MASTERS THESIS, 2011
OXFORD INTERNET INSTITUTE

ZARINO ZAPPIA

Participation, Power, Provenance: Mapping Information Flows in Open Data Development

MSc Social Science of the Internet

Participation, Power, Provenance: Mapping Information Flows in Open Data Development

Zarino Zappia

Oxford Internet Institute

August 2011

Thesis submitted in partial fulfilment of the requirement for the degree of MSc in Social Science of the Internet at the Oxford Internet Institute at the University of Oxford

9968 words excluding front matter and bibliography

Abstract

This thesis represents one of the first empirical studies of Open Data development in the UK and USA. Through content analysis of web apps and interviews with key players in the Open Data field, it analyses the range of uses to which Open Data is put, the motivations of both suppliers and developers of Open Data, and finally—most importantly—the role powerful intermediaries are having on these information flows.

It emerges that, although intermediaries are present in other media such as online news provision and web search, they are yet to take a hold in the Open Data sphere. The effect is twofold: it leaves Open Data development free and distributed, but facing tough issues of organisation and discoverability. Most Open Data developers work alone and surprisingly few combine datasets from different sources. The Open Data 'community' is in fact more a constellation of independent, task-focussed communities, with little overlap.

Thus, if power brokers are to appear, they will be sited between these communities: either as data sources with information suitable for multiple contexts, or as data developers with contacts and skills in a number of areas. In these key positions, such intermediaries would hold a good deal of power over which data are released and how they are used. While such figures are by no means a threat to Open Data development, this thesis lays important groundwork for more thorough future research into how intermediaries' power can be best applied to ensure the future of innovative Open Data participation.

For Kovinka.

With thanks to Viktor, my generous
interviewees, and the 66 crew.

Contents

List of Tables & Figures

1 Introduction

> "Government should be transparent.
> Government should be participatory.
> Government should be collaborative."
> *(President Obama, 21 January 2009)*

Over the last decade, the notion of 'Open Data' has been presented as a near-panacea for social collaboration and political accountability. In government especially, the online publishing of new and existing datasets has been hailed as a groundbreaking move for democracy, empowering citizens and making government processes more efficient and transparent (Shadbolt, 2011*a*; Orszag, 2009).

In reality, Open Data is only the latest in a line of public accountability systems, including Freedom of Information (FOI) legislation and Environmental Information Regulations (EIR). Much of this legislation required that governments—and in many cases private bodies—move beyond a purely reactionary role, toward providing information *ex ante* in a publicly accessible format. Meanwhile, legislation from the UK Advisory Panel on Public Sector Information and the European Public Sector Information Directive, to name but a few, cemented the place of data as a public good, whether it be from governments, public authorities, or international organisations like the World Bank and the OECD. In a time of financial and political uncertainty, free access to public data was seen as a means

to foster both trust in the government and renewed exploration of innovative, data-centric business models.

The recent announcement, however, of drastic cuts to the US Open Data budget and the proposed closure of its main public-facing portals, Data.gov and USAspending.gov (Yau, 2011), has caused division over how best to proceed with Open Data initiatives. Only now do we realise the true dearth of empirical research on exactly *how* Open Data is currently being used and on how *open* to collaboration and recombination it really is.

It is with this goal in mind—the conceptual, and later physical, mapping of the nascent Open Data landscape—that I embarked on an empirical study of Open Data development in the UK and North America. The Open Data field may be new, but that certainly does not preclude detailed study of its progress so far. Indeed, to gain a true understanding of the processes involved in today's Open Data movement, one must first look *back* to the much larger, much older Open Source movement, which itself has roots in the fundamentally libertarian beginnings of the Internet. Common to both Open Source and Open Data movements are the central tenets of shared knowledge and distributed information. Both were made possible by the Internet as a collaborative medium, and both inherited from that medium similar sets of practices and norms. To simply talk about Open Data is to ignore the bigger picture; one which has recurred again and again throughout the history of the Internet.

Likewise, dealing only with Open Government Data is to miss out a whole raft of alternate uses for Open Data development. Open Data (much like the Internet itself) originated as a technology for collaborative science research (CGED, 1995) – a purpose to which it is still put today. Its future, meanwhile, lies in the corporate

sphere (Shadbolt, 2011*b*) where shared information has the potential to both reduce barriers to entry for newcomers and also increase efficiency for existing players. Independent media are already finding their place as prospectors and aggregators of public and corporate data; be they traditional print- and screen-media corporations or networked members of the 'fifth estate' (Dutton, 2007; Sunstein, 2007). While there are important social and political debates to be had over the role of Open Government Data in particular, this thesis views Open Data as a wider technological movement towards openness, transparency and innovation, under which government[1], public service, and corporate data all fall.

The real question though, as Open Data begins to be used by more actors outside of government, is one of who holds the power in this growing information network. Theorists have already shown that other media, including the press, web search and news/political debate sites have become dominated by powerful 'intermediaries' – central participants with the power to dramatically shape information flows across the network (Hindman, 2009). In many cases this 'gatekeeping' has beneficial effects, such as bringing users more relevant content (Halavais, 2009), avoiding 'information overload' (Palfrey and Gasser, 2008), or providing a common space for collaboration and discussion (Barber, 1984). Critics, however, have highlighted the inherent tension between such centralising tendencies and the distributed, 'free' nature of the Internet (Sunstein, 2007). If intermediaries *are* emerging in the Open Data sphere, we should at least be aware of their existence, and their growing power over the data we can find and use.

[1]It is worth noting that, throughout this thesis, 'government' will be taken to mean not only the workings of congress, parliament and the state, but also more general public service institutions such as the National Health Service, the police force and local councils. This distinction will be of some importance when 'government' is used as a code in my content analysis.

Questions of power are closely tied to questions of trust. The (re-)aggregation of datasets through central portals and mashups, means developers need also be aware of the provenance of the data they find. Even with government data, the originating body is not always made clear, meaning developers cannot be sure about the motivations of the organisation or sub-contractor who gathered, compiled and released the data. The problem is further aggravated for the end-user when visualisations and websites often lack attribution to their creators, let alone the data sources that were used in their creation. If we are living, as Hal Varian points out, in a time of "data obesity" (Finn, 2011), users must not only be empowered with the data itself, but also with the ability to sift out the most trusted, authoritative data and treat the rest with a modicum of caution.

This thesis, as a record of my research, first outlines the theoretical context through which we can come to understand the processes and motivations of Open Data participants, before documenting and discussing my data collection methods and findings, and finally concluding with clear evidence of the structure of the Open Data community, the motivations of its members and, most importantly, the role that powerful 'intermediaries' such as government departments, international economic institutions and data-centric corporations will play in the flows of information across this emerging industry.

2 Literature Review

With such pressing questions of participation, power and provenance overshadowing Open Data technology, it is surprising so little research has been published on the topic. Existing studies generally fall into two camps: economic and policy studies into the impacts of public sector information (eg: Robinson et al., 2009; Pollock, 2009) or studies on the place of scientific open data as part of a wider 'Open Science' movement (eg: Murray-Rust, 2008; Guha et al., 2006). Many researchers approach the topic from a purely computing science perspective, discussing the actual technological structures and processes of Open Data, RDF, Triples and the Semantic Web (eg: Krummenacher et al., 2010; Berners-Lee et al., 2006).

What is missing, however, is a study which takes a step back from the intricacies of licensing models and triple stores, and instead explores more generally the actors and information flows which those legal and technical architectures propose to be governing. By understanding more of how Open Data is currently used, why, and by whom, future discussion on Open Data policy and technology can be all the more well-founded. The exploratory research described in this thesis is an attempt to do just that: to lay empirical foundations for further study.

In laying these foundations, it is best to consider what Open Data stands for: collaboration and innovation, enabled by the Internet (Shadbolt, 2011*b*). From this central aim, Open Data's cultural heritage in the Open Source movement

becomes clear. If we are to understand the motivations of Open Data providers and developers, then Open Source literature is surely a good place to start.

Like the Open Data movement, Open Source was fundamentally linked to the architecture of the devices and networks on which it ran. Kim (2003) argues that the patchwork nature of the early UNIX operating system necessitated Open Source collaboration, while Reynolds' conception of the Internet as a "horizontal medium" (Reynolds, 2006, p121) goes some way to explaining why an Open Source community was able to develop in the first place. Blogs, wikis and social networks have all been attributed to the success of Open Source collaboration (Bauwens, 2005; DiBona, 2006), although the community's roots lay in far more basic communication media such as bulletin boards, email and IRC (Hafner and Lyon, 1998; Raymond, 1999).

Technology is, as Weber (2004) notes, an enabler: increasing the diffusion of information and knowledge (von Hippel, 2005*a*) and acting as a "fundamental dimension of social change" (Castells, 2001, p155). However technology alone is not enough to enable groups of Open Source developers to collaborate and innovate. Protocols and standards apply just as much to developers as to the software they create. "Technical specs matter," warns Eugene Thacker (in Galloway, 2004, p*xii*), both "ontologically and politically." Taking inspiration from Deleuze's concept of control societies (Deleuze, 1992), Thacker and Galloway discuss how information and protocol can effectively shape and direct the progress of an otherwise decentralised group of actors. "Protocol is fundamentally a technology of inclusion," Galloway argues, "and openness is the key to inclusion" (Galloway, 2004, p147). Perhaps then, we can understand the data with which developers work not only as the raw material of their trade but also the glue which holds their community together? That is, of course, if an Open Data community exists at all.

O'Mahoney and Ferraro (2004), for example, confirm that frequent interaction between Open Source participants—both online and in person—builds trust, and indeed, the American and British Open Data scenes are frequently punctuated by physical and virtual 'hack days' in which groups of developers converge at a single venue (or multiple venues linked by video and social media streams) for 24–48 hours' intense software development (Kuk and Davies, forthcoming). Hack days can be seen as very physical manifestations of the 'communities of practice' surrounding Open Data development (Wenger, 1998). Such communities are characterised by knowledge sharing and the rapid propagation of information (Berdou, 2011) and by a meritocratic focus on value created (Weber, 2006). Communities of practice, and the associated concepts of 'egoless programming' and small, modular teams, have also been suggested as methods for overcoming "Brooks' Law" (Raymond, 1999; DiBona et al., 2006)[1]. The concept, I would argue, goes some way towards explaining how Open Source communities maintain their productivity even as their numbers grow larger and more complex. One of the aims of this research, then, must be to examine whether such communities also exist around Open Data.

In *Democratizing Innovation*, von Hippel (2005*a*, p165) discusses the structure of "information communities" which "rendezvous around an information commons" open equally to all participants. The notion bears obvious similarities to the Open Data movement, where both data and the knowledge of how to use it circulate between developers, often through a process of 'learning by doing' (von Hippel,

[1]Software engineer Frederick Brooks noted that "adding manpower to a late software project makes it later" (Brooks, 1995) – in fact, for each person added, the time taken to complete the project squares, since each developer must cope with each other developer's different coding styles and practices. This so-called 'Brooks' Law' poses a major challenge to corporations employing large teams of software developers, leading to coping strategies such as software modularisation and Agile development practices.

2005*b*). Indeed, a number of studies have shown that learning new skills is a main motivation for 80–90% of Open Source contributors (Lakhani et al., 2002; Ghosh et al., 2002).

The motivation of developers is a key step in understanding the processes of Open Data development. Literature on the Open Source movement suggests a range of possible motivations for participation. As Lerner and Tirole (2002) note, the common explanation of altruism only goes so far – they instead focus on 'signalling incentives' such as career advancement and ego-boosting. In a study by Kim (2003), one in five developers reported that their open source contributions led to a job, while other studies have suggested up to 25% contribute for improved job prospects (Deek and McHugh, 2008).

Interest and creativity are two often cited motivations, especially for 'hackers' in the traditional sense of the term[2]. Himanen (2001) points out that some of the Internet's key figures—including Vint Cerf and Linus Torvalds—were motivated, at least initially, by technological curiosity. Lakhani and Wolf (2005) go further, suggesting many Open Source developers may in fact be seeking psychological "flow states", where they can lose themselves in creatively solving heuristic coding problems (see also Csikszentmihalyi, 1996). Perhaps working with Open Data can fulfil a similar function.

Some agreement seems to have arisen, however, that social factors are among the Open Source movement's strongest motivators. In a study of over 2700 European developers, Ghosh (2005) reports that over 53% felt participating in the Open Source community was the strongest motivation to join. David and Shapiro (2008) agree

[2]Meaning a "tinkerer, problem solver, expert" (Raymond, 1999:2) rather than the malevolent 'crackers' more commonly discussed by the media.

that big projects attract 'social programmers', for whom the community-based intrinsic motivation is the major factor. Galloway (2004, p80), quoting Deleuze, adds that "technology is social before it is technological," and indeed a number of theorists have argued that technological innovation is a largely social practice (see, for example, Tuomi, 2002).

Interestingly, of the sparse research we do have on the motivations of Open Data developers (see Kuk and Davies, forthcoming), the principal motivator is split between ideological and need-based factors. Some participate out of an obligation "to show how government services could be run better or more efficiently," while others are driven by personal frustrations and the need for certain facts or visualisations (ibid). It is true much Open Data development has a political/activist undercurrent, but the movement can clearly go further than public service. As Bracking (2011) and Shadbolt (2011*b*) show, there is an increasing need for open *corporate* data in a number of markets. "Information," say Nye and Owens (1996, p35), "is the new coin of the international realm," and companies as diverse as Facebook[3] and Nike[4] are all hoping data flows will give them a competitive edge. "Better information makes better markets" (Shadbolt, 2011*b*) and, indeed, any study of current Open Data trends must deal with the growing tide of corporate data, as well as the more prevalent streams of government data catching all of the headlines.

Since governments, NGOs and corporations now rely so heavily on information to underpin their power (Mayer-Schönberger and Brodnig, 2001), understanding these information flows will ultimately lead to understanding which actors those

[3]Facebook, in April 2011, released detailed specifications and CAD drawings for its datacentres onto the Internet, for other companies to use and improve upon (see *http://opencompute.org*). Most of Facebook's competitors—like Google, for example—keep this data highly secret.

[4]Nike's recently announced 'Better World' program includes releasing information about their supply chain as Open Data, to encourage logistical and environmental innovation (see *http://www.nikebetterworld.com*).

flows empower. While previous scholars commented largely on government exploitation of information flows (Nye and Owens, 1996; Keohane and Nye, 1998), Hindman (2009) goes further, applying the concept of 'intermediaries' to search engines and media hubs like Google and the Huffington Post. Far from encouraging democracy, he argues, the Internet is instead enabling new 'winner-takes-all' networks and political elites, more concentrated than even the traditional print and broadcast media empires. Hindman even identifies such intermediaries in Open Source development, where gigantic projects such as Apache, Firefox and the Linux kernel attract an order of magnitude more developer attention than the rest (Hindman, 2007, 2009). The question is: To what extent are similar power-centres evident in Open Data development?

To understand the Open Data information space, and to locate the most powerful nodes within it, we must first research where the data is being shared, who is enacting it, and which routes it is taking through the network. With these objectives in mind, it soon becomes clear that research is required on three fronts, to answer the questions of:

RQ1. How is Open Data being used on the Web? What types of data are most popular? Is government data, for example, still leading the way, or have other data types started to gain a following? How many web apps combine multiple datasets, and where are these datasets from? Lastly, how far is the information travelling, and how distributed are the developers who use it?

RQ2. Who is releasing and developing with Open Data? How often do developers collaborate over Open Data? Does Open Data development follow the communal Open Source model of 'many eyes', or is it a more individual activity? Is there global collaboration, as Open Data proponents initially hoped? And on a

personal level, are the motivations of Open Data developers similar to those in the Open Source community, or do they have different goals in mind?

RQ3. Have dominant intermediaries emerged in Open Data development? Returning to a macro perspective, are there signs of the emergence of central participants, platforms, or meeting places in the processes of Open Data development? Where do developers find their data, and what is the role of government in supporting and even shaping Open Data development?

Before we discover the answers to these research questions, it is important to explain the methods I employed, and the reasoning behind their selection. As readers are no doubt aware, the Open Data community, like the Open Source community before it (Berdou, 2011; Feller et al., 2005; Dekkers et al., 2006), is notoriously hard to research. It is not, however, impossible. Gathering reliable findings depends on the selection of suitable measurement methods, properly honed to the research aims above. I would like to take a moment to explain my choices.

3 Methodology

With so little existing research on Open Data development, my aims were first and foremost to map the landscape and note patterns therein, from which fundamental theories could later be generated. While I would not go as far as Miles and Huberman (1994, p1) in describing qualitative data as "sexy," it was nonetheless clear that an inductive, qualitative approach would enable me to gain this deep exploratory understanding of both the Open Data field and the actors within it. Qualitative research has an inherent "emphasis on process" (Bryman, 2008, p388) and an ability to quickly generate new theories from experiential data (Miles and Huberman, 1994; Berg, 2009), leaving it well suited to this particular study's exploratory research questions.

It soon also became clear that one method would not gather the range of data necessary to answer those research questions. While content analysis of the structure and code of Web apps utilising Open Data would answer RQ1 and RQ2, it would give little insight into the motivations and community practices behind app creation. And while network analysis of the relationships between apps, developers and sources would help identify clusters and central players, it would do little to explain *why* those hierarchies exist in the first place.

Thus I combined two data collection methods (website cataloguing and interviews) and two qualitative analysis methods (content analysis and network analysis) into a thorough multi-method research design (Johnson et al., 2007; Morse, 2009). To

borrow Morse's (2010) notation, the study followed a simultaneous QUAL+qual process, with the core method, content analysis, informing the complementary method, interviews. The interviews, in turn, helped confirm the motivations and hierarchies behind the Web apps, providing vital answers to RQ2 and RQ3.

As Morse (2010) notes, content analysis and observational techniques like focus groups or interviews often work well together to provide thick description on the one hand and experiential data on the other. Although under-explored in traditional mixed-methods literature (Webb et al., 1966; Denzin, 1970), so-called *'within-methods'* qualitative research designs are becoming more popular thanks to their descriptive, contextual approach, and the possibility for triangulation of findings (Brannen, 2005). My combination of interviews and qualitative social network analysis, for example, is not without precedent and provided a useful "perception of the network from the inside" (Edwards, 2010, p24).

It is worth noting that, although other studies following Open Data developers have made effective use of participant observation at 'hack days' (eg: Kuk and Davies, forthcoming), the homogeneity of hack day attendees (who are often of similar proficiencies and from similar backgrounds) led me instead to a purposive sample of interviewees which would garner the experiences of a wider range of Open Data participants, including those *supplying* the datasets – participants largely absent from developer-centric hack days. Interviews also presented the advantages of being better suited to the reconstruction of past events and experiences (Bryman, 2008) and to guiding from the interviewer, especially in response to topics highlighted by the concurrent web app content analysis.

3.1 Content & Network Analysis

The difficulty of generating a sampling frame for uses of Open Data has already been noted (Dekkers et al., 2006; Kuk and Davies, forthcoming). Even when one concentrates on *websites* using Open Data (the most prevalent medium, and the one with arguably the lowest barriers to entry), a sampling frame remains elusive. Although there are some partial lists of Open Data apps (such as *data.gov.uk/apps*) they by no means represent the entire population. Likewise, the 'galleries' left behind after Open Data competitions also prove fertile sources for app references, albeit from a self-selecting audience of proud developers. Unlike hyperlinks, which can be back-traced through search engines by using the 'link:' search prefix, Open Data apps rarely link directly, in a search-engine-visible way, to their source datasets. This oversight in itself presents a major barrier to attempts tracing Open Data provenance, as I will revisit in the Discussion section.

With a representative sample of web apps unattainable, I gathered a cluster sample (Krippendorf, 1980) of links from the following aggregator sites to build a suitably large (and suitably diverse) corpus of apps for analysis:

- http://appsfordevelopment.challengepost.com/submissions
  (Apps based on the World Bank economic and development data)
- https://pub.needlebase.com/actions/visualizer/V2Visualizer.do?domain=Open-Data-Apps&query=Application
  (UK, USA and Canadian apps, mainly government, automatically scraped from official repositories)
- http://data.gov.uk/apps (UK government apps)
- http://warwickshireopendata.wordpress.com/app-gallery
  (Apps using UK local authority data)

The clusters were chosen to increase the variety of apps discovered. Together they provided examples of local government apps, national government apps, and non-governmental apps including development apps (mostly) created specifically for a competition. They also represented apps from different countries, and collated with different degrees of centralisation.

Data collection was specified such that apps without source or developer attributions were skipped, as were apps that required payment for access, or that simply no longer functioned. Non web-based apps (eg: those run locally as executables) were also skipped.

In total 175 apps were manually catalogued through a custom-built CAQDAS system which recorded manifest content such as the name, URL, scale, data type, creation date, last modification date, sources and developers of each app, along with a freeform qualitative description of the process of using the app from a visitor's perspective. The categories were guided by my research questions: codes for app types and locations clearly contributing to RQ1, codes describing developer sizes and locations aimed towards RQ2, and the relational data itself aimed at detecting hierarchies for RQ3. Since all the data were publicly accessible, there were no significant ethical issues in their collection.

The very act of recording the structure and content of the web apps could be seen as a process of content analysis, but once a sufficient range of apps had been collected, further analysis was performed on the codes collected in the database. Patterns in the codes were noted and considered in relation to the themes brought up in the concurrent interviews. To answer RQ2 and RQ3, the relational information between apps, sources and developers was extracted from the database—something that traditional qualitative coding software packages wouldn't allow—and visualised through network diagrams to reveal the distribution and interaction of actors

across the sample. Network summary statistics such as betweenness centrality and degree helped identify 'brokers' or intermediaries between different parts of the network (Burt, 1992), although as I will later discuss, their utility was limited by the network's sparseness and inherent biases.

3.2 Interviews

A purposive sample of interviewees was chosen, based partly on names emerging from the content analysis and partly on my own professional experience of the field. The aim was that a wide range of perspectives could be brought together both for comparison with the content analysis data (Krippendorf, 1980) and to stand in their own right. Interviewees were contacted by email, and if the response was positive, they were supplied with a consent form and given the option of being interviewed in person, over the telephone or via VoIP. Two interviews were carried out in person, two over the telephone and one via VoIP. A sixth was planned to take place in person but illness forced us to reorganise for a telephone interview. Bryman (2008) notes that qualitative telephone interviews are relatively uncommon, perhaps because of the lack of non-verbal cues, despite evidence that the medium makes no noticeable difference to responses (Sturges and Hanrahan, 2004). I found telephone interviewees went into more detail and spent more time talking than their in-person counterparts, but neither medium was significantly better or worse than the other.

Interviewees were offered anonymity in the final report, although none desired it. Similarly, all six interviewees agreed to audio-recording of their interviews. A list of the interviewees and reasons for their selection are presented in Table 3.1 for the reader's convenience:

Table 3.1: Interviewees

Beth Noveck	Former White House CTO and leader of the US Open Government Initiative, Beth now advises the UK Government on technology and participation.
Chris Taggart	The creator of websites OpenlyLocal and OpenCorporates, Chris also advises the London Datastore and sits on the UK Government's Local Public Data Panel.
Francis Irving	A former Open Source programmer and MySociety member, Francis now runs ScraperWiki, a platform for collaborative data collection and visualisation.
Matthew Somerville	Although also member of MySociety, Matthew is perhaps better know for his unusual uses of Open Data, such as his infamous Live London Underground Map.
Richard Taylor	Less of a front-line developer, Richard is a vocal commentator of local and national government, and a supporter of community Open Data initiatives.
Rupert Redington	With a background in education and Open Source technology, Rupert develops with Open Data as an individual and as part of a small web development company.

Interviews followed a rough interview guide (copies of which were supplied to interviewees in advance) again led by my research questions and on findings from the concurrent web app cataloguing process. The semi-structured approach struck a good balance between providing enough flexibility to follow up interviewees' points with additional questions, and producing answers largely comparable between interviewees (Bryman, 2008).

Completed interviews were transcribed, and then coded for recurring themes or concepts in a grounded theory-inspired process (Charmaz, 2006). The resulting categories (see Appendix) were chosen to be, as Rose (2007) puts it, "exhaustive, exclusive and enlightening" and to provide data suitable to answer my research questions. Patterns were noted and incorporated into future interview guides and the web app content analysis.

4 Findings & Analysis

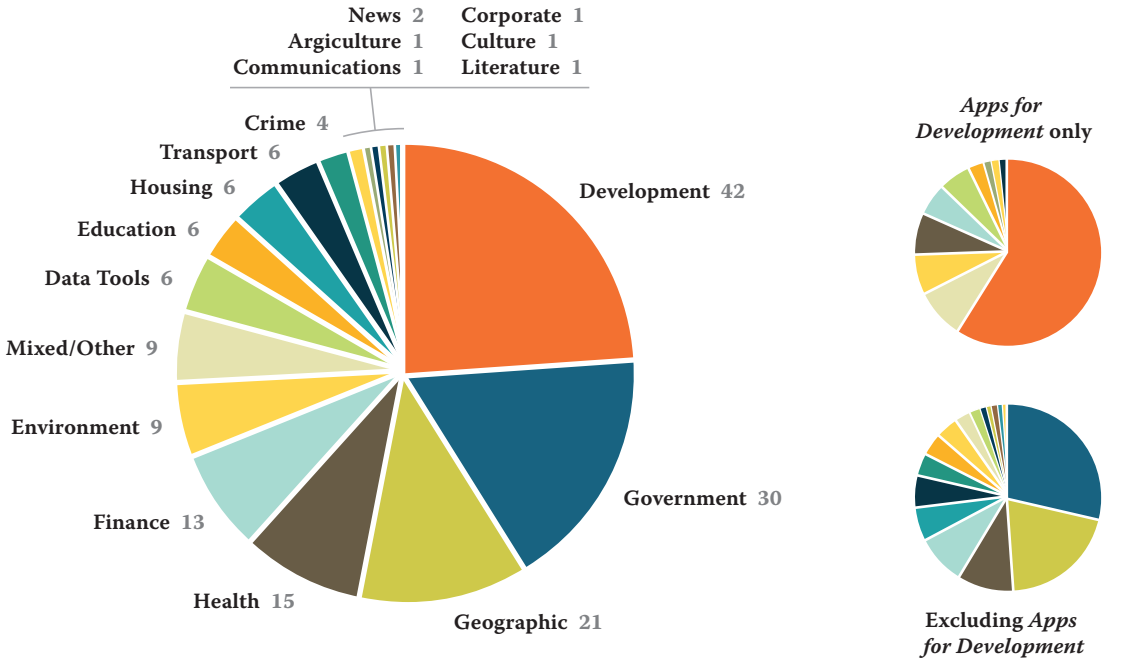
The content analysis of both web apps and interviews revealed a number of important patterns and themes, which I shall explore in the order of my research questions.

4.1 How is open data being used?

Of the 175 web apps analysed, just under a fifth fulfilled a government or public service remit, while a quarter were built around development data such as poverty and economic indices (Fig. 4.1). All of these 'development' apps used the World Bank's Millennium Development Goals dataset, and all but two were created specifically for the World Bank's *Apps for Development* competition in Spring 2011. Filtering out the 71 apps from the competition reveals a stronger trend towards governmental/public service and geographic data in the remaining apps. Indeed, mapping seems to be a very popular means of visualising local datasets, with just under a sixth of all surveyed apps utilising Google Maps functionality. Surprisingly few apps utilised Google's open source competitor OpenStreetMap, although its usage as a data source in a total of 8 apps made it the second most prevalent data source in the sample, after the World Bank's MDG dataset.

Aside from development, government and geographic apps, those dealing with health and financial data also proved modestly popular, along with apps focussing

Figure 4.1: Types of web app present in sample

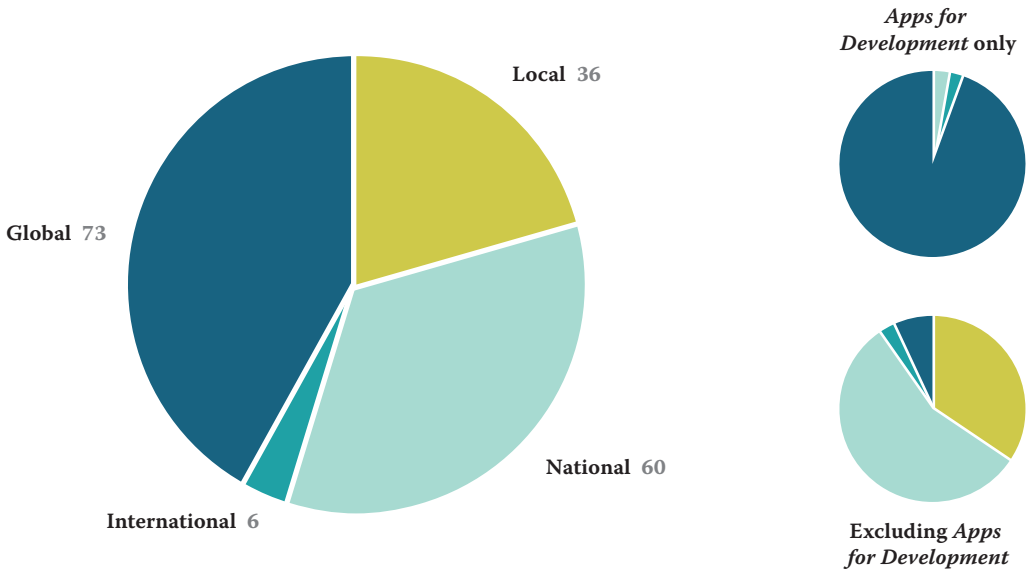


on everyday issues such as transport, housing, crime and education. Notably, the sample included one 'literary' app (an open compendium of Shakespeare's works) and one 'corporate' app (OpenCorporates), suggesting the novel uses to which Open Data could be put in the future. A number of the remaining apps were data *tools*, aiming to help users gather and visualise datasets of their own choice.

While Open Data is clearly being put to a wide range of uses, government data still retains its traditional popularity among Open Data developers. Interview responses confirmed this finding, with all six interviewees having dealt with government or public service data on multiple occasions. Although four of them showed excitement about future uses of corporate data in particular, it seems very few apps utilise it at the moment.

The 'scale' of each app was also recorded, to show the geographic area to which each applied (Fig. 4.2). Apps for local, national and global audiences were of

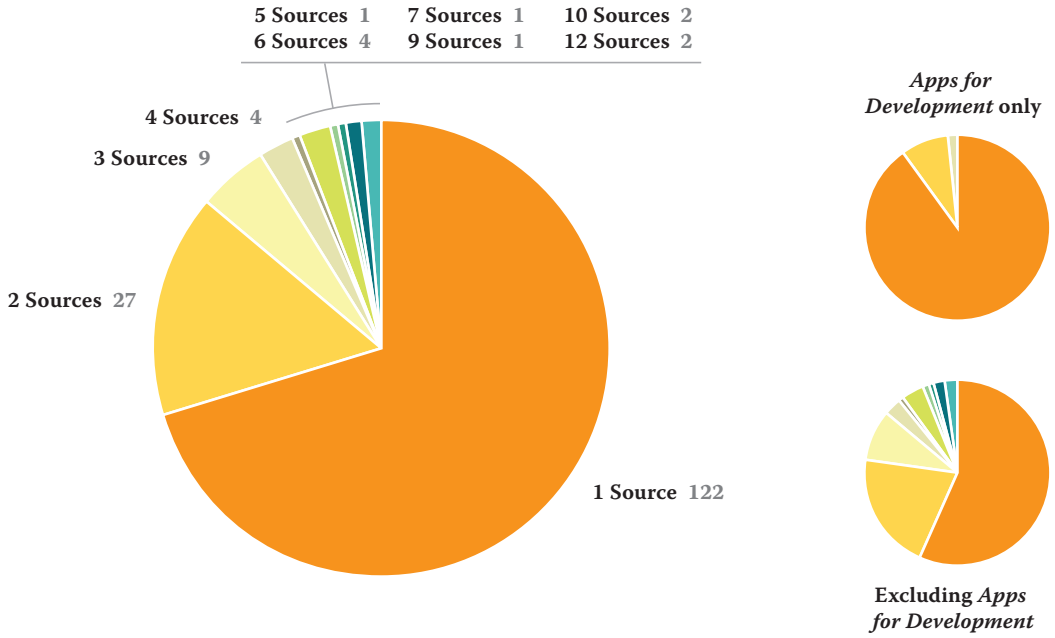
Figure 4.2: Areas to which web apps applied



roughly equal prevalence, although, if we again filter out those apps built for the *Apps for Development* competition, a pattern towards local and national data becomes much clearer. The sheer number of *Apps for Development* entries in my sample clearly masks the noticeably insular nature of the other 'organic' Open Data apps. Furthermore, apps of different scales tend to have different focusses: global apps tend towards exploring economic or development progress in a number of countries; national apps focus more on civic and government affairs; while local apps make increased use of mapping data to plot everyday statistics such as education, crime and health figures in a familiar context.

Data sources, meanwhile, are also frequently used in isolation (Fig. 4.3). Very few of the sampled apps combined two or more sources, with nearly three quarters only using data from a single dataset. Roughly an eighth used two source datasets. Interestingly, apps submitted to the *Apps for Development* competition combined far fewer sources, on average, than the other apps in the sample.

Figure 4.3: Number of data sources per web app



One of my aims was to measure the actual geography of Open Data flows, by recording the geographic location of data sources and developers. This results, perhaps unsurprisingly, in a concentration of participants in London, New York, Washington and Ottawa, and a long tail of sources and developers in a range of cities and countries around the world. Although the unavoidable bias towards the centres of economic and political power within my sampling frame has skewed the numbers, the presence of particular locations is still of some interest. For instance, despite Cambridge (UK) being home to a number of Open Data developers (including Rufus Pollock and the Open Knowledge Foundation – two of the most active developers in my sample), apps employing data from Cambridgeshire Council, or indeed any other source based geographically in Cambridge, were noticeably absent from my sample.

Of the thirteen source organisations with the most frequently used datasets, all were based in the UK or North America. With the exception of one dataset released

by the African Development Bank in Côte d'Ivoire, all of the data sources in my sample originated from organisations in Europe or North America. The location of developers was a little more distributed, with the World Bank competition in particular attracting developers from a number of African and East-Asian states.

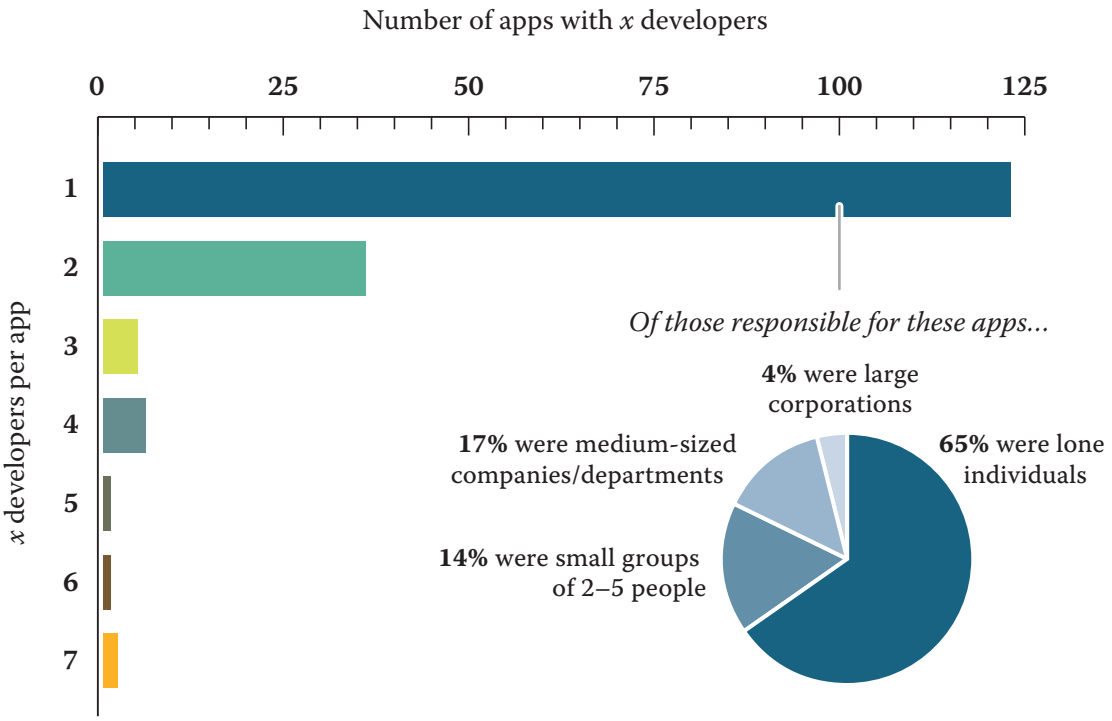
4.2 Who is releasing and developing with Open Data?

A good deal of Open Data's allure to governments thus far—and the Open Source movement's success over the past two decades—has been attributed to the collaboration that openness brings. "Given enough eyeballs, all bugs are shallow" as the Open Source proverb goes (Raymond, 1999, p41). How does this translate to the reality of Open Data use?

In short, not well. Developers in my sample were coded by 'size' – lone individuals, small groups (2-5 people), medium companies or university departments (5+ people) and large organisations (like Google and The New York Times). Overall, just under three quarters of developers in my sample were thus classified as 'individuals' (Fig. 4.4). Similarly, of the apps in my sample, the vast majority (nearly three quarters) had only one 'developer'. Further analysis reveals that, of those lone 'developers', two thirds were in fact individuals working alone on their app, with the remainder split equally between small and medium groups of developers. Thus, while collaboration is evident in my sample, roughly half of catalogued apps were developed by a lone individual[1].

[1]It is worth noting that non-contemporaneous collaboration may actually be evident through shared code and software libraries. Francis Irving in particular highlighted code sharing as one of the nascent Open Data community's greatest strengths. I, however, take a view similar to that of the Open Source literature (Berdou, 2011; DiBona, 2006; O'Mahoney and Ferraro, 2004) that collaboration in a physical, simultaneous sense is the more valuable construct for understanding developer communities.

Figure 4.4: Number of developers per web app



In the interviews, responses were split, with two interviewees noting a lack of community (especially when compared to the Open Source Software movement) and another three enthusing that communities *did* exist, albeit multiple and centred around tasks or interests rather than data.

Francis Irving and Rupert Redington highlighted the network of bloggers, politicians and groups like MySociety and the Open Knowledge Foundation, which form a sort of community or, in Rupert's terms, "loose alliance," around Open Data. "But at the same time," Francis admitted, "it doesn't feel quite like the open source community" – perhaps, he suggested, because Open Data developers lack a common 'enemy' against which to rally.

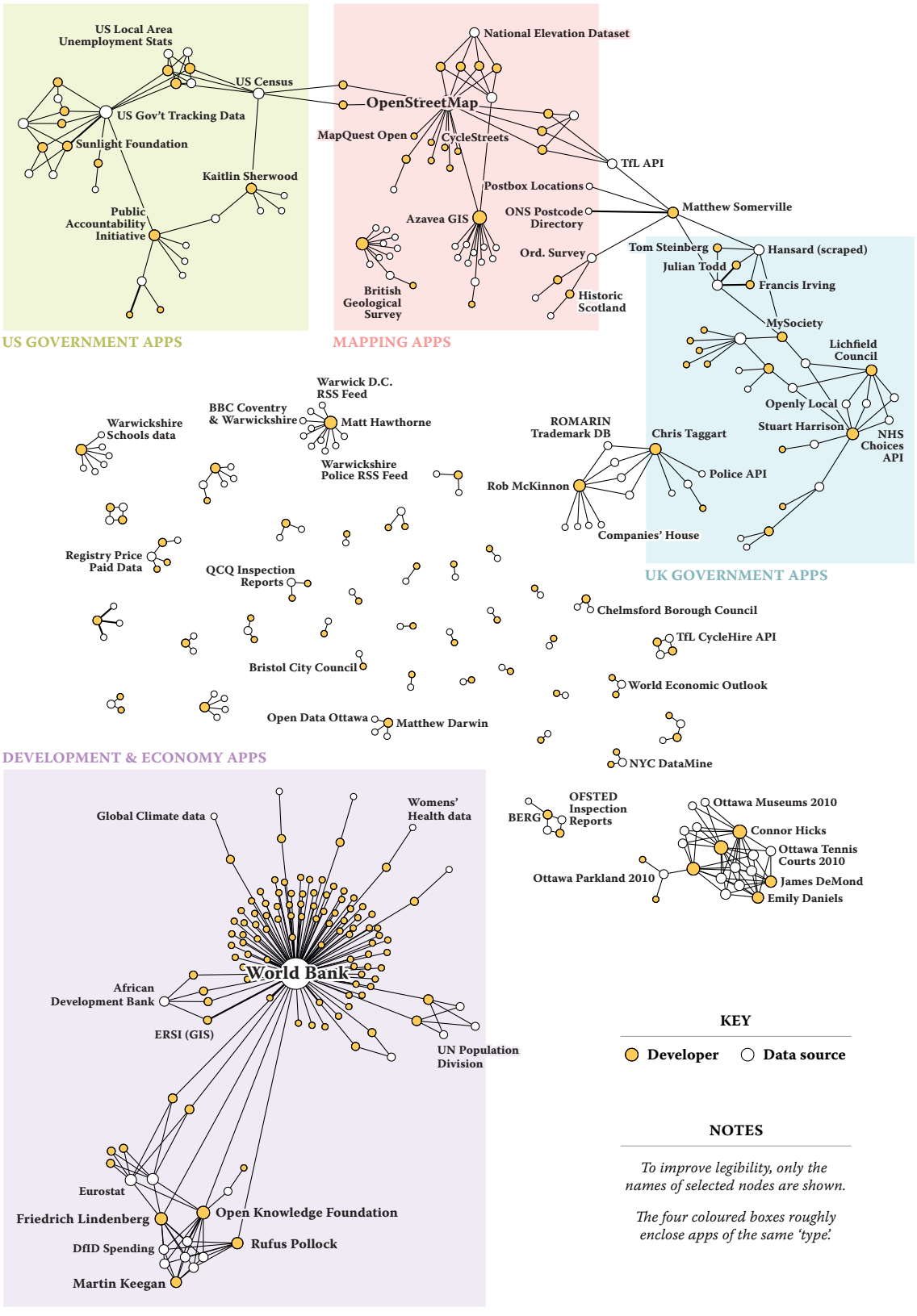
"There are lots of communities around things *of* Open Data," Matthew Somerville,

a developer and member of MySociety explained, "but not around Open Data *itself*." Richard Taylor, a local government activist, agreed that, at least in Open Government Data "everybody knows each other," helped in part by hack days, conferences and Twitter hashtags. "Lots of this stuff is inherently collaborative," he argued, "that's what the whole framework is about." He added, however, that work was also often done by individuals, such as Chris Taggart, who had "done his own thing" with corporate data. Chris himself concurred that "there is quite a community," of which government ministers and civil servants also formed a part. However, Open Data often retains this tendency towards individual or small-group development, with "a bunch of people going off and solving their problems, rather than coming together from disparate groups."

To better understand the interaction between Open Data developers, I constructed a two-mode network diagram of developers and sources within my sample (see Fig. 4.5, next page). My aim was to uncover why the interviewees were having such a hard time confirming or disproving this feeling of community. Developers and sources were represented as nodes, scaled according to their degree. Developers share edges with sources they have used. In effect, each edge represents a Web app from my sample. The nodes were laid out using Yifan Hu's Multilevel algorithm (Hu, 2005), to clearly separate out clusters of developers and sources.

Unsurprisingly, the network is sparse, since most apps in my sample shared only a few common developers or sources, if any at all. However, looking closer, sizeable clusters can be discerned, such as one of developers and sources dealing with the activities of the UK public sector, another cluster of USA government watchers, a third concentrated around international development and finance, and an array of smaller clusters based around local community data.

Figure 4.5: Developer & source network



The World Bank dataset (with the highest degree by far) attracts a dense cloud of developers, most of whom participated specifically for the *Apps for Development* competition, and very few of whom combined the dataset with others from different sources. The World Bank dataset does, however, form part of a more general 'economic' or 'development' cluster, which includes developers such as the Open Knowledge Foundation, and sources such as the European Commission (Eurostat) and the UK Department for International Development (DfID).

Further up the diagram, three clusters—one around UK government or public service data, one around mapped data, and one around US government data—are connected by relatively popular data sources such as US Census Data and the Transport for London API. My sample included no sources or developers common to both UK and US government clusters, suggesting a strongly national (rather than international) focus for government data use. Local datasets, meanwhile, form a disconnected array in the centre of the diagram, with few data sources being reused by multiple developers. It is in these local apps, however, that developers combined the most sources. One, for example, combined six datasets on Warwickshire schools to provide an at-a-glance summary of every school in the County. Another (a high-school project!) charted the location of twelve different types of child-friendly amenities on a map of Ottawa. A third provided news and events listings via a mashup of ten RSS feeds from sources such as the BBC, the local council, local police and regional newspapers.

Thus the situation seems to be closer to Matthew Somerville's observation: of a constellation of developers, many individual but some linked into communities of task-based groups, rather than one overall 'Open Data' community. These task-based groups, and Chris Taggart's observation of "people going off and solving their problems" also tied in with a central theme, in the interviews, of *motivation*.

A number of incentives and motivations for using Open Data were suggested during the interviews. Rupert Redington described much Open Data development as "21st Century pamphleteering" through which political agendas were often pushed. Chris Taggart, Francis Irving and Richard Taylor all expressed a more constructive aim to "show government what can be done" with Open Data, something to which Beth Noveck, recently appointed by the UK Government to increase participation, agreed. She hopes to push government to react more quickly to innovations suggested by citizens and developers, in the hope that their ideas could help the UK Government do its job more efficiently. Indeed, efficiency and transparency were commonly-cited motivations for governments and corporations to *release* data, as was the notion of using data to gain a competitive advantage in the marketplace.

Other interviewees, however, disagreed. "Nobody gets on their computer one day and goes: I need some data," explained Matthew Somerville, "You want to *do* something." Getting something done or fulfilling a task, then, is also a key motivator. Francis Irving began his interview by declaring "I don't actually care about Open Data." He cares instead about *how it can be used* to achieve change. The data is simply a means to an end, and that end—it appears—is very different for each community of developers or sources in the Open Data scene. In my sample alone, motivations for developers ranged from improving government processes and increasing government transparency, to fulfilling personal needs or requirements, and to creating new, more effective business models. Motivations for sources to release data included capitalising on the 'wisdom of the crowds', gaining a competitive advantage, improving efficiency and more generally expediting economic growth.

4.3 Have dominant intermediaries emerged in Open Data development?

Despite governments' seemingly powerful positions as leaders of the Open Data movement, it was the threat of *private* companies owning information that concerned three of my interviewees. Rupert Redington joked about "the tyranny of Ordinance Survey" while Matthew Somerville admitted his fear that "they are the owner of postcodes now." Matthew also discussed the lack of technical control that proprietary data can lead to, noting that tight commercial controls over civic data would would have stifled innovation on the early MySociety sites, had they not been given free access to otherwise expensive datasets through ties with the Department for Constitutional Affairs.

Francis Irving raised a more fundamental objection to Ordinance Survey's ownership of electoral district boundary data. "Heck! That's the boundar—that's our democracy!" he exclaimed, adding that "it was just so blatant, so clear; that should be public." He believes the problem of data ownership extends further than government, into modern Internet services. And worse, the freedom-fighters of the last few decades, the Open Source Software community, seem to be oblivious:

> "What they are missing is the new problem, which is the proprietariness off the top application layer [...] Facebook might make identity proprietary, potentially. They could own identity on the Internet."

He pointed to services like Unhosted as examples of technology empowering individuals by giving them control of their own data again. "That's where the cool shit is", he joked, "not making Gnome 3. Like, seriously!"

He agreed, however, that when it comes to Open—rather than personal—Data, "fewer people are interested" in questions of control. Rupert Redington similarly

suggested that teenagers in particular are more likely to worry about data privacy than data control. This no doubt ties into a general theme throughout all of the interviews of data being a minority interest, something for geeks and 'armchair politicians.' Beth Noveck, for example, admitted that Open Data "doesn't have the demonstration value [...] that other projects do" and that one of her aims in pushing UK Government to release more data is also to "make it concrete in very real ways to people" what role Open Data can play in their everyday lives.

Both Francis Irving and Richard Taylor discussed the importance of having a few observers, well-versed in data analysis, watching government on everyone's behalf, implying the *need* to watch government, as one of the most powerful players in the Open Data environment. Richard Taylor and Matthew Somerville, on the other hand, also noted the power new private sector data providers were accumulating: Richard pointing out MySociety's 'MapIt' service as an example of a positive intermediary, on top of which numerous other applications have been built; and Matthew discussing how the widespread online adoption of MusicBrainz and IMDB identifiers (ie: the specific numbers given to online representations of real world things like songs, albums, films and actors) could quickly cement them as intermediaries in their respective sectors.

Evidence of this 'cornering' of a specific market can be seen in the prevalence of Google Maps in my sample, although Google Maps itself is not an example of an Open Data intermediary in the strict sense of the term. That said, Google might actually constitute a data intermediary in another sense, thanks to its dominance of the web search market. Two of the interviewees stated that they often use Google, rather than government portals like Data.gov.uk, to find datasets, with Chris Taggart in particular adding that a general web search is often the only way he can find the highly specific data he requires.

Looking at the network of developers and sources that have been extracted from my web app content analysis offers some suggestions as to the existence of intermediaries. The World Bank has the highest degree (number of links to developers and sources) and the highest betweenness (more information has to flow through it to reach other parts of the network). Its extreme performance in these metrics, however, may be more down to sampling bias than actual power in the Open Data landscape.

After the World Bank, OpenStreetMap holds the next highest degree and betweenness, suggesting it too acts as an information broker between disperse areas of the network. Closely behind it, Matthew Somerville is identified as the most influential developer in the sample, with a very high betweenness despite his modest number of links to other developers and sources. He is clearly visible in the network diagram, connecting the UK Government group to the Mapping group. Perhaps he, as a bridge between two communities, could be seen as one of the sample's strongest intermediaries, at least in terms of social brokerage.

5 Discussion

The above analysis certainly raises some important points for discussion. Perhaps most obvious is the lack of discoverability for uses of Open Data. With the recent announcement of the US Government's plans to dramatically cut funding for their data-related sites, a number of critics highlighted the relatively low visitor numbers to sites like Data.gov and Data.gov.uk and argued for increased visibility (Yau, 2011). I would argue those critics were onto something but had focussed on the wrong part of data's use-chain. The issue is not with Open Data repositories having modest visitor statistics; such sites are meant only for developers, for people wanting to *use* the data before presenting it to the wider public. As Tom Steinberg put it:

> "There is no need for the Data.gov to be a big shiny, well trafficked site [...] Sites like Data.gov should be entirely honed to the needs of a small number of frustrated data seekers." *(Arthur, 2011)*

Although regular citizens may like to know sites like Data.gov and Data.gov.uk exist, they are not those sites' target audience; developers are. And developers, from the look if it, seem to know the sites exist – even if, like Chris Taggart, they usually end up finding their datasets through other means.

The issue is not over supplying developers with the tools of their trade: Budding Open Data developers with a passion, a need or a goal will *find* the data, just as

budding Open Source programmers will *learn* the necessary steps to submit patches to their favourite software or even develop new software of their own (Berdou, 2011; DiBona, 2006). The real issue, as the Open Source movement discovered, is making yourself known to *end users*. In the Open Source realm this task has been performed admirably by popular titles like Firefox and Wikipedia, but still the majority of Open Source products fall below the radar (Deek and McHugh, 2008). I worry Open Data products fare little better. While data journalists such as those at The Guardian and the New York Times have effectively brought solid data and clear visualisations to the masses, we now need similar intermediaries to provide personalised, engaging uses of Open Data to the wider population (Sunstein, 2007; Arthur, 2010). The success of services like FixMyStreet and OpenStreetMap belie a much longer tail of potentially useful but largely invisible Open Data apps. Finding a solution to this predicament is beyond the scope of this thesis, but this lack of discoverability must be overcome if Open Data is to fulfil its potential as a useful, empowering force in society.

Apps dealing with non-governmental Open Data, in particular, would benefit from greater publicity. We have already seen how much impact development competitions can have on participation: the World Bank's *Apps for Development* competition incentivised the creation of over 100 apps dealing with their Millennium Development Goals dataset, from a developer-base noticeably more geographically dispersed than any other in my sample. Similar events, well publicised, could quickly raise the profile of non-governmental Open Data, even to users outside the Open Data community. As my findings have shown, Open Data in a wide range of subjects—from health and development to financial and corporate data—is finding its way into apps and services. This is a trend that should be supported and encouraged.

Looking more closely at the users of Open Data themselves, a number of patterns became clear. Content analysis of my interviews resulted in two central theories, the first of which was that, despite the numerous incentives to both produce and develop on Open Data, it is still seen as an individual niche interest, with no real community to provide momentum. Content analysis of the Open Data webapps themselves revealed a similar story: the majority of developers work alone, and the majority of apps utilise only one data source. The situation is reminiscent of patterns of Open Source participation (Deek and McHugh, 2008; Weber, 2004; Lerner and Tirole, 2002), and a far cry from early Open Data proponents' hopes of developer collaboration and data combination.

Why are more developers not collaborating? Francis Irving and Rupert Redington, linking back to the issue of discoverability, suggested that collaboration was limited because very few people can see what you are working on, or what you have made. Sites like GitHub and Francis' company, ScraperWiki, hope to counter this by becoming central platforms on which developers can visibly collaborate.

The interviews also revealed a need for leaders or visionaries to champion the Open Data cause. It became clear from Beth Noveck's responses that she often finds herself acting as a persuader and a motivator – both for government to release more data, and for it to work more closely with the community. Emer Coleman, Director of Digital Projects at the Greater London Authority, and Tom Steinberg, the founder and director of mySociety, were also suggested as the sort of evangelists and visionaries the Open Data environment badly needs. Indeed, much of the literature argues that it was only through the leadership of characters like Richard Stallman and Linus Torvalds that the Open Source community really took hold (Raymond, 1999; Himanen, 2001).

The second major theory arising from the interview coding was a suggestion of government bureaucracy holding back the pace of development. Each interviewee argued that the government's infrastructure and corporate culture needed to change to make the most of Open Data. Richard Taylor expressed bemusement over bureaucratic loops and inefficiencies he had experienced in his dealings with local councils, while Matthew Somerville highlighted the poor quality and infrequency of Open Data he has worked with in the recent past. Government seems to acknowledge this, with Beth Noveck, in her interview, describing her plans to "move the default towards open" and improve the quality of UK Government data. In July this year, David Cameron expressed his party's intention to maintain "the most ambitious open data agenda of any government in the world," starting with increased publication of health, crime, education and transport data, as well as improved data quality across the board (Cameron, 2011). My sample included a good number of apps from those four areas, suggesting a need or desire amongst developers to use these more diverse datasets. Apps created from these datasets will also have the advantage of clearly relating to citizens' everyday lives, hopefully increasing the uptake of Open Data apps outside of its current user base.

It is easy, given the centrality of the UK and US governments in their countries' Open Data regimes, to assume governments must currently be the most powerful intermediaries in the Open Data environment. This would, indeed, be in stark contrast to previous literature which emphasised the intermediary role played by the formal and informal media (Hindman, 2009; Sunstein, 2007), perhaps suggesting a more hands-on role for governments and local authorities in interacting directly with their citizens. The reality, however, is not so clear cut. Government data is by far the most popular type, but no one government data source stands out as more influential or more widely used than any other. Government data portals such as Data.gov.uk—theoretically a point of centralisation in Open Data

information flows—are frequently bypassed by developers using search engines to locate datasets on the originating departments' or councils' websites. Indeed, if one organisation is to be earmarked as an intermediary in the procurement (and even display, through its popular Maps API) of Open Data, then it must surely be Google – much as it is on the Web in general (Halavais, 2009).

However, taking a step back from 'power' in the traditional sense, and instead considering power as brokerage (Burt, 1992) or social capital (Lin, 2001), new possible locations for intermediation are opened up. Developers with contacts in multiple task-based Open Data 'communities' (such as Matthew Somerville, in my sample) could grow to be seen as power-brokers, as could those organisations publishing the most diversely-used datasets (such as the World Bank or Transport for London). Indeed, with Open Data development proving more task-based than its Open Source predecessor (Kim, 2003; Ghosh, 2005), those organisations who can satisfy the most diverse needs will soon hold the most power across the fragmented Open Data landscape. Corporate data, especially, holds untapped potential in a wide range of applications, both for- and non-profit (Shadbolt, 2011*b*) – an idea which left some of my interviewees excited, and others distinctly uneasy. The take-away is clear however: with more involvement, corporations and NGOs, just as much as governments, could come to shape and define Open Data information flows. And that involvement can only be beneficial for us all.

6 Conclusion

This study, one of the first of its kind, has mapped out the key characteristics of the nascent Open Data sphere. Contrasting Open Data development to the Open Source movement that came before, we have seen the importance of architecture and shared norms in defining interactions, the role of 'hack days' and meeting places in generating communities of practice, and the factors that motivate developers to participate and collaborate in Open Source and Open Data projects. I also discussed the notion of information flows, and set out to map the interaction and data provenance in a non-random sample of British and North American Open Data web apps. By building up a relational database of web apps, developers and data sources, I was able to represent the often intangible relationships between different actors across the Open Data development network. Using additional data on the character and content of each web app, I was able to discern general patterns and unusual cases in what types of data are being used. By combining these findings with qualitative interviews from a purposive sample of key Open Data players—including developers, activists and government advisers—I was able to triangulate actions with motivations to build up a stronger picture of *who* these Open Data participants are, and *why* they participate.

This led to some very useful findings. Despite the media focus on Open Government Data, there is a wider ecosystem of non-governmental datasets in use, including apps dealing with data on health, development, education, crime,

and the environment. Maps surfaced as highly popular means of localising quite abstract data on a familiar scale, although commercial ('closed') map providers like Google Maps proved more prevalent than Open Data alternative OpenStreetMap, perhaps because of the technical knowledge required to use and implement the latter. We saw the huge effect Open Data competitions can have on increasing participation, albeit in somewhat of an artificial community, as opposed to the organic communities which have evolved around government data or mapping data, for example.

The developer and source communities in my sample were organised around tasks or goals—such as monitoring government or mapping local amenities—rather than around the technology of Open Data itself. My interviews confirmed this, with a number of interviewees noting the absence of a general 'Open Data community', but a proliferation of smaller, task-focussed communities. Organisations like MySociety and the Open Knowledge Foundation were suggested as important foci for developer engagement, and could indeed be considered 'intermediaries' given their simultaneous engagement with government, developers and the public.

We also, however, saw the shortcomings of the Open Data system. We noted how the vast majority of apps had only one developer, and the majority of developers worked alone. On a similar note, the majority of apps utilised only one data source, despite the possibilities Open Data offers for combining different datasets to gain new perspectives on the issue at hand. My interviewees closest to government confirmed that the state saw increasing collaboration as one of its key obligations, and it was aiming to do this through releasing more datasets, of a higher quality, more frequently than at present. My other interviewees, however, remained sceptical of how well and how quickly government could adapt to the fundamental change in corporate culture.

Lastly the question of power came to the fore, and I attempted to identify who or what holds the power in the Open Data environment through their brokerage of information flows. Certain central developers and data sources exhibited some of the key hallmarks of intermediaries – a pattern I expect will recur as the Open Data movement matures. In absence of the expected oligopoly of central intermediaries like governments or media multinationals, we instead saw a much more fragmented constellation of task-specific communities and lone individuals, out of which strong intermediaries have yet to form. The real power-brokers will be those developers or sources that begin to interact across these community boundaries. But as it stands, the movement is too nascent and the communities too disparate for such intermediaries to be discernible.

A Appendix: Interview Coding Scheme

Barriers to Participation

- Data.gov.uk is no good for power users
- Fear of upsetting government
- Government doesn't understand Open Data
- Governmental bureaucracy & outdated systems
- Governmental change is too slow
- Hard to know what data is available
- Hard to trace data provenance
- Need for more / better / fresher data
- Nobody sees what you make
- Proprietary / paid-for data
- Semantic web / linked data won't happen
- Very few success stories

Incentives for Participation

- Breaking the law for the greater good
- Changing society / influencing government
- Competitive advantage
- Economic growth
- Empowerment (through owning & using data)
- Getting stuff done (data as a means to an end)
- Helping government do its job better
- Improving corporate reputation
- Increased efficiency
- Innovations taken on by industry / government
- New business models
- Political "pamphleteering" / axes to grind
- "Show them what can be done"
- Transparency / trust
- "We can do it better"
- "Wisdom of the crowds"

The Role of Government in Open Data

- Closing the feedback loop / being more reactive
- Dedicated internal teams to make data available
- Government as data aggregator rather than host
- Government as gatekeeper
- Government as leading the way
- Government as playing catch-up
- Government infrastructure must improve
- Government must listen to user requests for data
- Government must support new projects (eg: MySociety)
- Open Data helps government do its job

The Role of Leaders / Visionaries

- Developer visionaries (eg: Tom Steinberg)
- Government evangelists (eg: Emer Coleman)
- Lack of vision right now
- Need to pressure government from below

Changing Attitudes

- "Bring me a belt to beat you with"
- Data needs to address citizens' needs
- Government attitudes/culture must change
- "Moving the default towards open"
- Perceived extra cost of producing (government) data
- Showing ordinary people how useful data can be
- "We own the government" – the Big Society

Centralisation / Intermediation

- Becoming an intermediary by setting a standard/protocol
- Centralised data portals are good
- Corporate ownership of data is bad
- Google as an intermediary
- Word of mouth – the user as intermediary

Open Data as a Communal/Individual Activity

- "A loose alliance" of disparate contributors
- Communities are based on needs not technologies
- Data is an individual activity
- MySociety / Open Knowledge Foundation as communities
- No *there isn't* an Open Data community
- Yes *there is* an Open Data community

Open Data as a Niche Interest

- "Nerdy" / "Geeky"
- Armchair politicians / Leaving the data to a select few representatives
- You need experts to understand data

Bibliography

Arthur, C. (2010), Analysing data is the future for journalists, says Tim Berners-Lee, *in* 'The Guardian', 22 November 2010. [Online] Last accessed: 19 July 2011.
URL: *http://www.guardian.co.uk/media/2010/nov/22/data-analysis-tim-berners-lee*

Arthur, C. (2011), Data.gov is a project for the few – but they really matter, says Tom Steinberg, *in* 'Technology Blog', The Guardian, 5 April 2011. [Online] Last accessed: 19 July 2011.
URL: *http://www.guardian.co.uk/technology/blog/2011/apr/05/data-gov-steinberg-questions*

Asay, M. N. (2006), Open Source and the Commodity Urge: Disruptive Models for a Disruptive Development Process, *in* D. C. Chris DiBona and M. Stone, eds, 'Open Sources 2.0: The Continuing Evolution', O'Reilly, Sebastopol, CA.

Babbie, E. R. (2010), *The Practice of Social Research*, 12th edn, Wadsworth, Cengage Learning, Belmont, CA.

Barber, B. (1984), *Strong Democracy: Participatory Politics for a New Age*, University of California Press, Berkeley, CA.

Bauwens, M. (2005), 'The political economy of peer production'. [Online] Last accessed: 19 July 2011.
URL: *http://www.ctheory.net/articles.aspx?id=499*

Berdou, E. (2011), *Organization in Open Source Communities*, Routledge, New York, NY.

Berg, B. L. (2009), *Qualitative Research Methods for the Social Sciences*, 7th edn, Allyn & Bacon, Boston, MA.

Berners-Lee, T., Shadbolt, N. and Hall, W. (2006), 'The semantic web revisited', *IEEE Intelligent Systems* **6**, 96–101. [Online] Last accessed: 19 July 2011.
URL: *http://eprints.ecs.soton.ac.uk/12614/1/Semantic_Web_Revisted.pdf*

Bracking, S. (2011), 'Data for Change', *Think Quarterly* **1**. [Online] Last accessed: 19 July 2011.
URL: *http://thinkquarterly.co.uk/01-data/data-for-change*

Brannen, J. (2005), Mixed methods research: a discussion paper, Discussion Paper NCRM/005, ESRC National Centre for Research Methods.

Brooks, F. J. (1995), *The Mythical Man-Month*, Addison-Wesley, Boston, MA.

Bryman, A. (2008), *Social Research Methods*, 3rd edn, Oxford University Press, Oxford.

Burt, R. S. (1992), *Structural Holes: The Social Structure of Competition*, Harvard University Press, Cambridge, MA.

Cameron, D. (2011), 'Letter to Cabinet Ministers on Transparency and Open Data'. 7 July 2011. [Online] Last accessed: 19 July 2011.
URL: *http://number10.gov.uk/news/letter-to-cabinet-ministers-on-transparency-and-open-data*

Castells, M. (2001), Epilogue, *in* P. Himanen, ed., 'The Hacker Ethic', Vintage, London.

Castells, M. (2009), *Communication Power*, Oxford University Press, Oxford.

CGED (1995), On the full and open exchange of scientific data, Executive Report, Committee on Geophysical and Environmental Data, National Research Council, Washington, DC. [Online] Last accessed: 19 July 2011.
URL: *http://www.nap.edu/readingroom/books/exch/exch.html*

Charmaz, K. (2006), *Constructing Grounded Theory*, SAGE, London.

Csikszentmihalyi, M. (1996), *Creativity: Flow and the Psychology of Discovery and Invention*, Harper Perennial, New York, NY.

David, P. A. and Shapiro, J. S. (2008), 'Community-based production of open-source software: What do we know about the developers who participate?', *Information Economics and Policy* **20**(4), 364–398.

Deek, F. P. and McHugh, J. A. M. (2008), *Open Source: Technology and Policy*, Cambridge University Press, New York, NY.

Dekkers, M., Polman, F., te Velde, R. and de Vries, M. (2006), MEPSIR: Measuring European Public Sector Information Resources, Parts 1 & 2, European Commission, Brussels, Belgium.

Deleuze, G. (1992), 'Postscript on the societies of control', *October* **59**, 3–7.

Denzin, N. K. (1970), *The research act: A theoretical introduction to sociological methods*, Aldine, Chicago, IL.

DiBona, C. (2006), Open Source and Proprietary Software Development, *in* D. C. Chris DiBona and M. Stone, eds, 'Open Sources 2.0: The Continuing Evolution', O'Reilly, Sebastopol, CA.

DiBona, C., Cooper, D. and Stone, M. (2006), Introduction, *in* D. C. Chris DiBona and M. Stone, eds, 'Open Sources 2.0: The Continuing Evolution', O'Reilly, Sebastopol, CA.

Dutton, W. H. (2007), 'Through the Network of Networks – The Fifth Estate', *SSRN eLibrary* . [Online] Last accessed: 19 July 2011.
URL: *http://ssrn.com/paper=1134502*

Edwards, G. (2010), Mixed-method approaches to social network analysis, Paper NCRM/015, ESRC National Centre for Research Methods Review.

Feller, J., Fitzgerald, B., Hissam, S. A. and Lakhani, K. R. (2005), Introduction, *in* J. Feller, B. Fitzgerald, S. A. Hissam and K. R. Lakhani, eds, 'Perspectives on Free and Open Source Software', MIT Press, Cambridge, MA.

Finn, H. (2011), 'Lunch With Hal', *Think Quarterly* **1**. [Online] Last accessed: 19 July 2011.
URL: *http://thinkquarterly.co.uk/01-data/lunch-with-hal*

Galloway, A. R. (2004), *Protocol: How Control Exists After Decentralization*, MIT Press, Cambridge, MA.

Ghosh, R. A. (2005), Understanding Free Software Developers: Findings from the FLOSS Study, *in* J. Feller, B. Fitzgerald, S. A. Hissam and K. R. Lakhani, eds, 'Perspectives on Free and Open Source Software', MIT Press, Cambridge, MA.

Ghosh, R. A., Glott, R., Krieger, B. and Robles, G. (2002), Free/libre and open source software: Survey and study, floss, final report, Technical report, International Institute of Infonomics, University of Maastricht, The Netherlands. [Online] Last accessed: 19 July 2011.
URL: *http://www.flossproject.org/report/FLOSS_Final4.pdf*

Grbich, C. (2007), *Qualitative Data Analysis: An Introduction*, SAGE, London.

Guha, R., Howard, M., Hutchison, G., Murray-Rust, P., Rzepa, H., Steinbeck, C., Wegner, J. and Willighagen, E. (2006), 'Blue obelisk – interoperability in chemical informatics', *Journal of Chemical Information and Modeling* **46**(3), 991–998.

Hafner, K. and Lyon, M. (1998), *Where Wizards Stay Up Late: The Origins of the Internet*, Simon and Schuster, New York.

Halavais, A. (2009), *Search Engine Society*, Polity Press, Cambridge.

Hildreth, P. and Kimble, C. (2004), *Knowledge Networks: Innovation through Communities of Practice*, Idea Group Publishing, London.

Himanen, P. (2001), *The Hacker Ethic*, Vintage, London.

Hindman, M. (2007), "Open Source Politics" Reconsidered, *in* V. Mayer-Schönberger and D. Lazer, eds, 'Governance and Information Technology', MIT Press, Cambridge, MA.

Hindman, M. (2009), *The Myth of Digital Democracy*, Princeton University Press, Princeton, NJ.

Hu, Y. (2005), 'Efficient and High Quality Force-Directed Graph Drawing', *The Mathematica Journal* **10**, 37–71. Also available online (Last accessed: 19 July 2011).
URL: *http://www2.research.att.com/~yifanhu/PUB/graph_draw_small.pdf*

Johnson, R. B., Onwuegbuzie, A. J. and Turner, L. A. (2007), 'Toward a definition of mixed methods research', *Journal of Mixed Methods Research* **1**, 112–133.

Jones, P. (2006), Extending Open Source Principles Beyond Software Development, *in* D. C. Chris DiBona and M. Stone, eds, 'Open Sources 2.0: The Continuing Evolution', O'Reilly, Sebastopol, CA.

Kane, P. (2004), *The Play Ethic*, Pan Books, London.

Keohane, R. O. and Nye, J. S. (1998), 'Power and Interdependence in the Information Age', *Foreign Affairs* **77**, 81–94.

Kim, E. E. (2003), An introduction to open source communities, Technical report, Blue Oxen Associates. [Online] Last accessed: 19 July 2011.
URL: *http://blueoxen.com/download/BOA-00007.pdf*

Krippendorf, K. (1980), *Content Analysis: An Introduction to Its Methodology*, Sage, Beverley Hills, CA.

Krummenacher, R., Norton, B. and Marte, A. (2010), Towards linked open services and processes, *in* A. J. Berre, A. Gomez-Perez, K. Tutschku and D. Fensel, eds, 'Future Internet-FIS 2010', Vol. 6369 of *Lecture Notes in Computer Science*, Springer-Verlag Berlin, Berlin, Germany, pp. 68–77. 3rd Future Internet Symposium, Berlin, GERMANY, SEP 20-22, 2010.

Kuk, G. and Davies, T. (forthcoming), 'Assembling open data complementarities for service innovation'.

Lakhani, K. R. and Wolf, R. G. (2005), Why Hackers Do What They Do: Understanding Motivation and Effort in Free/Open Source Software Projects, *in* J. Feller, B. Fitzgerald, S. A. Hissam and K. R. Lakhani, eds, 'Perspectives on Free and Open Source Software', MIT Press, Cambridge, MA.

Lakhani, K., Wolf, B., Bates, J. and DiBona, C. (2002), The Boston Consulting Group Hacker Survey, Release 0.73, Technical report, The Boston Consulting Group.

Lerner, J. and Shankerman, M. (2010), *The Comingled Code: Open Source and Economic Development*, MIT Press, Cambridge, MA.

Lerner, J. and Tirole, J. (2002), 'Some simple economics of open source', *Journal of Industrial Economics* **L**.

Lin, N. (2001), *Social Capital: a Theory of Social Structure and Action*, Cambridge University Press, New York, NY.

Lincoln, Y. S. and Guba, E. G. (1985), *Naturalistic Inquiry*, Sage, London.

Mayer-Schönberger, V. and Brodnig, G. (2001), Information Power: International Affairs in the Cyber Age, Working Paper RWP01-044, John F. Kennedy School of Government.

Mayer-Schönberger, V. and Lazer, D. (2007), The Governing of Government Information, *in* V. Mayer-Schönberger and D. Lazer, eds, 'Governance and Information Technology', MIT Press, Cambridge, MA.

Miles, M. B. and Huberman, A. M. (1994), *Qualitative Data Analysis : An Expanded Sourcebook*, 2nd edn, Sage, Thousand Oaks, CA.

Morse, J. M. (2009), 'Mixing Qualitative Methods', *Qualitative Health Research* **17**, 1523–1524.

Morse, J. M. (2010), 'Simultaneous and Sequential Qualitative Mixed Method Designs', *Qualitative Inquiry* **16**, 483–491.

Murray-Rust, P. (2008), 'Open Data in Science', *Serials Review* **34**, 52–64.

Neuendorf, K. A. (2002), *The Content Analysis Guidebook*, Sage, Thousand Oaks, CA.

Newman, M. E. J. (2010), *Networks: An Introduction*, Oxford University Press, Oxford. (Of particular interest: Chapters 3 & 4).

Nye, J. S. and Owens, W. A. (1996), 'America's Information Edge', *Foreign Affairs* **75**, 20–36.

O'Mahoney, S. and Ferraro, F. (2004), 'Hacking Alone? The Effects of Online and Offline Participation on Open Source Community Leadership'.

O'Reilly, T. (2006), The Open Source Paradigm Shift, *in* D. C. Chris DiBona and M. Stone, eds, 'Open Sources 2.0: The Continuing Evolution', O'Reilly, Sebastopol, CA.

Orszag, P. R. (2009), Open Government Directive, Memorandum for the Heads of Executive Departments and Agencies M-10-06, 8 December 2009, Executive Office of the President, Washington, DC. [Online] Last accessed: 19 July 2011.
URL: *http://www.whitehouse.gov/sites/default/files/omb/assets/memoranda_2010/m10-06.pdf*

Palfrey, J. and Gasser, U. (2008), *Born Digital*, Basic Books, New York, NY.

Polese, K. (2006), Foreword, *in* D. C. Chris DiBona and M. Stone, eds, 'Open Sources 2.0: The Continuing Evolution', O'Reilly, Sebastopol, CA.

Pollock, R. (2009), The Economics of Public Sector Information, CWPE 0920, University of Cambridge.

Raymond, E. S. (1999), *The Cathedral and the Bazaar*, O'Reilly, Sebastopol, CA.

Reynolds, G. (2006), *An Army Of Davids*, Nelson Current, Nashville, TN.

Ritchie, J. and Lewis, J. (2003), *Qualitative Research Practice*, Sage, London.

Robinson, D., Yu, H., Zeller, W. P. and Felten, E. W. (2009), 'Government Data and the Invisible Hand', *Yale J.L. & Tech* **11**, 160–175.

Rose, G. (2007), *Visual Methodologies*, 2nd edn, Sage, London.

Rubin, H. J. and Rubin, I. S. (2005), *Qualitative Interviewing*, 2nd edn, SAGE, Thousand Oaks, CA.

Searls, D. (2006), Making a New World, *in* D. C. Chris DiBona and M. Stone, eds, 'Open Sources 2.0: The Continuing Evolution', O'Reilly, Sebastopol, CA.

Shadbolt, N. (2011*a*), A year of data.gov.uk, *in* 'Data Blog', The Guardian, 21 January 2011. [Online] Last accessed: 19 July 2011.
URL: *http://www.guardian.co.uk/news/datablog/2011/jan/21/data-gov-nigel-shadbolt-government*

Shadbolt, N. (2011*b*), 'Open For Business', *Think Quarterly* **1**. [Online] Last accessed: 19 July 2011.
URL: *http://thinkquarterly.co.uk/01-data/open-for-business*

Silverman, D. (2006), *Interpreting Qualitative Data*, 3rd edn, Sage, London.

Silverman, D. (2010), *Doing Qualitative Research*, 2nd edn, Sage, London.

Souza, B. (2006), How Much Freedom Do You Want?, *in* D. C. Chris DiBona and M. Stone, eds, 'Open Sources 2.0: The Continuing Evolution', O'Reilly, Sebastopol, CA.

Sturges, J. E. and Hanrahan, K. J. (2004), 'Comparing Telephone and Face-to-Face Qualitative Interviewing: a Research Note', *Qualitative Research* **4**(1), 107–118.

Sunstein, C. R. (2007), *Republic.com 2.0*, Princeton University Press, Princeton, NJ.

Surowiecki, J. (2006), *The Wisdom of the Crowds*, Abacus, London.

Torvalds, L. (2001), Prologue, *in* P. Himanen, ed., 'The Hacker Ethic', Vintage, London.

Tuomi, I. (2002), *Networks of Innovation: Change and Meaning in the Age of the Internet*, Oxford University Press, Oxford.

von Hippel, E. (1988), *The Sources of Innovation*, Oxford University Press, Oxford.

von Hippel, E. (2005*a*), *Democratizing Innovation*, MIT Press, Cambridge, MA.

von Hippel, E. (2005*b*), Open Source Software Projects as User Innovation Networks, *in* J. Feller, B. Fitzgerald, S. A. Hissam and K. R. Lakhani, eds, 'Perspectives on Free and Open Source Software', MIT Press, Cambridge, MA.

Wark, M. (2004), *A Hacker Manifesto*, Harvard University Press, Cambridge, MA.

Webb, E. J., Campbel, D. T., Schwartz, R. D. and Sechrest, L. (1966), *Unobtrusive Measures. Non-reactive Research in the Social Sciences*, Rand McNally, Chicago, IL.

Weber, S. (2004), *The Success of Open Source*, Harvard University Press, Cambridge, MA.

Weber, S. (2006), Patterns of Governance in Open Source, *in* D. C. Chris DiBona and M. Stone, eds, 'Open Sources 2.0: The Continuing Evolution', O'Reilly, Sebastopol, CA.

Wenger, E. (1998), *Communities of Practice: Learning, Meaning, and Identity*, Cambridge University Press, Cambridge.

Wenger, E., McDermott, R. and Snyder, W. M. (2002), *Cultivating Communities of Practice*, Harvard Business School Press, Boston, MA.

Yau, N. (2011), Data.gov in crisis: the open data movement is bigger than just one site, *in* 'Data Blog', The Guardian, 5 April 2011. [Online] Last accessed: 19 July 2011.
URL: *http://www.guardian.co.uk/news/datablog/2011/apr/05/data-gov-crisis-obama*