Statistical Data Model Recommendations

Leigh Dodds, Open Data Institute, 9th November 2018

This report is the result of a short project conducted by the Open Data Institute on behalf of the Parliamentary Digital Service.

The primary goals of the project were to understand the current processes for collecting, managing and publishing statistical data by the House of Commons Library team and then recommend an appropriate data model for storing these statistics as Linked Data within the Parliamentary Data Platform (PDP).

The report provides:

- a summary of the current ways in which statistical data is collected, managed and published
- recommendations for how to express this data as Linked Data using the Data Cube ontology
- brief thoughts on creating a reliable import process for loading data into the PDP

Statistical data in the House of Commons Library

The statisticians working in the House of Commons library collect, manage and use statistics in order to:

1. publish a set of constituency dashboards that provide Members with a summary of key statistical indicators
2. publish both one-off and regularly updated research papers summarising key policy areas along with supporting statistics
3. answer ad hoc requests for data and information from Members

The team uses data from a variety of sources, including the ONS, Ordnance Survey, Department for Education and regulators like Ofgem.

Collection and processing of data

Statistical data is collected and processed in several different ways:

1. by taking existing statistical data from a single primary source, e.g. House Price Statistics from ONS
2. by aggregating existing statistical data from several primary sources, e.g. Constituency Population data by age from ONS, NRS, NISRA
3. by deriving new statistics from existing primary sources, to create, e.g. new indicators such as the recently launched City and Town Classification for Constituencies and Local

Authorities, or to re-project statistics to new geographies as done for the Broadband coverage and speed statistics
4. by generating statistics or data from other primary sources, e.g. Hansard
5. by collecting new primary data, e.g. constituency results and statistics for elections

In future, the statistical team hope to ingest more data from APIs, e.g. as exposed by ONS or EuroStat. Although the need to frequently combine data from several sources, and the lack of a common API across statistical publishers means creating a national dataset will often require more than just a simple import and transformation process.

Statistical geographies

The Commons Library primarily reports constituency level statistics. Data from existing sources is re-projected to constituency level where it is not directly available. New constituency-level indicators have also been created to address gaps in existing official statistics that focus on ONS or EU statistical geographies.

Over time the statistical team plan to publish statistics for smaller areas, specifically to the level of Middle Layer Super Output Areas (MSOA) which is widely available in existing datasets.

Publishing of statistical data

The Commons Library publishes statistical data in several different ways:

- as dashboards, using PowerBI which provides an interactive interface over a set of normalised, well-formed CSV files
- as CSV and Excel downloads, associated with individual research briefings
- as charts and tables within research briefings

The data is currently largely collected and normalised by hand, e.g. to prepare it for publication using PowerBI. The tabular datasets are often reorganised in order to generate the required dashboard interfaces, but are generally well-formed and machine-readable.

Over time, the statistical team expect to publish data in other forms, e.g. as more interactive charts and maps. This will require more functionality to query and manipulate the datasets. This is the primary driver for moving data from the existing content management system and PowerBI environment into the PDP.

As they are the most widely used and already existing in a well-formed, machine-readable format, in this report we focus on the statistical data used to populate the PowerBI dashboards.

Modelling statistical data

We recommend that the W3C Data Cube ontology[1] is used as the standard data model for representing statistical data in the PDP.

The currently unused Stats Series ontology closely aligns with the DataCube model, but doesn't address some cover all of the requirements necessary to correctly capture statistical data, e.g. describing the full range of measurements and dimensions associated with individual observations.

The Stats Series ontology does cover some additional requirements that are not covered by the Data Cube ontology. E.g. description of individual statistical releases, spatial coverage, etc. But these could also be addressed by using other vocabularies, e.g. the Data Catalog Vocabulary[2] (DCAT).

The following sections introduce the Data Cube ontology and illustrate how it would be applied to three example datasets that are currently included in the constituency dashboards.

We then identify the general steps required to define data cubes for any statistical dataset that may be published in the PDP.

A brief overview of the Data Cube ontology

The Data Cube ontology is a domain model for describing statistics. The model represents a statistical value, e.g. the population of an individual constituency as recorded in 2018, as an **Observation**.

An individual Observation consists of:

- values for one or more **dimensions**, that collectively describe what is being observed. E.g. the group of *people aged 18-24*, in *Aberavon* in *2018*.
- one or more **measures**, that record the statistical value being reported, e.g. the *number of people*
- optionally, one or more **attributes** that qualify the value being reported, e.g. a *unit of measurement* or a *quality indicator* ("estimated", "revised")

A statistical **Dataset** consists of a set of Observations.

All of the Observations in a Dataset will have a consistent set of dimensions, measures and attributes which are defined in a **Dataset Structure Definition**.

It is relatively straightforward to take a tabular statistical dataset and represent it as a set of Observations belonging to a Dataset. A Data Cube dataset can easily be queried to create multiple tabular views from individual Observations.

[1] https://www.w3.org/TR/vocab-dcat/
[2] https://www.w3.org/TR/vocab-dcat/

The dimensions and attributes in a dataset will all have a pre-defined set of values. These values will be drawn from one of the following:

- a generic SKOS vocabulary providing a set of values that are reused across multiple datasets, e.g. a standard set of age groups taken from a national census, or a set of economic indicators like SIC codes[3]
- a dataset specific SKOS vocabulary, e.g. in cases where a dataset uses a non-standardised set of age groupings
- instance data conforms to an existing ontology, e.g. geographic areas or constituencies, etc

Frequently, the values of a dimension will be organised into a hierarchy. This is commonly the case with geographic areas, e.g. NUTS area codes or the ONS statistical geographies.

A statistical dataset will often include observations that report values at several levels within an individual hierarchy, e.g. for a country, region and a constituency. This is necessary because small area statistics may be adjusted for anonymisation or reporting reasons and can't be aggregated to calculate the value for larger geographic areas.

Slices

The Data Cube ontology defines some additional classes that can help to organise and describe parts of a statistical dataset. These are useful, but optional.

A common requirement is the ability to describe a set of observations that describe a time series, e.g. the population of people aged 18-24 in Aberavon over time.

A **Slice** through a dataset identifies a set of Observations that have the same values for some dimensions. E.g. a Slice might identify observations that have specific values for place ("Aberavon") and an age group ("18-24"), but not time.

At present there isn't a clear need to define Slices for the statistical data to be imported into the PDP. But these may be necessary to:

- advertise useful subsets of the data to data consumers
- help organise data in a way that supports querying of the dataset
- attach commentary and notes that can be included in reports and visualisations

Dataset metadata

The Data Cube ontology defines a DataSet class but does not provide terms for capturing the additional metadata that associated with those datasets. A range of existing ontologies already address this use case.

---

3 https://www.gov.uk/government/publications/standard-industrial-classification-of-economic-activities-sic

We recommend using DCAT as a means of capturing the descriptive metadata for a dataset.

The DCAT standard is currently being revised[4] to include additional terms. This will includes terms to describe spatial[5] and temporal[6] coverage as well as provenance information[7]. Collectively, these cover some of the additional functionality covered by the Spatial Series ontology.

With regards to provenance, we recommend that data in the PDP should clearly identify where a dataset:

- is simply a copy of a third-party dataset
- has been derived by the Commons Library team from one or more primary sources
- is a new primary dataset that has been created by the Commons Library statisticians

In our brief requirements gathering exercise we didn't identify a clear need to manage multiple versions of statistical datasets, e.g. to distinguish between data reported as part of separate annual statistical releases.

Pending further analysis by the PDS team, our recommendation would be to initially maintain just a single current version of each dataset, replacing observations and updating the relevant statistical geographies as new data is added to the platform. Metadata about provenance and original sources can be used to reflect the versions of the primary sources that were used to create the version in the PDP.

Worked examples

The following sections provide some worked examples that illustrate how existing datasets published by the Commons Library can be mapped to the Data Cube vocabulary.

As noted in the Appendix, a set of supporting turtle files have been provided to illustrate the technical aspects of the mapping.

Broadband speeds

The Commons Library constituency data on broadband coverage and speeds is a relatively simple statistical dataset. It consists of:

- a single **dimension**: the *statistical area*, with data reported at two levels: for the constituency level and the whole of the UK
- seven different **measures**, including the "*Percent of lines with superfast availability*", "*Average download speed, in megabits per second*", "*Number of premises with connectivity below the Universal Service Obligation*"

[4] https://w3c.github.io/dxwg/dcat/
[5] https://w3c.github.io/dxwg/dcat/#Property:dataset_spatial
[6] https://w3c.github.io/dxwg/dcat/#Property:dataset_temporal
[7] https://w3c.github.io/dxwg/dcat/#Property:dataset_wasgeneratedby

- two **attributes** that are associated with constituency level observations: a *link to a broadband speed map* and a *link to a connectivity map*

The data doesn't include a time dimension as the data is only reported for a single year.

The Data Cube ontology recommends two approaches for datasets that include multiple measures. In the first approach every observation includes values for all measurements. In the second, each observation reports only a single measure. This means that multiple observations need to be created, one for each measurement and set of dimensions.

As the specification notes, there are design trade-offs to be made between these approaches. The most significant being in the volume of data stored. For simplicity, we recommend associating multiple measures with each observation. This reduces amount of data to be stored, will simplify data conversion and make the data easier to query.

This approach does requires associating a unit of measure with each Measure property, which slightly limits the ability to reuse properties across datasets. However our brief review suggests that many of these measures are already dataset specific, so this is a reasonable trade-off.

House prices

The Commons Library constituency data on house prices is slightly more complex than the broadband speed dataset, as it includes additional dimensions and geographic areas. The dataset consists of:

- two **dimensions**: the *statistical area* and *reporting date*. There are multiple statistical areas at which data is reported: the constituency, the region and the country as a whole
- five **measures**: "*median house price*", "*percentage change, one year*", "*percentage change, 5 years*", "*median salary*", "*salary to wage ratio*"

This dataset illustrates how the values of a dimension can refer to resources described elsewhere in the PDP. In this case via the *statistical area* dimension which will refer to Constituency Groups and other geographical areas.

The statistical areas used in the dataset include:

- Westminster Constituencies, e.g. Berwick-upon-Tweed (E14000554). These map to current resources in the PDP, e.g. the Constituency Group for Berwick-upon-Tweed
- European Electoral Regions, e.g. North East (E12000001)
- Country, e.g. England and Wales (K04000001)

The need for additional geographic locations is a common theme across all of these examples. In order to import just the existing datasets the PDP will need to be updated with additional geographic areas. As noted in the previous section, future work will require additional data.

It will also be important to include the hierarchical relationships between geographies to allow data users to easily find data for broader geographic areas. Comparison of constituency statistics against regional and national indicators is a common feature of the existing dashboards. In the current Commons Library datasets, these hierarchical relationships can be inferred from the CSV files. But in others additional data may be needed, e.g. by importing data directly from the ONS or elsewhere.

The PDS Place ontology can be used to describe these hierarchical relationships. While they are not often represented in statistical geographies, it may also be useful to include relationships like "touches" from the OS spatial relationship ontology. This can be helpful when querying for data from neighbouring statistical areas to support local comparisons.

Population by age

The Commons Library constituency data on population by age illustrates how dimensions in a statistical dataset may need to draw values from a SKOS ontology.

This dataset consists of:

- three **dimensions**: *statistical area*, *year*, and *age group*. As with the house price data, the statistical areas span multiple geographies. The age group dimension uses a set of categories to describe age ranges in the population
- two **measures**: "*population*", and "*percentage of people in constituency*"

The age range dimension refers to a controlled list of 13 categories which would be represented as a SKOS vocabulary. The values are:

    - (*decile ranges)*: 0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79
    - (*groupings*): 0-15, 16-64, 65+, 80+
    - total

This usefully illustrates that the controlled values used in dimensions will not necessarily fall into a simple hierarchy.The groupings may overlap so cannot be replaced by calculating aggregates during querying.

Reviewing the original sources, it looks like for this dataset, the Commons Library statisticians may be using their own groupings, as the ONS and NRS datasets are more granular. This also highlights the need to maintain custom SKOS vocabularies to support the published data.

Defining a Data Cube

As the worked examples hopefully illustrate, defining a Data Cube consists of:

1. identifying the dimensions used in the dataset and, where necessary, for each new dimension:
    a. declaring a new Dimension property with the appropriate name, etc

    b. declaring the range of the property (if it refer to existing resources) or the SKOS vocabulary from which it will draw values
2. identifying the measures used in the dataset and, where necessary, for each new measure:
    a. declaring a new Measure property with appropriate name, etc
    b. identifying the range of the property, e.g. as an XML schema data type
    c. identifying the units of measurement, e.g. “megabits per second”
3. identifying the attributes used in the dataset and, where necessary, for each new attribute
    a. declaring a new Attribute property with appropriate name, etc
    b. declaring the range of the property and/or the SKOS vocabulary from which it will draw values
4. ensuring that the values of dimension and measure property are all available, e.g.
    a. by defining or importing the necessary geographic area or other resources
    b. by defining or importing the appropriate SKOS vocabularies
5. combining the dimensions, measures and attributes into a Data Structure Definition that describes the structure of the Dataset.

The sample turtle files accompanying this report have been created by following these steps.

As the examples show, there will be some dimensions, such as statistical area, age-group, year or date that are likely to be repeatedly used across datasets. In other cases, the dimensions and measures will be dataset specific.

Just as there is a need to include additional geographic areas, there may be a need to create additional instance data to cover other types of resources.

For example if the educational attainment dataset were to provide data on individual schools, then this would require the PDP to include resources for each school[8].

Sketching a reliable import process

This steps involved in defining the schema and structure of a dataset, will be part of a broader process that will result in new RDF data being loaded into the PDP.

A future process might look like the following:

1. the Commons Library statisticians collect/aggregate/generate the statistics required to populate a constituency dashboard
2. the data is turned into a well-formed, denormalized CSV file which is loaded into PowerBI to create dashboards
    a. this step is iterated to ensure the tabular data can be used to generate the appropriate functionality

[8] Note: school level statistics are currently shown in the dashboard, but these don’t seem to be included in the dataset download

3. once a stable tabular format is available, the statisticians can define the necessary component properties, controlled vocabularies, etc necessary to convert the dataset into RDF, following the process outlined in the previous section
4. this configuration is used to drive a conversion tool, as part of the existing orchestration step, that will generate RDF to load into the PDP
    a. the conversion tool could use the Data Structure Definition to help validate the generated data and ensure it is consistent

To make the process reliable the tabular format for the statistical data and the Data Cube definition will need to be as stable as possible.

There are some existing tools that may help with the conversion of the CSV data into RDF. This includes Data Baker[9], which is currently used internally by the ONS. And Grafter[10] which is used to help generate RDF for a variety of data portals.

Summary

The RDF Data Cube vocabulary provides an existing ontology that can be used to describe and import statistical data into the Parliamentary Data Platform. The worked examples in this report show how three different datasets currently published by the House of Commons Library can be described using the RDF Data Cube ontology.

The ability to join data across dimensions (e.g. based on location) and across other data in the platform (e.g. relating statistics to activities by MPs) will provide a great deal of flexibility in querying across diverse datasets.

In exploring the mapping to RDF Data Cube, this report has made some assumptions about requirements, e.g. around the versioning of datasets and/or statistical geographies, which may need further review.

Additional work may need to be done to explore other areas in more detail, e.g. the appropriate level of provenance information to record around the processing of data. But this should require drawing on additional vocabularies, e.g. DCAT, and will not impact the core statistical data model which is the focus of this project.

The steps to map a statistical dataset to an RDF Data Cube is relatively straightforward. With some support and guidance, this task could be given to the Commons Library statisticians as part of the preparatory work required to publish data. Additional work will be required to ensure that the appropriate controlled vocabularies and geographic data is present in the platform.

[9] https://sensiblecodeio.github.io/quickcode-ons-docs/
[10] http://grafter.org/

Appendix: Example Files

Accompanying this report is a zip file containing several example files:

- Three turtle files that illustrate how to the worked examples in the report can be expressed as RDF. The files include data structure definitions as well as a couple of example observations
- Two diagrams that illustrate the RDF graph generated from these data. These diagrams have been generated from a slightly simplified version of the broadband speed worked example. One diagram focuses on the dataset and the Data Structure Definition, the second shows the observation data.