Creating a spatio-temporal "Data Feed" API for a large and diverse library of historical statistics for areas within Britain

Humphrey Southall[*1] and Michael Stoner[†1]

[1]Department of Geography, University of Portsmouth

January 9, 2015

Summary

The GB Historical GIS holds 14m. diverse statistical data values in a uniform structure linked to a geospatial ontology of reporting units and a domain ontology of statistical concepts. This paper describes the addition of a Linked Data API enabling programmatic access to this "big data" structure and discusses topical and spatial sub-setting.

KEYWORDS: Social statistics, Open Linked Data, Datacube Vocabulary, Historical GIS

1. Introduction

Numerous examples, notably Google's search engine, have shown the vast power of even quite simple analytic tools when accessing really large amounts of unstructured text. Similar approaches fail with statistics because individual data values lack intrinsic meaning, and the frameworks which give statistics meaning often divide them into silos, making automated analysis possibly only within individual silos, not across them. For example, the UK Data Service's catalogue holds consistent information about very many "studies", enabling users to download particular datasets. Their internal structure was hopefully documented by dataset creators to enable users to unpack them, but to the repository system data and documentation are essentially binary large objects, so you can't "analyse the Data Service" *in toto*.

Although the organisation of data archives into studies and datasets has changed little over forty years, their development of new metadata frameworks enables radically new approaches if unencumbered by legacy holdings. Preparations for the 2011 census included much discussion of a "data feed" Applications Programming Interface, leading to a Census Web Services working group; and, eventually, to the Office of National Statistics launching an experimental API (https://www.ons.gov.uk/ons/apiservice/web/apiservice) and the Data Service launching InFuse, an API-based replacement for Casweb. However, ONS's API is constrained by the organisation of the c. 8 billion 2011 data values into about 400 "Tables", a paper-derived concept now constraining the almost purely digital. The InFuse API remains private and undocumented.

We describe here a more integrated approach, influenced by these projects but, perhaps surprisingly, less constrained by legacy data: our Great Britain Historical GIS relies mainly on data we computerise ourselves from old paper reports. Two more closely related systems are the Irish government's census gateway (http://data.cso.ie/index.html; Maali *et al*, 2012), although our historical work must address long periods of time, diverse reporting geographies and diverse categorisations of similar things; and the Dutch CEDAR project (Meroño-Peñuela *et al*, 2014).

[*] Humphrey.Southall@port.ac.uk

[†] Michael.Stoner@port.ac.uk

2. Holding a large and diverse library of statistical data in one table

Our data library currently comprises 14,099,469 data values including 8,525,126 numbers drawn from all British censuses 1801-2001 down to parish level; 3,202,448 death counts categorised by district, cause, age and gender from vital registration reports 1851-1910; 70,580 vote counts covering every candidate in every constituency in every parliamentary election 1833-2008; and 90,958 counts from farm censuses 1866-1971: a fair summary of the quantitative history of Britain's localities. Each value forms one row in the central "data table" within our Postgres database, including 2,100,844 copies of the number zero, 1,136,484 copies of one, etc.

Any particular zero is given meaning by other data table columns, mostly identifiers given meaning by sub-systems: a date; an ID defining geographical coverage; a source identifier; an "acknowledgment ID" identifying transcribers; and a "cell reference" recording *what* the number measures by locating it within an n-dimensional hypercube, or "nCube".

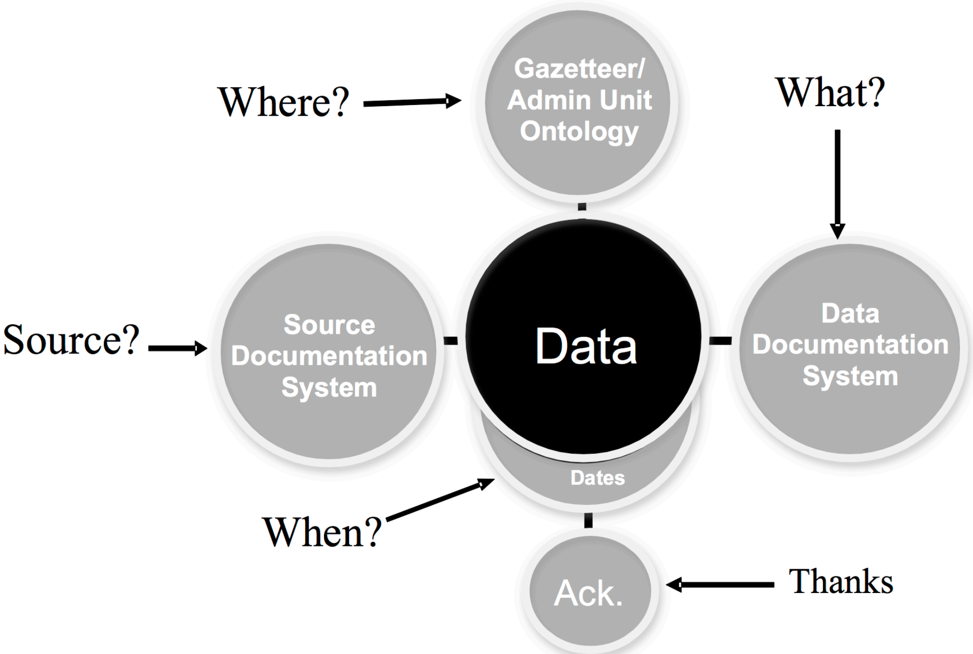
Where?
Gazetteer/ Admin Unit Ontology
What?
Source?
Source Documentation System
Data
Data Documentation System
Dates
When?
Ack.
Thanks

Figure 1 Simplified structure of Great Britain Historical GIS

This last is based on the Aggregate Data Extension to the Data Documentation Initiative (DDI) standard (Southall, 2011; http://www.ddialliance.org/) and enables *Vision of Britain* to automatically select the most appropriate way to visualise any given dataset (Southall, 2008). In brief, the dimensions of an nCube are defined by the variables and categories in the underlying microdata, such as age group, gender and occupation; nCubes and variables are organised within a hierarchy of topics; and labels and explanatory text can be held for all these entities. We have shown that an absolutely fixed set of database tables can not only hold any amount of data, but also cover an endlessly expanding geographical scope and set of topics.

3. Defining the API

We have also implemented a Linked Data API accessing the PastPlace gazetteer, returning information about both "places", such as Portsmouth, and administrative units such as Portsmouth Registration District and Portsmouth County Borough, using separate systems of numerical place and unit identifiers, and consequently two sets of Uniform Resource Identifiers (URIs) (Southall, 2012). The API can be searched using a placename or by specifying a bounding box.

Our new PastPlace DataCube API is based on the World Wide Web Consortium's Datacube Vocabulary (Cyganiak *et al*, 2014) and is implemented as a separate application running within Tomcat using Apache Jena, an open source Semantic Web framework for Java (https://jena.apache.org/). Jena serialises RDF graphs into different output formats including RDF/XML, Turtle, and Notation 3.

The sample below contains just one actual data value, 53,058, the total population of Portsmouth Registration District (unit 10154984) in 1841. That value appears as the bottom line, and dates and location are identified concisely so most output captures data semantics. Where possible we link to externally defined vocabularies, such as Dublin Core (e.g. dc.publisher) but for now much is self-defined (Kramer *et al*, 2012).

```
@prefix admingeo: <http://data.ordnancesurvey.co.uk/ontology/admingeo/> .
@prefix dc:    <http://purl.org/dc/elements/1.1/> .
@prefix gbhgis: <http://gbhgis.geog.port.ac.uk/> .
@prefix obs:   <http://obs.gbhgis.geog.port.ac.uk//uri/#> .
@prefix qb:    <http://purl.org/linked-data/cube#> .
@prefix rdfs:  <http://www.w3.org/2000/01/rdf-schema#> .
@prefix sdmx-measure: <http://purl.org/linked-data/sdmx/2009/measure#> .
@prefix sdmx-subject: <http://purl.org/linked-data/sdmx/2009/subject#> .

<http://dataset.gbhgis.geog.port.ac.uk/uri/#TOT_POP:now>
        a                      qb:dataset ;
        gbhgis:hgisMeaningDDI  "TOT_POP:now" ;
        dc:publisher           "gbhgis" ;
        dc:title               "Current Total Population" ;
        qb:slice               <http://gbhgis.geog.port.ac.uk/TOT_POP:now/1841> ,
                               <http://gbhgis.geog.port.ac.uk/TOT_POP:now/1851> ;
        qb:structure           <http://datasetdefn.gbhgis.geog.port.ac.uk/uri/#TOT_POP:now> .

<http://datasetdefn.gbhgis.geog.port.ac.uk/uri/#TOT_POP:now>
        a            qb:DataStructureDefinition ;
        dc:subject   "subjects"^^<java:com.hp.hpl.jena.rdf.model.impl.ResourceImpl> ;
        qb:SliceKey  "http://gbhgis.geog.port.ac.uk/sliceByAUO" .

<subjects>  sdmx-subject: "1.1" .

<http://gbhgis.geog.port.ac.uk/TOT_POP:now/1841>
        a                      qb:slice ;
        gbhgis:hgisMeaningDDI  "TOT_POP:now" ;
        gbhgis:ref-period      "1841" ;
        qb:Observation        "http://obs.gbhgis.geog.port.ac.uk//uri/#Observation-1032424" ;
        qb:sliceStructure      "http://gbhgis.geog.port.ac.uk/sliceByAUO" .

obs:Observation-1032424
        a                      qb:Observation ;
        gbhgis:ref-auo         "10154984" ;
        qb:dataset             "http://dataset.gbhgis.geog.port.ac.uk/uri/ - TOT_POP:now" ;
        sdmx-measure:obsValue  "53058"^^<http://www.w3.org/2001/XMLSchema#decimal> .
```

Figure 2 Sample output from prototype PastPlace datacube API

4. Subsetting mechanisms

The underlying database structure makes dumping out data values straightforward, but current software just outputs the whole data table in the above format – "big data" with a vengeance. The challenge is to provide useful sub-setting mechanisms, and here we draw on two distinct mechanisms already developed for the Vision of Britain download sub-system (www.VisionOfBritain.org.uk/data).

The first and spatial strategy starts with the client specifying a point coordinate and a broad statistical theme. The system returns a list of specific reporting units whose boundary polygons cover the point, and associated nCubes within the theme. The client can seek additional information about both units and nCubes, then extract data values for selected unit/nCube combinations, so obtaining local time series.

The second strategy starts with the client reaching an nCube by moving down the topic hierarchy or searching by keyword, the results being relevance-ranked based on where and how frequently the search term appears in the various metadata elements linked to the nCube. The system returns the "unit types", effectively GIS coverages, for which data exist within the nCube; the dates for which data exist; and the number of data values for each type/date combination. Extracting all data for specific combinations essentially populates a statistical map.

5. Conclusion

Current funding does not extend to developing client software. However, we get many data requests from academics, the media and government, with current projects for the EU and Greater London. We hope that providing programmatic access will enable them to start exploring the analytic potentials of a "big" and genuinely integrated dataset spanning all Britain's localities over two hundred years.

6. Acknowledgements

This research is part of an Arts and Humanities Research Council Big Data Project (award AH/L01002X/1).

7. Biography

Humphrey Southall is Professor of Historical Geography and Michael Stoner is Senior Research Associate at the University of Portsmouth.

References

Cyganiak R, Reynolds D and Tennison J (2014) The RDF Data Cube Vocabulary. World Wide Web Consortium, Cambridge MA (http://www.w3.org/TR/vocab-data-cube/)

Kramer S, Leahey A, Southall H R, Vampras J and Wackerow J (2012) *Using RDF to describe and link social science data to related resources on the Web: leveraging the Data Documentation Initiative (DDI) model.* Working Paper. Data Documentation Initiative, Ann Arbor, Michigan.

Maali F, Cyganiak R, and Peristeras V (2012) A Publishing Pipeline for Linked Government Data, in Semperl E et al (eds) *The Semantic Web: Research and Applications: 9th Extended Semantic Web Conference,* 778-92. Springer, Berlin.

Meroño-Peñuela A, Guéret C, Ashkpour A and Schlobach S (2014) CEDAR: The Dutch Historical Censuses as Linked Open Data. *Semantic Web Journal* (under review but online: http://www.semantic-web-journal.net/system/files/swj878.pdf).

Southall H R (2008). Visualization, data sharing and metadata, in Dodge M, McDerby M and Turner M (eds) *Geographical Visualization: Concepts, Tools and Applications*, 259-75. Wiley, Chichester.

Southall H R (2011) Rebuilding the Great Britain Historical GIS, Part 1: building an indefinitely scalable statistical database. *Historical Methods*, 44 (3), 149-59.

Southall H R (2012) Rebuilding the Great Britain Historical GIS, part 2: a geo-spatial ontology of administrative units. *Historical Methods*, 45 (3), 119-34.