



Economic and Social Council

Distr.
LIMITED

E/CONF.79/L.16
23 June 1987

ENGLISH ONLY

FIFTH UNITED NATIONS CONFERENCE
ON THE STANDARDIZATION OF
GEOGRAPHICAL NAMES
Montreal, 18-31 August 1987
Item 6 (c) of the provisional agenda*

TOPONYMIC DATA FILES

AUTOMATIC DATA PROCESSING (ADP) SYSTEMS

Update on the Geographic Names Information System

Paper submitted by the United States of America**

* E/CONF.79/1.

** Prepared by Roger L. Payne, United States Geological Survey.

The Geographic Names Information System is an automated system designed to be the official geographical names depository of the United States of America and to serve as a toponymic research tool. The computer-based system meets a broad spectrum of information and program needs including the national standardization of geographical names and furthering the efforts of the National Mapping Program. The purpose of the paper is to describe the changes in philosophy, methodology, and procedures during the past five years, and to comment on the future direction of development and application of the Geographic Names Information System.

INTRODUCTION

The need for an efficient means of processing domestic geographical names had been apparent since the mid-1950's. The concept of a centralized automated names processing system for domestic names in the United States began to crystallize about 1970. By 1974 initial data base planning, software and hardware selection, and prototype compilation of the names in one State were complete. Development of the Geographic Names Information System (GNIS) may be described by examining four separate but related areas: data gathering and encoding, nature of the data, hardware and software, and training and products.

DATA GATHERING AND ENCODING

The initial philosophy for identification of names to be gathered for a computer-based file limited the types or categories of names to be included to only the names within the purview of the United States Board on Geographic Names (USBGN). This limited compilation meant that, generally, only physical features and certain cultural or manmade features were included. Names bestowed by organizations with administrative responsibility were not within the purview of the USBGN and, therefore, were not included in the compilation of the prototype file. Completion of the prototype file was followed by a period of analysis and discussion among toponymists.

By the early 1970's, there was general agreement that a centralized automated system should be developed to house a national geographical names depository, and the range of possible applications should not be constricted by arbitrary data-gathering techniques. The United States Geological Survey (USGS), in cooperation with USBGN, approved the concept of the national names depository and in 1976 began systematic compilation of toponymic data.

The scale of the task required long-range planning. Encoding names from the large-scale topographic maps of USGS was determined to be the most efficient means of establishing the National Geographic Names Data Base, the largest of four data bases in GNIS. From 1976 through 1981 most of the names found on the approximately 55,000 topographic maps of the United States were identified, encoded, and added to the data base. Some

categories of names were excluded because more complete sources were available in digital form from other Federal agencies. This initial compilation effort is referred to as Phase I.

The second phase of compilation, termed Phase II, was begun in late 1981. Only about 50 percent of the known names are shown on USGS topographic maps; therefore, additional sources were to be checked during this second phase of data gathering. The original purpose of Phase II was to identify names that were not shown on USGS topographic maps but were published on other Federal maps and were used in documents. Even after an examination of these materials, it was known that the data base was still incomplete with respect to other known published names. At this point (1982), in order to meet the demands of the user community and to satisfy the requirements of public law 242-80 (U.S. Congress, 1947), the Phase II requirements were changed to include additional materials. These additional materials included maps and documents of the various States of the United States as well as nongovernmental sources of maps and texts that describe names. Additionally, because of requests from the user community, historical documents were also included. After the second phase of compilation is completed, most all published categories of named features will be present except roads and highways. The need for file completeness as well as accomplishment of the task in a reasonable time required a magnitude of effort beyond the means of the GNIS staff. Discussions with various toponymists and State officials led to the implementation of a program whereby interested State agencies and universities would assist in Phase II compilation of their respective State's geographical names file. Formalization of this program led to

contractual agreements for compiling names in support of the National Geographic Names Data Base. This Phase II compilation is being done in each State in accordance with the methodology and procedures established and set forth in The National Geographic Names Data Base: Phase II Instructions (Orth and Payne, 1984). After completion of this second phase of compilation of a State, the USGS publishes the State volume as part of the National Gazetteer of the United States of America (USGS, 1982) in cooperation with USBGN.

NATURE OF THE DATA

The nature and collection procedures of the data base have evolved and expanded during the past . years to accommodate the growing applications and needs of GNIS users. Originally, categories of information associated with names were mainly locative in nature and were coded in some cases for eventual interchange among the digital-data user community. To this end, the first phase of compilation included the following categories of data:

- o entity name - the single name reference of a particular feature
- o type of feature - a broad grouping of features with similar attributes primarily for purposes of data base search and retrieval
- o first-order civil subdivision - the counties or county equivalents in which the feature is located, represented as a five-digit code for digital data interchange

- o geographical coordinates - the location of the feature by latitude and longitude, represented to degrees, minutes, and seconds
- o topographic map - the topographic maps on which the feature is shown, represented as a four-digit code for inventory and digital data exchange
- o elevation - elevations published on the maps for named features

Evolution of the data base over the past 5 years centered on enhancing existing categories of data for publication purposes and on the addition of certain categories of information based on specific user requests and application requirements. Categories enhanced for publication and increased user friendliness include:

- o first-order civil division - the name of each civil division to supplement the five-digit code
- o topographic map - the names of the topographic maps on which each feature is located to supplement the map code
- o elevation - elevations for most features

The categories added because of user-application requirements include:

- o variant names - any other name (spelling or form) by which a feature was known in the past or may be currently known, recorded and cross-referenced to the primary or official name
- o public land survey system reference - the rectangular grid reference of the public land survey system in the United States where applicable

- o federal status - names listed as official by USBGN, official by an administrative organization, or unofficial because they are not within the purview of USBGN
- o size - the size of the feature (length, breadth, area)
- o name origin and etymology - the history of the feature and its name origin (if time permits during data compilation)
- o administrative responsibility - when applicable, the organization that has administrative responsibility for the feature and its name
- o population - population figures for incorporated places
- o zip code - the Zoning Improvement Plan (ZIP) postal code associated with populated places to facilitate data interchange and for integrating names data with other data bases
- o bibliographic code - a code that identifies the published source in which the name was found.

Undoubtedly, user application requirements will demand additional data categories in the future. The software that manages the data base easily accommodates natural evolution and expansion of data requirements, and any software enhancements or replacement will retain this vital characteristic.

TRAINING AND PRODUCTS

The initial standard product derived from Phase I of GNIS was an alphabetical listing of geographical names within a State. It was also

possible to generate specialized listings and to rearrange the order of the data. All listings during the first phase of compilation were batch generated (the request or retrieval had to be submitted into a queue for group processing by the computer).

In 1981, a major advancement in GNIS was achieved. Interactive access or real time capabilities were implemented. These new capabilities meant that the USGS geographic names staff could use a computer terminal located in their office to initiate retrieval questions and immediately display geographical name information. Over the next 3 years, interactive access capabilities were enhanced to include simultaneous user access to the same file, remote printing capability, and an electronic user-tracking system. Of special significance was the ability to access the data base from any major city in the United States or Canada through a local telephone number. From other cities, a toll free number was made available.

Since 1982 standard GNIS products have been available on microfiche and magnetic tapes. Also, paper listings have been upgraded from standard computer printouts to high-quality laser-printed, bound preliminary gazetteers. Recent improvements include the capability of providing information on floppy or soft diskettes for use with personal computers.

During the initial compilation period of the data base, the USGS established the National Cartographic Information Center for the purpose of distributing cartographic, geographic, and other similar information.

The responsibility for distribution of GNIS-related products naturally fell to that office, freeing GNIS personnel to concentrate on compilation and developmental research.

Effective distribution of GNIS-related information depends on the ability of information specialists to analyze toponymic requests, retrieve information from the data base, and format appropriate reports for distribution. Software previously had been developed to permit the generation of standard products, but training of personnel was required in toponymic analysis and problem solving to allow the production of specialized reports.

A 3-day formal training course was begun in 1983 to train information specialists at various USGS offices in analysis of toponymic requests, interactive data retrieval, data manipulation, and report generation. Information personnel now process the majority of public requests, with only occasional complex requests and problems referred to the GNIS staff for resolution. The training is supplemented by The Geographic Names Information System: Data Users Guide 6 (USGS, 1985). USGS information specialists are kept aware of changes and enhancements to GNIS through an electronic message and status system available through the GNIS menu system.

SOFTWARE AND HARDWARE

In establishing the automated names system, one of the first decisions to be made was whether to develop a data base management system designed specifically for a toponymic data base or to obtain a general system with most of the required features. The former option was very costly and

would require considerable time for research and development. The latter option was less costly and could be immediately implemented. A general text handling system was found to meet virtually all of the data retrieval and manipulation requirements, and the beginning of the automated names system was established in 1972. The data management system selected operates in a mainframe computer environment. There have been minimal changes in the hardware configuration since implementation, and these were mainly to take advantage of technical advances. The data base and retrieval software still reside in a mainframe environment.

The addition of data categories, enhanced user services, and more efficient means of data handling have required the development of ancillary computer programs. All of the existing additional software packages and programs developed over the past 5 years in support of user services are available to GNIS information specialists through one menu of options. The information specialists may solve problems interactively, format reports, generate GNIS data on magnetic tape, check GNIS file status and system messages, cancel errors, and view an electronic GNIS user's guide. When soft diskettes containing GNIS data are approved for sale and distribution, the ability to generate them will be added as an option of the menu.

The viability and success of any data base requires a program of maintenance. Such a program has been in place for the National Geographic Names Data Base for the past 3 years. In 1984, the USGS regional mapping centers were provided with software that allowed data entry and telecommunication of data to the GNIS staff in Reston,

Virginia. The data were then checked for syntax and adherence to policy and procedures before entry into the data base. The advent of personal computers has rendered that software obsolete, and the GNIS staff has, within the past few months, completed a data input program that operates in a personal computer environment. The program allows rapid data input as well as error checking, eliminating many data input errors and thereby enhancing data base integrity. This software is available to cooperators assisting in Phase II compilation, and it also provides an efficient means of incorporating the results of the activity of USBGM into the data base. Plans are being made to make this software package available to interested Federal and State agencies as part of an expanded maintenance program.

At present, research is being conducted in system development that may have far-reaching effects. A study is underway to decide whether to upgrade current data retrieval software in a mainframe environment or whether a dedicated super minicomputer would be more cost effective and efficient. Another area of research involves compact disk-read only memory (CD-ROM) technology. Disk readers could be placed at any user location with a disk of the entire data base that would periodically be updated. Implementation is straightforward and operating costs are nominal, but a good deal of research is still required.

System integrity and security have always been of primary concern. GNIS users are permitted only to view and manipulate the data. Only GNIS staff members are allowed to add, delete, or alter the data base

information. In 1986 a new computer security program was installed to further safeguard the integrity of the data base.

SUMMARY AND CONCLUSIONS

Data gathering techniques originally concentrated on information that locates named features and names found on the USGS topographic maps. Expanded application and user requirements have led to a second phase of compilation whereby other sources, including Federal, State, and local as well as other pertinent and historical materials, are researched to ensure the completeness of the data base as required by law and user application. To ensure timely completion of the project, a program has been established to utilize cooperators in compiling names information in various States of the United States.

Expanding use and requirements have led to the addition of certain categories of data and the enhancement of some existing categories. The data base and operating software are flexible enough to handle the addition of future categories of data as needed.

To meet the goals of product development and data distribution, a user-friendly menu environment was made available to USGS information specialists throughout the United States. This menu system allows various options for product generation and the solution of various toponymic problems. To augment the use of the interactive menu system, a formal training course has been provided to nearly all information distribution offices, and a users guide has been prepared to supplement the training.

The original data retrieval software has been augmented by ancillary software to meet expanding applications requirements and product development. Additional software is now available that allows various selected offices to input names information in support of the data base maintenance program.

Significant research currently involves questions regarding optimum hardware and software requirements. Additional research is being conducted with regard to new onsite read-only disks. Whatever course the new developments take, maintenance, data integrity, and security will always be of paramount importance in the decision-making process.

REFERENCES

- Orth, D.J. and Payne, R.L., 1984, The National Geographic Names Data Base: Phase II Instructions: Open File Report 84-036, 100 p.
- U.S. Congress, 1947, An act to provide a central authority for standardizing geographic names: U.S. Congress, 80th, 1st session, S. 1262 [Public Law 242].
- U.S. Geological Survey, 1982-, The National Gazetteer of the United States of America: U.S. Geological Survey Professional Paper 1200; New Jersey (1200-NJ), 1982; Delaware (1200-DE), 1984; Kansas (1200-KS), 1985; Arizona (1200-AZ), forthcoming.
- U.S. Geological Survey, 1985, Geographic Names Information System: Data Users Guide 6, 34 p.