A PROTOTYPE GEOGRAPHIC NAMES INPUT STATION

FOR

THE DEFENSE MAPPING AGENCY

Douglas R. Caldwell
US Army Engineer Topographic Laboratories
Fort Belvoir, VA
USA

Robert F. Augustine
Defense Mapping Agency
Hydrographic/Topographic Center
Washington, DC
USA

Diane E. Strife
IIT Research Institute
Annapolis, MD
USA

ABSTRACT

The collection, maintenance, and distribution of geographic place names information at the Defense Mapping Agency is accomplished through the Foreign Place Names File, an index card data base. An attempt to automate the file is currently underway, starting with the development of a prototype Geographic Names Input Station. In the short term, the station will be integrated with the gazetteer production flow and will be used to enter, process, and display Roman script foreign text information. Its more important long-term role, however, will be to serve as a testbed for the development of an all digital Foreign Place Names Information System.

THE DEFENSE MAPPING AGENCY FOREIGN PLACE NAMES FILE

The Defense Mapping Agency (DMA) collects, evaluates, and maintains names information on foreign places, as well as on undersea and extraterrestrial features. DMA supports the efforts of the Board on Geographic Names (BGN) by referring evidence for the existence and correct orthographic representation of the names to the Board for review and action, and maintains files of all BGN-approved foreign names. This names information exists as a continually expanding file which serves as a base to support the use and application of geographic names throughout the United States government, in numerous professional organizations, and for the public at large.

The file of foreign place names was established under the Department of Interior as a result of the creation of the BGN in 1890. Transferred to the Department of Defense in 1949, this collection is now called The Foreign Place Names File (FPNF). The file today includes BGN-approved names for just over 2.5 million features with primary and variant names stored on 4.5 million index cards. The file is stored in a series of power file cabinets and is organized alphabetically by country.

The principal information contained in the FPNF is the BGN-approved feature name or names, the feature designation, and locational data. Supplementary information includes a record of primary and variant names with the sources from which they were derived. Linguistic, geographic, and other pertinent notes are included as well.

Distribution of information from the FPNF has been labor-intensive and inefficient, considering currently available digital communications methods. For instance, inquiries about individual names from agencies outside DMA are handled manually by searching the index cards and responding by telephone or mail. Wider names circulation is accomplished through the publication of gazetteers in paper form. In fact, gazetteer production requirements drive (and limit) the names collection process. DMA has produced approximately 45 gazetteers since assuming the function in 1969, and until recently, the rate of production has been slow. Since 1975, only 14 gazetteers have been published. A more effective method of production was introduced in late 1980 which has increased the potential for names distribution throughout the user population by allowing DMA to produce as many as 10 or 15 gazetteers per year. However, since the current system still only allows for the addition of approximately 125,000 names per year to the data base, most gazetteers represent limited revisions. DMA is investigating state-of-the-art information processing techniques to improve the system and hopes to increase the FPNF from 2.5 million to 50 million names within a decade. Large strides in names data acquisition and processing methods must be made in order to approach this mid-range goal.

THE GEOGRAPHIC NAMES INPUT STATION

Introduction

DMA is currently sponsoring the development of a prototype Geographic Names Input Station (GNIS). The technical work is being carried out by the Department of Defense Electromagnetic Compatibility Analysis Center, Annapolis, Maryland, and supported by the IIT Research Institute (IITRI) under Contract No. F19628-80-C-0042. The US Army Engineer Topographic Laboratories (USAETL) supervises the project. The GNIS addresses basic research problems associated with the processing of foreign text, as well as immediate gazetteer production requirements. An all digital Foreign Place Names System will be developed from the FPNF, as well as from the knowledge developed during this initial work.

GNIS and the Foreign Text Problem

During the development of the GNIS, many problems common to any digital foreign text system were encountered, because the required set of alphabetic characters is not standard and exceeds both the basic English alphabet and standard American Standard Code for Information Interchange (ASCII) alphabet. Processing this information requires special techniques for data entry, storage, and display.

DMA has a large and diverse foreign language requirement. One hundred and sixteen native languages or transliterated forms, ranging from Afar to Yoruba, must be depicted in Roman script form. Over 60 percent of these languages require an extended character set with marks not found in the 26 characters (A-Z) of the English alphabet. These marks used to extend the English alphabet fall into three categories: diacritics, special characters, and special symbols. (Figure 1)

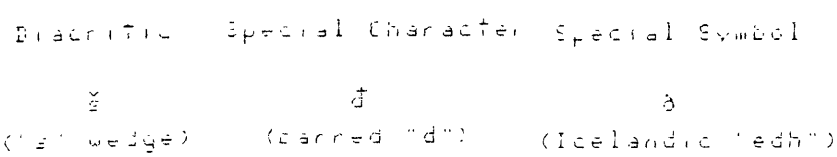

Figure 1. Sample Character Display From Matrix Printer

A diacritic is a mark that may be placed above or below an alphabetic character, but does not alter the basic character (e.g., "š"). A special character is an English alphabetic character with a superimposed mark (e.g., "đ"), and a special symbol is a non-English alphabetic character that may be used alone (e.g.,"ð"). DMA requires over seventy of these three types of marks in order to represent, as accurately as possible, the names of places and features in countries throughout the world.

An examination of the problems of input, storage, and display of the extended English alphabet text was necessary prior to selecting equipment and developing software. IITRI explored a number of options for data entry and storage and identified the keyboard as a critical element. There are two basic methods of data entry; via a standard,

preprogrammed keyboard, or via an expanded, programmable keyboard. The former approach is inexpensive, and it does not require the purchase of special hardware or any keyboard programming. With the standard preprogrammed keyboard, non-Roman marks are entered using standard codes in a predefined sequence. For instance, in the Central Intelligence Agency system, an "á" is entered as follows: 1) press "A" 2) press "%" and 3) press "a". The "A" denotes the acute accent, the "%" denotes a delimiter, and the "a" represents the letter associated with the diacritic. This approach is cumbersome if a large number of diacritics and specials symbols are to be entered because an operator either must use a look-up table containing the codes or prepare a keyboard overlay. Neither solution is desirable from the human factors standpoint.

IITRI selected a terminal with an expanded, programmable keyboard for the DMA names station. The keyboard has three sets of keypads, a main keypad and two outboard keypads. The main keypad contains the basic English alphabet, while the outboard keypads contain the extended alphabetic marks. To enter an "á" the operator follows the sequence: 1) press "a" on the main keypad and 2) press "´" on the outboard keypad. This approach streamlines data entry, since diacritics may be entered with a single keystroke. The need for a look-up table is eliminated; however, the keyboard is not entirely standard and has been programmed specifically to meet DMA's needs. Data storage is in ACSII format, so an "á" would be represented as "a$a"; where the first "a" represents the letter associated with the diacritic, the "$" denotes a delimiter, and the second "a" represents the acute accent.

Attempts to standardize data storage formats for interchange among users are in the beginning stages. The International Standardization Organization (ISO) has adopted ISO 5426 "Extension of the Latin Alphabet Coded Character Set for Bibliographic Information (1980)" and proposed ISO/GIS 6937/2 "Coded Character Sets for Text Communication Part 2: Latin Alphabet and Non-Alphabet Graphic Characters." Two deficiencies of the standards, however, are their lack of agreement and failure to include all the extended English alphabet marks required by DMA. Since no inclusive standard exists, translate tables must be developed to convert one format to another.

As with data entry and storage, a number of alternative approaches exist for hard copy data display. At DMA, the final production is accomplished with a Multiset III Phototypesetting system, so the GNIS printer addresses only interim hard copy. The critical option considered for hard copy was the choice of matrix versus non-matrix printers. Non-matrix printers use a system of preformed characters cast in metal or plastic, such as a daisy wheel, type element, or thimble. These produce high quality characters, but limit the selection to those available, since the production of new preformed characters is expensive. The use of non-matrix printing is not acceptable for the DMA environment for two reasons. First, consolidated character sets meeting DMA's requirements do not exist, and second, DMA's requirements may change as new transliteration schemes are approved. Matrix printers are an attractive alternative, offering reduced print quality but much greater character flexibility. With a matrix printer, each character is formed from a matrix of cells which may be "turned-on" in any combination. Although a 7 x 9 matrix is commonly used to generate

characters, DMA required an 8 x 16 matrix in order to accommodate the diacritics that must be placed above and below the English characters. A Florida Data printer was selected to satisfy this requirement. New characters may be added via a font editor as additional requirements develop. After addressing the foreign text problem, DMA was able to move towards integrating the GNIS with names production.

GNIS Hardware

The chief hardware components of the GNIS are a Plessey System 23VX with a PDP 11/23 microcomputer, an ECD Intelligent Terminal and a Houston Instruments digitizing tablet. The major computational and data storage element of the system is the Plessey PDP 11/23 micro system, using the RT-11 operating system. Mass storage consists of a 5 megabyte (MB) fixed disk and a 5 MB removable disk cartridge. A nine-track magnetic tape unit allows the user to load and transport data files. The tape drive, disk drive, and microprocessor are contained in the same compact unit which operates in a wide range of physical environments. Several hardware options are available, and the software configuration is easily upgraded.

Foreign-text entry and display are accomplished through the ECD intelligent terminal, which includes a video monitor, a microprocessor, a dual floppy disk drive, and a dot matrix printer. (Figure 2)

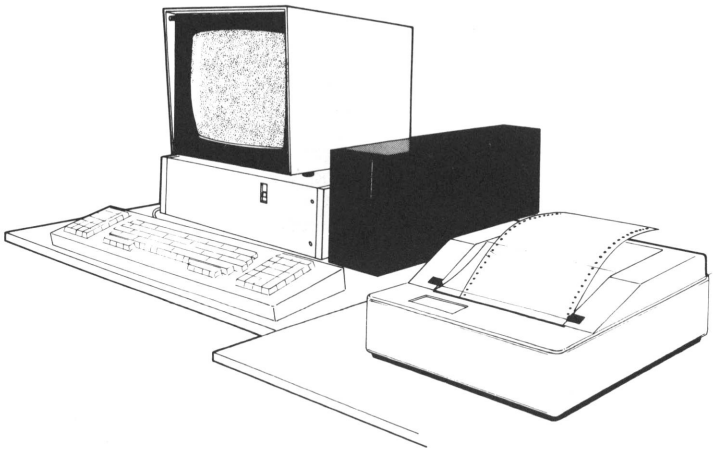

Figure 2. GNIS Work Station

The 6502 microprocessor-based intelligent terminal has a 64 kilobyte memory and 0.5 MB of floppy disk storage. The keyboard is a special configuration with two outboard keypads and programmable keys. The Florida Data dot matrix printer enables local listings of stored data. The ECD functions as the console to the PDP 11/23, as well as serving as an intelligent terminal, by means of a series of ONLINE/OFFLINE commands downloaded at appropriate intervals from the PDP 11/23. In the OFFLINE mode, the ECD operates under Translex, a display-oriented macro language that supports the many word processing features and is unique to the terminal.

The Houston Instruments digitizing tablet connects in-line by cable between the ECD and the PDP 11/23. It provides raw digitized data to the PDP 11/23 for processing.

The selection of hardware and subsequent software development was heavily influenced by the problems inherent in the digital manipulation of foreign text. Now that these problems have been solved, the system can be phased into the names production workflow.

GNIS and Gazetteer Production

As previously outlined, the current names processing system is driven by gazetteer production requirements. It is not surprising that the earliest gazetteer production was a manual process using handwritten copy and standard early printing techniques. Diacritics and special type characters were available only through special type order and were costly. The process by any standard was indeed slow. By approximately 1960, the revised process for creating gazetteers was to key names data directly from handwritten cards on to computer punch cards and store the information on magnetic tape. This revised process enabled some sorting and correction, but only uppercase alphanumeric characters could be produced on the printout, and special characters and diacritics had to be added by hand. The resulting hand-annotated printout was photographed and reduced. The photographs were used to create plates, and the gazetteer was then printed.

The current gazetteer production process makes use of the archival tapes created by the 1960's process. An archival tape without diacritics, special characters, or special marks is withdrawn from storage and translated into a format compatible with the Multiset III automated typesetter currently used by DMA. The information is processed into upper- and lowercase characters and printed on a high speed printer. The printout is then analyzed by toponymists familiar with the appropriate geographic area. If the tape is currently readable and technically complete, the printout is returned to the typesetter with the associated gazetteer along with toponymic and linguistic comments. The diacritics are then keyed by the typesetter into the digital file, and a second printout is generated. This printout depicts the diacritic as a special symbol beside the character that it affects (e.g., fore*^*t equals forêt).

The second printout is reviewed and then compared to the card file. Corrections are made by the toponymist manually on the printout. Previously unpublished feature names are manually keyed by the typesetter from file cards, and a third printout is created by the typesetter and reviewed by the toponymists. Final corrections are made by the typesetter, and the introduction and names list is printed in galley form. The galley proof is reviewed by the toponymist, minor corrections are accomplished, and the final pages are photographed. The resulting set of gazetteer negatives is forwarded for printing to the Government Printing Office. The digital file resulting from the automated typesetting may now be updated periodically to ensure that the file is current. It is important to note that this gazetteer file represents a Names Type File (NTF) only and does not include the historical or linguistic data that is associated with a primary name in the FPNF data base. The existing production system saves as much as 75 percent of the time required for manual keying, but requires extensive coordination and communication between the toponymists and the typesetters.

The prototype GNIS gazetteer production equipment provides direct contact between the toponymist and the digital data. All printout reviews of the current production system will be eliminated by the prototype system. The toponymist will review all files prior to printing. A digital file will be locally maintained for uninterrupted updating, and will serve as a primitive base for names information. It is estimated that the prototype system will reduce the project processing time by 50 percent and improve production from 125,000 to 250,000 names per year.

The GNIS has three major software programs; the NTF Tape Loader, the NTF Editor, and the NTF Tape Create. DMA currently has approximately 165 seven-track archival magnetic tapes containing only uppercase characters with no diacritics. In the GNIS production flow, DMA will geographically sort this tape data into 15 minute squares and transfer the information to a nine-track tape. The NTF Tape Loader program will read the nine-track tape into the Plessey system, convert uppercase characters to upper- and lowercase characters, and develop the names record format. Records may be accessed later through the NTF Editor for editorial corrections and geographic coordinate digitization.

The NTF Editor allows the toponymist to manipulate the names records built by the NTF Loader. Five main modules are accessed by the NTF Editor: Select, Digitize, Edit, Delete, and Add. The Select module accesses records sequentially or directly by record number. The Digitize module registers source maps, captures raw table coordinates, and converts the raw data to geographic coordinates. The geographics are then used to generate Universal Transverse Mercator (UTM) coordinates and Joint Operations Graphic (JOG) map series numbers. Upon completion of the digitizing, a proof plot may be generated. The Edit

module takes the ECD terminal OFFLINE. Using the word processing features of the terminal, a record can be edited. The unique characteristics of Translex allow the insertion and display of diacritical marks. When editing is complete, the user initiates a command which performs an integrity check on the record, places the ECD ONLINE, and sends the record back to the PDP 11/23. The Delete module allows the user to delete any selected record. The Add module accommodates the addition of new records for a geographic area. The user may add as many records as were specified in the NTF Tape Loader program.

NTF Create is the final program, and it produces a gazetteer tape or an intermediate tape for alphabetic sorting by the Multiset III. The sorted tape may be recycled through the NTF Editor for further modification, while the gazetteer tape may be used for the addition of typesetting information prior to printing.

GNIS and the Future

The prototype GNIS has deficiencies as a production system, but will be improved and serve as the foundation for a Foreign Place Names Information System. The two most obvious drawbacks of the prototype system are a lack of mass storage and an inability to perform local sorting. With the addition of a 300 MB disk, the system should accommodate all files except those for the largest countries. A larger capacity will also permit comprehensive sorting. Another desirable option would be the addition of a direct communications link with the Multiset III. This would eliminate the need to physically transfer data to the typesetter on tape. It is anticipated that these improvements will be made in a system upgrade.

DMA's long-term goal is the development of an all digital Foreign Place Names Information System. From the GNIS Names Type File, a data base will be developed to incorporate the information currently contained in the Foreign Place Names File, and additional information required by names applications specialists and map compilers for map production. In addition to integrating and consolidating names production within DMA, the system will be directly accessible to qualified users in the public and private sectors.