documentary accounts. The few that seek out individual voices in blogs, twitter feeds, interviews and other online venues have little evidence to draw from, especially in online spaces. Many of the sites dedicated to individual level interaction and social behaviour have expired and few archives of online Taqwacore content exist. Those that do are small (see The South Asian American Archive web archive of the online Tawqacore Magazine, saadigitalarchive.org/item/20120627-713), or are personal ad hoc archives stored on personal drives. A number are inaccessible to the public due to Institutional Review Board and other institutional research restrictions.

Murthy (2010) writes an exploration of participant observation in online cultural spaces during a particularly active time in Taqwacore's online cultural spaces, but relies heavily on interview data, and not on an exploration of web materials. Interview questions, when found in the literature, aggressively asked respondents to justify the subculture or resolve the conflict between punk and Muslim identities. This approach to interviewing belies researchers' assumptions about Taqwacore, punk and Muslim culture. All interview questions found in the literature explore definitions and meta-discussion about Taqwacore. None explore everyday practices, processes or people. None explore what people do in this subculture, as Furness (2012) suggests, only what they think the larger labels mean in light of assumed conflict. None describe web artefacts from the culture, content of web materials or presence on different web platforms.

Little effort has been made in digging through collections of archived websites to retrace the development of the scene in digital media. There is good reason to review Taqwacore-centred sites preserved in web archives. The scene was largely a diasporic collection imagined initially by music and literature, and solidified by belief and politics. Connections were most often made online over long distances. This kind of transmedia study, where a subculture has largely played out in digital media and virtual spaces, requires exploration into web archives to find past traces of the subculture online as it was when it was more active. This can lead to significant problems for the researcher.

The analysis to follow is based on the contents of a private collection. Objects in the private collection were captured between 2012 and 2015 by the author and a team of research assistants. Objects for inclusion were identified by strategic searches for web content using Google as a search engine. Research assistants began with basic

searches for Taqwacore materials using simple search terms (e.g. news about Taqwacore, names of bands, names of prominent figures in the scene, etc.). The results from those basic searches were used to identify additional search terms. Links on sites found also provided avenues for identifying additional materials. Websites identified as having content regarding Taqwacore were categorized according to type of content (e.g. scholarly, press, blog, first person vlog, etc.). Extensive field notes were written about objects, and about objects in relation to each other, as they were captured. Objects and field notes were captured and saved in a private group account on Diigo and on personal drives using Wget. The collection is not archival. It is inaccessible to the public, and not safeguarded for long-term preservation. The following analysis presents commentary on the rise and fall of Taqwacore as seen through the content ebbs and flows on the most prominent types of sites represented in what was collected.

The rise and fall of Taqwacore online

Taqwacore did not have a central space online until the feature film appeared on the scene in 2010, arguably popularizing the scene to the point at which it was driven into decline, or people began to feel it had 'sold out'. There are multiple layers of online public discourse regarding Taqwacore specifically, and Muslim punk in general. Mentions of Taqwacore, and discussions of Muslim punk appear on all kinds of site types around the web, including press, blogs, forums, YouTube and its comments, portals and academic sites. Each contains elements of the popular discourse around Muslim punk and Taqwacore. The most revealing discourse, the discourse that tells the richest evolution story of Taqwacore takes place in blogs, press sites and online discussion forums dedicated to the topic.

In the evolution of blogs that discuss matters related to Muslim punk, there are discussions about political leanings, world events and beliefs that are tied to the scene. There is also an attempt to find a way to describe the coherence punk can bring to the politics, events and beliefs discussed. Press sites offer a less raw representation often defining the 'movement' and passing judgement on its credibility. In Forums, there are discussions about shared experience and also personal experience, with generalities and representations seeking to make sense of personal experience in the face of larger social contexts and how they weave together.

Blogs

Much of the rich discussion of Taqwacore takes place in blogs maintained by the bands that make up the heart of the scene. Take, for example, Al-Thawra's band blog (http://althawrapunk.com/]). Al-Thawra is a Chicago-based band that got its start in 2006. They maintained a blog discussing everything from political revolution in the Middle East and the role of music in revolution, to band merchandise, bootlegs and the latest gig. However, this much-linked blog, which returns at the top of search results with numerous inlinks, has only six posts, and was only active during 2011.

The Kominas, a Boston-based band started in 2005, in contrast, have a more recent and active presence on the web. Their first appearance on the web was a release of two songs on MySpace. They readily adopted the Taqwacore label, and referred specifically to the sensibilities contained in *The Taqwacores*. The Kominas are the most active band in the live Taqwacore web, maintaining an active Facebook and Twitter presence featuring updates about politics, current events and the scene. Looking archivally, The Kominas have consistently maintained a presence in the Taqwacore web sphere over time, with their own presence as a band, in the blogosphere, in reviews and in the Press since 2010.

Beyond the blogs written by the bands themselves, most blog posts are one-off reviews of a single show, a record release or a review of the documentary film. These posts routinely dig no deeper than the surface, describing and attempting to define the Taqwacore scene, briefly mentioning the complexity of life for third culture kids (Pollock and van Reken, 2001). The authors of these posts are less careful and their reviews end with the tired witticisms of punk rock reviewers sarcastically questioning, and ultimately denigrating, a scene that has gotten too much popular attention to be legitimately punk. The posts tend to vacillate between exoticizing a seemingly conflicting combination of subcultures, or accusing it of being not 'punk' enough. Now, one question always comes up: What came first – forming the band, or reading Mike Knight's book, *The Taqwacores* (published in 2003), which gave the name to this movement?[1] The band blogs demonstrate a richer background of cultural elements that motivate the scene, and in some ways chronicle their presence in the scene. The highest levels of blogging activity can be dated to the most active moments of the Arab Spring, which garnered much attention from the Press.

Press

Despite the contention that mainstream recognition means the end of a subculture, there are articles from the commercial press trying to find a scoop on the existence of a subculture, to define and legitimize it. Their aim seems to be to make sense of, or find an angle with which to approach the events of the Arab Spring uprisings.

While many of the articles in the popular press highlight this scene as something interesting, and to be taken seriously, the coverage was generic and unoriginal. The pieces that added the most to the popular press discourse were those contributed by band members themselves, and appeared in the wake of the Arab Spring protests in Tahrir Square.

News articles regarding Taqwacore began in earnest with the screenings of the documentary at SXSW in late 2010, then gained momentum during the Arab Spring Egyptian Revolution centring on Tahrir Square in early 2011. Still, the press has kept Taqwacore present on the web more consistently than any other site producer within the scene itself. Coverage of specific bands, like The Kominas, and coverage of shows and festivals featuring Taqwacore bands have been covered consistently through 2015. As the Middle East continues to be a centre of attention for the global press, small alternative subcultures, like Taqwacore, maintain a presence as a way for the medium to place a human interest angle on their coverage.

Forums

The most robust discussion of Taqwacore, its roots, its direction and influence, and what holds it together, took place and still takes place in a variety of public forums. Wikipedia, YouTube, and Reddit are public spaces where the most robust public discussion of Taqwacore can be found.

A single Wikipedia editor contributed nearly all the updates to the Taqwacore article in 2011, adding nearly 26,000 bytes of information to an existing 30,000 byte article. The article is still being updated and maintained by a variety of Wikipedia editors during the writing of this chapter in 2015. Much of the backchannel chatter in the Wikipedia editing page revolves around the origins of Taqwacore, making the point that Muslim punk has been around as long as punk; it was simply given a name by Michael Muhammad Knight in his novel. The scene's origin story is an undercurrent of most discourse regarding Taqwacore.

One first-person vlogger, who posted a video on YouTube in 2010, captured in 2012, reacts to learning about Taqwacore. In the video, she is excited to learn that there are other people out there who merge

seemingly conflicting ideologies. The vlogger describes *The Taqwacores* novel, and her affinity for the subculture – how it matches her interpretation of Islam and ideologies of punk (and other subcultural music scenes), and how it can help one come to a new understanding of faith.

The comments in response to this vlog post are interesting. In one of the more rich and deep online discussions of Taqwacore, many YouTube commenters warn the vlogger not to reinvent Islam. Comments discuss hypocrisy and rules and indicate a right/wrong attitude towards rules rather than a continuum along a spectrum from right to wrong. This response shows Taqwacore to be something far more complex than is evidenced in other places online. Some time between personal archival capture in 2012 and the time of this writing in 2015, the video was made private, making it and its comments unavailable on the live web, and so inaccessible for archiving in public archives.

The most consistently ongoing and intellectual discussion of Taqwacore takes place on Reddit. Taqwacore appears frequently in other subreddits (subs) including /r/progressiveIslam, /r/lgbt, /r/debatereligion, /r/Islam, and others. Many of these mentions are instigated by a single user – /u/Taqwacore – who participates frequently in multiple subs of varying topics. The /r/Taqwacore sub, first captured by this study in 2012, contained a much more robust conversation regarding Taqwacore's larger cultural meaning than any other discourse online. The sub, however, was made private since being captured by this study and has not been captured by the Internet Archive. It therefore cannot be read unless the moderator approves private membership. Mentions in other publicly available subs address progressive Islam, weighing the benefits and drawbacks, debating the need for reform, and the efficacy of reform coming from a pairing with punk sensibilities.

Discussion of the scene in public forums is robust and recurrent. Threads of conversation retrace recurring lines of inquiry as new people join the conversation, but each cycle moves the discourse forward, and does so in spaces that are not necessarily dedicated to Taqwacore, but rather are interested in larger cultural spaces that may overlap with a small but active, alternative music scene. In online public forums, Taqwacore is settling in, and traces of it can be found woven into a broader cultural landscape.

Taqwacore's lost hype

Taqwacore, in its short-lived existence so far, has shifted and changed in many ways, both in response to the evolutionary trajectory of those in the

scene, but also in response to social and cultural shifts that have taken place globally since its early days in the early 2000s. So while Taqwacore as a scene has lost some of its initial popular hype, it has not fallen completely by the wayside. Rather, it has taken up residence as a niche element in a culturally diverse punk scene. Most recently it made an appearance in the form of one band's show – The Kominas – during Chicago's 2015 Black, Brown, and Punk Show Collective Annual Music Festival.

Methodological limitations on studying Taqwacore

Finding evidence of the Taqwacore scene online, especially evidence of its early development, is difficult. Discourse found on blogs reveals something of the early days of the scene, but do little to reveal much of how it has evolved. Discourse in the press is ongoing, but tends to retrace the same surface-level, outsider perspective of the scene. Public forums show a slightly more intimate engagement with the scene. But all of these sites are difficult to access. Over time, they have disappeared from the live web, been made private due to the potentially inflammatory nature of the discussion, and evaded capture in archives due to the relative insignificance or technical depth of the relevant content. Generally, online activity regarding Taqwacore in these public web spaces has waned as discussions moved to increasingly popular and closed spaces like Facebook and Twitter.

One of the more interesting findings in this study was methodological. Midway through the collection period, research assistants began to notice a stalling of new material to capture. Repeating basic searches resulted in material already captured in the collection. Few new avenues for new material presented themselves. By the end of the collection period, all results, regardless of creative search efforts, were redundant.

The materials collected in this study have much to tell us about this particular subculture. But it cannot tell us why it waned, or why what remains of it moved into more private spaces online. Traces left behind in captured cultural materials can hint at internal strife, or aggression from outsiders, but ultimately can only provide minimal insight. They cannot tell us if a blog was abandoned out of boredom on the part of the site producer, because of external pressures, or simply because an account expired, wiping the site from a platform.

Throughout the collection period, sites found via searches of the live web were cross-checked with the Internet Archive. None of the sites collected in this study were found when searched by URL in

the Internet Archive's Wayback Machine. Because the Internet Archive has the most broadly sweeping collection strategy of all the web archives, this absence indicates that there is little evidence of Taqwacore's cultural artefacts, possibly none at all, in publicly accessible, safeguarded archives meant to preserve cultural heritage long term.

There is much that is lost about a cultural moment such as this when its web presence is not captured and preserved in a web archive. If it is indeed the case, as this author found, that traces of Taqwacore have not been preserved in large publicly accessible archives, the findings of this study should be regarded as questionable, as they cannot be verified or replicated with an independent study using alternative means.

Conclusions

There is only so much we can come to learn using web archives to find evidence of small subcultures such as Taqwacore. As a small subculture shifts its values and moves its activities to more private spaces online, less and less of that subculture remains visible to those who may study it, whether through searches of the live web, or through objects included in archives.

In this study, researchers found that over the course of the data collection period, the web itself became the archival record. Absence of relevant material in official archives led the researchers to recursively search the live web for material. The longer they searched, the more they found that relevant materials had been created in the past and were simply lingering, by chance, on the live web. Some materials that had been found early in the collection period were found to have vanished when searched for later with no trace other than the site previously captured for this study. This left the researchers with little material upon which to base a study, and concern that the material that *did* remain of this once-active subculture was ageing and might easily be lost for good, and with little evidence to explain why.

Taqwacore is a particular hybrid youth subculture that has not been well-explored, for good reason. The failure to explore the subculture's presence online is partly due to it not being represented in web archives and partly to the diasporic nature of the community. Taqwacore's presence online is just as diverse, mobile, contingent and unlocatable as it is offline. The author is left to worry that perhaps it is also unarchivable, and so too easily forgotten.

11
Cultures of the UK web

Josh Cowls

Introduction

This chapter reports findings and insights from 'Big UK Domain Data for the Arts and Humanities' (BUDDAH), a project led by the British Library, the Institute for Historical Research at the University of London, and the Oxford Internet Institute at the University of Oxford. This project ran from January 2014 to March 2015. The primary aim of the project was to facilitate the use of a 65 terabyte dataset containing crawls of the .uk domain from 1996 to 2013. The crawls were conducted by the Internet Archive, which captures and archives web pages on a massive scale (Kahle, 1997). This dataset (hereafter 'the web archive') was acquired for the use of the British Library by Jisc, a charitable organization which facilitates the use of digital technologies in UK education and research. As becomes clear in the researchers' reflections in this chapter, the dataset shared many of the limitations and challenges of other web archives. Although enormous, the archive does not hold every site or page in the .uk domain, raising questions around the representativeness of the archive, and those resources that were captured are time-stamped in relation to the date they were captured, rather than originally created.

As part of the project, ten arts and humanities researchers were invited to use this web archive dataset to conduct cutting-edge research. The researchers, who received a bursary for their participation in the project, were all studying at a doctoral level or higher at the time of the project, and were thus experts in their respective areas of scholarship.

This chapter describes how the researchers utilized the dataset, and explores the findings which emerged from their research. The initiative was distinctive for several reasons. It represented a rare opportunity

for researchers with little or no existing expertise in the use of web archives to utilize this exciting but challenging source of data. Moreover, the process was structured so that development of the archive could proceed iteratively, in response to the researchers' feedback; regular meetings between the researchers and developers facilitated this process. This collaborative, iterative process was especially important for the development of 'Shine', the interface used to conduct full-text searches of the archive. Due to the scale and diversity of the data contained in the archive, the search interface became an essential tool for navigation of the archive. Yet, as will be seen, the search interface also served as a platform for *conceptual* navigation of web archive research itself.

The chapter begins with a description of the ten projects, briefly outlining the research foci, the approaches taken and the findings which emerged. These experiences are then synthesized in the discussion section, with a series of wider reflections on the challenges of conducting research using web archives, and the implications for arts and humanities scholarship that result from the use of this valuable but as yet underexplored resource.

Project summaries

Online reactions to institutional crises: BBC Online and the aftermath of Jimmy Savile

For her case study, Rowan Aust of Royal Holloway, University of London focused on the aftermath of a major scandal at the heart of the British Broadcasting Corporation (BBC). Following his death in 2011, a string of sexual abuse allegations emerged against the iconic broadcaster Jimmy Savile, who was famous for decades fronting BBC radio and television programmes and for his prolific charity work. Aust described the revelations about Savile as 'ruptur[ing] the stability' which undergirds cultural memory, and her research focused on understanding how the BBC as a prominent institution reacted to the allegations, analysing how content relating to Savile changed over time on the BBC's website (Aust, 2015: 3).

Aust began the research by conducting an iterative series of searches relating to Savile and the BBC within the archive as a whole and the BBC website in particular. Through comparison with the live site, this yielded a series of instances in which changes had been made as a reaction to the scandal. Aust found only one instance – an interview

with Savile on the long-running BBC radio series 'Desert Island Discs' – in which explicit reference had been made to the removal of online material relating to Savile. Elsewhere on the website, attempts to erase or modify content had been applied inconsistently and haphazardly. In some cases, links had been removed or broken; in others, content had been modified or greyed out.

Aust followed up this comparative analysis of primary sources by writing to the BBC's Controller of Editorial Policy. Through back-and-forth correspondence, she learned that the BBC had procedural guidelines for the removal of online content, but that these had only been implemented in 2014, well after the first allegations. Aust's research suggests that, contrary to the 'presumption that material published online will become part of a permanent accessible archive' described in these guidelines, at least some sensitive content has been quietly removed (2015: 8). Yet while attempting to control the narrative around the Savile case, the BBC appears to have been hamstrung by the size and scale of its online presence, which has perhaps proved too diverse and diffuse for a blanket policy of removal or modification to be effectively implemented. Aust's research is thus significant on its own terms – shining light on a serious and significant case – but it also holds lessons for the maintenance and modification of large institutional archives more generally, and the implications of this for cultural memory.

The web archive and Beat literature

For her project, Rona Cran of University College London sought to discover academic and public receptions to Beat literature in the UK as reflected in the web archive, and to establish whether web archives were therefore a useful research tool for literary studies more broadly. Overall, Cran found that the archive 'has great validity and enormous potential as a research tool for literary researchers', describing a 'liberating sense, when working within the archive, of exploring both the past and the future simultaneously – of entering uncharted territory whilst also rediscovering forgotten artefacts' (Cran, 2015: 3).

Yet while it clearly holds much promise for study in this area, Cran encountered a number of challenges with using the archive for conducting research. One was the 'geographic' limitation of the archive dataset: much relevant material was likely to have been published on domains other than .uk. Moreover, the data that was available in the

UK archive was 'far more fragmented and disparate' than in more consolidated and comprehensive literary collections elsewhere on the web (2015: 2). What is needed, then, is a process of 'foraging and sensemaking': the 'territory' represented by the web archive 'needs to be mapped' by scholars and the interested public (2015: 3).

Cran found, however, that the archive in its current messy, unmapped and arbitrary state in fact meshes perfectly with a sensitive understanding of Beat literature. Cran draws parallels between the writing styles of Beat writers – such as William Burroughs, whose novels 'read […] like the uncontrolled spewings of an ailing machine' and the haphazard nature of the dataset (2015: 5). Further, Beat writers 'treasured notions of fragmentation, ellipsis and inherent unknowability', which are 'positive aspects of the web archive in its current form' (2015: 6). From this perspective, Cran's research shows that not only can we learn much about the Beats from the web archive, so too can we learn plenty about the web archive from the Beats.

Revealing British Euroscepticism in the web archive

In his project, Richard Deswarte of the University of East Anglia sought to discover whether and how British Euroscepticism has been recorded in the archive. Deswarte started by creating a list of keywords relevant to Britain's place in the EU, including 'referendum', 'Eurosceptic' and 'UKIP' (the acronym for the Eurosceptic United Kingdom Independence Party), which were then searched for within the archive (Deswarte, 2015: 3). Unfortunately, the volume of results returned for these queries was enormous, which precluded closer analysis of all or even a meaningful sample of the available resources. Even when more filtered searching was conducted – for example by limiting the date range and removing the most prolific subdomains such as news sites – large numbers of results were returned.

Nonetheless, Deswarte was able to find a number of individual items of relevance to his research focus, including the full text of an old speech by UKIP's leader Nigel Farage and a series of documentary films. However, despite the academic value of these discoveries, as Deswarte notes, 'their discovery was serendipitous rather than based on a sound methodological approach to analysing the increasing mountains of materials' (2015: 4). Various challenges, including the abundance of data available, and issues over the quality and consistency of web page capture, precluded large-scale quantitative analysis to draw out more

general patterns regarding Euroscepticism, or to relate them to offline trends. In one instance, searching for 'UKIP' returned hundreds of results contained on the sports pages of a regional newspaper. It took considerable effort on the part of the researcher to establish that this was the result of a rolling news banner. This example demonstrates how seemingly minor elements on a web page can create major issues which are extremely time-consuming to weed out. Deswarte's project serves to show that, while web archives may host valuable material, locating this material, and relating it to broader societal trends across time, currently represents a major challenge for historians.

Searching for home in the historic web: An ethnosemiotic study of London-French habitus as displayed in blogs

Saskia Huc-Hepher, of the University of Westminster, used a number of web archives to conduct an ethnosemiotic case study – an approach combining ethnographic research and semiotic analysis – of the French community in London. Huc-Hepher sought to 'think small when handling big data [to] inject new, deeper meanings', by focusing on a small set of primary sources written in French, using 'the storytelling of individual lives serving as a guiding light' to illuminate the enormous web archive (Huc-Hepher, 2015).

Huc-Hepher searched the archive to develop a corpus of relevant sites: blogs written by French émigrés living in London. There were both advantages and limitations to this approach. The use of French-language search terms proved an effective filter for sites in the .uk archive. However, visual components of the blogs in the corpus – including fonts, photos and videos – were often deficient or broken, threatening to 'ultimately jeapardis[e] the very validity of the multimodal semiotic approach'.

Despite these shortcomings, Huc-Hepher was able to locate a number of blogs authored by French people in London across a range of archives. Due to the restrictions of the web archive – for example, the fact that it only contains .uk sites – only one blog was found here; other examples were identified in different web archives. Through the multimodal analysis, Huc-Hepher was able to assess a whole raft of components on each blog including colour palettes, the content and layout of banner images, typography and text. Crucially, through analysis of these blogs over time – using captures of the same blog at different times in different archives – Huc-Hepher was able to detect subtle but

meaningful changes in the emotional position of the blogger in relation to London. In many cases she observed a gradual integration of bloggers into their new environments, finding a 'half-way habitus' or 'hybridisation of habituation'. It is notable that Huc-Hepher was able to conduct a rich, illuminating analysis with only a small number of resources. This points to the contribution to research that even a single page or object can play, and thus reminds us of the importance of archiving as much as possible for the benefit of future research.

Capture, commemoration and the citizen-historian: Digital shoebox archives relating to POWs in WW2

In her research project, Dr Alison Kay of Northumbria University focused on using the web archive to locate and explore 'digital shoebox archives' – micro-collections and narratives of lived experience – relating to Prisoners of War (POWs) in the Second World War (Kay, 2015). At the core of Kay's approach was the iterative development of search strings which would return results relating primarily to Second World War POWs, especially in regard to personal narratives and commemoration. This involved various filtering techniques, including the exclusion of various domains, such as media organizations (like bbc.co.uk) and commercial sites (like amazon.co.uk) and proximity searches of key phrases, to limit irrelevant results. This strategy proved effective: without these filters, the number of results returned for one of Kay's basic search strings was an impenetrable 53,638; with them, it was a much more manageable 206. Overall, for 11 distinct search terms, this figure was 2,894. On one hand, this represented a sizeable decrease from the 24,727 pre-filtered results; yet on the other, it still remained too large for a researcher to single-handedly tackle in the course of the project.

Despite the volume of results and the limited time available to assess them, Kay's project offered an illuminating overview of the sort of valuable historical material captured by the web archive. By identifying a number of online projects gathering memories of war, and filtering to only include results from these domains, Kay was able to investigate what proportion of memories from the live web had made it into the web archive. The findings here varied by project. For the Wartime Memories Project, nearly 10,000 results were returned. This might represent up to two thirds of the 15,000 or so wartime stories and testimonies, though as Kay notes, duplicate captures may have increased this total artificially. For the BBC's People's War project, however, only 346 results were

returned – a tiny proportion of the 47,000 stories the project claims to have on the live web. Kay's research therefore suggests that, however imperfect and incomplete (or in the case of duplicates, 'over-complete'), the collection of 'shoebox' memories contained in the archive might well prove a valuable source of materials for historians and the public at large.

Digital barriers and the accessible web: Disabled people, information and the internet

In his project Gareth Millward, of the London School of Hygiene and Tropical Medicine, sought to investigate how information was presented on the internet in a format accessible to disabled users. In particular, it focused on the accessibility of information made available about, and by, disabled organizations themselves. This initial investigation began with a series of searches of the entire dataset, seeking to discover how well represented disability organizations were over time – this analysis found that overall, the Royal National Institute of the Blind stood above its peers in terms of references in the dataset (Millward, 2015). More generally, public-facing charities seem to enjoy better coverage compared with more focused lobbying organizations. Millward also investigated the extent to which disability organizations' websites adhered to the World Wide Web Consortium's Web Accessibility Initiative accessibility guidelines using code validation tools.

Yet in addition to these findings, Millward's report pointed to a series of challenges in conducting analysis of this sort. First, the names of disability organizations had a significant bearing on whether their reach could be accurately analysed: organizations with common names such as Scope and Mind returned a large proportion of irrelevant results, even with additional terms such as 'disability' included in the search string. Secondly, the code validation did not allow a like-for-like comparison between websites: as web pages became longer, the number of accessibility errors in the code would also increase. A final, more general challenge that Millward pointed to was the enormous size of the dataset and the amount of potentially relevant material. This necessitated a series of blanket decisions taken to try to reduce the size of the corpus to an extent that more sensitive analysis would be impossible. As such, while interesting discoveries could and did emerge from this approach, ultimately 'there was very little academic validity to the corpus, and it was difficult to defend the results as representative or in any way objective' (2015: 8).

Nonetheless, Millward sketched a number of future areas to extend this analysis. Instead of conducting large-scale searches of the whole archive, link analysis could be a more accurate way of assessing the relative reach and influence of different organizations. Qualitative analysis of individual websites as well as oral histories with the figures responsible for organizations' online strategy could augment quantitative findings. Moreover, the importance of this line of research is not in question: in the case of the RNIB, for example, we can see 'a continuity from braille through to web access as a core part of the charity's remit' (2010: 8). Engaging with web archives as a primary source for exploring this phenomena is therefore inescapable.

A history of UK companies on the web

Marta Musso, of the University of Cambridge, sought to track the diffusion of internet use among UK companies, with a focus on the period between 1996 and 1999, when websites of companies were generally in their infancy. Musso utilized a range of sources of data and approaches, comparing captures of company websites in the archive with contemporary versions on the live web, as well as conducting a questionnaire and examining records of website registrations and contemporary newspaper articles (Musso, 2015).

Musso began by attempting to sketch the gradual uptake of website registration by major companies in the early stages of the web. Combining information from various archival and other sources, she found that, despite a UK-centric sample of companies, many used the generic .com domain as opposed to UK alternatives. Moreover, over the course of the period she observed 'an overall tendency [among companies] to switch to a .com domain and to simplify the address altogether'. She also found that older and larger companies tended to register their company websites earlier than others, usually before 1996. Through responses to the questionnaire, she learned that BP registered their domain address bp.com as early as 1989, although this served merely as a placeholder for many years.

Following up these findings, Musso conducted a series of searches of the archive for words relating to business and commerce, which yielded a series of company directories from the period. These directories, however, only existed briefly; from around 2000 onwards, the sophistication of search engines had rendered them far less useful. Examining some of these sites in more detail, Musso observed that many of them were 'mimicking physical commercial spaces', borrowing elements from an

offline commercial setting – for example by using a 'front door' and user-friendly menus – to make the online browsing experience more familiar to users. Taking these findings together, Musso suggests that the early experiences of companies on the web reflects more general patterns in internet use, resulting in a 'shift from a private means of communication […] to a public space, a virtual reality in which everyone [can] have access to every space'.

The online development of the Ministry of Defence

In his project, Harry Raffal of the University of Hull explored the web archive to trace the development of the websites for Britain's Armed Forces – the Army, the Navy and the Air Force as well as the umbrella Ministry of Defence (MoD). Raffal began by creating a corpus of five iterations of each of the websites across the period 1996–2013 (Raffal, 2015). This corpus was thematically analysed by coding various elements on the page, chiefly text, videos, images and navigational elements. Raffal also conducted link analysis of the MoD site, in relation to a number of subsidiary recruitment and educational organization websites.

These analyses yielded various findings about the purposes behind the Armed Forces' web presence. Through thematic analysis Raffal found that, although the content and design of the MoD website has changed over time – in line with developments in web standards and trends more widely – the initial intentions for the Ministry's online presence – promoting a 'corporate image' and serving 'business and presentational needs' – remain largely in place. In the case of the Armed Forces, recruitment has remained a chief concern: especially in the case of the Army, which integrated television commercials with interactive website content, and shifted its recruitment terminology from the word 'career' to the more informal 'join'. Raffal contrasts this with the continued use of 'career' among the Navy and Air Force, for whom longer-term, more technical appointments remain the norm.

Raffal's link analysis also yielded interesting results. It highlighted an unexpectedly prominent role in the network of armed forces websites for the MoD's Supporting Britain's Reservists and Employers Agency (SaBRE). A high concentration of inbound links from local authorities and reserve associations suggested that SaBRE was 'achieving at least part of its remit as an organisation that aims to build support for members of the Armed Forces'. Overall, Raffal's research benefited from the

creation of a systematic, self-contained corpus and the utilization of mixed methods to uncover meaningful patterns among the UK Armed Forces' online presence.

Looking for public archaeology in the web archives

In her project Lorna Richardson, now of Umeå University, sought to explore representations of public archaeology in the web archive (Richardson, 2015). In common with other projects, the mass of resources available led Richardson to narrow and fine-tune her search criteria, from the well-known archaeological site Stonehenge to three slightly more obscure sites, which respectively returned substantially fewer results. Richardson employed the archive's n-gram function to display the results over time, and found that, for the 'Ness of Brodgar' site, an archaeological find in Orkney made in the late 1990s, 'the N-gram visualises beautifully the release of information as the archaeological site progressed in its discoveries'. Richardson is able to show here that information released from the archaeological site gradually makes its way onto archaeological websites.

Richardson notes that overall, there are hundreds of thousands of pages in the archive containing material potentially relevant to public archaeology. Richardson's approach to working through this mass of data evolved into a 'manageable scoping exercise for a handful of key archaeological sites and terms'. Indeed, this approach can be extended to web archive research more generally: as Richardson suggests, 'using an archaeological approach to explore, reconstruct and reimagine the technologies of past iterations of the World Wide Web' could improve academic understanding of how the archive can be used effectively.

Do online networks exist for the poetry community?

Helen Taylor of Royal Holloway, University of London conducted research into the presence of poetry networks in the web archive (Taylor, 2015). Poetry networks long predate the internet, allowing content to be spread informally between different locales; Taylor sought to discover whether online poetry networks 'exist in and of themselves, or whether the online presence of a group is merely a kind of "placeholder" or "directory"' (2015: 2). Taylor restricted her analysis to two poetry websites which represent both ends of this spectrum: The Poetry

Forum, a place for people to share and comment on their own poetry, and the Oxford University Poetry Society's website.

In her analysis of The Poetry Forum, Taylor observed a number of features which point towards it being a genuine community. The sign-up process, while optional, encourages contributors to have a profile – although these need not faithfully reflect their offline persona – promoting the sharing of original work and comment. Taylor surmises that the site 'is an example of how poetry networks do exist online […] this kind of interaction and exchange would not have been possible before the internet' (2015: 3). This contrasts sharply, Taylor finds, with the case of the Oxford University Poetry Society, wherein the site's 'online presence is only there in order to get you to engage offline' (2015: 4). Having taken a temporally representative series of captures, Taylor found that members' poems only seldom appear on the site, and even when they do there is no facility for comment or discussion. Taylor further speculates that there was a deliberate decision not to bring the discussion online, given the haphazard use of different URLs over time, and the frequent appearance of 'we have moved' messages.

By highlighting these two examples located in the archive, Taylor's research demonstrates the diversity of approaches taken to the creation and maintenance of poetry networks online. She concludes that, although the web has been transformative in facilitating connections between geographically diffuse participants, in circumstances where these virtual connections are not required, another website can play an entirely different role.

Discussion

Giving the ten researchers direct access to the data through the search interface, and closely involving them in the development process, yielded a range of insights regarding the utility of web archives for humanities research. Each of the researchers dedicated significant portions of their project reports to reflect on the potentials and pitfalls of conducting research using web archive datasets. In this discussion section, these reflections are synthesized, yielding three topics which are drawn out in greater depth: how the researchers conceptualized web archives; the strategies for research that were taken; and the importance of search tools for navigating the massive archive.

Conceptualizing web archives

Although prior to this project the ten researchers had had little or no experience with web archives, they came across many characteristics of this field that have already been highlighted by scholars elsewhere (Brügger, 2012; Schneider and Foot, 2004). For example, many researchers noted both similarities and differences between the web archive dataset and older, traditional archives. Richard Deswarte suggested that 'in many ways, the term "archive" is a misnomer', since what the researcher really faces is not a web page in its original form, but a reconstruction of a pre-existing web page – and often an incomplete one' (Deswarte, 2015: 6). In a similar vein, Alison Kay noted that 'as a historian I have to remind myself that the online web is gone. We [only] have representations'. Web archives might therefore have less in common with historical archives – which are typically text-based in nature – than with archaeological artefacts, a metaphor proposed by Lorna Richardson. A technical disjuncture between how a web page appeared on the live web and how it is rendered in a web archive thus exists alongside a temporal disjuncture between when an archaeological artefact was originally used and when it was found. Rona Cran also observed conceptual similarities between the web archive and the 'uncontrolled spewings of an ailing machine' characteristic to Beat literature (Cran, 2015: 5).

Yet this is not to overstate the conceptual departure from traditional archives represented by web archives. As Alison Kay noted, 'mass printing was worrisome in its volume in years past, in the way that the archived web is challenging today'. Indeed, just as with web archives, most historical archives are subject to some degree of arbitrariness regarding what is and is not preserved. In the case of web archiving, both technical and curatorial factors can affect what is kept and what is discarded. The web archive used here essentially represents the .uk portion of the far larger Internet Archive, which hoovers up vast tracts of the live web for archiving on the basis of links between sites. Yet even at this huge scale, there is a role for curation, as with the Archive's policy to respect the robots.txt protocol when crawling. Whether a page does or does not appear in the Internet Archive is therefore the result of both technical contingencies and curatorial considerations. In a certain respect, the fact that technical as well as more subjective factors affect what appears in a web archive 'could be regarded [as] a refreshing objectivising tool […] a means of making the final collection less a reflection of [the archivist] and more about the material itself', as Saskia Huc-Hepher suggested.

The researchers thus found that web archives, at least in their current state, represent a curious position in relation to previous sources of data. They are both similar to and distinctive from traditional historical archives, whilst also holding conceptual affinities to archaeological and literary traditions. The following section explores how these different perspectives on what the web archive represents fuels alternative approaches to utilizing it for research.

Strategies taken

The breadth of research interests pursued by the ten researchers on the project was reflected in the large degree of diversity in terms of the methodological approaches taken, despite the fact that all the researchers had access to the same data and the same tools. These diverse approaches can nonetheless be roughly clustered into two contrasting strategies, which can be labelled the 'part of the whole' approach and the 'whole of a part' approach. Should they tackle the entire archive dataset in all its enormity and complexity, using the search engine to isolate specific items across the archived web relating to their research (the 'part of the whole' approach)? Or should they restrict their research focus to a pre-defined, substantively meaningful subset of resources (the 'whole of a part' approach)? These strategies are not mutually exclusive, even in a single research project: many of the researchers ultimately used both in their own projects. But the distinction drawn here helps to illuminate the strengths and limitations of each approach for conducting valuable research.

An example of the 'part of the whole' approach is Richard Deswarte's investigation into Euroscepticism. One clear advantage of Deswarte's strategy of searching the archive as a whole is that results which emerge are representative of the archive itself, allowing longitudinal analysis of the prevalence of a given phenomenon (in this case Euroscepticism). Of course this is not to suggest that the archive itself is necessarily representative of society: numerous aforementioned issues, such as the inconsistent capture of pages, cast doubt on the true representativeness or reflectiveness of society in the archive. Furthermore, other practical issues limit the utility of this strategy. As Deswarte found, the number of results returned when searching the archive as a whole can be huge, particularly when researchers are seeking evidence of general phenomena (Deswarte, 2015). This strategy also gives rise to the 'so what?' problem described by Gareth Millward, who argued that 'only through disaggregating these results can we gain any real meaning that might be of use' to researchers (Millward, 2015: 5). With the net cast so wide,

trawling through the vast catch can simply be too time-consuming, a challenge also noted by Alison Kay and Lorna Richardson.

Seeking to collect and analyse a part of the whole archive is therefore replete with challenges. Other researchers, in contrast, adopted what might be called the 'whole of a part' strategy. This approach meant focusing squarely on a pre-defined set of resources – usually one or a small number of websites – and analysing them in their entirety. Research projects primarily using this approach include Harry Raffal's investigation of the Ministry of Defence websites, Helen Taylor's analysis of two distinct poetry networks, and Rowan Aust's investigation of the Jimmy Savile scandal as reflected on the BBC website. Again, there are obvious advantages to this approach. Researchers can use the filter-by-domain feature of the search interface at the outset, resulting in a far smaller set of results, which is likely to allow more sensitive, 'line-by-line' analysis of all the results returned. Yet though this approach may yield a tighter and more internally coherent group of resources on which qualitative analysis can be performed, care must be taken to highlight that the items selected for analysis are not representative of the archive as a whole.

Both the 'part of the whole' and 'whole of a part' strategies therefore have strengths and limitations for producing valuable research using the archive; these are summarized in Table 11.1. Overall, it seems sensible to make the decision of which strategy to adopt based on the nature of the research question. Where the research question centres on a broad social or historical phenomenon it may make most sense to pursue the 'part of the whole' strategy, all the while bearing in mind the significant challenge of scale that this approach often entails. In contrast, where the research is focused on a specific entity or event, particularly where this is associated a priori to a particular subdomain or website, the 'whole of a part' approach may work best.

This section has explored two contrasting ways of conducting valid research using large-scale web archive datasets. Of course, the two strategies presented here are not mutually exclusive, since many researchers utilized both strategies at different points. Nor are they strictly dichotomous: a 'web sphere' (Schneider and Foot, 2004) – the third largest of the five 'strata' of web analysis listed by Brügger (2012) – could be the unit of analysis in either approach, depending on the size and nature of the 'sphere' in question. Moreover, both strategies involve the use of searching the archive at some stage in the research process. In the following section, therefore, the purpose and process of searching is explored in more detail.

Table 11.1 Comparing strategies for web archive research

Approach taken	Summary of process	Key advantage	Key limitations
'Part of the whole'	Searching the entire archive for a broad historical/ social phenomenon (filtering occurs mostly a posteriori)	Ability to treat the archive as a whole and make definitive statements about the archive	– Archive not necessarily representative of society – Structured data important for quantitative analysis, yet web archive data is not (usefully) structured
'Whole of a part'	Searching a particular subdomain or website for a specific entity or event (filtering occurs mostly a priori)	Able to adopt a sensitive, grounded understanding of web pages and elements	– Difficult to make definitive statements about how resources analysed relate to or represent the archive as a whole – Meaningful findings are discovered serendipitously not systematically

The use of search tools

Whichever strategy the researchers employed, the enormous size of the database meant that researchers required a search interface through which to access, assemble and analyse the resources relevant to their research question. Shine, the search interface which the researchers used, was developed over the course of the project, largely informed by the researchers' experiences. The development and use of the interface inevitably opened up another raft of conceptual questions and challenges. Research does not take place in a vacuum – something that humanities scholars appreciate more than most – and many of the researchers noted that their previous experience with using search engines to navigate the live web affected their assumptions about searching web archives. Richard Deswarte coined the phrase 'Google mindset' to describe the set of expectations that researchers had about how the search interface would or should work (Deswarte, 2015: 9).

The core difference – straightforward in principle but disorientating in practice – relates to the ordering of search results. The algorithms developed by Google and other search engines are as lucrative as they are elusive, ranking billions of results by perceived relevance in a split second. In the case of the Shine search interface, the size of the index is enormous – with hundreds of thousands of results common for basic queries – but the ordering of results is far less sophisticated. Instead of 'relevance' as an option, researchers can order results by, for example, an item's title or the date it was crawled. Thus as Richard Deswarte pointed out, 'all of the results' – not just the first few – 'will potentially be of interest' (Deswarte, 2015: 7).

Viewed from a different perspective, however, some researchers found this limitation liberating and even empowering. Rona Cran took a 'deliberately unsystematic approach to the archive, by treating it as something akin to a vast bundle of unsorted papers rather than, say, Google'. In doing so, she 'was able to confront it with my own perspectives'. For Cran, this process 'heightened [the] intellectual integrity' of her study, since she was 'using processes of reasoning and selection which were unique' to her as a humanities researcher (Cran, 2015: 5). Thus through the limitation of the search interface her research expertise had fresh importance.

This sense of greater control over the research process was bolstered by later developments of the interface. As noted, this development process took place iteratively, directly in response to the feedback of the researchers. Most significantly, the ability to create a personal corpus of results – extracting individual results from search results into a persistent collection – was the new feature most requested by researchers across the project, and was well received when introduced. When combined with advanced search tools, which allowed researchers to search only particular domains or across narrower time periods, for example, it gave researchers a more powerful set of tools with which to tackle the data available.

However, even with the addition of these tools, saying anything definitive about the contents of the archive in general remained extremely difficult, particularly for researchers whose research questions were broadly conceived. One approach would be to create a small sample of the data, which could then be sensitively analysed by the researcher. Yet Richard Deswarte noted that while 'structured data mean[s] it is possible to make clear and academically justified decisions on the size and relevance of representative samples […] unfortunately and problematically the data in web archives is almost completely

unstructured, at least in terms of content' (Deswarte, 2015: 3). Another problem was explained by Gareth Millward, who suggested that it was in fact the relative scarcity of data that made it possible to answer a historical question in the traditional way. Typically, traditional historians 'identify a question and source base, go back to the archive, and then mine what [they] can until that vein is exhausted. This is [only] possible because we have a relatively small amount of evidence which has survived' (Millward, 2015: 6).

In summary, the researchers' experiences with the search function seem to have been in equal parts frustrating and empowering: frustrating, because of the lack of any substantive ordering through which researchers can get to grips with the voluminous resources available; yet empowering because, in the absence of any such pre-ordained notion of relevance, researchers must make more decisions based on their own domain expertise. These perspectives are not, of course, mutually exclusive: the most frustrating aspects of the experience could be mollified by, for example, more powerful and tailored search functions, whilst still allowing researchers the ability to make informed decisions about what to include in a corpus. Yet in addition to technical improvements, researchers and developers need to continue to engage critically with the utility of full text searching: as Richard Deswarte argued, 'its pre-eminence as the main approach to accessing web archives cannot remain unquestioned', and for Alison Kay, 'historians need to be contributing to discussions today about the sources of tomorrow' (Deswarte, 2015: 9).

Conclusion

Each of the ten case studies discussed here have moved the web archiving research front forwards, both in the specific areas they covered, and through the necessarily innovative methodological approaches they adopted. This reflects the initial aims of the BUDDAH project, which set out not only to 'highlight the value of web archives as a source for arts and humanities researchers', but also 'to establish a theoretical and methodological framework for the analysis of web archives' (Big UK Domain Data for the Arts and Humanities, n.d.). The previous section suggests that such a framework is indeed in development, albeit at a nascent stage. Moreover, the project demonstrated the importance of incorporating the perspectives of researchers at each stage of development of the dataset.

Yet this chapter has also made clear how much still remains to be done to ensure that the great potential of web archives as a source for arts and humanities research can be realized. The chapter has described the nascent strategies and techniques used by researchers on the project, but clearly many questions remain unresolved. These include, for example, how to handle messy and incomplete data; how web archive research is assimilated into the mainstream of a range of different disciplines; and how the results of search queries of the dataset can be meaningfully presented.

Since the conclusion of the research projects, both the underlying archive dataset and the Shine interface used to access it have been under continuous development, in large part in response to the researchers' experiences described here (Jackson, 2016). The interface is now accessible to everyone interested in conducting their own research about how UK society is reflected on the web between 1996 and 2013. The two main features currently offered are the search tool, which enables faceted browsing of the results, and a trend analysis tool, showing the relative appearance of a given word or phrase in the archive over the 18-year period. These tools, and more guidance, can be found at https://www.webarchive.org.uk/shine.

Crucially, however, those coming into contact with web archives in the future will not enjoy many of the advantages that researchers on this project benefited from, including contact with those developing the dataset, and the ability to share challenges and solutions as a group. As web archives continue to be developed, therefore, it is important that researchers as a user group are kept squarely in mind, even if they are not always in earshot. This chapter has illuminated many of the successes that researchers enjoyed, the challenges they faced, and most significantly, the ways in which they conceptualized and approached web archives as a source for scholarship. It is hoped, therefore, that this chapter – and the research projects that it profiles – can serve as a resource not only for scholars engaged at this emerging research front, but also for those involved at every stage in the development of web archives for research.

12
Coda: Web archives for humanities research – some reflections

Jane Winters

Introduction

For historians, and researchers in many other humanities disciplines, web archives remain largely an unknown, and certainly underused, primary source. Even within digital humanities, web archives as a focus for study have remained on the fringe, much more likely to be represented on the programme at events such as the ACM Web Science conference than the Annual Conference of the Alliance of Digital Humanities Organisations (ADHO).[1] There are many possible reasons for this – the particular focus of digital humanities, for example on textual editing; the difficulties of gaining access to web archives within national libraries and archives; the real and perceived technical barriers to working with this material; the paucity of digital skills training in the humanities generally; and simply the natural length of time it takes for new ways of researching to emerge and be recognized – but it is nevertheless a problem which needs to be overcome.

It is hard to imagine how one might study the history of the developed world[2] in the late twentieth and early twenty-first century without recourse to the archived web.[3] The traditional tools of the historian's trade – newspapers, letters, diaries, the records of government and business – are commonly, and in some instances now solely, online.[4] Some of these have been transformed – think of the relationship between, and intended audience for, a paper diary and a blog – while others are broadly similar in form and purpose but the method of delivery and consumption has changed. Our primary sources are increasingly on the web, whether we like it or not, and this is a trend which is unlikely to be reversed any time soon.

And time is important in another sense. The web was 25 years old in 2014, and an archiving process has been in place for almost 20 years, when the Internet Archive in the USA began its invaluable work, acknowledged elsewhere in this volume. For contemporary historians at least, this is beginning to look like a reasonable chronological span. The UK has traditionally adopted a 30 Year Rule in relation to the public release of non-sensitive government records, but in 2013 The National Archives began a move towards releasing records when they are just 20 years old, that is, the same age as the earliest instances of archived websites. It is becoming increasingly hard to argue that this is not material worthy of historical study.

So why, then, do web archives remain so persistently underutilized, so hidden from the mainstream of historical and digital humanities research? It was this mismatch between the clear value of web archives – for modern cultural, economic, political, social and technical history – and low levels of usage and awareness that we set out to address in the Big UK Domain Data for the Arts and Humanities (BUDDAH) project which has led to this present collection of essays.[5] The substantive research conducted during the project is described by Josh Cowls in his chapter on 'Cultures of the UK web', but the generation of this series of case studies was just one element of the project. We were concerned with the incubation of a community of humanities researchers who would move on from the project to advocate for the importance of web archives within their host institutions and among their disciplinary peers. The intention was not to transform them into 'web researchers' but to equip them to use web archives, and to encourage others to do the same. For most humanities scholars it will be a very long time before they transition to using solely digital sources, let alone solely born-digital sources, and for many this will never be the case. They will continue to mix and match, to compare and contrast, and to work with overlapping sets of material which contain subtly different information and are designed for subtly different audiences. Their research, however, will be impoverished if they are unaware of what web archives may contain – even if it is only to discount that information as unhelpful or unreliable.

Combining old and new approaches?

This is not, of course, to understate the challenges posed by web archives, as highlighted by many of the chapters in this volume. There are challenges arising simply from scale, or from the nature of the archiving

process, but there are also new conceptual challenges that will require innovative approaches and ways of thinking. Some of the problems are very familiar. For example, it is difficult to ascribe a clear date of publication to an archived web page. Even if all of the elements on a web page were captured at the same time, the date associated with them marks the point at which they were archived rather than the point of their formal publication (however we might think of this in an age of limitless editing possibilities and multiple versions). This seems far removed from the publication of a modern printed book for example, which will have an apparently clear date listed in the preliminary pages, or prelims.[6] But it is not uncommon even today to see bibliographic citations for serial publications along the lines of '2013 (really 2015)'. This indicates a discrepancy between the scheduled or official publication date and the date on which the book actually appeared in print. Which is definitive? And will the answer be the same in 50 years' time? Medieval manuscripts may be datable to, for example, only a rough 25-year period. Uncertainty about date is something with which historians have always had to deal. Not so is the presence in the archive of a 'web page' which never actually existed. The memento protocol pieces together the 'best' view of a page, bringing together elements from different archives captured at different points in time. The British Library home page on 20 July 2009, for example, may be assembled using 14 mementos from four separate archives, spanning four months. Despite the superficial similarity of the process to critical or scholarly editing,[7] this is a new phenomenon which is embedded within the archive itself and not imposed subsequently by one or more human editors.

It is clear that humanities researchers need to acquire new skills and develop new methodologies if they are to get to grips with web archives as a source, but much can be achieved either by repurposing and adapting existing analytical frameworks or simply by approaching digital data with the same critical eye that one might bring to incunabula or to early modern newsletters. Eric Ketelaar, for example, has argued persuasively that diplomatic, traditionally applied to medieval documents, may also be useful for the analysis of digital materials like web archives: 'The principle of provenance and other basic tenets of archival science can be put to new uses in the digital age'[8] (Ketelaar, 2007: 167–91). Existing methodologies may be adapted to accommodate different data structures and different signifiers of purpose, authority and authenticity, in combination with new tools, approaches and theoretical frameworks.

This, however, is to take a primarily micro-historical approach to the study of web archives, to search for stories about particular individuals, institutions or events. There is scope for the macro-historical too, as championed recently by Jo Guldi and David Armitage in their call to arms, *The History Manifesto*. This deliberately provocative book, which includes a chapter titled 'Big questions, big data', argues that

> Together, micro-historical work in archives and macro-historical frameworks can offer a new horizon for historical researchers who want to hone their talents of judging the flow of events and institutions across centuries and around the globe as well as a new opportunity to engage with the public. (Guldi and Armitage, 2014: Conclusion)

The key point here, and one which has been overlooked by some commentators, is the combination of approaches – there is room for what Tim Hitchcock has described as 'beautiful histories of small things' (Hitchcock, 2014) but also for the historian's macroscope (Graham et al., 2015). The data in which humanities researchers are most interested is characterized by complexity and mess because it reflects and records complex and messy human interactions. Hitherto unsuspected patterns emerge when it is analysed at scale, but these can only be tested by digging in to the data and understanding the individual elements which make up the whole.[9]

Nowhere is this approach more apposite than when working with web archives, as evidenced by the research presented in this volume. The histories of individuals and organizations, at least as they played out online, can be traced over the past 20 years. Conclusions may be drawn about how the culture of an institution has evolved; how a government department has interacted with the public (and what information it has deemed to be most important to communicate at particular points in time); how a small business has expanded and/or contracted; how an individual has reflected on their journey through illness or on their family life. Alternatively, wider social and cultural changes may be traced through the online development of a single organization. How have changes in design and technology influenced a company's web presence? What has been the effect of developments in e-commerce on its online services? How has the increased penetration of the web into everyday life affected the language used to communicate with users and consumers? How, if at all, has it accommodated social media and the

growing customer expectation of increased interaction, sometimes in real time?

However, it is also possible to study wider patterns and trends, for example to attempt to trace developments in language, to undertake complex network analysis or to track the movements of peoples and political ideas. There is no need to rehearse again here the many difficulties posed by web archives for this kind of research, but the fact that it is challenging does not undermine its enormous potential value. Even a very simple n-gram approach can produce immediately suggestive results, for example when identifying neologisms and the point at which they become widely adopted. In the UK, the *Oxford English Dictionary* (*OED*) produces an annual 'word of the year', chosen because it has risen to prominence in the previous 12 months, or might in some way be said to characterize that period. A comparison between some recent *OED* choices and instances of those terms in the archive of UK web space for 1996–2013 reveals that the selections of the dictionary's experts are mirrored (driven?) by online trends. In 2004, the chosen word was 'chav',[10] and the trends graph developed for the British Library's 'Shine' interface reveals a clear spike in mentions of the term in that year. In 2003 it appeared just 923 times, but in 2004 this figure jumped to 60,467 (an increase of 6,551%). In 2008 'credit crunch' was nominated and the pattern in the web archive was very similar (even if the order of magnitude differed): in 2007 there were 128,152 instances of the phrase, while in 2008 this rose to 1,555,960 (an increase of 1,214%). Interestingly, the web archive indicates a rather different fate for these two 'words of the year': 'chav', perhaps rather unfortunately, persists, but there is a sharp drop in instances of 'credit crunch'[11] relative to the archive as a whole after 2009. It would seem that it was specific to a particular moment and set of circumstances, or at least as it was used on the web. These are, of course, very simple examples found using a simplistic methodology, but nonetheless interesting.[12]

Moving beyond text (and search)

The digital humanities embrace a wide range of methods and sources, but much of the most innovative work to date has been concerned with the analysis of text.[13] Web archives contain a great deal of text, from formal publications and newspapers to material verging on direct speech (some social media), but the data is distinguished by its variety.

There are varieties of textual information – html pages, MS Word documents, PDF files – but there are other media too – videos, image files, sound clips, animated gifs. The web is becoming an ever more visual medium, with the dominance of services like YouTube and Instagram and the ease with which photographs and video can be captured and uploaded to the web from smartphones. Much of this data is beyond the reach of web archives as they currently exist. This is either because it falls outside a nationally-harvested country code Top-Level Domain (ccTLD) or because it is the property of a commercial service provider like Facebook. There is, however, a great deal that falls within the scope of the archives. A British Library visualization of popular image formats in the archive of UK web space for 1996–2010, for example, reveals that JPEGs alone account for 10% of the total crawl in 2010 (the figure is roughly consistent across the whole period). The analysis of non-textual big data at scale is a significant challenge that will only become more pressing as born-digital data becomes a focus of research. Traditional image databases, like the John Johnson Collection of Political Ephemera or the Warburg Institute Iconographic Database, rely on the generation of exhaustive metadata to support discovery, but this is not present for the bulk of the films, images and sound clips in web archives. The problem is particularly acute for platforms and services where the addition of metadata is largely optional and almost entirely uncontrolled. A very simple example serves to illustrate the problem. It is a truism that the web is overrun with pictures of cats. Searching the 'Shine' interface for 'cat' and limiting results to the content type 'image' produces 340,453 instances of the term. This would, of course, not be a sensible search to conduct, as is clear from an investigation of the first few images listed: the initial four are blocked because of robots.txt; the fifth is indeed a photograph of cat; but the sixth is a pair of Caterpillar boots, the seventh a 'music catalogue' gif, and so on. It is here that existing methods of interrogating data begin to break down.

The dominance of (a particular type of) search as a digital research method very quickly becomes problematic for web archives where, quite apart from difficulties arising from scale, the scope of a particular archive is unknown and the process of creation largely undocumented. Discovering what might be in the archive is often the primary objective – and this is not well served by keyword searching which produces a list of results unordered by anything other than date. For sound and image, moving or still, there is the extra limitation of poor or non-existent metadata for even crude keyword searching. The absence of metadata is a limitation here in another way too. Images are one of the elements of

a web page which the crawl process is more likely to fail to capture, and the absence of metadata or alternative text confers invisibility. This may be seen, for example, in the capture of the home page of the Institute of Historical Research (IHR), University of London in the Internet Archive from 1 December 2003. There are three broken image links at the top left of the page, but the associated text makes it clear that they are two logos – one for the IHR itself and one for the 'History' website – and a picture of the building that hosts the institute. For other missing images on the page, however, there is no associated text so it can only be guessed what might have been used to illustrate, in this instance, training courses for Latin and for Palaeography and Diplomatic.

(In)completeness and loss

Web archives raise questions of (in)completeness. Should we be trying to keep everything, particularly as existing methods of selection and cataloguing are not scalable? If we do not know what future scholars will be interested in, should we simply collect it all? And what do we mean by 'everything', when the web archiving process is marked by patchy data collection and loss? Web archives are, after all, only an often partial snapshot in time. Notions of comprehensiveness exist simultaneously in our consciousness with the counter-narrative that we are about to enter or have already entered a 'digital dark age' which will see the historical record lost for future generations. Of course, neither of these is true, but questions of survival and loss do seem to loom particularly large in relation to born-digital data, including web archives.

This seems to me to be an old problem of the survival of evidence filtered through a new expectation that it is somehow possible, even desirable, to keep everything. This is to ignore the fact that the primary sources we value so much from earlier periods have in large part survived through historical accident. One particular monastic library burnt to the ground while another did not; one individual was more diligent at keeping her correspondence than another; one national archive was bombed during the Second World War while another was spared when an incendiary device failed to go off. Medieval historians, for example, become used to working with and around gaps, to speculating about the representative nature of a particular set of records, to trying to reconstruct a legal code from vague references to it in other documents. Perhaps the difference in focus comes from our ability to know precisely what we do not have when we are dealing

with web archives. A missing image confronts us with a blank square on the web page; a broken link produces an error. To take one example, the first capture of the IHR website in the Internet Archive dates from 27 December 1996, but the website went live on 9 August 1993 (Segell, 1993: 4). We are immediately confronted with the fact that more than three years' worth of data no longer exists. Data loss is also a very real presence in our daily lives, whether it is the disappearance of whole services which once seemed essential or the failure to back up a much-used computer.

The susceptibility of the web to archiving may, however, lead to other kinds of gaps. It is noted on the British Library website, for example, that 'Where […] web crawling software encounters a login facility, it cannot access any material behind the login facility without the appropriate password or access credentials'. In practice this means that data of this kind is not captured, so openly published information is privileged in the archive. This has potentially fascinating implications for what will remain available to researchers in 10, 15 or 50 years' time. In an interesting reversal of previous patterns of data survival, might open data be more likely to persist than commercially managed and published digital material?[14] Publishing companies are, of course, taking steps to ensure the long-term availability of their outputs, but they are often working outside the national infrastructures that underpin web archiving. Which is the more likely to last, if we accept that the digital presents a sustainability challenge? And what of apps,[15] which are largely closed systems unsusceptible to archiving by national institutions, let alone the wealth of data published via social media platforms such as Facebook. This is truly vulnerable information, reliant on the self-interest of corporations for its maintenance (Webster, 2015). Might we be forced to rethink what we consider to be ephemeral?

Unlocking value

The sheer variety of information contained in web archives poses huge difficulties for researchers, but this mixture of formats and types, of the personal and official, of the public and private is precisely why they are such an important primary source for humanities researchers. It is at once possible to compare, for example, the official announcement of a government policy with its subsequent coverage in newspapers and other online media, and then with its discussion in online forums and selected

social media. Perhaps the policy is a controversial one which results in the creation of an online petition, which in turn triggers a debate in parliament.[16] All of this information may be found in the archive, even if it is not in any sense comprehensive or indeed easy to locate. Our stories and our histories are increasingly online, but the inherent ephemerality of the live web means that they only achieve any degree of permanence in web archives.[17]

There are clear technical and methodological hurdles facing researchers who wish to study these histories, but simply gaining access to web archives introduces an additional layer of complexity. Over the past two decades and more, researchers have become used to the widening of access through digitization and the increased availability of digital materials online. Hierarchies, of course, remain – notably between those within and outside well-resourced universities – but nevertheless more people have greater access to the historical materials held in our national memory institutions than at any point in our history. And that access is often international – the selective open UK Web Archive, for example, can be viewed from anywhere in the world. But legal frameworks which have failed to keep up with changing technologies and modes of communication mean that artificial barriers are being erected around web archives which are preventing the integration of their study into the mainstream of humanities research. In the UK, access to archived websites and electronic publications is severely restricted by legal deposit regulations, with the result that 'deposited works may not be made available online externally, including for readers logging in remotely. They can only be viewed on the premises of the six deposit libraries'.[18] Moreover, 'the 2013 Regulations stipulate that "A deposit library must ensure that only one computer terminal is available to readers to access the same relevant material at any one time"' (Netarkivet, n.d.); in other words, two people may not look at the same instance of an archived web page concurrently. In other countries which archive their ccTLD, access may be even more restrictive.[19]

All of this is against a background of increased expectation not just of open access to data but that there will be APIs which allow researchers to download and take away the material with which they choose to work. It is the portability of data, its separability from an easy-to-use but necessarily limiting interface, which underpins much of the most exciting work in the digital humanities. Web-based tools such as Voyant have brought quite sophisticated textual analysis within the reach of anyone who has access to data, but data from web archives can be very hard to come by. When it is available, as with the host link graph derived

from the UK domain dataset 1996–2010 used by Meyer and colleagues to examine 15 years of UK universities on the web in this volume, the results are fascinating and suggestive of numerous avenues for research. Initiatives like the Common Crawl, which provides 'an open repository of web crawl data that can be accessed and analyzed by anyone', are doing important work here too. Problems of access are not, of course, unique to web archives, but if it is made too difficult for researchers to engage with the data, they will turn elsewhere or simply rule out using web archives as a source.

A perception of difficulty is most damaging for those who might study web archives as just one of a number of primary sources, including printed newspapers, the paper records of government, film and television, and other kinds of digitized data. They are not concerned with the history of communication or technology, but with what the archived web can reveal about the development of a popular political movement, health scare or terror attack. These are the researchers whose work on the BUDDAH project has been admirably synthesized by Josh Cowls and, as noted above, it is typified by a mixture of both methods and sources. They do not have the time, or indeed the willingness, to develop the full range of skills that might be expected from a specialist; nor do they commonly have access to the high performance computing facilities that working with web archives may require. They are likely, however, to be the key to increasing familiarity with and usage of the growing volumes of data that archiving institutions are collecting and storing, often at considerable expense in a time of generally straitened finances. That is where a volume of this kind, which showcases innovative research using web archives and presents a range of use cases for different humanities disciplines, is so useful. If the BUDDAH project is any indication, it often only takes dipping a toe in the water for researchers to discover the value of web archives.

If web archives need to be integrated into established processes and workflows in order to become widely consulted, they also need to be considered in debates about approaches to working with born digital big data more generally. National libraries and archives are not simply responsible for archiving the web; they are increasingly having to deal with email archives, with institutional and departmental file systems, and with personal digital data. The difficulties of storing, preserving and making available these different types of data vary, as do the problems facing researchers who wish to study them, but there are commonalities too, which go beyond mere scale. One such is the question of how you protect individuals in this mass of information. It is not just individual

pages or documents which may be sensitive but the combination of those pages and documents, or even very small snippets of data. These may reveal a larger picture or more information about someone than they would either anticipate or be comfortable with. The reverse is a problem too – how do you securely identify persons of interest in large-scale and complex data where diversity in naming is almost systemic? This is to return to the requirement for new theoretical and methodological frameworks identified above, which these chapters, and the explicit connections between them, are helping to advance. Other interdisciplinary and international forums and networks are developing to consider these questions and, as a crucial first step, to articulate which problems are common to many forms of born digital data and which relate only or primarily to web archives.[20] It is a thriving and vital (in both senses) field of research.

At present, and necessarily, scholarly debate has tended to focus on the impediments to working with web archives, and on the sheer effort involved in making sure that this data is captured effectively. This edited volume has sought to move the discussions on, to make available the first fruits of research – and in an open access form which introduces them to the widest possible audience. It is a starting point, a signpost to future interesting locations which may be reached by more or less circuitous routes. As the roadmap becomes clearer, and the data begins to be better understood, it is to be hoped that the enormous richness of the archived web will come increasingly to the fore. Like any new area of investigation, any new type of primary source, it takes time before its full potential is realized. These chapters are a first, and fascinating, indication of what we might expect.

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 dens les années 1990' (2014–16) and 'De #jesuischarlie 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).
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.diplomacie.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 warcbase platform, a web archiving analytics platform led by Jimmy Lin (University of Waterloo) available at http://warcbase.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., Cowls, 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., Cowls, 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-mouse 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 websites. *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.linkeddata.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., Cowls, 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 of 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 journalisms: 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. Brugger (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. 967
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., Cowls, 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/aboutthebbc/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. Accessed 22 August 2015.

Bouquillion, P. and Pailliart, 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.

Curtil, 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.

Desautez, 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 'l'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=flteach&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=flteach&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

Alaithea (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://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.

Html_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/1PqSZmKwQwSIFrbmVievbStTbt7PrtsxNgC3W1oY5C4/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-Clio.
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/nettext. 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.
Gansuage, 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.themacroscope.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
cyberbalkinization, 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 Consortiuum), 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
Issuecrawler, 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

ISBN 978-1-911307-42-6
90000
9 781911 307426