Disambiguating geographical names in historical British censuses and travel writing*

Paula Aucott (corresponding author)
https://orcid.org/0000-0002-1637-2672
University of Portsmouth

Humphrey Southall
https://orcid.org/0000-0002-4406-1425
University of Portsmouth

Zarys treści: Niniejszy artykuł stanowi studium dwóch przypadków. Pierwszym jest nowatorska analiza nazw parafii w angielskich raportach spisowych, gdzie kompleksowe ujednoznacznienie wymagało dopasowania do polihierarchicznej ontologii jednostek administracyjnych (*Administrative Unit Ontology* – AUO). Drugim jest identyfikacja toponimów w dużej kolekcji historycznych pism podróżniczych dokonana poprzez analizę trasy podróżnika. Stworzone w wyniku tego spis AUO oraz szczegółowy wykaz miejsc to bogate struktury danych integrujące różnorodne źródła historyczne.

Słowa kluczowe: Wielka Brytania, Irlandia, spisy ludności, nazwy miejscowości, podróżopisarstwo, indeks geograficzny

Abstract: This paper presents two case studies. Firstly, a novel analysis of parish names in English census reports where comprehensive disambiguation required matching to a poly-hierarchic Administrative Unit Ontology (AUO). Secondly, identifying toponyms in an extensive collection of historical travel writing by following the traveller's route. Constructed alongside the toponym matching, the AUO and 'places' gazetteer are rich data structures that integrate diverse historical sources.

Keywords: Great Britain, Ireland, population censuses, place names, travel writing, gazetteer

Introduction

Historical texts and statistical tabulations almost always contain geographical names, not coordinates, so exploring past geographies digitally means converting these names into coordinate geometries, meaning points, lines or polygons. This is arguably true even when the historical source is a map, as limited topographical accuracy means we cannot simply overlay images of historical maps on modern digital mapping, we must link named features on the two kinds of mapping and use these as control points to 'rubber sheet' the historical map to fit modern geography.

Sometimes, such associations are genuinely unproblematic. For example, the string of letters 'Colwall' seems to refer to just one locality on the earth's surface, in England's county of Herefordshire; there is still some ambiguity but purely local, between Upper Colwall, Colwall Stone, Colwall Green and Old Colwall, which we do not address here. However, a given name may often refer to multiple very different locations. Further, as this article will discuss, dealing with historical sources means allowing for greater variations in how the name of a particular place is written, and in the administrative hierarchies often used to disambiguate common place names.

Place name ambiguity is an especially large issue for Britain, partly because we

* This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

have a thousand years' worth of written history, during which many toponyms have substantially evolved from their original Anglo-Saxon, Norse or Celtic forms. Further and more specifically, while British census reports go no further back than 1801, a similar period to those of, for example, the United States', US census geographies mainly concern a single system of never more than 50 states and 3,000 or so counties, while in Britain and Ireland the census geographer must work not only with circa twenty thousand parishes, as discussed below, but with multiple county-level and district level geographies, often existing in parallel; so even historical researchers are often unclear which kind of county a listing is referring to.

The next section reviews the existing literature on place name disambiguation, but this is mainly research by data scientists developing automated methods to convert toponyms contained within vast bodies of free text into the most probable locations. We are, instead, historical geographers engaged in substantive research and, in presenting the history of places to the general public, and often not just places in general, but the particular places each person lives in through our website 'A Vision of Britain through Time'.[1] Here, 'most probable' is not good enough: incorrect place attributions can and often have led to complaints, so if we cannot be reasonably certain the source is best not included.

Our aim is complete disambiguation, and this is illustrated in our two case studies. The first comes from our statistical research, focusing on the most geographically-detailed published tabulations, for the parishes of the United Kingdom, as listed in the reports of the Censuses of Population for England and Wales, Scotland and, until 1911, Ireland. Parishes are identified by names, not numbers, and these names are neither unique nor fixed over time, as shown by a detailed analysis of the names of English Civil Parishes. However, the tabulations are organised around administrative hierarchies and, despite these also changing over time, once the hierarchies are understood, it is possible to assemble data from successive censuses to construct population time series for essentially all parishes.

The second case study describes how over twenty thousand place references have been located within a corpus of twenty-five historical British travel writers, mostly from the eighteenth and nineteenth centuries, totalling nearly three million words. The methods used here were essentially manual, and the examples given explain why automated matching was rejected. In brief, unlike the census reports, town and village names lack hierarchical context, so we must instead follow the traveller's journey, and accept that references to places which do not form part of their sequence of visits may be unidentifiable.

The final sections bring together the approaches for disambiguating names in both the Census and travel writing in the more recently added information on Irish places. It discusses the advantages gained from our previous experiences and why this kind of detailed disambiguation is so important.

1. Literature review

The literature concerning the disambiguation of place names can be grouped into two parts of this process: those considering the identification of each place name within a body of text and then those taking the second step to find the location of that identified name.

With regards to the first of these steps, Hill's work with the Alexandria Digital Library initiative emphasised the need for consistent naming to act as the link between geographic representations, a role

[1] Great Britain Historical GIS Project, University of Portsmouth, 'A Vision of Britain through Time', 2003–24, https://www.visionofbritain.org.uk/.

that could be filled by a gazetteer.[2] Selection of a suitable gazetteer to use as this link is identified as an important factor in the success of the process.[3] There is no single gazetteer that would work as a disambiguation source in all instances. Indeed, most projects working in this area have used their own bespoke gazetteer created specifically for the project. Often, this is built by combining existing gazetteers covering specific localities with more well-known reference works, including the Getty Thesaurus of Geographic Names, Geonames, and, in more recent works, Wikipedia, Wikidata and DBpedia.[4] Examples of using these gazetteers all use at least some automation.

The most useful early discussion on disambiguating historical place names in a digital setting was by Smith and Crane.[5] Their collation of digitised versions of texts for the Perseus project, dating from the nineteenth century back to ancient Greece, led them to develop a two-step classification method. The first step separates out entities to identify the names and type of record ('Named Entity Recognition'), and the second step takes the identified place names and identifies the locations they refer to in a bespoke gazetteer. Their project combined various sources to produce a gazetteer that contained over one million place names, validating their claim that manually tagging place names would be impractical for large bodies of text.

Different approaches to matching the place names with a location have been developed, based on maps, context or annotated data.[6] Buscaldi and Rosso present their map-based matching, using the coordinates of the matched place names to determine the most likely match based on the distance to the centroid of all other place names given in the text, as successful at 85–96%.[7] However, they do point out that the gazetteer used may have candidate matches missing, thus even a good match is not necessarily the correct one. Rauch and colleagues created the Metacarta search engine software to disambiguate geographical terms in texts.[8] They used data mining techniques to train their gazetteer to identify the terms most likely to be geographical and applied a confidence measure which used the proximity of other locations already identified as place names to help determine the terms most likely also to be place names. Pouliquen and colleagues used the minimum distance between ambiguous and non-ambiguous place names to rank potential matches.[9]

[2] L. Hill, 'Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints', in *Research and Advanced Technology for Digital Libraries: 4th European Conference, ECDL 2000 Lisbon, Portugal, September 18–20, 2000 Proceedings 4* (Springer, 2000), pp. 280–90.

[3] E. Rauch, M. Bukatin, and K. Baker, 'A Confidence-Based Framework for Disambiguating Geographic Terms', *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References* (2003), pp. 50–54.

[4] M. Coll Ardanuy, C. Sporleder, 'Toponym Disambiguation in Historical Documents Using Semantic and Geographic Features', *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage* (2017), pp. 175–80; S. Overell, J. Magalhães, and S.M. Rüger, 'Place Disambiguation with Co-Occurrence Models', *CLEF (Working Notes)* (2006), https://ceur-ws.org/Vol-1172/CLEF2006wn-GeoCLEF-OverellEt2006.pdf (accessed on 27 Dec. 2024); J. Santos, I. Anastácio, and B. Martins, 'Using Machine Learning Methods for Disambiguating Place References in Textual Documents', *GeoJournal*, vol. 80 (2015), pp. 372–92; E.A. Sultanik, C. Fink, 'Rapid Geotagging and Disambiguation of Social Media Text via an Indexed Gazetteer', *Proceedings of the 9th International ISCRAM Conference – Vancouver, Canada, April 2012*, ed. L. Rothkrantz, J. Ristvej, and Z. Franco (2012), https://idl.iscram.org/files/sultanik/2012/212_Sultanik+Fink2012.pdf (accessed on 27 Dec. 2024); R. Volz, J. Kleb, and W. Mueller, 'Towards Ontology-Based Disambiguation of Geographical Identifiers', *I3: Identity, Identifiers, Identification* (2007), https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=a0fc2bba69e48a3bf310e882bdd9b8f8484b98c8 (accessed on 27 Dec. 2024).

[5] D. Smith, G. Crane, 'Disambiguating Geographic Names in a Historical Digital Library', in *Research and Advanced Technology for Digital Libraries: 5th European Conference, ECDL 2001 Darmstadt, Germany, September 4–9, 2001 Proceedings* (Springer, 2001), pp. 127–36.

[6] Ardanuy, Sporleder, 'Toponym Disambiguation.

[7] D. Buscaldi, P. Rosso, 'Map-Based vs Knowledge-Based Toponym Disambiguation', *Proceedings of the 5th Workshop on Geographic Information Retrieval* (2008), pp. 19–22.

[8] Rauch, Bukatin, and Baker, 'A Confidence-Based Framework', pp. 50–54.

[9] B. Pouliquen, M. Kimler, R. Steinberger, et al., 'Geocoding Multilingual Texts: Recognition, Disambiguation and Visualisation', *Language Resources and Evaluation Conference (LREC)* (2006), http://www.lrec-conf.org/proceedings/lrec2006/pdf/578_pdf.pdf (accessed on 27 Dec. 2024).

Although this experiment obtained a success rate of only 77%, the place names chosen were selected to be ambiguous. In comparison, with a more random selection of newspaper articles, the success rate increased to between 94% and 98%. However, the gazetteer used included only country and city names, thus significantly reducing the possible amount of ambiguity for place names to match to in the gazetteer.

In contrast, Smith and Crane used nearby places for context, and they located all possible places within an individual document by narrowing down available options by assigning a weighted probability of a gazetteer entry being a match to the place name.[10] Around 92% of place names initially matched multiple gazetteer entries, but the automated matching allowed for accuracy between 81% and 96%, depending on the text being analysed.

It has been demonstrated that progressively narrowing the criteria for matching leads to the best results, as does the expansion of the elements of the gazetteer.[11] So, it would include additional information beyond the basic name, category, and point coordinates to encompass alternative names, additional relationships to places both nearby and related in some hierarchical way, and boundaries.[12] Even so, their automated disambiguation was only successful in 82% of 346 place names. This was an improvement on the simple heuristic matching without using the contextual relationships, but significantly less than the 100% success achieved via manual disambiguation done by volunteers, although the volunteers failed to identify as place names the most obscure places they had not heard of.

Volz and colleagues emphasised the need to categorise what kind of disambiguation is needed for the name being identified.[13] These include different versions of a place name for the same location, different locations with the same place name, and names that may or may not be a place and relate to a location, like a person's name. Their use of an ontology allows for the utilisation of formal rules about relationships, more so than the map-based disambiguation methods.

Overell's use of Wikipedia combined with the Getty Thesaurus of Geographic Names to build a structured gazetteer led to a success rate of 89.6%.[14] Ardanuy and Sporleder used Wikipedia data supplemented with GeoNames to identify places, including many alternate spellings and names in nineteenth- and twentieth-century newspaper articles. They tracked the locations of potential matches in relation to other place names in the article to weigh the likelihood of the match.[15] Including GeoNames improved the matching simply because there were far more candidates in that source. Annotating the place names was done manually, but the matching was automated, and it achieved 81% success. However, correctly identifying the right location did not necessarily equate to precise coordinates.

The work most closely aligned with our approach was the Finnish place name ontology, which considers names in different languages and changes in names, boundaries and relationships to higher-level administrative units over time.[16] Like us, they extended their base gazetteer to

[10] Smith, Crane, 'Disambiguating Geographic Names'.

[11] Overell, Magalhães, and Rüger, 'Place Disambiguation'.

[12] I.M.R. Machado, R. Odon de Alencar, R. de Oliveira Campos Jr, and C.A. Davis Jr, 'An Ontological Gazetteer and Its Application for Place Name Disambiguation in Text', *Journal of the Brazilian Computer Society*, vol. 17, no. 4 (2011), pp. 267–79, doi:10.1007/s13173-011-0044-4.

[13] Volz, Kleb, and Mueller, 'Towards Ontology-Based Disambiguation'.

[14] S. Overell, 'The Problem of Place Name Ambiguity', *SIGSPATIAL Special*, vol. 3, no. 2 (2011), pp. 12–15.

[15] Ardanuy, Sporleder, 'Toponym Disambiguation'.

[16] Tomi Kauppinen, R. Henriksson, R. Sinkkilä, et al., 'Ontology-Based Disambiguation of Spatiotemporal Locations', *Proceedings of the 1st IRSW2008 International Workshop on Identity and Reference on the Semantic Web, Tenerife, Spain, June 2, 2008* (2008), https://www.researchgate.net/publication/220853817_Ontology-based_Disambiguation_of_Spatiotemporal_Locations (accessed on 27 Dec. 2024).

include these changes by adding different kinds of relationships between these units. However, their aim was to create an ontology to be used for disambiguation via a manual search interface rather than doing any automated matching.

The majority of these approaches use a range of source information, but their assessment of success is based on matching to available identified places. Overell commented that a 10% error rate in disambiguating place names is too high in cultural heritage environments.[17] This error rate is far too high for a public-facing service such as a website. While some consideration is given to changes over time, this is concentrated in the projects concerning ontological gazetteers, which can more easily accommodate such needs, and very little is given to places which are not initially included in the gazetteer as potential matches. Most work in this field concentrates on obtaining the highest possible success through automated methods without focusing on the difficult places that require more exacting and manual efforts.

2. Ambiguity in English parish names

For England, Wales and Scotland, the most geographically detailed census tables provided population counts for parishes. Although these were originally ecclesiastical units, each organised around a parish church, the priests and officials of the Church of England were also important government officials below the level of counties and, by the nineteenth century, parishes were the most important units of civil administration. By 1881, the census was reporting on a system of 'Civil Parishes', which then evolved separately from ecclesiastical parishes. Especially in the north of England, many of these Civil Parishes had previously been Chapelries or Townships within a large Ancient Parish.

Table 1 quantifies the ambiguity of parish names. It has been calculated from data held by the GB Historical GIS in its Administrative Unit Ontology (AUO), which has been matched to and extended from all census listings of parishes from between 1881 and 1971, so it contains many variant spellings of names.[18] The table below is limited to England, to exclude place names in the Welsh language, and to units legally defined as Civil Parishes, partly because coverage of pre-1881 censuses is less complete. The first numeric column covers all such units, while the second focuses on one particular census listing, for 1911, chosen because it was the first full transcription we made and, therefore, the most thoroughly checked.

The table clearly demonstrates great ambiguity across the country as a whole: even when we define toponyms as ambiguous only if they exactly match more than one parish (row 2), 25% of all parishes have ambiguous names, or 18% if we consider only the names used in 1911. Further, many other names are ambiguous if very small variations are allowed, as measured by the Levenshtein distance (rows 3 and 4): a Levenshtein distance of one means that one name can be turned into the other by adding, removing or changing one letter. The table also lists the 20 most common names in each sample (row 5), and it will be seen that most are short and end in 'ton', meaning 'farm' or 'hamlet' in Old English. This helps explain the Levenshtein results; for example, 'Marton' and 'Norton' have a Levenshtein distance of two.

In popular speech and writing, the most common way to disambiguate English village names has long been also to specify the county. There are two problems with this. The first is that within the period covered

[17] Overell, 'The Problem'.

[18] H. Southall, 'Rebuilding the Great Britain Historical GIS, Part 2: A Geo-Spatial Ontology of Administrative Units', *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, vol. 45, no. 3 (2012), pp. 119–34, doi:10.1080/01615440.2012.664101.

Table 1. The ambiguity of English Parish names

No.		All Civil Parishes		1911 Civil Parishes	
1.	Total number in England	15,598	Percentage	13,378	Percentage
2.	Share a name with at least one other parish	3,902	(25.0%)	2,453	(18.3%)
3.	Name within Levenshtein distance of 1 of another parish	7,085	(45.4%)	5,382	(40.2%)
4.	Name within Levenshtein distance of 2 of another parish	10,467	(67.1%)	8,477	(63.4%)
5.	Twenty most common ambiguous names (with frequencies)	Sutton (19) Newton (15) Middleton (14) Broughton (13) Norton (13) Preston (11) Denton (11) Carlton (10) Leigh (10) Upton (10) Horton (9) Walton (9) Milton (9) Wootton (8) Stoke (8) Hardwick (8) Weston (8) Marton (8) Elton (8) Thornton (8)		Sutton (17) Broughton (14) Middleton (14) Preston (14) Newton (13) Norton (12) Denton (11) Leigh (10) Upton (10) Milton (9) Marton (9) Thornton (8) Weston (8) Tunstall (8) Aston (8) Hardwick (8) Horton (8) Carlton (8) Bradford (8) Bolton (8)	
6.	Shares a name with another parish in the same county	1,041	(6.7%)	238	(1.8%)
7.	Shares a name with another parish in the same Poor Law Union/Registration District	270	(1.7%)	2	(0.0%)
8.	Shares a name with another parish in the same Local Government District	139	(0.9%)	0	(0.0%)

by the British censuses, at least three distinct systems of counties have been used: Ancient Counties, dating from the earliest times and the only counties used by the census up to 1841; Registration Counties, the primary units used from 1851 to 1911; and Administrative Counties, used from 1911 onwards, although significantly revised in 1973. What makes this especially confusing is that each system used a very similar set of county names. This is what creates examples such as Abington-in-the--Clay, Hertfordshire, listed in 1881, which was exactly the same parish as Abington Pigotts, Cambridgeshire, listed in 1951: this particular Hertfordshire was the Registration County, while this Cambridgeshire was an Administrative County; and 'in--the-Clay' and 'Pigotts' were two different ways to disambiguate the common name 'Abington'.

The other problem with counties is simply that, as shown in Table 1 (row 6), a substantial number of parish names were ambiguous even within a particular county, and our aim is to remove all ambiguity.

This means that full disambiguation requires that we also use the districts that

fell between counties and parishes in the hierarchy. While parishes were the dominant unit of village-level administration throughout the British Isles over a very long period, district-level geographies were more fluid. In different parts of England, the traditional districts were known as hundreds, wapentakes or wards, but from 1851, the census replaced these with a new set of Registration Districts, largely co-extensive with the Poor Law Unions established by the Poor Law Reform Act of 1834. These typically each consisted of a market town and its surrounding villages, but from 1911 onwards, the census prioritised a new local government geography, in which that market town would be an Urban District while the rural surroundings would be a Rural District of the same name. Larger towns would be Municipal Boroughs or County Boroughs, with greater powers, while London was divided into Metropolitan Boroughs.

Table 1 shows that there was still significant ambiguity across all Civil Parishes even if either Registration Districts or Local Government Districts are specified. However, this is almost entirely because there are many cases where the same settlement was covered by two different units, with the same name, at different dates. This is still ambiguous and explains why our detailed methods, described below, always include dates. The 1911 census listing is unusual because it includes both Registration Districts and Local Government Districts. Table 1 (row 8) shows that specifying the latter, in this one year, removes all ambiguity. Specifying the Registration District (row 7) leaves one ambiguous case: two different parishes called Eaton within Chester Registration District. There were, in fact, four parishes in the county of Cheshire called Eaton, and three of them are sometimes disambiguated as 'Eaton by Congleton', 'Eaton by Davenham', and 'Eaton by Tarporley': references to nearby towns, rather than to changing districts. Within the 1911 census listing, the two within Chester are disambiguated by being in different Registration sub-Districts, a level in the hierarchy not previously mentioned.

3. Disambiguating parish names in British census listings

Figure 1 shows part of one particular census table from 1891, for the county of Surrey, the Registration district of Epsom, Registration Sub-district of Carshalton and the parishes within it. It shows one example of 'Sutton', the most ambiguous parish name, as mapped in Figure 2.

The previous discussion has made clear why we need to consider all levels in these frequently changing hierarchies, and this

REGISTRATION DISTRICTS and Sub-Districts, and Civil Parishes.	AREA in Statute Acres.	HOUSES.						POPULATION.					
		Inhabited.	Uninhabited.	Building.	Inhabited.	Uninhabited.	Building.	Persons.		Males.		Females.	
		1881.			1891.			1881.	1891.	1881.	1891.	1881.	1891.
2. SURREY.													
30. EPSOM	43,882w	6,741	450	256	8,528	442	59	41,261	50,124	19,804	23,450	21,457	26,674
1. Carshalton	12,226w	3,176	290	189	4,190	275	32	21,118	26,108	9,932	11,763	11,186	14,345
Banstead †[a]	5,557w	305	17	–	438	8	–	3,826	4,560	1,703	1,932	2,123	2,628
Carshalton	2,926w	910	103	65	1,091	97	18	4,841	5,425	2,331	2,574	2,510	2,851
Sutton (*w.s.*)	1,836w	1,514	152	97	2,210	137	13	10,334	13,977	4,810	6,241	5,524	7,736

Fig. 1. Sutton, Surrey in the 1891 Census. Example from the parish level table. Source: Census of England and Wales. 1891. Area, Houses, and Population, vol. 2: Registration areas and Sanitary districts (London, HMSO, 1893), page 53 of Table 2

Fig. 2. Places called 'Sutton' in England and Wales. Source: Authors own work

led to the construction of our AUO gazetteer. This gazetteer is held in a relational database with four main tables which (1) identify over 90,000 administrative units, (2) list over 200,000 names for them, (3) record over 260,000 relationships between them and (4) document over 75,000 details of their exact legal status. These four are supported by many smaller tables defining types of unit, types of relationship and so on. The AUO is linked together by g_unit ID numbers, defined in the central master listing of units, and includes the sources of all information and as many dates of changes as possible. Figure 3 provides a simplified view of the main tables in the AUO with the 'units' table at the centre.[19]

Within this system, 'disambiguating' the name of a parish, or other unit, means associating it with one of these unique

[19] For further discussion of the details of the development of the AUO and this diagram, see: H. Southall, P. Aucott, 'Expressing History through a Geo-Spatial Ontology', *ISPRS International Journal of Geo-Information*, vol. 8, no. 8 (2019), p. 362, https://doi.org/10.3390/ijgi8080362.

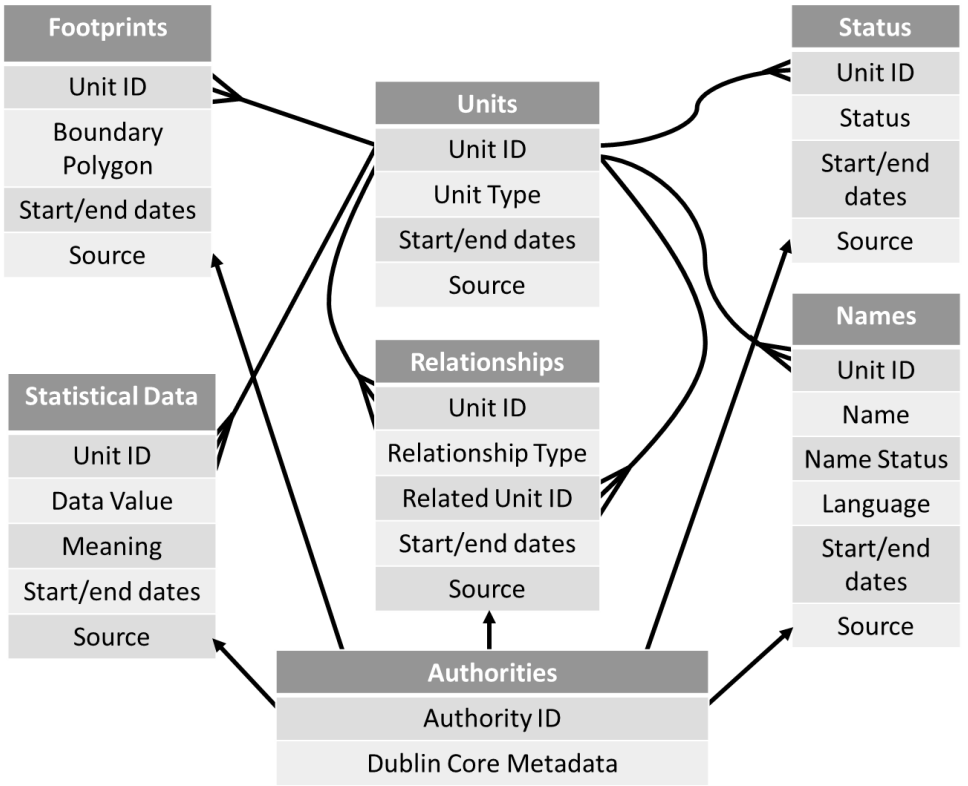

Fig. 3. Simplified Administrative Unit Ontology Data Model. Source: Authors own work

g_unit identifiers.[20] The AUO for Great Britain was initially constructed in 2002-3 from existing reference works: for England, from Frederick Youngs' *Guide to the Local Administrative Units of England*;[21] for Wales, from Melville Richards' *Welsh Administrative and Territorial Units*;[22] and for Scotland from an existing digital gazetteer of counties, parishes and burghs constructed by the Scottish Archives Network.[23] These sources provide systematic information on what units existed, their legal status and their hierarchical relationships, but the AUO was then greatly extended to meet our changing needs over time.

Within the names table of the AUO, each unit name identifies the language it is written in, whether it is the most recent officially recognised name for the unit 'Preferred' (of which there must be one), an earlier officially recognised name 'Official', an 'Alternate' one, an 'Abbreviated' one and so on, plus the source of the information and any known dates associated with that name. It is these unit names that we use to match against the geographical

[20] Southall, 'Rebuilding the Great Britain Historical GIS, Part 2'.

[21] F. Youngs, *Guide to the Local Administrative Units of England*, vol. 1: *Southern England* (Royal Historical Society, 1979), vol. 2: *Northern England* (Royal Historical Society, 1991).

[22] M. Richards, *Welsh Administrative and Territorial Units* (University of Wales Press, 1969).

[23] Scottish Archives Network, 'Scottish Archives Network (SCAN) Gazetteer', 2000, http://www.scan.org.uk/knowledgebase/index.htm. This website has now been superseded by 'ScotlandsPlaces', https://scotlandsplaces.gov.uk/.

```
update      ew1931_census_tab t set par_unit =
--      (1) Find the unique identifiers for the units:
(select     distinct u.g_unit
from        auo.g_unit u, auo.g_name n, auo.g_status s, auo.g_rel r
--      (2) Specify the query only applies where the unit type is Parish:
where       u.g_unit_type = 'PAR_UNIT' and
--      (3) Specify the dates of existence for the unit as a parish:
            (util.get_start_year(u.g_duration) <= 1931 or
            util.get_start_year(u.g_duration) is null) and
            (util.get_end_year(u.g_duration) >= 1931 or
            util.get_end_year(u.g_duration) is null) and
--      (4) Link to the names table:
            n.g_unit = u.g_unit and
--      (5) Specify what kind of name each has and the dates that name was used:
            (n.g_name_status = 'P' or n.g_name_status = 'O') and
            (util.get_start_year(n.g_duration) <= 1931 or
            util.get_start_year(n.g_duration) is null) and
            (util.get_end_year(n.g_duration) >= 1931 or
            util.get_end_year(n.g_duration) is null) and
--      (6) Specify the name should match:
            n.g_name = t.par_name and
--      (7) Link to the statuses table:
            s.g_unit = u.g_unit and
--      (8) Specify the status values the parish has and the dates they were used:
            (s.g_status = 'CP' or s.g_status = 'ExP' or s.g_status = 'PA') and
            (util.get_start_year(s.g_duration) <= 1931 or
            util.get_start_year(s.g_duration) is null) and
            (util.get_end_year(s.g_duration) >= 1931 or
            util.get_end_year(s.g_duration) is null) and
--      (9) Link to the relationships table:
            r.g_unit = u.g_unit and
--      (10) Specify the other unit it has a relationship with:
            r.g_rel_to = t.dist_unit
--      (11) Specify what kind of relationship and the dates that relationship existed:
            r.g_rel_type = 'IsPartOf' and
            (util.get_start_year(r.g_duration) <= 1931 or
            util.get_start_year(r.g_duration) is null) and
            (util.get_end_year(r.g_duration) >= 1931 or
            util.get_end_year(r.g_duration) is null)
--      (12) Specify that the parish unit not yet matched but the district unit is matched:
where       t.par_unit is null and
            t.dist_unit is not null and
            t.dup_par_flag < 3;
```

Fig. 4. Example code used for matching. Source: Authors own work

listing in the census parish tables to identify the individual administrative units.

While the AUO is an object of study in its own right, as shown in the previous section, the main reason for its construction was as a framework around which both census data and boundary polygons could be integrated. To match the names of the administrative units in a census listing such as Figure 1, we work systematically down the levels in the hierarchy, beginning with the counties, and working down through the districts to the parishes, and sometimes below. By working in this

systematic way, the relationships between the higher-level units and those below them in the hierarchy can be used to limit the available options for the smaller units.

All rows in the geographical census listing are checked against the unit names existing for the date of the Census within the AUO gazetteer to identify those that match. The example of code shown in Figure 4 is part of a much longer sequence of SQL queries that match the units within a Census listing using the names given, filtering down using the administrative unit hierarchy. The previous queries identify the Administrative Counties, then the Local Government Districts, after which this code matches the parishes that existed in 1931.

The query shown is divided up with numbered comments in italics to distinguish the different parts. Each part links to an AUO table and then specifies the matching against that particular table. For this query to work the name in the database table must exactly match the name as given in the census listing.

This approach works well for a significant proportion of the units listed in the census as they are unique. However, certain factors do affect the efficiency of the system. For instance, the spellings of names can vary, particularly in earlier censuses. Once the entire query has finished running and all units with 'Preferred' and 'Official' names matched, for those units where the name in the census listing was not matched, we would run the query again, but this time only on the remaining unmatched units using 'Alternate' names from the AUO.

Even with the logical progression of these matches, some unusual ones remain and have to be dealt with individually. For instance, sometimes a unit was officially created after the census was taken, but it still appeared in the geographical listing, like West Humberstone parish in Leicestershire which was created in 1892 but was listed in the 1891 Census. In other cases, an ancient unit was divided in two and split across a county boundary, so that both new civil parishes had the same name as the original and appeared in the same Registration sub-district in the Census. One example of this was Mollington parish, originally in Oxfordshire but after the split in 1889 also in Warwickshire.

Occasionally the way the information was presented changed, sometimes in conjunction with changes in boundaries. The Isles of Scilly were all listed as a single parish 'Scilly Islands' in the Census in 1881. Prior to that in 1871 the single parish of St Mary's was listed, but the statistics were divided up to provide individual data for each island. In March 1891 a Local Government Board Order led to each island becoming a separate civil parish, and thereafter the islands were listed separately as individual parishes. Even so there remain certain oddities in the census listings and other sources meaning automated matching of every unit is not possible. In these rare cases, we have to hard code the g_unit value match between the listing and the AUO.

Disambiguating names in the Census to clearly identify individual administrative units is essential to enabling any analysis of the statistics it contains through time. Whilst most of this identification can be automated once an extensive gazetteer like the AUO has been created, using progressively looser criteria to effect the match, it is unlikely to be comprehensive, and some manual decisions must still be made.

4. Locating places within travel writing

The previous section discussed sometimes locating over 20,000 geographical names in a census parish-level table, but with the large saving grace that the census tables were consciously designed to enable disambiguation through hierarchy once one understands the hierarchies. Historical travel writing poses entirely different

challenges, and what follows is not about automation, but instead explains why we have continued to use manual methods which depend on a historical geographer reading through each travel narrative.

The Great Britain Historical GIS and especially the website based on it, *A Vision of Britain through Time* ('Vision of Britain'), assembles 'geographical surveys of Britain'. Most obviously, that means the Census of Population, initiated in 1801, and the systematic topographic mapping of the Ordnance Survey, commenced in 1790. Various later sources, such as vital registration statistics, are almost as obvious, but what should be included for earlier dates? Some earlier statistical data sets exist at, approximately, the village level, but besides copyright issues with transcriptions, they mostly record just how much tax each locality paid.

This is one reason for our interest in 'travel writing', a very distinct literary genre which for Britain begins either with John Leland's *Itinerary*, from 1535 to 1543 but not published at the time, or William Camden's *Britannia*, first published in 1586. Although presented as narratives describing sequences of places, these early 'itineraries' do not seem to be accounts of particular journeys. By the eighteenth century, writers were more clearly describing actual tours and including dates. The other reason for our interest was the second author's previous research into nineteenth-century travelling artisans and political agitators, as recorded in working-class autobiographies and the radical press.[24] Our online collection includes some of these later writings, and in places, they are visibly mimicking earlier more literary travellers, as with *The Life and Rambles of Henry Vincent* (1839), written by the leading Chartist orator in the west of England.

Four of the best-known travel narratives, by William Cobbett, Daniel Defoe, Celia Fiennes and Arthur Young, were computerised for the project, but additional texts were then added by including public domain texts from Project Guttenberg, obtaining permission to use other transcriptions that were online elsewhere, and enhancing raw OCR output accessible via Google Books. Details of this assembly are not relevant here; the Guttenberg texts proved easiest to work with precisely because they were very 'plain text', with minimal formatting other than paragraph breaks. The current collection comprises twenty mainstream narratives plus six 'artisans and agitators', generally shorter.

Vision of Britain includes a Travel Writing section, which lists all narratives and allows each to be read chapter by chapter. However, our main aim was to make everything written about a particular town or village accessible from its 'place page', which also provides access to statistics and descriptive gazetteer entries. This was achieved by marking up the texts to include <placename> tags as defined by the Text Encoding Initiative (TEI). For example, this excerpt from the agricultural propagandist Arthur Young:[25]

> At Slabbard, in the way to Narbarth, rents are from 15s. to 20s. an acre

is marked up to become:

> At <placeName reg="Slebech" cnty="Pembrokeshire">Slabbard </placeName>, in the way to <placeName reg="Narberth"cnty="Pembrokeshire">Narbarth</placeName>, rents are from 15s. to 20s. an acre

[24] H. Southall, 'Mobility, the Artisan Community, and Popular Politics in Early Nineteenth Century England', in *Urbanising Britain: Class and Community in the Nineteenth Century*, ed. G. Kearns, Ch.W.J. Withers (Cambridge University Press, 1991), pp. 103–30; H. Southall, 'Agitate! Agitate! Organize! Political Travellers and the Construction of a National Politics, 1839–1880', *Transactions of the Institute of British Geographers*, vol. 21, no. 1 (1996), pp. 177–93.

[25] A. Young, *1776 Tour of South Wales and South Midlands, Selected from the Annals of Agriculture* (London School of Economics, 1932), https://www.visionofbritain.org.uk/travellers/Young/1 (accessed on 27 Dec. 2024).

In other words, we add a *regularised* form of the place name, in this example 'Slebech' for 'Slabbard', chosen as matching our gazetteer of British places, and then always add the name of a county, in this case 'Pembrokeshire', to provide disambiguation; this allows for potential extensions to the gazetteer adding ambiguity, as when a second 'Portsmouth', in Lancashire, was added. The marked-up text is then loaded into our Postgres database, but before it is presented on our website it is pre-parsed to replace the regularised name/county name pairings by the corresponding ID numbers from the gazetteer:

> At <placeName key="8820">Slabbard</place Name>, in the way to <placeName key="1118"> Narbarth</placeName>, rents are from 15s. to 20s. an acre

The pre-parser simultaneously builds a table, essentially a concordance, of all the place references, including the place ID, the exact location within the corpus of travel writing, and the particular form of the name used. That is linked to the gazetteer so that searches from the site's home page for 'Slabbard' will take users to the Slebech page.

Finally, as the text is included on the web page, it is run through a second parser, TagSoup, which converts it to pure HTML. The HTML includes both anchor points, so that users coming from place pages can be taken directly to the relevant place reference, and hyperlinks back to the place pages:

> <p>At Slabbard, in the way to Narbarth, rents are from 15s. to 20s. an acre

The above explains how we have integrated travel writing into the website via our gazetteer of British places, but says nothing about how we decided that 'Slabbard' equalled 'Slebech'; and note that a Google search for Slabbard leads to a suggestion we perhaps mean 'Svalbard'; to a Middle English Compendium at the University of Michigan which defines it as 'Someone slow or dull-witted'; to a Wikidata entry for the Dutch painter Karel Slabbaert; and, next, to our own place page for Slebech. With that exception, which obviously did not exist when the mark-up was done, there is nothing in the five pages of search results to connect the word 'Slabbard' to a place in Pembrokeshire; and as we will see this is not an especially extreme example.

The answer to the above question is that the connection was made by the authors, reading through the text and adding <placeName> tags manually, assisted by word processing editing features. The main basis for the decision is the sequence of places mentioned in the text. The text fragment used as an example above appears quite close to the start of Young's *A Tour in Wales*. Editing out much descriptive material, this reads as follows:

> OCTOBER 23, 1776, landed at Milford haven from Ireland ... To Haverford-West, the soil a rich reddish loam on slate and clay ... To Narbarth. Several cottages building in the Irish way, of mud with straw ... At Slabbard, in the way to Narbarth, rents are from 15s. to 20s. an acre; some rich meadows at 40s.

Milford Haven and Haverford West are well-known towns, and unambiguous; 'Narbarth' is an unusual spelling, but easily associated with a small town about 15 km east of Haverford West. Examining the road in between on a modern map, there are only three settlements larger than a farmstead: Slebech, Canaston, and Robeston Wathen, the first and last having churches.

In other words, we are dealing with a quite different kind of ambiguity from the previous section, and one far harder

to deal with through automation: while 'Sutton' matches to very many places, 'Slabbard' directly matches to none, and to make a match we have to allow for great variation in name forms. One reason is that the usual names of places have evolved substantially over time, as very extensively researched by the English Place-Names Survey. Two other factors are that travellers, by definition, lack local knowledge and may often be unclear about exactly where they are and that they may frequently only hear place names spoken, rather than seeing them written down. Arthur Young was an English traveller in Wales, and language and Welsh pronunciation may be an additional issue. Travellers often mention county names but were unlikely to know anything about the ephemeral systems of districts.

Once we read Young's whole narrative, the above example is straightforward to deal with. More problematic are place references that do not form part of the tour. This comes from Celia Fiennes's tour in 1698 and explains why, when in Carlisle, she chose not to proceed into Scotland:

> their miles are soe long in these Countrys made me afraid to venture, Least after a tedious journey I should not be able to get a bed I Could Lye in. It seemes there are very few towns Except Edenborough, Abberdeen and Kerk wch Can give better treatement to strangers, therefore for the most part persons yt 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.[26]

Edinburgh and Aberdeen are not too hard to identify, but where is 'Kerk'? 'Kirk' is Scots for church, so very many places contain it as part of their name, but none dominates. Falkirk? Kirkcaldy? This remains unresolved.

Fiennes never went any further into Scotland, but in other such cases, reading the whole narrative enables an editor to associate an off-route place-reference in one tour with a place actually visited on another tour. More straightforward ambiguity comes elsewhere in Fiennes' 1698 tour: there is no town or village in England usually called 'Norwitch', but Fiennes often adds a 't' to '-ich' names, so in the next quotation 'Ipswitch' and 'Norwitch' are obviously Ipswich and Norwich (while 'Berry' must be Bury St. Edmunds, 'Beckle' Beccles and 'Yarmouth' Great Yarmouth):

> To Beckle is 8 mile more wch in all was 36 miles from Ipswitch, ... This is a Little market town but its the third biggest town in ye County of Suffolke-Ipswitch, Berry and this ... At ye towns End one passes over the river Waveny on a wooden bridg railed wth timber and so you Enter into Norfolk ... Its from this town to Norwitch 12 miles, and its 10 to Yarmouth where they build some small shipps, and is a harbour for them and where they victual them.[27]

However, later in the tour the closeness to Manchester, and the references to the Earl of Warrington's estate at Dunham Massey and the county of Cheshire, mean that 'Norwitch' in the next quotation refers instead to the smaller town of Northwich:

> [After visiting Manchester] I went by Dunum the Earle of Warringtons house wch stands in a very fine parcke, ... Cross Little rivers, so to Norwitch wch is 14 mile. I Entred Cheshire 3 mile before I Came to ye town, its not very

[26] C. Fiennes, *Through England on a Side Saddle in the Time of William and Mary, Being the Diary of Celia Fiennes with an Introduction by the Hon. Mrs Griffiths* (Leadenhall Press, 1888), https://digital.library.upenn.edu/women/fiennes/saddle/saddle.html (accessed on 27 Dec. 2024).

[27] Ibid.

> Large, its full of Salt works the brine pitts being all here about.[28]

In the examples so far, it is clear which words are geographical names, but sometimes it is less obvious. Elsewhere in her 1698 tour, Fiennes refers to the spa town of Bath as 'the Bath', while 'the Cross bath' refers as usual to a particular bath:

> There is a very fine hall wch is set on stone pilllars wch they use for ye balls and dancing. This is the only new thing since I was at ye Bath before, Except the fine adornements on ye Cross in the Cross bath ... From the Bath I went westward to Bristol over Landsdown 10 mile, and passed thro' Kingswood.[29]

A particular issue is deciding whether geographical names refer to a place, or to a person by a place-based title: the Bishop of Chester, the Duke of Cumberland. This is a common challenge because these pre-1800 travellers were generally not staying in hotels, but rather using letters of introduction to obtain free accommodation from local aristocrats, who then appear in the narratives.

This made marking-up Boswell and Johnson's accounts of their tour of northern Scotland especially challenging, as within a single sentence the same word is used to refer both to a place, or an island, and to the 'laird', or main landowner, of the place:

> We had a very good ride, for about three miles, to Talisker, where Colonel M'Leod introduced us to his lady. We found here Mr Donald M'Lean, the young Laird of Coll (nephew to Talisker) [23 September 1773];
>
> Much time was lost in striving against the storm. At last it became so rough, and threatened to be so much worse, that Col and his servant took more courage, and said they would undertake to hit one of the harbours in Coll. [3 October 1773].[30]

In the first excerpt Boswell refers to 'Talisker' first as a place and then to its laird. In the second, 'Col' refers to Donald McLean, its laird, but 'Coll' to the island. The slight difference in spelling is insignificant as Johnson elsewhere refers to the island as 'Col'.

We should emphasise that we are not asserting that automating place name mark-up and disambiguation in these seventeenth and eighteenth-century texts is impossible, only that it clearly needed more than a competent database programmer. In practice, most of the time working on these texts was spent tidying up the output from optical character recognition systems, and the place name mark-up added some interest to that task.

5. Building unambiguous gazetteers for Ireland

Up to this point, 'disambiguation' has been misleadingly presented as being entirely about working with digital transcripts of historical documents, i.e. census reports and travel narratives, to match them to pre-existing gazetteers of administrative units and 'places'. As explained above, this is not wholly untrue for the parts of Great Britain, as the AUO does have a single main scholarly source for each of England, Wales, and Scotland. The initial gazetteer of 'places', which at its core is simply a list of preferred names and point coordinates, was created algorithmically from the AUO: first, each urban local government district was used to define a place, with a coordinate calculated as the mean centroid of associated boundary polygons;

[28] Ibid.

[29] Ibid.

[30] J. Boswell, *The Journal of a Tour to the Hebrides with Samuel Johnson, LL.D.*, 1784, https://www.visionofbritain.org.uk/travellers/Boswell (accessed on 27 Dec. 2024); S. Johnson, *A Journey to the Western Isles of Scotland*, 1775, https://www.visionofbritain.org.uk/travellers/Johnson (accessed on 27 Dec. 2024).

then as many other units as possible associated with each of these urban 'places' based on name similarity and proximity; and then a much larger set of rural 'places' was similarly created from the remaining parishes.

However, subsequent extensions and manual editing mean that what now exists is very different. For example, it became clear that Melville Richards' book essentially combined scholarly information about medieval Welsh units such as Commotes with listings of the units existing when he was writing in the 1960s: nineteenth-century Registration Districts and Poor Law Unions were absent, as was the substantially different local government district geography that existed before the County Reviews of the 1930s. Even when units were listed, the book did not include the many different spellings of Welsh place names encountered in census reports. With the benefit of hindsight, it would have probably been simpler to create a new Welsh gazetteer from scratch.

Precisely this has now been done for Ireland. For now, the focus is entirely on the period before the First World War, when the whole of Ireland was part of the United Kingdom. Our most important sources are the reports of the Census of Ireland, and especially the very detailed tabulations at parish-level published between 1821 and 1911, most of which we have now computerised. These tables cover first Ireland as a whole; then the four Provinces of Leinster, Munster, Ulster and Galway; then the 32 counties; then either or both of about 300 Baronies or, in later reports, about 160 Poor Law Unions; then about 2,500 parishes; and below these about 60,000 townlands, the smallest administrative sub-divisions, and varying numbers of "towns", described by the 1851 census as "a collection of twenty houses and upwards".

The largest problem here is that parish names are not unique within counties, let alone across Ireland; but many parishes were part of two or more baronies, and consequently their populations would be listed in several separate rows within each of the relevant baronies. The initial Irish AUO was therefore created instead from the listings in the *General alphabetical index to the townlands and towns, parishes and baronies of Ireland,* published with the reports to the 1861 census but based on the units listed by the 1851 census. From this, 2,421 parishes were added to the AUO, with 3,043 'IsPartOf' relationships to Baronies and 2,097 such relationships to Poor Law Unions.

As expected, this initial Irish AUO worked well with the actual 1851 parish tabulation, and only limited extensions were needed in matching to later tables. Further, while there was no modern authority list for historical Irish parishes, the Townlands.ie project had created digital boundary data for the parishes as they existed around the end of our period, which linked reasonably well to the AUO.

Working with the pre-1851 tabulations proved more problematic, probably reflecting the limited understanding census officials from England had of Irish geographies and place names. Firstly, very many parish names in 1821 were substantially different from those used in 1851, often seeming more anglicised. Secondly, a substantial number of parishes had completely different names, and could be identified only because another key source, Lewis's *Topographic Dictionary of Ireland*, listed both forms. Thirdly, while the 1821 and 1851 barony geographies are similar, many are listed as one barony in 1821 and as two half-baronies in 1851, or vice-versa. Fourthly, in 1821 the relationships between parishes and baronies was even more complex, so 419 additional 'IsPartOf' relationships had to be added to the AUO based on the 1831 parish table and its footnotes, and another 105 from the 1821 tabulation.

Given that the only census data reported for townlands were basic population totals, adding this very large number of units to the AUO would mostly just add confusion. Conversely, everything reported for parishes was also reported for the "Towns", and the latter relate to Ireland's small towns and villages far better than do the parishes, so 1,284 Towns have been added from the *General alphabetical index*. This was again complicated by many Towns being divided between more than one parish, and sometimes more than one barony. Another reason why 1821 parish names differed from 1851 was that they were more likely to match the main settlement, listed as a "town".

Previously basing "places" in Great Britain on administrative units had created substantial duplication which then had to be removed manually: for example, the Civil Parishes which were divided into Urban and Rural portions following the 1894 Local Government Act, such as Ledbury in Herefordshire; or small towns containing multiple ancient parishes, as with Sawtry St Andrew and Sawtry St Judith in Huntingdonshire.

For Ireland, therefore, we began with the 3,347 entries in Lewis's Topographic Dictionary, this count excluding 593 cross-reference entries for places with alternative names. At the time of writing, we have defined a total of 3,631 Irish places, and of these 3,081 (84.9%) are linked to a Lewis entry, and a further 513 (14.1%) lack a Lewis link but are linked to one of the shorter entries in Bartholomew's *Gazetteer of the British Isles* (1887), so only 37 places lack any descriptive text.

Having grounded our places in these contemporary descriptive gazetteers, we then systematically cross-matched them with administrative units in the AUO. Again, at the time of writing 3,365 Irish places (92.7%) are linked to at least one administrative unit, including 2,422 (66.7%) linked to parishes, 1,163 (32.0%) linked to towns, and 430 (11.8%) linked to both a parish and a town. The 266 places not linked to units are diverse, including 60 whose descriptions mention 'seat', meaning a large estate often belonging to aristocrats, and 38 mentioning 'island'.

One reason for including them was to enable our collection of travel writing to be linked into the system, and especially the lengthy tour of Ireland in 1776 by Arthur Young; note that his visit to Wales, discussed in the previous section, followed directly from his four months in Ireland. Youngs generally stayed not in inns but at the country seats of various gentlemen, and while he wrote very extensively on Irish agriculture, he also described many mountains, lakes and so on. For example, he spent several days in and around Killarney, and this led to four mountains, two islands and one peninsular being added as "places", generally linked to entries in the Bartholomew gazetteer.

The largest challenge in extending the 'places' gazetteer to Ireland was adding the point coordinates required by the data model. Coordinates for 2,240 (61.7%) were found manually, the most important sources being Wikipedia[31] (1,146) and GENUKI,[32] created cooperatively by family historians (626); this research also enabled us to link out to those sites. A further 281 places were located on Ordnance Survey maps of Ireland. These included many of the 'towns', especially those which did not have the same name as the parish containing them; given these were often tiny settlements, they could be hard to find. The remaining places were given point coordinates based on the centroids of linked units, generally parishes.

Having extended our gazetteer of administrative units and places to Ireland, how have historical sources been linked to them,

[31] Wikimedia Foundation Inc., 'Wikipedia', 2024, https://www.wikipedia.org/.

[32] GENUKI, 'GENUKI', 2015, https://www.genuki.org.uk/.

avoiding ambiguity? As explained, the Irish part of the AUO was constructed not directly from census reports but from the 1851 Index, which unsurprisingly proved an excellent match to mid-century census listings; and the extensions to include 1821 and 1831 have already been discussed. There are many examples of two parishes with the same name in the same county, but the nineteenth-century census listings always include the baronies or poor law unions containing each parish, and so avoid ambiguity.

The Lewis gazetteer also generally includes the barony or baronies in its entries for parishes. Linking in travel writing is generally more problematic, not so much following the traveller's route as identifying places unambiguously in the mark-up, given that the parsing system allowed for only a place name followed by a 'container', generally a county. Therefore, for Ireland, we systematically eliminated ambiguity in the places gazetteer by, where necessary, including the barony name within the master place name. This was needed for 68 places; for example, in County Wicklow, which had 91 places in all, we had 'Kilbride in Arklow' versus 'Kilbride in Lower Talbotstown', and 'Kilcommon in Ballinacor' versus 'Kilcommon in Newcastle'.

6. Conclusion

We must leave it to data scientists to discuss whether and how the travel writing discussed above could now be marked up by automated methods. Working manually on the mark-up took many days, but this was spread over several years; and to a historical geographer it was not entirely a hardship to have to carefully read these classic writings. We also emphasise that our concern with user complaints about any inaccuracies is far from hypothetical as complaints in the website's early years were frequent.

Working with the census parish-level tables had to be automated, as a single parish table for England and Wales contains, counting counties, districts and sub-districts, more toponyms than we have matched in the entire corpus of travel writing. However, the deterministic methods we have developed are grounded in a detailed understanding of evolving administrative geographies, and an extremely detailed knowledge base, the Administrative Unit Ontology, which was initially created by computerising traditional scholarly works, but was then extended by adding variant names and alternative hierarchies as we encountered them in the census reports.

This goes far beyond what could have been achieved through personal local knowledge alone, which is often limited to a very restricted geographical area: the only true local knowledge in this paper concerns the different parts of Colwall. Physical original sources are bound by the information contained within them, where and why they were created and how they survived and have been stored. In contrast, the consolidation of the information in the digitalised versions of these sources adds context to them. This allows for both greater cohesion through the internal disambiguation of places and more flexibility through the external presentation of the material, either as text through place-names or spatially through maps.

All of the work described here contributed to the construction of the Vision of Britain system, which is both a very large statistical database and mapping system and a large assembly of travel writing; but assembling all this content into a single system means it is also an enormously rich gazetteer, including toponyms assembled from all these sources.[33] That, in turn, makes it a very powerful tool for identifying and disambiguating toponyms in other sources, both names already in the

[33] The gazetteer beneath *A Vision of Britain* is directly accessible and searchable via https://www.visionofbritain.org.uk/expertsearch.

system and new variants. Critically, the overall table of names currently includes over 130,000 names, but all their associations with the over 20,000 places have been decided by an academic researcher, not an algorithm.

The need for disambiguation of place-names in our system is essential to both the smoothness of the user experience and the functionality of the website. While place-names in our statistical material takes the form of structured lists rather than free text, what we display has to be credible and clear, with no room for ambiguity in how we present our results to our users. Information is made less ambiguous by only offering users clearly defined places as far as possible. The considerable fluidity of British and Irish place-names over time makes the task far more complex than for nations with less historical change. By comparing different historical sources against each other we have built a continuous chronological sequence for British place-names, allowing us to ensure consistency in the quality of our output. ■

References

Ardanuy M.C., Sporleder C., 'Toponym Disambiguation in Historical Documents Using Semantic and Geographic Features', *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage* (2017), pp. 175–80.

Boswell J., *The Journal of a Tour to the Hebrides with Samuel Johnson, LL.D.*, 1784, https://www.visionofbritain.org.uk/travellers/Boswell.

Buscaldi D., Rosso P., 'Map-Based vs. Knowledge-Based Toponym Disambiguation', *Proceedings of the 5th Workshop on Geographic Information Retrieval* (2008), pp. 19–22.

Fiennes C., *Through England on a Side Saddle in the Time of William and Mary, Being the Diary of Celia Fiennes with an Introduction by the Hon. Mrs Griffiths* (Leadenhall Press, 1888), https://digital.library.upenn.edu/women/fiennes/saddle/saddle.html.

Hill L., 'Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints', in *Research and Advanced Technology for Digital Libraries: 4th European Conference, ECDL 2000 Lisbon, Portugal, September 18–20, 2000 Proceedings 4* (Springer, 2000), pp. 280–90.

Johnson S., *A Journey to the Western Isles of Scotland*, 1775, https://www.visionofbritain.org.uk/travellers/Johnson.

Kauppinen T., Henriksson R., Sinkkilä R., Lindroos R., Väätäinen J., and Hyvönen E., 'Ontology-Based Disambiguation of Spatiotemporal Locations', *Proceedings of the 1st IRSW2008 International Workshop on Identity and Reference on the Semantic Web, Tenerife, Spain, June 2, 2008* (2008), https://www.researchgate.net/publication/220853817_Ontology-based_Disambiguation_of_Spatiotemporal_Locations.

Machado I.M.R., Odon de Alencar R., de Oliveira Campos Jr R., and Davis Jr C.A., 'An Ontological Gazetteer and Its Application for Place Name Disambiguation in Text', *Journal of the Brazilian Computer Society*, vol. 17, no. 4 (2011), pp. 267–79, doi:10.1007/s13173-011-0044-4.

Overell S., 'The Problem of Place Name Ambiguity', *SIGSPATIAL Special*, vo. 3, no. 2 (2011), pp. 12–15.

Overell S., Magalhães J., and Rüger S.M., 'Place Disambiguation with Co-Occurrence Models', *CLEF (Working Notes)* (2006), https://ceur-ws.org/Vol-1172/CLEF2006wn-GeoCLEF-OverellEt2006.pdf.

Pouliquen B., Kimler M., Steinberger R., Ignat C., Oellinger T., Blackler K., Fuart F., Zaghouani W., Widiger A., Forslund A.-Ch., and Best C., 'Geocoding Multilingual Texts: Recognition, Disambiguation and Visualisation', *Language Resources and Evaluation Conference (LREC)* (2006).

Rauch E., Bukatin M., and Baker K., 'A Confidence-Based Framework for Disambiguating Geographic Terms', *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References* (2003), pp. 50–54.

Richards M., *Welsh Administrative and Territorial Units* (University of Wales Press, 1969).

Santos J., Anastácio I., and Martins B., 'Using Machine Learning Methods for Disambiguating Place References in Textual Documents', *GeoJournal*, vol. 80 (2015), pp. 372–92.

Smith D., Crane G., 'Disambiguating Geographic Names in a Historical Digital Library', in *Research and Advanced Technology for Digital Libraries: 5th European Conference, ECDL 2001 Darmstadt,*

Germany, September 4-9, 2001 Proceedings (Springer, 2001), pp. 127–36.

Southall H., 'Agitate! Agitate! Organize! Political Travellers and the Construction of a National Politics, 1839–1880', *Transactions of the Institute of British Geographers*, vol. 21, no. 1 (1996), pp. 177–93.

Southall H., 'Mobility, the Artisan Community, and Popular Politics in Early Nineteenth Century England', in *Urbanising Britain: Class and Community in the Nineteenth Century*, ed. G. Kearns, Ch.W.J. Withers (Cambridge University Press, 1991), pp. 103–30.

Southall H., 'Rebuilding the Great Britain Historical GIS, Part 2: A Geo-Spatial Ontology of Administrative Units', *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, vol. 45, no. 3 (2012), pp. 119–34, doi:10.1080/01615440.2012.664101.

Southall H., Aucott P., 'Expressing History through a Geo-Spatial Ontology', *ISPRS International Journal of Geo-Information*, vol. 8, no. 8 (2019): 362, https://doi.org/10.3390/ijgi8080362.

Sultanik E.A., Fink C., 'Rapid Geotagging and Disambiguation of Social Media Text via an Indexed Gazetteer', *Proceedings of the 9th International ISCRAM Conference – Vancouver, Canada, April 2012*, ed. L. Rothkrantz, J. Ristvej, and Z. Franco (2012), https://idl.iscram.org/files/sultanik/2012/212_Sultanik+Fink2012.pdf.

Volz R., Kleb J., and Mueller W., 'Towards Ontology-Based Disambiguation of Geographical Identifiers', in *I3: Identity, Identifiers, Identification* (2007), https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=a0fc2bba69e48a3bf310e882bdd9b8f8484b98c8.

Young A., *1776 Tour of South Wales and South Midlands, Selected from the Annals of Agriculture* (London School of Economics, 1932), https://www.visionofbritain.org.uk/travellers/Young/1.

Youngs F., *Guide to the Local Administrative Units of England*, vol. 1: *Southern England* (Royal Historical Society, 1979); vol. 2: *Northern England* (Royal Historical Society, 1991).

Internet sites

GENUKI, 'GENUKI', 2015, https://www.genuki.org.uk/.

Great Britain Historical GIS Project / University of Portsmouth, 'A Vision of Britain through Time', 2003–24, https://www.visionofbritain.org.uk.

Historic Environment Scotland, the National Records of Scotland and the National Library of Scotland, 'ScotlandsPlaces', 2020, https://scotlandsplaces.gov.uk/.

Scottish Archives Network, 'Scottish Archives Network (SCAN) Gazetteer', 2000, http://www.scan.org.uk/knowledgebase/index.htm.

Wikimedia Foundation Inc., 'Wikipedia', 2024, https://www.wikipedia.org/. ■

Ujednoznacznianie nazw geograficznych w historycznych brytyjskich spisach ludności i pismach podróżniczych

Streszczenie

Teksty historyczne i raporty statystyczne prawie zawsze zawierają nazwy geograficzne lub toponimy, a nie współrzędne, mapowanie zatem wymaga powiązania z jakąś formą wykazu nazw geograficznych. Historyk musi zdecydować, gdzie znajdują się opisywane miejsca, jaką lokalizację powiązać z konkretnym toponimem i czy obiekty o tej samej nazwie odnoszą się do tego samego miejsca. Niniejszy artykuł przedstawia dwa uzupełniające się studia przypadków.

Pierwszy dotyczy ujednoznacznienia nazw parafii pojawiających się od 1801 r. w brytyjskich raportach spisowych. Nowatorska analiza nazw parafii angielskich pokazuje, że większość z tych nazw była niejednoznaczna, co oznacza, że dwie lub więcej parafii miało takie same lub bardzo podobne nazwy. W większości wypadków niejednoznaczność ta jest usuwana poprzez podanie hrabstwa, ale pełne ujednoznacznienie wymaga również określenia pośrednich „okręgów". Ponieważ podział administracyjny hrabstw i okręgów uległ znacznym zmianom w ciągu ostatnich 200 lat, kompleksowe ujednoznacznienie wymagało z kolei dopasowania do wielohierarchicznej ontologii jednostek administracyjnych (AUO), w ramach której poszczególne parafie zachowały swoją tożsamość, podczas gdy ich nazwy i pozycja hierarchiczna ewoluowały.

Drugie studium to identyfikacja toponimów za pomocą wykazu nazw geograficznych

występujących w wyjątkowo dużej kolekcji obejmującej ponad trzysta lat historycznego piśmiennictwa podróżniczego; wykaz ten opracowano po części na bazie XIX-wiecznych opisowych wykazów miejsc. W tym przypadku poszczególnym odniesieniom do miejsc brakuje hierarchicznego kontekstu, a podróżnicy często niedokładnie podawali toponimy. Z tego powodu opracowanie takiego wykazu wymaga dokładnego prześledzenia trasy podróżnika, aby zidentyfikować kolejno odwiedzane przez niego miejsca.

Zamiast łączyć źródła z istniejącymi listami miejscowości, AUO i wykaz miejsc były opracowywane równocześnie z dopasowywaniem toponimów i uzupełniane o warianty toponimów ze źródeł. Finalna część artykułu opisuje ostatnie prace mające na celu rozszerzenie obu wykazów nazw miejscowych o Irlandię, prowadzone w dużej mierze w oparciu o raporty ze spisów powszechnych, pisma podróżnicze, a także dwa XIX-wieczne spisy nazw miejscowych. Efektem końcowym jest wyjątkowo bogata struktura danych, integrująca różnorodne źródła historyczne wokół identyfikatorów miejsc, które złożyły się na bazę internetową „A Vision of Britain through Time". ■

Paula Aucott – a Senior Research Associate who works on the Great Britain Historical GIS Project, based in the School of the Environment and Life Sciences at the University of Portsmouth. Her main interests focus on historical GIS, historical land use, gazetteers and administrative unit ontologies. Her publications arising from the project cover a range of topics, with the most recent pieces discussing urban landscape features on maps from around 1900, and redistricting historical election statistics to provide context for the latest UK general election.
(Paula.Aucott@port.ac.uk)

Paula Aucott – starszy pracownik naukowy pracujący nad projektem Great Britain Historical GIS Project opracowywanym w Szkole Środowiska i Nauk o Życiu na Uniwersytecie w Portsmouth. Jej główne zainteresowania badawcze koncentrują się na historycznym GIS, historycznym użytkowaniu gruntów, wykazach nazw geograficznych i ontologiach jednostek administracyjnych. Autorka wielu publikacji powstałych w ramach projektu, w tym najnowszych, podejmujących dyskusję nad cechami krajobrazu miejskiego na mapach z około 1900 r., a także analizę historycznych statystyk wyborczych w celu ukazania kontekstu dla ostatnich wyborów powszechnych w Wielkiej Brytanii.
(Paula.Aucott@port.ac.uk)

Humphrey Southall – professor of Historical Geography in the School of the Environment and Life Sciences at the University of Portsmouth. He has researched the history of Britain's north-south divide, building the Great Britain Historical GIS to assemble and map a mass of information about Britain's changing localities over the last three centuries. That led to the creation of the website 'A Vision of Britain through Time', giving the general public access to all this information. He has also written extensively on the organisation of historical geographical information and 'rich gazetteers'.
(Humphrey.Southall@port.ac.uk)

Humphrey Southall – profesor geografii historycznej w Szkole Środowiska i Nauk o Życiu na Uniwersytecie w Portsmouth. Badacz historii podziału Wielkiej Brytanii na północ i południe poprzez opracowywanie Great Britain Historical GIS, zbierającego i mapującego informacje o zmieniających się miejscowościach Wielkiej Brytanii w ciągu ostatnich trzech stuleci, co zaowocowało powstaniem strony internetowej *A Vision of Britain through Time*, umożliwiającej społeczeństwu dostęp do wszystkich tych informacji. Autor wielu publikacji poświęconych organizacji historycznych informacji geograficznych i szczegółowych wykazów nazw geograficznych.
(Humphrey.Southall@port.ac.uk)