



UCL

WORKING PAPERS SERIES

Paper 174 - Nov 11

**CensusProfiler – Creating
accessible Geovisualizations
of the Census of Population**

ISSN 1467-1298



UCL CASA Working Papers Series

<http://www.casa.ucl.ac.uk/publications/workingpapers.asp>

CensusProfiler – Creating accessible Geovisualizations of the Census of Population

Pablo Mateos¹, Oliver O'Brien ²

¹ Department of Geography, University College London

² Centre for Advanced Spatial Analysis (CASA), University College London

Tel: +44 (0) 7679 0500, Fax +44 (0)20 7679 0565

p.mateos@ucl.ac.uk; o.obrien@ucl.ac.uk

<http://pablomateos.com/>; <http://oliverobrien.co.uk/>;

KEYWORDS: Census, Population Geography, Geovisualization, Thematic mapping, Mashups

ABSTRACT

This paper describes the outcome of the Census Geographic Visualization (Censusgiv) research project; a web prototype that offers significant geographic visualizations of the most significant Census outputs termed “CensusProfiler” (www.censusprofiler.org). This website is accessible by any member of the general public and is focused on exploring Census data geographically with very easy navigability, based on open-source software for mapping mashups. There is a gap in the dissemination of Census outputs through meaningful geographic visualizations, within and outside the research community. Local government bodies, health authorities, NGOs and the general public are demanding this type of easily accessible geographic visualizations of the Census that are not sufficiently addressed by current 2001 Census websites. This research project renders visualization methods, Census data and intuitive mapping easily accessible to the social science research community and effects knowledge transfer to local government and public services, as well as to a lay audience, through an interactive web-based geographic visualization.

Contents

1. INTRODUCTION	4
1.1 Context and Justification	4
1.2 Aims and Objectives	6
2. EVALUATION OF ALTERNATIVE TECHNOLOGIES	7
2.1 Selection criteria	7
2.2 Evaluation of pre-existing technologies	8
2.3 Recommended solution	10
2.4 Functional design considerations	11
2.5 Technology Summary	11
3. SOURCE DATA	13
3.1 Census Data	13
3.2 Spatial Data	17
3.3 Metadata	18
3.4 Non-Census Data	19
4. CREATING THE MAP	22
4.1 Map Tile Specification – URL Parameters	23
4.2 Map Tile Rendering – Mapnik	25
4.3 Context Map	26
5. DISPLAYING THE MAP	27
5.1 Visualization – OpenLayers	28
5.2 User Interface – Non-mapping elements	29
5.3 Choropleth Map Colour Schemes	31
6. PERFORMANCE TESTING	31
7. SCALING	32
8. CONCLUSION	32
9. REFERENCES	34
10. APPENDICES	36
10.1 URL Parameters	37
10.2 System Specifications	40

1. INTRODUCTION

The innovation of digital globes, such as Google Earth or Microsoft Virtual Earth, and their 2-D representations such as Google Maps, Bing Maps or OpenStreetMap, combined with user generated content and services, has brought an explosion of interest by the general public in geographic information (GI) and its visualization over the last three years, a phenomena termed as the democratisation of GI (Nature, 2006). The combination of GI services is also known as mapping mashups, a term recently introduced to refer 'to hybrid web applications that combine data or software from two or more sources' (Monmonier 2007). Although Geographers have arrived late to this renaissance of geographic visualization (GVis) there is increasing realisation of the potential for 'geography' to provide the primary basis for innovative visualization and knowledge exploration (Dodge et al. 2006). Despite the recognised growing usefulness of GVis to make sense of increasingly large datasets, the means of access to meaningful geographic visualizations of the UK 2001 Census outputs and other population administrative datasets have failed to keep up with advancements in the 2001 Census output dissemination efforts.

This paper presents the final results of a research project known as Census Geographic Visualization (**Censusgiv**), funded by the Economic and Social Research Council (ESRC) under the Census Development Programme (grant code RES-348-25-0015). This project developed a novel and intuitive geovisualization framework of the next UK Census in 2011 as well as other population geography datasets, with the use of innovative geographic visualization using mapping mashups. The resulting prototype website www.censusprofiler.org aims to demonstrate the potential of such applications in the geographic visualization of the Census and other population datasets.

1.1 Context and Justification

The introduction of Google Earth and Google Maps in 2005, followed by equivalent geographic platforms in the Microsoft and Yahoo web portals, has brought an explosion of interest in local geographic information and associated services. This step-change over the last five years has only partially been harnessed by the academic and policy communities.

Geographic Visualization or 'Geovisualization' refers to the visual representation of spatial data. As a research tool it 'can be applied to all the stages of problem-solving in geographical analysis, from the development of initial hypotheses, through knowledge discovery, analysis, presentation and evaluation' (Buckley et al., 2000: 2). There is increasing realisation of the potential for 'geography' to provide the primary basis for innovative visualization and knowledge exploration (Dodge, McDerby and Turner, 2006). Despite the recognised growing usefulness of geovisualization to make sense of increasingly large datasets, the means of access to meaningful geographic visualizations of the UK Census outputs and other population administrative datasets have failed to keep up with advancements in data visualization in the Web 2.0 era.

In the United Kingdom (UK), the release of the 2001 Census nearly ten years ago led to the development of a set of online tools that provide much easier access to an increasing range of digital data products freely available from the Office for National Statistics (ONS), the General Register Office for Scotland (GROS) and the Northern Ireland Statistics and Research Agency

(NISRA). However, the means to visualize and analyse the spatial dimension of these datasets is still only available to a few privileged researchers with extensive knowledge of Geographic Information Systems (GIS), principally through the services provided to the academic community by the ESRC Census Population Programme data dissemination units. Geographic visualization of these datasets is currently only made possible by accessing the services of some of these units individually, and requesting specific census outputs at a series of fixed geographical scales to be mapped later in a separate GIS package.

Another option to access geographical visualizations of the census information is through the ONS Neighbourhood Statistics (NeSS) website (www.neighbourhood.statistics.gov.uk), as part of a governmental effort to provide a steady flow of population and administrative statistics by small area to support neighbourhood based policies. However, access to GVis of ONS Neighbourhood Statistics data requires a cumbersome selection process to specify the desired dataset, location and geographical scale, which recently required the user go through twelve different selection criteria screens before the first map ever appears on the screen (this process has subsequently been simplified). Moreover, if the user wants to make slight changes to the selected area or geographical scale of data aggregation, or select another dataset to visualize, he or she needs to start the whole process again. Furthermore, the mapping facility offers little contextual information that allows the user to intuitively relate the thematic content of the population data with the geographical context within which the processes that those statistics reflect actually take place. Finally, the expected technical facilities to zoom and pan across the area are also missing from this application.

These poor geographical visualization facilities of population data available to non-GIS specialists are in stark contrast with recent developments in web mapping characterised by the services introduced by Google since 2005. The diffusion of Google Maps and Google Earth has dramatically increased the use of geographic information (GI) among internet users and fuelled new ways of deploying GI in an effective and easy way that is rapidly and intuitively understood by any sort of public. What characterises this technology is that it is mainly built with a 'bottom-up' perspective, in which multiple users provide content, rather than based on a centrally prescribed 'top-down' approach to disseminate geographic information. Mapping mashup is a term recently introduced to refer 'to hybrid web applications that combine data or software from two or more sources' (Monmonier, 2007: 373). The most widely used mapping mashup technology is based on Google Maps, a web service that has 'dramatically raised users' expectations [with] its fluid movements, intuitive user experience and competent cartography' (Fairhurst, 2005: 57). Although Google Maps is not a complete GIS tool, the availability of an Application Programming Interface (API) stimulates users with basic programming skills to build their own applications using Google Maps as a visualization interface 'mashed-up' with their own geographic data. Therefore, the user's data is given instant geographical context through detailed satellite and aerial imagery, placenames, administrative boundaries, road and street networks and point of interest data, seamlessly available throughout the world. Before Google Maps became available, webmapping developers had to either acquire very expensive proprietary data, or sort out cumbersome copyright issues with geographic data for every single country or area of concern. Moreover, change of location or geographical scale of visualization is delivered by Google Maps and Google Earth in a seamless and smoothed fashion that remains the most successful feature for the. Other similar applications to Google Maps and Google Earth

are Microsoft Virtual Earth, Bing and ESRI ArcGIS Explorer. In the Open Source community the most successful platform with highly detailed cartography is OpenStreetMap (OSM) (www.openstreetmap.org), a service that will be fully described later in this paper.

Following this trend in the “democratisation” of geographic visualization, a Stanford graduate student launched a service called *gCensus* (gCensus 2008) that offers a flexible and intuitive way of visualising U.S. Census data. It uses a selection mechanism to extract selected census variables into KML files (Keyhole Markup Language) that can be then viewed in Google Earth. This has opened up intuitive explorations of census data for the general public, restoring the hidden potential of the census previously locked in tedious government reports and tables.

Therefore, in the late 2000s we argued that the current geographic visualization tools of the UK Census information available through the NeSS and CASWEB webmapping facilities present a series of technical as well as conceptual limitations. Since these websites were designed to map specific data tabulations extracted from these websites, as a subsidiary process to the main data distribution service, they present a rigid and cumbersome geographic visualization tool developed using the technology available in the late 1990s.

1.2 Aims and Objectives

Following from the described state of affairs (originally investigated in the summer of 2007), a team of geographers at University College London (UCL) decided to put together a funding proposal for a project that would help in remedying this situation. We contended that the future success of plans for wider dissemination of the 2011 Census outputs cannot rely on the aforementioned 1990s technologies. Instead, geographic visualizations of the Census should be built around a fresh approach that starts with a map (as opposed to a stream of linear queries), showing real geography to which people easily relate to, on top which they can visualize and intuitively interact with the different Census population dynamics, while experiencing seamless changes in geographical scale and location. The aim should be to provide a new web-based service that allowed any person to carry out hypothesis generation, knowledge discovery, analysis and evaluation through geographic visualization (Buckley, 2000). This can be effectively and inexpensively achieved through a mapping mashup approach as demonstrated by the current project that developed a software prototype in the run up to the 2011 Census.

As such, the Census Geographic Visualization or *Censusgiv* research project was developed around the following key research questions:

1. What should be the main features of a service to provide intuitive geographical visualizations of the Census data aimed at both specialist and general publics?
2. Could an approach to provide a geographical visualization service of Census data based on just navigating a map be popular and meaningful?
3. How would the functionality of such a service be designed?
4. What are the technological options to provide such a service using mapping mashups? What would be the final recommended solution?
5. What resources and timescales would be required for the complete roll-out of the proposed geographic visualization service as the 2011 Census data became available?

However answering these research questions presented a series of technical, logistical and legal challenges. These are all derived from the complex characteristics of the Census data, which comprises a large and complex statistical effort, summarised in thousands of tables at different geographical as well as thematic scales (Martin 2000). Furthermore, the provenance and copyright of the Census data itself as well as the underlying geographical units, impose certain restrictions in the method of dissemination and the technological solution to be finally adopted.

Our approach to answering these questions required the development of a web-based prototype, hereinafter called “**CensusProfiler**”, whose characteristics this paper aims to fully describe. The final result can be explored at www.censusprofiler.org.

2. EVALUATION OF ALTERNATIVE TECHNOLOGIES

A number of overall alternative technologies for the CensusProfiler website were evaluated. The first phase of the *Censusgiv* project consisted in reviewing each one of these key available geovisualization technologies for population datasets. In the period 2008-2010 the most appropriate technologies available consisted in graphic design solutions based on Adobe Flex or Flash-based applications and SVG-based vector mapping, a range of options based on KML vector mapping with Google Maps or Google Earth APIs, and the Open Geospatial Consortium standards known as WMS/WFS services. Such review was conducted through a formal and thorough evaluation that quantified their features, advantages and disadvantages, considering their potential adoption as a platform of choice to develop the proposed *Censusgiv* service. We consider that this evaluation, as a by-product of this project, will also be a useful for any future web cartographer in order to obtain a comprehensive review of the geovisualization methodologies available at the time.

2.1 Selection criteria

The main selection criteria upon which each of the technological solutions considered here has been evaluated are as follows:

1. **Scalability**; how easy is it to expand the platform in terms of additional datasets, geographical areas or number of users.
2. **Response time**; average waiting time between a user's request and the desired action being executed.
3. **Maximum number of concurrent users**; limitations in terms of simultaneous users connected and using the service at any one time.
4. **Data storage** and retrieval.
5. **Flexibility** of geovisualization; in terms of different options for the cartographic representation of statistical data.
6. **Ease of use** and simplicity in user interaction with the web interface.
7. **Intellectual Property Rights** (IPR) issues; how well are copyright and data protection limitations are handled.
8. **Cost** of development and implementation; in terms of software, hardware and development costs.

Each of these criteria was evaluated for each of the technologies available, before a final solution was recommended for final development. Some of the available options present a series of advantages over some of these criteria but disadvantages over some others.

2.2 Evaluation of pre-existing technologies

The following seven technological options were formally evaluated against the eight criteria listed above, with a final score reported here for each of them and the full evaluation scores shown in Table 1 and Figure 1.

Web Map Service (WMS)

Web Map Service (WMS) is one of the classic technologies for disseminating GI over web GIS solutions (Open Geospatial Consortium 2008). It uses client / server protocols and OpenGIS consortium standards to produce and serve individual cartographic layers that can be read by any client adhering to the WMS standard as a final thematic map in raster form (Tu et al. 2004). This technology has a series of limitations regarding response time and number of concurrent users, and because of the complexity of its development and configuration aspects. Its final score is only 16/100.

Web Feature Service (WFS)

Web Feature Service (WFS) is very similar in essence to WMS but based on serving GI in a vector format rather than raster maps (Open Geospatial Consortium 2008a). This technology is complex and costly to implement, and with a poor performance in terms of number of concurrent users and response time. Its final score is only 18/100.

GMapCreator

GMapCreator is a software tool developed at UCL's Centre for Advanced Spatial Analysis (CASA) that creates mapping mashups through the rasterisation of thematic maps in GIS format so that they can be overlaid on top of map tile providers such as Google Maps (Gibin et al. 2008). This offers a series of advantages for the *Censusgiv* project, amongst them, a safer position on Intellectual Property Rights (IPR) issues, being based on raster maps, and the ability to serve large numbers of concurrent users without large server requirements. Its main disadvantage is the relatively complex procedure to rasterise, store and serve the large number of datasets required by the Census. Its final score is 56/100.

KML via Google Earth / Google Maps

KML (Keyhole Mark-up Language) is a new XML-described, vector-based standard in geospatial visualization science (Goodchild 2005). By combining KML and Google Earth or Google Maps, a flexible, cheap and scalable solution can be provided, offering high quality web GIS capabilities for visualizing complex datasets such as the Census of Population. Amongst its advantages are its flexibility, storage/serving capabilities and ability to serve a large number of concurrent users. Amongst its disadvantages are the difficulty to serve complex vector polygons and the fact that all data are transferred to the user's browser/ Google Earth client, as plain text inside the KML file, which might infringe the copyright of Census data in this context. Its final score is 58/100.

'Traditional' Web GIS

'Traditional' Web GIS technologies are based on using proprietary GIS server software, such as ESRI ArcGIS Server. They have the potential to offer a broad variety of services for almost all spatially related applications of modern web environments (Alesheikh et al. 2002), but there is a significant lack of flexibility, scalability and number of concurrent users, as well as having a very high cost, although recent competition is starting to reduce costs. Its final score is 42/100.

Flash & Google Maps

Flash technology (Adobe 2008) can be used in conjunction with map image "tile" providers such as Google Maps. It presents significant disadvantages in terms of scalability, ease of navigation through geographical scales and cost of software and development. On the other hand, amongst its advantages, it provides an easy graphical user interface and can serve big number of concurrent users across a wide variety of platforms. Its final score is 45/100.

SVG & Google Maps

Scalable Vector Graphics is a vector-based technology that can be applied in this case on top of Google maps. This has a series of advantages in terms of response time and flexibility. Amongst its main disadvantages are scalability problems, since complex geographical datasets could reach large file sizes and a historic lack of documentation (as no longer supported by Adobe) especially in implementations with Google Maps. However it is undergoing something of a renaissance as computer power and web browser capability increases. Its final score is 45/100.

These seven technological options were formally evaluated by building a simple prototype for each of them and placing a request to serve an identical dataset, using the same server, and for a fixed number of 1000 users (1000 local requests). A final score was given to each technological option for each of the eight evaluation criteria listed above, and are presented in Table 1. These results could vary according to the server technology and capacity available, but it offers a meaningful insight on the potential capabilities and limitations of these technologies once rolled out into a full service.

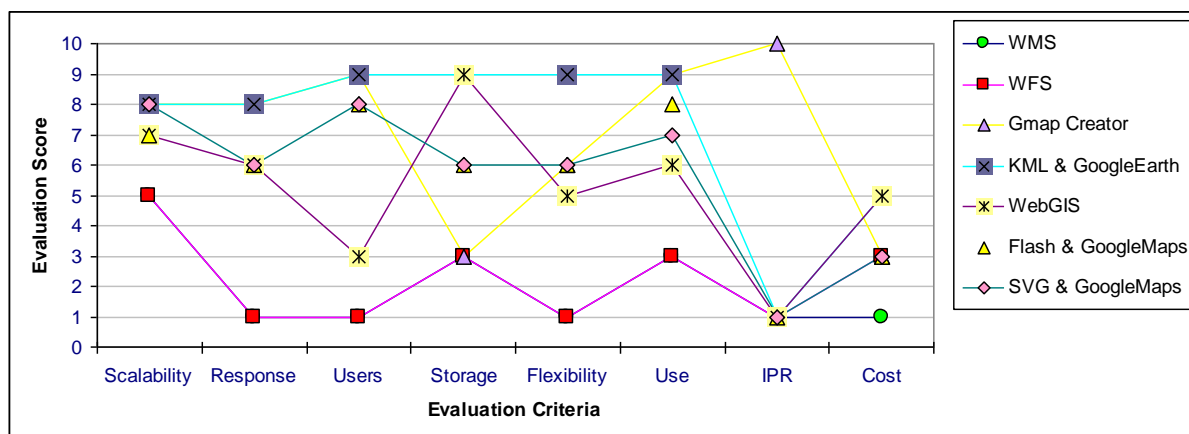


Figure 1: Comparison graph of the technologies evaluated for the Censugiv project.

		A	B	C	D	E	F	G
		WMS	WFS	Gmap-Creator	KML & Google Earth	WebGIS	Flash & Google Maps	SVG & Google Maps

1	Scalability	5	5	8	8	7	7	8
2	Response	1	1	8	8	6	6	6
3	Users	1	1	9	9	3	8	8
4	Storage	3	3	3	9	9	6	6
5	Flexibility	1	1	6	9	5	6	6
6	Use	3	3	9	9	6	8	7
7	IPR	1	1	10	1	1	1	1
8	Cost	1	3	3	5	5	3	3
total		16	18	56	58	42	45	45

Table 1: Comparison matrix of geovisualization technologies evaluated for the Censusgiv project. Seven technologies(A-D) were rated according to eight evaluation criteria (1-8) described in the text using a score from 1 (strong disadvantages) to 10 (strong advantages).

2.3 Recommended solution

The majority of the seven geovisualization options evaluated here, with the only exception of GMapCreator and WMS, are based on serving geographical boundaries in vector format and data attributes as “readable text”. This presents major issues for the *Censusgiv* project with regards to the IPR restrictions imposed upon the Census of population geographical boundaries at the time this project was evaluated (copyright partly owned by Ordnance Survey) and possibly datasets (Census tables). Therefore, geovisualizations based on the rasterisation of census data would be a safer solution on this sole criterion. However, raster solutions have their own disadvantages in terms of limited flexibility in updating new datasets, and more stringent requirements on the server’s storage and serving capabilities.

Looking at each of the eight evaluation criteria, it was decided that given the short timeframe to develop a solution (just over a year) as well as the cost restrictions and IPR limitations, these two last criteria should perhaps have more weight against the other six factors. Therefore, it was decided to base a solution around the Google Creator type of technology, rasterising maps into pre-rendered tiles. This was the solution with the second highest score but much better positioned in terms of IPR as well as other factors.

However, the GMapCreator technology, developed in 2005, had some disadvantages associated with its high dependence upon the Google API. This in turn placed a number of restrictions on the flexibility to alter the design of the geographic visualizations produced using this tool, as well as the rather “manual” procedure to rasterize each Census variable at a time.

Therefore, it was finally decided that a simple approach to improve existing knowledge within CASA would be most effective. The final solution was designed around a website showing raster-based choropleth (thematic) maps of census tables, using a set of open source technologies to replace Google’s mapping framework, primarily through the OpenStreetMap

community of developers. The details about the final website design are the focus of the rest of paper, and therefore duplication is avoided at this stage.

2.4 Functional design considerations

As stated in the project aims and objectives, it was important that the user was not initially faced with sets of queries or complex questions long before accessing a visualization, and hence a default population map would be shown on first loading the website. Furthermore, technical aspects such as the level of geographic units, the number and range of data intervals (groupings of data values), the colour scheme used, and the location within the UK would be provided by default rather than required to be initially selected (as is still common with most mapping websites).

As stated in the evaluation section, there was a concern over delivering geographic boundaries and census datasets directly to the user's computer, because of licensing restrictions¹. Therefore, it was desirable to restrict the data to the server and only serve rasterised imagery of the data to the user's computer (technically known as "the client").

The sheer number of census variables (over a thousand, when most common combinations of variables are considered), along with a large number of intervals, data classification schemes, colour combinations, and geographical level of disaggregation, meant that a dynamic solution was needed. This entailed giving the user some saying in the characteristics desired for the map to be produced "on the fly", as opposed to pre-creating all possible combinations of these variables in separate maps in advance. However the increased server computation required for this level of flexibility, needed to be balanced with the capacity of the website to handle a large number of concurrent users. This trade-off between flexibility and volume of users has been at the core of the project since its inception. The final website adopted the best possible solution with the resources available, but we are nonetheless certain that in the hands of the Office for National Statistics or indeed any other large institution the technology will perform much faster than in a single server running from a University.

2.5 Technology Summary

CensusProfiler presents a "slippy" map as the predominant element of the website. This characteristic is the "Google-style" dragging of the whole map as a single sheet, as opposed to the pre-2005 standard of clicking and waiting in around a map's edge to visualize the "next page". All user interface components appear as moveable panels, in a "toolbar" style overlaid on top of the map, that the user can place wherever is more convenient. The "context" layer is displayed using a black and white map, overlaid on top of the other map layers, showing some generic geographical features such as rivers and city names, in order to provide spatial context to the rather artificial Census geographical units. The actual Census data is finally displayed in a colour thematic map, layered underneath the context map. In this paper we will refer to the Census map as the choropleth map.

¹ The licensing restrictions have since been partially alleviated with the Ordnance Survey Open Data and UK Data Store initiatives during early 2010.

The choropleth map is made up of tiled square images, or *tiles*, which are created “on-the-fly” on a server, following a request made by the user’s browser – the “client”. The tiles are produced on the server by a service module that we have named *Chorogen* (for *CHOR*Opleth map *GEN*erator). The Chorogen service in turn is fed by a database containing spatial tables of Census data with which to produce the rasterised maps. The context black and white map also consists of tiles, although these are pre-created in order to save time.

We have decided to build the CensusProfiler prototype using the univariate (single-variable) tables from the 2001 UK Census. The main advantage of these tables is that they are readily available to the public and generally reflect the most popular Census variables.

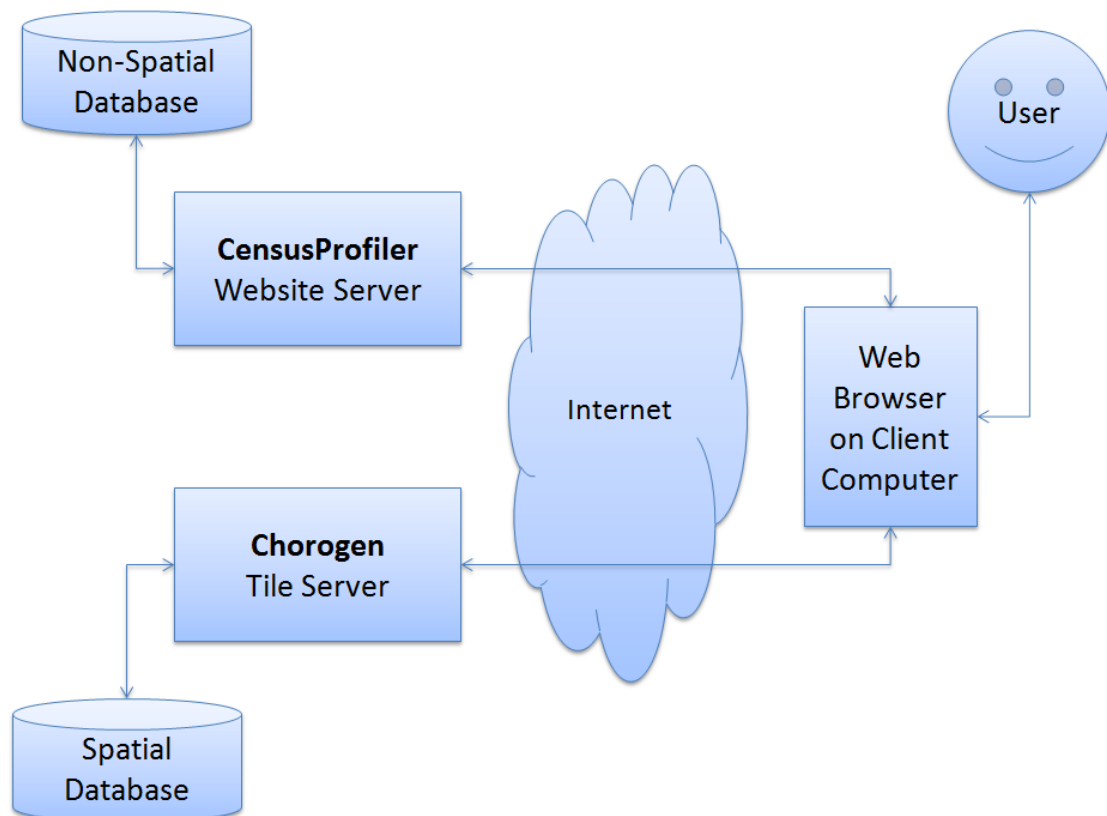


Figure 2: Simplified architecture diagram for CensusProfiler

3. SOURCE DATA

3.1 Census Data

The data tables visualized in this application are primarily from the UK 2001 census, specifically the univariate (UV) tables first published by the Office of National Statistics (ONS) covering England and Wales, and the General Register Office for Scotland (GROS), in 2003/4. The Northern Ireland census data sets are not included in this project but provision has been made for their future inclusion. Each Census univariate table contains a single variable recorded in the census, as opposed to combining multiple variables. Therefore, UV tables are made available at a high level of geographical and nominal granularity, for example ages of the population are available in one-year increments by Output Area geography.

The data was obtained on CD from the corresponding authorities and combined into a large database. The input data came as separate tables, each containing groups of univariate variables. Each input table had the following general table structure of fields or columns:

- Output Area: The code of the smallest geographical area for which data is made available in the Census
- Total: The total population of the OA or the universe population to which the variable or set of variables relate to (nr. of households, etc)
- Variable Value: The population count falling under the description for the variable in question
- Additional variables: Depending on the topic other univariate variables were listed under the same table. Each input table included at least one² and sometimes up to a hundred additional variables

Construction of the Census Data Tables

The data for all the 900 univariate Census variables at OA level was merged into a single large table, by concatenating the source data described in the previous section³. The final OA table structure adopted in the database is shown in Figure 3.

² The first of each of the UV dataset fields contains the overall population of that type (e.g. people, families), and ends in 01, followed by the actual variables themselves (there may only be one variable), suffixed with 02, 03 etc. For example, UV07 is the UV dataset for marital status. It contains a field UV0701 which is the population (in this case, people), followed by fields sequentially numbered UV0702 (single), UV0703 (married) and so on. The UV07 fields appear in the aggregated tables after the UV06 fields and before the UV08 fields.

³ During the initial stages of the project separate small data tables were used resembling the structure of the source tables. However, it was found that the performance of the map tile generator (Chorogen) was not significantly affected by the use of a single large data table for all the UV census data, while the selection and data management operation was greatly simplified.

UV004		UV0401		UV0402			
OA	UV005		UV0501		UV0501		
OA1	OA	UV006		UV0601		UV0602	
OA2	OA1	OA1	30	16			
OA3	OA2	OA2	20	3			
	OA3	OA3	25	10			
		OA4	28	12			

UV (OA)	UV0401	UV0402	UV0501	UV0502	UV0601	UV0602
OA1	30	28	30	3	30	16
OA2	20	27	20	0	20	3
OA3	25	24	25	5	25	10
OA4	28	18	28	0	28	12

Figure 3: Concatenating the source data tables into a single large table. Here, the UV04, UV05 and UV06 tables are shown being concatenated.

The Census data at Output Area (OA) geographical level, were then aggregated to two higher level geographies; the Lower-Level Super Output Area (LSOA) and Middle-Level Super Output Area (MSOA) levels. Such aggregation was performed using the lookup tables between these geographies, as provided by ONS and GROS through the National Postcode Directory (NSPD)⁴. Examples of the nested hierarchy of the OA/LSOA/MSOA geographies and their aggregations are shown in Figure 4. We acknowledge that aggregating OAs might not produce the same exact results than the figures provided by ONS and GROS directly at LSOA and MSOA level. However, in order to ensure the full correspondence between the three levels of geography used in the prototype as well as to simplify the analysis, we took this decision assuming that the rounding of small numbers introduced by ONS at OA level would be done in a way that the aggregated figures matched at higher geographical levels.

⁴ In Scotland, the Data Zones (DZ) and Intermediate Geography (IG) aggregated areas were used as these are the equivalents to LSOA and MSOA in England and Wales. It should be noted that the guideline minimum and averages population used by the GROS to create them differ slightly to England and Wales.

UV (OA)	UV0401	UV0402	UV0501	UV0502	UV0601	UV0602
OA1	30	28	30	3	30	16
OA2	20	27	20	0	20	3
OA3	25	24	25	5	25	10
OA4	28	18	28	0	28	12

↓

UV (LSOA)	UV0401	UV0402	UV0501	UV0502	UV0601	UV0602
LSOA1	50	55	50	3	50	19
LSOA2	53	42	53	5	53	22

↓

UV (MSOA)	UV0401	UV0402	UV0501	UV0502	UV0601	UV0602
MSOA1	103	95	103	8	103	41

Figure 5: Aggregation of the OA table into LSOA and MSA larger spatial units

This process resulted in three tables, one for each spatial unit, with each table containing all the census variables under the same exact table structure.

	area_code character variable	uv0010001 integer	uv0010002 integer	uv0010003 integer	uv0010004 integer	uv0030001 integer	uv0030002 integer	uv0030003 integer	uv0040001 integer	uv0040002 integer
1	00AAF00002	222	222	0	0	222	100	122	222	0
2	00AAF00003	149	149	0	0	149	75	74	144	3
3	00AAF00004	438	438	0	0	438	225	213	425	8
4	00AAF00001	264	264	0	0	264	150	114	273	3
5	00AAF00001	131	131	0	0	131	61	70	134	0
6	00AAF00013	251	251	0	0	251	111	140	263	5
7	00AAF00006	126	126	0	0	126	72	54	128	3
8	00AAF00014	239	197	42	0	239	126	113	246	0
9	00AAF00002	136	136	0	0	136	64	72	150	3
10	00AAF00005	164	164	0	0	164	60	104	164	0

	area_code character variable	uv0010001 integer	uv0010002 integer	uv0010003 integer	uv0010004 integer	uv0030001 integer	uv0030002 integer	uv0030003 integer	uv0040001 integer	uv0040002 integer
1	E01000001	1600	1530	70	0	1600	860	740	1597	12
2	E01000002	1470	1470	0	0	1470	783	687	1460	12
3	E01000003	1537	1393	144	0	1537	716	821	1576	14
4	E01000004	1505	1395	110	23	1505	923	582	1512	3
5	E01000005	1073	1073	0	0	1073	550	523	1064	14
6	E01000006	1571	1571	0	0	1571	784	787	1587	25

	area_code character variable	uv0010001 integer	uv0010002 integer	uv0010003 integer	uv0010004 integer	uv0030001 integer	uv0030002 integer	uv0030003 integer	uv0040001 integer	uv0040002 integer
1	E02000001	7185	6861	324	23	7185	3832	3353	7209	55

Figure 6: Examples of the database table structure for the three geographical levels

The three examples show part of (from top to bottom) the OA, LSOA and MSA tables. Each of the tables has the same set of fields (columns), containing all of the UV fields for all of the UV tables ordered in number sequence (UV001xxxx, UV003xxxx etc). Each record (row) represents a geographical area. The field names ending in "01" represent the population fields, the others are the actual metrics associated with this population.

The only issue in terms of geographic coverage is that many of the UV tables are not available for the whole of Great Britain, since there were two sets of Census questionnaires; one for England & Wales, and another one for Scotland. In some cases, equivalent tables with often only slightly different wording of variable names were available, while in other cases, questions were region-specific and there was no equivalent in the other regions (e.g. Gaelic speakers). In both of these cases, we decided not to combine similar variables into a single field, since this would require much more detailed knowledge of the context and history of each variable, and this felt outside the scope of this geovisualization project. As a result, some UVs in the CensusProfiler website can only be visualized for parts of the country. This is a problem that the Census organisations should take into account for the 2011 Census dissemination, and such alignment between the different UK Censuses should be consistently produced “in-house” rather by the outside public.

3.2 Spatial Data

The three levels of geography to be visualized in the CensusProfiler website; OA, LSOA, and MSOA, comprised the base set of spatial units through which all Census variables were to be represented. The boundaries for the OA spatial units were obtained from the Economic and Social Research Council (ESRC) Census Dissemination Units, in particular *CASWEB* at the University of Manchester and *Edina* at the University of Edinburgh. These boundaries were supplied as GIS files under the ESRI shapefile format. At a later stage in the project we decided to replace them with those available under *Ordnance Survey's Boundary Line* product as part of its Open Data initiative (Ordnance Survey, 2010). The Intermediate Geography spatial data for Scotland was not available from any of these sources but obtained directly from the Scottish Neighbourhood Statistics website (Scottish Neighbourhood Statistics, 2010).

Earlier in the project we found out that these Census area boundaries were perhaps too detailed for the purposes of visualising Census data for the UK. Therefore, cartographic generalisation was required in order to reduce the size of shapefiles of OAs with the ultimate purpose of obtaining very significant increases in the speed of the raster map generation with minimal decrease in visualization quality. Cartographic generalisation was performed on the OA shapefile using the *Simplify Polygon* tool's *POINT_REMOVE* algorithm in *ESRI ArcGIS's Data Management* geoprocessing toolbox (ESRI, 2009) applied to a maximum 25 metres deviation. The size of the spatial database was reduced by around 80% after generalization, resulting in substantially reduced disk access and processing time, significantly improving the overall performance of the service.

In order to amalgamate OAs into the higher level geographies, LSOA and MSOA, another GIS algorithm was applied. The *Dissolve* function in *ESRI ArcGIS Data Management* geo-processing toolbox tool was used to merge together the generalised OA spatial areas and form the corresponding LSOA and MSOA shapefiles, again using the National Postcode Directory lookup table between these geographical levels.

The resulting set of shapefiles were converted to a database compatible format, in this case as PostgreSQL database tables. The conversion was performed using the utility *ogr2ogr*⁵, from shapefile to SQL (structured query language) compatible table.

The spatial database was built in PostgreSQL using the *PostGIS* spatial extensions. *Generalised Search Tree* (GiST) spatial indexes were added to the three spatial tables to ensure that the geographic data could be quickly accessed by external queries made to the database. The final characteristics of the database spatial and Census data tables are listed in Table 2.

Spatial Area Unit	Number of Areas (Records)	Disk space used to store the Data Table	Disk space used to store the Spatial Table
OA	218,038	858 MB	124 MB
LSOA (England/Wales) & DZ (Scotland)	40,883	161 MB	47 MB
MSOA (England/Wales) & IG (Scotland)	8,429	33 MB	21 MB

Table 2: Size characteristics of the tables in the database

The final database contains three spatial tables and three Census data tables, whose size and number of records are described in this table.

3.3 Metadata

In order to provide meaningful explanations of the meaning of each univariate table in the CensusProfiler, and its overall context, a set of metadata was gathered for each variable. Such metadata assists the CensusProfiler user in selecting the Census variable they wish to map and provides information such as the geographical coverage of the variable.

The metadata and common names (long plain-English descriptions) associated with each variable were combined in a similar fashion as the actual Census data, into a single table. Such table was stored in a separate, non-spatial MySQL database, for fast performance and simplicity

⁵ A utility which is part of the Geospatial Data Abstraction Library (GDAL) and which allows conversion between various standard geospatial file formats, as specified by the Open Geospatial Consortium (OGR).

of integration between PHP and MySQL, being two of the open source components used to build the CensusProfiler.

The metadata table contained the UV code (as shown in the Census data tables), the UV long name, the UV group of variables to which it belongs (and the country coverage (some UV tables being only available for England & Wales, or Scotland). For each UV, the average value and standard deviation across Great Britain were calculated and added to the metadata table as two additional fields.

A natural two-level hierarchy of variables was therefore automatically generated by adding to each variable the original variable grouping in which the UV tables were provided (for example Age is a grouping of all variables representing individual years of age). To simplify the user interface, a third, “super-level” was added to further group this hierarchy, describing the broad characteristic of the table, such as “Education”. This was specified manually and added as an attribute to the metadata table.

The metadata table is held on a separate, non-spatial MySQL database. This database is held on a separate server than the spatial PostgreSQL/PostGIS database, containing the three Census data tables and the three geometry tables. This means that if and when the map tile generator service (Chorogen) is unable to create the imagery on an on-demand basis, the website can still function with tiles that have been cached as files on the server.

3.4 Non-Census Data

Apart from the Census UV data sets described above, additional Non-Census data were added as separate tables. This was done in order to demonstrate the value of mapping mashups to combine datasets from different sources and differing levels of geography, as demonstrated in the LondonProfiler website developed at UCL (www.londonprofiler.org). The additional datasets were the Index of Multiple Deprivation 2007 for England (CLG, 2007) available at LSOA level as a rank measure, and the Office of National Statistics’ Output Area Classification (OAC, 2007) available at OA level as a nominal classification. These variables were added to the two databases, again keeping the data and metadata split.

Finally, two cartographic tables were added to the MySQL database containing the colour ramps that can be selected in the website in order to generate the thematic maps. One table contains the colour ramp descriptions, suitable data types and number of intervals available, while the other table specifies the actual red-green-blue (RGB) colour combination that makes up each colour in the RGB scale.

Figure 7 shows the sources of the tables in both databases, the corresponding server used by each database and the resulting output from each server. Figures 8-10 show examples of the data and metadata in the tables.

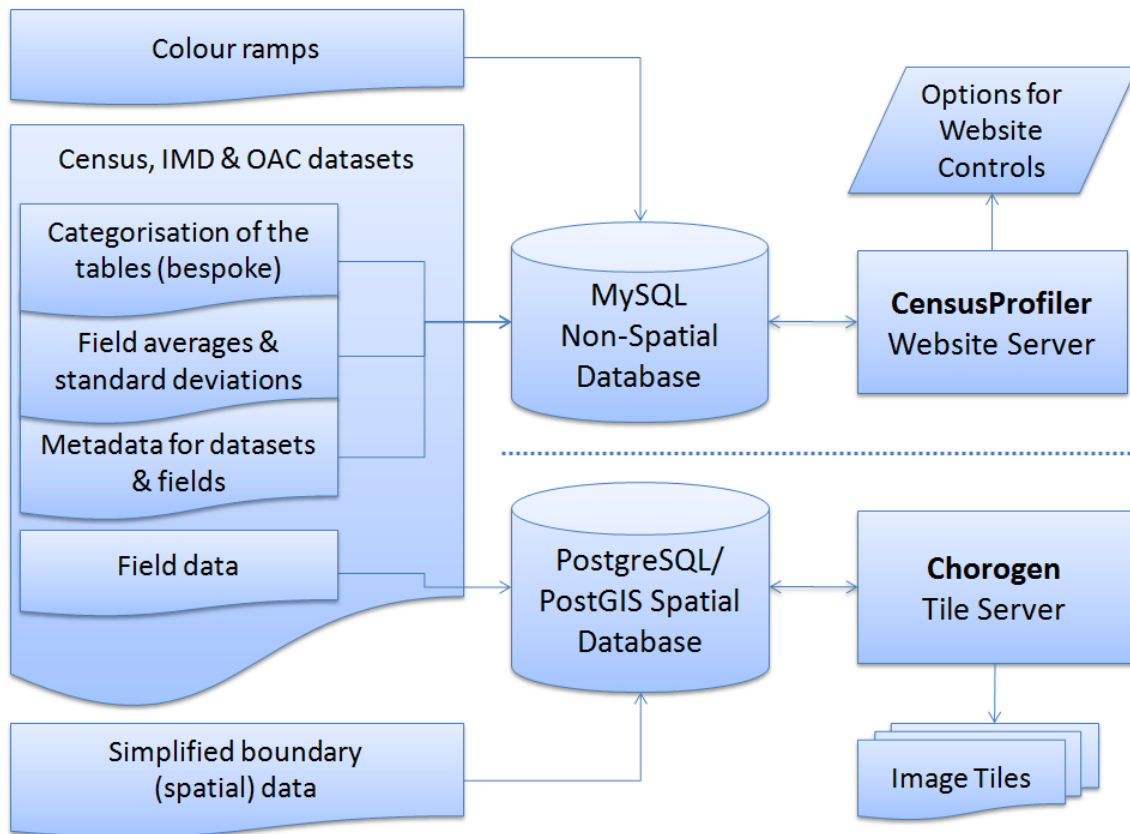


Figure 7: Data architecture diagram for CensusProfiler and Chorogen – server side

	area_code [PK] character	uv0010001 integer	uv0010002 integer	uv0010003 integer	uv0010004 integer	uv0020001 integer	uv0020002 integer	uv0020003 integer	uv0030001 integer	uv0030002 integer	uv0030003 integer	uv0040001 integer	uv0040002 integer	uv0040003 integer	uv0040004 integer
1	00AAFA0001	216	216	0	0				216	115	101	212	3	7	0
2	00AAFA0002	100	100	0	0				100	71	29	93	0	0	0
3	00AAFA0003	215	215	0	0				215	111	104	217	0	3	0
4	00AAFA0004	277	277	0	0				277	152	125	283	3	3	0
5	00AAFA0005	234	234	0	0				234	132	102	233	0	0	0
6	00AAFA0006	222	222	0	0				222	114	108	216	3	0	0
7	00AAFA0007	169	99	70	0				169	80	89	176	3	3	0
8	00AAFA0008	167	167	0	0				167	85	82	167	0	0	0
9	00AAFE0001	107	104	3	3				106	63	43	115	0	3	0
10	00AAFQ0001	131	131	0	0				131	61	70	134	0	0	0
11	00AAFQ0002	136	136	0	0				136	64	72	150	3	0	3
12	00AAFQ0003	230	230	0	0				230	91	139	233	3	5	4
13	00AAFQ0004	260	158	102	0				260	131	129	258	0	0	0

Figure 8: Example of the UV Census table data

The screenshot shows the phpMyAdmin interface for the 'censusgiv' database. The 'Metadata_Census' table is selected, and its structure is displayed. The table has 11 columns: TabName, Description, Caveat, Category, and regional columns for 'england', 'scotland', 'wales', and 'nireland', along with 'comments' and 'exclude'.

TabName	Description	Caveat	Category	england	scotland	wales	nireland	comments	exclude
UV001	TABLE UV01 TOTAL POPULATION	Cells in this table have been randomly adjusted to...	General	1	0	1	0		0
UV003	Table UV03 SEX	Cells in this table have been randomly adjusted to...	General	1	1	1	0		0
UV004	TABLE UV04 AGE	Cells in this table have been randomly adjusted to...	General	1	1	1	0		0
UV005	TABLE UV05 SCHOOLCHILDREN AND STUDENTS IN FULL TIM...	Cells in this table have been randomly adjusted to...	General	1	1	1	0		0
UV006	Table UV06 DEPENDENT CHILDREN	Cells in this table have been randomly adjusted to...	General	1	1	1	0		0

Figure 9: Example of the Metadata table (detail 1)

The screenshot shows the phpMyAdmin interface for the 'censusgiv' database. The 'Metadata_Dictionary' table is selected, and its structure is displayed. The table has 14 columns: Variable, Description, TabName, and various statistical measures (oa_avg, oa_sdp, lsoa_avg, lsoa_sdp, msoa_avg, msoa_sdp) along with regional columns for 'E', 'W', and 'S'.

Variable	Description	TabName	oa_avg	oa_sdp	lsoa_avg	lsoa_sdp	msoa_avg	msoa_sdp	E	W	S
UV0010001	ALL PEOPLE	UV001	0	0	0	0	0	0	[BLOB - 1B]	[BLOB - 1B]	[BLOB - 1B]
UV0010002	All people resident in households	UV001	0.98693524	0.05332466	0.98342546	0.0463818	0.98249546	0.03272809	[BLOB - 1B]	[BLOB - 1B]	[BLOB - 1B]
UV0010003	All people resident in Communal Establishments	UV001	0.01306443	0.05332427	0.0165743	0.04638178	0.01750432	0.03272807	[BLOB - 1B]	[BLOB - 1B]	[BLOB - 1B]
UV0010004	Of which People Sleeping Rough are	UV001	2.076e-05	0.00070876	1.834e-05	0.00029606	1.863e-05	0.00015195	[BLOB - 1B]	[BLOB - 1B]	[BLOB - 1B]
UV0030001	ALL PEOPLE	UV003	0	0	0	0	0	0	[BLOB - 1B]	[BLOB - 1B]	[BLOB - 1B]
UV0030002	Male	UV003	0.48405143	0.03851597	0.48544856	0.02201769	0.48578302	0.01435413	[BLOB - 1B]	[BLOB - 1B]	[BLOB - 1B]
UV0030003	Female	UV003	0.51594857	0.03851597	0.51455144	0.02201769	0.51421698	0.01435413	[BLOB - 1B]	[BLOB - 1B]	[BLOB - 1B]

Figure 10: Example of the Metadata table (detail 2)

4. CREATING THE MAP

As previously mentioned, the CensusProfiler website is based on a hybrid solution to serve raster maps, some of which are created “on-the-fly”, following a user’s request, and others pre-created and stored in the server.

The purposely-developed map tiling service used in CensusProfiler is called *Chorogen*. This allows the user to create custom-built raster choropleth maps for any of the approximately 970 variables available, combined with a wide range of alternative colour schemes, multiple number of intervals with which to group the raw data, and for three different spatial units. Moreover, all of this is done at various cartographic scales (or “zoom levels” in the mapping-mashup jargon) ranging from the whole of Great Britain to street-block level detail, with the number of tiles increasing quadratically with each zoom level. An approximation of the very large potential number of combinations possible is shown in Equation [1]

$$N = \sum_{uv=01}^{970} \sum_{b=3}^9 \sum_{z=6}^{16} T_z G_b C X_{uv} \quad [1]$$

Equation 1: Indication of the number of variables needed. N = number of tile images needed, T = number of tiles at this zoom level to cover the whole of the UK, z = zoom level, G = grouping, b = number of groups, C = number of colour schemes, X = census variables, uv = census univariate variable reference number

Because of the enormous number of possible combinations, each resulting in a unique map tile that could therefore be viewed, it is essential to be able to create the map tiles “on demand” in an efficient and timely manner, as opposed to pre-rendering these maps, as done in the LondonProfiler.org website that was built using the GMapCreator software. Therefore, we developed the *Chorogen* service in order to efficiently create custom-made maps in a bespoke manner. The website architecture also employs limited “caching” of the most popular map views, to avoid duplication efforts and accelerate response time, but for the vast majority of queries these are created “on-the-fly”. As described in the evaluation section, there are obvious trade-offs in this approach, between, on the one hand; the response time and the number of concurrent users, and on the other; the degree of flexibility provided to users in requesting custom-made maps. The current CensusProfiler prototype, just based on two servers, has limited capability to support a large number of concurrent users but provides much better flexibility than most other raster-based web mapping systems available. However, the technology is designed in a way that with additional computational power and server space a large number of users can be easily supported to create the multiple imagery on demand.

The Chorogen map tiling service provides one layer of imagery at a time, the combination of which are overlaid in CensusProfiler website, as shown in Figure 11.

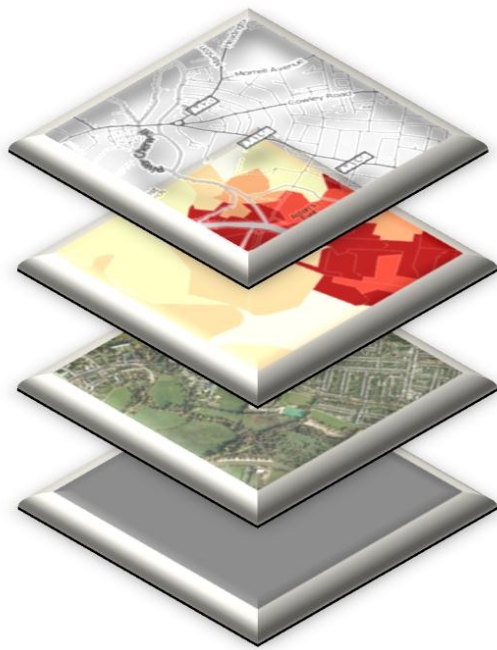


Figure 11: The layers of imagery that make up CensusProfiler.

The context layer is on the top, the Map Tiling Service (Chorogen) provides the choropleth mapping (second layer from the top) that “shines through” the context layer. It can be partially faded to show a base layer of aerial imagery below, or a grey base layer is present to show regions where no thematic data is available.

The Chorogen service receives a request with certain map characteristics, accesses the spatial database, retrieves the necessary information, and performs the rendering of a number of map tiles that fulfil the required extent and scale specifications. These two steps, a) the generation of the map tile specification followed by b) the actual rendering of such tiles, will be described in this paper under separate sections. Section 4.1 describes the former step while section 4.2 deals with the latter.

4.1 Map Tile Specification – URL Parameters

The request for a specific Census map starts with the user navigating to a specific geographical area (a) and level of zoom (b), selecting a Census variable to map (c), deciding the number of intervals (d) in which the data will be grouped and a colour scheme (e) to display the thematic map (for details see Section 5). Such request will therefore contain five parameters, labelled (a) to (e) above. These parameters are actually put together by the user’s browser, using an OpenLayers JavaScript map framework that will be discussed in detail later in this paper. The user’s computer requests the necessary choropleth map tiles through a URL (Uniform Resource Locator or web address) that contains the specified parameters. These are summarised as the geographical coordinates of the tile required, the zoom level resolution and the choropleth map parameters described above.

The following example of an actual URL request, corresponds to a map tile showing the male versus total population for part of the City of London using a red to blue bipolar colour scheme. The system is so flexible that if the reader actually clicks on this URL an “on-the-fly” request to

Chorogen will be generated to produce and serve such tile in the reader's browser, as shown in Figure 14.

http://tiler1.censusprofiler.org/standard_dev|uv_|mlo|UV0030002|UV0030001|0.48578302|0.48544856|0.48405143|0.01435413|0.02201769|0.03851597|cb|RdYlBu|na_1|8|0/13/4093/2723.png

The parameters in the URL contain all that is needed to specify the choropleth map tile required. Appendix A describes each of the parameters, their use and range of possible values. The request is automatically constructed by the JavaScript code running in censusprofiler.org on the user's computer internet browser and its structure is not actually seen by the user.

Since a thematic map in the CensusProfiler website will be necessarily composed of several map tiles, the JavaScript codes makes a single URL request for each of the map tiles required to fill the browser's window on the user's computer. This typically accounts for around 20 requests, while panning around the map fires off additional URL requests as the user seamlessly navigates around the map. Each URL request sent by the user's computer is received by a standard *Apache* web server, which has the *mod_python* module enabled and appropriate instructions for managing file directories. Through this process the map request from CensusProfiler is then passed to a *Python* script that executes it in the server.

The executing Python script takes the specified parameters, does some pre-processing to calculate the necessary spatial units to be used, and creates a series of programmatic value-colour rules each of these representing a colour "bin", to be used by the map rendering module, called *Mapnik* which will be described in the following section.

The RGB values for the colour ramps are created from the colour names using a python module called *colorbrewer* that implements the most commonly adopted colour schemes for thematic maps (see www.colorbrewer2.org). The colour rendering is described later in the next section.

If the request corresponds to a very detailed scale (lowest zoom level), the resulting map tiles are saved to the server's disk for caching purposes, and returned back to the user's browser.

```
[metric] < 0.452422025
[metric] >= 0.452422025 and [metric] < 0.46343087
[metric] >= 0.46343087 and [metric] < 0.474439715
[metric] >= 0.474439715 and [metric] < 0.48544856
[metric] >= 0.48544856 and [metric] < 0.496457405
[metric] >= 0.496457405 and [metric] < 0.50746625
[metric] >= 0.50746625 and [metric] < 0.518475095
[metric] >= 0.518475095

(select *, cast(UV0030002 as float)/UV0030001 as metric from gb_lsoa_2001
as b , uv_lsoa as d where d.area_code = b.area_code) as foo
[(215, 48, 39), (244, 109, 67), (253, 174, 97), (254, 224, 144), (224, 243,
248), (171, 217, 233), (116, 173, 209), (69, 117, 180)]
```

Figure 12: Example of the code generated to assign colour values to a map request

The URL example shown in this page creates the following colour value rules, and generates an SQL query and an array of RGB colour triples. These rules, query and colour values are passed programmatically into *Mapnik* by the *Chorogen* Python script

The overall workflow and modules involved are summarised through the chart shown in Figure 13, while an example of the SQL code to generate the colour scheme is shown in Figure 12.

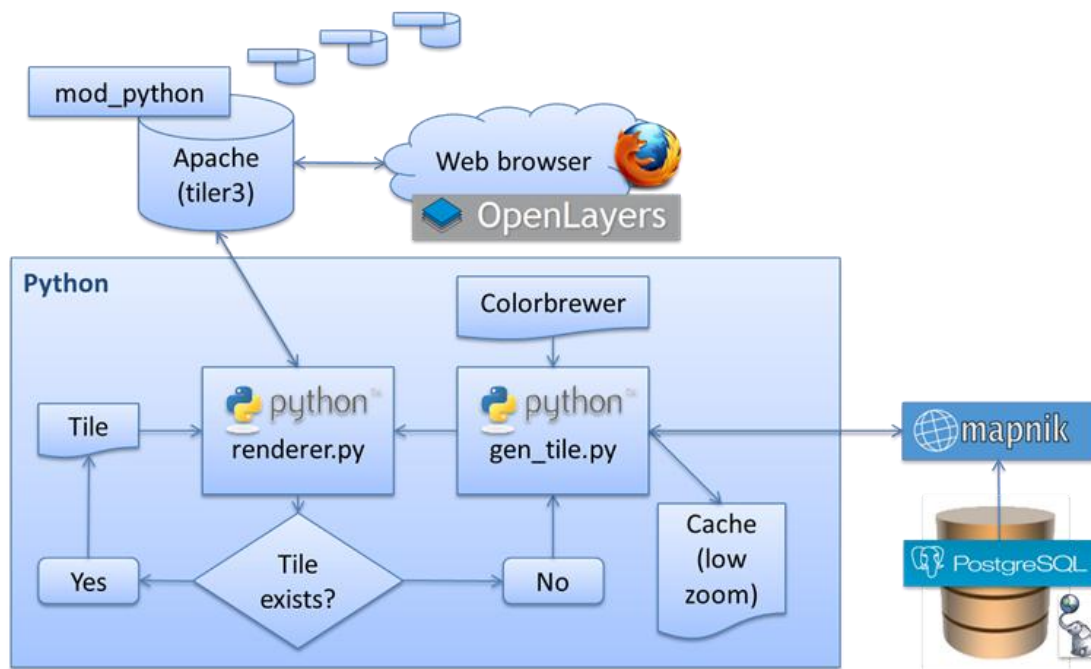


Figure 13: Technologies and workflow used for each URL request for a map tile

4.2 Map Tile Rendering – Mapnik

At the core of the map tile rendering service is the *Mapnik* engine. Mapnik is an open source map engine, with some advanced cartographic styling functionality. Its most high-profile use is as the “default” map creation for the popular OpenStreetMap (<http://www.openstreetmap.org>).

Mapnik has been designed with good integration with the Shapefile map format (a standard format for spatial data defined by ESRI) and PostGIS spatial databases as data sources. It has a powerful Python API, and features very fast performance. The cartographic design is configured by a *stylesheet*, containing pairs of layers and CSS-like layer styles, either specified in an XML file or built up programmatically in Python. It can output in a variety of formats, including, amongst others, PNG raster imagery (the format used in CensusProfiler) and PDF vector imagery.

Mapnik simply creates a map of a fixed size for a fixed geographic area using spatial and non-spatial data retrieved from a database. In our system design, a single call to Mapnik is made for each image tile. Each of these calls produces a small tile (in PNG format) that are then sent back to the user browser to be tiled together into a mosaic that forms the seamless map displayed in the user’s screen.

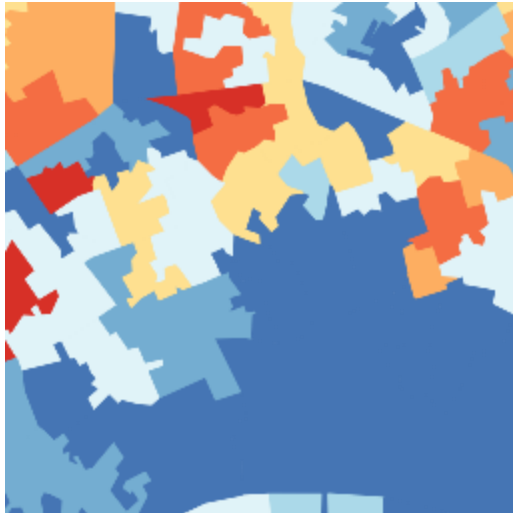


Figure 14: The map tile created from the URL request shown in the example mentioned in the text

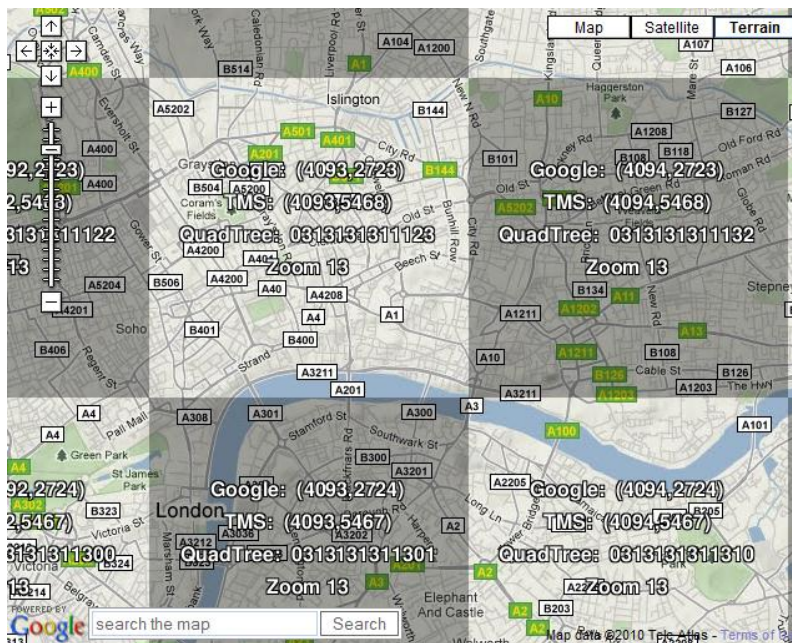


Figure 15: An example of Google image tile coordinates

At the centre of the image is Google title with XYZ coordinates 13/4093/2723, corresponding to part of the City of London, as in the example shown in Figure 14. Those of adjacent tiles are also partially shown (courtesy of Maptiler.org)

4.3 Context Map

The map tile rendering process described in the previous subsection, using Mapnik, was also used as a pre-rendering process to create the map tiles for the black and white “context layer” shown in Figure 11 as the uppermost layer, lying above the choropleth map layer. This allows the colour choropleth mapping to “shine through” the context layer. The process of creating the context layer involved retrieving a copy of the OpenStreetMap database of spatial objects, and a customised stylesheet that concentrates on linear features and points of interest. The OpenStreetMap database for Great Britain was “clipped” using a coastline boundary in order to remove off-shore features resulting in a more familiar looking British coastline.

The context layer is pre-rendered because there is only a single version of this layer for each area and zoom level, as opposed to the almost limitless number of tiles possible with the Census choropleth maps. The pre-rendering process⁶ is identical to how the choropleth maps are created, although instead of using the web server to create the tiles, they are sequentially rendered them for the whole of the UK. The customised stylesheet was programmed in a way that the context map would only show the most prominent features at a high zoom level (smallest cartographic scale), with increasing amount of geographical detail for lower zoom levels (larger cartographic scales). For example, cities, towns and villages are weighted by population, roads and streets by its level of hierarchy, and so on. The degree of detail shown in the final context map is automatically determined by the zoom level.

5. DISPLAYING THE MAP

This section describes the final ingredient in the overall CensusProfiler service, the geographic visualization of the maps created by the different services covered in the previous section. A screenshot of the CensusProfiler user interface, as it appears when first launched is shown in Figure 16.

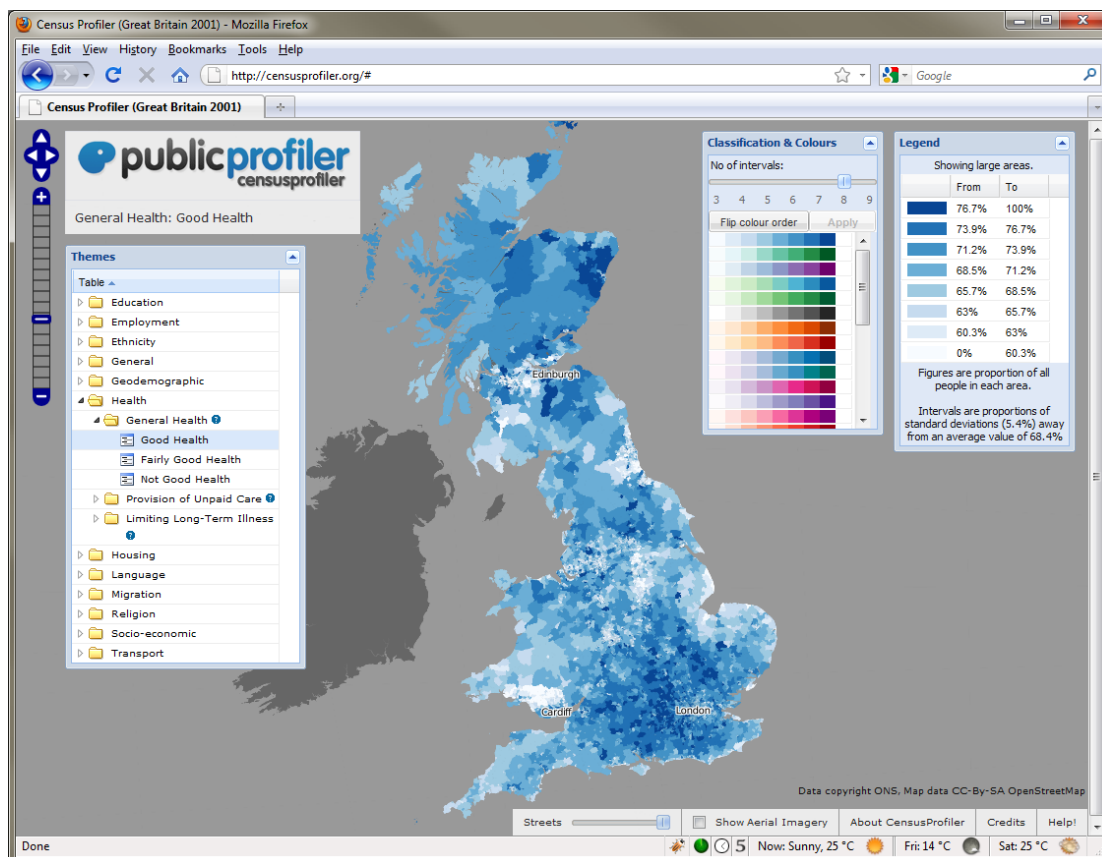


Figure 16: A screenshot of the censusprofiler.org website user interface

⁶ The script used is a customised version of the “generate_tiles.py” script used in the OpenStreetMap project and available on that project’s source code repository.

The CensusProfiler website is actually hosted on a different server to those creating and distributing the choropleth and context maps. The server is of a lower specification than the mapping servers, as it simply serves the static webpage content and manages a small database of metadata. Once the website is initially loaded onto the user's internet browser, there is no further communication with the CensusProfiler server. The dynamic elements of the website, such as requesting the imagery and updating the user interface elements in response to the feedback given, is purely carried out on the client's computer browser in Javascript. This off-loads server capacity and minimises data traffic between the client and the server.

The description of the Census and additional variables (the metadata) are loaded onto the user's browser when the website is first accessed⁷. The size of the complete set of metadata, including the average and standard deviation for all three spatial levels, is just around 200KB, being unnoticeable to the user.

5.1 Visualization – OpenLayers

We have also adopted open source technology to display the final map, request new mapping and other imagery from the tile servers, and handle mouse/keyboard input for navigating around the map. This technology is OpenLayers Javascript mapping API (<http://www.openlayers.org/>), which is implemented here unmodified, so it could potentially be served directly from the Openlayers website rather than from our server.

The feature-set of OpenLayers is similar to that offered by Google Maps, Microsoft Virtual Earth, Cloudmade's Web Maps Lite and other "slippy map" platforms. However, Openlayers offers a richer set of capabilities, particularly around vector display and interaction with more layer types, specially the integration of satellite and aerial imagery from OpenStreetMap, Google and Microsoft. For the CensusProfiler website, OpenLayers is used to provide the choropleth map and context map layers, using the "XYZ" layer type, and Google satellite Imagery, using a "Google" layer type (as shown in Figure 11). Vector features are not used in CensusProfiler.

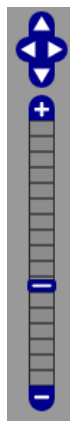


Figure 17 The standard OpenLayers navigation controls as used in the CensusProfiler.

⁷ The metadata is actually handled in the user's internet browser using a PHP script which creates json-style hierarchical Javascript arrays with the results of the user's queries. AJAX is not used in CensusProfiler, except internally by OpenLayers to load the maps.

5.2 User Interface – Non-mapping elements

The non-mapping elements of the user interface in the website are comprised of; a hierarchical selection tool with the Census variables, a colour ramp selector, a slider for the number of data intervals, sliders to select or fade in/out the aerial imagery and context layers, and a dynamically changing legend for the choropleth Census map.

This user interface was built in a consistent and polished fashion through the use of Ext JS, a JavaScript application framework produced by Sencha⁸. A publically available extension to Ext JS, *TreeGrid*⁹, combines the standard *Ext JS Grid* (a table) with an *Ext JS Tree*, to provide a hierarchical tree structure with multiple columns. This is used as the basic Census data selector as shown in Figure 18.

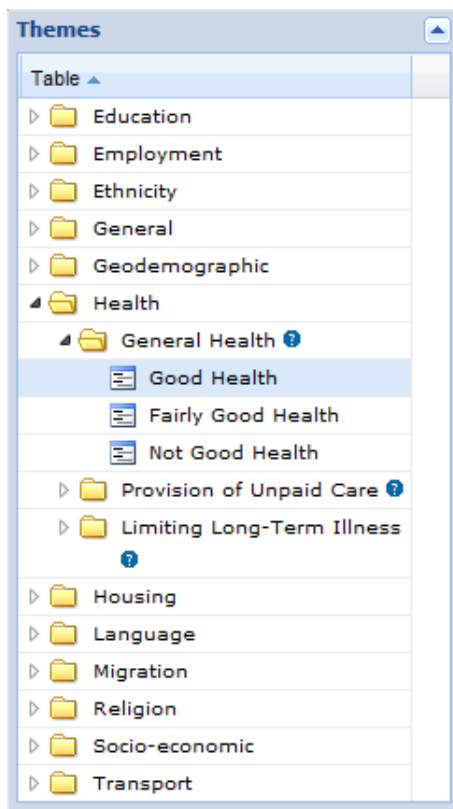


Figure 18 The hierarchical Census data selection tool in CensusProfiler.

A separate panel allows the user to select the colour scheme and number of intervals with which to classify the data. This panel was built with a customised *Ext JS Grid* is used, along with a slider and two buttons. The second button, with the label “Apply”, is used when the user wants to confirm a new colour and interval selection, for the map to be created and displayed. This “Apply” button was introduced in order to prevent overwhelming the server with multiple requests automatically generated by rapid adjustments to this dynamic interface.

⁸ <http://www.sencha.com/products/js/>

⁹ <http://extjstreegrid.sourceforge.net/>

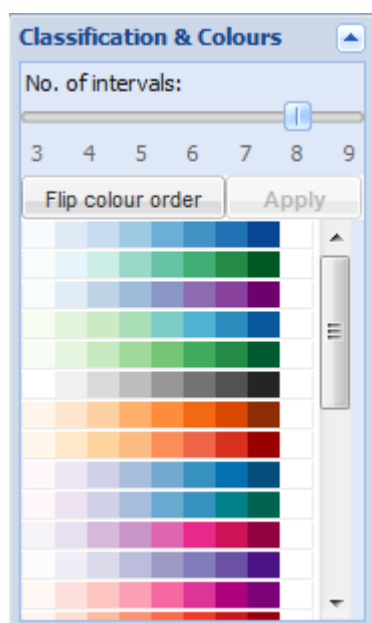


Figure 19 The colour scheme and number of intervals selection tool in CensusProfiler.

A dynamically changing map legend is produced every time the displayed Census variable, colour scheme and/or number of intervals change. This legend is also created using Ext JS by mimicking the logic used in the *Chorogen* map tile rendering service to assign colours to data values. The legend is designed to automatically update following the exact appearance of the map tiles served by Chorogen. Unfortunately, these colour rules had to be manually replicated between the Python scripting in Chorogen to the Javascript in the CensusProfiler website, rather than both sources using a common source of rules. No information from Chorogen itself is used to update the legend.

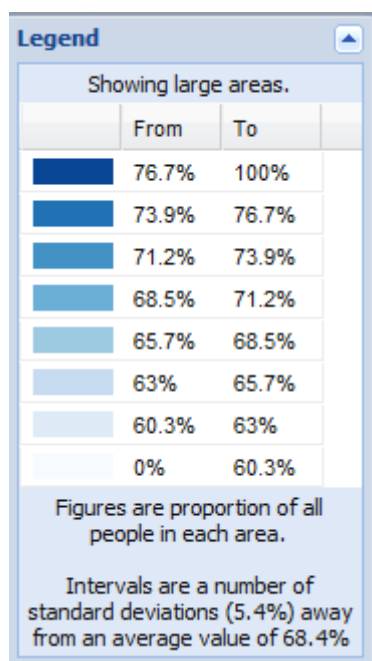


Figure 20 The map legend display in CensusProfiler.

All these three panel components (Data selector, Colour and interval selector, and Map Legend) are moveable and collapsible. Ext JS proved to be an extensive and powerful framework to produce these user interface elements, while more interesting features could be easily added to the website in the future.

5.3 Choropleth Map Colour Schemes

In order to provide a consistent, meaningful and pleasing set of colour ramps to display the Census choropleth maps, as well as making sure that these conform to established cartographic design principles, we used the well recognised (Harrower 2011) *Colorbrewer* palette (www.colorbrewer2.org)

Colorbrewer is used in two different parts of the CensusProfiler system. The Chorogen map tile server uses it through the *colourbrewer.py* module in Python. Furthermore, two database tables contain colour information from colorbrewer. These tables contain the RGB values for each set of colour ramps and possible number of intervals. These values are in turned used by the Colour and Interval selector on the user interface, to display the colour ramps available on the CensusProfiler website, and once selected they are passed to the Chorogen map tile server for rendering. These RGB values have been extracted from the spreadsheet available for download from the Colourbrewer 2.0 website¹⁰.

Apart from the Colorbrewer schemes, for the Output Area Classification (OAC) the standard “OAC colours” for consistency with other publications (except for the green colour for the “countryside group” which has been replaced with a less bright hue).

6. PERFORMANCE TESTING

Before the CensusProfiler website was made available to the general public, we conducted a test in order to ascertain the performance of the Chorogen dynamic map tile rendering. A web service called *OpenOrienteeringMap*¹¹ (OOM) was adapted for this purpose. This service allows anyone in the world to create a printable PDF map of the OpenStreetMap database but styled for orienteering. The map is created on-the-fly using the Chorogen service. In this case, the maps are not choropleth maps but the technique used to produce them, is almost identical, the main difference being the lack of dynamic colour ramp and number of data intervals. The OOM service was based on exactly the same Chorogen engine running on a different Chorogen server with similar specification to the one used by CensusProfiler. The OOM service was released in early 2010 to the general public and has been steadily used around the world to create orienteering maps. No significant performance hit on the Chorogen engine has been observed in over a year since this service was launched, a reasonable demonstration of good reliability for this architecture.

¹⁰ <http://colorbrewer2.org/>

¹¹ The service can be accessed at <http://oobrien.com/oomb/>

7. SCALING

The key performance bottleneck of CensusProfiler website is the Chorogen on-the-fly map rendering service, and within this service the database index read performance proved to be the slowest link in the overall chain of events. This hindrance could be mitigated in a number of ways, that are listed here for others to improve future versions of similar mapping mashup websites:

- Increasing the number of Chorogen servers with identical copies of the database tables would help to scale the system up to a large number of simultaneous users, without experiencing delays. This could be easily accommodated through hiring “Cloud Computing” type of services, such as Amazon Cloud or OpenStack in the open source community.
- Use larger spatial units, such as MSOAs for small-scale mapping, so that the user quickly receives a map while the finer level geography is rendered.
- Some further optimisation of the data tables can also be made. Replacing the alphanumeric keys with short integer numbers was shown to increase performance of the table indexes resulting in the Chorogen service running around 20% faster. This however came at the one-time expense of converting the keys and maintaining a lookup table for future debugging. Other table optimisation opportunities could also be found
- OpenLayers tends to over-request image tiles when the user quickly zooms in or out. In these cases imagery for the intermediate zoom levels are still requested, even when those levels are no longer visible in the user’s web browser. This could be remedied by building a filter for such cases.
- Increased caching of the most commonly used datasets, colour ramps and interval groups at the at the two lowest zoom levels. This would need to be scripted for each combination, it will be time consuming to build and extensive storage space would be needed of for a very large number of tiles.
- The use of 8-bit imagery (the PNG256 format) with 256 colours rather than the 32-bit imagery used by default. There would be a slight performance drop with the extra processing time required to write and transmit every image, but a considerably smaller bandwidth usage and cache storage requirements. These improvements would easily outweigh and the slight image quality lost, since these would just look marginally less attractive.

Despite these suggested improvements to scale up the service, the current server hosting the CensusProfiler website should be able to cope with a sustained large number of simultaneous users without issue.

8. CONCLUSION

Through this paper we have described the key characteristics of the CensusProfiler website, a prototype service to disseminate the Census of Population through meaningful geographical visualization. A full review of the different technological options available was conducted before justifying the development of a new approach to web mapping based on open source programs and geographic datasets.

The Censusgiv project has demonstrated that it is possible to create an automated web-based service to map the large number of census datasets giving the user a certain degree of flexibility over the characteristics of the desired map, such as colour schemes, number of intervals and geographic scale and extent. Furthermore, such service is provided through a very intuitive user interface that does not require any pre-existing knowledge of the Census administrative geography, types of Census tables, geographic information systems, data interval classification schemes, and so forth. The CensusProfiler website was developed using a raster-based approach to map tiles that are rendered on-the-fly. To our knowledge, this is one of the very first such services of its kind, at least in the area of Population Geography.

Our final recommendation to authorities in charge of National Statistics in various countries is that they can easily introduce substantial improvements in the way they disseminate the forthcoming 2010/2011 Census datasets, as well as other population datasets at small geographical scales. As demonstrated by this paper, this can be easily and cheaply implemented through the development of open source mapping functionality similar to the architecture proposed in this paper. Failing this, they should in any case always make such datasets easily accessible to the general public, through dynamic data cubes or other forms of data streaming in popular data formats, in order to allow the development of innovative geovisualizations by the web community.

The key legacy of the Censusgiv project will be the censusprofiler.org website that we hope will act as a beacon to the dissemination of the UK Census both before and after the 2011 Census is disseminated. We hope that this website, together with the technical documentation provided, will show the way in which future methods for the dissemination of Census information to the general public should be developed. This is very relevant to the ONS, the other UK Official Statistics institutions, as well as to the ESRC, and similar outfits in other countries. This technology will enable these institutions to maximise their large investments in data collection, storage and delivery for the next round of 2011 Censuses. Moreover, the effectiveness in which the proposed technology can deliver such geovisualizations to a mass audience completely based on Open Source software, will be key for these institutions at a time of economic hardship and public budget restrictions.

9. REFERENCES

- Adobe, 2008. Adobe Flash®. Available at: <http://www.adobe.com/> [Accessed December 4, 2008].
- Alesheikh, A.A., Helali, H. & Behroz, H.A., 2002. Web GIS: Technologies and Its Applications. In Geospatial Theory, Processing and Applications. ISPRS Commission IV, Symposium.
- Buckley, A. R., Gahegan, M. & Clarke K (2000) *Geographic Visualization as an emerging themes in GIScience*. University Consortium for Geographic Information Science (UCGIS), White Papers. Available at [http://dusk.geo.orst.edu/ucgis/web/emerging/Geographic visualization.pdf](http://dusk.geo.orst.edu/ucgis/web/emerging/Geographic%20visualization.pdf) (last accessed 23 January 2011.)
- CIBER (2008) *Information behaviour of the researcher of the future*. A report commissioned by The British Library and JISC. <http://www.bl.uk/news/pdf/googlegen.pdf> (last accessed 3 September 2010.)
- CLG (2007) *Indices of Deprivation 2007 – Communities and neighbourhoods*. Website: <http://www.communities.gov.uk/communities/neighbourhoodrenewal/deprivation/deprivation07/> (last accessed 5 July 2010.)
- Dodge, M., McDerby, M. & Turner, M., 2006. The Power of Geographical Visualizations. In *Geographic Visualization: Concepts, Tools and Applications*. New York: Wiley. Available at: http://media.wiley.com/product_data/excerpt/12/04705151/0470515112.pdf.
- Harrower, M. & Brewer, C. A. (2011) ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps, in *The Map Reader: Theories of Mapping Practice and Cartographic Representation* (eds M. Dodge, R. Kitchin and C. Perkins), John Wiley & Sons, Ltd, Chichester, UK.
- ESRI (2009) *ArcGIS Desktop Help 9.2 – Simplify Polygon (Data Management)*. Website: [http://webhelp.esri.com/arcgisdesktop/9.2/index.cfm?TopicName=Simplify Polygon %28Data Management%29](http://webhelp.esri.com/arcgisdesktop/9.2/index.cfm?TopicName=Simplify%20Polygon%20Data%20Management%29) (last accessed 5 July 2010.)
- Garrett, J.J., 2005. Ajax: A New Approach to Web Applications. Adaptive Path, 18.
- gCensus, 2008. gCensus: Free online GIS using Google Earth. Available at: <http://gecensus.stanford.edu/gcensus/index.html> [Accessed November 16, 2008].
- Gibin, M., Singleton, A., Milton, R. et al., 2008. Collaborative Mapping of London Using Google Maps: The LondonProfiler. UCL Working Papers Series.
- Goodchild, M., 2005. What does Google Earth Mean for the Spatial Sciences. In *Biennial Conference of the Spatial Sciences Institute*. pp. 12-16.
- Google, 2008. Google Earth. Available at: <http://earth.google.com/> [Accessed November 16, 2008].
- Martin, D., 2000. Towards the Geographies of the 2001 UK Census of Population. *Transactions of the Institute of British Geographers*, 25(3), 321-332.

MIMAS Census Dissemination Unit , 2008. CASWEB. Available at: <http://casweb.mimas.ac.uk/> [Accessed November 16, 2008].

Monmonier, M., 2007. Cartography: the multidisciplinary pluralism of cartographic art, geospatial technology, and empirical scholarship. *Progress in Human Geography*, 31(3), 371.

OAC (2007) *Output Area Classification User Group – OAC – What is the Output Area Classification?* Website: <http://areaclassification.org.uk/getting-started/getting-started-what-is-the-output-area-classification/> (last accessed 5 July 2010.)

Office for National Statistics, 2008. Neighbourhood Statistics - Home Page. Available at: <http://www.neighbourhood.statistics.gov.uk/dissemination/> [Accessed November 16, 2008].

Open Geospatial Consortium, 2008a. Web Feature Service | OGC®. Open Geospatial Consortium. Available at: <http://www.opengeospatial.org/standards/wfs> [Accessed December 4, 2008].

Open Geospatial Consortium, 2008b. Web Map Service | OGC®. Open Geospatial Consortium. Available at: <http://www.opengeospatial.org/standards/wms> [Accessed December 4, 2008].

Ordnance Survey (2010) *Boundary-Line – Britain's electoral and administrative boundaries*. Website: <http://www.ordnancesurvey.co.uk/oswebsite/products/boundaryline/> (last accessed 5 July 2010.)

Scottish Neighbourhood Statistics (2010) *Scottish Neighbourhood Statistics, Data Downloads*. Website: <http://www.sns.gov.uk/Downloads/DownloadGeography.aspx> (last accessed 5 July 2010.)

Tu, S., Flanagan, M., Wu, Y. et al., 2004. Design strategies to improve performance of GIS Web services. In *Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004. International Conference on Information Technology*.

10. APPENDICES

10.1 URL Parameters

The following table describes the parameters used by requests from the CensusProfiler website Javascript to the Chorogen server, in order to receive the choropleth mapping tiles.

URL component and delimiter	Specified by	Meaning
http://tiler1.censusprofiler.org/	CensusProfiler website.	The server which will provide the image. OpenLayers can allow multiple servers to be specified and will round-robin between them, allowing for a basic level of scalability.
,	CensusProfiler website.	The comma designates that a dynamic request follows, rather than a pre-existing static image which the Python script would simply retrieve and return to the browser. However the script will still use a cached image if available.
standard_dev	CensusProfiler website, based on the average value supplied by the database.	<p>The system of data classification to calculate the break points around which the map intervals should be built. CensusProfiler currently supports:</p> <ul style="list-style-type: none"> • standard_dev, which splits data into intervals of one or half standard deviations away from the mean (or straddling it) • geometric_small, which data into intervals that are $\frac{1}{4}$, $\frac{1}{2}$, twice and four times the mean. Suitable for ratios heavily skewed towards zero. • geometric_large, similar to geometric_small but for values skewed towards one. • classification_quan which assumes that values are ranked and splits them into equal population sized intervals • classification_qual which assumes a qualitative classification scheme and restricts the bin count and selectable colour types appropriately. <p>Each Census variable has been allocated a pre-specified interval classification method, which at present cannot be changed by the user.</p>
uv_	Database, modified by CensusProfiler.	The name of the table with the data. If an underscore is appended, as shown in the example here, the name of

		the spatial unit to be used will be appended on.
mlo	CensusProfiler website.	The areal unit desired (oa, lsoa or msoa) – if “mlo” is specified, this allows the python script to choose the unit based on the requested scale. Currently the CensusProfiler website will only pass “mlo” to ensure good tile generation performance.
UV0030002	Database, user specifiable.	The name of the column which corresponds to the field being measured.
UV0030001	CensusProfiler website derives this from the measured field by replacing the last four digits with “0001”.	The name of the column which corresponds to the “base” field which, by acting as a denominator, allows the results to be calculated as a ratio rather than raw counts. Not used the classification grouping methods.
0.48578302	Database.	Average of the field value ratios, for OA, LSOA and MSOA spatial units respectively. For fields which use classification groupings, the minimum value for that spatial unit - generally 1.
0.48544856	Database.	
0.48405143	Database.	
0.01435413	Database.	Population standard deviation of the field value ratios, for OA, LSOA and MSOA spatial units respectively. For fields which use classification groupings, the maximum value for that spatial unit.
0.02201769	Database.	
0.03851597	Database.	
cb	CensusProfiler website.	The family of the colour ramp to be used. Currently accepts “cb” (Colorbrewer) or “pre” (pre-determined)
RdYlBu	CensusProfiler website, user specifiable (although the range of values is restricted to a quantitative or qualitative set based on whether the classification_qual grouping method is being used.	The name of the colour ramp to be used. Currently accepts “oac” if the family is “pre”, or one of the Colorbrewer colour names.

na_1	CensusProfiler website.	This field was intended to specify other grouping methods but is not currently used.
8	CensusProfiler website, user specifiable.	The number of groups the data should be split into, each represented by a different colour on the choropleth map.
0/	CensusProfiler website, user specifiable.	Whether the colour ramp should be reversed.
13/	CensusProfiler website, user specifiable.	The zoom level (scale), using Google XYZ tile addressing.
4093/	CensusProfiler website, user specifiable.	The X coordinate, using Google XYZ tile addressing.
2723	CensusProfiler website, user specifiable.	The Y coordinate, using Google XYZ tile addressing.
.png	CensusProfiler website.	The format that the image should be in. Currently only “png” is supported, although using “new” instead ensures that a image is always freshly created, rather than using any cached image.

10.2 System Specifications

The systems used to host the CensusProfiler website has the following specifications

Chorogen Map Tiling Server

- Dell Precision 690 workstation with quad-core Intel Xeon 5300-series processor.
- Ubuntu 10.04 Linux operating system (desktop-optimised).
- 4GB memory. The full amount is not utilised.
- 1GB/s network connection.
- Apache 2.2 with mod_python.
- PostgreSQL 8.4 with PostGIS 1.4.
- Python 2.6 with Mapnik library and colorbrewer-0.1.0.
- Mapnik 0.7.1.
- Currently running from a single server for ease of configuration
- Multiple servers could be set up and requests round-robin between them – configuration for this would be set up in the CensusProfiler website, which is straightforward to do but less flexible, or by using multiple servers behind the same domain name – more advanced.
 - Perceived inefficiencies from using multiple servers to return imagery from similar areas and for similar attributes is actually very small. By far the greatest amount of time is normally spent accessing the data in the database. The time taken to render the map is small. The time taken to send and receive the data from the client is network dependent rather than server-dependent.

CensusProfiler Website Server

- Dell PowerEdge 2850 rack server with dual-core Intel Xeon processor.
- Ubuntu 10.04 Linux operating system (server-optimised).
- 2GB memory. The full amount is not utilised.
- 1GB/s network connection.
- Apache 2.2 with PHP 5.
- MySQL 5.0.
- OpenLayers 2.9.1.

Browser

- Javascript is required.
- A modern, standards-compliant browser is recommended.
- Tested on Firefox 3.6, Internet Explorer 6, 7 & 8, Chrome 4 and Safari 4.