Given and Family Names as Global Spatial Data Infrastructure

Oliver O'Brien and Paul Longley

4.1 Introduction

This chapter outlines an ongoing Consumer Data Research Centre (CDRC) project that has produced a large (c. two billion record) global database of people's names, together with the approximate locations of their bearers. Although in part a 'hobby' project, the work is being reconfigured into demographic profiles of people over a full range of scales. We discuss the value and provenance of different data sources, data extraction techniques and the tools used to assemble a truly global database. We also present illustrative results from the resulting dataset, at the global scale as well as for selected countries. Case studies are used to examine some individual countries, where simple analysis of publically available data samples can reveal internal migration patterns and sub-national variations in the popularity of both given and family names.

In important respects a name is at the same time the purest and most widely used form of personal data. It is a characteristic of a person that is typically assigned at birth and often not changed or adulterated during the bearer's lifespan except for reasons of marriage. Names data are shared between individuals for many reasons core to social organisation, and sharing often follows established social patterns. Names are ubiquitous across cultures and throughout the world — nearly everyone has a name and it is rarely if ever assigned at random. A person's full (given/fore- plus family/sur-) name may thus provide a direct indicator of its bearer's gender, ethnicity and religion. Changing fashions in given naming practices often render given names a reliable indicator of age.

Additionally, most names have geographic origins, some of which may be very specific and localised. The naming practices of most societies provide for inheritance of family names, and thus comparison of

present residence with historic name origins makes it possible to trace the probable migration histories of many individuals and their blood lines. It is even possible to establish links between long settled populations and their genetic make-up (www.peopleofthebritishisles.org). The accuracy with which this may be done does of course depend upon the event histories of the individuals whose names are inherited through the generations, and historically these have been overwhelmingly male – yet local marriage practices throughout history do in practice retain this signal through the generations.

Finally, the fact of geographic concentrations of clusters of surnames means that names may be associated indirectly with the economic fortunes of the areas in which their bearers live. Inheritance, along with the intergenerational inheritances of human and social capital, may also mean that a name provides clues to economic and social standing. This tendency may be reinforced if there are social dimensions to given naming practices within broader cultural, ethnic and linguistic groups.

Linking given names and family names together, and clustering, produces groups of names typically sharing demographic traits. Studying such groups, with appropriate control data, allows the geographer to assign profiles which then can be used with similar names to infer similar demographic characteristics.

The clustering process is dependent on having a large and geographically representative pool of given/family name pairs. This chapter describes an ongoing CDRC project that is gathering many name pairs, across the world, for such clustering purposes, as well as for a similar probabilistic classification process based on the fact that many names have strong country (and indeed intra-country) geographic profiles that are retained to this day, even in an era of greater increasing global mobility.

4.2 The Worldnames 2 project

The Worldnames 2 project is an outcome of a series of research grants and private initiatives stretching back to 2003 (see www.onomap.org), and attempts to collect and assemble as many name pairs as possible until global coverage is achieved. The names of resident individuals are associated with the country (and often region or locality) in which they reside. The results can be mapped in order to portray the geographic distribution of names at a range of scales, and can be analysed in order to link names and groups of names to places. It follows on from the Worldnames 1 project, which presents a website showing name distributions in around 25 mainly western countries, by expanding to cover almost every country of the world. A secondary aim is to allow public dissemination of the enlarged dataset via a more modern website than the existing Worldnames 1 platform (which is around 10 years old). The data are mainly collected from freely available governmental, administrative and other public datasets from the countries concerned – although some of the data have been obtained under licence with attendant restrictions upon reuse. The project is in an ongoing phase and this chapter describes the work to date, as well as outlining some early results and presenting a number of country-specific case studies.

4.3 Objectives, execution principles and simplifications

The general objective of the ongoing project is to obtain as many name pairs as possible, on a country-by-country basis that can be deemed to represent the established population of the country. In some senses the objective is to create a demographic framework for the world, in terms of the personal attributes that can be inferred from naming conventions at national, regional and local levels. We are in the process of creating a website that will

enable users to examine these traits of any name that is part of our dataset. Users will also be able to interrogate the data in order to identify the most prevalent surname by country lists or intra-country distributions of popular single-country-origin names.

As such, the project is clearly a Big Data project that is vulnerable to the vagaries of Big Data sources that are discussed at various points in this volume. Some of the data sources are acquired under licence, with restrictions upon how they may be redistributed, particularly those that formed part of the original Worldnames 1 project. The countries that have formed the focus of our renewed attention on the project are openly available on the web, without needing a login or subscription to access, and from the original source, rather than from other consolidator sites with related foci of interest. We also exclude datasets whose custodians did not, in our opinion, intend the data to be made available for wide public use, albeit in aggregate form. These judgments are inevitably subjective and our intention is to avoid any legal infringements arising from reuse of data for new purposes. In particular, we have avoided using data published by third parties without the consent of the original owner to this end. From this standpoint, newspaper republishing of time-restricted electoral lists would be considered to be valid but database dumps, obtained as a result of a breach of security or insider leaks, would not be used. This distinction is not always clear cut, and require decisions to be made on a case-by-case basis – for example, data obtained from the WikiLeaks service and similar investigative journalism projects, or those where the original source and authority to publish is unclear. To simplify processing a vast array of diverse datasets, a number of simplifications are applied. The western-style naming convention of a given (assigned) first name followed by a (typically hereditary) family name is assumed, with other name structures (e.g. Spanish double surnames) simplified. The 'Western Latin' alphabet is used, with accents and capitalisations removed and the only non-alphabetic characters allowed are apostrophes and dashes – these being combined anyway with pure-alphabetic variants. This is necessary to accommodate the inconsistent ways which names are stored on the official records are typically used in the project. For example, MacDonald can appear, in different datasets across different countries, as Mac Donald, MACDONALD, Mac.Donald and Mac-Donald. Other non-alphabetic characters, such as spaces and underscores, are replaced or removed as judged appropriate.

It is acknowledged that, with a project of this scale, using hundreds of diverse datasets, such simplifications will potentially obscure helpful demographic information; this is minimized where practical. We have retained the names in the original forms captured, however, in order to allow the incorporation of other accents in future spin out projects from the research.

4.4 Data acquisition and processing methodology

4.4.1 Search-based initial discovery

To ensure a reasonable level of quality and a high geographical and demographic representation for each country are maintained, we carry out the data collection manually, rather than creating a 'bot' or 'spider' to crawl the web automatically. This also presents opportunities to discover additional unindexed datasets with intelligent URL modification by the investigator.

This means that, for each of the ~200 countries of the world, a different collection process is employed, built up by starting from a set of common principles detailed below, but then refined as name data are discovered from the current active country.

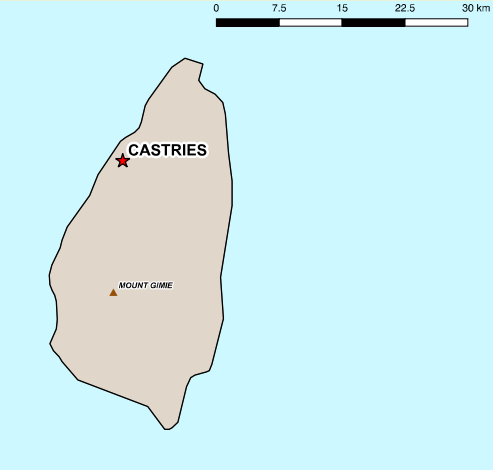


Figure 4.1
Map of St Lucia. St Lucia is approximately 40km in length (north to south) and 20km in width (east to west).

Case Study 1: Saint Lucia

Saint Lucia is an island nation in the Caribbean with a population of approximately 186,000 people (Figure 4.1). Our names data come from the polling lists published by the Saint Lucia Electoral Department at www.electoral.gov.lc/polling-list. Saint Lucia's top-level administrative areas are known as quarters, the constituencies are based on these quarters but with a number split or merged, to make 17 in total. These are then each further split into between 3 and 9 polling divisions, or precincts. The electoral data for each of the 84 precincts are listed on the website as paged tables, with a POST query needed to access each page.

The data listed on the tables include the given name, family name, street name, constituency, unique registration number, precinct and gender.

A python script was used to send POST queries and download the HTML tables, and extract the data from them using regex into a CSV file. The data are believed to be relatively up-to-date, as the most recent election was in 2016. 162,025 records were extracted in this way, at first glance matching well with official summary information for the 2016 election suggesting that there were 161,883 registered voters. However, a large number of duplicated records were found – where the same record would appear on multiple pages in a table for a single precinct. On de-duplicating, 129,685 records remained, representing around 70% of the 2016 UN estimate of 186,383 people. The reasons for this large discrepancy are not clear, but, if the official figures are at fault, it would go some way to explaining the apparently low turnout of 57% and that the numbers of reported registered voters have increased at a much larger rate than the population in general, since 2000.

70% of the 2016 estimate is a plausible percentage, as electoral lists are not population lists – they typically exclude young people and foreign residents. Voters for a general election in Saint Lucia must be at least 18 years of age and either a Saint Lucian or a Commonwealth citizen who has resided in Saint Lucia for at least seven years.

Because of the readily available geospatial boundary data for Saint Lucia's quarters, and the relatively small population of the nation as a whole, it was decided to sub-divide the population by a single level – quarters, but merged where the constituencies go across quarters, for the Worldnames 2 project – rather than further

Figure 4.2
Distribution of Charlemagne surname in St Lucia, with darker shading indicating higher proportions of the population bearing this surname.

dividing into constituencies, quarters or precincts, the former of which are not sufficiently different from the merged quarters and the latter of which have some very small populations. The constituencies and precincts also do not have readily available boundaries available.

This results in nine geographical areas, with the electoral population varying from 4,962 to 45,980.

The data were considered to be relatively 'clean', because of the direct extraction from HTML tables, rather than a structure-recreation approach that is often needed when extracting from PDFs.

As is characteristic for many Caribbean countries, Western given names have come into usage as family names. In Saint Lucia's case, the most common given family names show this trait: Joseph (4.6% – the most popular), Charles (2.1%), James, Alexander and Henry. There are 10,074 distinct family names. 7,119 of these are most common in Saint Lucia, compared to other countries for which there is currently data in the Worldnames 2 database. Hippolyte and Mathurin are two popular family names in Saint Lucia that are relatively unusual elsewhere.

Looking at the names data split by the merged quarters, a number of popular family names have strong variations through the nation. For example, Charlemagne is more popular on the western side of the island than the eastern side (Figure 4.2). The 100 people with this family name on the polling list for Choiseul quarter represent a frequency of 1.7%, while the 11 in Dennery quarter equate to frequency of just 0.12%. However, as these are small numbers in total, this may just be a statistical coincidence.

A small amount of logic, however, can be shared across countries. A number of countries in West Africa, for example, use the same off-the-shelf portal software for their government or public service websites, and so file locations discovered during the search for data for one country, can then be reapplied to additional countries using the same software.

The initial stage is to perform a simple search, typically using Google Search, for 'obvious' open-access lists of the greater part of a country's population. These may take the form of public versions of electoral rolls or civil registry lists. These may be posted both on a country's official portal, but also occasionally republished by citizens on private websites, for example the Chilean electoral roll can be freely obtained on disk and implicitly republished privately, but is not itself published online by the electoral authorities.

If a comprehensive list is not obtained, then it is necessary to search using distinctive surnames and then full names for any given country. It is necessary to 'seed' the search with certain key names which are popular in the country in question, but ideally not in neighbouring countries or other jurisdictions. To do this, reliable pan-national websites containing lists of famous people from a country, for example from Wikipedia, were used, as well as government and pan-national database-driven websites of elected government ministers or national election candidates – one type of data that are nearly universally published and that a number of separate projects, such as IFES Election Guide and International IDEA, are aiming to catalogue and maintain. The latter project also contains some information about the availability of online electoral rolls for each country and its approach to open data, including direct links to them where available.

As a rule of thumb in our research we deemed that at least three distinctive 'seed' surnames were required for any country. Google Search queries were then carried out using these surnames in conjunction with various combinations of numerical, country filter, data format and/or list keywords. Top-level country filters on Google Search were used, along with gov.xx and edu.xx second level domain filters (xx here the country's top level domain). These filters restricted results to being from subdomains of the domains specified and, due to the nature of Google Search's indexing, this more specific search often revealed additional results of interest. Format keywords can narrow large numbers of results returned to ones likely in the form of a downloadable, processable, list, for example 'pdf', 'xls' or 'xlsx'. The CSV format is probably the simplest and easiest list format to parse but is little used by non-technology focused websites. A small number of useful sources were found in the more modern JSON file format. Adding sequential numbers, e.g. 1345 1346 1347 1348 can both reveal list-focused results, and ones with a likely population of (in this case) well over a thousand names. Finally, the inclusion of other key words in the search for distinctive document classes can also be useful – as with 'cedula' (national identification document) for Spanish-speaking countries. Other more generic words were also useful in our searches, e.g. 'first name', 'given name', 'forename', 'last name', 'family name', 'surname', 'ID number', or 'candidate'. Translating these into different languages (typically using Google Translate) was also useful.

Somewhat counter intuitively, increasing the numbers of names in a query led to more search results being returned. This is another example, as mentioned above, of the heuristics applied in Google Search and other search engines, and the ways in which key words for websites are stored in the internal indices of the search engines.

Where surnames alone did not reveal useful datasets, use of distinctive individual full

names was useful, since this focused searches on websites with a record for that person amongst a larger list of bearers of less distinctive or unusual names, or with a search function or directory index (as discussed below) that revealed additional data files. As a general point, we avoided searches using famous names since these directed focus away from general population names.

It was often the case that, once identified, a relevant dataset offered coverage of only a (regional or sectoral) part of the population. In such instances the issue of coverage was addressed by amending the relevant part of the URL until full population coverage had been achieved – for example after using every government regional identifier within a given country. In some circumstances this was achieved by identifying the parent directory of a relevant dataset. We became used to anticipating abbreviations and invoking other trial and error procedures in this process.

The Internet Archive was also useful for retrieving files that no longer existed at their original locations but which were still revealed through stale indices in search engines and related weblinks. Additionally, websites occasionally change their domain names but fail to provide forwarding links from the old domain. Sometimes, missing files referred to from a Google search or external web links can be retrieved simply by modifying the domain name to the new one.

As well as Google, specialized search engines, such as Docs Engine, are useful. In particular, we found Docs Engine good at revealing lists of PDFs, Microsoft Excel (XLS) and Microsoft Word files which are not readily found with Google Search – particularly when searching for a single non-celebrity full name.

Where the information was available as numerous web pages rather than as documents, simple Python scripting was used to automate retrieval of large numbers of webpages, supplying appropriate GET/POST parameters on a consistent, sequential or known list basis, and simple processing and name extraction of the resulting HTML files retrieved. Where webpages listed a large number of documents, a bulk downloader browser extension ‘uSelect iDownload’ was used.

4.4.2 Targets

A minimum target number of names to be harvested was calculated for each country. In important respects, Web-based research often remains unreconciled with respect to established scientific apparatus of sampling and inference, in that:

1) Sample frames (in our case normally resident populations) are at best imperfectly defined. For example, not every country in the world has reliable and accurate procedures for measuring and monitoring population size.
2) There are multiple definitions of the eligible populations of any country. In our research, there is inherent ambiguity in the definition of ‘eligible’, and in practice falls back upon identifying the full names of as many of the ‘ordinarily resident’ population as possible. The conception and measurement of ‘ordinarily resident’ nevertheless varies between jurisdictions. For example, the UK uses different definitions of ‘ordinarily resident’ (for access to health services and formerly for tax purposes), ‘indefinite leave to remain’ and ‘right of abode’.
3) Closest approximations to ‘ordinarily resident’ populations may systematically exclude some groups that have distinctive naming practices, for example lists of eligible voters will exclude many recent migrants and others that have not yet attained citizenship, voting age or other requirements for voting. In some

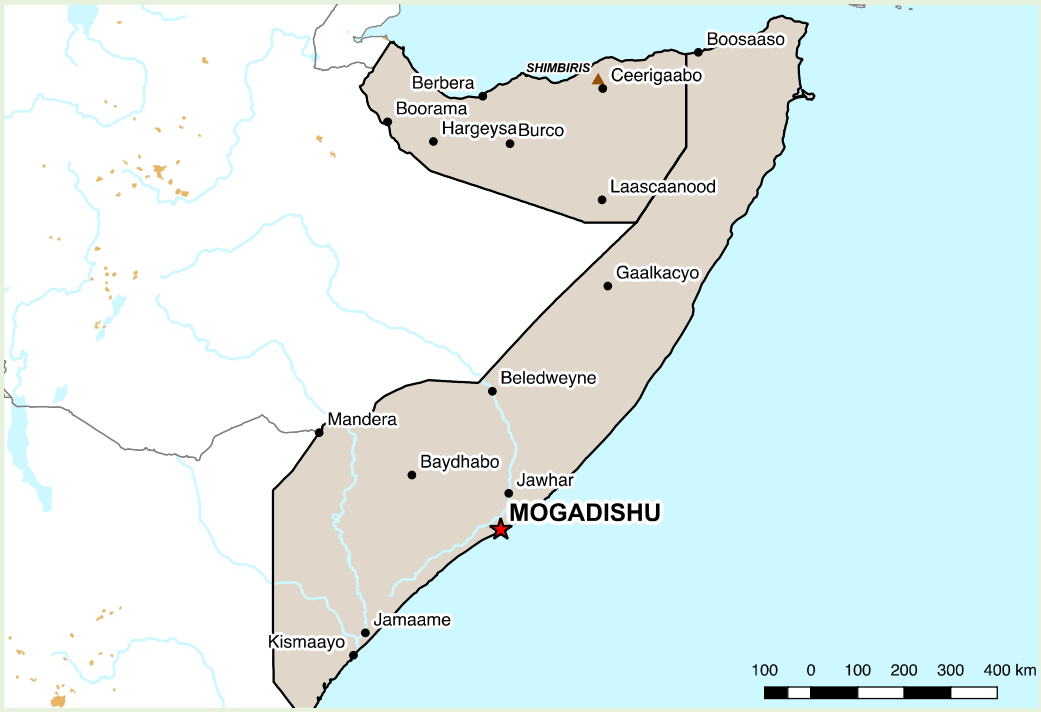


Figure 4.3
Map of Somalia.

Case Study 2: Somalia

There are no openly accessible names lists for Somalia (Figure 4.3) that provide widespread coverage of the country – not surprising in a country without an effective government controlling much of its territory. An additional problem, from the perspective of the Worldnames 2 project, is that documents are generally published in Somali, the native language, although this uses the Latin alphabet and so is relatively easily translated.

The best data source that was available was a number of PDF downloads from the Ministry of Education for the Federal Government of Somalia, at moesomalia.net/english/ (although the website has been updated since the data was extracted and so it is no longer available for direct download). The data files are lists of students that have received the national secondary school leaving certificate. The data include the full name of each pupil (with given and family names not split out), the mother's name, year of birth, roll number and certificate number (neither unique to the country), academic year, score and issue date. Somalian names do not tend to follow the Western structure where family names are passed down each generation; instead the family name of a child is typically the given name of the father.

This is likely to be a highly selective sample, for example, completing secondary school may not be possible in large parts of the country due to safety concerns, there may be a tradition of only boys or only girls attending school in some areas, and so on.

The data sources, combined, contain 7,834 name pairs, representing just 0.07% of the population based on the 2015 UN estimate of around 11 million. Combined with the likely demographic biases discussed above, this means that the Somalia dataset will likely give a poor profile of given name and family name distributions in the country, and as such is included simply because of the desire to have as many countries as possible represented in the database – poor data being better than no data at all.

countries, notably in the Middle East, migrant populations may be large relative to those that are longer settled and that may consider themselves the true 'ordinarily resident' population.

4) The non-availability of population registers of any kind will often necessitate the use of available substitute sources, for example particular age cohorts or occupational groups. In each case, we tried to use sources that were unlikely to introduce bias into the sample number and frequency of names harvested.
5) The sample size required depends upon the heterogeneity of the phenomenon (in our case, given and family names) that is being recorded. Populations with heterogeneous characteristics require larger samples, and the diversity of forenames relative to surnames may itself vary within a country. Diversity of forenames may also vary between age cohorts within a country in line with other secular trends. The sample fraction required between countries will thus vary, subject also to a minimum sample size.
6) For all of these reasons, our source data are unlikely to represent a purely random sample of the ordinarily resident population. In the absence of any reliable population sources, it is not possible to reweight names by probability of selection when synthesising the complete population.
7) The jurisdictional partitioning of some of the world is in flux, and for this reason it may be necessary to use data that do not pertain to current geographic boundaries.

In practice our data harvesting criteria were to identify:

- 10% of 2016 World Bank population estimates for small countries/areas (less than 10,000 population)
- 1,000 people, for medium countries/areas (10,000–1 million)
- 0.1%, for large countries/areas (1 million+)

The data was extracted by using Tabula to detect the table structure and data fields in the PDFs and output as a CSV. The data was then combined in Excel, with the column ordering lined up. The first word and last word of the name were interpreted to be the given name and family name respectively, a necessary simplification to homogenise name structures globally for the project. No sub-national information (e.g. school name) was available to allow a first-level administrative geography to be developed; in any case the sample size is too small to allow for meaningful geographical division.

'C/' is a popular prefix for Somali names. It is short for Cabdi, and is regarded as a prefix rather than a true first name, so was not used as a first name.

Mohamed was found to be the most common last name (i.e. family name for the purposes of the project), with 8.5% of names, followed by Ali (6.5%) and Ahmed (5.4%), with 901 distinct names out of the 7,834 in total. Mohamed was also the most common given name (8.4%) followed by Ahmed (4.3%) and Abdirahman (4.2%). The tradition of the father's given name becoming the family name of the child means that it would be expected that the top lists for family and given names would be similar.

Alternative spellings of Mohamed are also popular, with Mohmed, Mohamud, Maxamed, Mohamoud, Mohmud, Mohmoud, Mahad and Maxamuud all appearing in the top 60 most popular family names. Maxamed is the direct Somali spelling of Mohamed.

Mohamed Mohamed is the most popular name pair in the list, with 101 occurrences within the list of 7,834. There were 5,375 distinct name pairs.

The targets are based on the most recent available estimated population of the country or sub-country administrative area. This is generally sourced through simple web searches for the current population. The accuracy of the resulting data is not critical, as the target thresholds are themselves very approximate. The data generally come from the Indicators section of the World Bank's Data platform.
Our experience suggested that smaller countries are generally more open about publishing lists of people's names – possibly because they are inherently more open, but also possibly because their governments have not created elaborate data infrastructures. However, such smaller populations, even when considered in aggregate, are less useful for this project, as clustering and demographic prediction is only effective for high data volumes. Smaller datasets are also more prone to individual errors/omissions biasing the final result, so it is more important to have a greater proportion of the population covered. For this reason, the higher thresholds targets were necessary for such smaller countries. Conversely, for very large countries, even a relatively small sample will likely be representative for the country, at least for more common names.

A time/effort limitation guideline was also adopted, over and above the number target outlined above, to protect against unnecessary effort and diminishing returns. Many countries have vast numbers of datasets including people's names, of vastly varying quality and quantity. By contrast, other countries appear to have virtually no useable datasets on the web that contain useful ranges of first names and last names. In both cases, to provide an appropriate balance between effort and reward, a maximum of one person-day was employed to discover the datasets.

4.4.3 Data sources

At the time of writing this chapter, (September 2017) around 450 distinct sources have been used across approximately 170 countries. A single source has been used for most countries but this has not always been possible, for example Pakistan's very large population and lack of a comprehensive single source openly available on the web necessitated use of 24 sources in order to achieve the 0.1% threshold, both across the country as a whole and at Level 1 (Province) and Level 2 (Division) scales.

Where multiple sources were used, care was taken to try and avoid counting the same person twice, by looking for significant overlaps of names across sources. Where individual sources represent a very small proportion of the population, duplication concerns were, however, often disregarded, as they were expected to have only a minimal effect on the quality of the overall information about name distribution in the country.

As mentioned above, data sources generally need to be openly available on the web, or purchased for this project or its predecessor. We have generally only rarely used names derived from social media directories (by other projects) and have strictly avoided using data from commercial but otherwise similar pan-national projects such as Forebears (forebears.co.uk), Ancestry (www.ancestry.com) or Linked-In (www.linkedin.com).

Frequently used sources include:

- Electoral rolls (also known as voter lists or voter registers)
- Landline telephone directories (white pages)
- Government fund qualification records (e.g. rural hardship)
- School/national examination results/ candidates
- University matriculation/graduation/

admission lists
- Professional practice licences (doctors, lawyers, engineers)
- Candidates for elections (local/parliamentary)
- Official statistics from national statistics agencies (e.g. Census summaries)
- Government transparency employee and contractor pay lists
- Business owners (e.g. local service providers, tax registers)
- National service callup lists (jury service, military)

Less frequently used, but still useful sources, particularly for countries with a limited web presence or a culture of significantly restricting personal data publication:
- Government lottery/scheme winners (e.g. university laptops)
- Government honours lists/award winners
- Local government employee directories
- Public meeting minutes
- Private club member lists

Other less frequently used datasets, which potentially can come from super-national data sources:
- Player league tables (e.g. chess rankings)
- Match lineup lists of international footballers and other athletes
- Social network names (used by and supplied by other projects)
- Academic paper data releases and books
- Private insurance subscribers

The project does not republish full individual records (i.e. no disclosure of both the first name and last name of a single record) but anticipates that the publication of 'most popular full name' by region will be appropriate and possible. It will not publish personally identifying information (PII), at any level, by aggregating appropriately to ensure that the statistics published are not about a single person.

Sensitive personal information (SPI) datasets — for example medical records, or full CVs — are not collected. During the data collection process, such information, surprisingly, is encountered on the open web from time to time, but is discarded.

4.4.4 Data processing and georeferencing

For each country, files were processed once sufficient names had been retrieved, and then entered into a number of database tables — individual name pairs, aggregated tables by area for given names and family names separately, and general statistical tables.

The Tabula (tabula.technology) open source software was used to efficiently extract tables and lists of names from PDFs. Microsoft Excel, TextWrangler and a standardised set of SQL queries were also heavily used for names lists, particularly to extract the given and family name components from full names, strip initials, convert accents to an unaccented approximation, standardise apostrophes, remove/convert spaces, remove certain prefixes and suffixes (e.g. Most, Dr) and normalise the way the additional demographic information was specified across multiple datasets. For example, certain data sources omitted leading zeros for national IDs while others maintained them.

More sophisticated data cleaning was occasionally required. For example, a number of data sources were mis-encoded, or double encoded in error, requiring careful manual decoding. Transliteration websites were used, mainly for Cyrillic-alphabet to Latin-alphabet conversion, which is relatively straightforward.

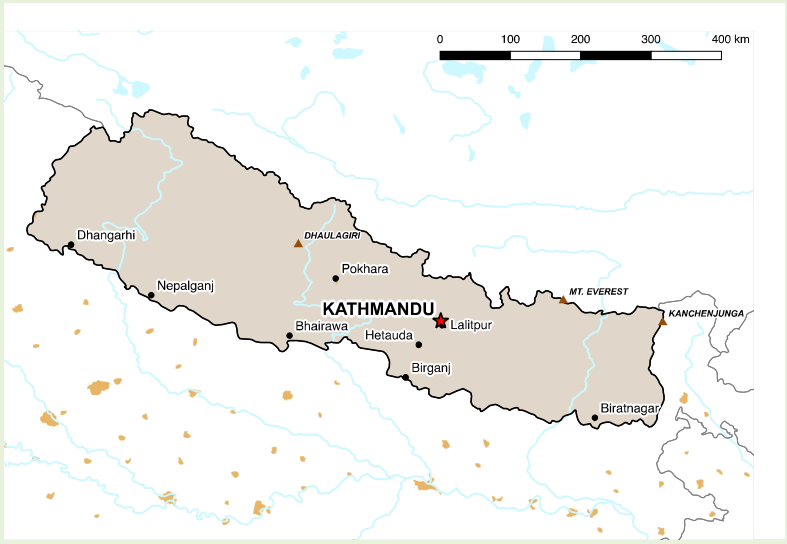
Geodata for Worldnames 2, for displaying name results in an online map on the project's website and helping understand geographic patterns of names, was mainly sourced from the GADM (www.gadm.org)

Case Study 3: Nepal

The Nepal data were built up from Police Clearance Report (PCR) Certificate Lists which are criminal record check application results, published in English by the Nepal Police as a PDF every few days. A complete list of nearly 1000 of these documents, published in the last three years, can be found at cid.nepalpolice.gov.np/index.php/pcr-certificate-list

By downloading the PDFs, converting them to CSVs using Tabula and combining them, 517,946 records can be retrieved, representing approximately 1.6% of the country's current population. Full names are listed, with a double space sometimes, but not always, separating the given name from the family name. Middle names are often present, but were disregarded, with the first word and last word forming the given and family name respectively, for Worldnames. The gender is also listed, along with a non-unique sequential list sequence number, passport number, district and a unique sequential dispatch number. The passport number is useful to de-duplicate the list (as someone may have applied more than once), while the district name can be used to build up the geography of the person's location.

Nepal (Figure 4.4) recently introduced a top-level administrative structure of seven provinces. However, geospatial data are more readily available for the previous 14 zones, which are split into 75 districts, and it is these latter two administrative areas that are adopted by Worldnames 2, particularly as the district is listed in the source data. One zone (with a single district) has only 1,132 names; however the rest of the zones are well populated in the dataset. Around 80% of the districts also have a population of at least 1,000 in the data, thus generally satisfying the target minimums discussed earlier in this chapter.



Nepal's most common family names are Tamang (4%), Gurung (3.6%) and Shrestha (1.7%), with 11,662 distinct family names in the data, while the most common given names are Ram (2.6%), Krishna (1.2%) and Santosh (0.93%), there being 37,141 unique given names detected. Ram Yadav was the most common name pair, with 1,193 occurrences, and there were 219,161 distinct name pairs.

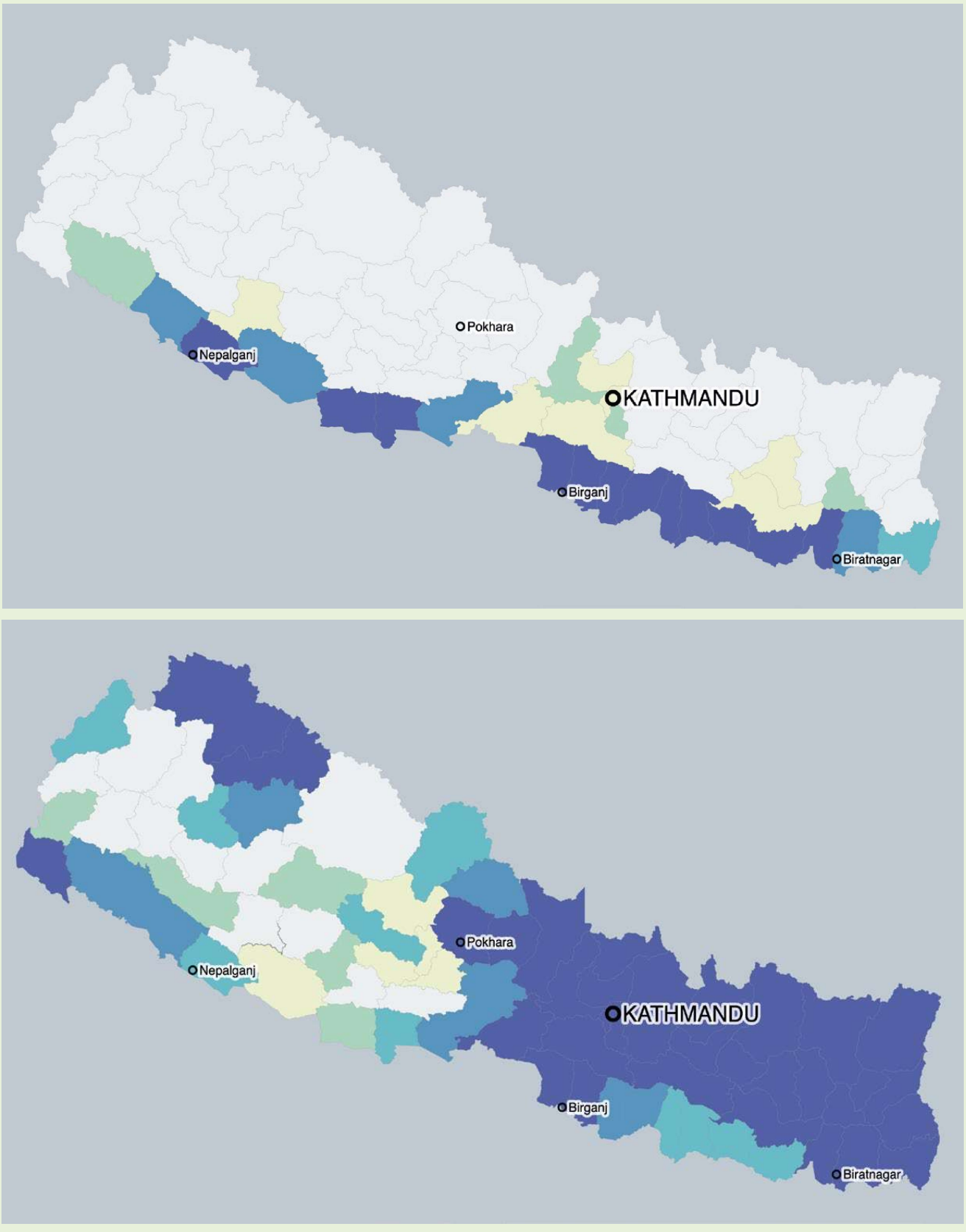
Looking at the sub-country level, in both Manang and Mustang districts, 61% of family names were Gurung, while 23% of the population in Siraha had the given name of Yadav. Analysis at the district level confirmed that this Indian-origin name is much more common along the southern border of Nepal (Figure 4.5).

Tamang, the most common family name, is much more common in the eastern part of the country and is almost absent further west (Figure 4.6).

Figure 4.4
Map of Nepal. Nepal is approximately 800km long (east-west) and 200km wide.

Opposite top:
Figure 4.5
Distribution of Yadav family name in Nepal, with darker shading indicating higher proportions of the population bearing this surname.

Opposite bottom:
Figure 4.6
Distribution of Tamang family name in Nepal, with darker shading indicating higher proportions of the population bearing this surname.


Pokhara
Nepalganj
KATHMANDU
Birganj
Biratnagar
Pokhara
Nepalganj
KATHMANDU
Birganj
Biratnagar

project. Some more recent (or legacy, where the names data were for jurisdictions that have recently changed) administrative boundaries were obtained from other projects – Natural Earth Data (www.naturalearthdata.com), OpenStreetMap data (osm.org) from Geofabrik Downloads (download.geofabrik.de) and the now discontinued MapZen Borders service (mapzen.com/data/borders), and various country-specific projects, often available through the GitHub (github.com) repositories. QGIS (www.qgis.org) was used to process and organise the metadata associated with the geodata.

This included the creation and population of a globally consistent ID for all countries and the first and second level subdivisions, where available and used for certain countries. While other projects maintain such a list (e.g. HASC codes and ISO 3166-2 codes), our own system was used for maximum flexibility, particularly as occasionally customized topologies and aggregations had to be employed, depending on the name data available. The system is based on the ISO 3166-1 country codes, an administrative level number and a padded integer code for the unit. Occasionally, the country's official codes were adopted for the latter part, where these were integer based.

For the world map of countries, a GeoJSON-format dataset from Natural Earth Data was used. Where countries had sub-country name data, the MapShaper website service was used to simplify the topology of the geodata, and one TopoJSON-format data file for up to two levels of sub-country area borders was created using it. TopoJSON (github.com/topojson/topojson) is a modern, flexible and highly compact file format.

4.5 Conclusion

The project thus far has collected individual data on approximately 1.7 billion individuals, plus a number of surname-only statistical breakdowns representing another 1.5 billion people (the majority of the names in this latter category being from China). Around 175 countries are represented, with an eventual aim to also include data for the remaining approximately 30 countries, albeit likely very simple statistical summaries of the most popular names.

As stated in the introduction, this project is quite different to other CDRC initiatives, both in its longevity (the first funding for this work was received in 2003), the persistently high levels of public interest that it has generated – recording nearly a million visits a year to Worldnames 1 for several years following its launch (worldnames.publicprofiler.org/webstats/index.html) and articles in large-circulation media such as the *Guardian* and *Daily Mail* newspapers, and the way that it has been conducted in spare time between funding streams. It is nonetheless important as the only CDRC project that purports to provide something approaching a global spatial data infrastructure, albeit founded on diverse, piecemeal and fragmented data sources.

This is, without doubt, a Big Data project – albeit one in which the search for and processing of appropriate data sources has been very labour intensive. We believe that this has implications for the wider practice – namely that Big Data have to be broadly understood before they are 'ingested', and that significant flaws in the content and coverage of data cannot be accommodated in subsequent analysis through blind application of sophisticated techniques. Spatial data are special by their very nature and geographic skills are foremost of those required to understand the possible sources and operation of bias in datasets such as Worldnames 2.

The greatest impact of the research to date has been upon the legions of amateur genealogists who are interested in understanding the geographies of their

origins across the widest possible range of spatial scales. But, as set out in our introduction, the work is of wider importance precisely because a name is a statement of a number of facets to our individual identities – ranging from cultural, ethnic and linguistic group, to age and probable social standing in the world. In our future research we hope to address the ways in which names provide indicators of the movement of populations through the generations – both by the contagious diffusion of a surname from its known point of geographic origin (nearly a thousand years in the case of Anglo Saxon surnames, but less than a century for much of Turkey, for example) and by hierarchical diffusion cascading through the increasingly interconnected system of world cities. From these standpoints, georeferenced names provide valuable indicators of the legacies of successive waves of global migration through to measures of the social progression of migrants relative to their source and host communities.

Further Reading

Brunet, G. and Bideau, A. (2000). Surnames. *The History of the Family,* 5(2), 153–160.

Cheshire, J. A., Longley, P. A. and Mateos, P. (2010) Regionalisation and clustering of large spatially-referenced population datasets: The Case of Surnames. *GIScience conference 2010.*

Longley, P. A., Cheshire, J. A. and Mateos, P. (2011). Creating a regional geography of Britain through the spatial analysis of surnames. *Geoforum*, 42(4), 506–516.

Mateos, P., Webber, R. and Longley, P. (2007). The cultural, ethnic and linguistic classification of populations and neighbourhoods using personal names. *CASA Working Papers Series* 116 Centre for Advanced Spatial Analysis, University College London.

Munzert, S., Rubba C., Meißner, P. and Nyhuis, D. (2014). Mapping the geographic distribution of names. In *Automated Data Collection with R.* Chapter 15, 380–395. John Wiley & Sons, Ltd.

Acknowledgements

The authors would like to thank Dr Muhammad Adnan for his work on the original Worldnames, which formed the basis of much of the research and development of this project.