Section 4 Regional and global initiatives

Chapter 10 – **Working with geonames.org**

Marc Wick

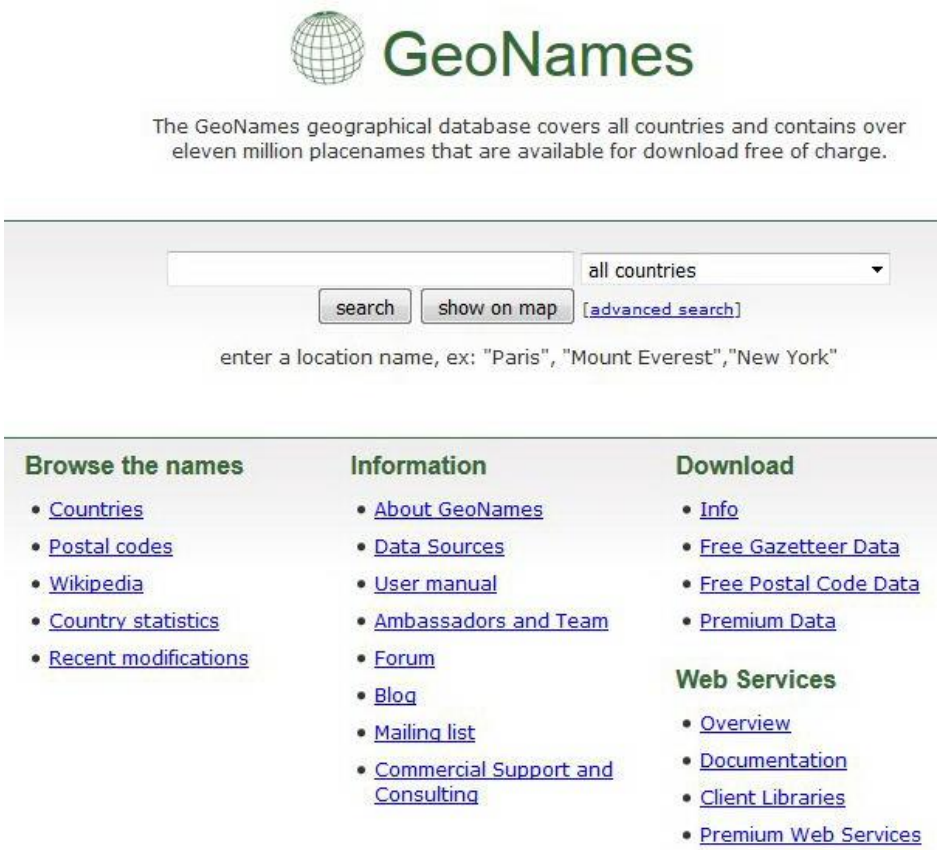

Figure 10-1 - GeoNames application

10.1 Introduction

This chapter will give an overview of the geonames.org gazetteer and how to work with it. We will introduce the data model and the classification system. We will describe the extract files, which can be downloaded to have a full copy of the gazetteer locally and also the web services, which can be used to build an application with the geonames.org ([1]) data set without having to download the data. We will list the most important sources and how data can be contributed to the gazetteer either in making edits directly with the wiki interface or with providing data sets to be imported into GeoNames.

10.2 What is GeoNames?

GeoNames is an open global gazetteer, a database with records of geographical features. At the time of writing (Summer 2016) it contains over 11 million features with additionally 11 million alternate names. The gazetteer data is downloadable in tab separated csv files in utf8 encoding under a liberal Creative Commons Attribution (cc-by) license. The only condition is to somehow give credit to GeoNames and in turn to the sources. It is left to the user how the attribution is implemented, it can be a link on a website, a phrase in the documentation or another form. GeoNames is aggregating data from many sources, the most important of which is the United States National Geospatial-Intelligence Agency. Other important sources are the national mapping agencies or the national statistical offices of all countries whenever they publish data compatible with the cc-by license. A wiki interface on the www.geonames.org website allows users to browse the data and to quickly add features, attribute values or correct errors. GeoNames is having over 160 000 users many of which contribute with a total of several hundreds of edits every day. Nevertheless, the main data volume are batch imports from the national mapping agencies. A quality assurance process in a monthly release cycle ensures the data quality and integrity, in particular with the wiki interface where everybody can contribute. Last but not least GeoNames offers an api with nearly 40 web services. The web services give direct programmatic access to the gazetteer data set. On the other hand, many services used internally to enhance the data when adding to the database are also made publicly available as a web service. Examples are various digital elevation services or time zone reverse lookup.

10.3 Data model

The data model of GeoNames is quite simple and consists of two main tables. A ***GeoName*** table with the feature attributes and a second table ***AlternateName*** with translations of the feature name. Each feature has a main toponym name, which is either in English or an

internationally understood name. Names in other languages are found in the alternatename table. The language is identified with the ISO language code [3].

GeoName – Feature Attributes

The most important table is the GeoName table. It contains the basic attributes of a geographic feature. The geonameId as the primary key is a sequence number and never changes. When inserting a new feature, the sequence is incremented and the new feature gets the next integer number. Each feature is represented by a lat/lng coordinate pair. The location of the lat/lng point is indicating the location on the globe, it may be the geometric centre, the location of the administration or some other position. The type of feature is described with a two-level classification system. The featureClass is a first rough classification and puts the features into one of nine classes. The nature of the feature is described in more detail with the assignment of one of the 660 feature codes on the second classification level.

GeoName	
geonameId	integer id of record in GeoNames database
name	name of geographical point (utf8) varchar(200), English or internationally understood name.
asciiname	name of geographical point in plain ascii characters, varchar (200)
alternatenames	alternatenames, comma separated, ascii names automatically transliterated,
	convenience attribute from alternatename table, varchar (10000)
latitude	latitude in decimal degrees (wgs84)
longitude	longitude in decimal degrees (wgs84)
feature class	char (1)
feature code	varchar (10)
country code	ISO-3166 2-letter country code, 2 characters [3]
cc2	alternate country codes, comma separated, ISO-3166 2-letter country code, 200 characters
admin1 code	fipscode, see exceptions below, see file admin1Codes.txt for display names of this code; varchar(20)
admin2 code	code for the second administrative division, see file admin2Codes.txt; varchar(80)
admin3 code	code for third level administrative division, varchar (20)
admin4 code	code for fourth level administrative division, varchar (20)
population	bigint (8 byte int)
elevation	in meters, integer
dem	digital elevation model, srtm3 or gtopo30, average elevation of 3''x3'' (ca 90mx90m) or 30''x30'' (ca 900mx900m) area in meters, integer. srtm processed by cgiar/ciat.
Time zone	the iana [2] time zone id (see file timeZone.txt) varchar(40)
modification date	date of last modification in yyyy-MM-dd format

Table 10-1 – GeoName table

AlternateName – name in other languages

A feature may have more than just a single name. It may have name variants in other languages, it may have had other names in the past, it may have short and long names and it may even be known with colloquial names. These names are modelled in the second table, the *AlternateName* table.

In order to identify the language an alternate name stands for the ISO 639 language code is used [3]. Furthermore, a couple of pseudo language codes describe other names like: '*post*' for postal codes, *'iata'*, *'icao'* and *'faac'* for the respective airport codes. *'link'* stands for an url pointing to a website. The most often used links are to the corresponding Wikipedia article. Over 500000 alternatenames are links to Wikipedia pages.

Four flags help to further describe an alternate name: short, preferred, historic or colloquial. *'Big Apple'* is a

AlternateName	
alternateNameId	the id of this alternate name, int
geonameId	geonameId referring to geonameid in table 'GeoName', int
isolanguage	iso 639 language code 2- or 3-characters; 4-characters 'post' for postal codes and 'iata', 'icao' and 'faac' for airport codes, fr_1793 for French Revolution names, 'abbr' for abbreviation, 'link' for a website, varchar(7)
alternate name	alternate name or name variant, varchar(400)
isPreferredName	'1', if this alternate name is an official/preferred name
isShortName	'1', if this is a short name like 'California' for 'State of California'
isColloquial	'1', if this alternate name is a colloquial or slang term
isHistoric	'1', if this alternate name is historic and was used in the past

Table 10-2 – AlternateName table

colloquial name for *New York City*, whereas *'Karl-Marx-Stadt'* is a historic name for *Chemnitz*.

Hierarchy - Administrative Hierarchy

The administrative division a feature belongs to is modelled with the attributes countryCode and adminCode1 to adminCode5. The countryCode attribute is the two character ISO country code and contains the countryCode of the country the feature belongs to. The adminCode1 point to the first order administrative division of the same country. The combination of countryCode, adminCode1 and adminCode2 gives you the second order administrative division. The administrative divisions are like all features part of the GeoName table, differentiated by their featureCode.

The hierarchical structure modelled with the adminCodes is treelike, every feature can have only one direct parent. For administrative hierarchies, this is normally sufficient.

For super-national features, a second countryCode attribute contains a comma separated list of all countries with a relation to the feature. The same for border features (mountains, lakes), which belong to more than one country.

Example for administrative hierarchy: Rome (geoNameId 3169070), the capital of Italy, has countryCode *'IT'* pointing to Italy (geoNameId 3175395), the adminCode1 is *'07'*, which points to the region *'Lazio'*. The second order administrative division is the province Rome with adminCode2 'RM'. On the third level, we have adminCode3 '058091' for the Commune (municipality) of Rome. The adminCode3 ' 058091' is the code used by the *Italian National Institute of Statistics [4]* for the third order administrative divisions.

Non-Administrative Hierarchy

The admin hierarchy cannot model all hierarchy types. An additional table contains relations between features outside of the administrative hierarchy. These relations are for instance relations between spot features and populated places, between neighbourhoods and cities or also regions consisting of administrative divisions. Example for tourism region: Spain has defined a couple of tourism regions which are defined as a group of municipalities. The Costa Brava (geoNameId 3127668) is made up of 26 municipalities.

10.4 Feature Classes and Feature Codes

FeatureClass	
A	Administrative features: country, state, region, ...
H	Hydrographic features: stream, lake, ...
L	Area features: parks, area, ...
P	Populated places features: city, village, ...
R	Road/Railroad features: road, railroad
S	Spot features: spot, building, farm
T	Hypsographic features: mountain, hill, rock, ...
U	Undersea features
V	Vegetation features: forest, heath, ...

Table 10-3 – Feature Classes

The *featureClass* is a rough categorization further enhanced by the *featureCode* which describes the feature in more detail. Each feature may belong to one of 660 feature codes. We distinguish populated places by size and function, whether a populated place serves as a seat of an administrative division or even as a country capital. Similarly, we separate protuberances into hill, mountain, peak, range, rock, pass or another of 99 feature codes.

FeatureClass			
		Number of Feature-Codes	**Number of Features**
A	Administrative features	24	357,767
H	Hydrographic features	134	2,134,794
L	Area features	49	379,109
P	Populated places features	18	4,349,577
R	Road/Railroad features	21	40,356
S	Spot features	244	2,276,788
T	Hypsographic features	99	1,516,560
U	Undersea features	63	14,476
V	Vegetation features	17	39,478

Table 10-4 – Feature Classes and its Feature Codes

Each feature is unique, there is only one entry in the GeoName table for a feature. Important to understand in this respect is that administrative divisions and populated places are considered two different concepts, each requiring its own feature in the gazetteer. Cities therefore often have two entries, one for the populated place and the other to represent the administrative division (municipality, commune, etc).

Most often used Feature Codes			
Number of Features	**Description**	**Feature Code**	**Feature Class**
3,982,992	populate place	PPL	P
856,325	stream	STM	H
382,137	mountain	MT	T
359,720	hill	HLL	T
320,048	farm	FRM	S
276,737	school	SCH	S
262,085	lake	LK	H
245,001	church	CH	S
241,238	hotel	HTL	S
194,191	intermittent stream	STMI	H

Table 10-5 – Most often used Feature Codes

10.5 Download – Extract Files

The GeoNames data is exported daily into a download directory where it can be downloaded for free. A username/password is not required for the data download.

The data model with the two main tables GeoName and AlternateName is reflected in the extract files. The main information is included in the two files allCountries.txt and alternatenames.txt, with the former being the export of the GeoName table and the later the export of the AlternateName table. The hierarchy information is found in the hierarchy.txt file.

The other files in the download directory are reference and convenience files.

10.6 Reference Files

A couple of files in the GeoNames download directory are reference data. These are lookup files for codes used by GeoNames. GeoNames is using the ISO 639 languages codes in the alternatename table to identify the language of a name variant. The timezone of a feature is identified by the iana timezoneId (see also the section 'TimeZone') [2]. The feature codes used by GeoNames are described in a handful of languages. Attributes specific to countries and therefore not part of the data-model for geoname

Reference Files	
iso-languagecodes.txt	iso 639 language codes, as used for alternate names in file alternateNames.zip
timeZones.txt	countryCode, timezoneId, gmt offset on 1st of January, dst offset to gmt on 1st of July (of the current year), rawOffset without DST
featureCodes_<lang>.txt	name and description for feature classes and feature codes in a couple of languages (bg, en, nb, nn, no, ru, sv)
countryInfo.txt	country information: iso codes, fips codes, languages, capital, ...

Table 10-6 – Reference Files

features are found in a separate countryInfo file. It contains various country codes, the languages spoken in the country, postal code format, internet top level domain and neighbouring countries.

Convenience Files

The two main files allCountries.txt and alternatenames.txt are quite big. The allCountries file is uncompressed about 1.3 GB large. For users who are only interested in a subset of the data a couple of convenience files are available.

Convenience Files	
XX.zip	features for country with iso code XX, see 'geoname' table for columns
Cities1000.zip	all cities with a population > 1000 or seats of adm div (ca 150.000), see 'geoname' table for columns
Cities5000.zip	all cities with a population > 5000 or PPLA (ca 50.000), see 'geoname' table for columns
Cities15000.zip	all cities with a population > 15000 or capitals (ca 25.000), see 'geoname' table for columns
admin1CodesASCII.txt	ascii names of admin divisions.
admin2Codes.txt	names for administrative subdivision 'admin2 code' (UTF8), Format: concatenated codes <tab>name <tab> asciiname <tab> geonameId

Table 10-7 – Convenience Files

These files are subsets of the allCountries file with the same file layout and attributes, but with fewer rows. Users interested only in a single country can download the features for this country in a file named with the iso two letter country code. Users only interested in major cities can download one of the three citiesXXX files.

The alternateName table has a couple of flags, which need to be evaluated in order to determine the best name variant for display in a particular language (see also the section 'Using GeoNames - Tips and Tricks - Which name to use'). For the first and second level administrative divisions, two convenience files (admin1CodesASCII.txt and admin2Codes.txt) contain the best name for display in English.

Modification Files

The download directory also includes a couple of convenience files with the modifications of the previous day. These may be used to update the local copy of GeoNames on a daily basis. The downside of this approach is that no file must be skipped.

We consider it preferable to download the full dataset periodically instead of the daily modification files.

Each feature has at least the main name in the geoname table, often additional name variants are found in the alternatename table. There might also be various name variants in a particular language and we need to decide which name to use.

Modification Files (With <date> = <yyyy-MM-dd>)	
modifications-<date>.txt	all records modified on the previous day
deletes-<date>.txt	all records deleted on the previous day, format: geonameId <tab> name <tab> comment. Duplicates usually have a delete comment of the form 'duplicate of <geoNameId>' with the geoNameId of the remaining feature.
alternateNamesModifications-<date>.txt	all alternate names modified on the previous day
alternateNamesDeletes-<date>.txt	all alternate names deleted on the previous day, format: alternateNameId <tab> geonameId <tab> name <tab> comment.

Table 10-8 – Modification Files

10.7 Using GeoNames - Tips and Tricks – Which name to use?

The recommended approach is to check the alternate name table for name variants in the desired language, ignoring the historic and colloquial names. From these

name variants, we then use the one marked as 'short' for display. If no name is available with the desired iso language code we see whether we have an alternate name without language code. Last but not least we use the name from the geoname table if we do not find an appropriate entry in the alternatename table.

Duplicates

Place names are not unique, they even tend to cluster in an area. People often complain about 'duplicates' in GeoNames because they erroneously assume that a feature name may only exist once per country and first admin division. Unfortunately, this is not the case and village names are often repeated in an area.

10.8 Webservice vs Data Download

Besides the freely available daily database extracts GeoNames also offers a wide range of web services in a freemium model. A number (currently 20000 credits) of requests per day is free, for higher usage or a service level agreement a premium offering exists.

The GeoNames web services have the advantage that they offer a ready-made and direct access to services based on the gazetteer data. Many additional services, internally used to enhance and verify the gazetteer data, are also available. Among these: elevation, time zone, Wikipedia, weather and street reverse geocoding services.

Provide Data – Contribute to GeoNames

Contributions to GeoNames can be made directly with the browser based wiki interface. Only a user account is needed. Experienced users may be upgraded to higher user levels to be able to also update critical attributes and features.

For contributing entire data sets it is preferred to import the data with a bulk import instead of the wiki interface. It is sufficient to make the data available in flat csv files, similar to the export files.

TimeZone

One of the attributes of a feature is the timeZoneId. It refers to the iana timezone project. Example Europe/Paris for the time in continental France.

The iana timezone project maintains all time changes since 1970 and is included in most programming environments. The iana timeZoneId is therefore sufficient to determine all time changes (daylight saving time) since 1970.

Countries used in GeoNames

GeoNames relies on the ISO [3] country codes. Each ISO country code has a corresponding entity in the GeoNames database.

The only exception so far is Kosovo, which does not yet have an ISO country code assigned. GeoNames is temporarily using XK, till an ISO country code will be officially assigned.

Admin codes used by GeoNames

On the first administrative level GeoNames is mainly using the US coding standard (formerly FIPS). On lower levels the national code, often provided by the national statistical office, is used.

10.9 Data Sources

GeoNames is accumulating data from a large variety of different data sources together with user contributions via the wiki style browser based edit interface. The largest single data source is the United States NGA GEOnet Names Server (GNS) from the National Geospatial-Intelligence Agency. [5] It contains data for all countries except the US. GeoNames is importing the NGA data set on a monthly basis.

As more and more national mapping agencies and national statistical offices become aware of the importance of free access to geographical data the number of national agencies offering data for free under an open license is growing year by year. Thus, continuously increasing the number of data sources GeoNames is integrating.

In developing countries where national mapping agencies do not have the same resources as in industrialized countries relief organizations working with GeoNames data often contribute their own data.

When periodically importing a data set from a data provider, GeoNames checks each modification with the modification history of the same attribute to make sure corrections done by users via the wiki interface are not directly overwritten and set back with the following data import.

10.10 References

[1] www.geonames.org, last accessed 09/2016

[2] Internet Assigned Numbers Authority (IANA), http://www.iana.org/, last accessed 09/2016

[3] International Organization for Standardization (ISO), http://www.iso.org/ last accessed 09/2016

[4] Italian National Institute of Statistics, http://www.istat.it/ last accessed 09/2016

[5] NGA GEOnet Names Server (GNS), http://geonames.nga.mil/gns/html/index.html last accessed 09/2016