Demo: Linked Open Statistical Data for the Scottish Government

Bill Roberts[1]

[1] Swirrl IT Limited http://swirrl.com

Abstract. This paper describes the approach taken by the Scottish Government, supported by Swirrl IT Limited, to publish a large and growing collection of statistical data as Linked Data. The system is operational and well used. The paper describes the features of the system and the approaches taken to various practical challenges in applying Linked Data to the task of sharing, disseminating and enabling effective use of public sector statistics.

Keywords: Linked Data, Statistics, Government Data

1 Introduction

This paper describes the approach taken by the Scottish Government, supported by Swirrl IT Limited, to publish a large and growing collection of statistical data as Linked Data. It is intended to accompany a demonstration at the Semstats 2017 workshop. The Scottish Government's statistics publishing system is known as 'statistics.gov.scot' and can be viewed online [1]. The system is operated by Swirrl's 'PublishMyData' software.

This paper and the associated demonstration describe an operational use of Linked Data, in particular the RDF Data Cube Vocabulary [2], for online dissemination of government statistical data. We describe the objectives of the system, the challenges faced and the approaches taken to address those challenges, thereby sharing some practical experiences of applying the RDF Data Cube approach at a relatively large scale.

Like all government organisations, the Scottish Government gathers statistics on many aspects of the state of the country, to assist in developing, assessing and implementing policies. The main objective of statistics.gov.scot is to make those statistics more widely accessible to those who need them. Many of the users are analysts in central or local government, but the data is also used by academic researchers, businesses, community groups and citizens.

The choice of Linked Data as a technology for this system arose from a number of objectives:

- To make good use of the Web as the dissemination mechanism
- To use existing open standards where possible
- To provide API access to all data, as well as downloads
- To support intercomparison of data from different sources and relating to different topics

However the majority of users of the system are unfamiliar with Linked Data as a technology, so it is necessary to hide some of the underlying details of the data representation while still aiming to take advantage of the strengths of Linked Data – such as providing unique web-accessible identifiers for all aspects of the data.

In the rest of this paper, we describe the data collection, the technical approach to delivering the data, the challenges faced and lessons learned and an overview of ongoing work to improve the system.

2 Characteristics of the data

At the time of writing (late July, 2017), statistics.gov.scot holds:

- 163 data cubes (plus some other supporting datasets)
- 77 million statistical observations
- 717 million triples
- 137 concept schemes

The system supports around 5000 unique users and 50,000 page views per month (not counting API use).

The size of the data collection is growing steadily and covers the range of topics that the Scottish Government is responsible for. (This covers the full range of normal functions of government, with the exception of aspects of fiscal and monetary policy, pensions and benefits and foreign policy, which are the responsibility of the United Kingdom government). Therefore the system holds data on demographics, crime, the economy, education, environment, health, transport and so on. The full list of datasets can be found at [2] and the datasets organized by themes is at [3]. The choice of themes was from a previously used Scottish Government classification: revision of these is anticipated, to align with an updated Scottish Government standard approach that is still being developed.

Almost all of the data cubes have dimensions including geographical area and time. The number of dimensions per data cube ranges from 2 to 6. Other commonly occurring dimensions and corresponding concept schemes are gender (52 datasets) and age (39 datasets) of the relevant parts of the population. Some dimensions and concept schemes are unique to particular datasets.

The system encourages but does not enforce re-use of dimension properties and concept schemes across datasets. There are a few concept schemes that cover overlapping concepts and which could and should be standardized and combined for better data comparability, for example two similar concept schemes for benefit types [4], [5]. All vocabularies defined in the system are listed at [6].

The geographical dimension of the datasets relates data to administrative or census-based areas. The smallest of these are known as 'data zones'. There are approximately 7000 data zones covering Scotland. These are designed to have approximately equal populations, therefore are relatively small in cities and large in rural areas. Other commonly occurring geographical areas are electoral wards (354 in Scotland) and council

areas (32 in Scotland). The total volume of data is dominated by the observations related to data zones.

The hierarchy of geographical areas is represented in the system and is used both to support automatic aggregation of data from small areas to larger areas, so reducing work for those responsible for preparing and loading data, and to support browsing and filtering the overall data collection.

Data referred to organisations (for example schools), as opposed to geographical areas, is now becoming a higher priority within the system and work is in progress (see Section 4) to extend some of the user interface facilities to provide better support for such data.

3 Technical approach

3.1 Data model

The data is stored in a triple store (using the Stardog database) and is organised according to a few simple conventions. Each dataset consists of one named graph containing the data contents and one graph containing the metadata. Each concept scheme or ontology is held in a separate graph. This is relied on in the approach to updating data.

The focus of the data model is the RDF Data Cube Vocabulary. Standard properties from that vocabulary are used for 'refArea' and 'refPeriod' dimensions. Other dimensions, measures and units are generally defined in the statistics.gov.scot namespace.

Time intervals are defined using the reference.data.gov.uk URI set [7]. Geographical area URIs are defined in the statistics.gov.scot namespace, based on the UK 'Government Statistical Service' codes [8], a comprehensive and well-maintained set of identifiers for administrative and statistical geography regions.

A challenge with representing the geography hierarchy arises because the relationships do not make a strict tree structure. Areas can have multiple 'parents'. One solution to this is representing the data as multiple trees, for example using XKOS. However our approach has been to treat the relationships as a graph structure. Each area is a member of one or more 'collections' of areas of a similar type, and has 'within' relationships to a number of larger areas. We link directly not just to immediate 'parent' areas but also to 'grandparents' etc.

Our approach to representing the semantics of measures and quantities is to use fairly generic measure properties, such as count, ratio, percentage; and to use the units to describe what is being measured (e.g people, households, years). The aim is to make it easier to know if measures in different datasets are directly comparable.

3.2 System architecture

Figure 1 summarises the architecture of the system. More detail on the functions of key components and design choices behind the architecture are discussed in the following sections.

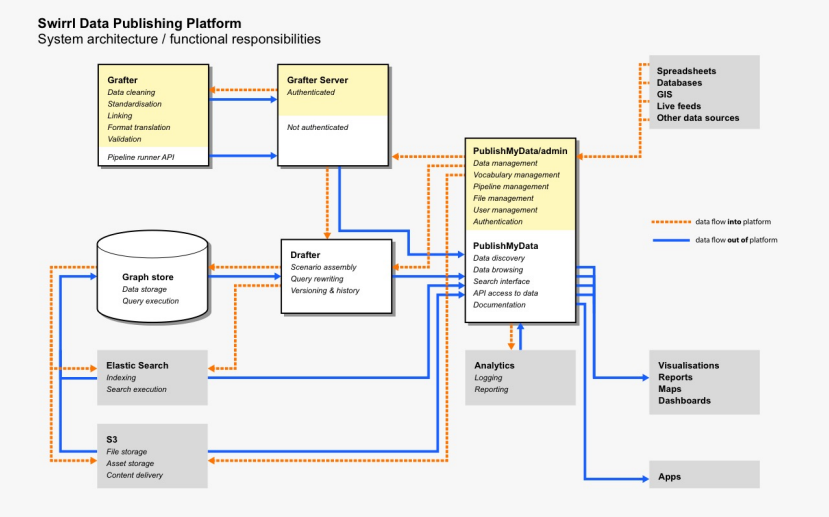

Figure 1 PublishMyData system architecture

3.3 Tools for data users

A challenge for the system design is the need to support a range of different user types, with varying levels of knowledge of the data and of technology. Possibly the most important user group is public sector analysts: who are relatively knowledgeable about statistical data, but know very little about Linked Data, HTTP or APIs or programming. They typically want to find data, view it online to assess relevance, then download it for use in simple reports, or analysis tools like Excel, R, or GIS.

The system also supports all the conventions of Linked Data, and a public SPARQL endpoint, to enable programmatic use of the data. A design challenge is that the information needed to support API users is confusing for those who just want user interface features and downloads. We have attempted to provide some separation of these two types of information in the user interface.

A key challenge is the large volume of data, both in terms of many diverse datasets and the size of some datasets. Users generally do not want to download a whole dataset. Many users (for example in local government) are only interested in one part of the country.

In the design of the system we have tried to provide tools allowing fine-grained selection of data, without needing sophisticated technical skills. Users can browse for datasets or find them via text search on dataset name and description. Users can also find data by searching or browsing the geographical hierarchy.

A user interface is provided for filtering multi-dimensional datasets, by fixing the values of dimensions until a two-dimensional table can be displayed. That tabular view can then be downloaded, or used to select a one-dimensional view with geography as the free dimension, allowing an automatically generated thematic map to be displayed, or leave time as the free dimension and show a time series chart. Figure 2 shows an example thematic map created by the system.

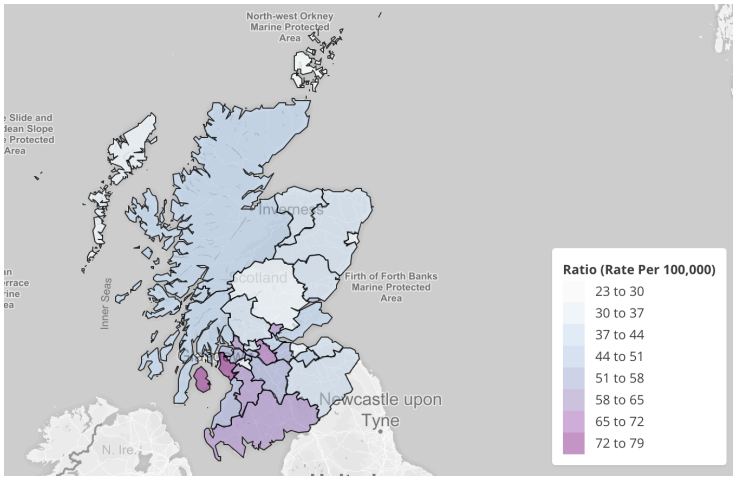

Figure 2 An example thematic map created by the system: deaths due to heart disease

The 'data cart' feature allows the user easily to create and download (as CSV or Excel) a custom table, with geographical areas as rows, and one-dimensional slices of various data cubes as columns. The views of the geographic hierarchy make it easy to select subsets of areas, for example 'all data zones in Glasgow'.

3.4 Tools for data managers

An important part of the system is the tooling provided to data managers, to assist them in maintaining and extending the data collection. Most of the data managers have little knowledge of Linked Data, and so tools have been provided to encapsulate the necessary knowledge of data modelling and RDF representation. Users create input data in the form of a table, with one statistical observation per row, and a column for each dimension, observation value or unit. This is uploaded and converted to RDF using a pre-defined transformation process, built using the 'Grafter' ETL library [9].

This approach has proved successful: the data managers find it relatively simple to produce the required input format, and it has enabled the Scottish Government to steadily expand the collection of data. One or two new datasets are added most weeks and there is an objective to add the data behind all releases of official statistics to the system by the end of 2017.

The 'Drafter' component illustrated in Figure 1 manages the workflow of loading and checking 'draft' data prior to publication. It applies access control to data for authenticated users, at the level of named graph. Different collections of named graphs are collected into 'draft sets' allowing multiple coherent updates to be prepared and reviewed at once. A subset of named graphs in the system are 'live' and viewable to external users without authentication.

Both the data import transformations and the editing of data via the Drafter system can be controlled by API as well as via the user interface, allowing automation of data loading processes.

4 Challenges and lessons learned

Many aspect of the system have been a great success: data managers can publish data with reasonable effort; the system is reliable with generally good performance; users can find and extract data flexibly; everything on the site has a URL that can be linked to, and data can be extracted by SPARQL to support external visualisations and applications.

Nonetheless, a number of areas have been identified where improvements would be beneficial. Many users report finding the system initially confusing or complex, though generally find it useful and powerful once some initial guidance or training is provided. Improvements to separation of the developer-focused and analyst-focused parts of the system are planned, as well as making text instructions clearer on user interface pages.

The value of visualisations as an easy and engaging entry point to data is recognized and development work is in progress to enrich this aspect of the system.

Another area of improvement is around systematic management of vocabularies, both to ensure better documentation and to encourage greater interoperability of different datasets.

The approach to data loading via a tabular template has worked well: improvements are planned to the level of validation of input data, tying in to the plan for stricter management of vocabularies.

The complexity of geography is a challenge and users need support. Improvements to the geography browsing and selection interface are in progress, with richer use of maps for viewing and selecting data.

Uptake of the SPARQL interface to data has been slow: the ambition is that external analysts and 'data intermediaries' will extract data in the system for use in their own visualisations and applications. The number of successful instances of this is currently small, though those which have been produced work well. The need to learn SPARQL appears to be a substantial obstacle, and so Swirrl (together with colleagues in the OpenGovIntelligence project [10]) are interested in developing a simpler JSON API to data cube data, designed to be more accessible to mainstream web developers.

References

1. http://statistics.gov.scot
2. http://statistics.gov.scot/data
3. http://statistics.gov.scot/def/concept/folders/themes
4. http://statistics.gov.scot/def/concept-scheme/benefit
5. http://statistics.gov.scot/def/concept-scheme/benefit-type
6. http://statistics.gov.scot/vocabularies
7. https://github.com/epimorphics/IntervalServer/blob/master/interval-uris.md
8. https://gss.civilservice.gov.uk/wp-content/uploads/2012/12/GSS-Geography-Policy-is-now-available.pdf
9. http://grafter.org
10. http://www.opengovintelligence.eu/