MTR 05B0000085

MITRE TECHNICAL REPORT

Design for an Integrated Gazetteer Database

Technical Description and User Guide for A Gazetteer to Support Natural Language Processing Applications

November 2005

Scott Mardis
John Burger

Sponsor:	DTO/NSA	**Contract No.:**	W15P7T-05-C-F600
Dept. No.:	G063	**Project No.:**	0706N7RZ

The views, opinions and/or findings contained in this report are those of The MITRE Corporation and should not be construed as an official Government position, policy, or decision, unless designated by other documentation.

MITRE
Corporate Headquarters
McLean, Virginia

Approved by:

Scott Mardis, Project Manager

Table Of Contents

Abstract

The Integrated Gazetteer Database (IGDB) is a compilation of open-source placename information in the form of an SQL database. It was designed as a full-globe meta-gazetteer supporting information system applications such as question-answering. This technical report documents the IGDB requirements and design. The schema and its intended usage are described in detail

1 Overview

The purpose of the Integrated Gazetteer Database (IGDB) is to provide a comprehensive database of geographic names (and related information) for use in text processing systems, particularly the Question-Answering systems of the ARDA AQUAINT program. It does not purport to be an original source for this type of information but rather an aggregator of it. Toward that end, the IGDB must:

- combine official comprehensive gazetteers of worldwide places, (its core data is a combination of the US NGA and USGS geographic names databases)
- aggregate names in both their original foreign-language form as well as latinate transliterations,
- provide a schema that provides a common view of the disparate sources from which it is drawn.

This resource should provide a stable base from which research on geographic place coreference and normalization can build.

2 Strategy

2.1 Do not add content

While it might be tempting to embellish or correct data from the source gazetteers, such effort is a costly distraction from the principle task of integrating other authoritative geographic information. We plan to strongly resist the temptation to be authors of original placename content in an effort either to improve the accuracy or the coverage of the IGDB. We will actively work to find additional sources of information where they would benefit the larger AQUAINT community. When errors are found in the constituent gazetteers, we believe the best course is to feed correction back to the producers of the source material so that the IGDB, as well as other consumers of the source, will be improved with future releases.

2.2 Accommodate regular updates

The utility of the IGDB depends on its ability to accommodate regular, perhaps frequent, updates of new placename information. In order to do this, the process for adding or updating information must be *well-specified*, *simple,* and *repeatable*. In this document we describe the general update (import) process for primary and secondary sources. Later on we will describe individual choices that have been made in regard to each specific source.

In general, there are two approaches to updating the IGDB from its primary sources. One, we can maintain a version of the original source material that matches the current IG, compare any updated version of the source with the saved original creating a list of modification to make; then make them. Or, two, we can compare an updated source against the existing IGDB itself, merging changes as they are discovered.

The advantage of the first method is that it may be faster under some circumstances to use non-DB comparison techniques to determine what changes have been made. Its disadvantages are several.

- It is essential to keep all versions of the gazetteers that are used to build the IG. Losing one will make this update approach impossible and will require the second approach to make further updates.
- If the format of the original source data changes substantially it may be very difficult to do direct comparisons.

Comparing the source gazetteers with the IGDB frees us of the administrative need to accurately match versions of the IGDB with its sources. It will also be easier to accommodate changes in the format of the sources. Additionally, we should be better able to use the database itself to maintain information about the change history. For these reasons, we have chosen to update the database using the second method, direct comparison between source gazetteer and the IGDB.

2.3 All processing must be automatic

The integration of each of the IGDB sources should be fully automated. Manual processes, given the size and complexity of the source gazetteers, could lead to an unwanted maintenance burden. If automated processing is difficult due to errors or inconsistencies in the sources, an automated patching method should be considered as any manual corrections might need to be applied with every update of a source.

2.4 Flexible representation

Different sources record varying information for each place. Sources may also have fundamentally differing conceptualizations of the relationships between places. For example, they may differ on what containment means. Because of these differences, we want to provide a simple and flexible unifying framework for combining sources in the IGDB. Specifically, we want multiple approaches to containment and coreference to be representable. We also want to be accommodate future extension such as incorporating geospatial polygon data or adjacency information.

2.5 Conservative coreference

In order to allow IGDB users flexibility in how the combined gazetteers are used, we want to make few irrevocable decisions when integrating most sources. Toward this end we have decided not to make the resolution of coreferring places a principal part of the IGDB. That is, if multiple sources contain information about a common place, they will be represented by multiple *place* entities in the IGDB. Coreference between sources (and also within a source, if necessary) will be accomplished with separate DB tables. This will permit both the IGDB team, as well as IGDB users, the ability to experiment with a variety of ways of computing coreferring geographic placenames. It also enables the database to contain multiple inconsistent coreference tables, allowing applications to choose between, perhaps, aggressive and conservative coreference strategies.

2.6 Make integrating new sources relatively easy

The IGDB's principal benefit is derived from its particular combination of source gazetteers and related information. In the long term, this benefit will grow primarily through the incorporation of more sources of a wider variety. All of the above strategies contribute in some way toward this longer term goal of creating an expanding source of placename information.

3 Design

The design of the database is described in the following four sections:

Sources – how they impact the design
Structure – the basic semantic concepts and their relationships
Schema – the specific DB schema
Details – important minutiae

3.1 Sources

We view each source to be added to the IGDB as either primary or auxiliary. Primary sources are those that identify unique places that are unlikely to be found in other sources. Auxiliary sources are those that provide names or additional information about places that are likely to be found in the primary sources.

When a primary source is added, each of its unique places is given a unique id in the IGDB. Coreference between individual places in primary sources is represented in tables separate from the main *place* tables. In this way, the IGDB provides a framework for utilizing the various primary sources either in their original separate forms, or in a combined form. Multiple methods of resolving coreference between primary sources can be experimented with and the results maintained in separate tables.

When auxiliary sources are added, it is assumed that each place in the source already exists in the IGDB. The information in the auxiliary sources annotates the main entries for places in the unified schema. Thus, coreference for places in auxiliary sources is resolved at the time it is added to the database. It is necessary, therefore, to determine how conflicts between information in auxiliary sources will be resolved at the time it is added. Such issues can be deferred for conflicting primary sources.

When planning the addition of a new source, it may be possible to consider adding it as either an auxiliary or primary source. In order to accommodate updates, it is best if primary sources maintain a stable unique place identifier across releases. This permits definitive matching of places across releases and simplifies the update process.

The content of each version of the IGDB can be described, therefore, in terms of its primary and auxiliary sources. The union of places in the primary sources represents the full set of places in the IGDB. Additional information is taken from auxiliary sources to augment the information know about that basic set of places.

3.2 Structure

Gazetteers and other geographic information sources vary somewhat in their use of terminology. Here we describe the way we will use these terms and the way in which they relate to concepts in the sources. Each of these concepts has a specific realization in the schema of the IGDB.

3.2.1 Places

A place is the primary concept in the IGDB. Anything having a relatively fixed location or extent, that is referred to by name can be a place. In general, the IGDB concept of place is derivative of the “place” concepts in the source material. Note that not all sources refer to this concept as “place”. The USGS gazetteer, for example, refers to each nameable place as a feature. In this document, and in all referring to the IGDB, we will refer to these nameable locations or areas as *places*.

When gazetteers are incorporated into the IGDB each unique place in the source gets a unique *place id* in the IGDB. These are the unique handles for referring to places with the database.

3.2.2 Names

All of the places in the IGDB can be referred to by one or more names. Places typically identified by description or address (e.g., the corner of 40th Street and California Avenue in Chicago, IL) are not represented in the IGDB or its sources.

Each place has a primary name, which is typically the official name given to a place by the appropriate government authority or by consistent popular usage. In general, most of our sources identify the primary name for each place. For sources that do not, we will document how the primary name was selected from those available.

The IGDB is designed to support collecting multiple names for each place, including names in non-latinate scripts. The present version of the IGDB includes small samples of Chinese and Arabic names derived from online bilingual resources..

3.2.3 Types

Associated with each place in a gazetteer is a type that indicates what sort of place is being named (e.g., city, mountain range, river, etc.). We will be using the geographic type taxonomy from the Alexandria Digital Libraries (ADL) project as a unifying taxonomy of place type. During the integration of each source, we will be mapping from the source's type system to the ADL type system. Our initial strategy is to use fixed maps for each source. These mappings are given in the *source*`FeatureTypeMappings` tables described with the Maintenance Tables in section 3.3. More sophisticated approaches may be considered later if deemed necessary.

3.2.4 Containers

One of the key relations between different places is that of containment. For example, Central Park is contained in New York City, which is contained in New York State, which is contained in the United States of America. Unfortunately not everything is so clear cut. There are actually two types of containment that people intermix quite readily: geospatial containment and geopolitical containment. For example, Kaliningrad is politically part of the Russian Federation but it is physically separated from Russia proper by Lithuania, Latvia and Belarus. Gazetteers generally encode geopolitical containment for governmentally administered regions. Because of this, caution should be used in drawing conclusions from containment data as containment may not be transitive as one would expect. The USA is "generally" contained by North America and Hawaii is contained by the USA, but Hawaii is not contained by North America.

Each gazetteer has its own approach to containment. The USGS gazetteer defines all of its features (places) in terms of the counties that contain it. As a result, some very large places, such as The Great Plains, are described as being in a great number of states and counties. Intuitively, one might general expect small places (counties) to be contained in large places (The Great Plains) and not the other way around. This problem is exacerbated in the USGS Gazetteer because it is a combination of data reported by each state. As a result, the granularity at which containment is reported varies from state to state.

As with other issues, we defer to our sources, following their containment strategies; which in aggregate might be better viewed as an *overlap* relation. For each source, we document how the containment relation is derived from the attribute given in the source.

3.2.5 Auxiliary Information

Information about a place other than its type, names, location, and containment, is considered to be auxiliary information. (E.g., population, shape, languages, etc.) Such information does not greatly impact the core unifying database schema and is maintained largely in its original form in tables that reflect each individual source's format.

3.3 DB Schema

This section describes the schema from a high-level, organizing its many tables into coherent groups that clarify the structure and function of each. The current schema can be found in the `igdbschema_`*`database`*`.sql` files in the distribution. Tables within the database can be divided into four relatively distinct groups:

Source tables – information direct from the original sources
Unified schema tables – the core schema unifying the disparate sources
Place code tables – standardized codes referring to places
Maintenance tables – information to support and record updates

Source tables are those that primarily contain information direct from the original sources with little modification other than form. For primary (and some auxiliary) sources, there are tables in the IGDB that, more or less, directly represent all the relevant information from that source. Because of a desire to maintain SQL table constraints, it is often necessary to use more than one table to represent each source. All tables related to a given source will be named with the same prefix (e.g., "usgs" for the USGS placename gazetteer). All primary sources have at least two tables:

an *auxiliary* table containing all of the core imported information from that source,

a *mappings* table relating identifiers (IDs) of places as given in the auxiliary table to the IGDB canonical place ID given in the `gazPlaces` table,

In the present release, the following tables are source tables:

`usgsAuxiliary`, `usgsMappings`,
`ngaAuxiliary`, `ngsMappings`,
`tipsterAuxiliary`, `tipsterMappings`

Unified schema tables constitute the core of the IGDB where the separate gazetteers are unified under a common type system and share an identifier scheme. The principle tables of the unified schema are:

`gazPlaces`, `gazPlaceNames`, `gazFeatureTypes`, `gazNamings`,
`gazContainers`, `gazEpochs`

`gazPlaces` – This table enumerates all the places identified by the *primary* sources. If a particular physical place is identified by two primary sources, then it will have multiple entries in the `gazPlaces` table. Coreference between places identified by `gazPlaces` records can be recorded in other tables as needed.

`gazPlaceNames` – This table enumerates all names of places. Each is represented in a unicode string encoded in UTF8 and normalized using the NFKD definition (compatible decomposition). All strings in the unified gazetteer appear only in the gazPlaceNames table and are referred to elsewhere only by the placeNameID. No two unicode-equivalent strings should appear in separate records of this table. As the NFKD specifies, all diacritics will be represented in compositional form using unicode combining characters, and all ligatures will be decomposed into their constituent characters.

The lang and script fields are used to record, where appropriate the language and script of the placeName. In general, it should only be necessary to so identify non-English languages and non-latinate scripts.

gazNamings – This table associate names with places using IDs to refer to records in each of the gazPlaces and gazPlaceNames tables. A primary name for a place is also identified directly in the gazPlaces table. All other names are associated with places relationally in the gazNamings table.

Places and names for them change over time. For example, the city of St. Petersburg, Russia was known as Leningrad, U.S.S.R prior to 1991. In some cases, information about the time domain of a place-name association is available in the source gazetteers. We will represent the restricted time-frame of valid reference using an epoch and associate the epoch with the place-name association in the gazNamings table. The **gazEpochs** table will be used to identify an epoch, indicating start and end dates (possibly underspecified) where possible.

gazContainers – This table records the containment relation for places in the gazPlaces table. The released version of the IGDB will have containment relationship only between places identified by the same source. That is, for example, the Manitoba place from the NGA and USGS sources will be contained by the respective Canada places of the same source. Collapsing the containment relation resulting from coreference of gazPlaces will be address in future releases or potentially not at all.

gazFeatureTypes – This table contains the main typology for places that unified the typing schemes of the constituent databases. It is based almost entirely on the Alexandria Digital Library (ADL) ontology for geographic places. When each source is integrated into the IGDB, a mapping from the source type system into the ADL type system is performed. The featureType field of the gazPlaces table refers to an entry in this table.

Place Code tables collect authorized identifiers and codes used by official agencies to uniquely identify geographic entities – principally countries and states (first order sub-nation administrative regions). We have attempted to collect several of the country and state code lists into a single table. Each mapping between a code and a place indicates the standard that defines the association.

gazStandards – This table lists the standards used by various official agencies and assigns each a standardID used in the gazStandardsMappings table.

gazStandardsMappings – This table lists standard codes for places. A given place may be assigned separate codes by different standards.

Maintenance tables are used in for update and administrative purposes. There is one primary maintenance table (gazUpdateSessions) and several support tables associated with various IGDB sources.

gazUpdateSessions – This table records specific information about an update event for the IGDB. Such an event is called a session and it is associated with information that uniquely identifies the source data used in the update. The table also records the time of the update, the user performing the update, as well as any additional comments.

source**Updates** – When information in a source auxiliary table is modified during an update session, that event is captured in a source-specific "update" table. Each record in the update table associates an update session with a place using the source-specific place ID of the corresponding auxiliary table.

source**FeatureTypeMappings** – Each feature-mapping table associates feature types as given in a particular source to the ADL feature types as given in the gazFeatureTypes table. Where possible we will try to maintain a strict mapping between source-specific feature types and the unifying ADL framework. In the future, we may determine that a more complex association between the type structures is necessary. Even under these circumstances we will maintain a feature-mapping table that contains the bulk of the declarative data describing the mapping.

Cumulatively, our present release has the following maintenance tables:

gazUpdateSessions,
usgsUpdates, usgsFeatureMappings,
ngaUpdates, ngaFeatureMappings,
tipsterUpdates, tipsterFeatureMappings

3.4 Details

Place identifiers

Most true gazetteers assign each place some type of unique identifier, typically an integer. The NGA and USGS use the positive and negative integers as identifiers. The unified schema of IGDB also assigns such identifiers: the *place ID*, usually referred to as the gazPlaceID within the schema. One of the important consideration in creating the IGDB was the need for an ID strategy that kept the IDs stable over multiple releases of the database.

Both the NGA and USGS gazetteers are released with IDs that are stable across releases (modulo necessary changes). This type of stability is important for human-language technology applications because it is necessary to annotate texts (manually and automatically) with proper place references that do not become obsolete with a new release of the gazetteer. It is, therefore, a necessary part of the IGDB that the gazPlaceIDs remain as stable as possible between releases. This is strongly facilitated by the stability of the IDs within the source gazetteers. Source gazetteers without stable identifiers (Wordnet) cause added difficulty for maintaining the stable ID property.

So, in considering the importation of any given source, it is necessary to identify the best candidate for use as an input identifier and apply a process that derives stable gazPlaceIDs from those. We have adopted a block allocation scheme for gazPlaceIDs that assigns a block or set of blocks to each primary source and maps contiguous blocks of the source ID space into a block of gazPlaceID space. In Section 5 we describe the details of the identifier scheme for each primary gazetteer.

4 Using the IGDB

4.1 Forms of the database

The IGDB could be used with several different DB management systems. At present, we distribute only distribution for MySQL on Windows and Linux platforms. A version for PostgreSQL could be made available if there were sufficient demand. You should obtain the distribution that matches the server and platform you intend to run.

4.2 Using MySQL

The IGDB is compatible with the MySQL database management system. Although the database has been tested to load against MySQL 3.23 and 4.0, UTF8 support does not exist in MySQL before version 4.1. Caseless string matching may be inaccurate when using databases that do not have full UTF8 support.

Scripts to initialize and load your MySQL database are included in the MySQL distribution. Consult the install documentation in your distribution to see how to load MySQL from your platform.

5 Source gazetteer strategies

The current version (IGDB 1.1) contains the following sources. Import issues specific to each are described in subsections below.

NGA Gazetteer	Primary
USGS Gazetteer	Primary
TIPSTER Gazetteer	Primary
NMSU Arabic Names	Secondary
Harvard U Chinese Names	Secondary

5.1 Primary Sources

5.1.1 NGA

Identifiers

The UFI (Unique Feature ID) is the best identifier of place within the NGA gazetteer. We will use these as the NGA source ID within the IGDB. Since UFIs are integers, the place ID (PID) for NGA entities will be allocated according to the general block allocation scheme described in section Section 3.4. At present two blocks of 10,000,000 identifiers are allocated to the NGA ID space.

5.1.2 USGS

Identifiers

The USGID is the best identifier of place within the USGS gazetteer. We will use these as the NGA source ID within the IGDB. Since USGSIDs are integers, the place ID (PID) for USGS entities will be allocated according to the general block allocation scheme described in section Section 3.4. The USGS ID require a single block of 10,000,000 identifiers.

5.1.3 CIA World Factbook

Placename information from the CIA factbook will be included in subsequent IGDB releases.

5.1.4 TIPSTER

Identifiers

The TIPSTER Gazetteer includes no place identifiers. It is arranged as a single file with each line containing information about some place. Because it is a static resource, unlikely to be updated, we have chosen to treat the line number (starting with 1) as a default source identifier for places within the TIPSTER Gazetteer. The place ID (PID) for TIPSTER entities will therefore be allocated according to the general block allocation scheme described in section Section 3.4. At present one blocks of 1,000,000 identifiers is allocated to the TIPSTER ID space.

5.1.5 Wordnet

Placename information from Wordnet will be included in subsequent IGDB releases.

Wordnet’s internal IDs (hereafter called byte position indicators - BPI) are based upon the byte-position of the synset record in Wordnet data files. As a result, the IDs for a synsets change across releases of Wordnet; they are not stable. We are presently working on a mechanism to maintain stable identifiers for the Wordnet synsets when imported into the IGDB.

5.2 Secondary Sources

5.2.1 Bilingual Onimastica

Generally, bilingual onimastica will be imported as auxiliary sources within the IGDB. They typically provide only additional names for places already existing within the main NGA and USGS gazetteers. At the time of import, it is necessary to identify the appropriate place or places that each given foreign-script name applies to. The typical process involves matching the given latinate placename against the names in the IGDB and controlling for given constraints such as the type of the entity, its size, and/or relationships to other entities.

This process is imperfect and may result in occasionally erroneous assignment of names to places. We seek feedback on the accuracy of these methods and will incorporate additional methods to improve accuracy where possible. The section below briefly describes issues related to specific onimastica.

Arabic-English

Our Arabic placenames are derived from two sources: and English-Arabic dictionary from arabeye.org and the Arabic-English dictionary from NMSU. From these dictionaries were first extracted all pairs with English terms that matched an entry in the TIPSTER gazetteer. The reduced list included many common nouns that also happen to be placenames of English-named places. In order to eliminate all possible common noun translations we further restricted our set to pairs that were reasonable transliterations of one another between Arabic and English. This might not have been possible if there were many cognates between English and Arabic.

We applied the Arabic names, so derived, to all places in the IGDB that had a primary or alternate name that matched the English term of the pair. This may have overgeneralized and we are looking for feedback in order to refine the import procedure.

Chinese-English

Our Chinese placename information is derived from the Chinese Historical GIS collection available from the Faculty of Arts and Sciences at Harvard University. This data is available online at http://www.fas.harvard.edu/~chgis/work/downloads/. Because this dataset was already restricted to placenames, no filtering was necessary to eliminate common nouns.

Our initial attempt at incorporating Chinese placename information (in the current release) applies the same process as our approach to Arabic. We have inspected the result and found a number of systematic errors which we plan to correct in future releases. In particular, some Chinese names include information about the type of entity being named (e.g., city) that are not present in the English term. It is inappropriate to apply these names to places of other types. Other errors occur because Chinese has alternate names for specific places (e.g., the country Burma – now Myanmar) that apply uniquely to one place and not to all of the given latin name.

Appendix A – Sample Queries

To find all places named “New York”:

```
select * from gazPlaces, gazNamings, gazPlaceNames
where gazPlaces.gazPlaceID = gazNamings.gazPlaceID
and gazNamings.placeNameID = gazPlaceNames.placeNameID
and placeName = 'New York';
```

To find all places that contain entities with primary name “New York”:

```
select * from gazPlaces as contained, gazPlaces as container,
     gazContainers, gazPlaceNames
where gazContainers.gazplaceid = contained.gazplaceid
and gazContainers.containerid = container.gazplaceid
and contained.primaryName = gazPlaceNames.placeNameID
and gazPlaceNames.placeName = 'New York';
```

To find the Chinese names for all places named “Beijing”:

```
select * from gazPlaces, gazNamings as egn, gazPlaceNames as epn,
gazNamings as zgn, gazPlaceNames as zpn
where gazPlaces.gazPlaceID = egn.gazPlaceID
and egn.placeNameID = epn.placeNameID
and gazPlaces.gazPlaceID = zgn.gazPlaceID
and zgn.placeNameID = zpn.placeNameID
and epn.placeName = 'Beijing'
and zpn.lang = 'ZH';
```

Distribution List

Internal

Dave Anderson
Donald Batkins
James Burnetti
David Day
Glenn Geoghegan
Rod Holland
Frank Linton
Inderjeet Mani
Marc Ubaldino

Project

John Burger
Scott Mardis

External

AQUAINT Phase 2 Executive Panel
AQUAINT Phase 2 Principal Investigators

Beth Sundheim, SPAWAR SSC
Doug Caldwell, US Army Topographic Engineering Center