# Linked Data in Government

**Nigel Shadbolt and Kieron O'Hara** • *University of Southampton*

Government data is powerful, plentiful, and relevant to citizens' concerns. Making it open supports transparency, crowdsourced data enhancement, and innovative service development. The authors review the state of linked open government data, in the context of the potential for the publishing organizations and the Linked Data Web itself, as well as the administrative and political issues raised.

**G**overnment data is powerful, plentiful, and relevant to citizens' concerns — ideal for demonstrating the "power of open." Making it open supports transparency, letting citizens hold governments to account; enables data enhancement by crowdsourcing improvements to the data; and facilitates the provision of innovative services, helping to drive economic growth. Resource constraints and inertia mean that much open data appears in the formats nearest to hand — such as PDF, Excel, and CSV — but linked data can extract maximum value, supporting reuse in unanticipated contexts.

In this department, Tom Heath has already explained linked data's potential.[1] If government data, with all its good provenance and authority, its open licenses, and its sheer quantity, can be harnessed to the Linked Data Web (LDW), so much the better. All well in theory, but as is clear from even a cursory examination of government portals, the amount of linked data is very small as a proportion of government open data worldwide, so the anticipated network effects have yet to fully emerge. Despite this, have there been any quick wins? What has it meant in practice?

## The Direction of Travel

Governments are enthusiastic. The US Digital Government Strategy commits to progress "from managing 'documents' to managing discrete pieces of open data and content, which can be tagged, shared, secured, mashed up, and presented in the way that is most useful for the consumer of that information" (www.whitehouse.gov/sites/default/files/omb/egov/digital-government/digital-government.html). Data.gov was established in 2009 as a central repository for open data and has fostered communities of practice using structured open linked data around general topics such as business, and specific problems such as restoring Mexican Gulf ecosystems. Scientists at Rensselaer Polytechnic Institute (RPI) have set up a portal at http://logd.tw.rpi.edu that provides open data converted to RDF and linked to other linked data resources such as DBpedia. The LOGD portal also includes demos, tools, and search.[2]

In the UK, data.gov.uk, launched in January 2010, contains government datasets from many domains. Public Data Principles (http://data.gov.uk/library/public-data-principles) enshrine the UK government's open data policy and contain a commitment to linked data.

The US and UK have made great strides in linked government data, but other nations are following suit. For example, the European Commission has an Interoperability Solutions program (http://ec.europa.eu/isa/) for public administration, and opened an open data portal at http://open-data.europa.eu/open-data to provide access to nearly 6,000 datasets (of which 97 percent are statistical data from Eurostat),

IC-17-04-Lnkd.indd   2 5/23/13   12:15 PM

with RDF metadata. The W3C has set up a Government Linked Data working group (www.w3.org/2011/gld/wiki/Main_Page), while Tim Berners-Lee's 5* rating system (see Table 1) is widely used (the system not only rates datasets, but also provides a roadmap for moving from open to linked data).

| Stars | Characterization | Example format |
|---|---|---|
| * | Data online under an open license | PDF |
| ** | As * and structured data | Excel |
| *** | As ** and nonproprietary formats | CSV |
| **** | As ***, using URIs as identifiers | RDF |
| ***** | As ****, linked to other data | RDF |

*Table 1. Tim Berners-Lee's 5*-rating system.[+]*

[+]http://5stardata.info

## Machinery

Particular information types can be expressed with linkable vocabularies. Multidimensional data types, including statistics, survey data, and spreadsheets, can be published as linked data with the W3C's Data Cube vocabulary. Other works in progress from the W3C include a vocabulary for describing data catalogues, and ontologies for organizations and people. Meanwhile, the EU's 2007 Inspire Directive (http://inspire.jrc.ec.europa.eu/) for representing spatial information in a common format to facilitate sharing now provides standards for metadata, data specifications, network services, and so on.

Key to linked data's power is using HTTP URIs as identifiers. Every item that a government will refer to — each post/zip code, school, fire station, road, company, job position, and so on — should have a single URI to facilitate consistent reference and linking. This will break data out of the silos in which it's too often trapped when multiple identifiers are coined by different units. Of course, this won't solve every identity issue,[3] but it will provide a simple way to begin the linking process. De-siloing data means that we have to think about what it means a little more deeply: this is no reason not to do it.

An official URI for each significant object or concept would liberate information from its silos, as well as carry a great deal of weight given the quantity and authority of government data available. Certain crucial data, including geodata, postcodes, businesses, and contracts, is core reference data — particularly useful for connecting other datasets together.[4]

This isn't to say that governments (or data users) should try to pretend that the linked data world isn't complex, with multiple URIs referring to the same objects — this is inevitable in a decentralized, webby world. In general, generating a new URI isn't the solution to a world with too many URIs, but governments and pan-government institutions such as the EU are well-placed to be authoritative with core reference data.

So, what difference has all this progress made? We illustrate some of the low-hanging fruit with a few plums from the UK (a list of UK government linked open data is at http://data.gov.uk/linked-data/who-is-doing-what).

## Realizing Open Data's Benefits

Linked data lets us realize open data's benefits with less effort from users. RDF supports several data models, so the data can be retrieved not only as RDF, but also as JavaScript Object Notation or CSV, for example. Just because data is stored as RDF, it doesn't have to be consumed as RDF, so users don't have to convert the data into the format that they need.

For example, the Environment Agency (EA) publishes weekly assessments of water quality in bathing areas, linking annual with weekly assessments and profiles of the locations (http://environment.data.gov.uk/lab/bwq-web). Users can interrogate data about water samples, see graphs through time about particular areas, and view assessments directly as linked data, RDF, or CSV (or a special-purpose language such as iCalendar), allowing mashups with other data — for example, about holiday schedules, accommodations, or air quality.

We could do this with any kind of open data, but using links makes it easier to discover related datasets or reconcile two datasets because URIs can disambiguate in a way that text labels can't. Moreover, the data user can be given provenance information, caveats to particular data values, definitions of terms, and versioning information more easily. In this way, linked data facilitates the open data agenda.

## Internal Efficiency

The UK Department of Communities and Local Government (DCLG) runs Open Data Communities (http://opendatacommunities.org/data), presenting its own statistics on local issues including finance, housing, and deprivation, and reusing Ordnance Survey geodata and statistics from the Office of National Statistics. The data drives the Local Authority Dashboard (http://opendatacommunities.org/dashboard), an application in which linked open data from DCLG can be explored, presented on maps, or graphed against national averages (see Figure 1).

The 74 datasets are linked data and presented via an open API, which also means a second gain — other groups can reuse the data together with their own to create new (not
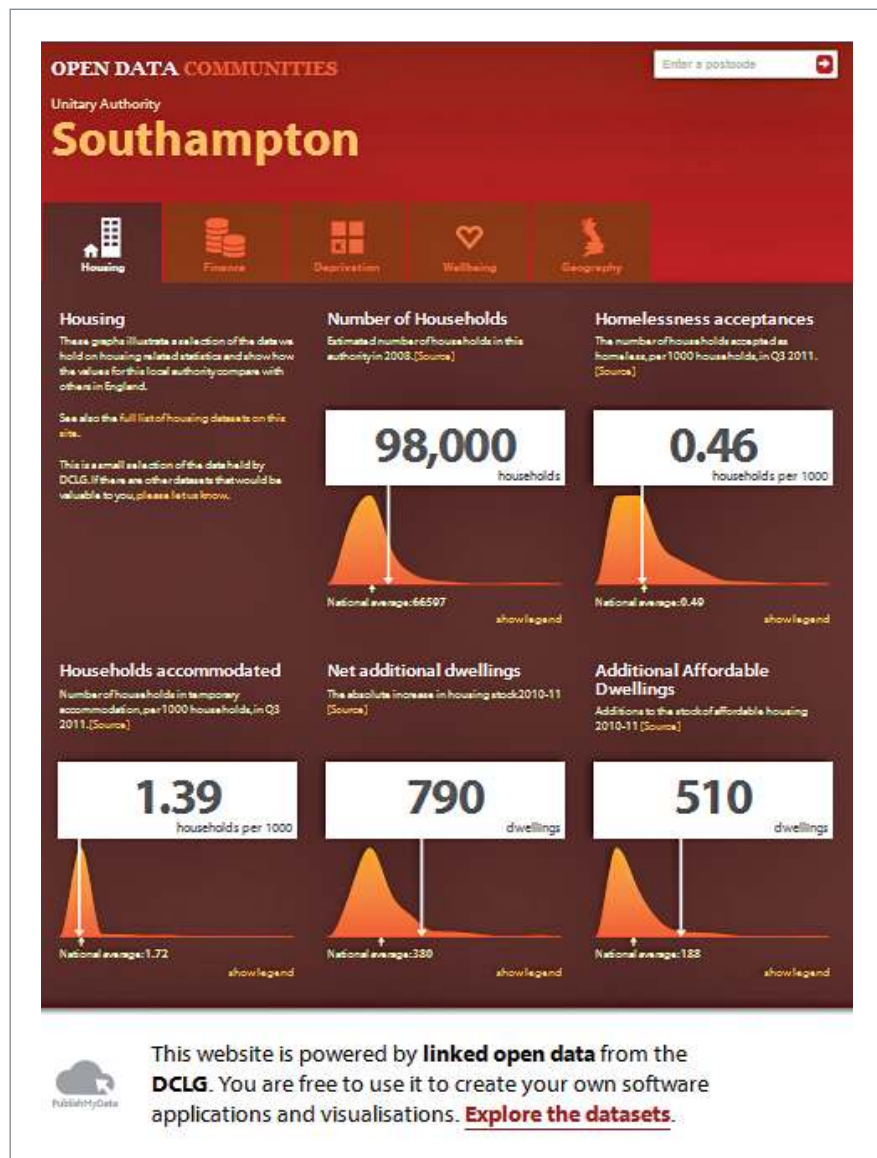
*Figure 1. The Local Authority Dashboard for Southampton. On the dashboard, data from the Department of Communities and Local Government can be explored, presented on maps, or graphed against national averages.*

necessarily open) applications. The UK Department of Health has used the data to match cancer patients' postcodes with the geography of deprivation, while the Lambeth in Numbers page (http://lambeth-in-numbers.co.uk/map) supports one London borough's public health strategy. Figure 2 shows the distribution of child obesity in Lambeth overlaid with food outlet types.

The third gain is the most immediate. Once the initial push occurs, linked data's benefits become clear not only to outsiders but also within the publishing organization. Those using data in the DCLG find it easier to consume their own linked data via the Local Authority Dashboard than integrate the data manually as they used to. The DCLG is a crucial beneficiary of its own linked data.

## Setting Authoritative Standards

Government data doesn't only describe the world. For instance, being on a state database of registered cars doesn't just tell you some facts about a car — it means that the car is registered. Government data offers great scope not just for enriching the LDW but also for providing it with standards.

UK legislation has been encoded as 5* data. Legislation.gov.uk, from the National Archives, uses the Crown Legislation Markup Language (CLML) to

- display legislation as required by law in a device-independent way, while
- representing changes in legislation over time as data, and
- supporting semantic representations of the content, facilitating intelligent search.

The legislation's presentation makes it clear whether it's currently in force or has subtly changed over time — aspects that aren't clear simply from reading the text (see Figure 3). Different data models cover the legislation as enacted, and as revised.[5]

Legislation.gov.uk is an important addition to the LDW. Legislation defines important concepts in the outside world — schools (27 types defined in UK law), companies, official posts, legal principles, and many others. Hence, the LDW can be underpinned with legally ratified definitions (from any jurisdiction), allowing automated checking of whether a particular concept is legally in force. Accountable systems can determine compliance with the law, and, as more nations put their legislation online as linked data, such systems could determine in which jurisdictions certain objects or practices conformed to law (or not).

## Prospects and Challenges

Given the exciting prospects, it's fair to ask how much linked open government data is actually out there. The proportion is low — even in the UK, of the 9,000 or so datasets on

data.gov.uk, just under half have been classified, and of those, fewer than 5 percent contained linked data. According to the latest version of the Linked Open Data project's linked data cloud (http://lod-cloud .net/), roughly one-sixth of the LDW stems from governments, so even an increase of linked data to 10 percent of open government data will have a dramatic effect.

As with all open data, concerns arise about quality. RDF publication is a demanding discipline (http:// pedantic-web.org) in which external linkage and vocabulary reuse correlate positively with good practice.[6] The full stack of Semantic Web technologies and protocols imposes a burden on units with complex data requirements and limited resources, whereas governments' priorities are to increase the amount of open data of whatever form, rather than to maximize linked data. Excel can be published by exporting data in use, and nonproprietary formats such as CSV generally require little more than one-off investments in plugins or converters. But linked data has resource and training implications: we must give thought to "slicing and dicing" data, assigning URIs, and developing or reusing ontologies. Links need to be made to other datasets, and broken links repaired. Small wonder that the low-cost 2* options sometimes look more attractive to cash-strapped managers than 5* data.

Yet quantity might sometimes trump quality for two primary reasons. First, many government statisticians privately admit that they have too much data to process. Publishing data increases the benefits it will provide and lets users critique and improve it. Crowdsourcing data improvement is a selling point for all open data, and third-party auditing of linked data depends on access.

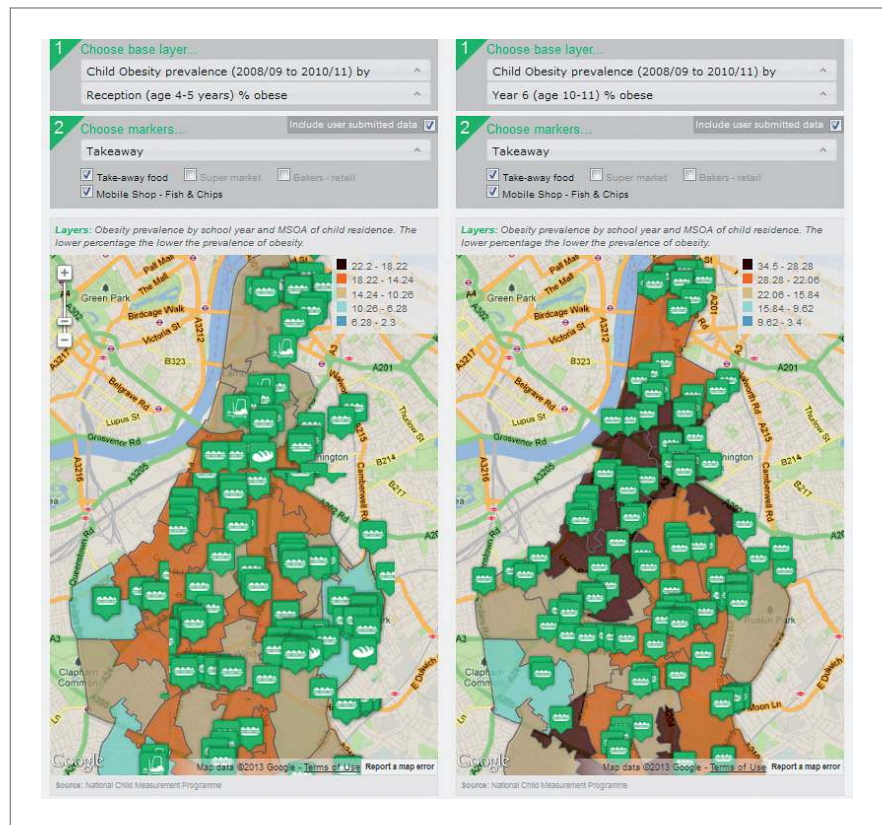Second, linked data's success depends on it being consumed and consumable. As noted, this is



Figure 2. The distribution of child obesity in Lambeth. This data is overlaid with food outlet types.

becoming increasingly important within governments independently of whether the data is open (and government is likely to be a major consumer of government data even in the open data era). Hence, real-world use cases are vital — examples on portals are fine for motivation, but engaging with users and identifying core reference data will be more important.

At the moment, pushes from the top are still needed. Until more data is linked to the LDW, the network benefits will remain nascent. Resources such as DBpedia are valuable and must be linked to, and link-creation services such as backlinking will be important. Dataset catalogues are vital for discovery, while generic metadata standards enable communities other than primary users to benefit.

Elsewhere, we've identified bottlenecks to bringing government data

to the LDW — discovery of open data, aligning ontologies between datasets, usable interfaces, and the low number of applications. Lightweight, pragmatic methods reduce administrative overhead, while better tools and interfaces should increase uptake.[4] We also need effective means to track provenance and protect citizens' privacy.

Linked open data will have important effects on political life, helping drive open participation and reconnect government and citizens. Inference will no longer be an oligopoly of governments and large corporations. Informed decision-making and service provisioning can be devolved to local governments, communities, the private sector, charities, nongovernmental organizations, and even individuals. Accountability is enhanced by investigating and tracking
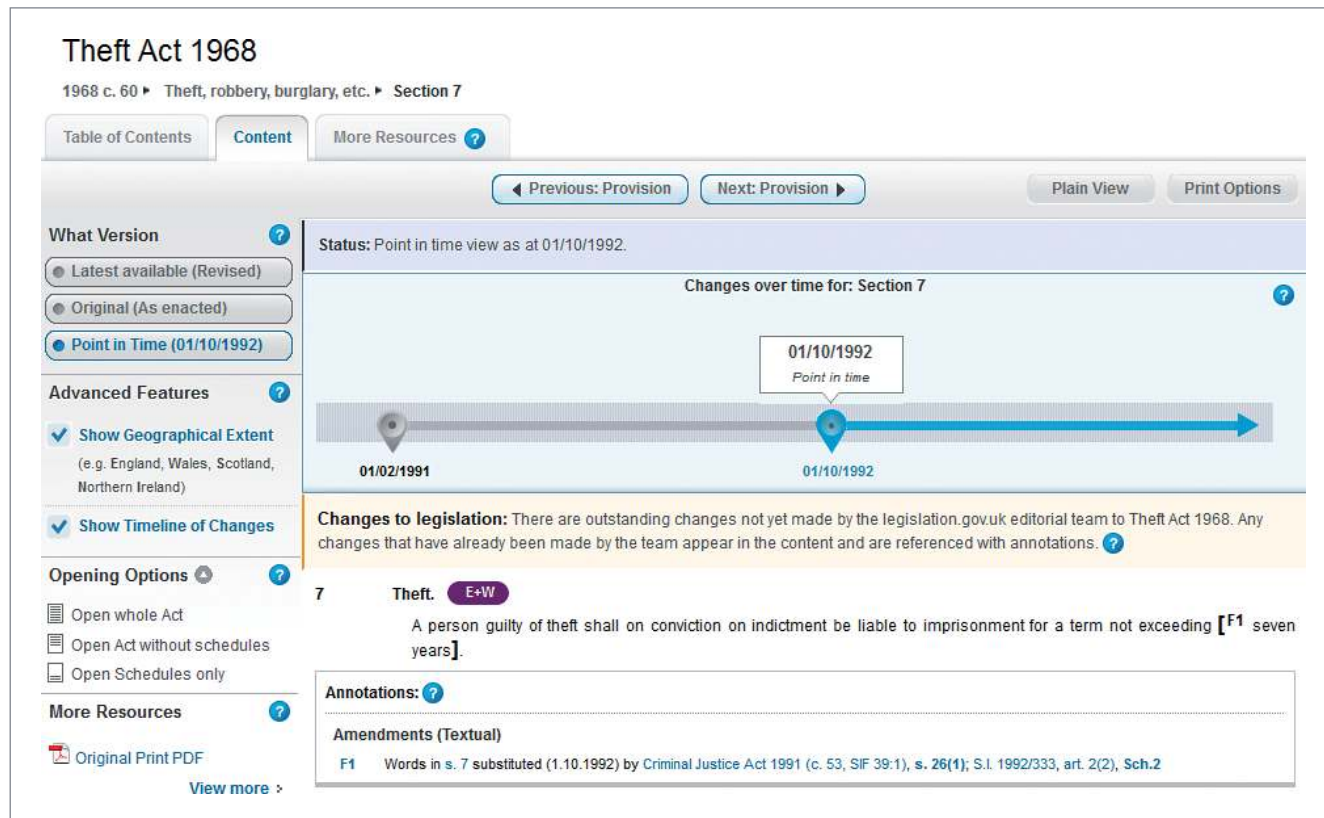
*Figure 3. Legislation.gov.uk showing a clause of the Theft Act 1968. The timeline shows when an amendment to that clause was enacted (in the 1991 Criminal Justice Act), and when it became law (in 1992). The purple lozenge at the beginning of the clause specifies that it applies only in England and Wales, not in Scotland or Northern Ireland.*

## The Open Data Institute

Linked open data in the UK is supported by a unique institution, the Open Data Institute (ODI; www.theodi.org). The ODI, part-publicly and part-privately funded, is a hub for connecting data supply and demand. It works with government officials to unlock supply, advising on policy and methods, and with start-ups and more mature businesses to unlock demand, organizing hack days and data analyses. One recent (nonlinked) example uncovered surprising and costly regional disparities in the UK between prescription rates of generic and proprietary drugs (www.prescribinganalytics.com). By intermediating between demand and supply, the ODI's ambition is to catalyze the development of innovative services based on the use of open data, and to help realize value by increasing the quantity of 5* data.

government performance. Linked data facilitates comparative studies and deep contextualization, although the skills required to query linked data and understand the outputs aren't yet widespread. Letting go of the narrative around data is a risk for governments, but publication need not preclude providing context — for instance, publishing a PDF report with the usual tables and charts to interpret the data (1*), while simultaneously putting out the raw data in CSV (3*) or RDF (5*) as a download for specialists.

It's important to realize that publishing linked data is neither a technical matter nor an administrative one, but a complex combination of the two. Methods, institutions such as the Open Data Institute (see the sidebar), and resources such as the LOGD portal at RPI are emerging to support this important shift toward democracy and the "power of open." ⊡

### References

1. T. Heath, "Linked Data — Welcome to the Data Network," *IEEE Internet Computing*, vol. 15, no. 6, 2011, pp. 70–73.
2. J. Hendler et al., "US Government Linked Open Data: Semantic.data.gov," *IEEE Intelligent Systems*, vol. 27, no. 3, 2012, pp. 25–31.
3. H. Glaser and H. Halpin, "The Linked Data Strategy for Global Identity," *IEEE*

*Internet Computing*, vol. 16, no. 2, 2012, pp. 68–71.

4. N. Shadbolt et al., "Linked Open Government Data: Lessons from Data.gov.uk," *IEEE Intelligent Systems*, vol. 27, no. 3, 2012, pp. 16–24.

5. C. Tullo, "Online Access to UK Legislation: Strategy and Structure," *Frontiers in Artificial Intelligence and Applications 236: From Information to Knowledge*, M.A. Biasiotti and S. Faro, eds., IOS Press, 2011, pp. 21–32.

6. A. Hogan et al., "An Empirical Survey of Linked Data Conformance," *J. Web Semantics*, vol. 14, July 2012, pp. 14–44.

**Nigel Shadbolt** is a professor of artificial intelligence and head of the Web and Internet Science Group in the Electronics and Computer Science Department at the University of Southampton, an advisor to the UK government, and chairman and cofounder of the Open Data Institute. His research interests include open data, the Semantic Web, and Web science. Shadbolt has a PhD in artificial intelligence from the University of Edinburgh. Contact him at nrs@ecs.soton.ac.uk.

**Kieron O'Hara** is a senior research fellow in the Web and Internet Science Group in the Electronics and Computer Science Department at the University of Southampton. His research interests include trust, privacy, open data, and Web science. O'Hara has a DPhil in philosophy from the University of Oxford. Contact him at kmo@ecs.soton.ac.uk.

cn *Selected CS articles and columns are also available for free at http:// ComputingNow.computer.org.*