

Rebuilding the Great Britain Historical GIS, Part 3: Integrating qualitative content for a sense of place

Humphrey Southall

Department of Geography, University of Portsmouth

Abstract

We describe the integration of old maps, descriptive gazetteers and a large library of travel writing into the Great Britain Historical GIS, presenting a range of approaches to geo-referencing diverse historical sources. While previous parts focused on legally defined administrative areas and statistical reporting units, these qualitative sources concern a less formal geography of “places”. We link these to administrative units in two ways: places are contained within units, but units are named after places and are consequently subsidiary to them. While rejecting existing gazetteer data standards, the approach aligns well with that of historical place-name researchers. The final section describes how the structure interacts with search engines to support a very popular web site for life-long learners.

Keywords: historical GIS; gazetteers; travel writing; historical maps.

Rebuilding the Great Britain Historical GIS, Part 3: Integrating qualitative content for a sense of place

Introduction

The main focus of historical GIS has been the creation of geographical frameworks for historical statistics, especially census data, reconstructing the changing boundaries of reporting units from states and provinces down to city blocks or even individual houses. However, recent years have seen growing interest in working with more qualitative material such as travel narratives. This trend is linked to the rise of “digital humanities” as a distinct discipline (Schreiber et al 2004; Cohen 2010), and the involvement of many historical GIS researchers in this new field (Jessop 2007; Bodenhamer et al 2010).

This is the third part of a three-part series describing the evolution of the Great Britain Historical GIS from a relatively traditional vector GIS, implemented using ArcGIS software and described in Gregory and Southall (1998), into a much more diverse geo-semantic structure. The first part (Southall 2011) maintained the original focus on statistical content but explored the new architecture we developed for capturing the meaning of those statistics, based on the work of the Data Documentation Initiative. The second part (Southall 2012) described the administrative unit ontology (AUO) which enables us to hold statistical content

for units with unknown boundaries or even locations, and to support a wide range of gazetteer searches.

However, both previous parts retained the traditional focus of historical GIS on statistics and reporting units, and said little about user interfaces. This final part begins by separately describing three types of qualitative content: historical maps, descriptive gazetteers and travel narratives; we also computerized introductions to census reports, but no attempt has been made to geo-reference these and they will not be further discussed, being held in the same database tables as the travel narratives.

Each account describes the main sources, how each is held in the system, and the associated web interface; these interfaces could all be separate web sites, but actually form parts of one large site, *A Vision of Britain through Time*.

Description of the travel writing leads into discussion of why and how all this content been linked together, and to the statistical content, by defining and constructing a high-level gazetteer of “places”. The final part of the paper describes how the semantic structure interacts with search engines to draw web users searching for information about named places to the site, and how this in turn creates the income needed to sustain the site:

www.VisionOfBritain.org.uk

While boundary mapping and the computerization of historical statistics was funded primarily as academic research, the web site and the qualitative content were mainly funded by the UK National Lottery. Their “Digitisation of Learning Materials” program had three aims: “to support lifelong learning through the provision of a range of specially-created electronic content; to digitise existing material, and to add and integrate new material ... [and] to base content on lifelong learning and education in its broadest sense, and not on the formal education curriculum” (Big Lottery Fund 2006, 4). However, over half the funding of £50m. went to consortia focused more specifically on “sense of place”: to a series of regional consortia creating web sites such as Staffordshire Past-Track (Staffordshire County Council 2003), and to the “Sense of Place (National)” consortium in which our main partner was the British Library (BL), who created the now-defunct CollectBritain web site. So how do you create a “sense of place” by assembling scanned images of historical sources into a web site?

Most projects in the program were based in local libraries, museums and archives, and focused on particular items in their collections with strong local connections: the “sense of place” was implicit. However, a national project lacking a physical collection of its own needed a more conscious strategy: our focus was on geographical surveys of the whole country rather than unique materials in local

collections; but our information architecture and user interface enabled users to access content from all the different kinds of survey via a single search.

Historical maps

“Geographical surveys of the whole country” obviously include the census, and various more specialized statistical surveys such as the annual Farm Census since 1866, and the Ministry of Labour’s Local Unemployment Index 1927-39. It equally obviously includes maps, and especially the work of the Ordnance Survey (OS). The costs of scanning and long-term storage inevitably limited our scope, so we focused on two sets of one inch-to-one mile (1:63,360) maps. Firstly, the *New Popular* series from the late 1940s. The main reason for choosing these was that they were both the first one inch maps to include the modern National Grid coordinate system and, when we were doing this work, the most recent to be out of copyright: digitizing these maps meant we could freely use the National Grid system without breaching OS copyright, and in particular could use these maps to geo-reference other sets. Secondly, the original First Series, published slowly between 1805 and 1891 as the OS worked its way from the south coast of England to the north of Scotland. The earlier sheets were periodically revised by *ad hoc* additions to the copper printing plates, without a clear set of “editions”, so the BL scanned for us the earliest such “state” in their collection for each sheet. They also scanned several less detailed topographic maps from similar dates,

enabling the Web Map Server described below to offer a full range of zoom levels.

Three other projects have extended the map library. Firstly, support from the Department of the Environment, Farming and Rural Affairs and its agencies, and the Frederick Soddy Trust, enabled us to computerize all the one inch maps published by the Land Utilisation Survey of Great Britain, a project based at the London School of Economics in the 1930s, coordinating fieldwork by schools around Britain (Stamp 1948). We were eventually able to include even the unpublished maps of upland Scotland they deposited with the Royal Geographical Society, so finally publishing the whole survey. Secondly, the European Union-funded QVIZ project added 1:500,000 military mapping of the whole of Europe, reaching Moscow, created in the early 1940s by the British General Staff Geographical Survey (GSGS), necessarily entirely by aerial survey. Thirdly, our Historic Boundaries of Britain project in 2007-9 added a large collection of administrative boundary maps, mostly acquired when the Office of National Statistics moved out of their London offices in 1997-8 and manually vectorized during the construction of the original ArcGIS system: it took another ten years before disk storage was cheap enough to put the map images online.

Most map library digitization projects have simply scanned maps and made them available online through image viewers such as MrSid and Zoomify, lacking

geospatial functionality. Our aim, however, was to make the geographical information in the maps accessible to people interested not in the history of cartography but in places. Scanning the maps was therefore only the first stage. We next cropped the sheets to remove all the marginal information, geo-referenced them by finding real world coordinates for multiple locations on each sheet, assembled each series into a single continuous mosaic, and finally re-projected them initially to the National Grid used in modern OS maps, and more recently to the European Terrestrial Referencing System (ETRS-89).

The end result is historical mapping that works like Google Maps: users can zoom in or out, seeing more or less detail; or they can move sideways without hitting the edge of a map, until they fall off the edge of Britain; and unlike Google Maps there is some ability to move in time, switching from modern maps from Open Street Map to 1940s maps, then back to the nineteenth century. The user interface is provided by OpenLayers, like Google Maps in being a Javascript toolbox working within the user's browser, but OpenLayers understands the Open Geospatial Consortium's Web Map Server (WMS) protocol. WMS requests for mapping of particular areas are sent to GeoWebCache on our server, which passes them on to Minnesota Mapserver software if the relevant area is not in the cache. Requests can be passed simply as URLs; the example below returns a 400 by 400

pixel image in PNG format covering a rectangle centered on Greenwich, from our nineteenth century mapping:

http://www.visionofbritain.org.uk/cgi-bin/mapserv?map=/usr/local/share/map-files/bound_map_page.map&layer=first_edition&mode=map&map_imagetype=png&mapext=3329113+2788182+3332566+2791635&map_size=400+400

Our WMS is a general solution to providing historical background maps for any British web site, figures 3 and 7 showing different applications within our site. However, many users need the original maps with all the explanatory text in their margins, and many maps are unsuited to inclusion in mosaics. We therefore have a separate library of unaltered images of individual sheets, implemented using IIPImage, an open source alternative to commercial image servers. The client portion of IIPImage works within browsers while the server portion manages map images held as multi-page TIFFs, which contain several different zoom levels forming pyramids. These maps are not geo-referenced in the same sense as those in the WMS, but we hold bounding box coordinates for every sheet within the main Postgres database and use these to provide a map-based search interface: as the user pans and zooms within an OpenLayers-based interface, the system lists the ten maps whose coverage comes closest to the area currently in the interface. This interface was developed independently of Klokan Technologies' similar MapRank Search system, which we are now using in the separate Old Maps Online project (Southall and Pridal 2012).

Lastly, we have recently added a download facility for historical maps involving a third format, high quality JPEGs being preferred to accelerate downloads. The download system inserts these into a Zip archive which also contains a small file containing the geo-referencing data from Postgres, usage notes and copyright information.

Descriptive gazetteers

We provide some information about even the smallest villages by including nineteenth century gazetteers, consisting of large numbers of very clearly separate entries, arranged alphabetically by the names of places: 55,516 entries from John Bartholomew's *Gazetteer of the British Isles* (1887); 29,411 from John Marius Wilson's *Imperial Gazetteer of England & Wales* (1872); 7,268 from Frances Groome's *The Ordnance Gazetteer of Scotland* (1882-5); and 3,939 from Samuel Lewis's *Topographical Dictionary of Ireland* (1837). Entries were formulaic: the place name; the type of feature; associated and containing administrative units; location relative to larger settlements, rather than a coordinate; and then a description whose length varies with importance. For example:

BROMYARD, a small town, a parish, a subdistrict, and a district, in Hereford. The town stands on the river Frome, 9 miles E of Dinmore r. station, and 14 NE of Hereford. It has pleasant, well wooded, hilly environs ... The property is much subdivided. ... (*Imperial Gazetteer*).

While both descriptive gazetteer entries and travel writings are rich in geographical names, geo-referencing them required different approaches.

Dividing the gazetteer text up into entries was essentially mechanical, and each entry is then held as a separate row in a single database table, **g_dgaz**. Entries for major cities are book length, the Groome entry for Edinburgh containing over 110,000 words including several poems and several statistical tables, so they are marked up internally using HTML. However, searching and referencing is supported by information extracted from the text and held elsewhere.

Three other columns within the gazetteer table hold: a numeric identifier for the entry; the “header”, containing the place name or names from the start of the entry; and the “feature type”, such as “a village” or “a river”. The header is the main source for a separate table, **g_dgaz_name**, linked via the identifier and supporting a simple place name search interface. For example, the header “CAISTOR, or Castor” is the source for two separate rows in **g_dgaz_name**, while text deeper within the entries has been harvested for additional variant names: “called by the ancient Britons Caer-Egarry; and by the Saxons Thong-Ceastre”. Searching on any of these names leads to a web page presenting the relevant entry. The feature type information has been systematically matched to the Alexandria Digital Library’s *Feature Type Thesaurus* (2002), enabling the search interface to offer narrowing by type.

Our original approach to geo-referencing gazetteer entries was by linking them to units in the AUO, the **g_dgaz_link** table defining many-to-many relationships by storing identifiers for both gazetteer entries and units, as well as a code recording whether an entry was about the unit or just for a place within the unit. Almost every gazetteer entry now has the second kind of relationship with an Ancient, Scottish or Irish county, enabling the search interface to also offer narrowing by area within Britain. This interface is accessible here:

<http://www.visionofbritain.org.uk/descriptions>

Because the gazetteer entries have a very regular structure, it was possible to write software for most of the above tasks: separating the text into entries; identifying the header and feature type; identifying directly associated units from the place name and feature type, so linking the Bromyard example above to each of the parish, sub-District and Registration District of Bromyard; identifying county names, and so linking Bromyard to Herefordshire. None of this was perfect, but we have done a substantial amount of further manual editing.

[Figure 1 appears near here]

Although the gazetteers were funded as a resource for local historians, linkage to the GIS creates analytic potentials. For example, Mills and Short (1983) used the *Imperial Gazetteer* for a local study of the distribution of “open” and “closed”

parishes under the Settlement Acts (Holderness 1972). Figure 1 replicates this nationally, phrases such as “the property is considerably subdivided” indicating open parishes, “the property is divided among four” indicating close. It confirms that the industrial north was more “open” and the grain-growing belt between Dorset and Norfolk more “closed”. The largest limitation is that relevant phrases exist in the entries for only a little over half (54%) of all parishes.

Travel writing

Historical travel writers are far less formulaic. We computerized just four texts with lottery funding: William Cobbett’s *Rural Rides*, describing journeys between 1821 and 1826; Daniel Defoe’s *Tour thro’ the whole island of Great Britain*, written in the 1720s; Celia Fiennes’ *Through England on a Side Saddle*, from the late seventeenth century; and Arthur Young’s *Tours in England*, written between 1776 and 1791. However, the collection has been substantially extended with relevant texts computerized elsewhere, now including twenty books written as tours plus our own special collection of six first person accounts written by tramping artisans or political agitators (Southall 1991; Southall 1996). One particularly notable addition is William Camden’s *Britannia*, the first county-by-county survey of Britain and by itself over half a million words.

These texts are continuous narratives and the embedded references to particular places are not necessarily in order of visit, or even to places visited on the

particular journey; for example, James Boswell mentions London in every chapter of his *Tour to the Hebrides* despite the journey being entirely within Scotland.

Given the relatively small number of books, designing a database structure and basic web interface was unproblematic. Information about each book as a whole is held in the same **g_authority** table used by the statistical database and Administrative Unit Ontology to identify sources, but using additional columns going beyond the Dublin Core standard. The text is held essentially as HTML, and we divide each book up into “selections”, usually the chapters of the original printed book. These are held as rows in the **g_text** table, which also holds census reports. Within the web site the “Travel writing” home page lists the books in a grid, with icons that in most cases contain a portrait of the author; the collection of “artisans and agitators” has a separate tab; and a third tab provides simple full text searching. Each book then has a contents page, including a short introduction by us to the author, with links to the pages presenting “selections”:

<http://www.visionofbritain.org.uk/travellers>

We have created the largest online collection anywhere of British historical travel writing, and the interface described so far enables each and every book to be read from start to finish. However, the real challenge was to make descriptions of particular towns or villages quickly accessible. We had already geo-referenced the descriptive gazetteer entries by linking them to the AUO, but this approach could

not be taken with our travellers: when Edwin Russell, a trade union organizer, visited Bromyard in 1872 and described it as “a small old town, which has almost grown out of remembrance” he was not visiting the parish, or the sub-district or the district, but a place which was all of these and none.

The travellers were therefore linked in to the rest of the system via our “places” gazetteer as described below, using placeName tags as defined by the Text Encoding Initiative (TEI; Sperberg-McQueen and Burnard 2002); for example, here is Celia Fiennes’ idiosyncratic verdict on Scotland:

```
It seemes there are very few towns Except
<placeName reg="Edinburgh" cnty="Scotland">Edenborough</placeName>,
<placeName reg="Aberdeen" cnty="Scotland">Abberdeen</placeName>
and Kerk w<sup>ch</sup> Can give better treatement to strangers,
therefore for the most part persons y<sup>t</sup> travell there go from
one Noblemans house to another. Those houses are all Kind of Castles and
they Live great tho' in so nasty a way as all things are in even those
houses one has Little Stomach to Eate or use anything, as I have been
told by some that has travell'd there, and I am sure I mett with a sample
of it enough to discourage my progress farther in Scotland. I attribute
it wholly to their sloth for I see they sitt and do Little.
```

The addition of these tags was done manually, given the many unusual forms of names and the need to avoid marking up the many persons with territorial titles, e.g. “Duke of Liverpool” (Southall 2003). The “reg” attribute is defined by TEI and holds a “regularized” version of the name, so “Edinburgh” rather than “Edenborough”. These names are not necessarily unique in the gazetteer, so we also include a “cnty” attribute, although in this example we define the two major

cities as both being within Scotland as a whole. “Kerk” is a third town we cannot identify.

We load text in essentially this form into the **g_text** table, but into the **raw_text** column. We then run a specially written pre-parser which copies the text into the **g_text** column, taking each `placeName` tag in turn and matching the `reg/cnty` pairs against the **g_place** table. Where it succeeds it replaces the attributes within the tag by two new attributes, so the Fiennes example begins:

```
It seemes there are very few towns Except  
<placeName key="16316" anchor="5">Edenborough</placeName>
```

The “key” attribute is defined by TEI and in our implementation holds the place identifier for Edinburgh, while the “anchor” attribute simply holds a sequence number: this is the fifth place reference that has been inserted within this particular “selection”. For each match, the pre-parser also writes a new row into the **g_text_link** table which is effectively a place-name concordance to the travel writing collection, storing the place identifier, the particular place name that appears and the location within the text, defined by “authority” and “selection” identifiers, and the “anchor” values.

When being presented on the web site, the text is further converted by an on-the-fly parser implemented using open source TagSoup software (<http://ccil.org/~cowan/XML/tagsoup>) which inserts conventional hyperlinks to

the relevant place pages, and also an HTML “name” enabling direct links to this point in the text:

```
It seemes there are very few towns Except  
<a name=pn_5 href='../place/place_page.jsp?p_id=16316'>Edenborough</a> ,
```

The web page also includes a small map of Britain showing the places mentioned in the current selection, which is created by joining the concordance table to the places gazetteer.

These procedures were designed to support analysis as well as presentation. In particular, while nineteenth and twentieth century Britain were subject to repeated statistical surveys, almost the only geographical surveys we have from the eighteenth century are these travel writings; so they provide unique insights into early industrialization. For example, here is Thomas Pennant noting the impact of new markets on the Scottish highlands in 1769:

at the four fairs in the year, held at *Kinmore*, above sixteen hundred pounds worth of yarn is sold out of *Breadalbane* only: which shews the great increase of industry in these parts, for less than forty years ago there was not the lest trade in this article. (Pennant 1800, 105)

Defining “places”

Part two of this paper described how we moved away from a conventional GIS architecture organized around polygons for administrative areas to an ontology of

named entities and relationships. Initially, however, these entities were still all administrative units. “Places” were added at a late stage in our lottery-funded work for two reasons. The impossibility of linking place names within the travel writing collection to specific administrative units has already been noted, but the larger reason was that focus group testing of early versions of the *Vision of Britain* web site showed that users were confused by the large numbers of units associated with many places.

For example, searching for “Newport” returns 51 British units, which include eleven units named after the market town in Shropshire, ten for the industrial city in Monmouthshire and ten for the Isle of Wight’s capital. The Shropshire units include an ancient Parish and Borough; a Registration District and sub-District; Urban and Rural Sanitary Districts, and later Local Government Districts; an Ecclesiastical Parish; a Rural Deanery; and a Constituency.

We therefore defined places around these groupings, naming each place after a “seed unit”, then assigning additional units to the same place based on matching names and either overlapping boundary polygons or explicit relationships. The first set of seed units were all urban Local Government Districts existing in 1911. Then, after associating all other possible units with these, the second set of seed units were all remaining urban Local Government Districts; and the third and largest set were all Civil Parishes existing in 1911, adding the majority of

villages. This was hurried work to support travel writing mark-up and the web site launch, so there we had to rest. Our “places” were a shallow overlay on a system primarily concerned with administrative units. One major limitation was that while every settlement in England of much size had given its name at least to a parish, the same was not true in Scotland. Further, there was no hierarchy of places, only of units, so navigation of the site by users and, as discussed below, by Googlebots worked poorly. Even so, adding places greatly improved usability.

More recently much work has been done to improve the places gazetteer to better integrate the system’s qualitative and quantitative content. One aspect was systematically ensuring that every unit of a given type was linked to a place, manually checking difficult cases; for example, every Ancient Parish listed by Youngs is so linked with one exception, a second Cheshire “Overchurch” supposedly south of Chester, which we and the Cheshire Record Office are agreed is an error by Youngs (*Northern England*, 30). Another was defining additional “places” based on mentions by travel writers or the existence of descriptive gazetteer entries above a certain length. The main table of geographical names has been systematically extended to include place names appearing in gazetteer entries or travel writing.

[Figure 2 appears near here]

So what is a “place”? As shown in figure 2, they exist in a separate database table from administrative units, with just three required values: an ID number, a name and a point coordinate. This matches most commonsensical notions of a gazetteer but differs from formal definitions of digital gazetteers, because our places have no types. The gazetteer content standards developed by the Alexandria Digital Library (2004) and the Open Geospatial Consortium (2006) require that each entry have a feature type, either general like ‘manmade features’ or relatively specific like ‘seaplane bases’.

This approach is very natural if a gazetteer is seen as an alphabetical inventory of items within a GIS, or features on a topographical map. However, a specifically historical gazetteer exists primarily to associate together different instances and variants of the same place-name in textual sources, and over historical time geographical features, especially man-made ones, come and go while names endure, although the precise forms of names tend to evolve. Firstly, English places were often originally named after landscape features such as fords, or clearings in woods; but Oxford has long had a bridge. Secondly, although gazetteer feature type thesauri treat “administrative areas” as a category of feature they exist in law not the landscape. Thirdly, the historian’s concern is less with “features” than with events, such as battles, and the ASDL Thesaurus’s “historical sites” term is deeply problematic. Our “places” are best seen as bundles of

references and figure 2 shows how they link together names taken from administrative units, from descriptive gazetteers and from travel writing; we are working on methods for also harvesting and referencing names from historical maps.

The philosophy behind our approach is further discussed in Southall, Mostern and Berman (2011). While it differs markedly from the approach taken by the Alexandria Digital Library it is arguably closely aligned both with how the Survey of English Place-Names define a place (Watts et al 2004, preface) and with our descriptive gazetteers; for example, the *Imperial Gazetteer* describes Clun in Shropshire as being “a river, a small town, a parish, a sub-district, a district, and a hundred”.

[Figure 3 appears near here]

The detailed implementation of “places” reflects a concern for computational performance and conceptual simplicity; as discussed below, most users of our web site arrive first on a “place page” such as figure 3, so it is important that these appear quickly even when the site is under heavy load, and that it be easy to understand. One source of efficiency is that the “places” table in the database holds all the information needed to create place pages, including the location and a second copy of the text of the most relevant descriptive gazetteer entry.

While the AUO has a separate table of relationships and can consequently record an infinite variety of hierarchies, the places table itself holds a fixed set of relationships each with a specific use within the web site. Each of our detailed “places” is located within a county and a nation, each of these being also defined as a place. Within the “nation” of England, for example, these essentially colloquial “counties” typically have three or four associated county-level units within the AUO of different types, the three different “Cambridgeshires” being discussed in part 2, but the “place counties” generally inherit the Ancient Counties’ boundaries. This simple hierarchy is used to define a geographically hierarchic crumb trail on the web site, and for this purpose poly-hierarchies would be confusing. This, for example, is the crumb trail appearing on our page presenting a population time series for Newport Urban District in Shropshire, both telling a user exactly where they are within the site and, as each element is a hyperlink, enabling them to back out: “Total Population” is the name of the nCube and “Population” is the statistical theme, as discussed in part 1; “Newport UD” takes users to the unit home page; the remaining links take them to the relevant place page or the overall home page:

Home / Britain / England / Shropshire / Newport / Newport UD / Population / Total Population

The place table also holds four other specific relationships. Firstly, each place has a named “container”, mostly identical to the county but, for example, identifying

the Yorkshire Ridings and so providing greater disambiguation when marking-up travel writers. Secondly, we identify the modern local authority containing the place, permitting a direct link to the unit whose redistricted census data provides the clearest overview of long-run trends. Thirdly, a manually-defined “see also place” is used mainly to link very minor settlements to the nearest village for which a substantial amount of text exists. Lastly, a formal hierarchy of “nearby” places has been constructed algorithmically, using data on locations and a single place “population” defined as the maximum total population among all linked units for any dates. The algorithm is constrained to assign the place ID of each higher level place to a maximum of ten lower places, a limit following from SEO considerations as discussed below.

As discussed in part 2, administrative units can be located with greatly varying precision: about half our units have boundary polygons, most of the rest have an inferred point coordinate, but some have no location at all. However, all “places” have a point coordinate and nothing more. These coordinates were originally computed in 2004 as the mean centroid of the seed unit’s boundary polygons but increasingly they are defined manually from where the place name appears on historical maps, and we aim to extend this via crowd-sourcing. The places table identifies the map layer within our historic map server on which the place name appears, so for “bigger places” we display less detailed maps. This approach is

both computationally quick and captures reasonably well an inherently “fuzzy” notion of place: the fuzziness of “Cambridgeshire” has been documented, while we include not so much rivers as river valleys, and mountain ranges not mountains.

One notable consequence of our structure is a novel method for sorting place name search results by likely relevance. Although we could sort places by approximate population, we actually sort them by the number of times the specific name string exists for each place. Essentially this query lies behind searches from the *Vision of Britain* home page:

```
vob=> select p.g_place, p.g_name, p.g_container,  
vob->         count(n.g_name) as freq  
vob-> from   g_place p, g_name n  
vob-> where  p.g_place=n.g_place and n.g_name='NEWPORT'  
vob-> group by p.g_place, p.g_name, p.g_container  
vob-> order by freq desc;
```

g_place	g_name	g_container	freq
630	NEWPORT	SHROPSHIRE	13
1121	NEWPORT	MONMOUTHSHIRE	12
177	NEWPORT	HAMPSHIRE	12
294	NEWPORT PAGNELL	BUCKINGHAMSHIRE	8
6839	NEWPORT	ESSEX	8
13788	WALLINGFEN	EAST RIDING	4
21030	NEWPORT	DEVON	4
8390	NEWPORT	PEMBROKESHIRE	4
17409	NEWPORT ON TAY	FIFE	3
21029	NEWPORT	CORNWALL	3
21031	NEWPORT	SOMERSET	3

26493	NEWPORT	GLOUCESTERSHIRE	2
25079	NEWPORT	NORTH RIDING	2

This has two advantages. Firstly, the total number of attestations of a name in our large corpus of texts, from both administrative units and geographical writing, may be a better guide to a place’s historical importance than a population count. Secondly, this method means we rank a more important place matched on an uncommonly used name below a less important place matched on its most commonly used name. NB in the above example the count is of the name in the **g_name** table, which the query requires to be precisely “NEWPORT”, but the name returned is the single name held for the place in **g_place**; which in the case of Wallingfen is quite different.

Serving a mass audience

An anonymous reviewer of part 2 suggested we should “comment on how much training it will take for off-site people to access [our] HGIS”. As discussed in the introduction, the system was developed to underlie the web site *A Vision of Britain through Time*, targeted primarily at “life-long learners”, which in practice means not students in schools or colleges but users of libraries and archives, and especially those interested in local and family history. This is not an audience who can be “trained” in any conventional sense, and most research independently so we could not rely on teachers or librarians to direct them to our web site: it needed to be both intuitive to use (Krug 2005) and “findable” (Walter 2008).

Making such a complex body of information “intuitive” to access was challenging, but the priority previously given to minimizing the number of underlying database tables helped greatly, leading naturally to our information being presented via a fairly small number of web page types. The largest architectural issue to emerge in initial user testing was the confusing variety of historical units. Grouping them as “places” has already been discussed, but of course the very complex history of British administrative areas is inherently confusing.

Making the site “intuitive” also means meeting user expectations. If we were to work well as a source of local information, our home page had to have a prominent form users could type place names into; in a UK context, this form also needed to understand postcodes, and translate them into coordinates. Simply by not having such a form on their home page, the majority of historical GIS web sites are failing a large potential audience, whereas our unified place-names table comes into its own. Similarly, online interactive mapping needs to offer the same controls as Google Maps for panning and zooming, as those are what a mass audience now expects, and consequently “intuitive”; fortunately OpenLayers provides exactly this.

However, making the site “findable” has proved even more important. Most people today, even academics, find information primarily via internet search

engines; and Google is used for 90% of web searches in the UK and 65% in the US (Kiss 2012).

“Search engine optimization” (SEO) has two sides, one of which has a very bad reputation: adding information to web pages, often concealed from ordinary users, that mislead the software “bots” which index the web for search engines; or using another kind of bot to plant irrelevant links to your site around the web. Search engines will blacklist sites using these techniques, if detected. “Findability”, however, is primarily about enabling bots to accurately index content; and for a site with large amounts of specialized content, this can be very effective.

Unfortunately, conventional GIS-driven web sites are doubly impenetrable to bots, and their place-specific content un-findable. Firstly, bots find web pages primarily by following hyperlinks, and when they encounter any kind of form they stop. This means that most database-driven sites cannot be indexed, including standard gazetteers. Secondly, bots index text and ignore images; and technically a web page with an interactive map implemented in HTML, like Google Maps or OpenLayers, is a large form consisting of graphics not text; embedded interactive maps using Flash or similar technologies are still worse. National Lottery funding came with strict rules on “accessibility”, meaning access by disabled users which in practice meant the blind and partially-sighted. This may seem a vast distraction for an online GIS of any kind, but meeting these

requirements had large benefits for findability: a site that works well with the screen reader software that the blind use instead of conventional browsers will necessarily work well with Googlebots.

Here the merits of a geo-semantic approach are overwhelming relative to the geo-spatial, as our system consists not primarily of a set of nameless polygons but of named entities systematically linked by explicit relationships, each of which is exposed as a hyperlink. The original 2004 web site worked well with Google, as the row of links appearing on all pages included a link to the then-root unit, representing the British Isles, with further links on to all other units and, via them, to pages for statistical nCubes and ultimately to the millions of pages for individual statistical data values. The revised 2009 site works better because unit pages are subsidiary to place pages, and those are organized into a hierarchy using the algorithmically constructed “nearby” relationships, described above: places are limited to ten “nearby” places so that all lower places can be listed at the bottom of each place page without including so many links as to confuse both users and bots. They start down this hierarchy from the Great Britain place page, which is linked to from the main menu bar appearing at the top of every page.

[Figure 4 appears near here]

Figure 4 shows the results of searching google.co.uk for information about each of the 188 Ancient Parishes in the county of Herefordshire, using the search string

“history of <parish name> Herefordshire” and considering only the top-ranked result. Such requests for “local knowledge” will not lead to major commercial sites, and the other results were mainly local sites constructed by amateur historians and parish councils. Wikipedia would perform much better with requests for information about towns, but most villages either have no Wikipedia article or only a minimal stub entry. Herefordshire was chosen because it is the author’s home county and in one sense is atypical: no material from the *Victoria County Histories* is online via the University of London’s British History Online site. For other counties that site also performs well, with text-heavy pages organized into an easily navigable geographical hierarchy, albeit one organized around the historical system of Hundreds.

In October 2012, the Vision of Britain site had 209,735 visits. 14% of these started with the user typing in the address, or more probably following a bookmark; 9% followed a link in another web site, most commonly Wikipedia which contains 6,895 links to Vision of Britain; and 77% arrived via a search engine, with Google by itself supplying 66% of all visitors. Google Analytics provides data on the search strings used. The most common was ‘Vision of Britain’, followed by ‘old maps’, but much more importantly 97,762 different search strings were used, the vast majority containing specific geographical names. This is a classic example of serving the “long tail”: on the web, the largest

audience is often for highly specialized kinds of information which cannot economically be served by traditional publishing methods (Anderson 2006).

Measuring web site usage is problematic. Counts of “hits” are easily manipulated, as each graphic image within a web page is a separate hit. Counts of pages viewed have been made obsolete by AJAX (Asynchronous JavaScript and XML) techniques, by which more content is sent to the user without a new web page being created, and our OpenLayers map viewer uses just this mechanism. Further, off-loading most searching to Google reduces page counts: users of our own home page see that and possibly a list of alternative matches, but most of our visitors arrive directly on a geographically-specific page, most often a “place page”; and as that provides the location and a short description even visits that end after a single page view are not necessarily unproductive. Consequently, numbers of unique users per month are the most commonly quoted usage statistics.

[Figure 5 appears near here]

It is surprisingly hard to obtain usage statistics for historical web sites created by academic projects, although it seems generally agreed that few sites have more than ten thousand unique users monthly. One reason is probably that most such sites are parts of larger university sites, and university IT staff are uninterested in detailed usage. Another requirement of lottery funding was that we report such usage data, but we found that neither of the universities that have hosted the site

had expertise in analyzing the copious but obscure log files generated by the web server. This explains figure 5: the gaps in early years reflect logging failures, ending with a complete shift in 2007 to instead using Google Analytics, which works by our embedding special tags within web pages. This is a free service providing many different views of usage including detailed maps of user locations. Since we switched to Analytics, 5,251,191 unique users have visited the site, 78% from the UK and 8% from the US. Unfortunately it cannot tell us how many were academics, how many in schools, etc.

One reason for adopting a geo-semantic approach was a consensus when we were applying for lottery funding that the computer hardware needed to operate an open access GIS-based web site was unaffordable, unless we had so few users that we clearly failed to meet lottery expectations. Even with large limits on geo-spatial functionality, the site until recently needed substantial dedicated servers: originally a Sun V880, then a Sun T5440 from 2009 to 2012, but currently an eight-core x86 server. The first two involved substantial hosting costs, initially met from development grants and by the British Library; but since 2009 we have had to be self-supporting.

Costs have been met partly by licensing data, primarily vectorised parish boundaries to companies selling information on legal liability for repairs to parish church chancels (National Archives no date; Southall 2013). This arcane legal

obligation is being reformed after 2013, so we have sought to expand income from the site itself, without restricting access. This includes three affiliate relationships with commercial sites, each of which pays us a percentage of any earnings from users we refer to them. Each partner site is historical, and each is geo-referenced so each referral link includes a coordinate: Cassini Publishing offer reproductions of historical maps covering the location; Ancestral Atlas are a specialized social network for genealogists, linking members not by shared ancestors but by ancestors having a shared birthplace; and the Francis Frith Collection have over 120,000 geo-referenced photographs of Britain, taken between c. 1860 and 1970. Frith is perhaps especially interesting, as they have enabled us to add a Historical Photographs page as an additional type of place-specific page, onto which they stream images directly from their servers to our users, who can buy high resolution copies but view medium resolution images for free.

[Figure 6 appears near here]

However, much the largest source of income via the site is Google advertising: we include special static HTML code within our pages which defines areas to contain advertising; Google's systems then sends specific advertisements directly to users to fill these areas, varying both with what Google knows about the particular user and with the place-specific content of our page. Google allow us to block both

specific advertisers and whole categories of advertiser. Income depends on users clicking on the advertisements and Google provide no predictions of likely income, but figure 6 presents our actual experience, showing both that income is substantial relative to hosting costs and that it scales automatically with increased site usage. We have recently added similar advertising to two other historical sites, *Old Maps Online* and *Bomb Sight*, without matching results, so a place-specific site appealing to local and family historians may be a particularly effective advertising vehicle. Income is paid into the bank account we specify, without further administration by us.

Conclusion

The potential for historical GIS to provide a framework for diverse multimedia content has been widely discussed but little developed: online academic resources are overwhelmingly focused on statistics and boundaries, and on interactive mapping derived from them. Meanwhile, online historical map collections have been created mainly by map librarians, mostly without geo-referencing even as a finding aid (Southall and Pridal 2012). Historical writing lives in a third silo, and while the Text Encoding Initiative provides mechanisms for geo-referencing text, as discussed above, they have been little used by that community, even for travel and topographic writing (Southall 2003). Lastly, the most widely used online resource for finding out “what places are like” is almost certainly Wikipedia,

roughly one-third of whose entries include a geographical coordinate, but its historical content is patchy, idiosyncratic and often inadequately referenced, although seldom actually wrong.

The Great Britain Historical GIS and the web site *A Vision of Britain through Time* that accesses it therefore appear to be unique in combining extremely diverse content with a rigorous formal geographical structure and large numbers of users. However, to achieve this we had to abandon packaged GIS software and traditional GIS data models for an approach more geo-semantic than geo-spatial: the previous part justified this through the uncertainties of historical knowledge, steadily increasing as we move further back in time; this final part adds to this the inherent fuzziness of geographical concepts as they appear in texts and discourse, again more easily represented in words than as coordinates. Even traditional maps are better at capturing this fuzziness of “place” than GIS, using a variety of text sizes and fonts when positioning place names.

[Figure 7 appears near here]

The first two parts of this paper emphasized data modeling without discussing usage. This final paper placed much greater emphasis on one particular use, our web site; so is the overall structure useful for anything else? Several answers are possible. Firstly, while the web site permits only those enquiries coded into it, mostly local in focus, at the other extreme is a database command line where

almost anything can be asked; and the relatively small number of database tables used to hold most content gives this great power albeit at the price of a high level of abstraction. Secondly, more conventional download interfaces have been created by the national data services and ourselves, as discussed in part 1. Lastly, while most users want data for a single locality, using the system as a rich gazetteer rather than a GIS, figure 7 shows our statistical mapping at work, zooming in on just a few parishes from the 15,000 or so in the national map. This is a very similar application to Social Explorer (Beveridge et al no date), but note the use of historical mapping as a backdrop, combining quantitative and qualitative.

However, there are two limitations. The first is that while the resource as a whole is pervasively geo-referenced, systematic analysis requires a broader representation of meaning; and while the Data Documentation Initiative has enabled us to create essentially a domain ontology for statistical concepts, our textual content lacks a similarly broad semantic mark-up, so the analysis of “Open” and “Closed” parishes involved scripts containing many separate specific strings identified through trial and error, and there is no easy way of finding all travellers’ descriptions of, for example, early industrial sites. Secondly, there needs to be some way of extracting a wider range of content in an analyzable format. We believe that Linked Data provides a way forward, well suited to the

semantic structures already built, and we have started to explore concepts and build interfaces (Southall et al 2011; Kramer et al 2012).

Acknowledgments

Map scanning was mostly by the British Library and National Library of Scotland. All textual sources described here, unless otherwise noted, were scanned by the Centre for Data Digitisation and Analysis (CDDA) at the Queen's University, Belfast, and converted by them to full editable text using optical character recognition, and much manual checking and correction. We have benefited immensely from assistance from innumerable librarians and archivists who loaned materials for scanning. Other researchers allowed us to use their digital transcriptions including Bruce Gittings (Edinburgh University; Groome's *Ordnance Gazetteer of Scotland*), Derek Rowlinson (LibraryIreland; Lewis's *Topographical Dictionary of Ireland*), Dana Sutton (formerly of UC Irvine; Camden's *Brittania*) and Project Guttenberg (travel writers).

References

- Alexandria Digital Library. 2002. *Feature Type Thesaurus*. Santa Barbara: University of California
(<http://www.alexandria.ucsb.edu/~lhill/FeatureTypes/ver070302>, accessed 16 Dec 2012).
- Alexandria Digital Library. 2004. *Guide to the ADL Gazetteer Content Standard*, version 3.2. Santa Barbara: University of California
(<http://www.alexandria.ucsb.edu/gazetteer/ContentStandard/version3.2/GCS3.2-guide.htm>, accessed 16 Dec 2012).
- Anderson, C. 2006. *The Long Tail: Why the future of business is selling less of more*. New York, NY: Hyperion.
- Beveridge, A.A., Lacevic, A., Weber, S., and Segall, J. No date. *Social Explorer*.
(<http://www.socialexplorer.com>; accessed 16 Dec 2012).
- Big Lottery Fund. 2006. *Digitisation of Learning Materials and Community Grids for Learning: final evaluation findings* (Big Lottery Fund Research, Issue 26). London: Big Lottery Fund (http://www.biglotteryfund.org.uk/research/-/media/Files/Publication%20Documents/er_eval_digi_final.ashx, accessed 16 Dec 2012).

Bodenhamer, D.J., Corrigan, J. and Harris, T., eds. 2010. *The Spatial Humanities: GIS and the future of humanities scholarship*. Indianapolis: Indiana University Press.

Boswell, J. 2004. *The Journal of a Tour to the Hebrides with Samuel Johnson, LL.D.* Oxford, Mississippi: Project Gutenberg.

Cobbett, W. 1932. *Rural Rides*. Letchworth: Temple Press.

Cohen, P. 2010. Digital Keys for Unlocking the Humanities' Riches. *New York Times*, 17th November 2010.

(<http://www.nytimes.com/2010/11/17/arts/17digital.html>, accessed 16 Dec 2012)

Defoe, D. 1927. *A tour thro' the whole island of Great Britain, divided into circuits or journies*. London: JM Dent.

Fiennes, C. 1888. *Through England on a Side Saddle in the Time of William and Mary*. London: Field and Tuer.

Holderness, B.A. 1972. 'Open' and 'Close' Parishes in England in the Eighteenth and Nineteenth Centuries. *Agricultural History Review*, 20 (2), 126-39.

Jessop, M. 2007. The inhibition of geographical information in digital humanities scholarship. *Literary and Linguistic Computing*, 23 (1), 39-50.

Kiss, J. 2012. Who controls the internet? *The Guardian*, 17th October 2012 (<http://www.guardian.co.uk/technology/2012/oct/17/who-rules-internet>, accessed 16 Dec 2012).

Kramer, S., Leahey, A., Southall, H.R., Vampras, J. and Wackerow, J. 2012. *Using RDF to describe and link social science data to related resources on the Web: leveraging the Data Documentation Initiative (DDI) model*. Working Paper. Ann Arbor: Data Documentation Initiative. (<http://dx.doi.org/10.3886/DDISemanticWeb01>, accessed 16 Dec 2012)

Krug, S. 2005. *Don't Make Me Think!: A common sense approach to web usability*. Second edition. Berkeley: New Riders.

Mills, D.R. and Short, B.M. 1983. Social change and social conflict in nineteenth-century England: The use of the open-closed village model. *Journal of Peasant Studies*, 10 (4), 253-62.

Maron, M.L., Kirby Smith, K., and Loy, M. 2009. *Sustaining Digital Resources: An on-the-ground view of projects today*. Joint Information Systems Committee. (http://sca.jiscinvolve.org/files/2009/11/sca_ithaka_sustainingdigitalresources_fullreport_with-casestudies_uk.pdf, accessed 16 Dec 2012)

National Archives. No date. *Chancel repair liabilities in England and Wales*. Legal Records Information Leaflet 33. Kew: The National Archives

(<http://www.nationalarchives.gov.uk/documents/research-guides/chancel-repairs.pdf>, accessed 16 Dec 2012)

Open Geospatial Consortium. 2006. *Gazetteer Service - application profile of the web feature service implementation specification*. Wayland, Massachusetts: Open Geospatial Consortium. (http://portal.opengeospatial.org/files/?artifact_id=15529, accessed 16 Dec 2012)

Pennant, T. 1800. *A Tour in Scotland*. Fourth edition. London: Benjamin White.

Schreiber, S., Siemens, R., and Unsworth, J., eds. 2004. *A Companion to Digital Humanities*. Malden, MA: Blackwell

Southall, H.R. 1991. Mobility, the Artisan Community, and Popular Politics in early nineteenth century England. In *Urbanising Britain: class and community in the nineteenth century*, edited by G. Kearns and C.W. Withers, 103-20. Cambridge: Cambridge University Press.

Southall, H.R. 1996. Agitate! Agitate! Organize! Political travellers and the construction of a national politics, 1839-1880. *Transactions of the Institute of British Geographers*, N.S. 21, 177-193.

Southall, H.R. 2003. Defining and identifying the roles of geographic references within text: examples from the Great Britain historical GIS project. In *HLT-*

NAACL 2003 Workshop: Analysis of Geographic References, 69-78. Edmonton:

Association for Computational Linguistics.

Southall, H.R. 2011. Rebuilding the Great Britain Historical GIS, part 1: building an indefinitely scalable statistical database. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 44 (3). 149-159.

Southall, H.R. 2012. Rebuilding the Great Britain Historical GIS, part 2: a geo-spatial ontology of administrative units. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 45 (3). 119-134.

Southall, H.R. 2013, in press. Applying historical GIS beyond the academy: Four use cases for the Great Britain HGIS. In *Rethinking space and place: New directions with historical GIS*, edited by A. Geddes and I.N. Gregory.

Indianapolis: Indiana University Press.

Southall, H.R., Mostern, R., and Berman, M. 2011. On historical gazetteers. *International Journal of Humanities and Arts Computing*, 5 (2), 127-45.

Southall, H.R., and Pridal, P. 2012. Old maps online: enabling global access to historical mapping. *e-Perimtron*, 7 (2), 73-81.

Sperberg-McQueen, C.M. and Burnard, L., eds. 2002. *TEI P4: Guidelines for Electronic Text Encoding and Interchange*. XML Version. Oxford, Providence, Charlottesville, Bergen: Text Encoding Initiative Consortium.

Staffordshire County Council. 2003. *Staffordshire Past Track* (<http://www.staffspasttrack.org.uk>, accessed 16 Dec 2012).

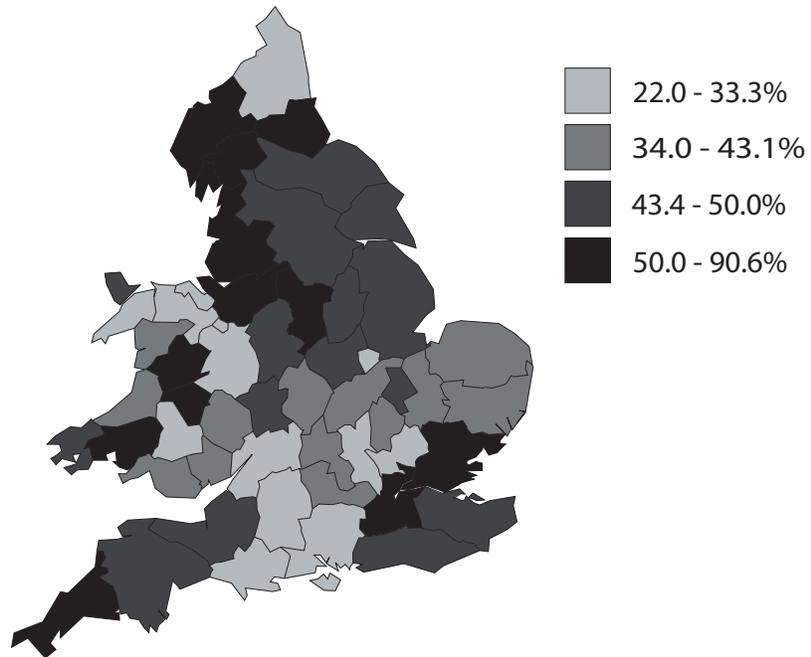
Stamp, L.D. 1948. *The Land of Britain: Its Use and Misuse*. London: Longmans.

Walter, A. 2008. *Building Findable Websites: Web standards, SEO, and beyond*. Berkeley: New Riders.

Watts, V., Insley, J. and Gelling, M., eds. 2004. *The Cambridge Dictionary of English Place-names*. Cambridge: Cambridge University Press.

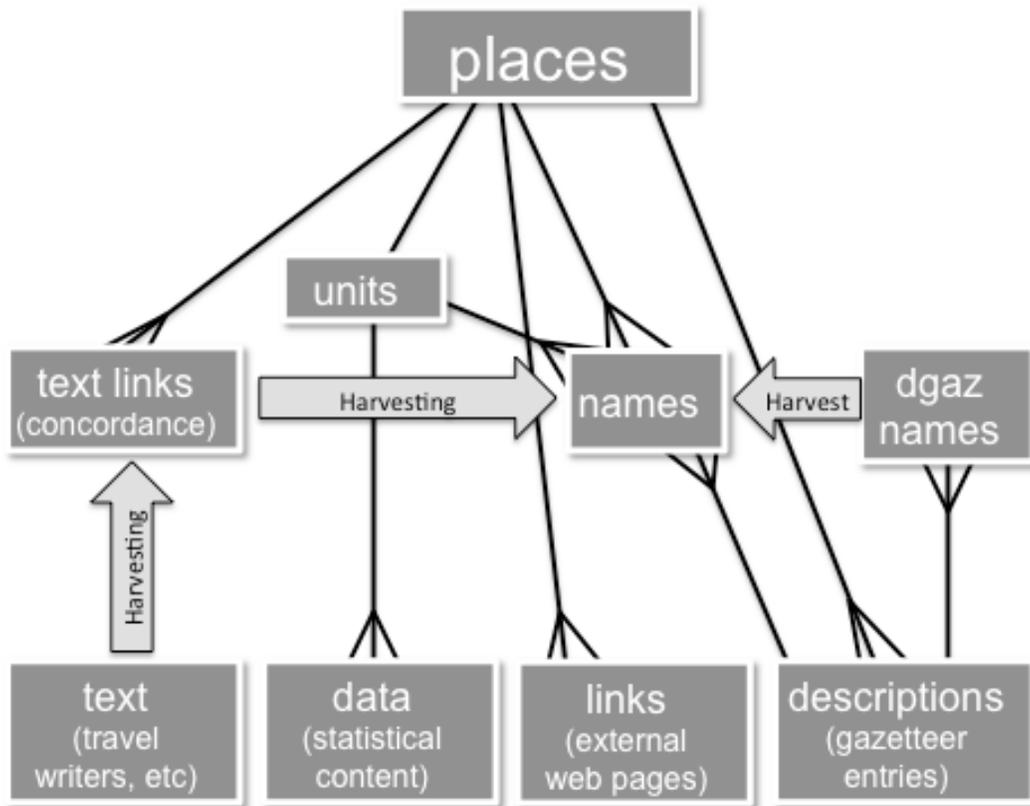
Young, A. 1932. *Tours in England and Wales, selected from the Annals of Agriculture*. London: London School of Economics.

Figure 1: Percentage of parishes whose property was in “many hands”:



Source: *Imperial Gazetteer of England and Wales* (1872-4)

Figure 2: Integrating “place” information with the AUO



**Figure 3: “Place page” for Greenwich from
A Vision of Britain through Time**

A vision of BRITAIN
THROUGH TIME
A vision of Britain between 1801 and 2001.
Including maps, statistical trends and historical descriptions.
Help using this website

Home Places Statistical atlas Historical maps Census reports Travel writing Learning zone Data access

Home / Britain / England / Kent / Greenwich

Place:
Greenwich Kent

Location

Historical places and writing

Historical photographs

Units and statistics

Related websites

Place names

AdChoices ▾
1:1250 Scale Maps
Buy Planning 1:1250 Maps Today View And Print Planning Maps Now!



Click on the map for other historical maps of this place

In 1887, John Bartholomew's *Gazetteer of the British Isles* described **Greenwich** like this:

“ **Greenwich**, parl. bor. and par., Kent, on river Thames, 4 miles SE. of London Bridge by rail -- par., 1741 ac., pop. 46,580; parl. bor. (which includes the pars. of Greenwich, Deptford St Nicholas, Charlton next Woolwich, and Kidbrooke), 3838 ac., pop. 65,411; 1 Bank. Market-days, Wednesday and Saturday. Greenwich possesses several important mfrs. including telegraph works, engineering works, chemical works, &c. The Royal Observatory has a world-wide celebrity; English geographers calculate longitude from its meridian. ...

Figure 4: Source of first ranked results from searching google.co.uk for “history of <name>“ for all Herefordshire ancient parishes

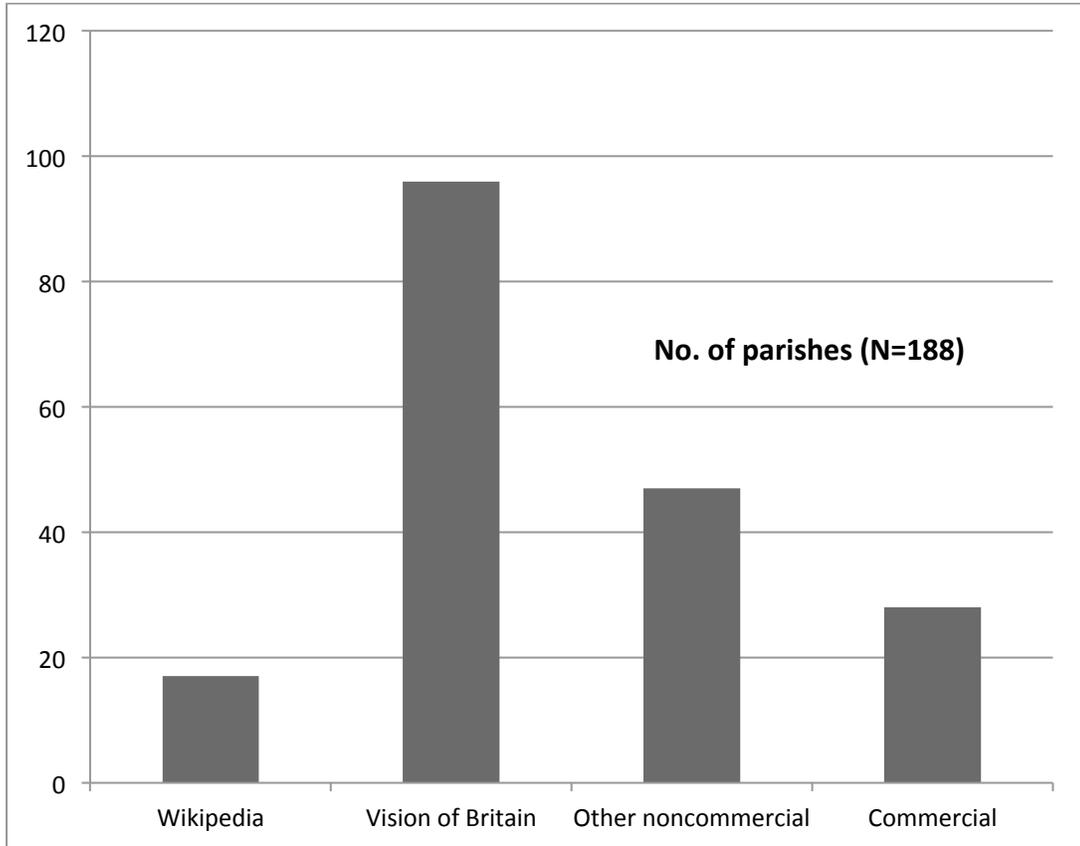


Figure 5: Vision of Britain Unique Users per Month, 2004-12

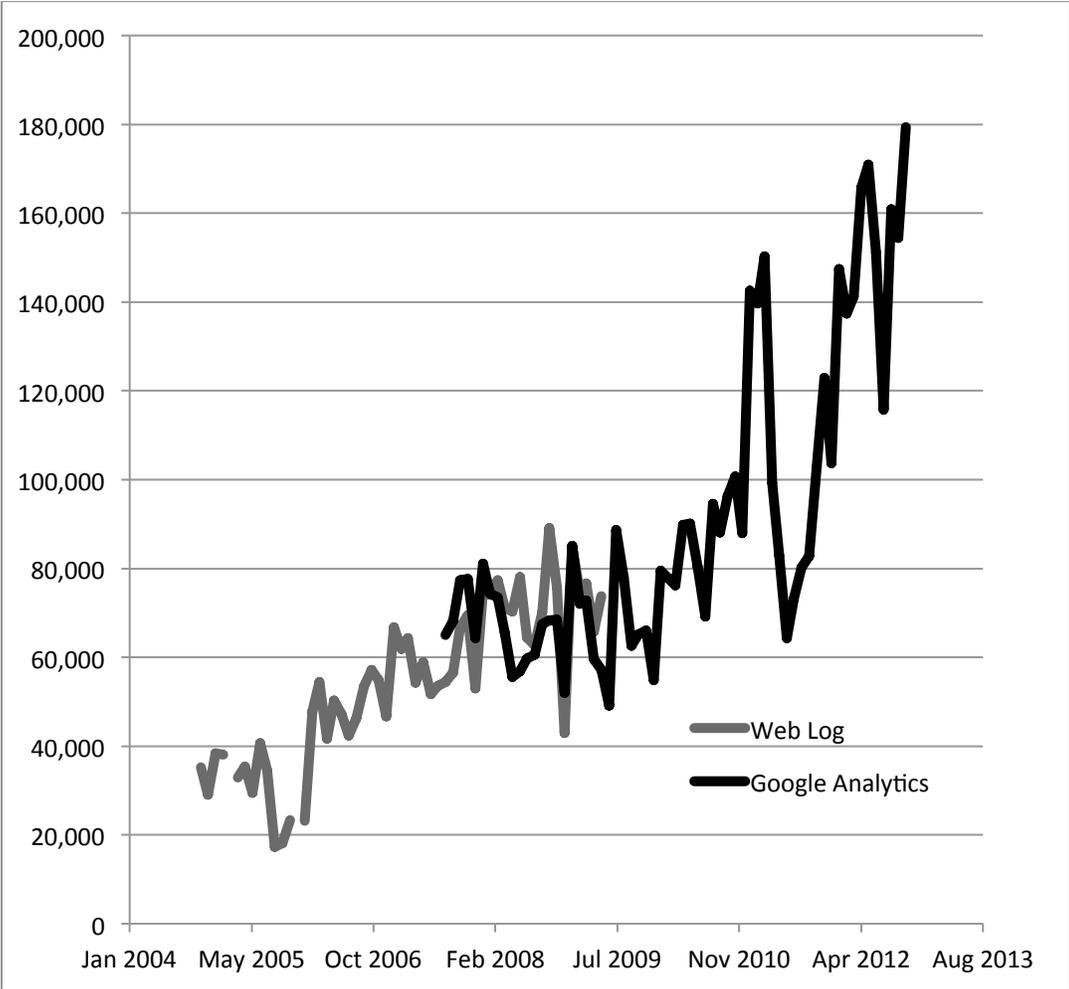


Figure 6: Relationship between usage and AdSense income, 2009-12

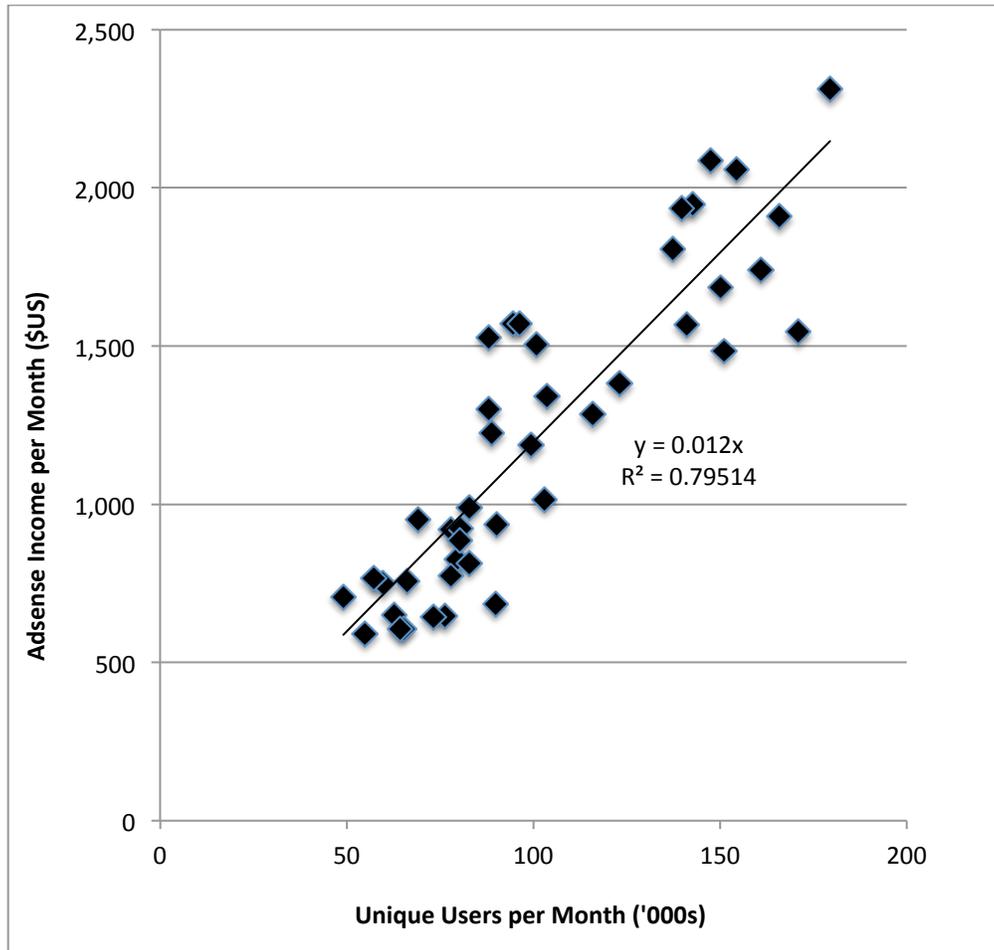
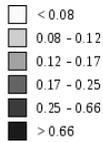


Figure 7: Parish-level Population Density in 1911 for the Portsmouth and Southampton area, presented within *A Vision of Britain through Time* using OpenLayers software and overlaid on GSGS mapping from the WMS

Population Density (Persons per Acre)



Our figures for the area of units as measured in acres are always numbers that were listed in census reports, while all areas measured in hectares have been calculated by us from our boundary maps. A small amount of the variation in persons per acre therefore results from changes in how accurately the country had been surveyed. However, the main reason is, of course, population growth and sometimes decline. Spectacular changes in density from one census to the next are usually the result of major boundary changes.

Rate definition

Options (reloads page)

Available rates:

Unit type:

Date:

