Linking UK Government Data

John Sheridan
The National Archives
102 Petty France
London SW1H 9AJ
john.sheridan@nationalarchives.gsi.gov.uk

Jeni Tennison
The Stationery Office
Mandela Way
London SE1 5SS
jeni.tennison@tso.co.uk

ABSTRACT

What does it take to create a web of linked government data? With the launch of data.gov.uk the UK Government has been finding out. This paper sets out the case for using Linked Data standards for publishing open government data and describes some of the benefits. It explains how Linked Data standards uniquely allow governments to publish data responsibly and why responsible data publishing is so important to the open government data movement. The paper goes on to explain how the Linked Data world was not quite ready for the large-scale adoption of these standards by a major government, leaving much to be done to develop practical approaches and patterns for the publishing of government data. From URIs, to provenance and versioning, through to statistics and geographic information, much thinking and work has been done. In each case the emphasis has been, not on research, but designing simple repeatable patterns, supported through tools. This work has also involved and building understanding and capability amongst officials from across government departments and agencies.

It explains why the government's use of linked data standards was not universally welcomed and was even greeted by antagonism from some. Learning from this feedback the paper describes how we are now using linked data standards to enable government as a platform, commoditising the process of creating APIs to meet the needs of a wide range of data consumers, from business, academia and the developer communities.

Categories and Subject Descriptors

H.4 [**Information Systems**]: Information Systems Applications

General Terms

Design, Standardization

Keywords

Linked Data, eGovernment

1. INTRODUCTION

There are many different ways of putting data on the web. It has been possible for governments to publish data using the Internet for over 30 years, long before the web was invented, by providing access to flat files over FTP. What is distinctive about the web is HTTP and the web's linking ability. In 2009 governments around the world started to move decisively towards publishing increasing volumes of government data on the web, perhaps most notably with the launch of data.gov in the United States. In the UK, Sir Tim Berners-Lee and Professor Nigel Shadbolt were appointed as advisors to the Government to spearhead work in government. The UK Government's data website, data.gov.uk, aimed at developers, was launched with a commitment that the government would use W3C standards and in particular adopt linked data standards and approaches for publishing UK Government data on the web.

The open government data movement provides a golden opportunity for linked data advocates to prove the value of standards such as RDF, OWL and SKOS. The experience from the UK is that the linked data community was not quite ready for a major government to start creating a web of linked government data. Whilst the standards are mature, capable and powerful, much work needs to be done to translate those standards into simple and repeatable publishing patterns that government departments and agencies can adopt, use and implement.

The UK Government aims to be a responsible publisher of linked data – indeed that is an important part of the motivation for using linked data standards. To do this it is addressing questions such as how to handle versioning and provenance information. The government is developing publishing patterns that ease the process of publishing data in linked data form. More specifically it is thinking through the implications of linked data standards in two important domain areas, statistics and geo-spatial information.

The divergent needs of government data publishers and government data consumers is also becoming apparent. Many developers want to download government datasets or, better, to have programmatic access to data through RESTful APIs. The demand from data consumers for government data in linked data form, particularly those outside of academia, is limited to a growing minority. The majority of developers want easy and immediate access to data in simple-to-use formats. Despite the inherently RESTful nature of linked data, SPARQL Endpoints and RDF are viewed with suspicion. The data.gov.uk mailing list received numerous posts from developers wanting to have programmatic access to government data but not wishing to use either SPARQL or RDF. To bridge the gap between responsible data publishing and easy data use, the UK government is using linked data as an underpinning technology, as a bridge that enables data publishers to meet the diverse needs of data consumers.

2. OPEN GOVERNMENT DATA

There is a global movement of governments and local authorities starting to put their data on the web. Open government data projects have sprung up in countries around the world from the United States, Australia and New Zealand to The Netherlands, Sweden, Spain, Austria and Denmark, not to mention an

LDOW 2010, April 27, 2010, Raleigh, North Carolina.

increasing number of city- and local-authority-based initiatives from Vancouver to London. The policy objectives achieved through open government data vary, including increasing transparency and democratic accountability, supporting economic growth by stimulating new data-based products and services, and improving how public services are delivered.

In the UK, the government has set out clear public data principles. These state that the government will make public data available in machine-readable formats, published using open standards and released under an open licence. The UK government has gone further, committing that it will make any raw dataset available in linked data form.

There are a number of advantages for using linked data standards for publishing open government data. The most important benefit for publishers of government data is how linked data standards enable departments and agencies to publish their data responsibly. This is because each fact or data point is associated with a URI and that URI can be resolved. The publisher determines what information is returned when a request is made and can serve whatever additional context or provenance information they deem necessary. For example, if the 2002 figures can't be compared with those in subsequent years, the publisher can say so, for every data point. The data can be copied, adapted and re-used, but the publisher always controls what is returned when each URI is dereferenced. This is an important benefit over interchange formats such as CSV or XML where data can be changed or context lost as it is passed from hand to hand or system to system. For government officials worried that their datasets will be dumbed down to create the machine readable form and then used in incorrect or even misleading ways, linked data provides a real boon. It leaves the publisher ultimately in control of their data in a unique way, whilst enabling very flexible consumption and re-use.

There are other benefits for governments wishing to publish data. Linked data is based on open standards. This aligns it well with the UK Government's commitment to open standards in *Open Source, Open Standards and Re-use*[1].

Linked data enables the government to publish its data in a very modular way, benefiting from a 'small pieces loosely joined' approach to government data. This is important as the government is itself a complex and highly distributed set of organisations. The most useful data about schools, for example, will be the combination of information from a number of different departments and agencies. Each organisation can publish its own data but using linked data the information can also easily be combined. Neither the government nor data consumers need everything to be planned in advance, the data web can evolve, as the web of documents has.

Rather than create many different bespoke APIs, which would prove time consuming and potentially expensive, linked data technology offers a way of providing flexible and easy programmatic access to data. Moreover, linked data is also very portable, not locking the government in to a particular vendors technology platform or approach.

3. DESIGN PATTERNS

The four Linked Data principles provide some very clear guidance: that HTTP URIs should be used to name real-world things, that they should be resolvable to representations of RDF graphs, and that they should link to other resources. However, they naturally remain silent on the details. Some of these gaps have been filled by work such as Cool URIs for the Semantic Web [2] and Best Practice Recipes for Publishing RDF Vocabularies[3] but even these rightly leave developers with a lot of choice about how to approach the publication of linked data.

This choice is good in many ways, and accurately reflects the relative immaturity of the field. However, in the context of encouraging government departments to publish their data, it can also be confusing. To help publishers get up and running quickly, with the minimum of effort, we have adopted a policy of providing clear guidelines and recommendations. These are not intended to constrain publishers (they are not *rules*) but to ease the path to publication by giving clear directions along the way. By providing a level of consistency in approach, we also hope to make it easier to create tools and to help consumer developers know what to expect from government linked data.

These guidelines are led by and refined by experience, and are thus at different levels of maturity. There are three sets in particular that are worth highlighting here: in the design of URIs, the approach to versioning and the provision of provenance information.

3.1 URIs

Some of the earliest work centered on the creation of patterns for URIs, culminating in the publication of Designing URI Sets for the UK Public Sector[4]. These married guidance based on the Linked Data approach, usability guidance, and practical constraints particularly regarding the impermanence of many department-based domain names. The result is a slash-based scheme that includes four main patterns:

- `http://{sector}.data.gov.uk/id/{concept}[/{identifier}]` for real-world things such as schools and roads
- `http://{sector}.data.gov.uk/doc/{concept}[/{identifier}]` for documents about those things
- `http://{sector}.data.gov.uk/def/{scheme}[/{concept}]` for vocabularies, classes, properties, concept schemes and concepts
- `http://{sector}.data.gov.uk/data/{dataset}[/{part}]` for datasets and the graphs they contain

The first pattern results in a `303 See Other` response to an equivalent URI using the second pattern. For the latter three patterns, suffixes are used to indicate particular formats for the returned documents.

We have also framed general guidance about these URIs, such as:

- using natural identifiers within URIs where possible
- considering the persistence of URIs over time
- designing URIs for things that are not ultimately controlled by data.gov.uk

[1] http://www.cabinetoffice.gov.uk/media/318020/open_source.pdf

[2] http://www.w3.org/TR/cooluris/

[3] http://www.w3.org/TR/swbp-vocab-pub/

[4] http://writetoreply.org/ukgovurisets/

3.2 Versioning

Versioning is particularly important with government data. It's important to be able to relate a given event to a local authority that was disbanded in April 2009, and to relate that local authority to the one that has taken its place. It's important to be able to track shifting classifications and coding schemes as these have implications to how we interpret statistics about crime or health over time. While in general consumers will be interested in the current state of the world, it's particularly important for policy makers to look back into the past and project into the future.

We also have to deal with different sets of information about a given resource updating at different times, and information from different sources, modified at different times, potentially overlapping with each other. For example, a school's name might be recorded in five different databases, all exposed as linked data and updated at different intervals. How can we determine which reflects the current name of the school?

These considerations have led us to adopt named graphs as a mechanism for annotating sets of statements with information about their validity over time, their authoritativeness, and other named graphs in the same series. While many sources may provide information about a given resource, only one should provide authoritative information about a particular property of that resource, such as the school's name. These graphs can be combined to give slices of information at a particular point in time.

3.3 Provenance

Alongside the requirement for handling changing information, the UK government needs to provide information about the source of the information that it publisher as linked data. This includes the provenance of the data itself but also, crucially, the ways in which it has been manipulated en route to the final consumer of the data.

There are going to be many different ways in which linked data is generated and published, including:

- publication of RDF-based representations from standard relational databases, as just another output format
- generation of RDF based on transformations from other formats, such as CSV files, which is then served purely as RDF
- programmatic, on-demand generation based purely on the resource URI

Named graphs are again vital for associating metadata, this time about the provenance of information, to sets of triples. We are currently working on developing the patterns to represent the complexity and variety of provenance of government linked data, in coordination with the W3C Provenance Incubator Group[5].

4. KEY CONTENT AREAS

From our earliest forays into linked data, it was clear that there were two areas which deserved special attention: statistics and geo-spatial information. Practically every interesting dataset contains statistics of some description, whether it's the number of vehicles passing a particular point or the number of pupils of a certain age within a school; and references a location or an area in the real world.

Unsurprisingly, both areas have large existing communities of interest and standard approaches to modelling and representing their data. But in both cases, the advantages of a linked data approach have been recognised very quickly.

4.1 Statistics

Statistics are a vital source of data. While RDF does not provide succinct representations of multi-dimensional hypercubes, representing statistical information using linked data provides three important benefits:

- the ability to slice hypercubes in ways not anticipated by their original publishers
- links into the wider cloud that provide extra contextual information for the statistics
- the ability to make annotations at various levels, from whole datasets down to individual observations

The UK government has led work to align the major existing standard for publishing statistics, SDMX[6], with Linked Data, leading to work on both extending SCOVO and mechanisms for accessing these statistics using web-based APIs.

4.2 Geo-Spatial Information

The UK is under a statutory obligation to implement the European INSPIRE Directive[7], which seeks to ensure that European countries are able to exchange spatial information. One of the features of this directive is the requirement to provide identifiers for and a resolution mechanism for spatial objects. This dovetails with the opening up of geographic information within the UK, particularly that provided by the Ordnance Survey.

The UK has decided to use Linked Data to fulfill these requirements: spatial objects will be identified through HTTP URIs, and the resolution mechanism will be the standard web architecture. Particular issues that we are working through include:

- the correspondence between real-world things and the spatial objects that represent them
- the representation of phenomena such as boundaries that both change over time and are available at varying resolutions
- the representation of geometries within RDF, whether as literals or as resources

5. LINKED DATA FOR DEVELOPERS

Linked data standards are very powerful but for many developers RDF and SPARQL are new technologies. Although the data is machine readable, the standard formats such as RDF/XML and Turtle are impenetrable without special parsers, making it hard to use by non-experts. Representing data as graphs rather than using the more familiar paradigms of trees or tables adds another obstacle. To consume linked data effectively the data consumer has to think differently, constructing both a new mental model of the data and, very often, starting to use a new or unfamiliar code library to work with and manipulate it. The experience from the data.gov.uk mailing list is that even experienced developers baulk

[5] http://www.w3.org/2005/Incubator/prov/

[6] http://www.sdmx.org

[7] http://inspire.jrc.ec.europa.eu/

at the learning curve required to fully exploit a SPARQL endpoint.

When the UK Government started publishing data using linked data standards, those already familiar with the technology cheered. A wider group of developers approached the technology with an open mind but many floundered. It was too hard to find out what data was available and how it had been modelled. This is made harder by difficulties getting overviews of RDF vocabularies and the possible properties that a given resource might have.

We provided example SPARQL queries, but important features such as aggregation of results through summing and counting were missing both from the current SPARQL standard and from the then data.gov.uk implementation. Tutorials available elsewhere on the web were sparse and not developer friendly. There was no easy to follow progression path from the world developers knew to the world of linked data. The choice was either to make the leap and invest considerable time and energy to use linked data, or to be left frustrated. Far from enabling access to open government data, linked data standards looked like they getting in the way!

The last few years have seen an explosion in web services provided through RESTful APIs. These APIs are defined either through URI patterns or queries and return data in simple XML or JSON formats that are easy to process on both servers and clients.

The UK government has therefore been supporting work to create middleware, based on a standard configuration format, that can sit above SPARQL endpoints to:

- provide simple JSON and XML views of linked data
- provide simple URI-based searching, filtering and sorting of linked data
- support the creation of flexible domain-specific APIs by data publishers

To do this, we have written a specification for the API features the middleware will provide and supported the creation of initial implementations.

All UK Government linked data can now also be available through RESTful APIs. The ultimate goal here is to commoditise the production of APIs through linked data in a way that is both simple for the publisher and valuable to the consumer, to demonstrate that publishing in linked data provides benefits that vastly outweigh the costs. The case for linked data is transformed by this approach. Powerful in their own right, linked data standards now also provide the UK Government with a basis for the rapid creation of APIs.

As a result of this work, the UK Government has now embraced linked data both as the most effective route for providing programmatic access to data both natively and in the easy to consume formats that developers already know, such as JSON. Linked data standards provide an underpinning technology that can enable other forms of programmatic access to data.

6. CONCLUSIONS

The UK Government is making a serious attempt to create a web of linked government data as part of the wider linked data cloud. . There are important benefits for governments by using linked data standards for data publishing. For data publishers in government, linked data standards mean they can publish their data responsibly. For data consumers, linked data standards mean they can re-use government data flexibly and easily, for example through APIs.

Adopting Linked Data within the UK government has been an exercise in balance:

- between the largely academic advocates of linked data and the pragmatic concerns of data consumers
- between providing publishers both helpful patterns, guidance and the flexibility to move outside their bounds when necessary
- between the need for a centralised, single point of access to government information, that makes it easy to find and use, and its distributed publication
- between providing core resources on which we can build while recognising that growth will only come from data holders publishing thier own data on the web

To overcome the bootstapping problem an important focus has been on realising immediate benefits from the use of linked data standards as well as the longer term gains; jam today as well as jam tomorrow. Using linked data as an underpinning technology for creating APIs is an important approach, embracing the needs of the widest range of data users, not just those familiar with linked data.

There are major opportunities for linked data standards with government data, particularly for statistical and geo-spatial information. There is much still to be done and more to learn about the implementation of linked data standards for government data. We believe the practical application of linked data standards by the UK government has strengthened the case for linked data whilst highlighting some weaknesses in the maturity of implementation approaches, which we have worked to resolve.