GEOGRAPHIC NAMES INFORMATION SYSTEM: AN AUTOMATED PROCEDURE OF DATA VERIFICATION

By Roger L. Payne
U.S. Geological Survey
523 National Center
Reston, Virginia 22092

ABSTRACT

The U.S. Geological Survey has researched and developed an automated geographic names data base to help meet national needs and to provide informational and technical assistance to its mapping program. The Geographic Names Information System (GNIS) is currently capable of providing data for approximately two million names and their related features in the United States and its territories. Cartographers can retrieve, manipulate, and organize name information to meet many mapmaking requirements. An automated mapping procedure has been developed to verify data submitted during the initial compilation of GNIS. This procedure has been expanded to assist the new provisional mapping program of the USGS, and it has proven to be valuable preliminary research for a program of automated typesetting and name placement. Also, research is being conducted toward the eventual incorporation of information from the System into the Survey's Digital Cartographic Data Base.

INTRODUCTION

The Geographic Names Information System (GNIS) is a multi-purpose data system designed to meet major national needs. Versatility is a basic tenet in the design of the system which affords great utility to a wide variety of users. Users may be grouped into two categories: (1) those who use the information directly for research or reference, and (2) those who reformat retrieved data for individual or specialized use. Information from the system may be retrieved, manipulated, analyzed, and organized to meet the general or specialized needs of a wide variety of users involved in research or application. There are currently five data bases in the system:

(1) Geographic Names Data Base
(2) Topographic Maps Data Base
(3) Designator/Generic Data Base
(4) National Atlas Data Base
(5) Board on Geographic Names* Decisions Data Base.

This paper will refer to the Geographic Names Data Base throughout. The purpose of this paper is to provide an overview of the development of the Geographic Names Data

*The U.S. Board on Geographic Names was created in 1890 and established in its present form by Public Law in 1947, and is authorized to establish and maintain uniform geographic name usage throughout the Federal Government.

Base, to describe the method used to verify the data compiled from U.S. Geological Survey topographic maps, and to identify and project some cartographic applications of the Geographic Names Data Base.

DATA BASE DEVELOPMENT

Research and initial compilation of GNIS was begun in 1968. GNIS data were collected as time permitted, resulting in the completion of data compilation of geographic names for the Commonwealth of Massachusetts and the State of Rhode Island as well as the automation of the data embodied within U.S. Geological Survey Professional Paper 567, The Dictionary of Alaska Place Names (Orth, 1967). In 1976 the geographic names in Kansas and Colorado were compiled and verified as a pilot project to determine the feasibility of compilation on a national scale. After analysis and a favorable evaluation of this pilot project, name data for the remaining States and territories were compiled from 1978 through 1981. This period of initial compilation is referred to as Phase One, consisting of the input to the data base of all geographic names and related information found on the U.S. Geological Survey's 7.5-minute topographic map series except roads and highways, communication towers, and triangulation stations. The 7.5-minute series was used because it was the largest scale available. In the absence of published 7.5-minute maps, 15-minute, 1:62,500-scale topographic maps were utilized, and when there was no coverage by either scale, the 1:250,000-scale topographic maps were used. Compilation of geographic names and attribute information was completed by a private contractor and the collected data were delivered to the Geological Survey on magnetic tape on a State-by-State basis. Geographic coordinates of named features were digitized to the nearest second of latitude and longitude. Coordinates were taken at the center of areal features and the mouth of linear features and were called primary coordinates.

If a feature was not contained entirely within the bounds of one map geographic coordinates (called secondary coordinates) were also digitized at some point on or along the feature for every map through which the feature passed. Other attribute information such as feature class, county, elevation, and map was keyed from annotated topographic maps directly to a magnetic disk.

METHOD OF DATA VERIFICATION

To ensure the integrity of data received from the contractor, names contained on 10 percent of the 7.5-minute maps used in compilation were verified. Data verification required the retrieval of names and their attributes from the data base for direct comparison to the 7.5-minute topographic map.

Comparing name information contained in the data base in the form of a computer printout can be a time consuming

process, mostly by time spent locating the name on the appropriate map and computing the geographic coordinate by hand for comparison with the digitized coordinate appearing in the printout. To minimize the time spent in this vital process, a form of automated name placement was used based upon the location of the feature according to the digitized geographic coordinates and relative to the position on the particular 7.5-minute map.

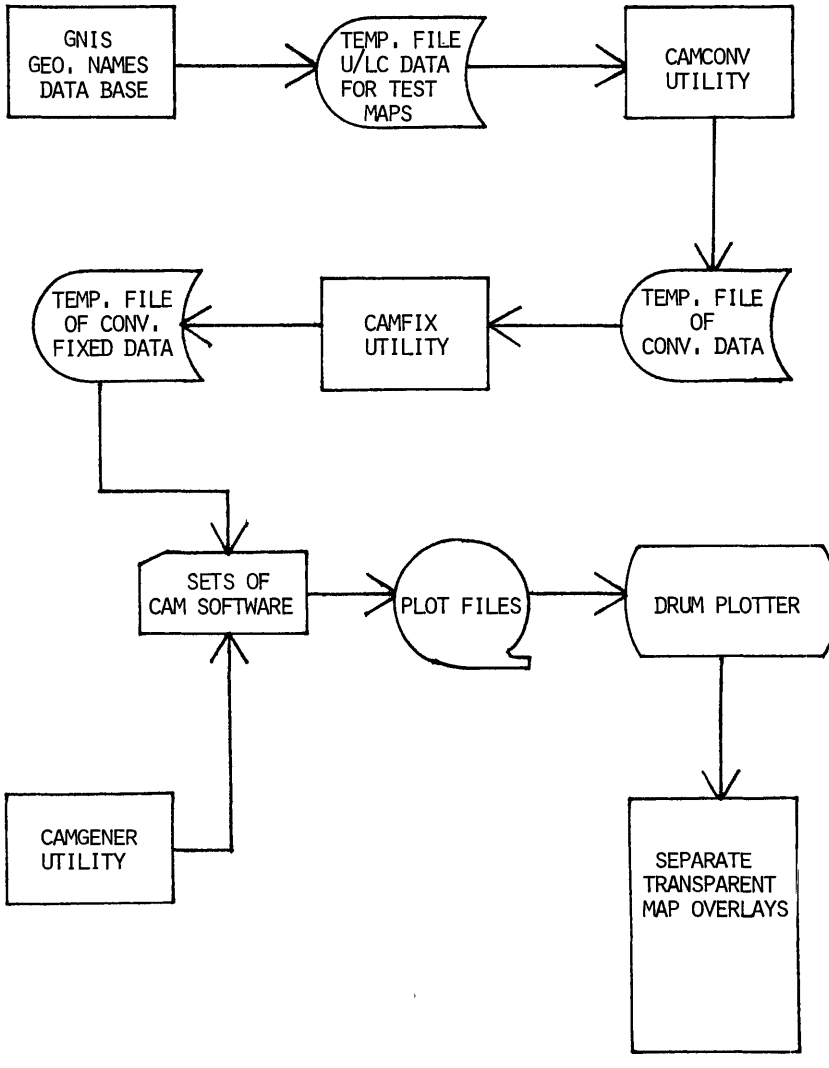
The automated name placement was accomplished through the use of the Computer Automated Mapping (CAM) software package. Although CAM is designed to be most effective when used at smaller scales, its use in plotting point data is also accurate at the scales of 1:24,000 and 1:62,500. Using the digitized coordinates as input to CAM, a selected symbol (*) was plotted to indentify the primary coordinates of each feature on a transparent map overlay while a different symbol (Ø) was used to indicate the source of linear features.

A computer program was written to generate individual sets of CAM software peculiar to the maps selected for sampling. The southeast corner coordinate of the desired maps were supplied and the utility software generated the necessary center point, boundary coordinates, map projection, and other individual map information required by CAM to produce the transparent map overlay.

The data base was then queried using the GNIS data base management system and data were retrieved for the sampled topographic maps and written as one temporary data file. The selected GNIS data were then passed through utility software to convert the name information from upper/lower-case to all uppercase. This conversion was necessary because the available version of CAM could utilize only uppercase information. The previously generated CAM software was used to select data from the temporary file according to map specifications and generate sequential files on a magnetic tape for each map selected. Any number of tapes may be generated simultaneously using this process. The resulting plot tapes contained plotting information and geographic name data for the sampled maps. A flow diagram of the data verification procedure is shown in Figure 1.

IMPLEMENTATION OF THE VERIFICATION PROCEDURE

A separate map overlay plot for each map selected for testing was generated on the drum plotter. The transparent map overlay with the plotted information was then laid directly over the corresponding 1:24,000- or 1:62,500-scale map. The researcher could locate the named feature at a glance because the geographic name was plotted directly beside the symbol where the digitized geographic coordinate of the feature was plotted. One could then determine by the placement of the symbol if the digitized coordinates were correct or within the accepted tolerance of ± 5 seconds of latitude or longitude. In this fashion the laborious



CAMGENER:	Generates sets of complete CAM software for individual test maps.
CAMCONV:	Generates all uppercase data from upper/lowercase data because CAM requires uppercase data.
CAMFIX:	Moves the printing of geographic names to avoid overprinting--does not move the symbol which indicates the actual point of the digitized coordinate.

Figure 1.--The automated data verification procedure, which may be accomplished simultaneously for any number of tapes.

task of locating the name on the map was minimized. An accompanying printout allowed other attribute information to be verified.

Often, a particular area or section of a topographic map would have a dense pattern of geographic names even at the large scales being used. The problem of overprinting of geographic names required additional attention because the available version of CAM could not test for overprinting. To increase the utility of the verification process, software was developed to overcome this problem. The area within ±2 seconds of latitude and the area within 15 seconds of longitude was tested for overprinting. If the printing of two or more geographic names were found to be in conflict then the printing of the names was moved ±3 seconds of latitude. There was no longitudinal shift. The placement of the symbol remained at its exact location. This procedure was effective in almost all instances and allowed for increased readability and continued efficiency in locating the named feature. The overprinting utility was the last operation performed on the selected data before input to CAM.

For the purpose of accepting data supplied by the contractor, a weighted point system for all possible errors was developed. Each plotted map was scored according to the weighted point system and the number of error points was calculated against the total number of names on the map. In order to be acceptable, the error rate per map could not exceed 5 percent.

ADDITIONAL APPLICATIONS

Elements of this geographic name data verification procedure are being used by the U.S. Geological Survey's National Mapping Division to assist in its mapping program. One application is the use of transparent map overlays of geographic names data as guides in field research of geographic names. Also, subsets of the names data may be retrieved for specific applications; one may wish to plot only the names of streams or often only names that required action by the U.S. Board on Geographic Names. An additional option is being incorporated into the procedure. While not necessary for testing the validity of contractor data, some researchers have expressed an interest in plotting secondary geographic coordinates. Software is being developed to incorporate secondary coordinates into the plotting procedure. This will allow researchers to visualize the general trend or scope of large features.

Phase Two of the development of the data base includes inputting to the system names and related information found on all available source maps and texts, but not recorded on USGS topographic maps. Phase Two compilation is complete for the States of New Jersey and Delaware. As more States complete Phase Two compilation the map overlays resulting from the verification process will be

invaluable in determining names for features in areas not mapped in great detail.

The procedures developed for testing the validity of GNIS data have proven worthwhile as a preliminary tool in the National Mapping Division's current research into automated name placement. Research is also being conducted toward the compatibility of GNIS information with that of the Digital Cartographic Data Base for eventual incorporation of selected GNIS data.

SUMMARY

The initial compilation of the geographic names data base required a completely automated procedure for verifying geographic names data digitized by a private contractor. The volume and nature of the data required the concatenation of utility programs with GNIS and CAM in order to automate the complex procedure required to verify the data. The products were transparent overlays containing symbols at the point where the primary geographic coordinates were digitized and the geographic name printed beside the symbol. The use of transparent map overlays provided for rapid and efficient location of the named feature with absolute verification of the coordinates, and the accompanying printout provided verification of attribute data. The procedure developed for testing geographic names data has provided valuable research and assistance in the development of automated name placement and has identified some possible compatibility problems in incorporating geographic names data into the Digital Cartographic Data Base.