

Encapsulating DDI Metadata within a Spatio-Temporal Data Feed for British Historical Statistics

Humphrey Southall & Michael Stoner
(Great Britain Historical GIS/
University of Portsmouth)

(gbhgis@port.ac.uk)

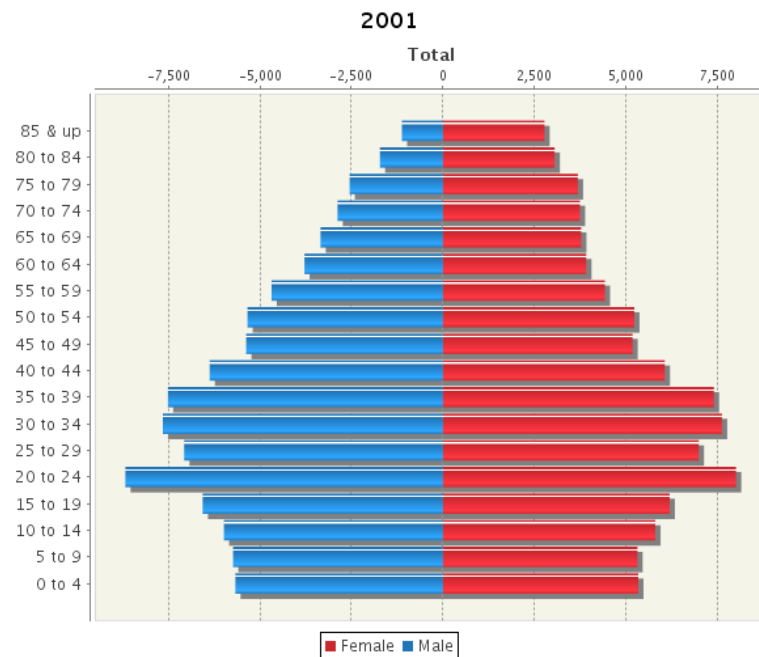
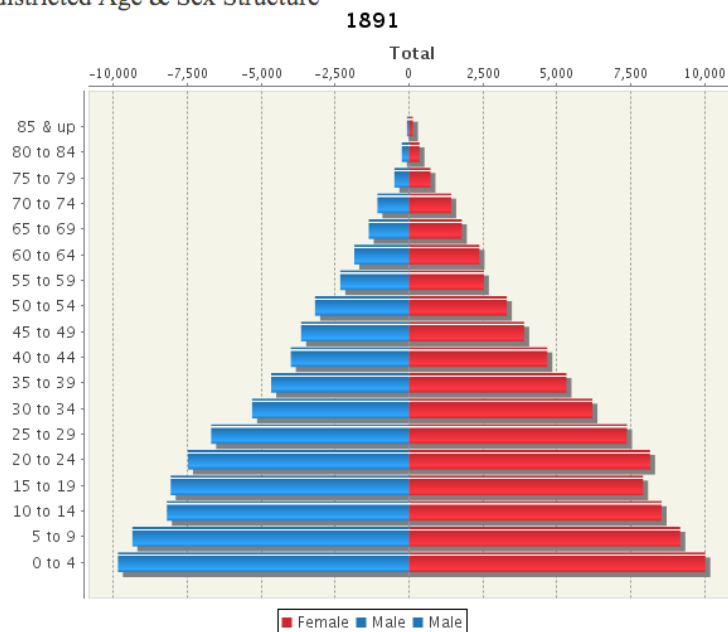
Statistical data feeds:

- ONS experimental **OpenAPI**
 - Launched November 2014
 - <https://www.ons.gov.uk/ons/apiservice/web/apiservice>
 - Access data values within each census table
 - ONS also have easy to use tools for selecting and visualisation
 - ONS no longer reports aggregated data totals from the previous census with current census reporting areas.
- **InFuse:**
 - New MIMAS/UK Data Service interface, which sits on top of a currently private API
 - Combine, and link data, across a complete census.
 - Programmatic disclosure rules

GBH GIS Data

- Integrated data structure for 14 million data values
 - Allows redistricted data to be presented as a time series, e.g. Portsmouth Pop. Pyramid:
http://visionofbritain.org.uk/unit/10168181/cube/AGESEX_85UP

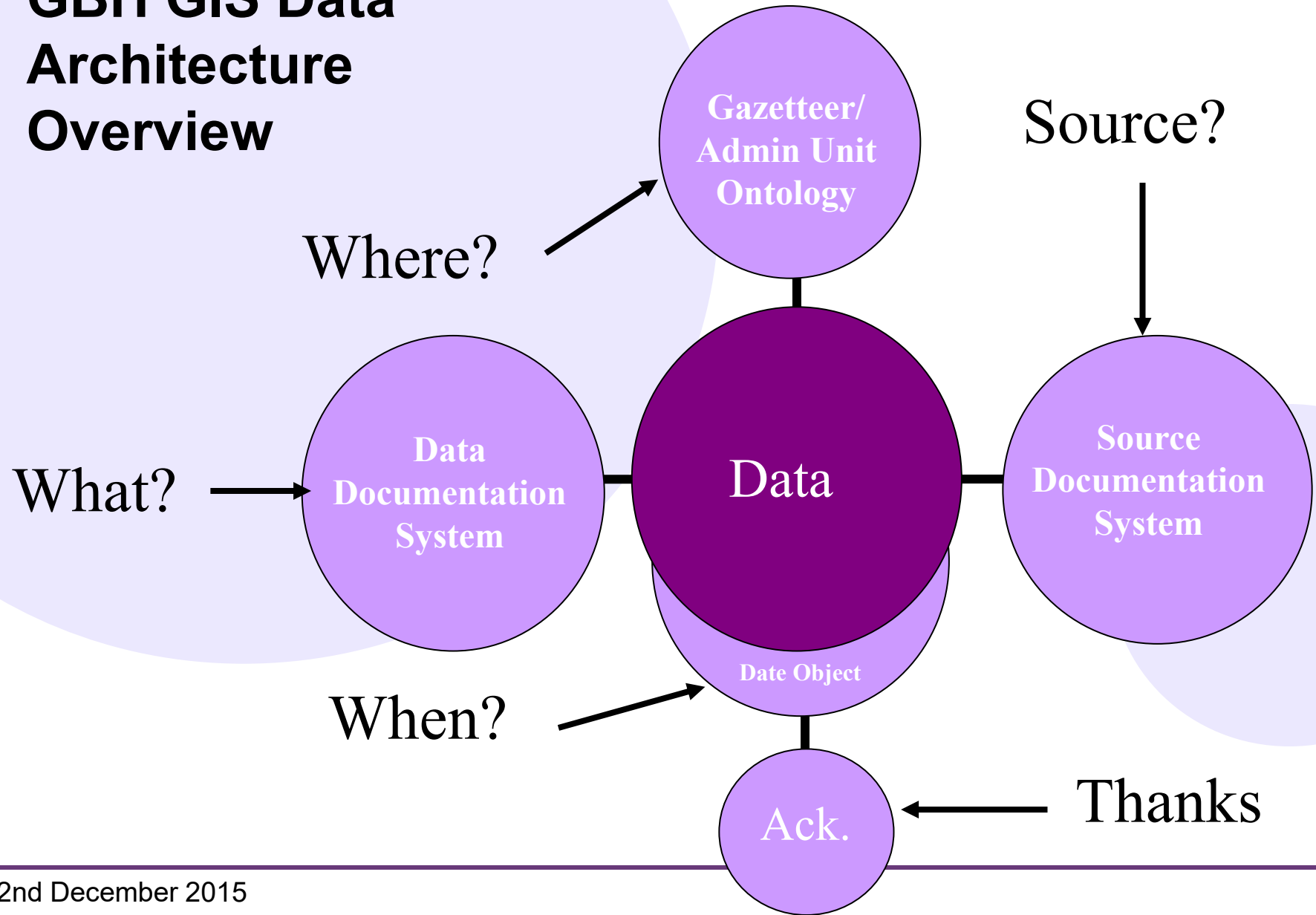
Redistricted Age & Sex Structure



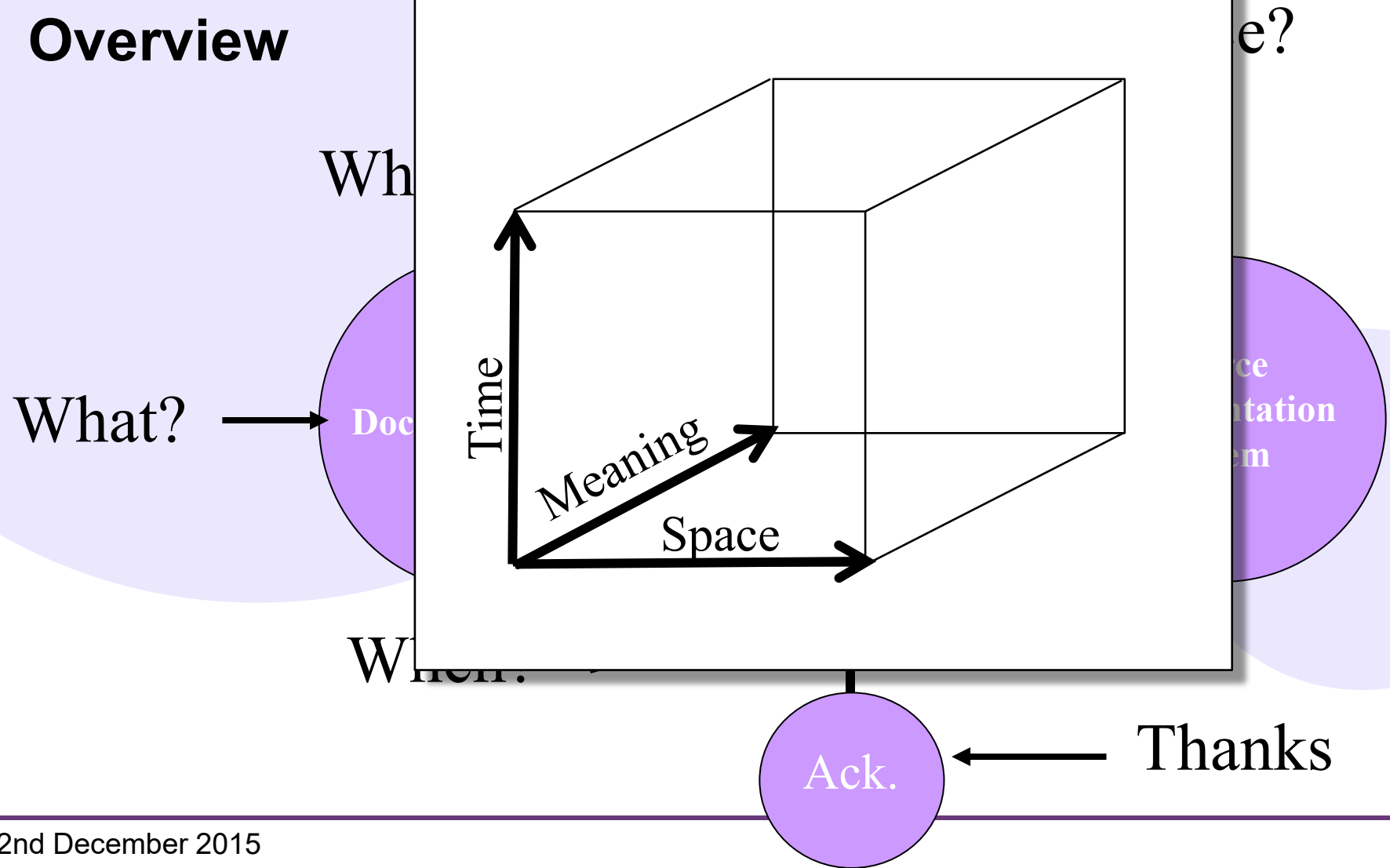
GBH GIS Statistical Content 14 million data values:

- Data from every Census of Population 1801-2001
 - Parish-level population counts
 - Age by sex by district
 - Occupations, industries, housing conditions, etc ...
 - Working now on adding 2011 data
- Vital statistics
 - Decennial data: cause of death by age by sex by district
- Votes for each candidate in each constituency in each election 1832-2005
- Farm census: crop acreages and numbers of animals, by county, at least decennial since 1869
- Unemployment and poor law data, by town/district
 - ... a diverse but very *ordinary* set of social statistics

GBH GIS Data Architecture Overview



GBH GIS Data Architecture Overview



Detailed structure of g_data table:

- Includes all columns except some very implementation-specific housekeeping columns
- Taken from current PostGIS implementation
 - \d g_data
- In previous Oracle version dates implemented as objects

Topic	Column Name	Data type
Row ID	g_seq	integer
Data value	g_data	numeric
Behaviour	g_data_precision	character varying(3)
	g_data_status	character varying(1)
	dominant	character varying(1)
Where?	g_unit	integer
	g_unit_type	character varying(16)
	g_rel_to	integer
	g_type_period	character varying(1)
What?	cellref	character varying(48)
When?	start_year	integer
	start_date	date
	end_year	integer
	end_date	date
	end_date_decimal	numeric
Source?	g_authority	character varying(20)
	g_auth_note	character varying(400)
	src_tab_panel	integer
	src_tab_seq	integer
	src_tab_col	integer
	src_tab_row	integer
Thanks	g_data_credit	character varying(24)

Documentation for individual data value

Behind the Vision of Britain web site is a true **data library**, a collection of individual statistics each with its own **metadata** to tell users what it means. These metadata also control how the system visualises statistical data through graphs and maps. This is what we know about the particular data value you selected:

Attribute	Value	Meaning												
Identifier	3526823	This number locates the particular value within the data table; it is like an accession number in a conventional library.												
Value	182	This is the data value itself.												
Date(s) covered	January 1st, 1901 to December 31st, 1910	We store dates within a <i>date object</i> , which can hold anything from a simple year value, for a census, to a period defined by two calendar dates.												
Administrative area covered	10154984	Portsmouth RegD/PLPar/PLU												
Meaning (DDS cell reference)	CoD_DS_1900s:male/25_34/pulmon	Cell within nCube N_CoD_DS_1900s: Decennial Cause of Death by Age <table><tr><th>Dimension</th><th>Variable</th><th>Category</th></tr><tr><td>1</td><td>Sex</td><td>Male</td></tr><tr><td>2</td><td>Age in mainly 10-year bands to age 75</td><td>25 to 34</td></tr><tr><td>3</td><td>Cause of Death (RG 1900s DS definition)</td><td>Pulmonary tuberculosis</td></tr></table> Universe: All deaths	Dimension	Variable	Category	1	Sex	Male	2	Age in mainly 10-year bands to age 75	25 to 34	3	Cause of Death (RG 1900s DS definition)	Pulmonary tuberculosis
Dimension	Variable	Category												
1	Sex	Male												
2	Age in mainly 10-year bands to age 75	25 to 34												
3	Cause of Death (RG 1900s DS definition)	Pulmonary tuberculosis												
Source	RG_DS	Registrar General, <i>Decennial Supplement</i> (HM Stationery Office) ,												
Status	P	Public. Visible in both graphics and tables.												
Precision	E	Exact												
Created by	cause_reg	This is the name of our script which inserted this data value into the main data table. For non-derived values, this is usually done by copying data values across from our more conventional GBHDB database, and the scripts are usually named after the table in GBHDB they copy data from.												

This table lists, in its "value" column, a single row from our data table. The values are mostly identifiers and codes which are given meaning by our main sub-systems, and the "meaning" column reaches into these to tell you more.

What kind of Data Feed? <http://www.w3.org/TR/vocab-data-cube/>

Sample Data Cube RDF

- NB links to external ontologies esp. admin units

```
eg:dataset-le1 a qb:DataSet;  
  rdfs:label "Life expectancy"@en;  
  rdfs:comment "Life expectancy within Welsh  
    Unitary authorities - extracted from  
    Stats Wales"@en;  
  qb:structure eg:dsd-le ;  
  .
```

```
eg:o1 a qb:Observation;  
  qb:dataSet eg:dataset-le1 ;  
  eg:refArea ex-geo:newport_00pr ;  
  eg:refPeriod <http://reference.data.gov.uk/  
    id/gregorian-interval/  
    2004-01-01T00:00:00/P3Y> ;  
  sdmx-dimension:sex sdmx-code:sex-M ;  
  sdmx-attribute:unitMeasure  
    <http://dbpedia.org/resource/Year> ;  
  eg:lifeExpectancy 76.7 ;  
  .
```

```
eg:o2 a qb:Observation;  
  qb:dataSet eg:dataset-le1 ;  
  eg:refArea ex-geo:cardiff_00pt ;  
  eg:refPeriod <http://reference.data.gov.uk/  
    id/gregorian-interval/  
    2004-01-01T00:00:00/P3Y> ;  
  sdmx-dimension:sex sdmx-code:sex-M ;  
  sdmx-attribute:unitMeasure
```

Phase 1: Dumping all the data as RDF

- Completed autumn 2014
- Defining basic format
 - Broadly based on W3C Data Cube Vocabulary
 - Additional details follow Irish census gateway <http://data.cso.ie/index.html>
- Main issue is choosing vocabularies to link to:
 - RDFS Schema (W3C)
 - SDMX (Statistical Data Metadata eXchange)
 - GB Historical GIS (we need to implement this!)
- NB this stage is of limited usefulness by itself, as result is a single file of > 100 Gb

Sample GBHGIS RDF

```
@prefix sdmx-subject: <http://purl.org/linked-data/sdmx/2009/subject#> .
@prefix qb:          <http://purl.org/linked-data/cube#> .
@prefix sdmx-measure: <http://purl.org/linked-data/sdmx/2009/measure#> .
@prefix gbhgis-ontogly: <http://data.pastplace.org/ontogly#> .
@prefix rdf:          <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix marcrel:      <http://id.loc.gov/vocabulary/relators/> .
@prefix rdfs:         <http://www.w3.org/2000/01/rdf-schema#> .
@prefix gbhgis:       <http://data.pastplace.org/> .
@prefix admingeo:     <http://data.ordnancesurvey.co.uk/ontology/admingeo/> .
@prefix sdmx-dimension: <http://purl.org/linked-data/sdmx/2009/dimension#> .
@prefix foaf:         <http://xmlns.com/foaf/0.1/> .
@prefix dc:           <http://purl.org/dc/elements/1.1/> .
```

<Observation-1110279>

```
    a                qb:Observation ;
    gbhgis-ontogly:PastP_Unit      "http://data.pastplace.org/unit/10076810" ;
    gbhgis-ontogly:hasUnitType     "MOD_DIST" ;
    gbhgis-ontogly:ref-period      "1951" ;
    qb:dataset                     "WORKING_AGE:male/total" ;
    sdmx-measure:obsValue          "37949"^^<http://www.w3.org/2001/XMLSchema#decimal> .
```

<Observation-990403> a

```
    qb:Observation ;
    gbhgis-ontogly:PastP_Unit      "http://data.pastplace.org/unit/10076810" ;
    gbhgis-ontogly:hasUnitType     "MOD_DIST" ;
    gbhgis-ontogly:ref-period      "1881" ;
    qb:dataset                     "INDUSTRY_GEN:ag" ;
    sdmx-measure:obsValue          "1832"^^<http://www.w3.org/2001/XMLSchema#decimal> .
```

<Observation-853267> a

```
    qb:Observation ;
    gbhgis-ontogly:PastP_Unit      "http://data.pastplace.org/unit/10188982" ;
    gbhgis-ontogly:hasUnitType     "MOD_DIST" ;
    gbhgis-ontogly:ref-period      "1881" ;
    qb:dataset                     "AGESEX_85UP:male/75_79" ;
    sdmx-measure:obsValue          "161"^^<http://www.w3.org/2001/XMLSchema#decimal> .
```

<http://data.pastplace.org/datacube?units=10188982,10076810>

Phase 2: Adding a sub-setting mechanism

- Currently implemented:
 - Geography:
 - One or more specific unit IDs (reporting areas)
 - One or more unit types (broadly, GIS coverages)
 - nCubes references
 - All components of a classification or cross-classification
 - Cell references
 - Specific measures, e.g. TOT_POP:now or AGESEX_85UP:male/70_74, Or 'Sheep'
 - Date range

To be added:

- Rates
 - I.E. extracting pairs of data values which we have associated in a rate definition, e.g. TOT_POP:now / AREA_HECTARES:total

Phase 3: Adding a discovery mechanism

- So you can find out relevant nCubes, cells, unit types and units!
- Broader queries based on date ranges, statistical keywords and “geography”
- Geography can be queried by:
 - Bounding box: specific units which overlap
 - Point: specific units containing
 - Name: units named after place
- Geography problems:
 - We have nearly 200 “kinds” of unit in system!
 - 22% of our units with statistics have point but no polygon
 - Units named after place don’t always contain it!

Phase 3: Discovery mechanism

- Spatial strategy

- The API takes a point coordinate and a broad statistical theme, the system returning the set of reporting units whose boundary polygons cover the point, with thematically relevant nCubes.

- Themes are: Population, Industry, Social Structure, Learning & Language, Agriculture & Land Use, Life & Death, Work & Poverty, Housing, Roots & Religion, Political Life

- http://data.pastplace.org/search?q=50.8193761,-0.3765256&theme=T_POP&format=json

```
{"gUnit":10056524, "uName":"WORTHING", "uType":"District",  
"nCubeId":"N_AGESEX_85UP", "nCubeTitle":"Redistricted Age  
& Sex Structure", "startYear":"2001", "endYear":"2001"...
```

- Data values selected via themes or nCubes and units:

- http://data.pastplace.org/datacube?theme=T_POP&units=10056524
 - http://data.pastplace.org/datacube?ncubes=N_AGESEX_85UP&units=10056524

Phase 3: Discovery mechanism

- Thematic strategy
 - allows the user to reach nCubes through free-text searching of labels and descriptions within the metadata.
 - <http://data.pastplce.org/search?metadata=SHEEP&format=json>

```
{
  "nCUBE": "N_AGCEN_SHEEP_GEN",
  "nLabel": "Generalised Sheep Categories as reported by the Agricultural Census.",
  "score": 36
} ...
{
  "nCUBE": "N_OCC_RAW1881",
  "nLabel": "Detailed 1881 occupational statistics, by gender",
  "score": 6
},
{
  "nCUBE": "N_OCC_RAW1841",
  "nLabel": "Unclassified occupations, categorised by sex & age",
  "score": 6
},
{
  "nCUBE": "N_AGCEN_ANIMALS",
  "nLabel": "Types of Animal reported by the Agricultural Census",
  "score": 6
}
```

Geospatial Ontology of Places

- Search API for places matching “Portsmouth” reveals places which are associated with units
 - <http://data.pastplace.org/search?q=portsmouth&format=n3>

```
@prefix place: <http://data.pastplace.org/place/> .
```

```
place:72259 a                gbhgis-ontogly:PastP_Place ;
    rdfs:comment              "Portsmouth.-- parl. and mun. bor., naval
station, seaport..."^^xsd:string ;
    rdfs:label                "PORTSMOUTH"^^xsd:string ;
    rdfs:seeAlso              http://www.wikidata.org/entity/Q72259 ...
gbhgis-ontogly:hasNearbyPlace place:776556 , place:985594, ..., ;
gbhgis-ontogly:hasUnit       unit:12833027 , unit:10214877 ,
unit:10187084 ...> ;
    geo:lat                   "50.7952777777778"^^xsd:double ;
    geo:lon                   "-1.08888888888889"^^xsd:double ;
    owl:sameAs              <http://dbpedia.org/resource/Portsmouth> ,
<http://sws.geonames.org//about.rdf> .
```

Update to latest Pelagios <https://github.com/pelagios/pelagios-cookbook/wiki/Pelagios-Gazetteer-Interconnection-Format>

Geospatial Ontology of reporting units

- Or you can search for units directly by name
 - <http://data.pastplace.org/search?q=portsmouth&format=n3&type=unit>

```
@prefix unit: <http://data.pastplace.org/unit/> .
```

```
unit:10200246 a                gbhgis-ontogly:unit ;
    rdfs:label                "PORTSMOUTH"^^xsd:string ;
    gbhgis-ontogly:hasUnitStatus "AP/CP"^^xsd:string ;
    gbhgis-ontogly:isPartOf    unit:10543411 , unit:10187084 ,
```

```
unit: ... .
gbhgis-ontogly:hasName
    a                owl:ObjectProperty ;
    rdfs:comment      "A name for the queried place."^^xsd:string ;
    rdfs:domain        gbhgis-ontogly:PastP_Place ;
    rdfs:label        "has name"^^xsd:string ;
```

```
gbhgis-ontogly:hasUnitStatus
    a                owl:ObjectProperty ;
    rdfs:comment      "The status the assc. unit has in the Past
Place gazetteer."^^xsd:string ;
```

API examples

- Everything about the modern Portsmouth unit
 - <http://data.pastplace.org/datacube?units=10168181>
- Combine Search and Data:
 - Units named after a place(including narrowing to a specific unit type)http://data.pastplace.org/datacube?q=portsmouth&unittypes=ADM_CNTY
- Population of Portsmouth aged 70-74
 - http://data.pastplace.org/datacube?units=10168181&cellref=AGESEX_85UP:male/70_74
- Single data value 1881 census:
http://data.pastplace.org/datacube?units=10168181&cellref=AGESEX_85UP:male/70_74&yearfrom=1881&year=1881

Software

- <http://data.pastplace.org/datacube> & <http://data.pastplace.org/seach>
 - Java accessing PostgreSQL/PostGIS DB via Mybatis for creating Java Objects
 - Jena for RDF representation and serialization
- www.pastplace.org
 - Mobile first responsive Web App built on
 - Nginx
 - NodeJS,
 - Webpack
 - Babal for ES6
 - ReactJS
 - Not currently using LD formats from API, need to investigate Jena's 30% overhead compared to gson serialization.

What's Next?

- Improve the unpacking of the data into more detailed DataCube Vocabulary. We would like to use the HASSET thesaurus (ELSST) but we have currently a sticking point on specific license terms
 - <http://www.data-archive.ac.uk/find/hasset-thesaurus>
- Enhance the ontology,
 - Link countries to FAO Ontology <http://www.fao.org/>
 - The 200 Unit types and status value. Scope for linking unit types to wikidata entry.
- Experiment with the Linked Data API in the PastPlace Web App
- Experiment with the API to create Client JS visualizations for the nCube statistics

Web sites, etc

- Great Britain Historical GIS Project
Geography Department,
University of Portsmouth, UK
- GBHGIS Project website:
<http://www.port.ac.uk/research/gbhgis/>
- Vision of Britain website:
<http://www.visionofbritain.org.uk>
- PastPlace www.pastplace.org (not launched)
- Research funded by the:
 - Arts and Humanities Research Council: “Frameworks and tools for statistical big data in the humanities”
- Contact us: **gbhgis@port.ac.uk**



Revisiting the earlier API examples & any question? **gbhgis@port.ac.uk**

Discover themes data.pastplace.org/search?q=50.8193761,-0.3765256&theme=T_POP&format=json

Portsmouth data.pastplace.org/datacube?units=10168181

Search and access Data data.pastplace.org/datacube?q=portsmouth&unittypes=ADM_CNTY
Celref data.pastplace.org/datacube?units=10168181&cellref=AGESEX_85UP:male/70_74

Single data value
data.pastplace.org/datacube?units=10168181&cellref=AGESEX_85UP:male/70_74&yearfrom=1881&yearto=1881

Discovering ranked nCubes from metadata search:
data.pastplace.org/search?metadata=SHEEP&format=json

Discovering nCubes from location and thematic search:
data.pastplace.org/search?q=50.8193761,-0.3765256&theme=T_POP&format=json



The RDF Data Cube Vocabulary

W3C Recommendation 16 January 2014

This version:

<http://www.w3.org/TR/2014/REC-vocab-data-cube-20140116/>

Latest published version:

<http://www.w3.org/TR/vocab-data-cube/>

Implementation report:

http://www.w3.org/2011/gld/wiki/Data_Cube_Implementations

Previous version:

<http://www.w3.org/TR/2013/PR-vocab-data-cube-20131217/>

Editors:

[Richard Cyganiak](#), [DERI](#), [NUI Galway](#)

[Dave Reynolds](#), [Epimorphics Ltd](#)

Contributors:

[Jeni Tennison](#)

Please refer to the [errata](#), a list of issues with this document discovered after publication.

This document is also available in this non-normative format: [diff to previous version](#)

The English version of this specification is the only normative version. Non-normative [translations](#) may also be available.

Copyright © 2012-2014 [W3C](#)® ([MIT](#), [ERCIM](#), [Keio](#), [Beihang](#)), All Rights Reserved. [W3C liability](#), [trademark](#) and [document use](#) rules apply.

Abstract

There are many situations where it would be useful to be able to publish multi-dimensional data, such as statistics, on the web in such a way that it can be linked to related data sets and concepts. The Data Cube vocabulary provides a

Other work on statistical APIs:

- World Wide Web Consortium (W3C)'s RDF Data Cube Vocabulary
 - <http://www.w3.org/TR/vocab-data-cube/>
 - Provides framework rather than full specification
- Irish government census gateway
 - <http://data.cso.ie/index.html>
- Dutch CEDAR project
 - <http://www.cedar-project.nl/>
 - Live server at: <http://lod.cedar-project.nl/>

