Open Data: Infrastructures and ecosystems

Tim Davies
University of Southampton
tim.davies@soton.ac.uk

ABSTRACT
Strong claims have been made for the potential benefits to be derived from government open data initiatives: from open data fuelled economic growth, to stronger democratic accountability and improved delivery of public services (33; 16; 11; 22; 15; 21). Activists have called on governments to free our data(1) and provide raw data now (3). Open data initiatives are conventionally presented with the primary role of governments being to remove the legal and technical barriers that have previously restricted access to data, with any action to realise benefits from that data being driven by actors outside government. However, this paper challenges that representation on two grounds. Firstly, it highlights that a number of significant open data initiatives involve the strategic creation of new datasets and data infrastructures, drawing on, but not solely consisting in, existing government data. Secondly, it argues that the successful realisation of impacts from open data relies on more than the dataset, involving the mobilisation of a wide range of technical, social and political resources, and on interventions beyond dataset supply to support coordination of activity around datasets. Drawing on case studies from the UKs open government data initiatives and the International Aid Transparency Initiative, this paper highlights some of the interventions that may be necessary to support realisation of impact from open data initiatives.

Author Keywords
open data; open government; ecosystems; social science;

General Terms

OPEN DATA INITIATIVES: CONTEXT AND CONCEPTUALISATION

Open government data initiatives are focussed on increasing the online accessibility and re-usability of government data. This involves addressing the public availability of data; the use of open formats and standards to publish data; and the adoption of licensing frameworks which facilitate data re-use (10; 29; 26). Since the launch of Data.gov in the United States in May 2009, over one hundred local and national open government data initiative have emerged across the globe[1], with an increasing number in developing countries (11). Initiatives vary significantly, from being led centrally by a government, or by a particular departments, to those initiatives primarily led by grass-roots campaigners outside of government, pursuing advocacy for more open data. Open data initiatives frequently involve diverse stakeholder groups including bureaucrats in pursuit of policy innovation; transparency activists with an ideological or specific policy interests in openness; technologists interested in the continued computerisation of government; and companies seeking economic gain from open public sector information[2] (8).

The existence of such a broad coalition, including actors from across the political spectrum, is in part enabled by a common rejection of proprietary management of government data, and a common belief that government data can act as a raw material, or platform (30), to build with and upon (c.f. Krikorian and Kapczynskis excellent analysis of similar properties in Access to Knowledge movements (**(year?)**)). The narrative of open data initiatives as focused on unlocking potential, whilst being essentially agnostic about the sorts of potential unlocked (democratic, administrative, economic etc.), allows open data initiatives to secure widespread support. However, this narrative can lead to the implicit or explicit assumption that the potential is a direct property to the datasets being released - and that opening access to specific datasets is in itself the key to unlocking almost unlimited potential. Furthermore, the dataset-centric and use-agnostic nature of open data initiatives drives a focus on data catalogues as the primary governmental output from open data initiatives: with crude counts of datasets published acting as a measure of progress.

This standard model of an open data initiative has a number of serious weaknesses. Firstly, many initiatives that follow it are confronted with frustration from data users who find that existing government datasets dont, in the forms in which they are made available on data catalogues, meet their needs. Secondly, many initiatives face a dearth of high-profile or sustainable uses of the data they release (27). Even when initiatives step in to stimulate demand for data with hack-days and competitions, many datasets remain without visible re-use[3], or with applications built on top of the data failing to make

[1] http://datos.fundacionctic.org/sandbox/catalog/faceted/ Accessed 18th January 2011

[2] In the European context where moves to liberalise the Public Sector Information market have been underway since the early 2000s, Open PSI may sometimes be used as a synonym for open data

[3] Though I have argued this at least some of the reasons that data re-use is not made visible is because initiatives have adopted a narrative of data for developers, failing to recognise, and invite reports of data use, from non-developers (8).

WebSci12,

it beyond the prototype stage to become scaleable businesses or public services (20). These may be labelled transitional problems that will be resolved as technologies are developed to better manage data, and as developers and companies become more aware of the potential of open data (c.f. Pollocks claim that We are still at the beginning in (25)). However, this paper argues that these gaps between the promise and reality of open data should encourage us to find a richer conceptualisation of what successful open data initiatives look like. By looking beyond the surface narratives to see how many open data projects initiatives in practice we can see that they commonly involve more than the simple release of datasets. They frequently require efforts to establish rich new open data infrastructures and standards, and to actively encourage the development of an ecosystem of open data use. This richer conceptualisation may lack the simplicity (and political attractiveness) of the standard dataset-centric model, but its articulation can better guide political and technical decision making in the development of successful open data initiatives.

OPEN DATA INITIATIVES IN ACTION: STANDARDISATION AND MOBILISATION

Robinson et. al. argue that the role of government should be providing a simple, reliable and publicly accessible infrastructure that exposes underlying [government] data (35, p.161), with any use of that data left entirely to [p]rivate actors, either nonprofit or commercial. In many open data initiatives, the data to be exposed is not some single pre-existing dataset, but consists in many disparate datasets from different agencies or authorities, drawn from different legacy systems in different formats. As a result, providing a simple and reliable infrastructure involves more than the release and aggregation of those datasets.

For example, the UK government has required local authorities to publish all their spending transactions above 500 online as open data (7). Whilst there are possible uses of this data in its raw form as initially published by authorities in a collection of monthly spreadsheets or PDF files on their own websites (e.g. simple citizen fact-finding that alters the balance of power (8); and altering the incentives of decisions makers (34)), most uses of this data, whether democratic, administrative or capitalist, require some sort of standardised dataset bringing together data from different local authorities. Although the minimal publishing of existing datasets allows third-parties to fill the gap between supply and demand by creating aggregation services (such as OpenlyLocal.com which aggregates local authority spending data, manually creating scrapers to deal with differences in the incoming data, or the ambitious OpenSpending.com project which provides tools for mapping spending data to a common model), open data initiatives generally lead to pressures for the creation of new data standards, and ultimately, new components of the data apparatus of the state. In the case of spending data, this has involved the creation of guidance for authorities that sets out a standard set of fields for spend data and advocates the use of established third-party categorisations of data as opposed to internal codes(24), as well as the development of a government-supported pilot to aggregate spending data as linked data[4]. The launch in Spring 2012 of a UK Government Standards Hub[5] to select or develop open data and software standards for government underlines the growing awareness that making (open) data work effectively requires standardisation efforts.

Whilst Robinson et. al. (35) argue that governments should solely focus on the provision of data, having no role in the creation of tools for working with data, in practice many open data initiative do involve either the direct creation of tools or services for working with data, or action to catalyse tool creation through hack days and app competitions (for example, the Rewired State events in the UK, Apps for Democracy project in the US, and the World Banks various open data app competitions). This may be because the open dataset is, in effect, a new dataset from which government itself can benefit, and thus internal-focussed tools are created, or because of a need to have tools that visibly demonstrate the impacts of sharing open data in order to secure ongoing political support for open data initiatives.

In the International Aid Transparency Initiative (IATI) action to catalyse data use has been particularly important. IATI is a multilateral open data initiative seeking to make data on aid activities for government and multilateral aid donors available as open data using a common XML standard (17; 18). IATI involves both a political process to secure a commitment from donors to publish their data to a common standard, and a technical work-stream establishing an infrastructure for access to the published data. Without tools that help visualise and present published IATI data, donors find it hard to see the impacts of their efforts to make the data available, but without good coverage of aid flows in the IATI open datasets (something that will only occur when a critical mass of donors publish IATI data) private actors have limited incentives to provide such tools[6]. As a result, between the release of version 1.0 of the IATI XML standard in early 2011, and the Fourth High Level Forum on Aid Effectiveness in Busan, Korea in November 2011 when politicians would be making key decisions about support for the initiative, partners in IATI put considerable effort into ensuring not just the data, but tools, convertors and visualisations for working with it were available. This process not only helped mobilise political support for the initiative, but in the process of using the data, actors involved in IATI were able to identify additional assets that data users may need beyond simply the provision of datasets, such as meta-data in the IATI Register data catalogue to indicate when data was last updates, and API access to code-lists to support interpretation of the available data.

Some government datasets (e.g. core reference datasets such as geo-data) are likely to see immediate demand when released as open data. For many others the picture is more complex. The dynamics of shifting from proprietary management

[4]ESD Toolkit Council Spending Linked Data Pilot - http://doc.spending.esd.org.uk/Spending/Default.aspx Accessed 18th January 2012

[5]http://standards.data.gov.uk - Accessed 26th January 2012

[6]In Robinson et. al.s Invisible Hand model, this might be seen as temporary market failure, but within the political arena its a failure that could significantly set back or undo an open data initiative.

of isolated and overlapping datasets across government to the creation of simple, reliable and publicly accessible infrastructures of open data mean that simplistic free our dataset models of open data is likely to be ineffective in driving the sorts of economic growth and political accountability hoped for from open data. However, this should neither lead to a rejection of the move towards open data, for which a powerful normative case remains, but it should lead us to think more carefully about the different interventions involved in successful execution of an open data initiative. In the following section we develop in more depth the notion of infrastructure building, and we introduce the idea of fostering an open data ecosystem to help identify and evaluate possible strategies that government and non-government open data initiatives can adopt in seeking the realisation of the promised benefits of open data.

INFRASTRUCTURES AND ECOSYSTEMS

The deployment of twin concepts of infrastructure and ecosystem to describe technical and sociotechnical systems has widespread precedent (e.g. (9; 23)). This pairing allows a distinction to be made between infrastructure as the basic physical and organisational structures and facilities needed for the operation of a society or enterprise (31) (often centralised, standardised and managed by some small set of agents), and the emergent, autonomous and self-organising components of an ecosystem, linked together in local and global feedback loops and developing according to local specialisations and adaptation rather than top-down design. Whilst some computing and web science uses of these terms might move beyond their use as metaphors, focussing on the development of technological systems that directly mimic biological ecosystems (6), or using infrastructure to refer to some domain-specific artefacts, this paper draws upon infrastructure and ecosystem as metaphors. This supports the identification and distinction of different activities and artefacts associated with an open data initiative; and it provides a linkage point connecting discussions of open data with existing historical and normative political, legal and social writings on infrastructures and ecosystems.

Figure 1. is a representation of the infrastructure and ecosystem of artefacts already created and under development as part of the International Aid Transparency Initiative. The infrastructure (in the grey box at the bottom of the diagram) consists in those technical components and processes which the core Initiative is taking direct responsibility for, whereas the ecosystem is made up of a series of interrelated tools and services that rely on one or more elements of the infrastructure either directly, or through intermediary tools and services, for their sustained operation. The infrastructure involves more than raw datasets: it involves providing metadata about the datasets to allow users to identify recently updated data and to easily find data relating to a particular country, or from a specific aid door; as well as providing progra-

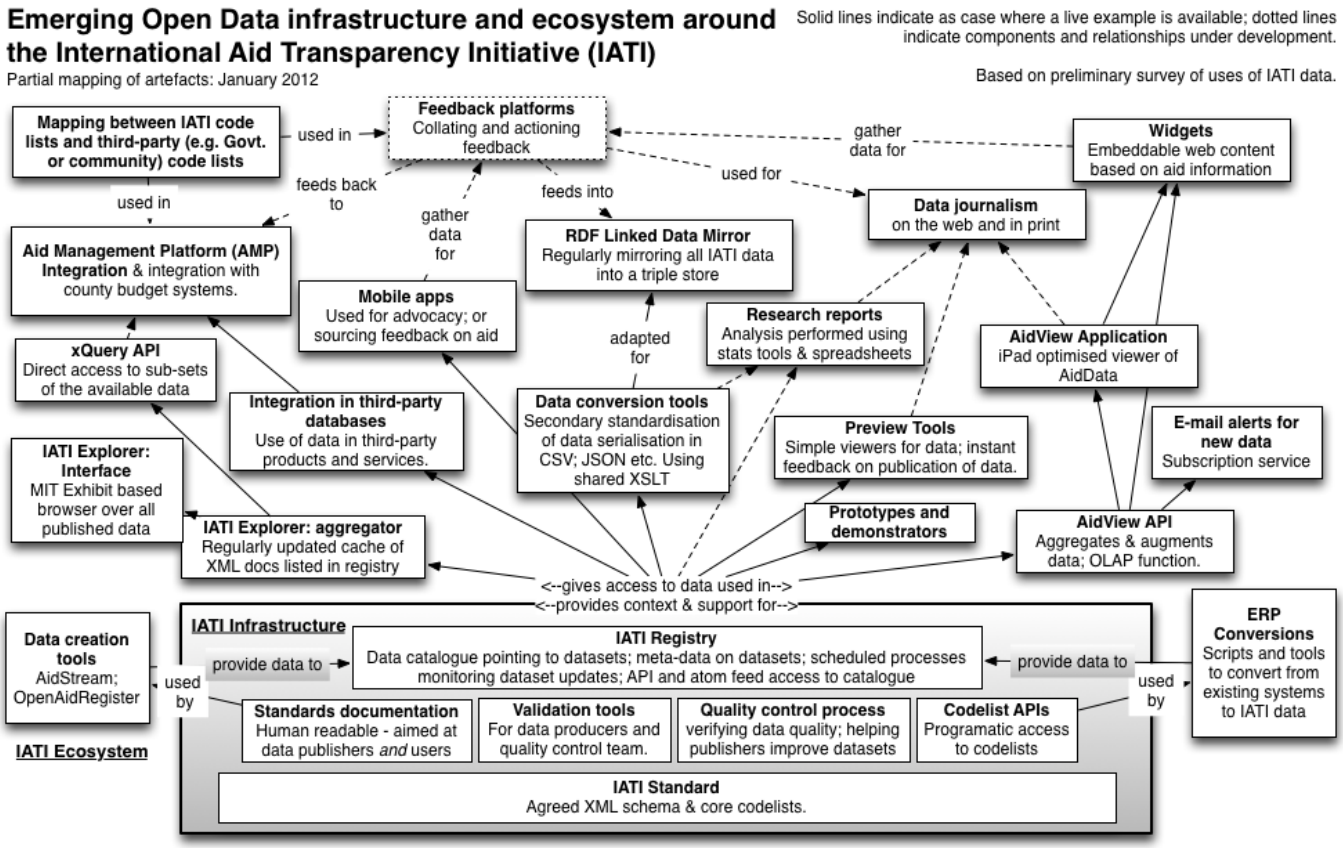

Figure 1. Schematic representation of International Aid Transparency Initiative Infrastructure and Ecosystem.

matic access to contextual reference data such as codelists. Building a stable infrastructure has also involved developing automated validation tools and a manual data quality control process that can help donors publishing open aid data to ensure their data does comply with the standard. If the infrastructure of IATI were simply the raw datasets, each user of the data would need to take care of key tasks of cleaning up the data, interpreting the XML standard to identify where datasets differ from it and choosing whether or not to add special case handling to manage quirks in incoming data, or whether to report errors to the dataset owner to ask for corrections. Placing a responsibility for data quality and context at the infrastructure level of an open data initiative avoids significant duplication of the groundwork effort needed before open data use can even get started.

The figure points, however, to some early duplication in the IATI ecosystem layer: both the AidView API and the IATI Explorer which draw directly on the registry of available data, are based on instances of the eXist XML Database (28) which work by aggregating together IATI datasets and then providing some form of API access to the data (one based on a RESTFUL API with pre-determined parameters running custom queries; the other offering generic xpath and xquery access to a store of IATI XML documents on aid projects); and scripts have been written to convert IATI datasets relational databases, providing an API layer on top of these. This might be seen as the sort of duplication that was prevented by the focus on a rich open data infrastructure, and could be taken to suggest API access to the IATI data should be part of the core infrastructure as well. However, on closer investigation, each API serves a slightly different purpose and exercises different selection criteria in the data it includes. For example, the AidView API performs currency normalisation to data and pre-computes OLAP Cubes of summary data (12); IATI Explorer provides access to the raw data without any conversion - acting primarily as a convenience layer for users who want to fetch information from across multiple datasets. These components of the ecosystem (that come to be relied upon by other data uses) specialise, supporting a particular subset of the full range of possibilities that the open data can support. The infrastructure seeks to support the widest possible range of uses of the data, whilst the ecosystem allows niche uses to emerge.

INTERVENTION IN ECOSYSTEMS

Should open data initiatives intervene in the ecosystems that develop on top of the open data infrastructures they provide? Robinson et. al.s (35) argument would suggest the answer is no: governments should provide open data infrastructures and then get out of the way. Contrary to this, the authors experience with IATI suggests that open data initiatives can benefit from greater co-operation and interaction between the core team of an open data initiative, and actors developing the ecosystem. Firstly, an effective infrastructure requires ongoing resources, and showing visible impacts of data is important to secure those resources. In the early stage of an initiative, direct investment, or offering active support to those who can develop uses of the data may be important. Secondly, users of data will generally seek to convert data into the formats they are most comfortable working with. This leads to lots of alternative formats and ways of representing the data. Actively engaging with this as a process of secondary standardisation, helping in the development of different serialisations of a dataset, can address possible fragmentation of the ecosystem. In the case of IATI, this has involved establishing a collection of open source XSLT (XML Style Sheet Transforms) for converting IATI XML into a range of formats[7]. Thirdly, actors from an open data initiative can work to ensure elements of the ecosystem are visible and shared. For example, by providing space for sharing source code and scripts for working with the open datasets, and linking to this from the data catalogues where new users will start accessing the data. Without this, many potential users of open data are likely to find it hard to identify tools within the ecosystem that they could use, or that have already added value to data that they could benefit from. This intervention may be supported by improved design of open data catalogues to facilitate display of downstream uses of data, and the development of a user-friendly provenance infrastructure for open data to enable better tracking of open data re-use[8]. Fourth, an open data initiative may wish to provide or broker investment and support to ensure key areas of an ecosystem can be made sustainable. As Figure 1. shows, in just a short period of time, elements of the IATI ecosystem have emerged that depend upon other elements. If a tool like the AidView API were to cease to exist, then other tools and uses of the data would be negatively impacted. Elements within an ecosystem can adopt a number of sustainability strategies themselves, from charging for the value they add to a dataset, to operating as grant or voluntary-funded public goods. Where ecosystem elements adopt a proprietary strategy for their sustainability, the range of uses of the value added data may be limited, and so in the interests of promoting the maximum possible uses of an open data infrastructure, offering direct support to some elements of the ecosystem my be a legitimate action for an open data initiative to take.

The strategies that open data initiatives will need to adopt to engage with the autonomous actors in an open data ecosystem are different from those that can be employed in developing in infrastructure. They need to be based on actors from the core of an initiative operating as equal partners in developing the ecosystem, not as top-down directors. This requires community building and facilitation skills, and may require significant culture change for some government data owners. It is possible that in some contexts, key roles developing an ecosystem around an open dataset will be taken up entirely by third-parties, but initiatives may still have a role critically reviewing the ecosystem that develops to ensure the full range of possible data uses, democratic as well as economic, are supported.

[7] Using XSLT makes the mapping between the IATI XML standard and other formats available in a wide range of programming languages that have XSLT support

[8] See the emerging suite of Provenance recording and access standards under development at W3C http://www.w3.org/TR/prov-primer/ for promising work in this area.

PRACTICAL AND NORMATIVE METAPHORS

Infrastructure and ecosystem are not only descriptive metaphors. Economists, historians, complexity scientists, social scientists and political theorists (amongst others) have all applied time to thinking practically and normatively about these concepts, and their insights may prove germane in exploring open data. To give just a few examples, economists have established models for exploring state investment into public goods like street lights and law enforcement, or for managing private financing, and state regulation, in open infrastructures like the telecommunication network. Such models can support a rational assessment ofwhere open data infrastructures are best developed as public goods, or where market and regulation based approaches to ensure their provision are appropriate. This might be contrasted with current UK decision making around Public Sector Information (PSI) which Saxby argues is based more on legacy compromises than clear strategy and economic analysis (36; 37; 38). Alternately, we can look to historians and political scientists who have traced the way in which national infrastructures have moved in and out of the public domain over time, and how this has impacted upon national commercial and democratic ecosystems. For example, road, rail, communication, and even mapping infrastructures often have their genesis in military needs, developed to enable state control of a territory, or planned to reward certain political constituencies (2; 14). They impact upon trade and upon which areas of a country have access to political power. Such cases will find analogies in government datasets: highlighting that many of the legacy datasets out of which open data infrastructures will be built are significantly non-neutral (39; 5), and drawing attention to power dynamics that may be implicit in the datasets being made open. This consideration is no less relevant in the UK or US than in open data initiatives emerging in Kenya, Ghana or Moldova. Ecological and complexity science thinking about how systems of dependent data use will respond to shocks in supply of their inputs can help us think about the sustainability of the open data ecosystems that are being developed. Many other disciples will be able to bring questions and contributions to our understanding of open data through the linking metaphors of infrastructure and ecosystem.

For web science, concerned with both the stability of the technical architectures of the web, and normative considerations of building a pro-social web (13; 4; 40), the challenge of integrating such insights with a technical understanding of how data flows between infrastructure and ecosystem components is a key challenge for future research.

CONCLUSIONS AND FUTURE RESEARCH

Limor Peer, discussing academic institutional open data repositories, has argued that open data requires effort, investment of resources, and planning. By itself, it does not enhance knowledge. (32). This paper echoes Peers finding, and develops a framework for thinking about the different sorts of effort and investment required. By looking behind generalised claims about open data to explore a number of high profile initiatives in detail, it has argued that we must go beyond the dataset if we are serious about the realisation of social, political and economic value from open data. Showing that open data initiatives split into two parts: the infrastructure that supports the widest possible range of uses, and the ecosystem in which specialised communities of use emerge, it has shown that open data initiatives frequently need to adopt strategies of co-ordination to provide a stable infrastructure, and strategies of collaboration and community building to support the mobilisation of the political, social and technical resources that will see enable open data to drive diverse impacts.

References

Arthur, C., and Cross, M. Give us back our crown jewels, 2006.

Beniger, J. R. *The control revolution: Technological and economic origins of the information society*. Harvard Univ Pr, 1986.

Berners-Lee, T. Tim Berners-Lee on the next Web, Feb. 2009.

Berners-Lee, T., Weitzner, D. J., Hall, W., O'Hara, K., Shadbolt, N., and Hendler, J. a. A Framework for Web Science. *Foundations and Trends in Web Science 1*, 1 (2006), 1–130.

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

Briscoe, G., and De Wilde, P. Digital ecosystems: evolving service-orientated architectures. 2009.

Cameron, D. Letter to Government departments on opening up data — Number10.gov.uk, May 2010.

Davies, T. *Open data, democracy and public sector reform: A look at open government data use from data. gov. uk*. Practical Participation, 2010.

Dini, P., Rathbone, N., Vidal, M., Hernandex, P., Ferronato, P., Briscoe, G., and Hendryx, S. The Digital Ecosystems Research Vision : 2010 and Beyond. 2005.

Eaves, D. Three Laws of Open Data (International Edition), Nov. 2009.

Gigler, B.-S., Custer, S., and Rahemtulla, H. Realizing the Vision of Open Government Data: Opportunities, Challenges and Pitfalls (Long Version). Tech. rep., Open Development Technology Alliance, 2011.

Gray, J., Chaudhuri, S., Bosworth, A., Layman, A., Reichart, D., and Venkatrao, M. Data cube: A relational aggregation operator generalizing group-by, cross-tab, and subtotals. *Data Mining and Knowledge Discovery 53* (1997), 29–53.

Halford, S., Pope, C., and Carr, L. A Manifesto for Web Science. In *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line* (2010), 1–6.

Hewitt, R. *Map of a Nation: A Biography of the Ordnance Survey* . Granta Books, 2010.

HM Government. *Open Public Services White Paper*. HM Government, 2011.

Hogge, B. Open Data Study. 2010.

IATI. International Aid Transparency Initiative: Accra Statement. Tech. Rep. September 2008, International Aid Transparency Initiative, 2008.

IATI. International Aid Transparency Initiative Concept Note - April 2009.

Krikorian, G., and Kapczynski, A. *Access to Knowledge in an Age of Intellectual Property*. Zone Books, 2010.

Kuk, G., and Davies, T. The Roles of Agency and Artifacts in Assembling Open Data Complementarities, 2011.

Kundra, V. Digital Fuel of the 21st Century: Innovation through Open Data and the Network Effect. 2012.

Lathrop, D., and Ruma, L. *Open Government: Collaboration, Transparency, and Participation in Practice*. O'Reilly Media, 2010.

Li, T. Services ecosystem: towards a resilient infrastructure for on demand services provisioning in grid. In *Proceedings. IEEE International Conference on Web Services, 2004.*, IEEE (2004), 394–401.

Local Government Group. Local transparency : a practitioners guide to publishing local spending data. Tech. Rep. November, Local Government Group, with the Local Public Data Panel, London, 2010.

Maguire, S. Can Data Deliver Better Government ? *The Political Quarterly 82*, 4 (2011).

Malmud, C., O'Reilly, T., Elin, G., and And others. See website for full list. 8 Principles of Open Government Data, Dec. 2007.

Mcclean, T. Not with a Bang but a Whimper The Politics of Accountability and Open Data in the UK. 2011.

Meier, W. eXist : An Open Source Native XML Database. 169–183.

OKF - Open Knowledge Foundation. Open Knowledge Definition, 2006.

O'Reilly, T. Government as a platform. 1 ed. O'Reilly Media, Feb. 2010.

Oxford English Dictionary. "infrastructure n.". In *OED Online*. Oxford University Press, 2011.

Peer, L. Building an Open Data Repository: Lessons and Challenges. 2011.

Pollock, R. The Economics of Public Sector Information, May 2009.

Prat, A. The more closely we are watched, the better we behave? In *Proceedings of the British Academy*, British Academy (2006).

Robinson, D., Yu, H., Zeller, W., and Felten, E. Government data and the invisible hand. *Yale Journal of Law & Technology 160* (2009), 160–175.

Saxby, S. The development of UK government policy towards the commercialization of official information. *International Journal of Law and Information Technology 4*, 3 (Dec. 1996), 199–233.

Saxby, S. UK Public Sector Information and Re-Use Policy - A 2008 Analysis. *Access to Public Sector Information: Law, Technology & Policy 1* (July 2010).

Saxby, S. Three years in the life of UK national information policythe politics and process of policy development. *International Journal of Private Law 4*, 1 (2011), 1–31.

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

Shadbolt, N. Research Roadmap: Fundamental Research Questions and Perspectives in Web Science, 2008.