Locating Past Places in Britain: Creating and evaluating the GB1900 Gazetteer

Authors: Paula Aucott and Humphrey Southall

Abstract

The GB1900 project used crowd-sourcing to transcribe all text from the second edition County Series six inch to one mile maps of Great Britain, published between 1888 and 1914, a total of c. 2.55m. geo-located text strings. These locate almost every farm and about half of all street names. The paper describes the final datasets, and how they were created. It then presents a detailed comparison with five other freely-available gazetteers of Britain: Geonames, the US government's NGA gazetteer, the Ordnance Survey's 50k and Open Names datasets, and the English Place Name Survey's DEEP project. Comparisons are presented at national level and, more qualitatively, for an area of eastern England. The results demonstrate both GB1900's greater volume of geo-located entries and its ability to locate places and features identified in other historical sources beyond administrative hierarchies: this is the most detailed historical gazetteer, certainly for Britain and possibly for anywhere. The final online system is described, including its integration of place name histories from DEEP.

Keywords: GB1900, locations, local history, gazetteer, place names

Introduction

Human activity is contingent on both time and place, so in humanities research we constantly need to know not only when events occurred but where. Historical documents are overwhelmingly text based and what makes them geographical is the place names (toponyms) within them. There is also a large need from family historians to locate ancestral origins. Both types of enquiry require gazetteers, turning place names into locations which can be found on maps, and we often need to know not only the current names of places but also past forms. The creation of these historical gazetteers is a central task of the geohumanities (Southall, Mostern, and Berman 2011; Berman, Mostern, and Southall 2016).

This paper presents the new gazetteer created by the GB1900 project, and compares it with other freely available gazetteers of Britain. GB1900 applied crowd-sourcing to transcribe all text strings appearing on six inch to the mile (1:10,560) maps of Great Britain, a total of c. 2.55m. strings, each with a coordinate, although these include many labels which are not place names. Detailed accounts of the project's origins and the crowd-sourcing software (Southall et al. 2017), and of the work of the online volunteers and their motivations (Aucott, Southall, and Ekinsmyth 2019) have already been published. The present paper focuses on the GB1900 place name gazetteer, while a further paper will present analyses of the non-place name data as evidence of past physical and cultural landscapes.

Our evaluation of the GB1900 gazetteer is through comparison with five other gazetteers of Britain, and this paper consequently also includes the first detailed investigation of the dataset created by the DEEP (Digital Exposure of English Place names) project. The initial section summarises the history of detailed mapping of

Britain and the recording of place names, leading up to the GB1900 project. We then detail the further work to turn the transcription outputs into final datasets. The third section presents a mainly quantitative comparison of the ‘Abridged‘ GB1900 data set, focused on place names, with the other gazetteers of Britain. The next section presents a more qualitative assessment, by focusing on a smaller area of eastern England. The fifth section describes our online version of the gazetteer, including how place name histories from DEEP have been integrated into the database. A concluding discussion argues the advantages of gazetteers rooted in specific historical sources.

Creating the historical record of British place names

The UK national mapping agency, the Ordnance Survey (OS), began mapping Britain in 1791 at one inch to one mile (1:63,360) scale, but this proved inadequate for railway construction and so the original County Series six inches to one mile (1:10,650) maps were published between 1842 and 1882. This was the largest scale at which the OS ever published paper maps covering all of Great Britain, the next scale, 25 inches to one mile (1:2,500) being limited to settled and farmed areas. However, the initial County Series used a different prime meridian, effectively a different map projection, for each county (National Library of Scotland n.d.). Creating the most detailed possible digital seamless historic map of Britain therefore requires the second edition of the County Series, published between 1888 and 1914, as exemplified in Figure 1.

The earliest place name inventories were itineraries not gazetteers: lists of places along routes, rather than systematic coverages of areas. Camden’s *Britannia* (1586) was structurally still an itinerary, particularly following rivers, but systematically covering every county, and enumerating current and past toponyms. Camden based this on having ‘conferred with most skillfull observers in each county, I have studiously read over our owne countrie writers, old and new…. I have had conference with learned men in other parts of Christendome‘ (Camden 1610, sec. Introduction). Immediately following Camden’s survey, John Speed (1552-1629) published a set of maps showing each of Britain’s counties, at scales enabling essentially every village to be shown, and that in turn led to John Adams’ *Index Villaris* of 1680, essentially a single alphabetical listing of c. 24,000 locations, each including a latitude and longitude (Adams 1680).

The Ordnance Survey followed in this tradition, its surveyors being instructed to systematically gather the names of features and places from local people (Harley 1971). The maps themselves are therefore the most authoritative record of Britain’s places and their names, although it should be noted that map-makers have generally copied names from earlier editions (Crone, Campbell, and Skelton 1962), so most but not all the names on the second edition County Series will come from the first.

The English Place Name Survey (EPNS) began in 1923 (see https://www.nottingham.ac.uk/research/groups/epns/survey.aspx). It creates place name histories, tracing individual names back through ever-earlier documents to forms which have meaning in Anglo-Saxon, Norse or Celtic languages, which can therefore be used as evidence of early landscapes and settlement patterns. This meticulous research requires expertise in the above languages, so the EPNS works through county teams and is still incomplete.

The EPNS began each county survey by systematically gathering names from the six inch maps (Smith 1954), but did not systematically include coordinates. The desire for an equally detailed gazetteer of Welsh place names led to the Cymru1900Wales project, led by the National Library of Wales and the Royal Commission on Ancient and Historical Monuments of Wales (Ell, Hughes, and Southall 2016), applying crowd-sourcing to transcribe all text appearing on the second edition County Series maps.

Cymru1900 launched in October 2013, then essentially relaunched as GB1900 in September 2016 as a collaboration between the original Welsh partners, the National Library of Scotland who provided digital mapping covering all of Britain not just Wales, and the University of Portsmouth who provided software revisions and hosting. GB1900 inherited software and all existing volunteers and transcriptions from Cymru1900 but, obviously, extended the geographical scope.

Volunteers were asked to transcribe all text within the maps except purely numeric data, such as ’spot heights’ giving elevations, and distances. Rather than attempting to identify the locations of features being named, volunteers were instructed to give each text string the location of the bottom left corner of the first character of the string. No attempt was made to capture variations in fonts and text sizes. The system also allowed volunteers to separately record additional names for features from personal knowledge, or add personal ‘memories‘, but these capabilities were little used, and obviously could not be confirmed by other volunteers.

The GB1900 Gazetteer

The crowd-sourcing process ended January 2018, by when it was almost impossible to find new text to transcribe or existing transcriptions needing confirmation. By then 2,656,830 initial transcriptions had been made, plus 2,618,533 confirmatory transcriptions; some of the former were mistakes incapable of confirmation, more than balancing those needing multiple confirmations to reach consensus. These data were held within the MongoDB system underpinning the web site, accessible from the GBH GIS’s main Postgres database via a foreign data wrapper.

Three final datasets have been created and made available for download. Firstly, the ‘final raw dump‘, a zip archive containing the final state of all tables from MongoDB, except that detailing individual volunteers. Secondly, the ‘Complete GB1900 gazetteer‘, a single listing of points and coordinates. Thirdly, the ‘Abridged GB1900 gazetteer‘, containing the same columns as the complete dataset but with common non-place names removed. The raw dump is made available under the simplest Creative Commons license (CC0), enabling anyone to use it as they please, the only limitation being that they may not call the result the GB1900 gazetteer. The other two datasets are under Creative Commons – Attribution – Share alike (CC-BY-SA) licenses. All can be downloaded from:

http://www.pastplace.org/data#gb1900

The Raw Dump is precisely the content created by the main transcription process and contains four data files plus four additional files of documentation. The core data file, gb1900_locations, has 2,666,341 rows, one for each location or ‘pin‘ created in the GB1900 system. Each row includes a WGS84 coordinate, the unique 24-character hexadecimal pin ID used internally by the crowd-sourcing software, and a new and simpler seven-digit ID number, based on the order in which pins were created. The gb1900_transcriptions file contains 8,043,297 text strings, each linked to a pin: this count is misleading, as one of the changes made in evolving from Cymru1900 to GB1900 was to replace a seldom-met requirement for three independent transcriptions for each pin, with no immediate checking of whether they matched, with two being sufficient provided they matched; but this had to be implemented by programmatically inserting a third matching string. The other two tables in the Raw Dump are much smaller, but are the only way in which the ‘user contributions‘ have been preserved: gb1900_memories has 399 rows, and gb1900_alt_names has 1,970 rows.

Creating the Complete Gazetteer from the Raw Dump files began with automated cleaning within the database. That worked by creating a new data set in which all three transcriptions were added to the locations as separate columns, then compared. Mostly they were identical, so a single canonical text string was easily identified. Elsewhere, small differences were deemed insignificant. This included ignoring double spaces and spaces at the beginning or end of a text string, and standardising common abbreviations for generic features. For example, any string in which the only two letters were an ‘F’ followed by a ‘P’, regardless of case, spaces or punctuation marks, were standardised to ‘F. P.’, the abbreviation for footpath, even though this occasionally reflected variations in the original maps rather than in volunteers’ interpretations.

At this stage we also excluded any purely numeric data which should not have been transcribed, and some duplicate pins, where two closely adjacent pins were separate attempts to transcribe the same string: pairs of pins were merged into one where they contained exactly the same text string as a neighbouring confirmed pin and they were located within 10 metres of one another. The process was repeated at 20, 40 and 60 metres, but with progressively more checking that they were not common abbreviations likely to appear within close proximity, such as ‘P’ for pump.

27,400 locations remained needing manual checking. Firstly, those still with only one transcription, which by this stage probably meant there was no actual feature on the map. One source of these was that in the County Edition areas falling across county boundaries were covered by sheets from the sets for both counties, and the mosaics used for Cymru1900 and GB1900 sometimes differed in the version used. This meant that some unconfirmed transcriptions inherited from Cymru1900 could not be confirmed in GB1900, but manual checking was required to delete them. Secondly, all cases were manually checked if only two non-matching transcriptions had been made, or there were three transcriptions which all differed. We also manually checked all locations where the agreed version did not contain an accented character, but a third transcription did, as many volunteers ignored the accents when transcribing Welsh and Gaelic names, even though buttons for adding these special characters were provided on the transcription form.

All these cases were extracted into spreadsheets and emailed out to volunteers who had offered to further assist. Each was checked by at least two volunteers against the National Library of Scotland web site presenting the original mapping, and the results compared; any remaining discrepancies were resolved by the lead author.

One further common error was breaking up long labels describing railway lines into multiple strings, as illustrated in Figure 2: ‘CAMBRIAN RAILWAY’ appears above the line and ‘KERRY BRANCH’ below. In the worst cases each word of the railway label had been separately transcribed. The authors corrected these c. 850 railways labels.

The third dataset, the ‘Abridged GB1900 Gazetteer‘, the focus of the remainder of this article, contains the same columns as the Complete version, but with most non-place names removed. This was done by ranking all unique strings in the Complete gazetteer in descending order of frequency and then working down manually. The five commonest strings are ‘F. P.’ (meaning Foot Path; 306,583 occurrences), ‘W’ (Well; 190,979), ‘P’ (Pump; 115,877), ‘F. B.’ (Foot Bridge; 74,514) and ‘Spring’ (46,876), so just removing these cuts over a quarter of all rows from the Complete data set. The commonest strings still included are ‘Manor House’ (1,617 occurrences) and ‘Manor Farm’ (1,496). All other strings retained appear less than a thousand times, and currently all strings appearing at least 25 times have been considered for exclusion. Street names are retained, however common, so there are 454 ‘High Street’ entries, while church names including saints’ dedications are excluded. Some categories of unique strings were also removed, such as most containing ‘found’, for instance ‘Human Remains Roman Coins &c. found here A. D. 1886’. The end result is that the abridged dataset contains 1,097,123 rows out of the complete dataset’s 2,552,459 (43.0 per cent).

Both datasets include, in addition to the agreed transcription and both the original and simplified unique identifiers, the location given both as latitude and longitude (WGS84) and as Ordnance Survey National Grid coordinates, and the name of the nation (England, Wales or Scotland), modern local authority and modern Civil Parish containing the location. These names were added by point-in-polygon database look-ups from the coordinates, and the parish boundaries were those provided for download by the UK Data Service in June 2018, representing English and Welsh parishes as defined for the 1991 census and Scottish parishes as defined in 2001. The downloads for the Complete and Abridged datasets both include, in addition to the data themselves as CSV files, the Creative Commons license and a Read Me file.

Based on our repeated visual inspections of both the source maps and the text files created by GB1900, we believe the latter are a comprehensive and accurate transcription of the text in the former, the main limitations being that distinctions between upper and lower case letters cannot be relied on, and the coordinates are sometimes imprecise.

Alternative place name gazetteers for Great Britain

How useful is the GB1900 gazetteer? This is best answered through comparison with other gazetteers, specifically five existing freely downloadable gazetteers under open licenses, as listed in Table 1.

The DEEP gazetteer

The Digital Exposure of English Place-Names (DEEP) project was funded by Jisc in 2011-13 to computerise all completed volumes of the EPNS, discussed above, excluding county introductions and volumes on single cities. Digitization combined optical character recognition with much manual work. DEEP created the online 'Historical Gazetteer of England's Place-Names'' (http://www.placenames.org.uk/), but this went offline in 2018. Ell, Hughes and Southall (2016) described the project but little detailed documentation was ever published. However, the underlying data are available through Jisc at this site, under a Creative Commons non-commercial attribution 4.0 licence:

http://mads.digitalresources.jisc.ac.uk/mads2017/

A total of 66 files are available for download, each corresponding to a particular EPNS volume. Together they comprise 9,812,355 lines of XML, based on the Library of Congress's Metadata Authority Description Schema (MADS:

http://www.loc.gov/standards/mads/) format but with no more specific documentation. What follows is based on our constructing an actual gazetteer from these files, which may or may not differ significantly from that which was behind placenames.org.uk.

Each XML file consists entirely of a series of entries for different places, demarcated by the <mads> tag, and in total there are 539,372 such entries. Entries can include one or more instances of various optional elements, so our main 'deep_places' table has four child tables. Firstly, 'deep_locations' (53,655 rows) holds geographic coordinates from four different sources. Secondly, 'deep_names' (820,556 rows) holds a primary name and any number of variant 'names' for each place. Thirdly, 'deep_attestations' (380,051 rows) details the historical sources from which names are drawn, and a given name can be attested to by multiple sources. Finally, 'names' are often lengthy lists of toponyms, so the data also include 391,177 'search terms' in the 'deep_searchterms' table, identifying individual names from within those lists for use in searching.

Each entry includes a place ID such as 'epns-deep-86-a-parish-000077', and these in practice identify the type of feature, here a parish. The MADS file for every volume begins with an entry for the county covered, and every other entry then includes a 'related entry' creating a hierarchy which always links back ultimately to the county, generally by way of a parish. Table 2 lists the overall frequency of different 'place types', showing that 70 per cent are the names of individual fields within farms, although none of these field names have either locations or supporting attestations, and they are listed for only sixteen counties. 'Mapped names' will generally have been transcribed from maps and especially early six inch maps, so although most lack locations, most can be assigned to a parish with known boundaries and then matched to locations in the GB1900 data, as described below. ‘Sub-Parish’ covers villages and hamlets, generally including the settlements parishes are named after so these names appear twice, while ‘Sub-County’ refers to the ancient system of districts, including Hundreds and Wapentakes.

An obvious limitation of the DEEP data is that they are based on an incomplete survey. They inevitably do not cover Wales or Scotland, but Figure 3 shows that eight counties (Cornwall, Hampshire, Herefordshire, Kent, Lancashire, Northumberland, Somerset and Suffolk) are also completely absent, and another six counties are incomplete (Durham, Leicestershire, Lincolnshire, Norfolk, Shropshire, and Staffordshire).

Coordinates are given in these digital files from four sources, The English Place Name Society themselves (EPNS), Geonames as described below, Unlock the Edina service which offered geo-referencing of place names and geographic data searching and ran until 2016 (https://edina.ac.uk/unlock), and the Key to English Place Names resource (KEPN), an AHRC funded project in 2004-5 making available place name elements and their meanings through an online searchable database of English place names (http://kepn.nottingham.ac.uk/). As Table 2 shows, parishes and sub-parish settlements generally have coordinates but most ‘mapped names’ and all fields lack them.

Geonames and the NGA gazetteer

Perhaps the best known global gazetteer of modern place names is Geonames (http://www.geonames.org/). Geonames has been assembled from many different sources, and then extended through crowd-sourcing: anyone can add entries. This makes it very large, but means that data quality may vary: Ahlers (2013) analysed

Geonames data for central America, finding mis-allocation of Feature Codes, duplication of features and significant variation in locational accuracy, depending on the original data source.

Geonames has at its core two US government datasets, the US Geological Survey's Geographic Names Information System (GNIS), covering the United States (https://geonames.usgs.gov/), and the National Geospatial-Intelligence Agency's Geographic Names Database (NGA Gazetteer), covering everywhere else (http://geonames.nga.mil/gns/html/index.html). Although the NGA Gazetteer provides only half as many entries for the UK as Geonames, it is separately included in our comparison as potentially more accurate and consistent.

Table 3 compares Geonames with the NGA Gazetteer. Geonames does not identify sources for individual entries, but two factors almost entirely explain why Geonames is almost twice as large as the NGA Gazetteer. Firstly, Geonames includes many more administrative areas, overwhelmingly parishes, Britain's most detailed administrative geography. They are identified not as parishes but as type 'ADM3' (3,870 features, also including Districts) if they are within Unitary Authorities which are contained within England, Wales or Scotland, or as 'ADM4' (7,730 features) if they are contained within Districts within Counties within those nations. Parishes are generally centred on and named after villages, separately identified as 'PPL' (Populated Places). In practice this means that a very large number of feature name/location pairs are duplicated. This is also true of GB1900, as both villages and parishes are named on the six inch maps, and of DEEP.

Secondly, Geonames include seven times as many 'Spot' entries as NGA. This may be where crowd-sourcing has the largest impact, and it is particularly notable that

8,275 (39 per cent) of the Feature Class ‘Spot’ locations are classed as ‘Hotels’. They also include 2,788 railway stations, versus 858 in the NGA Gazetteer. Conversely, numbers of ‘Populated Places’ and physical landscape features are very similar. Geonames does identify 2,236 Castles (Feature Code ‘CSTL’), but only 304 churches (Feature Code ‘CH’), which are arguably much the most common type of historic building as there is at least one in almost every village.

Ordnance Survey 50K and Open Names gazetteers

The Ordnance Survey (OS) have made two large gazetteers covering Great Britain freely downloadable. In 2010, they made available a gazetteer based on names appearing on their 1:50,000 map series, sometimes called the 50K Gazetteer, but in October 2017 they announced that this was being replaced by OS Open Names. The present situation is somewhat curious, as online re-sellers offer the 50K Gazetteer for £220, but a Linked Data version remains available from the OS under the UK Open Government License, which was designed to be compatible with Creative Commons. What follows uses a less verbose version of the 50K Gazetteer downloaded in 2010.

Table 4 is based on the ‘Feature Codes’ in the 50K gazetteer, and shows that the majority of features are effectively un-typed. It is unclear why two different Feature Codes (‘X’ and ‘O’) are needed to indicate this, but the lack of typing must be related to the data having been harvested from digital topographic mapping in which most names are labels for areas not symbols. The other major limitation is that coordinates are accurate only to 1 kilometre.

The ‘replacement’ Open Names gazetteer initially appears far larger, but as Table 5 shows the large majority of the entries are postcodes, the UK equivalent of zip codes, or street names. Although it lacks the farms in the 50K gazetteer, the number of settlements and landscape features is still very substantial, and Table 6 provides a more detailed breakdown of the ‘Local Types’ within the Type ‘PopulatedPlace’. This contains twice as many such features as Geonames, partly because it includes suburban areas within towns. 30.9 per cent of Open Names settlements are linked to corresponding Geonames entries, and 35.7 per cent to dbpedia, but only 21.4 per cent to both.

While GB1900 includes no feature classification, some features can be grouped based on their names. For example, 109,193 GB1900 entries end with a space followed by ‘Road’, ‘Street’, ‘Lane’, ‘Rd.’ or ‘St.’, far more than any other gazetteer analysed. While GB1900 has far fewer hotels than Geonames (1,290 in both GB1900 datasets), it includes many more churches (17,795 in the Complete gazetteer) while 66,151 entries end in ‘farm’, compared to only 100 farms in Geonames. OS Open Names identifies 3,248 railway stations, and 133 railway labels, while 4,078 GB1900 Complete entries end in ‘Railway’ or ‘Ry.’, and 5,611 end in ‘station’ or ‘Sta.’, although these include coast guard stations, police stations and so on.

Assessment of the accuracy of local area names

The three Norfolk hundreds of Holt, North Erpingham and South Erpingham were chosen for a more qualitative assessment because they are covered by a relatively recent EPNS volume (Sandred 2002) and because the lead author is familiar with the area. The hundred boundaries used for this procedure were created by merging constituent parishes from the Great Britain Historical GIS project, although Field Dalling and Horstead with Stanninghall parishes needed to be added to include all

DEEP entries. Data from the other five gazetteers were then included if their locations fell within the boundary polygons for the three hundreds, as modified.

GB1900 has 6,357 points located within the study area. Classifying them from their names, 302 are named as farms, 231 as some kind of house, 35 as lodgings and hotels and 72 as public or brew houses. Two are Urban Districts, Cromer and Sheringham, and 98 are parish names. Although parish names should be easily identifiable, as they are always printed in upper case, in practice only about half those within the study area had letter case correctly transcribed. Figure 4 shows the distribution of all GB1900 points relating to administrative unit names and identifiable building locations, including ruins or the sites of former buildings (farms, houses, lodgings, churches and chapels). The map excludes all other physical features whether natural or man-made. Even displaying only these points the coverage is dense.

6,763 DEEP 'places' were within the study area but only 210 (3 per cent) had coordinates. In most cases, these were the parishes and the identically-named main settlement within each parish, classed as 'Sub-Parish'. The ten other places with coordinates were 'Mapped Names', including six lost settlements.[1] The majority of other entries are identified only as being somewhere within a parish, reflecting the contents of the EPNS volume. For example, 284 places are identified as within Aldborough parish, but all the modern (27) and medieval (1) field names and the minor places (14) such as buildings, copses and lanes have no associated co-ordinate, and therefore cannot be mapped.

All other gazetteers include coordinates for every record. Geonames contains fewer points than DEEP. The majority of records are classified as ADM3 (1, North Norfolk

District) or ADM4 (78) plus settlements given a ‘Populated Places’ Feature Code (67). The remaining 36 points consist of natural features (4), buildings; railway stations (8), hospitals (4), large houses (2) and castles (7), hotels (2), air fields (4) and miscellaneous other features like a park and a pier (5). The NGA Gazetteer includes even fewer features, almost all of which are ‘Populated Places’ (64) plus just a few ‘Spot’ locations and one physical feature.

The OS 50K Gazetteer also includes fewer features, with limited and more confused Feature Coding. Only four entries have the ‘Towns’ Feature Code: Cromer, Sheringham, Holt and Aylsham. Most other features are classified under ‘O’ (Other - 183) or ‘X’ (All other features - 132). The majority of ‘O’ entries match parish names, although they include eight road names, while ‘X’ includes 46 hill names not listed under Feature Code ‘H’ (Hill or mountain - 1).

In contrast the OS Open Names data include 5,503 entries for the study area, but 46.8 per cent are postcodes. The remaining 47 ‘Other’ Type points are miscellaneous properties including two producing electricity and the rest are educational and medical facilities. Of those identified under Type ‘TransportNetwork’ just four are not road numbers (17), road names (1,858) or railway related (15). Only 126 points (2.3 per cent) identify the names of settlements which can be divided into Local Types; towns (4), suburban areas (6), villages (83), hamlets (29) and other settlements (4). 862 entries are related to landscape and water features, a far greater number than those given in Geonames. Figure 5 depicts the three hundreds with generalised boundaries and a combination of all the geo-located names in DEEP, Geonames, NGA Gazetteer and the OS 50K, plus the 126 ‘PopulatedPlaces’ in OS Open Names. The location of points from these datasets correspond well with one another.

Table 7 presents a statistical comparison between all six datasets. DEEP has the most records, but few are spatially located. Similarly, while Geonames clearly identifies and distinguishes between administrative units and settlements, for this area it provides fewer geo-located entries than DEEP. OS 50K has a small number of entries with limited classification while OS Open Names has good coverage across the area, but removing postcodes eliminates almost half the records. Railway-related points are few, and while there are plenty of road identifiers this dataset is perhaps most helpful for natural landscape features.

Analysing a single parish manually shows even more clearly the differences between GB1900 and DEEP, as shown in Figure 6. Sheringham was chosen, partly because it developed into an urban settlement and partly because, unusually, it has three sub-parish entries in DEEP with which other entries are associated. The parish name plus the sub-parish entries for both 'Sheringham' and 'Upper Sheringham' are in GB1900, only the 'Lower Sheringham' settlement name is missing. This is presumably because that settlement had developed to such an extent it was no longer referred to in this way by the early twentieth century, while Upper Sheringham was still distinct. Additional map evidence from around the turn of the twentieth century confirms this.[2]

Overall there were 116 entries in the DEEP dataset associated with this parish. GB1900 Complete had 157 points identified within the modern parishes of Sheringham or Upper Sheringham. Of the 45 DEEP entries for field names in Sheringham parish only 'Gibbet Plantation' could be found in GB1900, and this name also describes a landscape feature. Encouragingly, of the 67 DEEP 'mapped name' entries, all but five did match GB1900. Three of these unmatched entries relate to route-ways, 'Butts Lane', 'Limkiln Lane' and 'Holway Road', each of which had two

GB1900 entries because these linear features were each named twice. Also missing from GB1900 were ‘Potter’s Kiln’ and ‘Elcot House’, although there was an ‘Elcot’. In addition six matches had slight variations in spelling or punctuation, but clearly related to the same feature ‘Bullock[‘]s Carr’, ‘Bunker[‘s] Hill’, ‘Golboro['] Spinney’, ‘Howe’s Hill [Tumulus]’, ‘North St[reet]’ and ‘South St[reet]’. The result of mapping these matched features together is shown in Figure 6.

All the DEEP matches are included in the abridged version of GB1900 and of the 91 GB1900 entries without a DEEP match, only 25 are included in the abridged version. These include administrative labels for the Urban District of ‘Sheringham’ and one label for ‘Upper Sheringham’ because GB1900 has two entries (one each for the village and the parish (created in 1901) whereas DEEP only has one entry. There are also five buildings, plus a lifeboat house and a water works, eleven transport links which were all road names except one railway line label, two names associated with the neighbouring parish of Beeston Regis, ‘Sheringham Wood’, the aforementioned ‘Elcot’ and the site some stone querns were discovered.

GB1900 as an online gazetteer

Our project partners are already using GB1900 place name data. The Royal Commission on the Ancient and Historical Monuments of Wales have used the 125,000 names collected during the earlier Cymru1900 project, with other data, to create a List of historic place names for Wales (https://historicplacenames.rcahmw.gov.uk/). Property developers are effectively required to consult this list, as it is now a statutory requirement to consider historically appropriate place names whenever new developments are being planned in Wales. The National Library of Scotland have integrated GB1900 data into their geo-referenced maps explorer (http://maps.nls.uk/geo/explore/), offering a more detailed gazetteer for searching the map interfaces than was previously available.

The Great Britain Historical GIS are relaunching the web site *A Vision of Britain through Time* as PastPlace.org, with searchable GB1900 data as a major focus (Great Britain Historical GIS Project 2017). This version of GB1900 is based on the abridged dataset but enhanced in two ways. Firstly, each GB1900 entry provides links to the pages for various administrative units whose boundaries contained the location: to the modern local authority, the historic county and parish, and in England and Wales to the nineteenth century Registration District; this last is important to genealogists seeking to locate vital registration records for their ancestors. It also links to the four nearest 'places', generally meaning towns or villages for which text from historical gazetteers and travel writing is available.

Secondly, entries have been matched to corresponding entries in the DEEP gazetteer. This has been done by first using the hierarchy within DEEP to associate each lower-level entry with a DEEP parish entry, and then matching DEEP parishes to entries in the GBH GIS Administrative Unit Ontology (AUO), either via the hierarchy or matching parish coordinates from DEEP to GBH GIS boundary data (Southall 2012); the penultimate column in Table 2 shows results from this. Then, having already matched GB1900 entries to historic parishes, they were further matched to DEEP entries based on both containing parish and name matching, while excluding ambiguous cases.

Matching to the AUO was complicated by EPNS teams using a mixture of historical and relatively modern geographies: many of the Hundreds appearing had been abolished through mergers in pre-modern times, while many of the parishes listed were created through mergers in the twentieth century, or are groupings of actual parishes which have never existed as legal entities, such as ‘Lydiard’ in Wiltshire, combining the actual parishes of Lydiard Millicent and Lydiard Tregoze. Table 2 shows that the large majority of DEEP ‘Sub-county’ and ‘Parish’ entries are now matched to the AUO, although this required significant manual work.

Automated matching of individual GB1900 and DEEP entries is similarly problematic, and on-going. We have also done more manual matching of DEEP Sub-Parish entries, as these provide the richest historical information. The final column in Table 2 shows that the majority of DEEP ‘Mapped Names” and ‘Sub-Parish’ entries are now matched to GB1900 entries.

The search interface can be accessed at:

http://www.pastplace.org/expertsearch#gb1900

Visitors can include wild cards within their search terms, and narrow searches by county. Figure 7 shows how results are presented against the background of mapping, supplied by the National Library of Scotland. Here the initial view is of all the locations in Britain matching the search, but zooming-in on a particular location displays the County Series mapping. Clicking on a particular location selects it, and displays the results shown on the right. In this case, information from DEEP is included, listing earlier names by which this village was known and the dates and source from which these ‘attestations’ were drawn.

The site provides resolvable Uniform Resource Identifiers (URIs) for the GB1900 gazetteer via the simplified seven-digit numeric identifiers described earlier:

http://www.pastplace.org/gb1900/1474716

There are limited benefits to making GB1900 available in a format based on RDF (Resource Description Framework), or via a Linked Data API, if it consists simply of a large number of place names, coordinates and unique identifiers. However, Pelagios Commons have very recently funded a new small project to publish the AUO as Linked Data, and we hope to be able to include GB1900 within that, exploiting the GB1900-AUO linkages described above.

Discussion

We have argued elsewhere for ‘spinal gazetteers’, but this term has often been misunderstood (Ell, Hughes, and Southall 2016, p. 156). We are certainly calling for something more specific than just ‘a really large and really important gazetteer’, and if anything we are calling for smaller, not larger gazetteers. The main reason for this is seen above: most gazetteers contain multiple instances of more or less the same place name in more or less the same location. In most cases, they can be associated with different geographical features, in some sense, but geographical names encountered in historical texts can rarely be clearly associated with a particular feature. In a true spinal gazetteer, this ambiguity is removed: if the same name appears more than once, each instance should be in a quite different location and identify a quite different ‘place’.

We began developing just such a spinal gazetteer for Britain by grouping together the many different administrative units named after the same place held in the AUO; the small town of Sheringham discussed above is not an especially good example, but the AUO identifies a parish, a manor, a Registration sub-District and an Urban District, each with different boundaries. Focus group testing showed that this was confusing for most users, especially where there were both multiple settlements of the same name and multiple administrative units named after each settlement. Grouping units into ‘places’ enabled a two-stage search process, users first selecting a place and then a unit, and we were then able to also link in more qualitative sources, such as travel writing, which could be linked only to places, not units. We currently define 22,311 ‘places’ within Britain, versus 71,468 administrative units (Southall 2014), and this reflects much manual editing to locate and remove duplicates, making the spinal gazetteer smaller and better.

Our evaluation of alternative gazetteers suggest much work to make gazetteers larger but not necessarily more useful. The Ordnance Survey’s Open Names gazetteer initially appears vast, but is mostly postal codes and a street directory, in significant ways less useful than their earlier 50K gazetteer. Geonames appears to be almost twice as large as the NGA gazetteer from which it partly derives, but much of the additional content are parishes which duplicate settlement names, while people seeking hotels are probably better off with Trip Advisor. Similarly, the field names which form 70 per cent of the DEEP Gazetteer lack locations and attributions, so what is their value in a gazetteer?

One virtue of GB1900 is that it is based on the names appearing on a single but very detailed set of maps. This inherently limits duplication, although as with Geonames and DEEP the names of parishes largely duplicate the names of the main settlements. We are exploring whether parish names can be identified and potentially filtered out, through a combination of automated checking based on whether or not transcriptions are in capital letters, and manual checking based on the detailed information held in the AUO about which parishes existed circa 1900. Duplication also occurs through linear features being named at multiple points along their routes: railways, roads, rivers. The Complete GB1900 gazetteer also, of course, includes many items which are not place names at all, but they are relatively easily filtered out.

Another way gazetteers vary is in whether, and how thoroughly, they include a classification of features. Here, confusingly, the terminology varies greatly: the OS 50K uses just ‘Feature Code’, Geonames uses ‘Feature Class’ and within that ‘Feature Code’. NGA similarly uses ‘Feature Class’ divided into ‘Feature Designated Codes’. OS Open Names uses completely different classification names, ‘Type’ and within that ‘Local Type’. DEEP does not have a column specifying a feature classification, but the place IDs effectively provide a typology, even though there is significant variation in usage between the volumes. GB1900 lacks feature types, reflecting the lack of symbology in the County Series maps, but most features can be assigned to a broad classification. Of course, a true spinal gazetteer is inherently untyped as it is concerned not with features but with a more abstract notion of ‘place’. The OS50K gazetteer reminds us that many if not most names on many topographic maps are not linked to features and so fit uneasily into a typology.

Finally, this paper has explored the potential for integrating different gazetteers. This should never mean simply massing them together into a single vast list of place names and coordinates: in particular, all six gazetteers each include the name of every town and significant village at least once, but never with the exact same coordinate and often with slightly different coordinates, greatly complicating automated matching or elimination of duplicates. This is arguably why only OS Open Names makes any attempt to align itself with other gazetteers, and has achieved this for only 45 per cent of entries.

More specifically, work with the DEEP XML files began simply to enable a comparison, but has developed into a new project to integrate GB1900 and DEEP. This paper began with a historical account which noted that the English Place Names Survey begins each local survey by gathering place names from six inch maps, and that the Cymru1900 transcription project originated as an attempt to replicate the English survey for Wales, but working somewhat more speedily. The slowness of the survey's methods means that we can add data from DEEP for only about a quarter of Britain, but for those areas we are adding arguably the most thorough historical survey of place names made anywhere in the world, to the most detailed specifically historical gazetteer. This integration helps remedy DEEP's greatest weakness as a gazetteer, its lack of coordinate data. It is also only possible because both data sets are under Creative Commons licenses.

Returning to GB1900, it clearly cannot match the most detailed resources created by the UK national mapping agency: the Ordnance Survey's MasterMap system contains c. 450 million labelled features, but is not historical and available only at high cost (Ordnance Survey, n.d.). Setting MasterMap to one side, GB1900 identifies more 'places' and a greater number of overall locations, especially natural features and individual buildings, than any other freely-available data set, and is also the most detailed specifically historical gazetteer of Britain, or arguably of anywhere else. Grounded in a particular historical source, but now enhanced with information from DEEP, it will be a key reference aid, and organising framework, both for academic historical researchers, and a wide range of amateur family and local historians.

Acknowledgments

The work reported here would have been impossible without the support of our GB1900 project partners, and the work of the GB1900 project volunteers. Particular thanks to the 22 volunteers who assisted with the final manual checking, reported here. Thanks also to Chris Fleet and the National Library of Scotland for the supply of the six inch map images.

End Notes

[1] These lost settlements have been identified further by referencing external sources. The Norfolk Heritage Explorer website (Norfolk County Council 2017) describes:
‘Shipden’ near Cromer: http://www.heritage.norfolk.gov.uk/record-details?MNF11727-Site-of-Shipden-medieval-village,
‘Rippon Hall’ in Hevingham parish: http://www.heritage.norfolk.gov.uk/record-details?MNF7653-Rippon-Hall-or-Catte%27s-Hall,
‘Bolwick’ in Marsham parish: http://www.heritage.norfolk.gov.uk/record-details?MNF7485-Undated-mound-and-possible-site-of-Bolwick-deserted-medieval-settlement,
‘Southgate’ in Cawston parish: http://www.heritage.norfolk.gov.uk/record-details?MNF14398-Sygate-or-Southgate-deserted-medieval-village,
Blomefield’s (1807) detailed description of Norfolk includes ‘Crakeford’ near Banningham (pp. 326-330): https://www.british-history.ac.uk/topographical-hist-norfolk/vol6/pp326-330, and ‘Mortoft’ in Heydon parish (pp. 241-253): https://www.british-history.ac.uk/topographical-hist-norfolk/vol6/pp241-253#fnn4

[2] County Series 1:10560 first edition County Sheet for Norfolk published in 1888 names the settlement Lower Sheringham, while the first revision published in 1907 names it just Sheringham.

References

Adams, John. 1680. *Index Villaris: Or, an Exact Register. Alphabetically Digested, of All the Cities, Market-Towns, Parishes, Villages, the Hundred, Lath, Rape, Ward, Wapentake, or Other Division of Each County [Etc].* London: Sawbridge and Gillyflower.

Ahlers, Dirk. 2013. “Assessment of the Accuracy of GeoNames Gazetteer Data.” In *GIR’13 - 7th Workshop on Geographic Information Retrieval*, edited by Ross Purves and Chris Jones, 74–81. ACM Digital Library. https://doi.org/10.1145/2533888.2533938.

Aucott, Paula, Humphrey Southall, and Carol Ekinsmyth. 2019. “Citizen Science through Old Maps: Volunteer Motivations in the GB1900 Gazetteer-Building Project.” *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 52. https://doi.org/10.1080/01615440.2018.1559779.

Berman, Merrick Lex, Ruth Mostern, and Humphrey Southall, eds. 2016. *Placing Names. Enriching and Integrating Gazetteers*. Bloomington: Indiana University Press.

Blomefield, Francis. 1807. *An Essay Towards A Topographical History of the County of Norfolk: Volume 6*. London: W. Miller.

Camden, William. 1610. “Introduction: The Author to the Reader.” In *Britain, or, a Chorographicall Description of the Most Flourishing Kingdomes, England, Scotland, and Ireland*, edited by Philemon Holland, English tr. http://www.visionofbritain.org.uk/travellers/Camden/1.

Crone, Gerald Roe, E M J Campbell, and R A Skelton. 1962. “Landmarks in British

Cartography." *The Geographical Journal* 128 (4): 406–26. https://doi.org/10.2307/1792037.

Ell, Paul S, Lorna Hughes, and Humphrey Southall. 2016. "Digitally Exposing the Place Names of England and Wales." In *Placing Names*, edited by Merrick Lex Berman, Ruth Mostern, and Humphrey R Southall, 146–62. Enriching and Integrating Gazetteers. Bloomington: Indiana University Press.

Great Britain Historical GIS Project. 2017. "PastPlace". Great Britain Historical GIS, University of Portsmouth. Accessed April 2, 2019, http://www.pastplace.org/

Harley, J B. 1971. "Place-Names on the Early Ordnance Survey Maps of England and Wales." *The Cartographic Journal* 8 (2): 91–104.

National Library of Scotland. n.d. "Ordnance Survey Maps". Accessed April 2, 2019. https://maps.nls.uk/os/index.html.

Norfolk County Council. 2017. "Norfolk Heritage Explorer: Norfolk Historic Environment Record: Parish Summaries". Accessed April 2, 2019. http://www.heritage.norfolk.gov.uk/parishes.

Ordnance Survey. n.d. "OS MasterMap." Accessed May 11, 2019. https://www.ordnancesurvey.co.uk/business-and-government/products/mastermap-products.html

Sandred, Karl Inge. 2002. *The Place-Names of Norfolk. Part Three. The Hundreds of North and South Erpingham and Holt*. Nottingham: English Place-Name Society.

Smith, A.H. 1954. *The Preparation of County Place-Name Surveys*. London: English Place-Name Society.

Southall, Humphrey. 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): 119–34.

———. 2014. "Rebuilding the Great Britain Historical GIS, Part 3: Integrating Qualitative Content for a Sense of Place." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 47 (1): 31–44. https://doi.org/10.1080/01615440.2013.847774.

Southall, Humphrey, Paula Aucott, Chris Fleet, Tom Pert, and Michael Stoner. 2017. "GB1900: Engaging the Public in Very Large Scale Gazetteer Construction from the Ordnance Survey 'County Series' 1:10,560 Mapping of Great Britain." *Journal of Map & Geography Libraries* 13 (1): 7–28.

Southall, Humphrey, Ruth Mostern, and Merrick Lex Berman. 2011. "On Historical Gazetteers." *International Journal of Humanities and Arts Computing* 5 (2): 127–45.

Tables

Table 1: Gazetteers included in comparative analysis

Name	Coverage	Rows	Downloaded from	Date
GB1900 Abridged	GB	1,097,123	http://www.pastplace.org/data/#gb1900	N/A
DEEP	Most of England	539,372	http://mads.digitalresources.jisc.ac.uk/mads2017/	21/03/2019
Geonames	UK	63,106	http://www.geonames.org/	18/1/2019
NGA	UK	32,444	http://geonames.nga.mil/gns/html/namefiles.html	11/01/2019
OS 50K	GB	259,057	http://data.ordnancesurvey.co.uk/datasets/50k-gazetteer/downloads (Linked Data version)	30/06/2010
OS Open Names	GB	2,915,336	https://www.ordnancesurvey.co.uk/business-and-government/products/os-open-names.html	30/01/2019

Table 2: Frequency of different place types in the DEEP gazetteer

Place Type	**Total**	**Have Coordinate**	**Have Attestation**	**Have Both**	**Matched to AUO**	**Matched to GB1900**
Field Name	378,543					534
Mapped Name	145,242	8,561	87,155	8,028		81,201
Sub-Parish	8,195	8,190	8,073	8,068		6,357
Parish	6,634	6,634	122	122	6,400	577
Sub-County	461		424		458	
County	66				66	
Below Sub-County	48		28			
Local District	34		29			
DB Hundred	31		31			
[Ten Other Types]	118	63	53	10	6	
TOTAL	539,372	23,448	95,915	16,228	6,930	88,669

Table 3: Feature Class Frequencies for UK from the NGA Gazetteer and Geonames

Feature Class	**Description**	**Geonames**	**NGA Gazetteer**
A	Administrative	11,888	872
H	Hydrographic	5,610	5,634
L	Area or Localities	850	321
P	Populated Places	18,334	17,285
R	Road/Railroad or Transportation	422	12
S	Spot	21,106	3,089
T	Hypsographic	4,587	5,131
U	Undersea	22	24
V	Vegetation	287	76
Total		63,106	32,444

Table 4: Feature Code Frequencies in OS 50K Gazetteer

Feature Code	**Meaning**	**Frequency**
X	All other features	128,655
O	Other	41,219
FM	Farm	34,726
W	Water feature	24,423
H	Hill or mountain	14,518
F	Forest or wood	8,706
A	Antiquity (non-Roman)	5,252
T	Town	1,259
R	Antiquity (Roman)	237
C	City	62
Total		259,057

Table 5: Type frequencies in OS Open Names

Type	**Frequency**
Hydrography	24,951
Landcover (e.g. woods)	118,395
Landform (e.g. hills)	76,470
Other of which are Postcodes	1,731,929 1,697,220
PopulatedPlace	42,930
TransportNetwork of which are Roads	920,661 916,068
Total	2,915,336

Table 6: Detailed Local Type classification of ‘PopulatedPlace’ in OS Open Names

Local Type	**Count**	**Link to Geonames**	**Link to dBpedia**
City	64	58	28
Hamlet	12,826	1,775	2,480
Other Settlement	2,725	610	668
Suburban Area	10,824	1,362	2,361
Town	1,358	1,313	1,273
Village	15,133	8,130	8,534
Total	42,930	13,248	15,344

Table 7: Feature classification comparison of Norfolk study area in all six gazetteers

Dataset	**Total Records**	**Total Points**	**Settlement**	**Administrative Unit**	**Natural Landscape Feature**	**Transport**	**Other**
DEEP	6,763	210	103	97	0	0	10
OS Open Names	5,503	5,503	126	0	862	1,894	2,621
OS 50K	481	481	4	0	55	11	411
Geonames	182	182	67	79	5	12	19
NGA	72	72	64	0	1	4	3
GB1900	6,357	6,357	-	100	921	1,096	4,240

Figure captions

Figure 1: Excerpt from Ordnance Survey second edition County Series Six Inches to One Mile map, showing part of Sheringham, Norfolk

Figure 2: Name of the railway company and branch line split by the railway track

Figure 3: Coverage of all DEEP point locations for place name entries in England

Figure 4: GB1900 locations for named points relating to administrative units and buildings in three Norfolk hundreds

Figure 5: Place name location points from other gazetteers in three Norfolk hundreds

Figure 6: Spatial comparison of gazetteer entries for the parish of Sheringham

Figure 7: GB1900 search results page within A Vision of Britain through Time

Figures

Figure 1

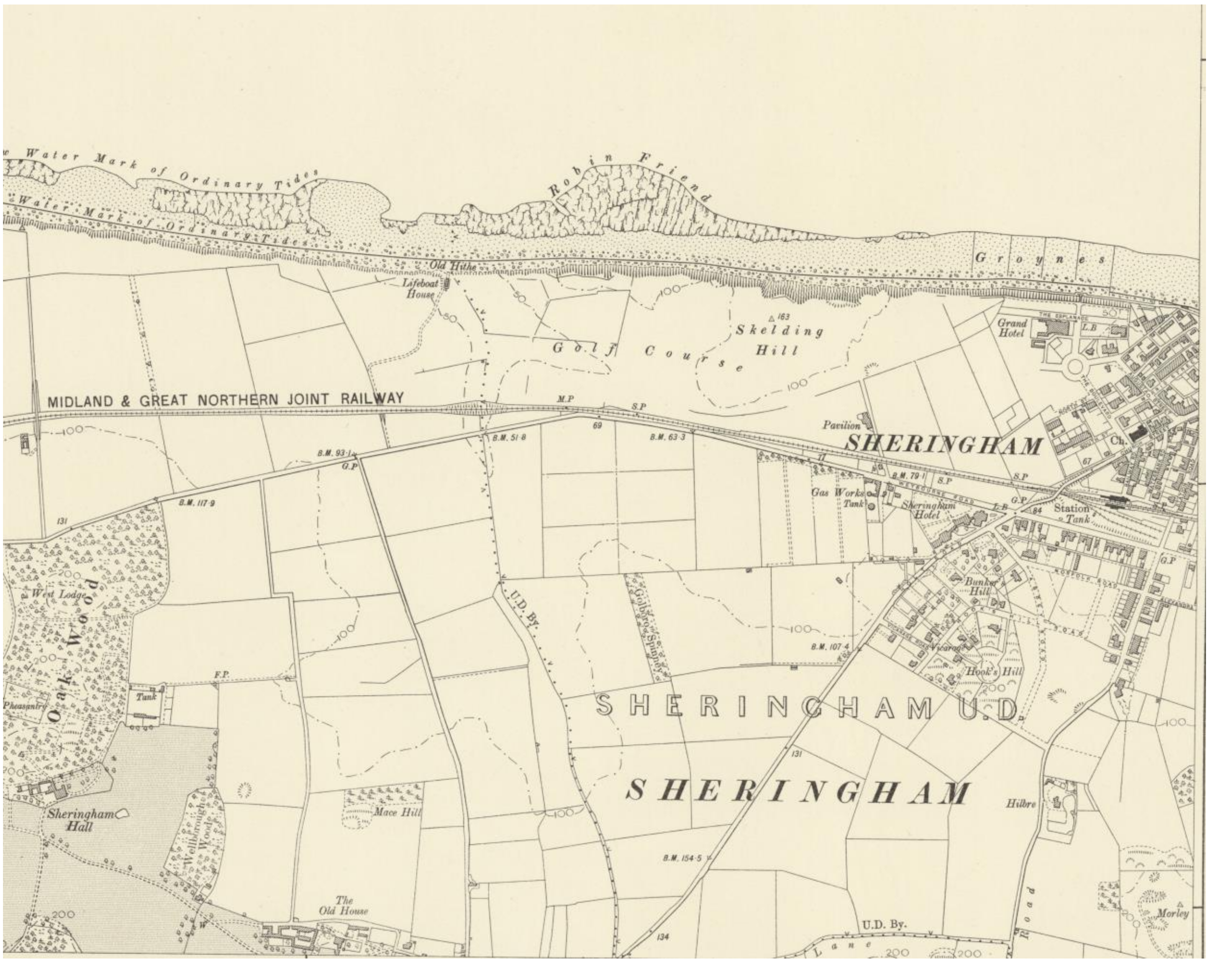


Figure 2

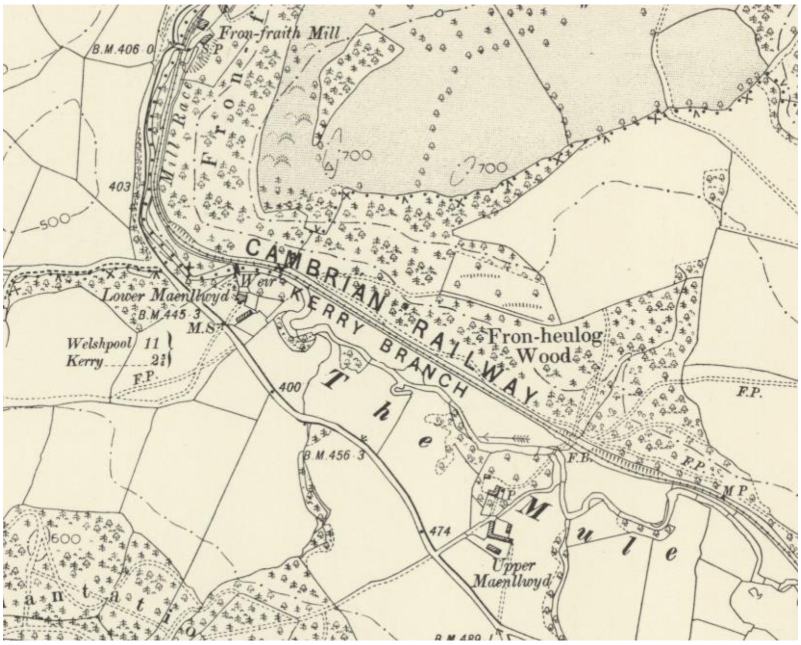

Fron-fraith Mill
S.P.
B.M.406·0
Mill Race
Fron-f
403
500
700
700
CAMBRIAN RAILWAY
KERRY BRANCH
Weir
Lower Maenllwyd
B.M.445·3
M.S.
Welshpool 11
Kerry 2¼
F.P.
Fron-heulog Wood
F.P.
400
The Mule
B.M.456·3
F.B.
F.P.
M.P.
P.
474
Upper Maenllwyd
600
antatio
B.M.489

Figure 3

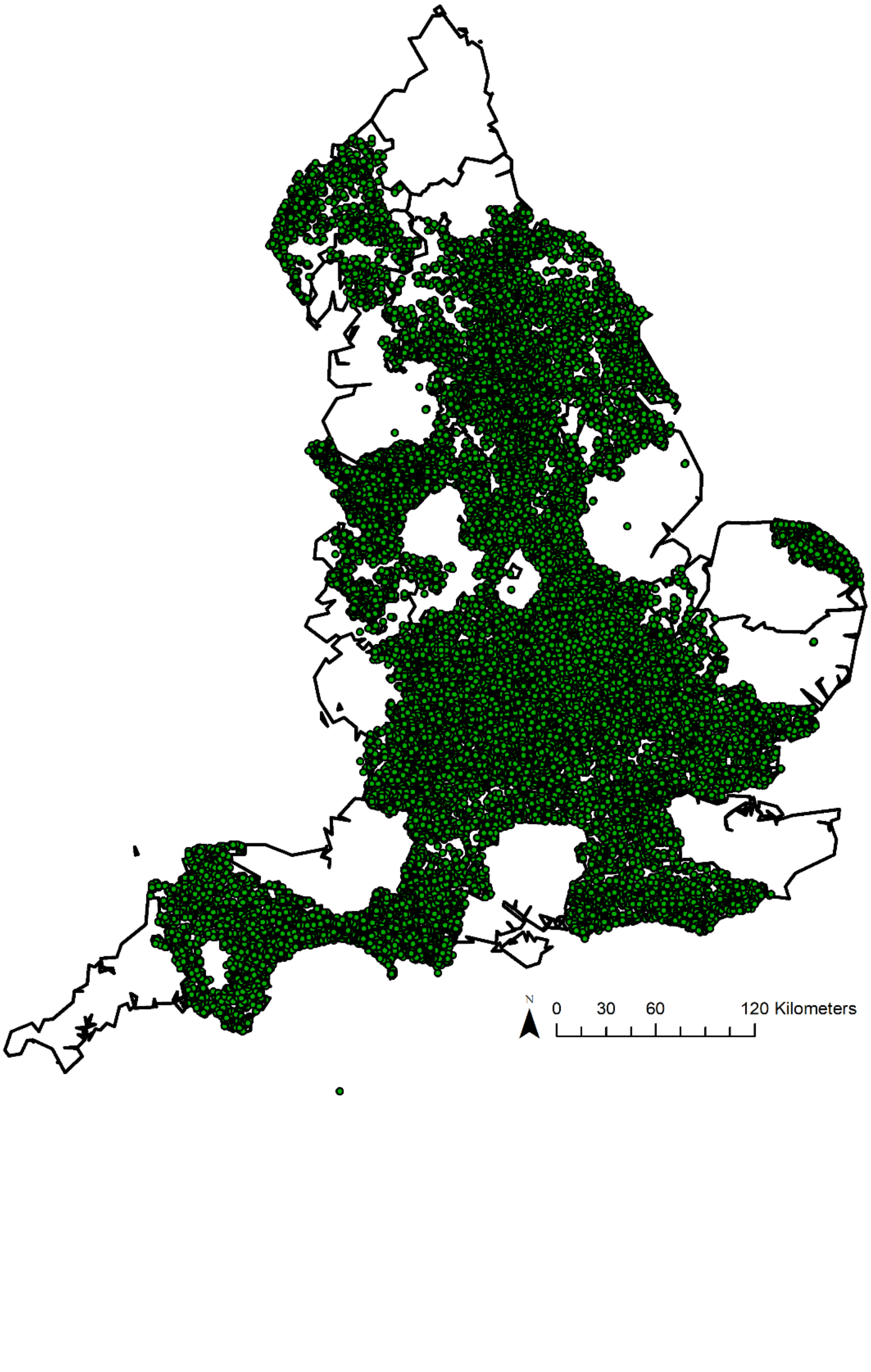


Figure 4

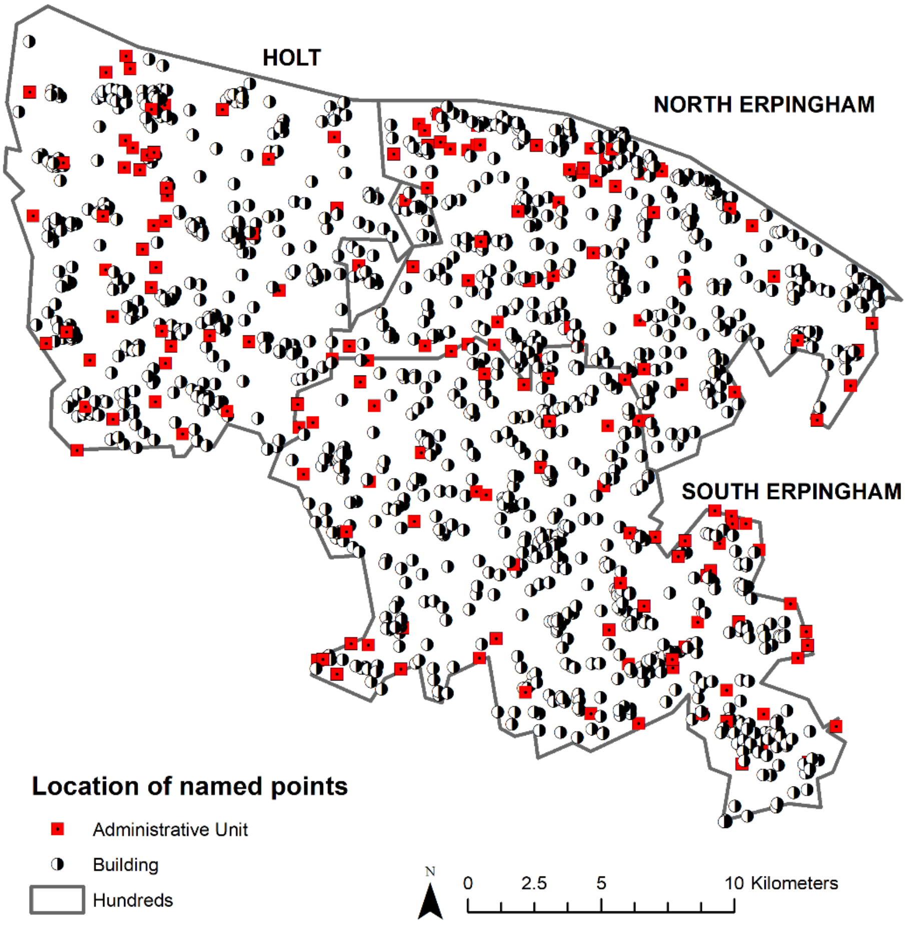

HOLT
NORTH ERPINGHAM
SOUTH ERPINGHAM
Location of named points
Administrative Unit
Building
Hundreds
N
0
2.5
5
10 Kilometers

Figure 5

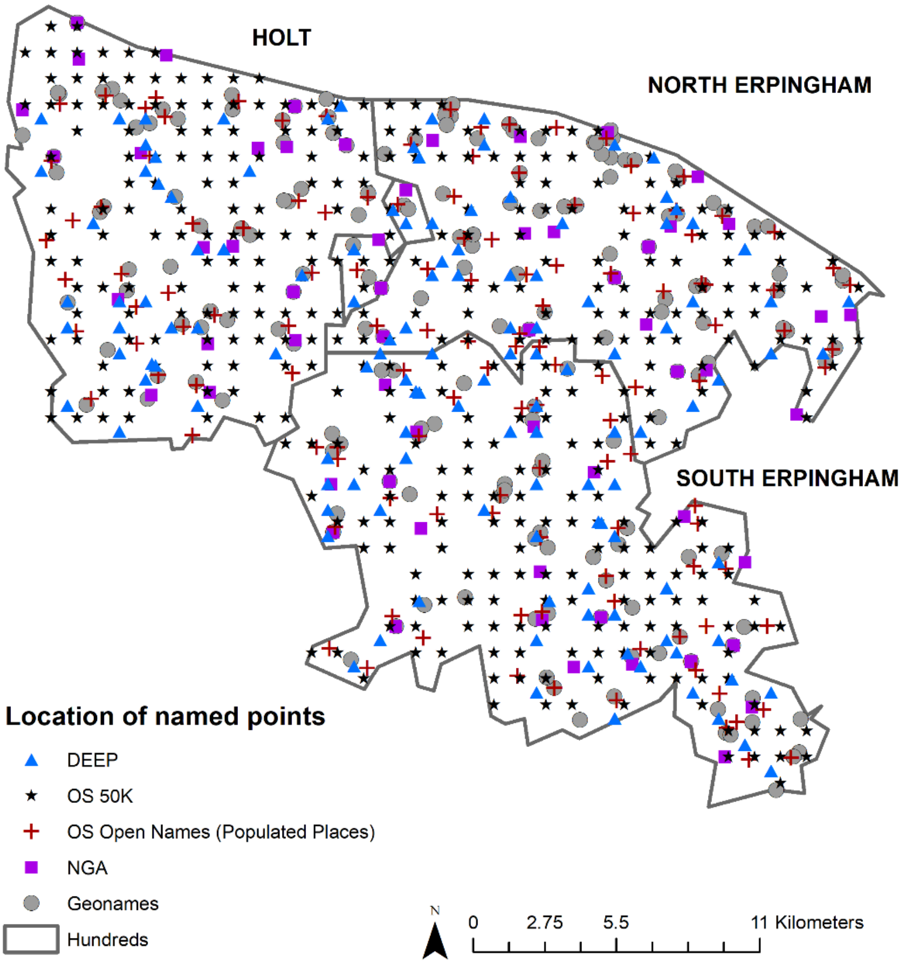

HOLT
NORTH ERPINGHAM
SOUTH ERPINGHAM
Location of named points
DEEP
OS 50K
OS Open Names (Populated Places)
NGA
Geonames
Hundreds
N
0
2.75
5.5
11 Kilometers

Figure 6

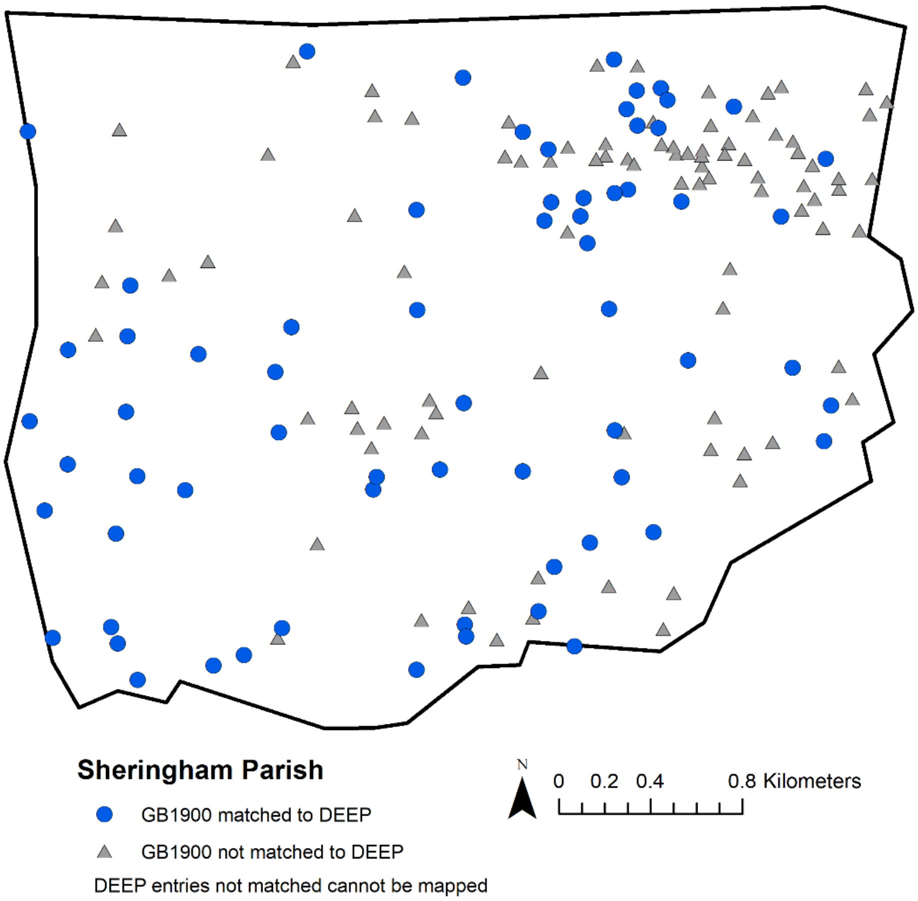
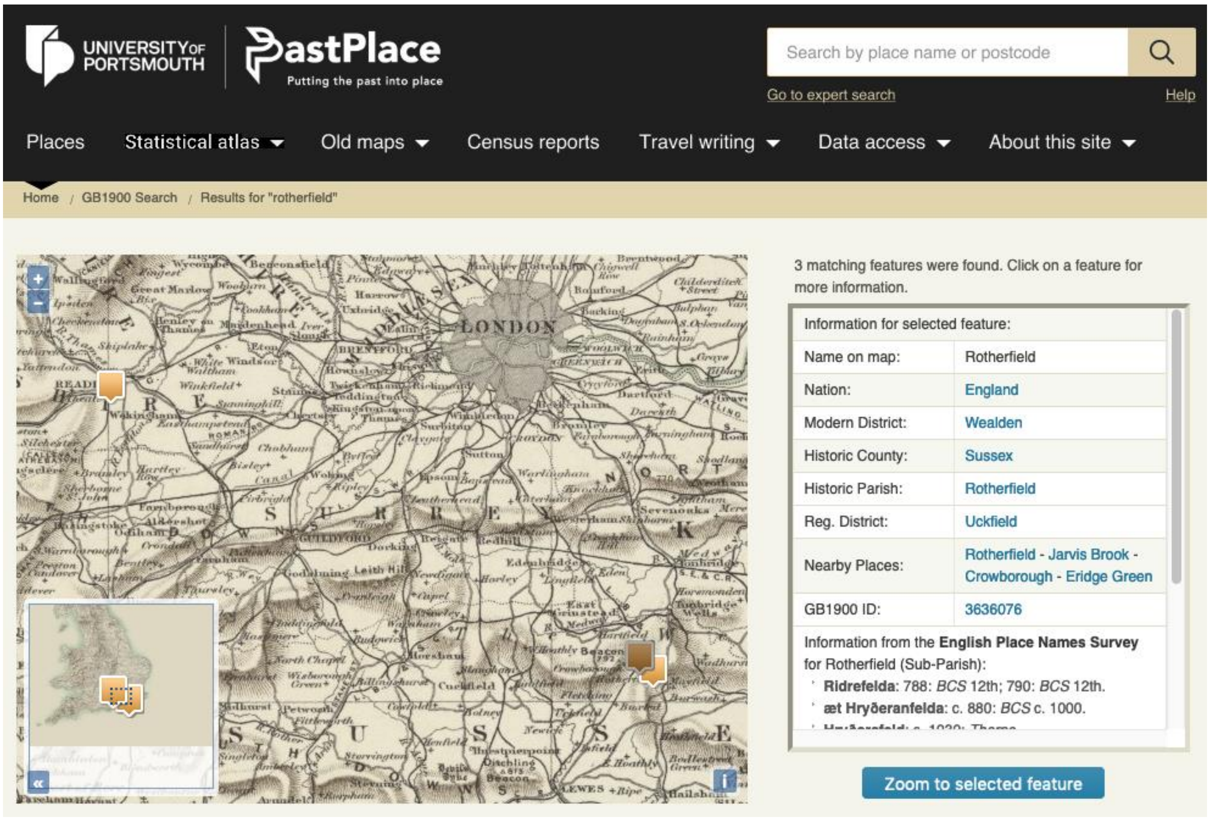

Sheringham Parish
● GB1900 matched to DEEP
▲ GB1900 not matched to DEEP
DEEP entries not matched cannot be mapped
N
0
0.2
0.4
0.8 Kilometers

Figure 7



3 matching features were found. Click on a feature for more information.

Information for selected feature:

Name on map:	Rotherfield
Nation:	England
Modern District:	Wealden
Historic County:	Sussex
Historic Parish:	Rotherfield
Reg. District:	Uckfield
Nearby Places:	Rotherfield - Jarvis Brook - Crowborough - Eridge Green
GB1900 ID:	3636076

Information from the **English Place Names Survey** for Rotherfield (Sub-Parish):

- **Ridrefelda**: 788: *BCS* 12th; 790: *BCS* 12th.
- **æt Hryðeranfelda**: c. 880: *BCS* c. 1000.