The current push to make public sector information available as raw data in open digital online forms will radically transform research. Discuss, with reference to a number of case studies

Tim Davies (tim@practicalparticipation.co.uk)
September 2010

Interest in 'opening up' public sector information (PSI) on the Internet has been growing in recent years (Aichholzer & Burkert, 2004), gaining a significant boost in Europe with the 2003 PSI Directive (European Commission), and gathering pace in the last twelve months, particularly in the United States and United Kingdom with high-profile central government open data initiatives (e.g. the launch of Data.gov.uk and Data.gov). However, whilst much has been written about the economic impacts of opening access to PSI (Pollock, 2009; Uhlir, 2009), and strong claims have been made about democratic potentials of open data (Mayo and Steinberg, 2007;), consideration of how open data impacts on academia have been limited, with benefits to researchers being little more than a footnote in studies, if noted at all (Sharif, 2009). This paper explores the extent to which the 'revolutionary' potential of open raw PSI could lead to radical transformations in how social scientific research is conducted and communicated, with a specific focus on the UK context.

Given the relative novelty of large quantities of PSI being available as open raw data (Data.gov.uk launched with just 1126 datasets in September 2009, and many of those datasets contained only aggregate statistics rather than raw

data[1]), this paper necessarily takes a theoretical and speculative approach, looking at prior studies and papers in neighboring domains, and drawing upon emerging trends from primary exploratory research and from online resources (grey literature). The paper starts by clarifying definitions of PSI as open, raw data, before exploring relationships between data and research. After considering a number of case studies relating to research with PSI, open data and linked data, preliminary conclusions about the likelihood of radical change in research practice are advanced.

PSI

The OECD define PSI as information *"dynamic and continually generated, directly generated by the public sector, associated with the functioning of the public sector…"* (Vickery & Wunsch-Vincent, 2006), and the European Commission PSI Directive (2003) notes this includes: *"social, economic, geographical, weather, tourist, business, patent and educational information"* (§4). In the EU context PSI has historically been kept solely for use by government and it's agencies (PIRA, 2000; Weiss, 2004), or has been available only in costly printed or proprietary formats, licensed to re-users under restrictive terms via contractual relationships and trading funds (De Saulles, 2005; Pollock, 2009). The MEPSIR study (Deckers et. al., 2006) identifies six key domains of PSI: Business Information; Geographic Information; Legal Information; Meteorological Information; Social Data; and Transport Information. Undoubtedly much of this

[1] Based on analysis of September 25th 2009 database dump of the CKAN platform hosting dataset metadata: http://ckan.net/dump/hmg.ckan.net-20091125.csv.gz (Accessed 12th April 2010)

information is of interest and use to researchers. PSI, defined to exclude personally identifying information, should be distinguished from administrative data and micro-data, which may include disclosive information (Elias & Jones, 2006).

Whilst open data initiatives mean that many PSI datasets are available to the public for the first time, researchers in recognized institutions have long had access to PSI datasets, including through the Administrative Liason Data Service (ADLS) and the UK Data Archive (UKDA). However, the push towards open data introduces two key differences: opening up PSI removes the privileged research access that academics formerly had to this data; much PSI is being placed online in new raw and linked data formats, affecting the research potential of that data. We now turn to an account of what is meant by open data, before looking at different data formats being adopted for sharing open PSI.

Open

Just as advocates of free software have to distinguish between different models of free, often done using Stallman's famous 'Free as in beer vs. Free as in speech' dichotomy, (Stallman quoted in Williams, 2002), so too those exploring 'open data' need to distinguish different forms of openness. Understandings of Open Science, the dominant, though not only, paradigm in the generation of public knowledge since the mid-1660s (David, 2004) stress the communal nature of scientific enquiry, with results shared 'universally' to all *peers* regardless of their personal attributes, and with *all* findings and methods disclosed regardless of

whether a hypothesis was proven or disproven. This is a very specific form of openness between intellectual and organizational peers, developed for the purpose of allowing trust in scientific findings, by facilitating 'organized skepticism' (Merton, 1973 in David, 2004). Open Science has generally not been seen as undermined by boundaries that have the practical impact of restricting access to findings or methods only to those within a given domain of science. By contrast, advocates of Open Access (OA) argue that openness involves the removal of boundaries (e.g. payment of subscriptions, or requirements of institutional affiliation) that prevent any individual from accessing content. The OA movement is integrally an Internet based movement, evident in the Bethesda Statement on Open Access Publishing (2003[2]) that talks of publishing through 'digital media', and the Budapest Open Access Initiative (2002[3]) which discusses convergence between Open Science principles and the affordances of digital networked technologies. However, OA principles such as the Bethesda Statement often impose limitation on re-use of published works, not least restricting production of 'hard copies' to personal use only – a concession of the economic interests of print-based journal publishers. Some OA licenses take a different track, *pushing against* the interests of publishers, and incorporating non-commercial terms in their licenses. OA advocate John Willinsky seeks to provide a progressive compromise position in his 'access principle' (2006) replacing a descriptive definition of OA with an ideological commitment to *"...extend the circulation of [a] work as far as possible and ideally to all who are interested in it*

[2] http://www.earlham.edu/~peters/fos/bethesda.htm#participants - accessed 2nd March 2010
[3] http://www.soros.org/openaccess/read.shtml - accessed 2nd March 2010

and may profit by it." Whilst more conservative than some OA principles (with the cautious caveat 'as far as possible'), Willinsky's principle introduces active encouragement to disseminate open works, rather than simply allowing that the works may be shared.

Open Knowledge arrangements (OKF, 2006), such as those compliant with the Open Knowledge Definition (OKD), move far closer to the 'Free as in speech' model, where open is taken to equate to placing content into, or close to, the public domain. The OKD (Version 1.0) is summarized by the statement *"A piece of knowledge is open if you are free to use, reuse, and redistribute it"*[4] and includes terms concerning access to works, royalty-free redistribution, derivative works, open technical standards without digital rights management, and non-discrimination in licensing of a work.

Talk of open data needs to be clear whether it is dealing with open 'as in Science', 'as in Access', or 'as in Knowledge'. Both open science, and OA environments have struggled with sharing data (as opposed to published analysis of data), for reasons including cultures of data ownership, technology (e.g. the difficulty of archiving persistent data alongside journal articles), and legal restrictions. A number of the advisory council members overseeing the OKD have been involved in articulating explicit 'open data' principles for 'open science' arguing in the Panton Principles (Murray-Rust et. al, 2009) that *"data related to published*

[4] http://opendefinition.org - accessed 2nd March 2010.

science should be explicitly placed in the public domain[5]*"*. Science Commons also encourage a 'public domain', rather than intellectual property based model, for open data.

When PSI is provided as a research input under OKD terms, it is necessarily open access, and, providing published research is similarly open access and open about it's methodology, open science is supported and extended beyond restricted academic domains. Whilst licensing terms of PSI data projects such as data.gov.uk are still under development, and the current 'Crown Database Rights' and 'Crown Copyright' models are somewhere between OA and OKD approaches, for the purposes of research (where misrepresentation of data, prohibited by Crown Copyright, is ruled out by the commitments of scientific practice) we can consider a majority of open PSI as Open Knowledge.

Raw Data

Data is commonly processed between collection and release, and in the case of PSI, data is often released only after considerable manipulation. Some data processing aims correct errors in the dataset and control for measurement errors; other processing prepares data for specific uses: for example, removing disclosive information and converting individual observations into aggregate figures. In such aggregations, information from the original dataset is lost, and certain future uses of the data are precluded. Much of the data released through data.gov.uk is aggregate data, but a number of 'raw' datasets are also available –

[5] http://pantonprinciples.org/ accessed 2nd March 2010.

providing data points for each observation originally taken, rather than aggregates.

The call for "Raw Data" is one that has been popular with technology developers (Robinson et. al. 2010), but which researchers may have a more cautious approach to. Raw data is not always the most useful data (Gray, 2009). Whilst web developers building mapping mash-ups with PSI may not be concerned with occasional outliers and erroneous points in a dataset, social scientific researchers are likely to be concerned to clean and check raw data carefully. However, both developer and researcher will be interested in the data structure – as how data is structured can affect ways it can easily be used. In the case of data.gov.uk, particular effort is taking place, with the involvement in the project of linked data advocates Nigel Shadbolt and Tim Berners-Lee, to publish RDF linked data (Berners-Lee, 2009; Shadbolt et. al, 2006; Alani et. al., 2007). Unlike data tables that use data rows and named columns, RDF models data through 'subject, predicate, object' triples, where any element of the triple may be either a literal string, or a URI. Linked data URIs are dereferenceable. A HTTP request to a URI should return useable data about the subject of that URI: machine readable data when requested. Given URIs could exist anywhere in the Internet, this allows the distributed linking of datasets. When common schema and ontologies are used to model data, and common URIs used to refer the same subjects in multiple datasets, machine assisted information retrieval and processing, and in some cases, logical inference, becomes possible across datasets without requiring that they be integrated first.

The use of raw *linked data* for the provision of open PSI has interesting implications for researchers. Firstly, there is limited familiarity amongst researchers, particularly social science researchers, with the processes involved in using RDF data and querying RDF data stores. This may be expected to change over time, but as Dutton and Meyer (2009) have shown, the development and distribution of digital research skills throughout the social scientific research community can be a slow process, as generations of researchers trained before the advent of certain technologies move through their careers. Secondly, and at a more theoretical level, the choice of data structures and ontologies for modeling raw data can have a profound impact on the questions that can be asked of it (Bowker, 2000). Government data already has certain biases present by virtue of having been collected for governmental, rather than research purposes. These biases can be compounded by choices about semantic structures applied in RDF models, and choices over which fields are treated as mandatory or optional. The third issue for research use of raw linked data is trust. Whilst meteorological and traffic data may be released entirely 'raw', or with minimal transformations applied, much 'social' data will have been heavily processed. In a web of linked data, knowing who has been involved in prior manipulation of data, and the extent to which their processing can be trusted as supporting, rather than undermining, effective social scientific research, is key. However, the trust layer of the semantic web stack (Berners-Lee, 1998) is one it it's most underdeveloped (though heavily researched) aspects, and at present, no clear trust, audit or provenance model is applied to UK open PSI data.

Data and Research

Having access to data is fundamental to carrying out research (Elias, 2007), but, as Cole et. al. note (2008), social scientific data is often expensive or difficult to generate. Secondary research allows analysis of data created by other projects (Carmichael, 2008) and open raw PSI can provide a potentially rich resource for secondary research.

The relationship between researchers and their data varies between fields and disciplines. In some fields, small qualitative datasets gathered and analyzed by a single researcher predominate, whereas in others, vast shared datasets, generated through research technologies and instrumentation, are available for analysis and as inputs for computer aided research by globally distributed research teams. Sawyer (2008) articulates a distinction between 'data-rich fields', such as astrophysics, biology and ecology, and 'data-poor fields', including humanities and social science. He argues that in data-rich fields: *"pooling and sharing of data is expected"* resulting in scholars developing common understandings of the datasets; *"form(s) drive methods"* and the easy availability of data reduces the likelihood of additional data collection during research; and there are *"few(er) theoretical groups"*, as shared data forces out empirically unsupported positions, but at the same time, different positions become more entrenched as their conflicts originate in differential interpretation of shared data, rather than disagreement over the data itself. By contrast, in data-poor fields Sawyer argues that *"data is a prized possession"*; *"access to data drives*

methods"; and there are *"many theoretical camps".* Whilst we might question Sawyer's precise claims, for which he provides little empirical justification, there is clearly a relationship between data availability and research practices. Findings from research into digitization projects underline this. Meyer et al. (2009) report in the evaluation of five digization projects that, as digitized resources became available, a noticeable shift towards quantitative methodologies could be observed in proposed research papers. However, the amount of data availability is not the only thing which affects it's use. The 'contents' of data matters as well – a point missing from Sawyers analysis. Where physical sciences are generally dealing with data about physical phenomena, social scientists are often dealing with data about people, and potentially sensitive data if it's use could facilitate discrimination against groups, or harm to individuals (Solove, 2005). These ethical differences between uses of data in different fields can have significant impacts on how data is handled and used.

Whilst in an ideal process of social scientific work, facts are produced based on rigorous analysis of data, in ways that can be repeated and are open to Popperian falsification, when data is not published along with findings, open science cannot easily be practiced. Latour and Woolgar (1986) have suggested that ideal processes are rarely followed in practice, and thus increasing transparency of research may either require changes to research, or chances to claims about what makes research valid (Busek, 2008).

We now turn to three case studies relating open, linked and raw data to research.

Case Study 1: Traffic Data

One early UK government experiment with publishing open raw PSI was the release, through the Directgov Innovate website, of a single dataset containing locations of cycle accidents across the country between 2005 and 2007 (Clarke, 2009). The data was quickly used to produce mash-up maps showing cycle accident hotspots and to suggest 'safer cycle routes', and has since been cited extensively in discussions of government open data. However, the accuracy of the mash-ups can be questioned. Simply placing markers on a map where accidents have occurred without taking into account the severity of those accidents or drawing on background information about normal traffic levels at accident spots to show statistically significant high levels of accidents, can give a distorted account of where the safer cycle routes are. Whilst datasets providing traffic counts for road across the UK have since been released through Data.gov.uk[6], such data has not yet been combined with the accident counts in any publically available papers or mash-ups.

The STATS19 dataset from which cycle accident data was derived was already available to researchers through the UKDA, so was not new PSI, although it's open release allowing for it, and derived data from it, to be freely redistributed was novel. Past use of STATS19 has found its accuracy to be questionable (Gill et. al., 2006), a finding echoed by cyclists discussing the released dataset in online fora (Cycle Chat, 2009). The dataset relies on reports from police forces, using a

[6] See http://data.gov.uk/dataset/gb-road-traffic-counts (Accessed 12th April 2010)

three-category classification of accidents into fatal, serious and slight injuries, where the last two categories are assigned according to subjective police officer judgments with no test of coder reliability. Whilst no social scientific data is perfect, and data errors are inevitable in any large dataset, the contrast between the standards applied to data gathered as PSI, and then released for research, as opposed to data gathered directly for research must be noted. For example, the DVLA estimate that 11.4% of their vehicle records, and 26.18% of their driver records contain errors (Watson, 2010). Although this is population data, as opposed to sample data, an error rate of over 25% may present problems to many research applications.

Lyons et. al. (2008) have found that more rigorous understandings of traffic accidents can be obtained through the comparison of multiple datasets including hospital admission records. They describe future analysis which draws upon *"pseudonymised data linkage"* and suggest that *"increasing availability of the casualty postcode in the STATS19 data allied to increasing availability of electronic emergency department data should facilitate more accurate large scale police–health linkage studies."* However, in drawing upon potentially disclosive micro-data such as emergency department statistics Lyons et. al. have quickly moved beyond the straightforward use of PSI, and any datasets that result from such work are likely to again be subject to restrictions on distribution rather than to exist as open datasets.

Case Study 2: Clear Climate Code

Our second case studies comes from outside social sciences, and outside academia to explore how advocates of open science and accessible climate science have responded to a long-available data series – the GISS-TEMP (Goddard Institute for Space Studies Surface Temperature Analysis) analysis from NASA[7]. GISS-TEMP uses public domain PSI detailing observations from climate stations across the world, and generates monthly climate models of global temperature using a complex series of COBOL, C and Python programmes. In 2009 a small group of programmers decided to re-implement the GISS-TEMP code purely in python, with the goal of making it easier to understand: increasing the transparency of, and consequently confidence in, the GISS-TEMP results. This Clear Climate Code project[8] has successfully generated accessible code that, drawing on the same data as GISS-TEMP generates equivalent analysis. The project has also led to changes in the NASA GISS-TEMP code, after the re-implementation identified bugs in the original[9]. The existence of both open data, and open source code, allowed actors outside a narrow community of scientists to engage in the practice of open science, and to address wider relationships between science and society.

Case Study 3: Edubase and Educational Research

PSI regarding schools and educational attainment provide another example of data previously available under restricted conditions to researchers, but now widely available through open data initiatives. In particular, the EduBase dataset

[7] http://data.giss.nasa.gov/gistemp/ - Accessed 12th April 2010
[8] http://clearclimatecode.org/ - Accessed 3rd March 2010
[9] http://clearclimatecode.org/finding-bugs-in-gistemp/ Accessed 3rd March 2010

detailing schools across England has been converted into RDF linked data for Data.gov.uk and efforts are underway to link this to a range of further datasets including OFSTED school inspection data and pupil attainment data[10]. This involves creation of, and agreement on, unique identifiers for specific schools, and their reconciliation across government datasets. The modeling of Edubase in RDF also makes linkages with recently released geodata (Ordnance Survey, 2010) that can be linked to statistical information on Wards and other geographical areas. Whilst such linkages may have been achieved in past research, they were only achieved using proprietary tools and datasets. The ability of query across these linked datasets and, using semantic web agents (Brent, 2008) to draw in information from the wider linked web of data (where government minted URIs may become key nodes), introduces new possibilities for social scientific research.

Discussion

Does the release of open raw PSI herald the shift of social science from data-poor, to data-rich, with the resulting changes in disciplinary practice that Sawyer's (2008) theory would predict? And can the creation of large linked semantic social datasets provide the forms of research technology that Collins (1994) suggests could enable social science to become a 'high-consensus, rapid-

[10] Authors notes from presentation by Jeni Tennison, Linked Data advisor to Data.gov.uk at OUCS Linked Data and Practical Semantic Web Workshop, March 2010. http://www.oucs.ox.ac.uk/rts/events/linked-data.xml Accessed 29th March 2010.

discovery' science? Revolutions are easy to predict in a synthetic analysis, but a grounded analysis must be more cautious.

It has already been noted that few open PSI materials were unavailable to researchers in the past, or at least, to researchers interested enough to seek them out. However, this does not mean that providing PSI as open raw data has no impact on the conduct of research. Firstly, as datasets formerly governed by restrictive agreements come to fall under Open Knowledge compliant licenses, any open access research based on this data is now potentially subject to review from a far wider community of peers. As the GISS-TEMP case shows, 'amateurs' from outside the academic field may take advantage of open data to explore the findings of academic study, to identify errors, and to extend, complement or disseminate academic findings. Research into physical science research practice and commercial problem solving finds that openness of both problem statements and data can support discovery of innovative solutions to entrenched problems, with many of the solutions coming from outside the 'home' field of the data or problem (Lakhani et. al., 2007; Lakhani, 2009). Not only may the same phenomena of a 'computational turn' that Meyer et. al (2009) identify from humanities digitization projects occur to some extent within social scientists research based on PSI, but actors from outside the social research fields may become involved in the analysis of social data, bringing different paradigms and approaches to it's analysis. Open PSI repositories are not organized according to academic silos. That geodata or weather data of interest to climate researchers is available for discovery in Data.gov.uk next to schools and education data and

local health statistics has the potential to challenge established research and data silos, much as Meyer and Schroeder (2009) suggest silos of disciplinary knowledge have been challenged by the growth of search based information retrieval across a web of documents. Open PSI does not make social sciences newly data-rich creating data-sharing via Sawyers suggested logic. However, as government is forcing certain datasets to be open it is setting new norms that, alongside requirements by funding bodies for publically funded data to be deposited in data repositories, point towards a more open future for social scientific data. A more open future is not, however, a foregone conclusion. It has already been noted in the first case study that when research draws upon PSI alongside other data sources, it may end up dealing with personal data. Ohm (2009) has shown that even apparently anonymised data can, when combined with other data, lead to the exposure of private information. Against a pressure towards more open data, we may then expect both normative and legal frameworks around privacy to exert an opposite pressure to keep research datasets derived from PSI private in future, for fear of unintended privacy violations.

A second impact of open PSI on research concerns use of secondary data. Cole et. al (2008) note that even minor restrictions, such as delays between registering for data and gaining access can be *"a significant barrier to use"*, and thus we may expect the increasing availability of open raw PSI to increase researchers uptake of data for secondary analysis. Sawyer's (2008) argument would suggest this may result in less primary data collection being carried out, with research

drawing on pre-existing resources. Whilst Heaton (in Carmichael, 2008) suggests the secondary analysis is often useful not as the totality of a project, but as a complement to original qualitative and quantitative research. A number of advocates of data repositories cite their role in teaching: giving students direct experience of real data for secondary analysis (Corti & Bishop, 2005; Cole et. al. 2008), but if training students with skills for secondary analysis displaces the development of skills for primary data collection then a potential knock on impact onto the collection of primary data may be observed. Whilst a full normative analysis of the impact of increased provision of open raw PSI is beyond the scope of this paper, should research become more reliant on government collected information to the detriment of primary research (a possibility compounded by the status of government as both data provider and funding provider in many cases) then the increased nodality (Hood and Margetts, 2007) this would give government in the production of social knowledge does give cause for concern, particularly when one key functional role of social research must be to support public discourse around the success, failure and future direction of government programmes.

Actual changes in the level of the use and citation of open raw PSI are something which will need to be tracked over coming years, an issue complicated in the bibliographic record by the lack of clear and widely agreed standards for data citation (Green, 2009), albeit with some efforts taking place to address this gap (Altman and King, 2007). However, it should be noted that Altman and King's proposed data citation standard relies upon a computer fingerprint of a dataset,

assuming it to be a static collection of quantitative information – rather than a dynamically developing set of RDF statements spread across multiple locations, or a real-time PSI data series. The development then of research technologies for working with open PSI may require some further attention.

For some, mention of RDF linked data encoding social facts holds out the possibility of new computational research technologies, enabling the forms of high-consensus, rapid-discovery science that Collin's discussed (2004). Whilst exploration of RDF and semantic data has been widely explored in academic communities, from biology (e.g. OpenFlyData, Zhao et. al., 2009) to archaeology (e.g. CLAROS, Kurtz et. al., 2009), widespread uptake of *linked data* has not yet been seen in the social sciences, and so the forms of open raw PSI effort described in the Edubase case study point to new potential here. However, the potential of linked data is distinct from the potential of a semantic web – a potential that is widely dismissed in any case by social scientists who reject the idea of uniform ontologies, and argue that multiple paradigms (Kuhn, 1962) are used to describe the same data in different research contexts (Brent, 2008), frustrating semantic computation across datasets. Within a linked data approach, linkages are made not for semantic reasons, but for functional use – and the articulation of relationships between entities in PSI datasets is driven not by research ontologies, but by day-to-day concerns of government administration. Statistical computation across open PSI may allow more rapid and high-consensus discovery of more social facts, but, as Hollingsworth (2008) notes,

theory is required before data can become knowledge, and open PSI promises little in the way of social theory-generating research technologies.

The considerations so far have focused on how the practice of research *within the academic community* may be transformed by open raw PSI. However, the most radical transformations for research may come about through the changing structure of communities of knowledge production brought about by widespread access to both data and tools of data analysis. The growth of open access data visualization tools such as IBM Many Eyes (Viégas et. al., 2007) which allow sophisticated exploration of public datasets increases non-academic engagement with data, as do mash-up projects such as the cycle accident maps of the first case study. Rather than being established through academia controlled peer-reviewed journal articles, widely held social 'facts' can be established through direct appeals to 'transparent' and auditable analysis of data. Of course, without critical appraisal of the data, what appears to be transparent analysis may be anything but – and this highlights a potential epistemic role for the academic researcher – contributing analysis and annotation of open PSI to both directories of data, and to communities where data is being used and manipulated. The cloud of online knowledge (Meyer and Schroeder, 2009) potentially becomes less the result of scholarly work, and more the result of agile and independent data analysis publishing works online. Researcher responses to such shifts are likely to be complex: whilst one possibility is increased researcher involvement in curating open datasets – for example, applying critical statistical analysis skills to convert a dataset from cycle accident incidents, into a measure of accident

liklihood given usually traffic flows – if such activities are not rewarded by academic incentive structures then their occurrence is not likely to be widespread. And as Meyer and Schroeder have found (ibid.), change in research practices generally takes considerable time to diffuse across different disciplines.

Conclusions: norms, technologies, practices and relationships

The current push towards open PSI must be understood in the wider context of open data movements. An academic 'movement' (Frickel and Gross, 2005) for open data, following on from the OA movement, has benefited from open PSI. Government open data initiatives have arguably helped shift norms towards opening access to data, and the normative sustainability of closed-data open-science is being undermined. However, research technologies and academic practices are likely to lag far behind the cutting edge of open, raw, PSI use. Skills and practices take time to diffuse within and across disciplines, acting as a constraint on innovative developers – particular in the absence of funding arrangements or incentive structures to change behavior. The real revolutions are arguably not in research technology or practice, but in the relationship between social science and wider society and the role researchers will need to play to remain relevant as open tools and open data allow a far greater range of actors to offer evidence-based answers to key social questions.

Bibliography

2009. How safe is your route? - Cycle Chat. Available at: http://www.cycle-cafe.net/forums/showthread.php?t=29659&page=4 [Accessed April 12, 2010].

2010. The Data Deluge. *The Economist*.

Aichholzer, G. & Burkert, H., 2004. *Public sector information in the digital age : between markets, public management and citizens' rights*, Cheltenham: Edward Elgar.

Alani, H. et al., 2007. Unlocking the potential of public sector information with semantic web technology. *Lecture Notes in Computer Science*, 4825, 708.

Altman, M. & King, G., 2007. A proposed standard for the scholarly citation of quantitative data. *D-lib Magazine*, 13(3/4), 1082–9873.

Arzberger, P. et al., 2004. Promoting access to public research data for scientific, economic, and social development. *Data Science Journal*, 3(0), 135–152.

Berner, 2009. Putting Government Data online - Design Issues. Available at: http://www.w3.org/DesignIssues/GovData.html [Accessed March 4, 2010].

Berners-Lee, T., 1998. Semantic web road map.

Berners-Lee, T., 2009. *Tim Berners-Lee on the next Web*, Available at: http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html [Accessed November 27, 2009].

Bowker, G.C., 2000. Biodiversity datadiversity. *Social Studies of Science*, 643–683.

Brent, E., 2008. Artificial Intelligence and the Internet. In N. Fielding, R. M. Lee, & G. Blank, eds. *The SAGE handbook of online research methods*. SagePublications Ltd.

Busek, E., 2008. Knowledge and Democracy: Is Freedom a Daughter of Knowledge? In N. Stehr, ed. *Knowledge and democracy : a 21st-century perspective*. Somerset, N.J.: Transaction ; London.

Carmichael, P., 2008. Secondary Qualitative Analysis Using Internet Resources. In N. Fielding, R. M. Lee, & G. Blank, eds. *The SAGE handbook of online research methods*. Sage Publications Ltd.

Clarke, P., 2009. Pedalling some raw data... |. Available at: http://innovate.direct.gov.uk/2009/03/10/pedalling-some-raw-data/ [Accessed April 12, 2010].

Cole, K., Wathan, J. & Corti, L., 2008. The Provision of Access to Quantitative Data for Secondary Analysis. In N. Fielding, R. M. Lee, & G. Blank, eds. *The SAGE handbook of online research methods*. Sage Publications Ltd.

Collins, R., 1994. Why the social sciences won't become high-consensus, rapid-discovery science. In *Sociological Forum*. pp. 155–177.

Corti, L. & Bishop, L., 2005. Strategies in teaching secondary analysis of qualitative data. In *Forum: Qualitative Social Research*.

Crabtree, J. & Chatfield, T., 2010. Mash the State. *Prospect*, (167).

Cummings, J.N. & Kiesler, S., 2005. Collaborative research across disciplinary and organizational boundaries. *Social Studies of Science*, 35(5), 703.

De Saulles, M., 2005. e-Government and the Re-use of Public Sector Information. In *5th European conference on e-government: University of Antwerp, Belgium 16-17 June 2005*. p. 121.

Dekkers, M. et al., 2006. MEPSIR Study - 2006. Available at: http://ec.europa.eu/information_society/policy/psi/actions_eu/policy_actions/mepsir/index_en.htm [Accessed December 21, 2009].

Department for Transport, H., 2008. Reported Road Casualties Great Britain: 2008 - Annual Report. Available at: http://www.dft.gov.uk/pgr/statistics/datatablespublications/accidents/casualtiesgbar/rrcgb2008 [Accessed April 12, 2010].

Dutton, W.H. & Meyer, E.T., 2009. Experience with New Tools and Infrastructures of Research: An Exploratory Study of Distance From, and Attitudes Toward, e-Research. *Prometheus: Critical Studies in Innovation*, 27(3), 223.

Eccles, K., Meyer, E.T. & Madsen, C., 2009. Digitisation as e-Research infrastructure: Access to materials and research capabilities in the Humanities. In 5th International Conference on e-Social Science. Cologne.

Eisen, M. & Salzberg, S., 2009. Open Access: The Sooner the Better. *Science*, 325(5938), 266-266.

Elias, P., 2007. The National Strategy for Data Resources for Research in the Social Sciences. Available at: http://eprints.ncrm.ac.uk/449/ [Accessed April 24, 2010].

European Commission, 2003. *Directive 2003/9 8/EC of Parliament and Council on the re-use of public sector information*, Available at: http://ec.europa.eu/information_society/policy/psi/docs/pdfs/directive/psi_directive_en.pdf.

Evans, J.A. & Reimer, J., 2009. Open Access and Global Participation in Science. *Science*, 323(5917), 1025-1025.

Fensel, D., 2005. *Spinning the semantic Web: bringing the World Wide Web to its full potential*, The MIT Press.

Frickel, S. & Gross, N., 2005. A general theory of scientific/intellectual movements. *American Sociological Review*, 70(2), 204.

Gill, M., Goldacre, M.J. & Yeates, D.G., 2006. Changes in safety on England's roads: analysis of hospital statistics. *British Medical Journal*, 333(7558), 73.

Gray, D.E., 2009. *Doing research in the real world*, Sage Publications Ltd.

Green, T., 2009. We Need Publishing Standards for Datasets and Data Tables. , 22(4), 325-327.

Hollingsworth, J.R., 2008. Introduction to Part 2. In N. Stehr, ed. *Knowledge and democracy : a 21st-century perspective*. Somerset, N.J.: Transaction ; London.

Hood, C.C. & Margetts, H.Z., 2007. *The Tools of Government in the Digital Age* 2nd ed., Palgrave Macmillan.

Jones, P. & Elias, P., 2006. Administrative data as a research resource: a selected audit. *Economic & Social Research Council Regional Review Board Report 43*, 6.

Kuhn, T.S., 1962. *The structure of scientific revolutions*, Chicago ; London: University of Chicago Press.

Kurtz, D. et al., 2009. CLAROS-Bringing Classical Art to a Global Public. In *2009 Fifth IEEE International Conference on e-Science*. pp. 20–27.

Lakhani, K.B., 2009. *Innocentive.com (A)*,

Lakhani, K.R. et al., 2007. The value of openness in scientific problem solving. *Boston: Harvard University*.

Latour, B. & Woolgar, S., 1986. *Laboratory life: The construction of scientific facts*, Princeton Univ Pr.

Lyons, R.A. et al., 2008. Using multiple datasets to understand trends in serious road traffic casualties. *Accident Analysis & Prevention*, 40(4), 1406–1410.

Mayo, E. & Steinberg, T., 2007. *Power of Information Taskforce Report*, Available at: http://www.cabinetoffice.gov.uk/reports/power_of_information.aspx [Accessed December 1, 2009].

Meyer, E.T. & Schroeder, R., 2009. Untangling the web of e-Research: Towards a sociology of online knowledge. *Journal of Informetrics*, 3(3), 246-260.

Murray-Rust, P. et al., Panton Principles. Available at: http://pantonprinciples.org/ [Accessed March 4, 2010].

Newbery, D., Bently, L. & Pollock, R., 2008. Models of public sector information provision via trading funds.

Obama, B., 2010. Memo from President Obama on Transparency and Open Government. In D. Lathrop & L. Ruma, eds. *Open Government*. Available at: http://www.whitehouse.gov/open/documents/open-government-directive [Accessed April 18, 2010].

Ohm, P., 2009. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. *SSRN eLibrary*. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1450006 [Accessed March 4, 2010].

OKF - Open Knowledge Foundation, 2006. Open Knowledge Definition. Available at: http://www.opendefinition.org/ [Accessed March 4, 2010].

Olson, G.M. & Olson, J.S., 2000. Distance matters. *Human-computer interaction*, 15(2), 139–178.

Ordnance Survey, H.O., 2010. News Release: Ordnance Survey launches OS OpenData in groundbreaking national initiative - 01 April 2010. Available at: http://www.ordnancesurvey.co.uk/oswebsite/media/news/2010/April/OpenData.html [Accessed April 12, 2010].

PIRA International Ltd, 2000. *Commercial exploitation of Europe's Public Sector Information*,

Pollock, R., 2009. The Economics of Public Sector Information. Available at: http://econpapers.repec.org/paper/camcamdae/0920.htm [Accessed March 4, 2010].

Robinson, D.G., Yu, H. & Felten, E.W., 2010. Enabling Innovation for Civic Engagement. In D. Lathrop & L. Ruma, eds. *Open Government: Collaboration, Transparency, and Participation in Practice*. O'Reilly Media.

Sawyer, S., 2008. Data Wealth, Data Poverty, Science and Cyberinfrastructure. *Prometheus*, 26(4), 355–371.

Scott, J.C., 1998. *Seeing like a state*, Yale University Press New Haven, CT.

Shadbolt, N., Hall, W. & Berners-Lee, T., 2006. The semantic web revisited. *IEEE intelligent systems*, 21(3), 96–101.

Sharif, R.M., 2009. Maximizing the Value of Public Sector Information for Scientific and Socioeconomic Development in Africa | www.KMAfrica.com - Knowledge Management Africa KnowledgeHub. In KMAfrica. Available at: http://www.isivivane.com/kmafrica/?q=group.governance.maximizing.the.value.of.public.sector.information.for.scientific.and.socioeconomic.development.in.Africa [Accessed April 13, 2010].

Solove, D., 2006. A taxonomy of privacy. *University of Pennsylvania Law Review*, 154(3), 477–560.

Stehr, N., 2008. *Knowledge and democracy : a 21st-century perspective*, Somerset, N.J.: Transaction ; London.

Uhlir, P., 2009. The Socioeconomic Effects of Public Sector Information on Digital Networks: Toward a Better Understanding of Different Access and Reuse Policies: Workshop Summary. Available at: http://www.nap.edu/catalog.php?record_id=12687 [Accessed September 30, 2009].

UK Data Forum, 2009. *UK Strategy for Data Resources for Social and Economic Research 2009-2012*, Available at: http://www.esrc.ac.uk/ESRCInfoCentre/NDS/.

Vickery, G. & Wunsch-Vincent, S., 2006. *Digital broadband content: public sector information and content*, Organisation for Economic Co-operation and Development.

Viégas, F.B. et al., 2007. Manyeyes: a site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics*, 13(6), 1121–1128.

Watson, A., 2010. Statistics on the accuracy of DVLA records - WhatDoTheyKnow.com. Available at: http://www.whatdotheyknow.com/request/statistics_on_the_accuracy_of_dv [Accessed April 1, 2010].

Williams, S., 2002. *Free as in freedom*, O'Reilly Media, Inc.

Willinsky, J., 2006. *The access principle: the case for open access to research and scholarship*, MIT Press Cambridge, MA.

Wuchty, S., Jones, B.F. & Uzzi, B., 2007. The increasing dominance of teams in production of knowledge. *Science*, 316(5827), 1036.

Zhao, J. et al., 2009. OpenFlyData: The Way to Go for Biological Data Integration. In *Data Integration in the Life Sciences*. pp. 54, 47. Available at: http://dx.doi.org/10.1007/978-3-642-02879-3_5 [Accessed April 12, 2010].