Importing the US Geographical Survey Data into a Relational DBMS

Technical Report: Version 1

Peter McBrien
Dept. of Computing, Imperial College,
180 Queen's Gate, London SW7 2BZ
pjm@doc.ic.ac.uk

Tuesday 9th April 2002

1 Introduction

This report documents the relational database version prepared by the author of certain parts of the data prepared by US Geographical Survey called the (Geographical Names Information System). As for the US 1990 Census data relational database import presented in [McB02], the database is mainly intended to be used for project work in the **Dept. of Computing (DOC)**, Imperial College, that requires a medium sized 'real world' database to work on. However, others may find the database of use for similar work, and the Java program used to build the relational database may prove useful to anyone who wishes to import the US Geographical Survey data into a relational database.

Note that like almost all real world data, there are errors and inconsistencies in the USGS . As far as is practical, the relational database version of the data has attempted to preserve the original structure and content of the data made available from the USGS as a set of flat files.

2 Summary of Data

This section presents a summary of the information present in the relational database after importation using the Java program related to this report. An overview is given in ER notation in Figure 1. The details presented below of the various attributes is largely a copy of the information presented in the readme files on the ftp GNIS download site, altered to reflect the manner in which Java program processes the information.

2.1 feature

Obtained from http://mapping.usgs.gov/pub/gnis/us_concise, the feature table contains around 40,000 rows, detailing major geographical features in the US, such as lakes, cities, *etc.*

- name: the common name of the feature

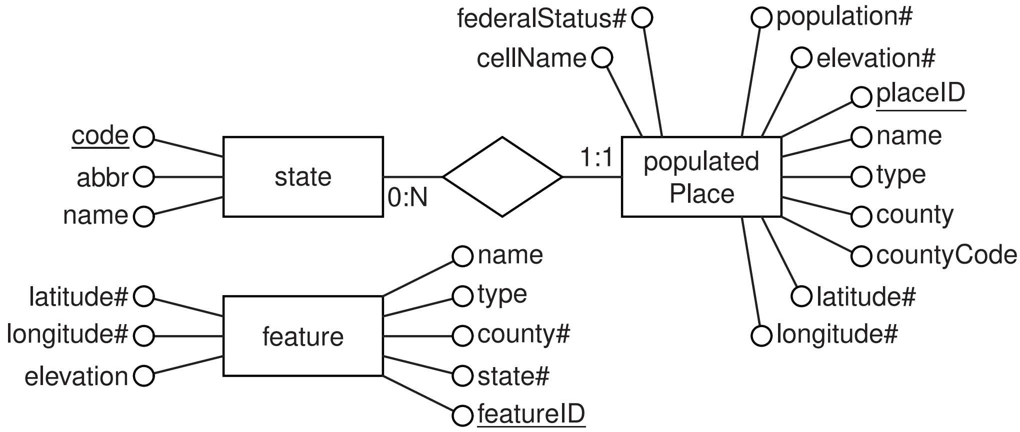

Figure 1: ER representation of the USGS data

- type: *e.g.* lake, ppl, mount. Note that the original data file has one row with this field set as empty, which is ignored by the import program.
- county: name of county in which the feature is present
- state: name of state in which the feature appears. The notion of state is rather loose, and includes such entries as 'Canada-U.S.' for areas of water between the US and Canada. Also, each state appears as an entry in features, and has this field null. Thus, there is no foreign key from this field to the name field of state, which one might expect to have been present.
- latitude: in degrees, + or - denoting N or S latitude, respectively
- longitude: in degrees, + or - denoting E or W longitude, respectively
- elevation: in feet
- placeID: some places have the same name and type, and since all the other fields may be null, there is no relational key that can be formed from any combination of the attributes. Thus a unique identifier is generated by the import program. If one wants to get a true version of the original data one should ignore this field.

2.2 populatedPlace

Obtained from http://mapping.usgs.gov/pub/gnis/Pop_places_deci, the populatedPlace table details all settlements in the US that have any official status, and contains about 164,000 rows.

- name: name of the populated place.
- type: *e.g.* lake, ppl, mount
- county: name of county in which the feature is present
- countyCode: FIPS code for county
- stateCode: FIPS code for state
- latitude: in degrees, + or - denoting N or S latitude, respectively
- longitude: in degrees, + or - denoting E or W longitude, respectively

- elevation: in feet
- population
- federalStatus
- cellName
- placeID: some places have null for latitude and longitude, but there are also places with the same name and type within the same country. Thus there appears to be no key that can be formed from the fields in the data file, and instead an unique identifier is generated by the import program. As for the features table, if one wants to get a true version of the original data one should ignore this field.

2.3 state

This is *not* provided by the USGS, but is a listing prepared by the author, of those states of the US and its protectorates, and includes all those states which appear in the populatedPlace table. It can be obtained from http://www.doc.ic.ac.uk/ pjm/testdbs/states.txt

- code: State FIPS two digit code
- abbr: State FIPS two letter code
- name: Textual name of state

3 Using the Imperial College version of the Database

A version of the database is maintained in a Sybase DBMS within DOC, for use only by members of the department. Any users of the database are expected not to make excessive use of the server.

To use the database, called `usgs`, you should use the jConnect JDBC driver (available from the Sybase web site) to connect with the Sybase database, and use the following details in a `Connection` class:

- `url`: jdbc:sybase:Tds:uranium.doc.ic.ac.uk:4100/usgs
- `username`: guest
- `password`: guestuser

References

[McB02] P.J. McBrien. Importing the US 1990 census data into a relational DBMS. Technical report, Dept. of Computing, Imperial College London, 2002.