ONS Linked Statistical Data Pilot – Final Report

Author: Bill Roberts, Swirrl IT Limited
Reviewer: Ian Coady, Office for National Statistics
Date: 20 May 2016
Version: 2

Contact: bill@swirrl.com, 07717 160378

Executive Summary

This is the final report of the 'RDF Data Cube Pilot Project', a collaboration between ONS and Swirrl IT Limited, with the following objectives:

- investigate the use of the RDF Data Cube vocabulary (a formal recommendation of the World Wide Web Consortium) as a potential data dissemination format for the statistical data published by ONS
- assess the potential of this approach to extend access to ONS data, particularly in machine readable ways, and to enhance the ability to interconnect related data from different sources
- investigate the potential to automate the process of extracting and transforming data from existing ONS systems to an RDF Data Cube dissemination format

A small but varied sample of ONS datasets was chosen to form the basis of the pilot. The ONS team (led by Darren Barnes) used the Databaker tool to convert those from their source formats to a standardised CSV format, with one statistical observation per row. This stage of the process managed the interpretation of varied and sometimes complex spreadsheet structures in which the data is handled internally.

The output of Databaker was passed to Swirrl, who created a generic data transformation process using the open source Grafter library, to convert from the CSV format into the RDF Data Cube format.

This process was successful in general, but issues were encountered leading to recommendations for alterations and extensions to the Databaker output format, to support easier and more automated processing by the Grafter CSV to RDF transformation.

The data resulting from this process was loaded into Swirrl's 'PublishMyData' software platform, which has pre-existing facilities for viewing, selecting, visualising or downloading data in the form of RDF Data Cube.

The pilot project led to a series of recommendations: simple adjustments to the output of Databaker to enable an automated process of data transformation and publishing; consideration of performance for browsing and display of the largest datasets in PublishMyData; and steps towards data standardisation and sharing of reference data to enable greater interoperability.

Overall the project demonstrated the feasibility of this approach to data publishing. Its potential value as an additional method for distributing the data that ONS collects could therefore be realised in an efficient way. Further steps are recommended to investigate and demonstrate how this approach can promote more effective interconnection of statistical data from multiple sources to support analysis of complex issues.

Background and Plan

ONS wants to investigate the potential for publishing statistical data as Linked Data, using the RDF Data Cube approach. This could complement existing ways that the ONS distributes the data it collects and has the potential to increase the value of ONS data to users in the public, private and charity sectors.

Linking data has the power to greatly improve our analysis and understanding of data, and this is increasingly recognised at national and international levels through documents such as the Cabinet Office's Data Sharing Discussion document or Scottish Government's strategy for improving data access and analysis for statistics and research.

This joining up of data has previously been done through methods such as data linkage[link to definition of data linkage] but increasingly the diversity and complexity of administrative and big data sets compared to traditional surveys and census data requires new thinking to allow these datasets to be queried and exploited in a way that is not always possible with more traditional techniques.

There is also a need to join statistical tables to fully understand data and to allow users of statistical data to innovate and to build new platforms for analysis that move beyond the existing relationships between statistical tables.

Linked data offers one methodology for exploiting these datasets by adding semantic structure to large datasets and by consolidating distributed datasets into a single queryable resource.

The need to structure data to query and join it also makes statistics a useful case study as statistical tables and their associated geographic and non-geographic classifications means that - at least in theory – they could be mapped against a semantic vocabulary with relative ease. The project described below was established to test this theory and to identify what issues and methodologies could be provided to support other National Statistical Institutes in the adoption of linked data.

ONS already provides reference data (rather than statistics) on administrative and statistical geography as Linked Data, forming an important set of reference points that enables statistical data to be joined and compared. This project explores the potential for extending that linked data publishing into a representative selection of statistical datasets from ONS.

Swirrl has applied this approach for other UK public sector clients, including DCLG, Hampshire County Council and the Scottish Government, as well as developing new tools and best practices as part of the EU funded OpenCube project. The experience gained through those activities has been used to inform the approach taken in this project.

The main objectives of this pilot project are:

- to demonstrate and learn from the process of transforming a selection of ONS datasets to Linked Data, using the RDF Data Cube vocabulary and publishing those datasets as Linked Data.
- to investigate the data flow and work processes in getting data from ONS back-end systems to a Linked Data website: in particular how the ONS Databaker tool set can form part of the end to end process, in concert with other software components.
- To test the processes and flows on a sample of ONS datasets and publish these temporarily through the PublishMyData platform

Databaker is a software tool developed on behalf of ONS. It was designed to help routinely convert spreadsheets to a consistently structured, flattened CSV format, with one statistical observation per row.

Internally, ONS has distinct streams of statistical production and each one may produce spreadsheets for publication in different styles, based on the needs of their differing user communities. Furthermore because ONS is greatly concerned with preventing misuse or misinterpretation of the data by the community, a lot of effort is put into formatting, to encourage the correct interpretation. The various statistical producers may regularly alter the format of their own statistical outputs - separating out tables, clarifying column headers etc. - and so ONS wanted a tool that could help quickly repeat the transformation process for subtly different inputs.

Databaker provides a useful interface for a powerful spreadsheet querying language. This language uses robust descriptions of the data that help ONS adapt to slight changes in spreadsheet formatting - so instead of specifying a row and column it might specify 'the first row in bold' for the header row and 'all subsequent rows and columns containing numbers, up until the first blank row' to specify the data within the table. This 'broad strokes' approach is robust even if more tables are added to the same worksheet or if the number of blank rows above or between tables changes.

Summary of approach

The pilot investigated the application of Linked Data to representation of statistical data and, in particular, the use of a World Wide Web Consortium standard vocabulary called RDF Data Cube.

Linked Data is an approach to publishing and sharing data, first proposed and advocated by Sir Tim Berners-Lee[1]. It builds on the basic mechanisms of the web. Entities of interest (real world things, abstract concepts, data points) are assigned identifiers in the form HTTP URIs. Those URIs can be looked up in the

[1] https://www.w3.org/DesignIssues/LinkedData.html

normal way (whether through a browser or some other piece of software) and return a description of the relevant entity. That can either be a web page or one of a range of machine readable formats, usually based on RDF (the 'Resource Description Framework'). Because every entity has a web identifier, it is easy to link from one thing to another (hence the name 'linked data') and this provides a very useful mechanism for connecting data from different sources, for referring to authoritative data maintained by others, or for associating data with metadata or documentation.

Linked Data is sometimes known as '5-star data, based on a scale of 1-5 for assessing the degree of openness and usability of data. The Cabinet Office white paper of 2010 'Open Data: unleashing the potential'[2], advocated that where possible government should aim to release data following the 5-star approach.

The RDF Data Cube Vocabulary[3] is a formal recommendation of the World Wide Web Consortium for the representation of multidimensional statistical data as RDF. It draws heavily from the ISO standard SDMX, already in use by many statistical organisations. Data in the form of RDF Data Cube is available from a range of public sector organisations in the UK, including the Department for Communities and Local Government, the Environment Agency, the Scottish Government and a number of local authorities.

The following ONS datasets were selected for the pilot:

- Small area population estimates: by output area, individual year of age and gender, 2014
- Business financial data
- Employment data by Workplace Zone
- Annual Survey of Hours and Earnings data: earnings by parliamentary consituency, 2013
- Annual Survey of Hours and Earnings data: hours by parliamentary consituency, 2013

Each of these was provided to Swirrl in a CSV format, as produced by Databaker.

Swirrl carried out the following steps on these datasets:

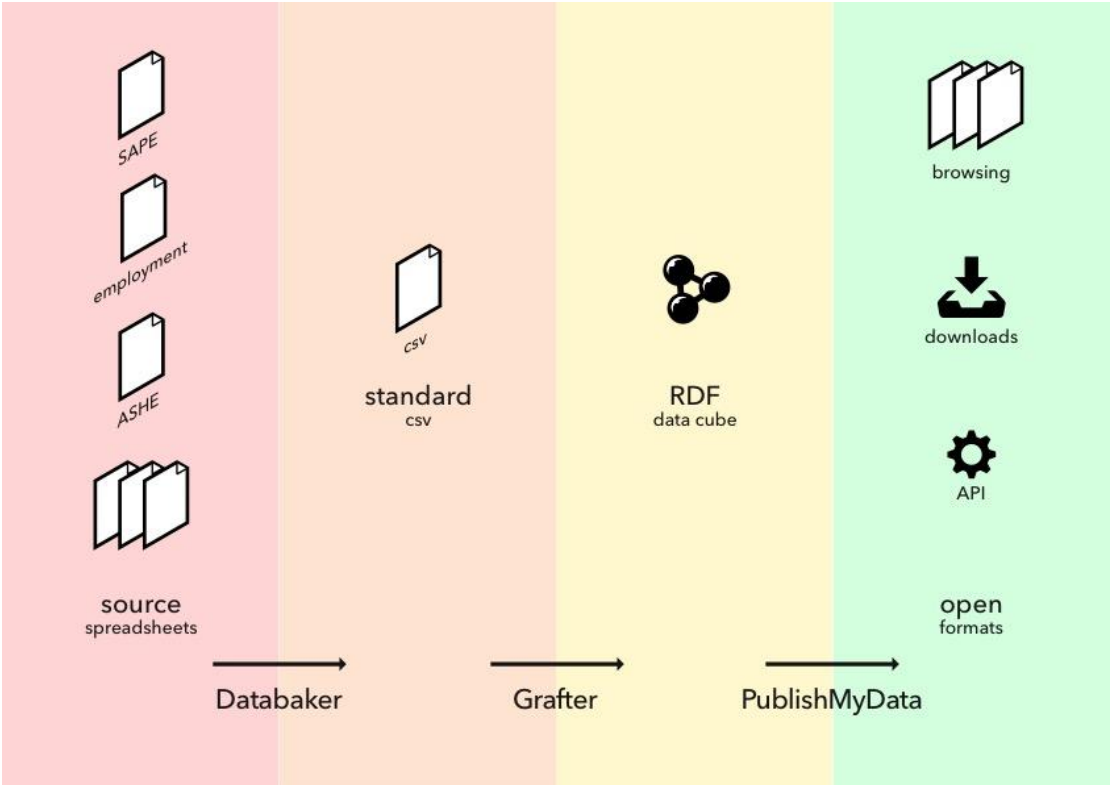
- used the Grafter software library to create software to transform the data from Databaker CSV format to RDF Data Cube. The code for these transformations is available at http://github.com/swirrl/ons-graft
- set up a server on Amazon Web Services, and installed an instance of Swirrl's PublishMyData platform for linked data publishing. That is available online at http://ons-pilot.publishmydata.com
- ran the Grafter transformations to obtain RDF representations of each dataset
- loaded the data into the PublishMyData platform

---

[2] https://www.gov.uk/government/publications/open-data-white-paper-unleashing-the-potential

[3] https://www.w3.org/TR/vocab-data-cube/

This process is illustrated in the following diagram:



Grafter (http://grafter.org) is an open source software library, written in the Clojure programming language, designed for a range of 'extract, transform, load' data integration activities. It is optimised particularly for the kind of data cleaning and transformation challenges encountered when converting data from typically tabular data sources to the graph structured data in RDF format: so is ideal for this pilot and for the transformation of Databaker outputs to RDF Data Cube.

The objective was to develop a transformation process that was generic enough to work with any dataset in the Databaker output format, so to provide ONS with a re-usable tool that could be used in future with other datasets. The five datasets listed above provided a test dataset to assess to what extent this had been achieved.

The output of the process was data in RDF format, following the RDF Data Cube vocabulary, a formal Recommendation of the World Wide Web Consortium. This is described in detail at https://www.w3.org/TR/vocab-data-cube/. For examples of the data produced, see the online data as hosted at http://ons-pilot.publishmydata.com.

Once in the standard RDF Data Cube format, the data then becomes suitable for use in with a range of standardised tools, including the SPARQL query language and a range of data filtering and visualisation approaches.

New identifiers created during the data transformation process were created in the domain http://statistics.data.gov.uk, to be consistent with the existing geography linked data published by ONS. These followed common conventions. Observations and datasets used URIs beginning:

http://statistics.data.gov.uk/data

Properties - for dimensions, measures, attributes etc – used URIs beginning:

http://statistics.data.gov.uk/def

Concepts and concept schemes also used URIs beginning

http://statistics.data.gov.uk

Existing identifiers were used for geographical areas, from the ONS geography dataset at http://statistics.data.gov.uk, and for time periods, using identifiers from http://reference.data.gov.uk

Issues Encountered

Making the pipeline generic

We have attempted to make the Grafter pipeline for converting Databaker outputs to RDF Data cube as generic as possible. With some minor alterations or additions to the format of Databaker output this should be achievable.

The issues encountered that required a degree of customisation of the pipeline for each different dataset were as follows:

1. some variation in the naming and capitalisation of column headers between the different files (possibly reflecting different versions of Databaker?).
   For example, “dim_id_1” vs “Dim1id” vs “Dim_ID_1”;
   “unit_of_measure_eng” vs “unit of meas eng” vs “"Unit_Of_Measure_Eng"
2. no definition of measure type (eg ‘count’) or measurement unit (eg ‘people’) in some cases
3. no definition of the reference period in the Databaker output file in some cases (eg WPZ)
4. no definition of the reference area in the Business Finance dataset.
5. dataset metadata, such as a name for the dataset, information on licence, publisher, contact details, description etc

The overall degree to which the code was customised for each dataset was relatively minimal and can be seen in the code files listed below. Addressing the above issues should be relatively straightforward and a generic Databaker to RDF Data Cube convertor using Grafter seems definitely achievable.

https://github.com/Swirrl/ons-graft/blob/master/src/ons_graft/input_data/ashe.clj

https://github.com/Swirrl/ons-graft/blob/master/src/ons_graft/input_data/business.clj

https://github.com/Swirrl/ons-graft/blob/master/src/ons_graft/input_data/sape.clj

https://github.com/Swirrl/ons-graft/blob/master/src/ons_graft/input_data/wpz.clj

A standard list of metadata items could be put together based on existing standard approaches to metadata, such as DCAT and taking into account the current work on 'StatDCAT-AP'.

A standard format and mechanism for associating the CSV file and its metadata should be established. The recommendations of the W3C 'CSV on the Web' working group should be closely considered for this requirement, see:

https://www.w3.org/TR/2015/REC-tabular-metadata-20151217/
https://www.w3.org/TR/tabular-data-model/#locating-metadata

Standardisation of the RDF data representation

There are a number of choices to make in designing the details of the data representation and the greater the degree of standardisation achievable between different data publishers, the greater the opportunities for combining data from different sources.

The choice of the RDF Data Cube approach is a significant step towards standardisation: it is a formal 'Recommendation' (i.e. a standard) of the World Wide Web Consortium, which comes with clear documentation and a set of tests in the form of SPARQL queries which can be used to determine whether a dataset is a well-formed RDF Data Cube.

However, in some aspects of the data representation, the RDF Data Cube offers several ways of doing things. Different choices of approach by different publishers make it more difficult for software tools (eg for visualisation or analysis) to work reliably with data from different sources.

These issues were addressed in a paper written by the OpenCube consortium and presented by Bill Roberts to the Semstats 2015 workshop: see http://semstats.org/2015/content/challenges-on-developing-tools-for-exploiting-linked-open-data-cubes/article.pdf

Discussion at that workshop led to the conclusion that supplementing the RDF Data Cube vocabulary with a best practice document or 'application profile' of some sort would be beneficial. There was discussion of using the existing W3C Community Group on Semantic Statistics (https://www.w3.org/community/semstats/) as a means of focusing efforts towards that goal, but to date there has been no concrete action since the October 2015 workshop. Swirrl is involved in a new EU project called 'OpenGovIntelligence' (www.opengovintelligence.eu), which includes 4 of the partner organisations from OpenCube. The OpenGovIntelligence project includes a specific task and associated budget to try to achieve this goal, so could be a useful ally to UK-based efforts to standardise the approach to Linked Data for statistics.

The other key issue around standardisation and interoperability is the choices of identifiers – for:

- dimension properties
- measure properties
- reference areas
- reference periods
- concept schemes and concepts for codelists and classifications

In many cases, each statistics publisher ends up creating their own URIs for these, even though the concepts they refer to may be identical to those used in data from other publishers.

The state of the art for reference areas and reference periods is relatively good: with the ONS URI set for administrative and statistical geography, based on the 9 character GSS codes, being a de facto standard, as is the use of the reference.data.gov.uk URIs for time intervals.

The GSS code based URIs are a good example of taking an existing well maintained set of codes and using them as the basis for a URI set. The ONS could play an important role by taking other existing sets of classifications and publishing those as maintained Linked Data URI sets or concept schemes, for use across the UK public sector and perhaps beyond.

There are some examples, such as age ranges and gender, which appear very frequently in statistical data and which could benefit greatly from standardisation on URIs for dimension properties and for the possible values.

Namespacing

If ONS adopts a Linked Data approach for publication of a wide selection of its data, some thought will be required into name-spacing and how dataset identifiers are assigned.

A standard starting point for dataset URIs is {domain}/data/{dataset-identifier}

Within the {dataset-identifier} part, a system of some sort will be required for generating unique and preferably human-readable identifiers for all published datasets.

Data marking

There are some standard data markers defined in the RDF Data Cube vocabulary, based on terms used in the SDMX 'CL_OBS_STATUS' code list. These can be associated with an observation as an 'AttributeProperty'. For example:

http://purl.org/linked-data/sdmx/2009/code#obsStatus-E - estimated value
http://purl.org/linked-data/sdmx/2009/code#obsStatus-I - imputed value
http://purl.org/linked-data/sdmx/2009/code#obsStatus-M - missing value

But that code list does not cover all cases in common use in UK government statistics and a standard concept scheme for use by ONS and others would be very beneficial.

Dataset size and processing performance

The size of datasets resulting from this process was influenced heavily by the size (and hence number) of geographical areas for which the data was presented.

This means that the Small Area Population Estimates, provided for all Output Areas in England (and also broken down into individual years of age hence more than 90 categories, and for gender categories of All/Female/Male) leads to a very large dataset – approximately 500 million RDF triples in total.

Other datasets were very much smaller: around 1 million triples for the Employment dataset (provided for around 50,000 workplace zones) and for the ASHE hours dataset (provided for 646 different geographical areas). The ASHE pay dataset also had 646 different areas, but was broken down in other ways, leading to around 3 million triples.

The smaller datasets were straightforward to process and load into the PublishMyData platform.

The SAPE dataset took many hours to process in Grafter, but (after some initial bugs were ironed out) can be handled row by row, so despite the large dataset size, does not lead to excessive server memory requirements.

Loading the SAPE data to PublishMyData was time consuming but not problematic. The large size of this dataset means that some options for viewing the data in the PublishMyData user interface led to timeouts. If datasets of this size are likely to be common in the broader ONS data collection, then some redesign and optimisation of the data retrieval methods in the data presentation layer may be required to ensure good performance.

In addition to the datasets themselves, the database contains definitions of vocabularies and codelists used in the data and some basic reference data on each of the geographical areas.

The data could be enhanced by enriching the geographical data with the kind of information already published by ONS at http://statistics.data.gov.uk, enabling additional ways of comparing and aggregating data about different places. This would enable existing data browsing and selection features of PublishMyData to be exploited.

Unused columns in Databaker output

The example datasets used as source data in this pilot include many empty columns. It is possible that all of these are necessary to cover the range of cases found in other ONS data, in which case it makes sense to retain them: it is straightforward for the Grafter code to ignore those columns. However it may be worth reviewing the Databaker output format to determine whether any columns are unnecessary.

Use of commas as thousand separators

In the WPZ datasets, numbers greater than 999 included commas as separators, for example "1,234". While it is possible to parse these in Grafter, it would be preferable if numbers were presented with no such separators.

Consistent capitalisation of data labels

This is a part of the broader question of consistent codelists, but the input data used in the pilot uses various approaches to capitalisation of the dimension names and values provided, for example 'Male' and 'age'. A standard approach would reduce the chance of accidental inconsistencies in data.

Different dimensions appearing in one dimension column

In one dataset, the ASHE 'hours' dataset, a single dimension-ID column in the Databaker dataset contained several different values for the dimension. In the Grafter transformation, the entries in this column are translated into dimension properties in the RDF Data Cube representation.

It is a requirement for RDF Data Cube that every observation in the cube has a value for every dimension property, so this mixing of dimension property names would lead to an invalid cube.

The options to address this are:

- split the data into multiple discrete data cubes – so that within each cube the rule about all observations have the same set of dimensions, but the different cubes could have different sets of dimensions
- model the data differently, to move that information from a dimension property to a dimension value
- expand the data cube to add extra dimensions, so that all observations have the same set of dimensions

The approach taken in the pilot was effectively the first of those above, but rather than produce multiple cubes, we took a single well-defined data cube through the full process, identifying the 'Hours Statistics' as the most important of the various dimension properties that appeared in a single Databaker output column.

Project outputs

The project led to the following main outputs:

- a demonstration online data publishing site: http://ons-pilot.publishmydata.com/
- a generic data transformation pipeline, using the Grafter library, for converting CSV output of Databaker into RDF Data Cube
- a set of recommendations to help improve the ability to automate the data transformation process and to make it easier to combine ONS data with data from other organisations

Recommendations

Based on the issues described in the previous section, we propose the following recommendations .

1. ONS should review the Databaker output format to tidy the various minor issues raised above
2. ONS (with input from Swirrl) should determine a standard list of metadata items and a mechanism for associating them with the CSV file produced by Databaker – together with the previous point this would enable a generic Databaker to RDF Data Cube convertor to be implemented in Grafter
3. Swirrl should update (and further test) the Grafter pipeline developed for the pilot (after steps 1 and 2 are complete) so that it can become a generic tool for ONS to convert the output of Databaker into RDF Data Cube, creating a flexible and efficient two-step process to preparing ONS data for publishing as Linked Data.
4. Swirrl should review the data retrieval and presentation process in PublishMyData to ensure good performance for very large datasets, such as SAPE.

5. ONS should encourage and contribute to an inter-organisation effort to develop an ‘application profile’ for RDF Data Cube, recommending a specific approach to the use of dimensions, measures, units etc
6. In collaboration with other UK public sector organisations, ONS should publish and maintain a series of the most commonly used codelists in Linked Data form, supporting interoperability of data between different organisations. This should be supplemented with a vocabulary of corresponding dimension properties.