Coordinates Series A, No. 6

An Analysis of Toponymic Homonyms in Gazetteers: Country-Level Duplicate Names in the National Geospatial-Intelligence Agency's Geographic Names Data Base

Persistent URL for citation: http://purl.oclc.org/coordinates/a6.pdf

Douglas R. Caldwell and James A. Shine

Douglas R. Caldwell (e-mail: Douglas.R.Caldwell@usace.army.mil) is a cartographer and geospatial analyst at the US Army Engineer Research & Development Center, Topographic Engineering Center, Research Division, Information Generation and Management Branch, 7701 Telegraph Road, Alexandria, VA 22315.

James A. Shine (e-mail: James.A.Shine@usace.army.mil) is a mathematician at the US Army Engineer Research & Development Center, Topographic Engineering Center, Research Division, Information Generation and Management Branch, 7701 Telegraph Road, Alexandria, VA 22315.

Date of Publication: 08/20/08

Astract: Place names are the most common way we identify geographic features. When place names are unambiguous, they can georeference features, locating them uniquely on the globe. The problem with place names is that they are often not unique; each place may have many names and many different places may have the same name. This paper studies the issue of identical names which refer to many different places, i.e., toponymic homonyms. Our country level analysis, using the National Geospatial-Intelligence Agency's Geographic Names Data Base, lays the foundation for future systematic analysis of the toponymic homonym problem. To better understand the scope of the problem, we evaluated the number of toponymic homonyms, toponymic homonyms as a percentage of all names, the maximum number of places referenced per toponymic homonym, and the 90th percentile of the toponymic homonym count. Finally, we calculated a measure of toponymic homonym complexity

Keywords: gazetteer, grounding, disambiguation, geoparsing, toponym, homonym, place name, georeferencing

Introduction

Chicago, Mount Everest, Albania, the Indian Ocean … place names describe our world and bring to mind images of places near and far. They are the most common way we identify geographic features (Hill 2006, 91). When place names are unambiguous, they can also be used to georeference features, assign them coordinate locations, and locate them uniquely on the globe. The problem with place names is that they are often not unique; each place may have many names, and many different places may have the same name.

The term used to define an individual place name that refers to many places is a homonym (Kadmon 2000, 308; Randall 2001, 103), which we refer to as a toponymic homonym, to distinguish it from other homonyms. This analysis focuses on the problem of toponymic homonyms, where different places share the same name. Our analysis is based on the National Geospatial-Intelligence Agency's Geographic Names Data Base (version as of October 2, 2007).

The concept of toponymic homonyms can be clearly understood by looking at the name 'Paris.' The National Geospatial-Intelligence Agency identifies 25 populated places or administrative places matching the exact name 'Paris' worldwide. There are many more names which contain 'Paris' as a part of the name, such as 'Puertas de Paris' in Nicaragua.

Toponymic homonyms complicate place name-based geographic information search and retrieval applications, particularly geoparsing applications, which involve "recognizing place references in text and associating geospatial coordinates with them" (Hill 2006, 100). The resolution of a name to a specific feature and location is termed grounding (Leidner et al. 2003, 31) or disambiguation (Hu and Ge 2007, 117; Smith and Crane 2001, 129-131).

The disambiguation of names in text references involves verbal cues to limit the search space. These include administrative hierarchical identifiers, feature types, and relationships to other features. The sentence, "Everyone should visit the chateau at Vaux-le-Vicomte, 40 kilometers south of the capital city of Paris, France." contains administrative hierarchy clues, i.e., this Paris is in France; feature description clues, i.e., Paris is a capital city, not just any city; and relationship clues, Vaux-le-Vicomte is 40 kilometers away from Paris in a southerly direction.

Administrative hierarchy information, which is little used in most geospatial applications, is particularly important for grounding toponymic homonyms. As you move from a global level, through first order administrative regions, to second order administrative regions, and further on down the administrative hierarchy, there will be fewer occurrences of a specific name in an area. For example, when looking at populated place and administrative types of names in the National Geospatial-Intelligence Agency's (NGA) Geographic Names Data Base (GNDB), the name San Antonio refers to 1406 locations globally, 415 locations in Mexico, and 29 locations in Chiapas, Mexico.

Despite a general recognition of the toponymic homonym issue, research on the scope of the problem remains limited. Smith and Crane provide continent level statistics on the percentage of toponymic homonyms using the Getty Thesaurus of Geographic Names. The values range from a low of 16.6% of the names in Europe up to 57.1% of the names in North America and Central America. (Smith and Crane 2001, 131) Hu and Ge document similar

information for Australian names at the national and territorial level with data from the Gazetteer of Australia and the Postcode Datafile (Hu and Ge 2007, 126-127). They calculated that 13.34% of the toponyms for the country of Australia were ambiguous or toponymic homonyms. These research results provide a preliminary view of the toponymic homonym problem, but there remains a gap in our understanding, specifically a lack of information globally at the country level. In addition, metrics beyond the percent of toponymic homonyms are needed to better understand the nature of the problem.

Country-Level Duplicate Names in the Geographic Names Data Base (GNDB)

Our analysis required decisions regarding (1) the source of the place names, (2) the granularity of the analysis, (3) the specific feature types to include, and (4) the handling of diacritics. These decisions narrowed the scope of the study and reflected the types of cues normally involved in geographic search and geoparsing applications. Results will vary depending on the choices for these factors. Thus, our study represents one of many possible views of the problem and a single snapshot in time.

As mentioned above, we evaluated the names contained in NGA's GNDB as of October 2, 2007. The GEOnet Names Server (GNS) provides access to the GNDB, which is "the official repository of foreign place-name decisions approved by the US BGN [Board on Geographic Names]."[1] Foreign places are considered to be those outside of the United States and its territories, excluding Antarctica. The GNDB provides nearly global coverage.[2]

Typically, when Americans speak of a place in a foreign country, they use the name followed by the country, with no additional hierarchy or feature type information. We say Stockholm, Sweden; not Stockholm, Stockholms Län, Sweden (capital of a political entity). To simulate this usage, a country-level approach was taken for the analysis, where we examined toponymic homonyms for geopolitical entities with unique country codes in the GNDB. These geopolitical entities include countries, dependencies, and areas of special sovereignty.[3] For simplicity, these are referred to as countries throughout the paper.

The analysis focused on place names associated with the human terrain, rather than all place names. Place names from the "Administrative and Population Names Feature Classification Codes" [4] were extracted from the database, while names for natural features, such as mountains, rivers, and lakes, were not included. This reflects the common situation where a user is looking for a name associated with population, i.e., knows some information about the type of feature associated with the place name.

The GNDB includes versions of each name with and without diacritics. While the BGN retains diacritics in the official versions of names where appropriate (Flynn 2007, 1), non-diacritic versions of the names were used in this analysis because they reflect common usage in the United States. According to Dillon:

> "More perhaps than in other European countries (including the United Kingdom), people in the United States are generally unfamiliar with diacritics and their use and resist employing them in a domestic context. This applies also to U.S. government employees, many of whom do not know how to create diacritic symbols using the keyboards at their workplace. Indeed, certain systems of communication used regularly by U.S. government personnel, such as telegrams, are technically incapable of using diacritics at all. The result of this is that, too often, U.S. government employees and the American public at large simply remove the diacritic symbols and substitute unfamiliar letterforms with familiar ones." (Dillon 2002, 2-3)

The use of non-diacritic versions of place names has the effect of slightly increasing the number of occurrences of specific toponymic homonyms, i.e., ‘San José’ and ‘San Jose’ would not be considered as different names, but would both be evaluated as ‘San Jose.’

To summarize, the study looked at the toponymic homonym problem using NGA’s GNDB as the data source, the country as the level of granularity, population and administrative feature types, and names without diacritics.

Our analysis went beyond the simple examination of toponymic homonyms as a percentage of all names. First, we wanted to obtain a sense of the magnitude of the problem, so we looked at raw counts of toponymic homonyms. Second, we followed the previous research and looked at toponymic homonyms as a percentage of all names. Third, we studied the worst cases for toponymic homonyms to get a feel for most extreme situation. Fourth, we looked at the overall pattern of toponymic homonyms using the 90th percentile values of toponymic homonym counts to better understand the distribution. Finally, we took the first, second, and fourth measures for each country and combined them into an overall score using a simple scoring system.

Number of Toponymic Homonyms

The first step in the analysis was to examine the number of toponymic homonyms, i.e., the raw count of the number of toponymic homonyms for each country. This gave us a feel for the magnitude of the problem at the country level.

Countries with No Toponymic Homonyms

In the GNDB, there are 33 "countries" which do not have any toponymic homonyms. For the most part, these are islands with a limited number of names. Many of them are not, strictly speaking, countries as commonly understood, i.e., independent states in the world. They include other types of geopolitical entities with unique country codes in the GNDB, as described in the preceding section. All have 116 or fewer place names.

Country	
Anguilla	Juan De Nova Island
Aruba	Monaco
Ashmore And Cartier Islands	Montserrat
Bassas Da India	Niue
Bouvet Island	Norfolk Island
British Indian Ocean Territory	Pitcairn Islands
British Virgin Islands	Saint Pierre and Miquelon
Christmas Island	South Georgia and The South Sandwich Islands
Clipperton Island	Spratly Islands
Cocos (Keeling) Islands	Svalbard
Coral Sea Islands	Tokelau
Europa Island	Tromelin Island
French Southern and Antarctic Lands	Turks and Caicos Islands
Glorioso Islands	Tuvalu
Heard Island and Mcdonald Islands	Vatican City
Isle of Man	Western Sahara
Jan Mayen	

Table 1. Countries Without Toponymic Homonyms

Countries with Toponymic Homonyms

The GNDB contains 340,360 toponymic homonyms that meet our criteria when analyzed at the country level. In this case, we are counting the number of unique toponymic homonyms. For example, there may be 468 occurrences of the name Hoseynabad in Iran, but the Hoseynabad counts as a single toponymic homonym. For countries with toponymic homonyms, the number varied from one name in Dominica, the Falkland Islands, Gaza Strip, Gibraltar, Marshall Islands, Saint Helena, and Swaziland to 35,392 names in Russia. The median value for countries with at least one toponymic homonym was 474.50.

The geographical distribution of this pattern shows three areas of higher values running from the northwest to the southeast across Eurasia, Africa, and the Americas. Eurasia has Russia, China, Iran, Indonesia, and Afghanistan [5] in the top five, with Germany, France, Poland, Belgium, and Sweden, North Korea, Taiwan, the Philippines, South Korea, Thailand, Burma, and Vietnam in the top 25. Outside of Eurasia, Mexico ranked in the top ten and Brazil and Columbia in the top twenty-five.

Figure 1. Map of Number of Toponymic Homonyms.

Rank	Country	Count	Rank	Country	Count
1	Russia	35392	14	Poland	5930
2	China	19363	15	South Korea	5559
3	Iran	17386	16	Thailand	5405
4	Indonesia	16504	17	Burma	4844
5	Afghanistan	10461	18	Brazil	4614
6	Pakistan	9221	19	Ukraine	4358
7	Germany	8454	20	Belgium	4181
8	Mexico	8193	21	Peru	3978
9	North Korea	7744	22	Colombia	3914
10	Taiwan	7615	23	Vietnam	3850
11	Turkey	7527	24	Nigeria	3742
12	Philippines	6030	25	Sweden	3599
13	France	5983			

Table 2 Countries with the Most Toponymic Homonyms

Because a count is sensitive to the number of names in a country, countries with fewer total names would be expected have lower counts as a general rule.

Toponymic Homonyms as a Percentage of All Names

Given an understanding of the absolute number of toponymic homonyms within each country, the next step in understanding duplicate name problem was to evaluate toponymic homonyms as a percentage of all names. This was calculated by dividing the number of toponymic homonyms by the total number of unique names in the country and multiplying the result by 100.

The values range from a low of less than 1% in Dominica and Swaziland to a high of 27.0% in Belgium. Thus, toponymic homonyms do not represent a large portion of the unique names for any country. The median value is 8.83%.

Figure 2. Map of Toponymic Homonyms as a Percentage of All Names.

Slightly fewer than half of the geopolitical identities identified in Table 3 "Top 25 Countries Ordered by Toponymic Homonyms as a Percentage of All Names" also appear in Table 2 "Countries with the Most Toponymic Homonyms." This indicates a positive relationship, albeit somewhat weak, between the two measures.

The Faroe Islands, with a total of 455 toponymic homonyms, stand out in this Top 25 list with a high value of 25.06%. This is contrary to the usual pattern for islands, which generally have smaller total numbers of unique names and few or no unique names with multiple occurrences. Other countries with lower total counts, but higher percentages, are found in Central and South America (Venezuela, Honduras, Costa Rica, and Panama), the Caribbean (Dominican Republic and Cuba), Europe (Bosnia and Herzegovina and Liechtenstein), and Africa (Burundi, Sierra Leone, Equatorial Guinea, and Madagascar).

Rank	Country	Count	Rank	Country	Count
1	Belgium	27.00	14	Cuba	19.64
2	Faroe Islands	25.06	15	Mexico	19.60
3	Venezuela	24.66	16	Sierra Leone	19.50
4	Honduras	24.37	17	Liechtenstein	19.36
5	Colombia	23.97	18	Philippines	19.27
6	North Korea	22.24	19	South Korea	18.85
7	Burundi	22.16	19	Costa Rica	18.85
8	Equatorial Guinea	22.06	21	Thailand	18.78
9	Bosnia And Herzegovina	22.04	22	Indonesia	18.46
10	Dominican Republic	21.73	23	China	18.44
11	Taiwan	20.54	24	Panama	17.80
12	Madagascar	19.68	25	Turkey	17.64
13	Brazil	19.65			

Table 3. Top 25 Countries Ordered by Toponymic Homonyms as a Percentage of All Names.

Maximum Number of Places Referenced Per Toponymic Homonym

Given an understanding of both the number and percentage of toponymic homonyms, the next step was to understand the worst case toponymic homonym for each country. This measure identifies the maximum number of different places referenced by a single toponymic homonym within each country. This was calculated for countries with at least one toponymic homonym.

The values range from a low of two in 29 countries (see Table 4) to a high of 468 in Iran for the toponymic homonym Hoseynabad (see Table 5). The median value is 13.5, which means that over half of the countries have more than 13.5 as the maximum count of their unique names having multiple occurrences.

Country	
Antigua And Barbuda	Macau
Andorra	Maldives
Bahrain	No Man’s Land
Botswana	Nauru
Cayman Islands	Suriname
Cook Islands	Marshall Islands
Djibouti	Saint Kitts And Nevis
Dominica	Seychelles
Falkland Islands	Saint Helena
French Polynesia	San Marino
Gibraltar	Saint Vincent And The Grenadines
Gaza Strip	West Bank
Jersey	Wallis And Futuna
Kuwait	Swaziland
Liechtenstein	

Table 4. Countries Having a Maximum Count of Two for Toponymic Homonyms

Figure 3. Map of Maximum Number of Places Referenced Per Toponymic Homonym.

Rank	Country	Name	Count	Rank	Country	Name	Count
1	Iran	Hoseynabad	468	14	Burma	Ywathit	129
2	Indonesia	Krajan	448	15	Thailand	Ban Mai	125
3	Mexico	San Antonio	415	16	Venezuela	San Antonio	125
4	Russia	Aleksandrovka	406	17	Ukraine	Mikhaylovka	122
5	Spain	Santa Maria	370	18	Mozambique	Joao	102
6	Brazil	Boa Vista	257	19	Germany	Berg	97
7	Colombia	La Esperanza	254	20	China	Taiping	96
8	France	Saint-Martin	230	21	Czech Republic	Nova Ves	90
9	Philippines	San Isidro	220	22	North Korea	Sinhung-ni	89
10	Turkey	Yenikoy	190	23	Cuba	San Jose	86
11	Madagascar	Tanambao	172	24	Hungary	Ujtelep	83
12	Peru	Santa Rosa	150	25	Pakistan	Tarkhanwala	82
13	Poland	Nova Wies	132				

Table 5. Top 25 Countries Ordered by Maximum Number of Places Referenced Per Toponymic Homonym.

Each country in the above list may have multiple toponymic homonyms, but each homonym refers to at most two locations.

90th Percentile of Toponymic Homonym Count

Since there is only one worst case toponymic homonym per country, additional analysis was needed to understand the statistical distribution of references per toponymic homonym. Toponymic homonym counts are non-normally distributed and strongly positively skewed. The mode or most frequently occurring value for every country with toponymic homonyms is two.

One way to view the distribution is to look at the percentile ratings. For each country with toponymic homonyms, the percentile ratings for the toponymic homonym counts were evaluated from the 75th percentile to the 100th percentile. For example, the results for Bosnia and Herzegovina are shown in Table 6. Looking at the 75 percentile value, this can be interpreted as follows: 75% of the names with multiple occurrences have four or fewer occurrences.

Bosnia and Herzegovina	
Percentile	Count
75	4
80	5
85	5
90	7
95	11
100	68

Table 6. Percentile Counts from 75% to 100%
of Toponymic Homonyms for Names in Bosnia and Herzegovina.

Further analysis focused on the 90th percentile value, as this was where the spread in percentiles begins to separate, and the first time a country's maximum value was greater than 10. Globally, the patterns are similar to previous patterns, with high values across Central and northern South America, Madagascar, and the Far East. Some new countries appeared in the Top 25 list, including the Central and South American countries of El Salvador, Guatemala, Nicaragua, Bolivia, and Ecuador; the European countries of Portugal, Austria, Greece, and Romania; and the Central Asian countries of Turkmenistan and Kazakhstan. These represent countries with a lower counts and percentages, but generally higher values of references per individual toponymic homonym.

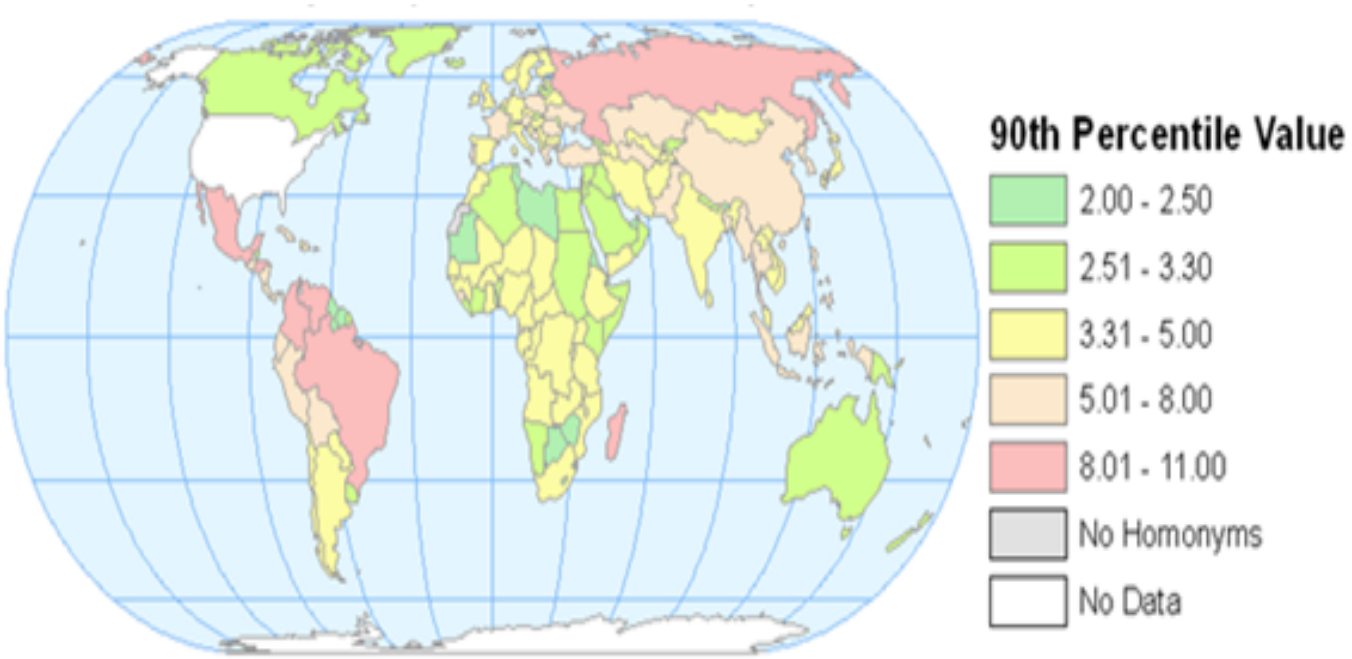

Figure 4. Map of 90th Percentile of Toponymic Homonym Count.

Rank	Country	Count	Rank	Country	Count
1	Brazil	11	21	Turkmenistan	6.3
1	Honduras	11	22	Austria	6
1	Mexico	11	22	Bolivia	6
4	Colombia	10	22	Burma	6
4	Madagascar	10	22	Costa Rica	6
4	Venezuela	10	22	Ecuador	6
7	Russia	9	22	Equatorial Guinea	6
8	Indonesia	8	22	Greece	6
9	Portugal	8	22	South Korea	6
10	Sierra Leone	8	22	Kazakhstan	6
11	Cuba	7.3	22	Peru	6
12	Guatemala	7.2	22	Pakistan	6
13	Bangladesh	7	22	Poland	6
13	Bosnia And Herzegovina	7	22	Panama	6
13	China	7	22	Romania	6
13	Dominican Republic	7	22	Philippines	6
13	Salvado	7	22	Thailand	6
13	North Korea	7	22	Taiwan	6
13	Nicaragua	7	22	Ukraine	6
13	Turkey	7			

Table 7. Top 25 Countries Ordered by 90th Percentile of Toponymic Homonym Count. Due to tie values, there are thirty-nine countries in the list.

Composite Score

As the final step in the analysis, we calculated a simple composite score from three of the four previous measures: (1) Number of Toponymic Homonyms, (2) Toponymic Homonyms as a Percentage of All Names, and (3) 90th Percentile Count of Number of Occurrences per Toponymic Homonym. The Maximum Number of Places Referenced Per Toponymic Homonym was not included, because it refers to a single name within a country. The composite score provides a general measure for estimating the duplicate name difficulties for individual countries.

There are many possible methods for calculating a composite score. Due to the non-normal distributions of the various measures, we used a nonparametric scheme, simple ranking. For each measure, the 214 countries which had at least one toponymic homonym were ranked from high to low. For example, for the Number of Toponymic Homonyms, Russia, with the highest value of 35,392 was given a score of 214, and the group of countries with the lowest value were given a score of 1. This ranking was repeated for each measure. The totals for the three measures were then added together to give a total, which was normalized.[6] The resulting composite score could theoretically have a maximum value of 100, but the actual maximum value was 96.42 for Mexico.

A map of the composite rankings is shown in Figure 5. Once again, we see a pattern of countries with high composite scores across Central America and South America. Interestingly, this is bordered by a region of extremely low values in northeastern South America in Suriname and Guyana. Other high value areas include a belt across Europe and Asia, countries of the Middle East, and a small belt across south central Africa. Not surprisingly, many of the countries with lower values include islands with both small numbers of named features and few duplicate names.

Figure 5. Map of the Composite Scores

Top 25 Countries Ordered By Composite Score					
Rank	**Country**	**Count**	**Rank**	**Country**	**Count**
1	Mexico	96.42	14	Philippines	89.70
2	Colombia	95.48	15	South Korea	89.08
3	Brazil	95.32	16	Thailand	88.61
3	Russia	95.32	17	Dominican Republic	87.83
5	North Korea	95.16	18	Portugal	87.68
6	Indonesia	95.01	19	Cuba	87.36
6	Venezuela	95.01	20	Burma	86.58
8	China	93.60	21	Pakistan	85.49
9	Honduras	93.14	21	Sierra Leone	85.49
10	Madagascar	92.98	23	Poland	85.02
11	Turkey	91.89	24	Peru	84.09
12	Bosnia and Herzegovina	91.58	25	Bangladesh	83.46
13	Taiwan	91.11			

Table 8. Top 25 Composite Scores

Summary

This analysis represents an initial attempt to describe the scope of toponymic homonyms at the country level, given a place name and country. Five measures were assessed: (1) Number of Toponymic Homonyms, (2) Toponymic Homonyms as a Percentage of All Names, (3) Maximum Number of Places Referenced Per Toponymic Homonym, (4) 90th Percentile Count of Number of Occurrences per Toponymic Homonym, and (5) a composite of measures 1, 2 and 4.

Number of Toponymic Homonyms

Thirty-three countries lacked any toponymic homonyms, but these countries had 116 or fewer total unique names. Not surprisingly, countries with larger areal extents and countries with more names in the GNDB tend to have larger numbers of toponymic homonyms, with Afghanistan, Indonesia, China, Iran, and Russia having both the largest number of unique names and over 10,000 toponymic homonyms.

Toponymic Homonyms as a Percentage of All Names

Of the Top 25 countries with the most toponymic homonyms, 22 have a score for Toponymic Homonyms as a Percentage of All Names of between 10% and 20%. This measure also brings to light countries with fewer toponymic homonyms, but higher percentages of toponymic homonyms. The top 11 countries for this measure all have scores over 20%, with Belgium at the number one spot with a score of 27.00% while the tiny Faroe Islands, with only 455 total unique names, have a score of 25.06%.

Maximum Number of Places Referenced Per Toponymic Homonym

The Maximum Number of Places Referenced Per Toponymic Homonym only tracks the worst case, not the distribution of multiple occurrences. Three of the top five countries (Russia, Iran, and Indonesia) with top counts of toponymic homonyms also appear in the top five of the Maximum Number of Places Referenced Per Toponymic Homonym. Belgium, Vietnam, and Nigeria are among the top 25 countries with high numbers of toponymic homonyms, but very low maximum values (below 50).

90th Percentile Count of Number of Occurrences per Toponymic Homonym

The distribution of counts for toponymic homonyms is distinctly non-normal, with all countries having a modal value of 2 and strong positive skew. After looking at the percentile distributions for the Number of Toponymic Homonyms, the distribution of the 90th Percentile was chosen for further analysis, as the spread among countries increases at this point. Of the 6 countries ranked with the top 4 rankings, the Central and South American countries of Brazil, Columbia, Honduras, Mexico, and Venezuela dominate, all with scores of 10 or higher.

Composite Analysis

The Composite Analysis provided a simple composite score using three of the four previous measures: (1) Number of Toponymic Homonyms, (2) Toponymic Homonyms as a Percentage of All Names, and (3) 90th Percentile Count of Number of Occurrences per Toponymic Homonym. The ranks for the three measures were then

added together to give a total, which was then normalized. The Composite Analysis identified distinct geospatial patterns of countries where the toponymic homonym problem is more significant. These countries occur across areas in Central America and South America, the Middle East, and the Far East. The countries without few toponymic homonyms are typically islands and other countries with few names.

Conclusions

This initial analysis of toponymic homonyms at the country level is intended to serve as a foundation for further systematic analysis. It has shown the value of using multiple measures, rather than simply providing toponymic homonyms as a percentage of all unique names, as has been done in previous studies. Use of counts of toponymic homonyms gives an absolute measure of the magnitude of the problem, and an understanding of whether 2, 200, 2000, or 20000 names are involved. Analysis of the distribution of toponymic homonyms provides an indication of the number of names associated with each toponymic homonym, providing a feeling for shape of the distribution and an understanding of whether toponymic homonyms generally have few or many names associated with them. Looking at the maximum number of names for a toponymic homonym is less useful for understanding the wider problem of duplicate names, but identifies the worst case for a country. Finally, our composite measure provides a useful indicator of the expected difficulties in dealing with toponymic homonyms on a per country basis for the domain of administrative and populated place names.

The results of this study should be of value to those involved with place-based information retrieval, particularly applications like geoparsing. Not only do the results indicate areas where toponymic homonyms are more prevalent, they point to the need for context beyond the name, country, and feature type to support precise geospatial information retrieval.

Future Work

Our study also points out opportunities for expanded study of toponymic homonyms. The results presented here represent specific types of data in a single gazetteer at a fixed point in time. The many possibilities for additional research include: looking at additional measures; different gazetteers, including gazetteers of the United States; other administrative hierarchy levels, including first- and second-order administrative divisions of countries; the extent of the problem within non-jurisdictional geographic place names, not just populated places and administrative names; and names with diacritics. In addition, further analysis is needed to address issues outside the scope of the present research, including the development of a better understand the relationships between the measures and underlying causes of the duplicate name problem, as well as potential solutions for limiting the impact of the toponymic homonyms in geospatial queries.

Acknowledgements

The authors would like to thank Mr. Rick Joy, Team Leader for Geographic Evidential Reasoning; Ms. Valerie Carney, Chief of the Information Generation and Management Branch; and Dr. Eric Zimmerman, Chief of the Research Division, all at the Topographic Engineering Center, Alexandria, VA, for their support. They would also like to especially thank David Allen and the anonymous reviewers for their valuable comments and suggestions.

Notes

1. Quoted from http://earth-info.nga.mil/gns/html/whatsnew.htm#C3, accessed on August 18, 2008.

2. A list of countries and associated country codes for names in the Geographic Names Data Base can be found at. http://earth-info.nga.mil/gns/html/namefiles.htm. This was accessed on August 18, 2008.

3. The definition of geopolitical entities used in the GNDB can be found at http://earth-info.nga.mil/gns/html/help.htm. This was accessed on August 15, 2008.

4. These feature categories include the following designation codes: first-order administrative division (ADM1), second-order administrative division (ADM2), third-order administrative division (ADM3), fourth-order administrative division (ADM4), administrative division (ADMD), leased area (LTER), political entity (PCL), dependent political entity (PCLD), freely associated state (PCLF), independent political entity (PCLI), section of independent political entity (PCLIX), semi-independent political entity (PCLS), parish (PRSH), territory (TERR), zone (ZN), buffer zone (ZNB), populated place (PPL), seat of a first-order administrative division (PPLA), capital of a political entity (PPLC), populated locality (PPLL), abandoned populated place (PPLQ), religious populated place (PPLR), populated places (PPLS), destroyed populated place (PPLW), section of populated place (PPLX), and Israeli settlement (STLMT).

5. These higher counts reflect NGA's emphasis in collecting more names in countries where the United States has specific interests. For example, the number of names collected in Afghanistan is higher relative to the country's size and population than other countries.

6. To normalize the data, the sum of the scores was divided by the maximum possible score and this result was multipled by 100. The potential range of values was thus between 0 and 100.

Bibliography

Crane, Gregory. 2004. "Georeferencing in Historical Collections." *D-Lib Magazine*, May. http://www.dlib.org/dlib/may04/crane/05crane.html (accessed August 14, 2007).

Dillon, Leo. 2002. *Recent Discussions in the United States Board on Geographic Names Concerning the Creation of Anglicized Exonyms.* Berlin: United States Board on Geographic Names. http://unstats.un.org/unsd/geoinfo/N0243895.pdf (accessed September 24, 2007).

Flynn, Randall. 2007. *Principles and Policies: Foreign Geographic Names*. Washington, DC: Board on Geographic Names. Agenda Item 5.1, 23rd BGN/PCGN Conference, April 23 – May 3, 2007.

Hill, Linda L. 2006. *Georeferencing: The Geographic Associations of Information*. Cambridge, MA: MIT Press.

Hu, You-Heng, and Linlin Ge. 2007. "A Supervised Machine Learning Approach to Toponym Disambiguation." *In The Geospatial Web: How Geobrowsers, Social Software, and the Web 2.0 Are Shaping the Network Society*, ed.

Arno Scharl and K. Tochtermann, 117-128. London: Springer.

Kadmon, Naftali. 2000. *Toponymy: The Lore, Laws, and Language of Geographical Names*. New York: Vantage Press.

Leidner, Jochen, G. Sinclair, and B. Webber. 2003. *Grounding Spatial Named Entities for Information Extraction and Question Answering*. Ed. A. Kornai and B. Sundheim. HLT-NAACL 2003.

Randall, Richard R. 2001. *Place Names: How They Define the World--And More*. Lanham, MD: Scarecrow Press

Scharl, Arno. 2007. "Towards the Geospatial Web: Media Platforms for Managing Geotagged Knowledge Repositories." In *The Geospatial Web: How Geobrowsers, Social Software, and the Web 2.0 Are Shaping the Network Society*, ed. K. Tochtermann and Arno Scharl, 3-14. London: Springer.

Schilder, F., Y. Versley, and C. Habel. 2004. *Extracting Spatial Information: Grounding, Classifying and Linking Spatial Expressions. Proceedings of the Workshop on Geographic Information Retrieval at SIGIR 2004*. http://www.geo.unizh.ch/~rsp/gir/abstracts/schilder.pdf (accessed August 14, 2007).

Smith, David, and Gregory Crane. 2001. "Disambiguating Geographic Names in a Historical Digital Library." In *Research and Advanced Technology for Digital Libraries*, 127-136. Heidelberg: Springer Berlin. http://www.springerlink.com/content/h7em0v5803e6h7yb (accessed September 18, 2007).

Stewart, George Rippey. 1970. *American Place-names; a Concise and Selective Dictionary for the Continental United States of America*. New York: Oxford University Press.

Stewart, George Rippey. 1975. *Names on the Globe*. New York: Oxford University Press.

Stewart, George Rippey. 1982. *Names on the Land: A Historical Account of Placenaming in the United States*. San Francisco: Lexikos.

Wacholder, Nina, Yael Ravin, and Choi Misook. 1997. *Disambiguation of Proper Names in Text. ACl Anthology: A Digital Archive of Research Papers in Computational Linguistics*, April 31. http://acl.ldc.upenn.edu/A/A97/A97-1030.pdf (accessed August 14, 2007).

Top | MAGERT Home | Coordinates Home