# On building a high performance gazetteer database

**Amittai E. Axelrod**
MetaCarta, Inc.
875 Massachusetts Ave., 6th Flr.
Cambridge, MA, 02139
amittai@metacarta.com

## Abstract

We define a data model for storing geographic information from multiple sources that enables the efficient production of customizable gazetteers. The GazDB separates names from features while storing the relationships between them. Geographic names are stored in a variety of resolutions to allow for i18n and for multiplicity of naming. Geographic features are categorized along several axes to facilitate selection and filtering.

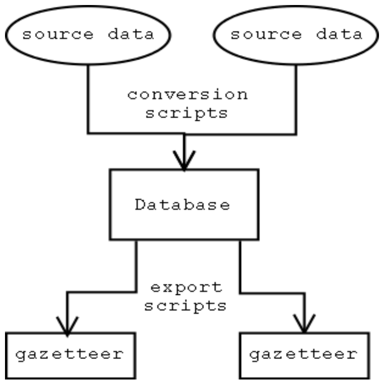


Figure 1: The gazetteer production process

## 1 Introduction

We are interested in collecting the largest possible set of geographic entities, so as to be able to produce a variety of extremely comprehensive gazetteers. These gazetteers are currently produced to search for both direct and indirect geospatial references in text. The production process can be tailored to produce custom gazetteers for other applications, such as historical queries.

The purpose of the MetaCarta GazDB is to provide both a place and supporting mechanisms for storing, maintaining, and exporting everything we know about our collection of geographic entities.

To produce a gazetteer from various data sources, we make use of a database, the *GazDB*, as well as two sets of scripts: *conversion scripts*, to transfer the data from its source format into the GazDB, and *export scripts* to output data from the GazDB in the form of gazetteers. The interaction between these elements is illustrated in Figure 1.

Geographic input data is collected from multiple (not necessarily disjoint) sources, each with their own peculiar format. As such, the conversion scripts must perform some amount of normalization and classification of the input data in order to maintain a single unified repository of geographic data. However, in order to justify the overhead of consolidating all the data into a single entity, it must be possible to output all of it into multiple gazetteers designed for different goals.

It should also be possible to perform filtering operations on the gazetteer entries, such as comparing entry names against common-language dictionaries. This can be used determine whether occurrences of gazetteer names in documents are geographically relevant (Rauch et al., 2003).

This is the task for the export scripts. However, in this paper, we shall focus on the heart of the system, namely the GazDB. Section 2 describes how the GazDB relates geographic names and features. In Section 3 we describe how the GazDB handles ambiguities and inconsistencies in geographic names. Finally, in Section 4 we outline the classification and storage system used for geographic features.

## 2 Gazetteer entries in the GazDB

The most basic form of a gazetteer entry consists of a mapping between a geographic name and a geographic

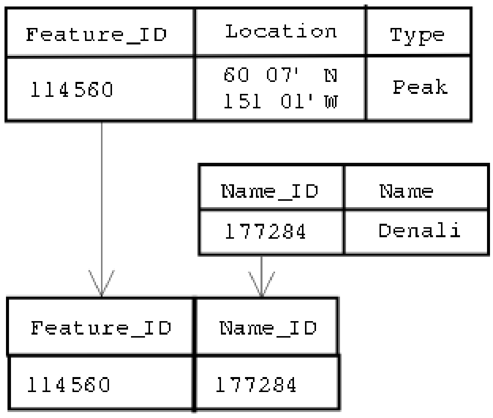


Figure 2: Relating features and names in the GazDB

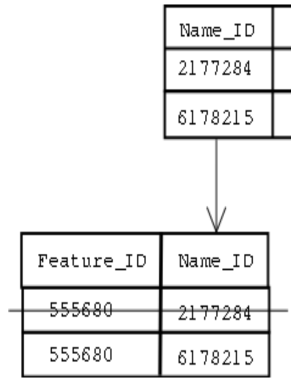


Figure 3: Updating a name in the GazDB

location. The Alexandria Digital Library Project (Hill, 2000), however, defines a gazetteer entry as also requiring a type designation to describe the entity referred to by the name and location. Because a geographical type designation classifies the physical entity rather than the name assigned to it, we think of gazetteer entries produced by the GazDB as relating geographic names and geographic features (which have inherent types). We will separately discuss geographic names and geographic features in greater detail later, and focus on the stored relations between them first.

A naive approach to creating a gazetteer is to maintain a flat file with one gazetteer entry per line, as follows:

| | |
|---|---|
| Boston | 42° 21'30"N, 71° 4'23"W |
| Cambridge | 42° 23'30"N, 71° 6'22"W |
| Somerville | 42° 23'15"N, 71° 6'00"W |

This schema is overly simplistic because it supposes a one-to-one mapping between geographic names and features, when in reality many geographic features have more than one name commonly associated with them. For instance, the tallest mountain in North America is unambiguously referred to as either *Mount McKinley* or *Denali*. Using this gazetteer, recording both names for the mountain would result in the creation of two entries. This is highly impractical on a large scale due to space requirements and the complexity of systematically updating or modifying the gazetteer.

The GazDB uses the well-known relational approach (Codd, 1970) to store the geographic data for the gazetteer. To do so, we separate the notion of a geographic name from the geographic feature that it represents. We maintain distinct tables for locations and names– mappings between names and locations are stored in a third table, keyed by the unique numerical identifiers of both the name and the location, as shown in Figure 2. This system enables the GazDB to support both many-to-one relations between names and features, as in the case of *Denali* and *McKinley*, and one-to-many relations such as *London* being the name of both a city in Britain and a town in Connecticut.

In the GazDB, several other relational tables are used to store numerical data associated with the known geographic features. For example, population data is kept in a separate table that links census figures with the ID's of entries in the feature table. This is useful because it facilitates queries to be performed only on inhabited places. Elevation data is stored in a similar manner.

As gazetteers get updated, corrections are often made to the name or to the feature data. To update a name, we formally abandon the old ID, create a new name entry, and update the name↔feature mapping table by replacing the old name ID with the new one, as in Figure 3. We repeat this process for each table in the GazDB that refers to the old ID– this is simple, because the tables are indexed by ID. Updating geographic locations or numerical data in the GazDB is done in an identical manner.

The GazDB also includes a table for storing detailed information about the sources of the data in the GazDB– for instance, *"NIMA GeoNet names datafile for Afghanistan (AF), published November 8 2002"*. Every element in the GazDB is then associated with the appropriate entry in the source table. This enables the accountability of all entries in the GazDB, preventing the appearance of "mystery data". The source table also allows easy, systematic, source-specific modifications of the GazDB's entries to keep pace with frequently updated datasets, thereby maintaining the freshness of the GazDB's data.

The GazDB also includes a complete log of all updates to the database tables and entries. Because data rows are abandoned but not deleted during updates, it is possible to recreate the state of the database prior to any particular set of updates.

The flexibility of the relational design also allows the inclusion of new kinds of data that were not thought of or not available in the original schema. For instance, one could add yearly precipitation data for geographic locations by creating an additional table mapping locations to rainfall amounts, without the need to re-ingest the data already in the GazDB.

The GazDB also maintains a historical geographical record by capturing temporal extents for mappings – i.e. the city at 59° 54'20"N, 30° 16'9"E would be associated with the names:

- *St. Petersburg* from 1991-present day
- *Leningrad* from 1924-1991
- *Petrograd* from 1914-1924

The GazDB can thus export temporally-sensitive gazetteers customized for use in historical documents.

# 3 Geographic names

Geographic names present a number of challenges to a gazetteer. These include issues inherent to translation and transliteration of foreign names, mediation between repeated entries and multiple sources, and the (in)accuracy of placename specifications.

## 3.1 Resolution of names

The first hurdle is internationalization (i18n). Differences between character encodings and display capabilities result in some names taking on a variety of forms (e.g. printing *São Tomé* as *Sao Tome*). Although the printed forms of the name are not character-identical, the name itself has not changed from its original representation.

To resolve this, the GazDB defines and stores a *geographic name* as a triple: [canonical name, display name, search name], with each element at a different level of resolution. The canonical form of the feature's name is kept as a 16 bit string (Unicode / UTF-8), the display form is 8 bits (ISO 8859-1), and the search name is 7-bit uppercase ASCII. These resolutions are appropriate for different purposes: wide characters are necessary for Chinese/Japanese/Korean (CJK) content, the display name is a necessary compromise given the default display capabilities of Internet browsers, and the search name is necessary given the data entry capabilities of the default (US-ASCII) keyboard. We henceforth use the term *name* to implicitly refer to this triple.

We also support Soundex and Metaphone geographic name searches at a 7 bit resolution, by storing the hash codes in separate tables within the GazDB.

However, there are cases when variances in a name arise due to multiple transliteration, rather than character encodings, as in the case of *Macau* and *Macao*. As such, we further define a *spelling* of a geographic name to be a similarly constructed triple of [UTF-8, 8859-1, ASCII] encodings, with the added restriction that while the authoritative name is directly associated to a geographical entity, a spelling is only directly associated to a name. Thus while *Macao* is a spelling variant of *Macau*, and *Macau* is the name of a city in Southern China, nonetheless *Macao* is not considered to be a GazDB name proper for the city.

## 3.2 Authoritativeness

The GazDB also makes a distinction about the authoritativeness of names. We view a placename as an information resource in and of itself, independent of the feature that it names. This is analogous to the Unicode standard, where the name of a character is treated as an information resource independent of the glyph it corresponds to.

There are multiple names that refer to the same geographic feature but are neither spelling variants of another nor are they seemingly derived from one another, such as *Holland* vs. *The Netherlands* or *Nihon* vs. *Japan*. Because of this, we define and maintain *alternate names* for each *authoritative name*. Each geographic entity is permitted to have only one authoritative name, but that authoritative name can have several more informal alternate names associated to it. Both alternate names and authoritative names can have variant spellings.

Conflicts between authoritative names from different sources are inevitable. However, we cannot independently determine the proper solution in an objective way because we are not a mapping agency– we seek to use geographic data, not produce it. Without being able to take our own measurements, resolving these discrepancies must therefore be done on the basis of the perceived trustworthiness of the sources providing the data. The GazDB's source data consists of many sources that can be trusted to varying degrees. We put the highest trust in the Geographic Names Information System (USGS, 2003) data and the GEOnet Names Server (NIMA, 2003) data, and mediate the incorporation of all the other sources accordingly.

To enforce the distinction between the authoritative and the alternate versions of a name,, and to emphasize the authoritative name, we speak of "names" referring only to the authoritative name. For all others, we speak of "alternates" and "spellings".

## 3.3 Explicitness

Lastly, the GazDB distinguishes fully specified geographic names, such as *New York City, New York, USA* from their short forms such as *New York City* or even the more colloquial yet ambiguous *New York*.

The GazDB maintains a taxonomy of geographic features, consisting of an administrative hierarchy of the world. The administrative hierarchy serves to locate geographic entities by country, then state, county, and so forth. This is based upon both the FIPS 10-4 citeFIPS and the ISO 3166-2 (ISO, 1998) codes. However, these standards often disagree and update infrequently, so we base ours upon the Hierarchical Administrative Subdivision Codes (HASC) system (Law, 1999). Using this taxonomy, we can specify geographic entities by name and by their location within the political divisions of the world. The GazDB is capable of maintaining multiple taxonomies for geographic entities, such as one based upon physical features (for instance: "*Mont Blanc* is a mountain in the *Alps* which are in *Europe*", in addition to "*Mont Blanc* is a mountain in *France*"), however these have not yet been completed.

We define as an *authoritative title* the unambiguous list of hierarchical administrative regions that contain the geographic entity. Here *New York State, United States* would be the authoritative title, such that the sequence *New York City, New York State, USA* unambiguously refers to a single geographic entity. The *authoritative title* is the ordered sequence of the authoritative names for the list of hierarchical regions that contain the feature, so it is easy to compute from a hierarchical region tree in the GazDB. Other titles can be computed by using variants or spellings of the containing regions' names, or by omitting some of them (*New York City, USA*, for example).

We have thus imposed an order on the GazDB geographic names: each feature can have one primary (most authoritative) GazDB name and some alternate GazDB names. Each GazDB name, both primary and alternate, can have multiple spellings associated with it. All of the above are available at all three encoding resolutions.

This ordering allows the GazDB to classify geographic names along three orthogonal scales: general/vernacular vs. authoritative; raw (original character encoding) vs. cooked (character-set- and transliteration-normalized); and implicit (short form) vs. explicit (long form). This allows us to export, on an as-needed basis, multiple gazetteers from the GazDB at different name resolutions.

### 3.4 Language information

The multilingual support in the GazDB goes beyond the use of Unicode. To map different name entries to geographic features for different languages, we also maintain within the GazDB a detailed list of the world's languages (Grimes and Grimes, 2000), and associate all names and descriptions with their language.

The GazDB can keep one authoritative name (but arbitrary numbers of associated spellings, variants, and titles) per language in the world for any geographic feature. Therefore, given authoritative sets of raw geographic data in a foreign language, the GazDB could produce a gazetteer in that language. By matching gazetteer entries by feature, the GazDB could potentially issue a multilingual gazetteer as well. Of course, obtaining the large, accurate, geographic datasets in foreign languages required for this purpose is a major ongoing undertaking– one that we make no claim to have completed!

## 4 Geographic features

As mentioned in Section 2, a geographic feature includes both a geographic location and some categorization of what is situated there. The GazDB classifies geographic entities along 3 orthogonal scales: spatial representation, functional class, and administrative type. These classifications allows users to better restrict gazetteer queries, perhaps via pull-down menus, for more relevant results.

### 4.1 Spatial representations

Simple point/bounding-box categorization does not accurately depict the topological footprint of most features (Hill et al., 1999). Points do not represent the geographic extents of locations, and bounding boxes misrepresent features by oversimplifying the shape. Of particular interest is the ability to categorize geographic entities with "fuzzy boundaries", such as the extent of wetlands, or disjoint regions, such as an archipelago. The GazDB classifies features by their footprint into 6 major types (each with numerous subtypes):

1 point – 0-dimensional (approximated to a point, e.g. a factory gate or a well)

2 line – 1-dimensional (e.g. a road or power line)

3 area – 2-dimensional without clearly defined boundaries (e.g. wetlands)

4 point-area – a 2-D region with clearly defined boundaries (e.g. county or lake)

5 cluster of point-areas – e.g. an archipelago

6 probability density distribution – a feature that shifts over time, e.g. ice packs

0 unknown/unclassified

### 4.2 Functional classes

Many features, particularly structures, can also be described by their functional class:

1 building – a man-made structure

2 campus – a feature that contains a number of buildings on open space, such as a military base.

3 field – a feature that predominantly open space without structures, such as a cemetery.

4 city

0 unknown/unclassified

### 4.3 Administrative types

We also distinguish administrative types:

1 international organization – encompasses multiple countries

2 nation

3 province – first-order administrative subdivision within a nation

4 county – first-order administrative subdivision within a province

5 smaller than county – anything below second-order subdivision within a nation

0 unknown/unclassified

It is worth reiterating that these categorizations are deliberately broad and are used for filtering purposes only. The GazDB maintains a complete hierarchical tree of all the administrative subdivisions within a country and the geographic entities contained therein, without any depth limitations.

### 4.4 Using feature categorization

The particular categories and classifications are specified for a number of reasons:

To facilitate Knowledge Representation within the GazDB by axiomatizing how we classify data. We currently have no ontology for the geographic entities, but we leave open the option to add one to our taxonomies.

To reduce the need for human training, such that an average user of the gazetteer can have reasonable expectations of what each category includes based on intuition.

User convenience: the categories in the appropriate pull-down menu should be ones useful to a user.

To make querying more efficient: for example, we can use axiomatic expectation to assume a polygonal feature to only match other polygons.

### 4.5 Storing geographic locations

A major advantage that coordinate systems have over naming systems is that, given an appropriate method, it is possible to convert from one coordinate system to another with reasonable accuracy. As such, the GazDB currently only stores geocoordinates in decimal degrees (albeit in two versions: one high-precision, and the other rounded for display purposes). However, the conversion and export scripts are already prepared to handle a wide variety of coordinate systems, such as Degrees-Minutes-Seconds (DMS), Military Grid Reference System (MGRS), Universal Transverse Mercator (UTM) coordinates, to name a few.

The GazDB scripts can also convert between map projections, but so far it is only done to convert source data into the GazDB standard format.

## 5 Conclusions

Maintaining a large-scale gazetteer database is a nontrivial task. Nonetheless, we have created a gazetteer database containing tens of millions of entries collected from several large gazetteers (each with their own format, encoding, classification, and field conventions), and providing output in several highly compressed binary formats. We believe that the problems we have encountered in designing and building the GazDB are not unique to us, but rather, they are inherent to the task. We therefore hope that others can use the solutions proposed here to some advantage.

## Acknowledgements

We would like to thank Dr. András Kornai for invaluable ideas and support, Dr. Michael Bukatin for technical assistance and caffeine, the anonymous reviewers for providing useful comments, and lastly, Kenneth Baker and Keith Baker for their roles in the development of this project. Thank you.

## References

Edgar Frank Codd. 1970. *A relational model of data for large shared data banks.* Communications of the ACM. 13(6):377–387.

US Department of Commerce, National Institute of Standards and Technology (NIST). 1995. *FIPS PUB 10-4: Countries, dependencies, areas of special sovereignty, and their principal administrative divisions.* `http://www.nima.mil/gns/html/fips10-4.html`

Barbara F. Grimes and Joseph E. Grimes. 2000. *Ethnologue. Volume 1: languages of the world.* SIL International.

Linda L. Hill, James Frew, and Qi Zheng. 1999. *Geographic Names: The implementation of a gazetteer in a georeferenced digital library.* D-Lib Magazine. 5(1).

Linda L. Hill. 2000. *Core elements of digital gazetteers: placenames, categories, and footprints.* In J. Borbinha & T. Baker (Eds.), Research and Advanced Technology for Digital Libraries : Proceedings of the 4th European Conference, ECDL 2000 Lisbon, Portugal. (pp. 280-290)

International Organization for Standardization. 1998. *ISO/IEC 3166 ISO ISO 3166-2:1998 Codes for the representation of names of countries and their subdivisions – Part 2: Country subdivision code.* Published by International Organization for Standardization, Geneva.

Gwillim Law. 1999. *Administrative subdivisions of countries.* McFarland & Company, Inc. http://www.mindspring.com/~gwil/statoids.html

National Imagery and Mapping Agency. 2003. *GEOnet Names Server (GNS).* http://www.nima.mil/gns/html/index.html

United States Geological Survey. 2003. *Geographic Names Information System (GNIS).* http://geonames.usgs.gov/

Erik Rauch, Michael Bukatin, and Kenneth Baker. 2003. *A confidence-based framework for disambiguating geographic terms.* Published in this volume.