INNOVATIONS IN DIGITAL RESEARCH METHODS

PETER HALFPENNY
ROB PROCTER

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC | Boston

1

INTRODUCTION AND OVERVIEW

PETER HALFPENNY AND ROB PROCTER

1.1 INTRODUCTION

The dramatic increase over the last two decades or so in computing power, in wired and wireless connectivity, and in the availability of data has affected all aspects of our lives. Our aim in this book is to provide an accessible introduction to how social science researchers are harnessing innovations in digital technologies to transform their research methods. In this chapter we provide an overview of how and why e-Research methods have emerged, including an account of the drivers that have motivated their development and the barriers to their successful adoption. The chapters that follow examine how innovations in digital technologies are enabling the emergence of more powerful research infrastructure, services and tools, and how social science researchers are exploiting them.

1.1.1 Digital Data

As everyone exposed to the Internet is aware, the amount of digital data available is expanding very rapidly, both through the digitization of past records and by the accretion of 'born digital' materials that are in machine-readable form from the outset. The digital universe – the data we create and copy annually – is estimated to be doubling in size every two years and projected to reach 44 trillion gigabytes by 2020 (where a trillion is a million million, or 10^{12}) (IDC, 2014). For social scientists, the predictions that more data will be generated in the next five years than in the entire history of human endeavour is both an opportunity and a challenge.

Today, vast amounts of data are generated as people go about their daily activities, both data that is deliberately produced and that which is generated by embedded systems. For example, use of public services is captured in administrative records; in the private sector, patterns of consumption of goods and services are captured in credit and debit card records; patterns of personal communications are captured in telephone

records; patterns of movement are logged by sensors, such as traffic cameras, satellites and mobile phones; the movement of goods is increasingly tracked by devices such as radio-frequency identification (RFID) tags; and the advent of the 'Social Web' has led to an explosion of citizen-generated content in blogs and on social networking sites.

Currently, these data sources are barely exploited for social research purposes. The potential benefits to researchers are enormous, offering opportunities to mount multidisciplinary investigations into major social and scientific issues on a hitherto unrealizable scale by marshalling artificially produced and naturally occurring 'big data' of multiple kinds from multiple sources. However, exploiting these digital data sources to their full research potential requires new mechanisms for ensuring secure and confidential access to sensitive data, and new analysis tools for mining, integrating, structuring and visualizing data from multiple sources.

1.1.2 e-Infrastructure

Since the beginning of the new millennium, a world-wide effort has been underway to create the research infrastructure and to develop the research methods that will be needed if the 'data deluge' is to be harnessed effectively for research. A new generation of distributed digital technologies is leading to the development of interoperable, scalable computational tools and services that increasingly make it possible for researchers to locate, access, share, aggregate, manipulate and visualize digital data seamlessly across the Internet on a scale that was unthinkable only a decade or so ago.

e-Infrastructure comprises the information and communication technologies (ICTs) – the networked computing hardware and software – and the digital data that are deployed to support research. A very broad definition has been adopted by Research Councils UK (2014), which spells out more fully the components that are brought together:

> e-Infrastructure refers to a combination and interworking of digitally-based technology (hardware and software), resources (data, services, digital libraries), communications (protocols, access rights and networks), and the people and organisational structures needed to support modern, internationally leading collaborative research be it in the arts and humanities or the sciences.

This definition highlights the complexity of e-Infrastructure and, correspondingly, the enormity of the socio-technical efforts required to efficiently integrate distributed computers, data, people and organizations in order to deliver tools and services that scientists can readily adopt to their advantage in pursuing their research. (In the US, the term cyberinfrastructure is more commonly used than e-Infrastructure.)

e-Research is the generic term that has been coined for the innovations in research methods that are emerging to take advantage of this new and vastly more powerful e-Infrastructure. Similarly, e-Social Science is the research facilitated by the e-Infrastructure. The 'e' in all these terms is short for 'electronic', although it is sometimes rendered as 'enhanced'.

The scope of the book is the application of e-Research methods across the social sciences, including both quantitative and qualitative data collection and analysis. The aim is to introduce the reader to the application of innovative digital research methods throughout the research lifecycle, from resource discovery, through the collection, manipulation and analysis of data, to the presentation and publication of results.

1.2 BACKGROUND

1.2.1 e-Science

Over the period 2001 to 2006, the UK Government invested £213m in an e-Science programme (Hey and Trefethen, 2004). The overall aim of the programme was to invent and apply computer-enabled methods to 'facilitate distributed global collaborations over the Internet, and the sharing of very large data collections, terascale computing resources and high performance visualizations'.[1] The funding was divided between a 'core programme', focused on developing the generic technologies needed to integrate different resources seamlessly across computer networks, and individual Research Council programmes specific to the disciplines they support. The Economic and Social Research Council (ESRC) allocation was £13.6m over the five years, with the major part of this investment devoted to setting up the National Centre for e-Social Science (NCeSS). The Centre had a distributed structure, with a coordinating Hub responsible for designing and managing the programme and eleven large three-year projects devoted to developing innovative tools and services and applying them in substantive fields of inquiry.

The ambition of the overall e-Science programme was to promote the adoption of innovations in digital infrastructure to facilitate bigger and faster science, with collaborators worldwide addressing major research questions in new ways. The initial technical focus was grid computing, driven by a set of 'middleware' standards. These are the shared protocols required for the development of sophisticated software to enable large numbers of distributed and heterogeneous computer systems to be linked and inter-operate, thereby providing researchers with seamless, on-demand access to scalable processing power to handle very large-scale datasets, regardless of the location of the researchers or the data. This model of e-Infrastructure was particularly appropriate to particle physics and such challenges as weather prediction and earthquake modelling. Advances in these areas are dependent on collecting and marshalling data on a vast scale and having huge computing resources to analyse it, accessible by large networks of research teams distributed across the world.

However, the grid computing blueprint for e-Infrastructure proved slow to mature, sometimes difficult to deploy in practice and it did not always offer the most appropriate solutions to scientists' requirements. Meanwhile, other technologies emerged

[1]www.epsrc.ac.uk/about/progs/rii/escience/Pages/intro.aspx. (All URLs were accessed on 17 Dec 2014.) Terascale computing achieves speeds of teraflops, where a teraflop is a trillion floating point operations per second.

and alternative solutions to the demand for scalable computing and data storage, such as cloud computing, became available. Alongside this was the flowering of the lightweight systems that are loosely collected together under the title of Web 2.0 (O'Reilly, 2005). While these are technically less powerful than grid-based systems, their relative simplicity – both in terms of implementation effort and ease of use – made them attractive to researchers who did not need sophisticated tools and services, and who were deterred from using grid services by their complexity and the perceived barriers to access. Moreover, many of these Web 2.0 tools and services are freely available on the Internet, and users can find help in adopting them in numerous online forums and support groups. They have been widely taken up because of their ability to deliver easy-to-use services via simple protocols and familiar Web-based user interfaces, and they provide flexible solutions to at least some researchers' needs for advanced computing tools and services. Accordingly, across the sciences the notion of grid computing being at the core of e-science gradually gave way to a wider understanding of e-Infrastructure, embracing a broad range of computing software and services that support the everyday work of scientists.

1.2.2 e-Social Science

From the start of the e-Science programme, the ambitions of grid computing were less matched to those disciplines subsequently encouraged to join the e-Science bandwagon, including the social sciences, where a mixture of numerous quantitative and qualitative methods is used to pursue relatively small-scale issues. These disciplines have very few generic problems requiring complex middleware to coordinate huge distributed computing and data resources. What requirements they do have were already – before the e-Science programme was initiated – well-served by established commercial and open-source packages to, for example, computer-assist personal interviewing, deliver Web-based surveys, manipulate and statistically analyse quantitative data, sort and code qualitative data, and visualize findings in tables, graphs and network diagrams. Moreover, competition between the commercial package vendors seeking sales to the market research industry as well as to the social research community maintained a flow of updates, including integration of different tasks from around the research cycle into single packages. Similarly, much of the open-source software continued to develop through the efforts of often very active and technically adept support groups.

As the NCeSS research programme unfolded within the changing technical environment, instead of focussing on grid computing, e-Social Science broadened out to include a diverse range of initiatives exploring how computer support and networking, as well as new sources of data including that harvested from the Web, could be used in new ways to capture people's views and map their behaviours and their networks. These projects included an exploration of new forms of digital data, such as mobile phone logs and GPS to track people's interactions (see Chapter 9); the creation and exploitation of metadata (that is, data about data, such as its provenance) to facilitate the sharing and reuse of research data (Edwards et al., 2011); linking data about individuals from different sources and the confidentiality and ethical issues

that this raises (Duncan et al,. 2011); webometrics, that is, measuring the number, types and patterns of hyperlinks in the Web (Thelwall, 2009); creating maps of geo-referenced data to reveal patterns such as the location of crime hotspots (Hudson-Smith et al., 2009); large-scale social simulations of, for example, the demand for housing in a city and how it changes over time (Birkin et al., 2010); parallelization of statistical routines to make more efficient use of computing time (Das et al., 2010); enabling researchers to collaborate in marking up videos to highlight significant aspects of the social interactions they record (Fraser et al., 2006); mining large bodies of unstructured text for patterns (Ananiadou et al., 2009a; 2009b); and developing software for delivering behavioural interventions over the Internet (Webb et al., 2010). Many of these initiatives will be further described in the chapters that follow.

As these examples reveal, the e-Social Science programme became highly disparate, expanding to include an increasingly wide range of emerging digital technologies, and drawing on many of the new forms of digital data that were becoming increasingly accessible. The various projects demonstrated that a modest input of technical support could ease existing research processes. This proved particularly productive when there was very close engagement between computer scientists and social scientist users in order to track and respond to changing requirements so that research practices and computing tools could co-evolve. However, successful co-production requires that effective local support structures are established and delivered 'at the elbow' of the users (Procter et al., 2013a). This leads on to the wider issue of user adoption, and the barriers to and facilitators for this.

1.2.3 User Adoption

As we noted earlier, the adoption of innovations in research methods and tools has been on a smaller scale to date than the e-Research vision initially anticipated. e-Science's radical ambitions for transforming everyday research have been tempered in the light of growing evidence about the very real barriers slowing widespread adoption of advanced tools and services across the science community. This extends to the social sciences too. We have already noted that computer packages to support most tasks in the social science research cycle were available before the e-science programme was launched. What many social scientists seek are more efficient or user-friendly versions of these existing digital tools rather than a transformation in their approach facilitated by novel e-Infrastructure, and they have often lacked the resources or incentives to take up the new methods that it offers.

Although a small cadre of 'early adopters' – mostly involved in the e-Social Science research programme – have been keen to experiment with innovations and to take risks, adoption of even the broader e-Infrastructure by the wider social science research community has been handicapped by a complex of factors (as has e-Research as a whole: see Voss et al., 2010; Procter et al., 2013a). These include a lack of awareness of the opportunities e-Infrastructure provides; problems in translating innovations in one field into benefits for one's own research; risk aversion; and levels of IT support that are often dictated by institutional policies and priorities rather than individual researcher needs. Late adopters are often resistant to training and require shallow

learning curves if they are to invest in new skills and adopt new ways of working. They may feel they can achieve their career goals – publications and promotions – using the tools with which they became familiar as graduate students. This environment is not conducive to the wide uptake of innovative tools and services or the pushing of boundaries.

Another factor hampering uptake is the uncertain path of technological innovation, which affected the whole of the UK e-Science programme from its launch in 2001. During the early stages of any innovation, the existence of competing technical solutions can be a disincentive to adoption. The emergence of alternatives to grid computing middleware, such as Web 2.0 tools as noted above, introduced uncertainty about the future direction of e-Infrastructure technology development. Studies of previous infrastructure innovations suggest that technological uncertainty may deter some potential users from engaging, at least until a clear technical winner has emerged (Edwards et al., 2007). This uncertainty has been amplified over the last decade as publicly funded research services have faced competition from commercial suppliers, for example, in the provision of cloud computing, with infrastructure, platforms and software all offered to users as subscription services. While this relieves users of the cost of support and maintenance, they lose control over the development path, which is driven by commercial priorities.

A further uncertainty in the future trajectory of emergent e-Infrastructure is its sustainability, that is, the resource-intensive path from research, through software development to delivery of services and support to users. To illustrate: even the more tractable new users will adopt new tools and services only when these are 'hardened' to production level, that is, become easy to use, stable, reliable, documented, maintained and fully supported. This requires that software development pathways be created that ensure that e-Infrastructure is able to move beyond the research stage, that is, beyond proofs of concept, demonstrators and prototypes, to production level tools and services. It is ease-of-use and the utility of e-Infrastructure, and its contribution to advancing social scientists' own substantive research that would persuade them to adopt new ways of working.

The achievement of sustainability is adversely affected by several aspects of the current academic reward system. One is the distinction between 'pure' computational research and 'applied' software development, with the former bringing rewards for 'proof of concept' software innovations but the latter – involving re-building the software to make it robust and efficient – being little rewarded within academia, to the extent that there are few developers to be found even in computer science departments, let alone social science departments. Yet without significant development work most 'proof of concept' innovations – such as those emerging from the e-Science programme – are unusable except in the hardware and software context in which the researcher constructed them. Earlier in this chapter, the advantage of software co-production was noted, but this requires collaboration not just between computer scientists and social scientist users, but also the addition of developers to the team, who can re-build innovative tools so that they become project-independent.

There is a similar distinction between both research and development on the one hand and service delivery on the other. The latter requires documentation, online or face-to-face support, FAQs, software maintenance, bug fixes, distribution, porting to new operating systems and so on. Service delivery to support e-Infrastructure is essential for effective and widespread use of e-research resources, but has little place in academia except in a very few specialized units.

Given the co-ordinated efforts of computer scientists, developers and service providers needed to deliver e-Infrastructure that can be readily deployed by users, and the lack of such organizational and human resources in many academic departments, it is not surprising that researchers tend to restrict themselves to the sorts of social science that can be achieved through an unsystematic mix of existing technologies with which they are most familiar.

The next section introduces the materials in the following chapters, which are designed to increase awareness of the opportunities that e-Infrastructure offers to transform social research. We begin with chapters focused on understanding the potential and challenges of new sources of social data for social research, while not forgetting that much can yet be done to enhance the use of more conventional data sources, such as surveys. We then turn to examining innovations where e-Research offers tools that open up new opportunities for social research across a broad range of topics. All of our contributors make clear in their individual chapters that they are aware of the issues around research ethics posed by new sources of social data and more powerful tools for analysis. Such is the importance of this topic that we include a chapter devoted entirely to it. Finally, this book had its genesis, in part, as a response to Savage and Burrow's widely cited paper, 'The Coming Crisis in Empirical Sociology' (2007). We believe that the chapters in this book present plentiful evidence that innovations in digital research methods have the potential to radically transform academic sociology, and we thought it appropriate to let one of the paper's authors have the last word on whether this transformation represents a crisis or an opportunity to be seized.

1.3 THE CHAPTERS

The chapters in this volume have been selected to provide an informative introduction to innovations in social science research methods and tools, along with a review of issues and challenges that remain to be resolved if researchers are to enjoy the full benefits of the innovations.

The chapters reflect the various ways in which social science research has changed under the influence of both new sources of social data and innovations in research infrastructure and tools. The social sciences are known for diversity of methods, and their quite different ideas about how to study and make sense of the social world. One fundamental distinction is what is often referred to as the quantitative-qualitative divide and another is between the use of primary and secondary data. Innovative digital tools have the capacity to blur both distinctions, as several chapters reveal.

Chapter 2: The Changing Social Science Data Landscape

This chapter reviews the new sources of social data being made available by a combination of new data services and changes in government policy on access to administrative records. It also notes the rapid expansion of born digital and big social data – of which social media comprise but one, admittedly high profile, example. In the chapter, Purdam and Elliot examine how access to the new data opens up new opportunities for social researchers and, drawing on an eight-point typology of new kinds of social data, they present a series of real world examples to illustrate how the social sciences can benefit from them. They also discuss some of the potential challenges for social researchers of using these new sources of social data, such as variable data quality, questionable generalizability and representativeness, and restrictions on free access to some kinds of social media data, and they explore the implications of these and other challenges for the practice of social science research.

Purdam and Elliot argue that the almost effortless capacity to collect new kinds of social data poses the risk that researchers will neglect theory in favour of more data-driven methods. They also speculate on how access to social data in real time ('datastreams') might lead to a blurring of the boundaries between research and policy intervention. Finally, in what is a recurring theme throughout this book, they examine some of the ethical issues that accompany the use of new forms of social data in research.

Chapter 3: Exploiting New Sources of Data

In this chapter, Elliot and Purdam take up the methodological challenges, outlined in Chapter 2, that researchers face if they are to make effective use of new sources of digital social data. They employ a series of case studies of research, including election campaigns, civil unrest, migration and mobility, and health and well-being, to illustrate how methodological innovations, such as crowd-sourcing, may be mobilized to meet the challenges.

Opinions on the value of new forms of social data have divided academic social researchers, with some taking the view that the discipline is on the threshold of a renaissance, including opportunities to study the social world in real time. Others dismiss such claims as naïve at best and – at worst – sacrificing methodological robustness and validity for convenience. Regarding the latter, numerous critics have raised concerns that the sheer volume of new forms of social data will make computational methods increasingly attractive to researchers and lead them to ignore the risks of relying on computer power to drive their analyses. One such risk is that posed to the verification and repeatability of results, which arises from using complex, and sometimes proprietary, algorithms that lack transparency; for example, the operationalization of statistical formulae in the packages researchers use are hidden from them. Another risk is that posed to meaningful understanding of social phenomena by the lure of spurious correlations thrown up by over-reliance on inductive methods.

Mindful of these problems, Elliot and Purdam argue that the solution is a middle course, combining new and conventional sources of data in a robust, mixed methods approach that bridges data- and hypothesis-driven traditions. There are, of course, obstacles to be overcome. New sources of data such as social media may increase threats to privacy, and Purdam and Elliot call for more research into ways of countering these threats through improved methods for data anonymization and a new ethical framework. In relation to the latter, they note that it is time that citizens realized the value (economic and social) of their own data and, equally importantly, they argue that commercial interests must not be allowed to constrain researchers' access to new forms of social data.

Chapter 4: Survey Methods: Challenges and Opportunities

In this chapter, Murphy considers the future for data collection and using survey methods in the context of new sources of digital social data and technical innovations in research methods and tools. He sets the scene by discussing current challenges for survey research, such as declining response rates in traditional face-to-face, telephone and mail surveys, alongside the opportunities that technical innovations provide for enhancing the quality and efficiency of survey research methods. Drawing on a selection of major social science surveys, Murphy offers examples that point toward the continuing importance of survey-based methods in the social sciences.

Murphy observes that, despite the proliferation of born digital data, recent years have nevertheless witnessed an explosion in the quantity and diversity of data generated through survey research. This has been facilitated by developments in e-Infrastructure, an example being the unprecedented opportunities for the recruitment and retention of respondents afforded by the public's mass adoption of email and, subsequently, social networking sites. Similarly, survey researchers have benefited from the increasing availability of paradata, that is, data about survey transactions and interactions with respondents, which can be used to gain insights into their motivations and the meaning behind their responses. Such e-Infrastructure affordances have made significant advances in the capture, analysis and dissemination of survey data possible. They lead Murphy to argue that, contrary to predictions that new sources of data will make surveys redundant, they offer ways both to make surveys more effective tools and to meet the challenges that have threatened their value. For example, the availability of administrative data and methods for matching it with survey data hold great promise for minimizing respondent burden and cost. In a different vein, Murphy observes that virtual worlds, such as Second Life, offer new ways of conducting interview-based surveys.

Nevertheless, Murphy reminds us that social media brings new challenges, in particular, the problems of bias through samples of unknown representativeness, and quality assurance. The prospect of using social media as a substitute for traditional surveys – for example, the use of Twitter as a way of measuring public opinion through sentiment analysis – is often heralded as a sign of their imminent demise. Murphy, however, warns of the dangers of relying on such data where there is '… no

standardization or check on the validity of the information being shared'. He argues, instead, for more research into the value of Twitter as a means to recruit respondents, citing as an example a recent study where it was used in diary data collection. Finally, Murphy discusses the potential of mobile devices for SMS-based survey delivery, noting its efficacy for administering them at predetermined times or in the context of specific events or – when used in conjunction with GPS – specific places.

Murphy's conclusion is that survey methods are continuing to play a major role in social research, and pessimism about their survival is misplaced. This role, however, is increasingly being shaped by people's use of communication technologies. Given the rapid pace of innovation of these technologies, the future for survey methods remains hard to predict.

Chapter 5: Advances in Data Management for Social Survey Research

As argued in the previous chapter, despite the availability of new sources of social data, making optimal use of more conventional data sources such as surveys remains of critical importance to social research. However, using survey research data can present major challenges for data management. For example, pursuing a particular research question may require linking different datasets, extracting variables, combining them and recoding their values before statistical analysis can start. In this chapter, Lambert argues that data management practices have failed to keep pace with these challenges and explains how e-Research can advance the state of the art, drawing on examples of working with quantitative datasets generated through social surveys taken from the DAMES (Data Management through e-Social Science) project.[2] He argues that enhanced facilities for file storage and linkage, for using metadata to describe data, and for the capture of data preparation routines ('workflows') can raise standards in data management and help researchers share their experience and expertise with one another. (Exercises illustrating each of these facilities can be found at the book's website.)

Lambert concludes by examining the prospects for the adoption of more advanced data management tools and practices. Using an example where 'bottom-up' and 'top down' innovation processes might successfully complement one another, he notes how the push from journals and funding agencies for researchers to publish metadata about their data management is likely to have a decisive influence.

Chapter 6: Modelling and Simulation

Quantitative simulation and modelling are perhaps the most obvious examples of the potential for e-Research methods and tools to revolutionize the study of complex socio-economic problems, and their applications are becoming increasingly widespread. New sources of data and more powerful computational resources have made possible the development of more complex and sophisticated techniques and, of

[2]www.dames.org.uk

course, larger-scale models. As Birkin and Malleson point out in this chapter, while modelling and simulation in the social sciences have been around for fifty years, prompted by an earlier wave of innovations in computation, recent advances in both data and computation are now having a profound effect.

This chapter provides an introduction to the state of the art in four model classes that are of particular interest to social scientists – systems dynamic models, statistical and behavioural models, microsimulation models and agent-based models. Examples are presented of each of these classes – a retail or residential location model (spatial interaction model or mathematical/systems dynamic model); a traffic behaviour model (discrete choice or statistical model); a demographic model (microsimulation model); and a crime model (agent-based model). Birkin and Malleson observe that while building ever more sophisticated models of social systems has never been easier, the task of demonstrating that such models faithfully represent an underlying social reality remains the key challenge. They then relate some experiences and lessons from building a prototype social simulation infrastructure capable of providing support for the whole research lifecycle, and they stress, in particular, the importance of model reproducibility, reusability and generalizability. They conclude with a summary of some of the – as yet – unexploited opportunities for social simulation presented by new sources of data (e.g., using mobile phone data to update in real time models of population movements) and the challenges (e.g., data ownership and ethics) that will have to be met if these are to be realized.

Chapter 7: Contemporary Developments in Statistical Software for Social Scientists

In this chapter, Lambert, Browne and Michaelides examine the prospects of the quantitative social sciences being in a position to exploit the power of new social data, computational resources and tools to achieve advances in statistical analysis. They review the range of statistical software packages currently available to social researchers and the factors influencing their patterns of adoption. They illustrate their review with examples of the application of statistical methods in domains such as education, health inequalities and epidemiology. They argue that the profusion of statistical tools, while having the benefit of offering choice to researchers, nevertheless raises significant barriers, both social and technical (and, indeed, socio-technical), that need to be addressed if the power of the tools is to be fully exploited by the social science research community.

Regarding social barriers, the authors note that in the UK there is a lack of capacity in statistical skills within the social research community. Regarding technical barriers, they observe that the proliferation of statistical tools has been at the cost of inter-operability and has created a situation that they describe as 'balkanization'. This can deter researchers from using the tool most appropriate for a particular analysis – rather than the one they are most familiar with – and may also inhibit experimenting with new tools. Echoing the concerns raised by Purdam and Elliot, they also point to problems with transparency, replicability and robustness of statistical

analyses using computer packages whose algorithms are not accessible to the user. Drawing on the principles of e-Research for their inspiration, Lambert et al. conclude by presenting some ways of overcoming the social and technical barriers, which they exemplify through their efforts to develop Stat-JR and eBooks, new tools for statistical analysis that promote inter-operability between analysis packages and sharing through better documentation of analysis routines.

Chapter 8: Text Mining and Social Media: When Quantitative Meets Qualitative and Software Meets People

Text mining has developed dramatically in recent years in its power to analyse and extract information from very large bodies of unstructured text. Its applications are motivated by a growing awareness that researchers need more powerful tools in order to benefit from rapidly increasing amounts of textual data being generated through the proliferation and unprecedented levels of take up of Web 2.0 technologies. Chief among these are blogs and social media ('micro-blogs'), the latter exemplified by the rise of platforms such as Facebook and Twitter.

In this chapter, Ampofo, Collister, O'Loughlin and Chadwick explore how text mining using natural language processing (NLP) techniques can provide qualitative social researchers with powerful analytical tools for extracting information from this unstructured data, including harvesting data and analysing it in real time. They survey the range of research tools for text mining, broadly defined, available both in the academic and commercial spheres. People's use of social media is seen by many researchers as providing an ideal source of data through which to monitor rapidly changing situations, hence, it has come to particular prominence during civil unrest (e.g., the so-called 'Arab Spring') and natural disasters (e.g., Hurricane Sandy). Beyond these inherently unpredictable phenomena, one of the most popular emerging applications of social media analysis lies in the tracking of public opinion through the application of NLP-based techniques such as sentiment analysis. These techniques have the capacity to generate results in real time, which offers intriguing possibilities for both commercial and academic research.

To illustrate the potential and challenges of using text mining techniques in social research, Ampofo, Collister, O'Loughlin and Chadwick present overviews of two projects. The first is a study of social media during the televised debates between political party leaders in the 2010 UK general election campaign. The second is also drawn from this election campaign and focuses on the reporting of accusations of bullying against then-Prime Minister Gordon Brown in the British media. The application of NLP-based text analysis tools to social data is still, in many respects, in its infancy. With this thought in mind, the authors conclude by outlining the ontological challenges (echoing the reservations that Elliot and Purdam set out in Chapter 3) and the technical challenges of mining text in social research settings. They note, in the case of social media, increasingly restrictive access policies, and they also consider the ethical implications of text mining used as a social research tool.

Chapter 9: Digital Records and the Digital Replay System

As many of the contributors to this book recognize, the capacity to capture behaviour through the 'digital footprint' that people generate as a by-product of their everyday activities has the potential to transform the practice of empirical social science. In this chapter, Crabtree, Tennent, Brundell and Knight examine how new tools for data collection and analysis make it possible to exploit this data. Their discussion focuses in particular on the development of 'digital records' that enable social science researchers to combine novel and heterogeneous forms of digital data, such as video, text message logs and GPS data, with more traditional and established forms, such as audio recordings and transcriptions of talk.

The authors describe the Digital Replay System (DRS), an open source, extensible suite of interoperable tools for assembling, synchronizing, visualizing, curating and analysing digital records.[3] In Chapter 5, Lambert presents solutions to the data management problems attendant in the use of conventional kinds of social data such as surveys. From this perspective, DRS can be viewed as a prototype for meeting the data management and linking challenges presented by novel sources of social data. Crabtree and his co-authors provide a step-by-step exposition of several different examples; these include capturing rich accounts of people's physiological reactions while on a fairground ride, a corpus linguistics perspective on visitors' interactions in an art gallery, and disaster mapping and management. Collectively, these examples illustrate how the use of a system like DRS can enable the assembly of digital records capturing a wide range of interactions between people that are a by-product of their use of various digital devices, and make them available for subsequent visualization, curation and analysis. Finally, the authors consider future developments, particularly the prospects for making use of mass participation in social science research through the use of mobile devices for the crowd-sourcing of data.

Chapter 10: Social Network Analysis

The distinctive contribution of social network analysis (SNA) to social research is its stress on the importance of studying the structure of relationships between people rather than considering them as unconnected individuals. Like many of the other advances in research methods covered in this book, SNA is a mature methodological tool. Arguably, it owes its rise to greater prominence in recent years to two factors. One is that, as with many other established social research methodologies, e-Infrastructure has extended the scale and complexity of what is achievable, in this case by providing SNA with new and more powerful means to capture social network datasets, analyse them and visualize the results. The second factor is that many of the new types and sources of digital social data – such as hyperlink networks (the structures of links between websites) and social networking sites such as Facebook and Twitter – are inherently relational.

[3]http://thedrs.sourceforge.net

In this chapter, Ackland and Zhu review the history and methodological principles of SNA, and survey several of the research tools now available for SNA data collection, analysis and visualization. They draw on examples of studies of Facebook, Twitter, Flickr, online newsgroups and websites to illustrate contemporary and arguably the most prominent uses of SNA – to study people's behaviour in social networking sites. Ackland and Zhu go on to discuss two key ontological questions associated with SNA as a research methodology. The first is its 'construct validity', an issue that has potentially major implications. Simply put, the question is: do the social structures observed in, for example, Facebook, have real-world analogies or are they properties only of the online world, entirely unrelated to its real world counterpart? If the answer is no, then arguably, for all the talk about the opportunities for social research offered by new sources of social data, the impact in terms of increased understanding of social phenomena will be very limited.

Ackland and Zhu's second question relates to debates about the capacity of social research methodologies to distinguish between causality and correlation. Here, they offer a somewhat more optimistic prognosis, observing that data generated through people's activity on, for example, social networking sites, is rich and time-stamped, allowing for more fine-grained analysis, while the sites themselves can be thought of as natural research instruments, ideal for carrying out large scale experiments.[4] Like other contributors to this volume, they conclude with a warning about the pitfalls for researchers of relying on data sources, such as Facebook, that are proprietary and whose access is subject to terms and conditions that may change at any time.

Chapter 11: Visualizing Spatial Data and Social Media

As earlier chapters have emphasized, the social data landscape is changing at an ever-increasing pace. The ways in which data is visualized has always played an important role in its analysis and in the presentation of results, and the ever-increasing volumes of data raise new challenges for visualization methods and tools. In this chapter, following a brief history of geographic information systems (GIS), Batty and his colleagues describe new ways of visualizing social data, with a particular emphasis on mapping. They argue that Web 2.0 mash-ups, layering geographically tagged social data on top of digital maps, enable quick and simple visualization of data, presenting research outcomes in ways that can be easily understood by diverse audiences.

Many of the examples the authors present emphasize how much researchers can achieve using simple, generic technologies and services such as Google Maps and Fusion Tables. Helpfully, Batty and his colleagues at UCL's Centre for Advanced Spatial Analysis (CASA) have packaged these services into useful tools (such as

[4]As the recent controversy over the Facebook experiment conducted by researchers at Cornell and the University of California, it is essential to think very carefully about the ethical implications of conducting such studies. See www.theguardian.com/technology/2014/jul/02/facebook-apologizes-psychological-experiments-on-users

MapTube[5]), which not only enable the geo-mapping of datasets with a few button clicks, but also provide ways for researchers to share and re-use each other's efforts.

Another way in which advances in visualization techniques have harnessed the increase in computer power and new sources of data is the creation of fly-through, 3D models and visualizations of, for example, urban environments. More mundanely, but perhaps of greater value to researchers and planners involved in urban science, and the latest of many research areas predicted to be transformed by the advent of big data,[6] are CASA's 'city dashboards', which integrate diverse sources of data to create a real-time visualization of the state of the city and its inhabitants. Example applications include visualizing in real-time the state of mass transit systems. Such tools can provide powerful and intuitive front-ends to the simulations and models presented in Chapter 6, allowing, for example, exploration of the impact of closure of parts of the system.

Batty and his co-authors stress the importance of crowdsourcing and 'citizen science' for creating resources accessible to the public and illustrate this with the example of Open Street Map, a free map of the world.[7] They conclude with some thoughts on the future of visualization as a tool for social scientific investigation and understanding. They predict the emergence of radically different kinds of tools that make use of more abstract forms of visualization, with an increasing emphasis on the use of non-spatial data as the way forward for understanding how social systems function.

Chapter 12: Ethical Praxis in Digital Social Research

Current approaches to ethics no longer seem adequate for twenty-first century social research. We have already noted the concern registered by the authors of preceding chapters about the privacy and confidentiality threats raised by the proliferation of social data. There is an emerging consensus that a new ethical framework for the conduct of social research is necessary in order to protect citizens from harm but, as yet, there is little agreement on what changes it should embody, and how it should be promulgated and enforced.

In this chapter, Jirotka and Anderson examine the ethical issues raised by e-Research methods and what steps the social research community might take to address them. They use three case studies to illustrate the issues. The first describes a flagship UK e-Science project eDiaMoND and the process of gaining ethical approval for its work. The second concerns a recent controversy regarding social science researchers' use of Facebook data called the 'Harvard Meltdown'. The final case study is about developing prototype assistive technology for vulnerable people. Jirotka and Anderson draw several conclusions from these studies: managing ethics in large scale, multi-disciplinary research projects is particularly difficult and some of the founding

[5]www.maptube.org

[6]For example, the Center for Urban Science and Progress (CUSP). See cusp.nyu.edu

[7]www.openstreetmap.org

principles of research ethics, such as informed consent, can be burdensome; protecting the identity of sources using conventional techniques for anonymization is becoming progressively less reliable as more and more information about subjects and settings becomes openly available via the Web (identification is always possible given enough correlated data); consenting to take part in research must be done in a principled way and, having consented, participants must have the power in practice – and not just in principle – to withdraw it; and finally, where a project involves interventions in people's lives, researchers must consider what may happen once the project finishes.

They conclude with a discussion of the ethics of big social data. They underline the importance of the well-rehearsed arguments about threats to privacy and confidentiality. They ask what rules should apply to the use of social media in research: does publishing thoughts and opinions in public render informed consent irrelevant? However, their key insight goes further: it questions whether the lure of big social data is persuading researchers to relax their professional judgment about what conclusions are warrantable from the data. Jirotka and Anderson's fundamental argument is that we need to bring ethical considerations into the heart of how we conduct research, from the point where decisions are being made about research goals, through to the collection and analysis of the data and the making sense of the findings.

Chapter 13: Sociology and the Digital Challenge

This final chapter examines the implications of massively increased computational and data resources for social research methods, including the impact on its established practices and future of its disciplines. In it, Savage returns to themes that he and his co-author, Burrows, first raised in their subsequently much-cited paper, 'On the coming crisis of empirical sociology' (Savage and Burrows, 2007). His aim, in part, is to ground expectations of the changes in social research that may follow from digital innovations and, not least, to question their inevitability. As the contributions of the authors of the chapters in this volume convincingly demonstrate, the future of digital sociology is contested: they all agree that the discipline is undergoing a sustained period of innovation, but its future direction is unknown. Together, they make a powerful case for Savage's assertion that the future of digital sociology is not a given, but lies in the hands of current and subsequent generations of practitioners.

1.4 FUTURE DIRECTIONS

1.4.1 Technical Developments

The other chapters in this book, described above, confirm that e-Research has moved on from an early focus on grid computing to encompass a very diverse set of tools, some of which are enhancements of previous software and others that are entirely new. A factor that suggests that this diversity will persist and even grow is the lack of central co-ordination and oversight. In the UK, the national e-Science Centre, which was the hub for the core programme, ceased operating in 2011, as did the

NCeSS Hub in 2010. Other national centres still exist, for example the New Zealand eScience Infrastructure (www.nesi.org.nz), as do several international initiatives, such as the Open Grid Forum (www.ogf.org) and the European Grid Infrastructure (www.egi.eu). The emphases of these centres and programmes, however, are largely high performance computing, providing cloud services and codifying grid standards; areas of limited relevance to the social sciences. Outside these programmes, technical developments are either mostly modest refinements to existing tools, updates to commercial packages driven by competition for market share, or the adoption and adaptation of whatever generic or specialized tools and services researchers find can smooth the path of their own research. The future path of technical developments is therefore impossible to predict, though the drive to harness computing power to enable better research is unlikely to abate.

1.4.2 The Data Deluge

As reiterated in most of the chapters in this volume, we live in an information age characterized by a deluge of digital data (Hey and Trefethen, 2004; Hey, Tansley and Tolle, 2009). The chapters set out many of the potential research benefits to be obtained by collecting and analysing artificially produced and naturally occurring big data of many kinds from numerous sources. However, these benefits will only be realized if the wealth of data is managed in ways that ensure that it is discoverable, accessible, usable and re-usable. Indeed, research data management was a cornerstone of the original e-Research vision.

Accordingly, national e-Research programmes to innovate research methods, tools and infrastructure have devoted significant efforts to raise awareness among stakeholders that research data is a vital resource whose value needs to be preserved for future research by the data originators and by others. Achieving this requires that the data be systematically organized, securely stored, fully described, easily locatable, accessible on appropriate authority, shareable, archived and curated. Fulfilling all of these research data management tasks is a complex socio-technical challenge that stakeholders, whether they are research funders, higher education institutions (HEIs), publishers, researchers or regulators, are currently ill prepared to meet (Procter, Halfpenny and Voss, 2012). There are, as yet, no widely-agreed, mature solutions that can be implemented across all the various platforms that researchers use. Moreover, given the combination of the data deluge and a world recession, the scale of the tasks is increasing while the financial and therefore human resources to undertake the tasks are shrinking.

Ensuring the implementation and sustainability of data preservation will need to take on board the prospect of research becoming more collaborative and research teams being more widely distributed, as signalled in the e-Research vision of researchers world-wide addressing key challenges in new ways. The implications for data management services are summarized in a report from the Department for Business, Innovation and Skills (BIS) in the UK, which concluded, 'A federated infrastructure will be essential to exploit existing and future investments [in data] effectively' (Business, Innovation and Skills, 2010, 9). If such a federated infrastructure is to be

achievable, then establishing effective inter-institutional service models will take on increasing importance. HEIs and other research organizations will need to develop strategies and infrastructure solutions that enable the federation of individual data repositories and the virtualization of data services. This will add a further layer of sustainability issues, the opportunities, costs and benefits of such collaborations will need to be carefully examined, and HEIs (both large and small) will need to develop competencies in managing services that span administrative and funding boundaries. In the current competitive environment, with universities locked in a zero-sum struggle for resources, there is little incentive to put effort into the inter-institutional cooperation required.

The term *big* social data serves to draw attention to three salient dimensions that define new forms of social data: volume, variety and velocity, the last reflecting its often real-time and rapidly changing character. Developments linked to the emergence of big social data are happening continually and we cannot be certain what impact such data will have on research processes. It is possible that it will promote the use of new computational social science methods in place of more traditional quantitative and qualitative research methods. It might also influence thinking and re-orientate social research around new objects, populations and techniques; network analysis offers an example here. The analysis of social processes as they actually happen is bound to give researchers insights and interesting avenues to explore that are absent from the often post-hoc reconstructions of events that are available via traditional research instruments and datasets.

Big social data will inevitably force us to rethink the role of academic social scientists. One way forward would be for them to actively seek collaborations with groups, both professional and lay, involved in doing various kinds of 'practical, everyday sociology'. An example of collaboration with professionals might include assisting journalists[8] who increasingly find themselves needing to analyse large datasets in order to report news stories.[9] Examples of collaborating with lay people include 'citizen social science' where members of the public can assist with research through crowd-sourcing data (as illustrated in Chapter 9), by participating in analytical work (Procter et al., 2013b), and even by taking a role in the setting of research agendas (Housley et al., 2014). These examples suggest possibilities for forging a new relationship between academic social science and society at large, a 'public sociology' (Burawoy, 2005), where social scientific knowledge is co-produced by a wide range of stakeholders (Housley et al., 2014) and is subject to greater public oversight and accountability. Initiatives in other discipline areas might provide models for how to proceed in the social sciences: see, for example, the Public Laboratory for Open Technology and Science (http://publiclab.org/), whose 'goal is to increase the ability of

[8]See, for example, the 'reading the riots' project, Lewis et al. (2011).

[9]This has given rise to the new specialism of 'data journalism'. News media organizations have also been at the forefront of experiments in citizen journalism and crowdsourcing data analysis. For an example of the latter, see www.theguardian.com/news/datablog/2009/jun/18/mps-expenses-houseofcommons

underserved communities to identify, redress, remediate, and create awareness and accountability around environmental concerns.'

Finally, as is noted in several of the chapters that follow, big social data has given fresh stimulus to debates about research ethics (see e.g., boyd and Crawford, 2012), much of which focuses on the issue of people's right to privacy but which also raises questions about the role and status of academic research. At the same time, we must not lose sight of the broader issue of the ethics of research and innovation (see e.g. Stahl, Eden and Jirotka, 2012, and Chapter 12 in this volume).

1.4.3 Collaboration

e-Research was conceived from the very beginning as a collaborative activity that would combine the abilities of distributed and complementary groups of researchers in order to achieve research goals that individual researchers or local groups could not hope to accomplish. With this in mind, the concept of the 'virtual research environment' (VRE), 'collaboratory' (cf. Olson, Zimmerman and Bos, 2008) or 'gateway' was another widely promoted element of the e-Research vision. VREs were seen as a way to support collaboration and provide integrated, shared access to resources throughout the research lifecycle, starting with literature searches and ending with the publication of results and curated datasets. In one system, accessible by all team members, a shared bibliography would be assembled. A joint laboratory notebook would be kept which would document all the research procedures undertaken. Data would be stored along with metadata recording the operations it had been subject to, and reports would be written collaboratively, with all versions archived, and publications prepared. Once again, experience has shown that the initial vision had to be tempered. VREs exemplify what happens when 'top-down' innovation programmes meet 'bottom up' processes through which individuals and groups of researchers experiment with whatever new technologies are at hand. They often prefer to work out their own – often ad-hoc, bespoke but nevertheless effective – solutions that match their needs and level of technical competence rather better than complex, all-embracing offerings whose adoption might lead to having to abandon favoured tools. A prosaic example is the use of an email list and attachments or freeware such as Dropbox[10] to share documents, rather than struggle to implement a VRE across different institutions' computer systems and seek local support in its use. Similarly, Web 2.0 has provided a host of applications that can be easily adopted to support various stages of the research cycle, such as switching from email attachments to an Internet file hosting and synchronizing service like Dropbox or Google Drive. Those VREs that have survived the turbulence of constant technological innovation and rapidly changing standards tend to be associated with 'big science' projects, such as climate change, and benefit from long-term funding arrangements.[11]

[10]www.dropbox.com

[11]See, for example, the Extreme Science and Engineering Discovery Environment (XSEDE) www.xsede.org/web/guest/gateways-listing

1.4.4 Scholarly Communications

Nowhere is this tension between top-down and bottom-up innovation processes in science more clearly evident than in scholarly communications. The past decade has seen the emergence of new ideas about the practice of scholarly communications, with talk of a 'crisis in publishing' and weaknesses in the peer-review system. One outcome is the notion of 'Open Science' (Neylon and Wu, 2009) with its advocacy of more open scientific knowledge production and publishing processes (Berlin Declaration, 2003; Murray-Rust, 2008). This has been inspired by discourses developed in 'Free/Open Source Software' and 'Creative Commons' movements (Lessig, 2004; Benkler and Nissenbaum, 2006; Elliott and Scacchi, 2008). Web 2.0 is widely seen as providing the technical platform to enable these new forms of scholarly communications and bring about a 're-evolution' of science (Waldrop, 2008).

Web 2.0 brings the promise of enabling researchers to create, annotate, review, reuse and represent information in new ways, promoting innovations in scholarly communication practices – e.g. publishing 'work in progress' and openly sharing research resources – that will help realize the e-Research vision of improved productivity and reduced 'time to discovery' (Arms and Larsen 2007; Hey et al., 2009; Hannay, 2009; De Roure et al., 2010). However, despite this increasing interest in Web 2.0 as a platform and enabler for e-Research, understanding of the factors influencing adoption, how it is being used, and its implications for research practices and policy remains limited. Recent studies suggest that there is considerable reluctance – even suspicion – to adopt new forms of scholarly communications among many academics, who fear that this will mean the end of the 'gold standard' of peer-review and the undermining public trust in science (Procter et al., 2010a; Procter et al., 2010b). Equally, it would be a mistake to ignore the capacity of established academic publishers to shape the emerging scholarly communications landscape so as to preserve their role as gatekeepers (Stewart et al., 2012). The future of scholarly communications may, after all, not be so radically different from its recent past.

1.4.5 The Future

The vision that motivated the e-Science programme in the UK and analogous programmes elsewhere was that grid computing-based infrastructure comprising computer power, big data and collaborative teams would transform science. Over the past decade this has morphed into a much more complex e-Infrastructure made up of a plethora of only loosely related tools and services taken up to different degrees and in different combinations and with different levels of enthusiasm even within the same field, allied with rapidly accreting digital data of new types and old. The e-Research facilitated by this maelstrom is transforming social science research, but in unpredictable ways, with many socio-technical barriers to be overcome before its full potential is realized. The aim of this book is to whet the appetite of social researchers to encourage them to explore how innovations in digital research methods might enable their research to advance in ways not possible otherwise.

1.5 ONLINE RESOURCES

Many of the examples of e-Research methods presented in this book already have online resources associated with them. To make these more accessible to readers, we have created a companion website.[12] This provides easy access to this content, including in-depth case studies, datasets, research workflows, tools and services, publications and links to the authors' own websites.

1.6 BIBLIOGRAPHY

Ananiadou, S., Weissenbacher, D., Rea, B., Pieri, E., Vis, F., Lin, Y-W., Procter, R. and Halfpenny, P. (2009a) 'Supporting frame analysis using text mining', *Proceedings of 5th International Conference on e-Social Science, Cologne, June*. Available from http://wrap.warwick.ac.uk/52916 (accessed 12 Dec 2014).

Ananiadou, S., Okazaki, N., Procter, R., Rea, B. and Thomas, J. (2009b) 'Supporting systematic reviews using text mining', in P. Halfpenny and R. Procter (eds) Special Issue on e-Social Science, *Social Science Computing Review Journal*, 27(4): 509–23.

Arms, W.Y. and Larsen, R.L. (2007) *The Future of Scholarly Communication: Building the Infrastructure for Cyberscholarship*. Report of a workshop held in Phoenix, Arizona April 17–19. Sponsored by the National Science Foundation (NSF) and the Joint Information Systems Committee (JISC). DOI: http://dx.doi.org/10.3998/3336451.0011.102.

Benkler, Y. and Nissenbaum, H. (2006) 'Commons-based peer production and virtue', *The Journal of Political Philosophy*, 14(4): 394–419.

Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (2003) Conference on Open Access to Knowledge in the Sciences and Humanities. Berlin, October. Available from http://openaccess.mpg.de/Berlin-Declaration (accessed 12 Dec 2014).

Birkin, M., Procter, R., Allan, R., Bechhofer, S., Buchan, I., Goble, C., Hudson-Smith, A., Lambert, P., DeRoure, D. and Sinnott, R. (2010) 'Elements of a computational infrastructure for social simulation', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1925): 3797–3812.

boyd, D. and Crawford, K. (2012) 'Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon', *Information, Communication & Society*, 15(5): 662–79.

Burawoy, M. (2005) 'For public sociology', *American Sociological Review*, 70: 4–28.

Crabtree, A., French, A., Greenhalgh, C., Benford, S., Cheverst, K., Fitton, D., Rouncefield, M. and Graham, C. (2006) 'Developing digital records: early experiences of record and replay', *Journal of Computer Supported Cooperative Work,* 15(4): 281–319.

Das, S., Sismanis, Y., Beyer, K.S., Gemulla, R., Haas, P.J. and McPherson, J. (2010) 'Ricardo: integrating R and Hadoop', in *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, New York: ACM. pp. 987–98.

Department for Business, Innovation and Skills (2010) *Delivering the UK's e-Infrastructure for Research*. Available from www.rcuk.ac.uk/RCUK-prod/assets/documents/research/esci/e-Infrastructurereviewreport.pdf (accessed 6 Dec 2014).

[12]https://study.sagepub.com/halfpennyprocter

De Roure, D., Goble, C., Aleksejevs, S., Bechhofer, S., Bhagat, J., Cruickshank, D., Procter, R. and Poschen, M. (2010) 'Towards open science: the myExperiment approach', *Concurrency and Computation: Practice and Experience*, 22(17): 2335–53.

Duncan, G, Elliot, M J. and Salazar, J.J. (2011) *Statistical Confidentiality: Principles and Practice.* New York: Springer.

Edwards, P. Jackson, S., Bowker, G. and Knobel, C. (2007) *Understanding Infrastructures: Dynamics, Tensions, and Design*, final report of the workshop History and Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures, National Science Foundation. Available from http://deepblue.lib.umich.edu/handle/2027.42/49353 (accessed 15 Dec 2014).

Edwards, P., Mayernik, M.S., Batcheller, A., Bowker, G. and Borgman, C. (2011) 'Science friction: data, metadata, and collaboration', *Social Studies of Science*, 41(5): 667–90.

Elliott, M. and Scacchi, W. (2008) 'Mobilization of software developers: the free software movement', *Technology and People*, 21(1): 4–33. Available from www.ics.uci.edu/~wscacchi/Papers/New/Elliott-Scacchi-Free-Software-Movement.pdf (assessed 09 April 2015).

Fraser, M., Hindmarsh, J., Best, K., Heath, C., Biegel, G., Greenhalgh, C. and Reeves, S. (2006) 'Remote collaboration over video data: towards real-time e-social science', *Journal of Computer Supported Cooperative Work*, 15(4): 257–79.

Halfpenny, P., Procter, R., Lin, Y. and Voss, A. (2009). 'Developing the UK e-Social Science Research Programme'. In Jankowski, N. (ed.) *e-Research, Transformation in Scholarly Practice*, Abingdon: Routledge.

Halfpenny, P. and Procter, R. (2010) 'The e-Social Science research agenda', *Philosophical Transactions of the Royal Society A*, special issue on e-Science, 368: 3761–3778, August.

Hannay, T. (2009) 'From Web 2.0 to the global database'. In Hey, T., Tansley, S. and Tolle, K. (eds) *The Fourth Paradigm: Data-Intensive Scientific Research*. Redmond, WA: Microsoft Research. pp. 215–20.

Hey, T. and Trefethen, A. (2004) 'UK e-Science programme: next generation grid applications'. *International Journal of High Performance Computing Applications,* 18(3): 285–91.

Hey, T., Tansley, S. and Tolle, K. (eds) (2009) *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, WA: Microsoft Research.

Housley, W., Procter, R., Edwards, A., Burnap, P., Williams, M., Sloan, L., Rana, O., Morgan, J., Voss, A. and Greenhill, A. (2014) 'Big and broad social data and the sociological imagination: a collaborative response', *Big Data & Society*, *1*(2): pp. 1–15.

Hudson-Smith, A., Batty, M., Crooks, A. and Milton, R. (2009) 'Mapping for the masses: accessing Web 2.0 through crowdsourcing', in P. Halfpenny and R. Procter (eds) Special Issue on e-Social Science, *Social Science Computing Review*, 27(4): 524–38.

IDC (2014) *The Digital Universe of Opportunities*. Executive Summary available at www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm.

Lessig, L. (2004) *Free Culture: How Big Media Uses Technology and the Law to Lock Down Culture and Control Creativity*. New York: Penguin Press.

Lewis, P., Newburn, T., Taylor, M., Mcgillivray, C., Greenhill, A., Frayman, H. and Procter, R. (2011) *Reading the Riots: Investigating England's Summer of Disorder*. Guardian Newspapers/LSE. Available from http://eprints.lse.ac.uk/46297/1/Reading%20the%20riots(published).pdf (accessed 09 April 2015).

Murray-Rust, P. (2008) 'Chemistry for everyone', *Nature*, 451: 648–51.

O'Reilly, T. (2005) 'What is Web 2.0?' September. Available from www.oreilly.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html (accessed 01 April 2005).

Olson, G.M., Zimmerman, A. and Bos, N. (2008) *Scientific Collaboration on the Internet*. Cambridge, MA: MIT Press.

Neylon, C. and Wu, S. (2009) 'Open Science: tools, approaches, and implications', *Pacific Symposium on Biocomputing*, 14: 540–4. Available from http://psb.stanford.edu/psb-online/proceedings/psb09/abstracts/2009_p540.html (accessed 09 April 2015).

Procter, R., Williams, R., Stewart, J., Poschen, M., Snee, H., Voss, A. and Asgari-Targhi, M. (2010a) 'Adoption and use of Web 2.0 in scholarly communications', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *368*(1926): 4039–4056.

Procter, R., Williams, R. and Stewart, J. (2010b) *If You Build It, Will They Come?: How Researchers Perceive and Use Web 2.0*. Research Information Network. Available at http://wrap.warwick.ac.uk/56246/1/WRAP_Procter_If%20you%20build%20it%20will%20they%20come.pdf (accessed 28 Jan 2015).

Procter, R.N., Halfpenny, P. and Voss, A. (2012b) 'Research data management: opportunities and challenges for HEIs', in G. Pryor (ed.) *Research Data Management*. London: Facet Publishing. pp.135–50.

Procter, R., Voss, A. and Asgari-Targhi, M. (2013a) 'Fostering the human infrastructure of e-research', *Information, Communication & Society*, 16(10): 1668–91.

Procter, R., Housley, W., Williams, M., Edwards, A., Burnap, P., Morgan, J., Voss, A. and Greenhill, A. (2013b) 'Enabling social media research through citizen social science'. *ECSCW 2013 Adjunct Proceedings*, 3.

Research Councils UK (2014) *e-Infrastructure*. Available at www.rcuk.ac.uk/research/xrcprogrammes/otherprogs/einfrastructure (accessed 28 Jan 2015).

Savage, M. and Burrows, R. (2007) 'The coming crisis of empirical sociology', *Sociology,* 41(5): 885–99.

Stahl, B., Eden, G. and Jirotka, M. (2012) 'Responsible research and innovation in Information and Communication Technology: identifying and engaging with the ethical implications of ICTs', in R. Owen, J. Bessant and M. Heintz (eds), *Responsible Innovation.* Chichester: Wiley & Sons. pp.199–218.

Stewart, J., Procter, R., Williams, R. and Poschen, M. (2013) 'The role of academic publishers in shaping the development of Web 2.0 services for scholarly communication', *New Media & Society*, 15(3): 413–32.

Thelwall, M. (2009) 'Introduction to webometrics: quantitative web research for the social sciences', *Synthesis Lectures on Information Concepts, Retrieval, and Services* 1(1): 1–116.

Voss, A., Asgari-Targhi, M., Procter, R. and Fergusson, D. (2010) 'Adoption of e-Infrastructure services: configurations of practice', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *368*(1926): 4161–76.

Waldrop, M. (2008) 'Science 2.0: great new tool, or great risk?' *Scientific American*. Available from www.sciam.com/article.cfm?id=science-2-point-0-great-new-tool-or-great-risk (accessed 12 Dec 2014).

Webb, T., Joseph, J., Yardley, L. and Michie, S. (2010) 'Using the internet to promote health behavior change: a systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy'. *Journal of Medical Internet Research*, *12*(1), e4.