Linked Data and Government

PAUL MILLER, The Cloud of Data, UK

Abstract

Governments around the world are increasingly moving to open up access to non-personal data, and high-profile examples of this trend include the United States' data.gov and the Canadian city of Vancouver. Building upon - and moving beyond - this trend is the broader concept of Linked Data. Built upon established specifications from the Semantic Web community and strongly backed by Web inventor Sir Tim Berners-Lee, Linked Data offers a vision in which public sector information is unambiguously published to the Web in a manner that makes it easy for conforming systems elsewhere to incorporate - and understand - the data.

Key words

Linked Data, Open Data, Open Government, Semantic Web

About the Author

Paul Miller [Paul Miller, Cloud of Data] is an independent consultant, focused upon helping clients to realise the benefits of emerging trends such as Cloud Computing and the Semantic Web. Previous positions have included a senior role at UK technology company Talis, and the post of Director at the Common Information Environment, a consortium of UK public sector organisations including the BBC, the British Library, the Joint Information Systems Committee (JISC) and the National Health Service (NHS). He has served in a variety of advisory and monitoring capacities, including the Executive Committees of the Dublin Core Metadata Initiative (DCMI) and Consortium for the Computer Interchange of Museum Information (CIMI).

Paul writes for ZDNet on the Semantic Web [The Semantic Web | ZDNet], and routinely presents and chairs panels at industry events. He holds a Doctorate in Archaeology from the University of York.

Introduction

A wealth of valuable data is collected and stored in the systems of complex organisations such as Europe's public sector institutions, frequently underutilised for a multitude of reasons from institutional inertia to technological complexity. As budgets contract and competitive pressures increase, the timely and effective exploitation of data is becoming an increasingly important characteristic of the successful organisation; and the public sector is certainly no exception. From efficiently transparent reporting to data-driven internal decision making and the cost-effective nurturing of new avenues for growth, collaboration or differentiation, there is increasing value in effectively exploiting data to further the institutional mission.

World Wide Web inventor Sir Tim Berners-Lee[1] and others compellingly describe the value in moving from today's 'Web of Documents' toward a 'Web of Data' in which much of the data we already curate is made available – via the simple architecture and technologies of the Web itself – for manipulation by computers. Pages on the web meant for reading by people, they argue, would gain structure such that whilst you or I might *read* a postal address off the screen as we do today, software would see the same page and offer to calculate a route to that address, add it to your address book, provide a synopsis of local responses to the last census, and more. Relevant data from your institution would be available alongside that from other bodies, powering a range of applications for staff, citizens, monitoring agencies, industrial partners and more; the value locked up inside institutional systems would be made available to drive efficiency in today's procedures whilst creating the opportunities for tomorrow's.

This vision of a 'Semantic Web[2]' has been discussed for years[3], but a combination of political and commercial will, community readiness, technological capability and openly available data has led to a recent leap forward[4] in adoption of one particular aspect of that vision under the banner of Linked Data[5].

This concept of Linked Data is attracting attention in quarters unfamiliar with the formal Semantic Web community from which it emerged. Recent announcements[6] from the UK's previous Prime Minister saw his Government join existing implementers as diverse as the BBC, Thomson Reuters, Tesco, Best Buy and Johnson & Johnson, and local, regional and national governments elsewhere are increasingly seeking to understand the same opportunity.

Four simple principles, or rules, laid down by web inventor Sir Tim Berners-Lee describe the practicalities of Linked Data, and implementers have been quick to apply these in exposing large collections of data for use and reuse, facilitated by the underlying structure of the web itself. In a world in which no single database is comprehensive, the value of being easily able to link related assertions from across diverse data silos is proving compelling.

---

1. http://en.wikipedia.org/wiki/Tim_Berners-Lee
2. http://en.wikipedia.org/wiki/Semantic_Web
3. http://www.scientificamerican.com/article.cfm?id=the-semantic-web
4. http://www.ted.com/talks/tim_berners_lee_the_year_open_data_went_worldwide.html
5. http://en.wikipedia.org/wiki/Linked_data
6. http://webarchive.nationalarchives.gov.uk/+/http://www.number10.gov.uk/Page22218

Linked Data

The concept of Linked Data has been embraced by a particular set of the Semantic Web's enthusiasts and by a growing cohort of potential beneficiaries, predominantly those active in research, media or government. From modest beginnings, Richard Cyganiak's *Linking Open Data Cloud* diagram7 now represents over 13 billion RDF statements8 from across a growing network of participating sites. This diagram only scratches the surface, in all likelihood missing a number of poorly publicised resources as well as the related work being done behind the firewalls of organisations such as pharmaceutical giant Johnson & Johnson. The 3,000 datasets9 currently catalogued on the UK Government's data.gov.uk site are also amongst those not yet represented, having been published since the diagram was last updated.

Tim Berners-Lee's Linked Data Principles

As web inventor and W3C Director Sir Tim Berners-Lee notes in his *Design Issues for Linked Data*10,

"The Semantic Web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other, related, data."

This straightforward realisation is expounded in a set of four deceptively simple 'rules' or (as Berners-Lee prefers) 'expectations of behaviour.' Ultimately these lie behind everything that might be described as Linked Data, whether out on the open web for all to see, or locked away in a Computer Science laboratory or behind the firewall at a Pharmaceutical company or Bank.

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
4. Include links to other URIs, so that they can discover more things.

Whilst the exact wording of these statements has changed slightly since first expressed in 2006, and there remains some question as to the strength of the requirement for specific standards, the acronyms mask a simple yet powerful set of behaviours;

1. Name objects and resources, unambiguously;
2. Make use of the structure of the web;
3. Make it easy to discover information about the named object or resource;
4. If you know about related objects or resources, link to them too.

---

7. http://richard.cyganiak.de/2007/10/lod/
8. http://esw.w3.org/TaskForces/CommunityProjects/LinkingOpenData/DataSets/Statistics
9. http://data.gov.uk/data/all
10. http://www.w3.org/DesignIssues/LinkedData.html

There is a widely held presumption amongst many of Linked Data's most persuasive advocates that the standards (such as RDF11 for modelling and syntax or SPARQL12 for querying) referred to by Berners-Lee are prerequisite for sharing or consuming Linked Data. Whilst the power of these standards delivers the richest set of capabilities today – with every indication that tool development and the ongoing standardisation process will increase this still further – there is also value in a more permissive reading of Berners-Lee's rules.

There is much to gain in embracing the philosophy behind these rules, separately to adopting the standards and specifications required to realise their full potential. Unambiguous identification of resources across the web, easily parsable descriptive information, shared terminologies comprising web-addressable terms, and unambiguous links to related resources deliver real value, as do microformats13, RDFa14 markup in web pages and other 'simpler' approaches. Whether this leads towards 'Linked Data' in a formal sense or not perhaps remains unclear, yet may ultimately prove unimportant.

'Linked Data' and 'Open Data'

There is some confusion evident in the way that the terms 'Linked Data,' 'Open Data,' and 'Linked Open Data' are used, often almost interchangeably. The early Linking Open Data project15 did much to exacerbate this trend, as it grew beyond its original scope to embrace data that were not technically Open.

For clarity, 'Linked Data' should normally be presumed to respect Berners-Lee's four rules. 'Open Data' is harder to pin down with precision, but could usefully be considered to cover data respecting the terms of the Open Knowledge Definition16. This definition comprises 11 clauses providing detail around the core premise that 'open' data should be freely available online for use and re-use. A number of licenses have been found to be conformant with the Open Knowledge Definition, and should be used where feasible in order to unambiguously assert that data are being made available for re-use.

Linked Data *may* be Open, and Open Data *may* be Linked, but it is equally possible for Linked Data to carry licensing or other restrictions that prevent it being considered Open, or for Open Data to be made available in ways that do not respect all of Berners-Lee's rules for Linking. In order to avoid confusion, the terms 'Linked' and 'Open' should be used specifically and with care.

---

11. http://en.wikipedia.org/wiki/Resource_Description_Framework
12. http://en.wikipedia.org/wiki/SPARQL
13. http://en.wikipedia.org/wiki/Microformat
14. http://en.wikipedia.org/wiki/RDFa
15. http://esw.w3.org/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
16. http://www.opendefinition.org/okd/

The work of groups such as the Open Data Commons17 is relevant in developing licenses appropriate to the use and reuse of data, and should be evaluated for use within the public sector. The recently updated PSI Licence18 from the UK Government's Office of Public Sector Information (OPSI) is one useful example of the ways in which these generic principles may be enshrined for use within Government.

Government Linked Data

The UK offered an early and prominent example of Linked Data in Government, with a concerted Linked Data effort forming a core part of their data.gov.uk initiative from the outset. In the United States, the data.gov site began by simply making large quantities of data 'open,' but with far greater reliance upon formats such as Adobe's PDF. Researchers at Rensselaer Polytechnic Institute19 (RPI) made some progress in converting data published on the site to the Semantic Web's Resource Description Framework (RDF) format, sharing results via their Data-gov wiki20. In May 2010, the main data.gov site was relaunched, incorporating and building upon much of RPI's work within data.gov itself.

In a second Design Issues note, Tim Berners-Lee specifically addresses *Putting Government Data Online*21 and identifies a number of the simple steps that can be taken (as in the United States) before completely embracing a Linked Data approach.

From small towns and local authorities22, through cities such as Vancouver23, Manchester24 and San Francisco25, to countries such as the United States, Government in all its forms is increasingly embracing an open agenda and publishing data freely online. Despite early successes in the UK, high-profile evangelism from the likes of Sir Tim Berners-Lee, and clear *interest* at many levels, the work of transforming Government systems and practices to embrace the full reality of a Linked Data approach to data production, dissemination and consumption moves more slowly.

Those with an interest probably need to embrace the reality of *open* (and compliance with PSI legislation is a large step in the right direction) before exploring the added possibilities - and responsibilities - of pushing out Linked Data upon which the systems of unknown third parties might increasingly come to depend.

17. http://www.opendatacommons.org/
18. http://www.opsi.gov.uk/click-use/psi-licence-information/index
19. http://www.rpi.edu/
20. http://data-gov.tw.rpi.edu/wiki/The_Data-gov_Wiki
21. http://www.w3.org/DesignIssues/GovData.html
22. http://worktogether.org.uk/2010/04/13/lichfield-district-council-%E2%80%93-open-election-data-project-case-study/
23. http://data.vancouver.ca/
24. http://www.futureeverything.org/news/opendata1
25. http://datasf.org/