

AD-A163 042

THE GEONAMES PROCESSING SYSTEM SYNOPSIS(U) NAVAL OCEAN
RESEARCH AND DEVELOPMENT ACTIVITY NSTL STATION MS
G LANGRAN SEP 8 NORDA-125

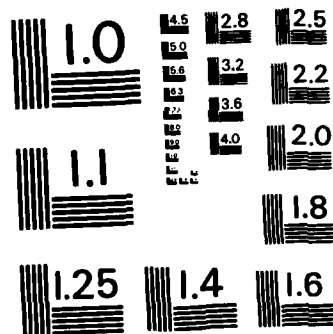
1/1

UNCLASSIFIED

F/G 5/2

NL

								END						
								FILED						
								DTA						

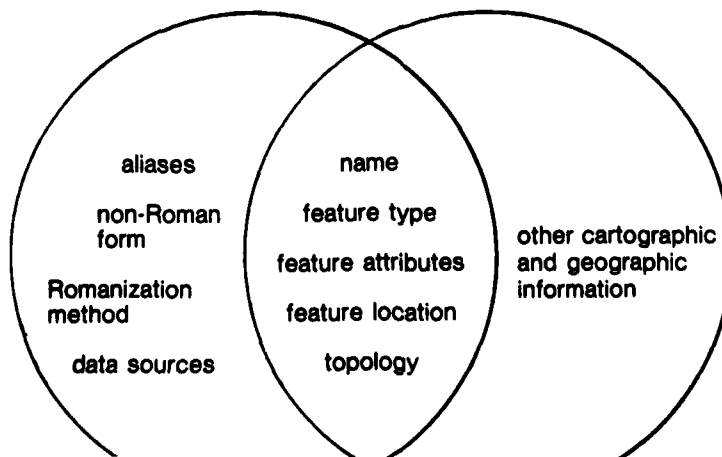


MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS - 1963 - A

AD-A163 042

TOPONYMIC NEEDS

CARTOGRAPHIC NEEDS



EXECUTIVE SUMMARY AND TABLE OF CONTENTS

INTRODUCTION

v

DMA has recognized a need for digital procedures to store, retrieve, and edit geographic names data and to prepare names for product generation. DMA's stated goal is a 50-100 million name digital data base with subsystems to capture names, edit and format names data, and prepare names overlays for maps. NORDA began a geonames processing system design study late in FY82. This report summarizes NORDA's study findings.

PART ONE: SUBSYSTEM DESCRIPTIONS AND RECOMMENDATIONS

1-1

The four sections in Part One discuss four geonames processing functions in turn: names data capture, data base management, editing and formatting, and map names processing. Each subsystem is described, unresolved issues are highlighted, and recommendations are made. → See p. 10

1.0 AUTOMATED ALPHANUMERIC DATA ENTRY SYSTEM (AADES)

1-1

Overview

1-1

Operational Alternatives

1-1

Technical Alternatives

1-3

Recommendations

1-5

2.0 GEOGRAPHIC NAMES DATA BASE (GNDB)

2-1

Overview

2-1

Interface to Feature Data Bases

2-1

Size Management

2-2

Distributed vs. Centralized Hardware

2-2

Data Base Structure

2-3

Quality Control

2-3

3.0 ADVANCED SYMBOL PROCESSING (ASP)

3-1

Overview

3-1

ASP Requirements

3-1

ASP Configuration

3-1

Recommended Software Upgrades

3-2

4.0 ADVANCED TYPE PLACEMENT (ATP)

4-1

Overview

4-1

Applicable Technology

4-1

Recommended Approach

4-1

PART TWO: OVERALL SYSTEM ISSUES

Part Two discusses concerns common to all four geonames processing subsystems. Section 5 defines interfaces between the subsystems. Section 6 weighs the technical possibilities for handling non-Roman alphabets and ideographs. Personnel and space requirements for a comprehensive names processing system are estimated in Section 7.

5.0	SUBSYSTEM INTERFACES	5-1
	Functional Interfaces	5-1
	Data Link	5-2
6.0	NON-ROMAN SCRIPT PROCESSING	6-1
	Background	6-1
	Electronic Processing of Ideographs	6-1
	Recommendation	6-3
7.0	PERSONNEL AND SPACE	7-1
	Overview	7-1
	Systems Support	7-1
	Applications	7-2

PART THREE:	SUMMARY OF RECOMMENDATIONS	8-1
-------------	----------------------------	-----

Part Three summarizes the recommendations made in this report.

Keywords:

- *Names Data Base*
pursue high-volume data entry in an R&D environment, focusing on OCR to improve capture rates;
- establish a working group at DMAHTC to establish requirements and priorities, and work toward an operational concept;
- handle non-Roman characters as bit maps;
- design the names data base as a subset of the feature data base.

8.0	SUMMARY AND CONCLUSIONS	8-1
APPENDIX A:	BIBLIOGRAPHY	A-1
APPENDIX B:	ORIGINAL REQUIREMENT STATEMENTS	B-1

FIGURES

<u>Figure</u>		<u>Page</u>
i-1	Names data sources	vi
i-2	System overview	vii
2-1	Relationship of toponymic and cartographic data requirements	2-2
4-1	Recommended ATP process flow	4-2
5-1	Interfaces between the systems	5-1
7-1	Personnel requirements	7-1

TABLES

<u>Table</u>		<u>Page</u>
1-1	Data entry timing	1-2
1-2	Average map timing	1-2
1-3	USGS data capture rate	1-4
2-1	Defining query requirements and priorities	2-2
5-1	Standard data transfer record	5-2
7-1	Systems support space estimate	7-2
7-2	Applications group space estimate	7-2



Accession For	
NTIS CRA&I	☒
DTIC TAB	☐
Unannounced	☐
Justification	
By	
Distribution /	
Availability Codes	
Dist	Avail and/or Special
A-1	

ACKNOWLEDGMENTS

This work was sponsored by DMA under Program Element 64701B, with subtask title, "Geonames Processing System;" Mr. Dennis Franklin of DMAHQ/STT was the project manager.

INTRODUCTION

The Defense Mapping Agency (DMA) has recognized a need for digital procedures to store, retrieve, and edit geographic names data and to prepare names for product generation. A small subset of world geonames and named feature attributes have been digitized by DMA and other U.S. and foreign government agencies. DMA's phototypesetters accept ASCII-coded digital data, and a prototype foreign text processor was designed and installed at DMAHTC. This text processor, the Names Input Station (NIS), enables keyboard entry and edit, and hard- and softcopy display of text with diacritics and special symbols. While these devices form a basis for further development, alone they improve DMA's digital capabilities only marginally, since interfaces are sketchy, data coverage is poor, and the NIS is operable only sporadically. Toponymists still work manually with analog media, and map type is produced by stick-up, fastened by hand to film.

Evidently, the primary requirement is for names data in digital form. DMA's stated goal is a 50-100 million name digital data base. Names data sources are shown in Figure i-1. Analog geoname sources include DMA's 4.5 million foreign placename cardfile (the FPNF) and a vast archive of maps and charts. The FPNF is being digitized by typed entry. Unfortunately, neither the FPNF nor the other available digitized names data provide the locational accuracy required for map names processing. Maps and charts alone provide adequate high-resolution feature positions. Today's technology allows location to be digitized from graphic sources using cursor and digitizing table, with text entered by voice or keyboard. These methods are slow and painstaking; a 50-million name data base would require many man years to build this way. Thus, a high-volume data entry method from graphic sources must be developed. Section 1 discusses the issues surrounding this system, which is called the Automated Alphanumeric Data Entry System (AADES).

Once digitized, names and their associated data will be stored in a data base. Associated data include named feature types and attributes, and toponymic information. Toponymic information includes variant spellings and aliases, data source(s), and for many names, the Romanization method and non-Romanized form. It must be possible to query the data base by area (all names in France, all names falling on a stated map sheet), by feature type (all rivers in Africa, all populated places in Japan), and by the name itself (provide all known information on a given name). Important data base design issues are its size, which is large enough to introduce technical risk; its interface to DMA's cartographic data bases; and its architecture, which will be a tradeoff between speed and flexibility. The Geographic Names Data Base (GNDB), which would be used to store and retrieve such information, is discussed in Section 2.

A text processor that handles diacritics and special symbols is urgently needed at DMA. Lack of reliable equipment has impeded toponymic work at DMAHTC and slowed the Foreign Place Name File's digitization. Excellent alternatives are now available for assembling off-the-shelf technology, but proposed interim deliveries were rejected in favor of integrated system delivery at a later date. Because this requirement clearly does not share the research status of the other geonames processing subsystems, it appears that it will be handled by procurement within the production center. The Advanced Symbol Processor (ASP), which will fulfill stated requirements, is described in Section 3.

The final geonames processing requirement is for digital map type composition and placement. Interactive systems for map type placement are available off-the-shelf. Names can be typed from

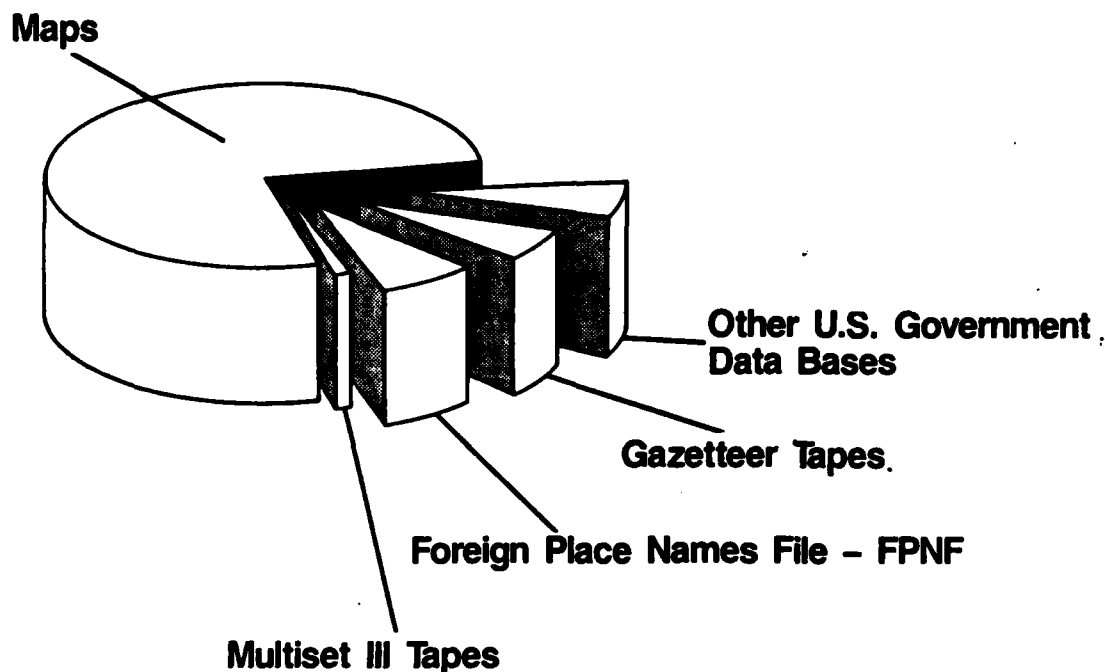


Figure i-1. Names data sources.

the keyboard of an interactive graphics workstation and placed on the map using a cursor and software commands. The USGS uses such a system for its provisional mapping program. A number of upgrades are possible that reduce the cartographer's role in type positioning, but all are contingent on digital data availability. Advanced Type Placement (ATP), used to automate map name production, is discussed in Section 4.

Figure i-2 shows the conceptual flow of names data through the four subsystems. Names data from maps and charts enter the system through the AADES, then are stored in the GNDB. ATP and ASP prepare and format names for products and other names applications. There is feedback to the data base from applications processing.

NORDA began a geonames processing system design study late in FY82. A number of reports have been generated to describe system design plans (Brown et al., 1983; Langran et al., 1984; Campbell et al., 1984). This report summarizes NORDA's study findings. Part One contains four sections that describe each subsystem in turn, discussing unresolved issues and recommending solutions. Part Two discusses concerns common to all four subsystems: interfaces between the subsystems, non-Roman script processing, and personnel and space requirements. Part Three summarizes the recommendations made in the report and highlights unresolved issues. DMA's original requirement statements are in Appendix B.

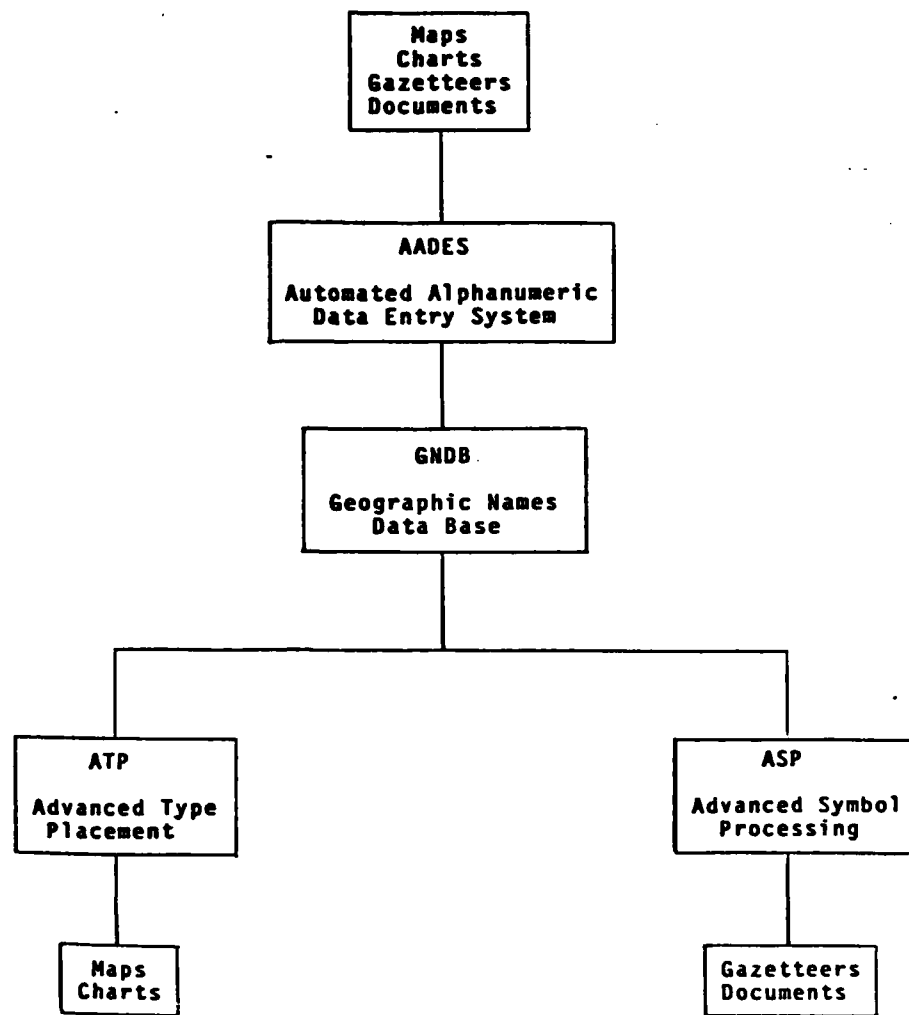


Figure i-2. System overview.

PART ONE: SUBSYSTEM DESCRIPTIONS AND RECOMMENDATIONS

The four sections in Part One discuss the four geonames processing subsystem's functions in turn: names data capture, data base management, editing and formatting, and map names processing. Each subsystem is described, unresolved issues are highlighted, and recommendations are made.

1.0 AUTOMATED ALPHANUMERIC DATA ENTRY SYSTEM (AADES)

Overview

AADES must provide DMA with high-volume names data entry from maps and charts. Required names data include the name itself, its country (or countries), the source document and source date, and the named feature type with associated position and attributes. Toponymic information (Romanized form or Romanization method, variant spellings, and aliases) is added if available. Otherwise, it is added at a later processing stage. Map series and single maps of many scales and projections from DMA and other sources will be input.

Speed is essential to completely load the names data base by the end of this century. A 50-million-name data base will require nearly 1500 man years to load if each name requires 3 minutes to capture (Table 1-1). Clearly, small speed increases produce major cost savings. Data capture decisions will be a tradeoff between development and personnel costs. By investing time and money on faster data capture procedures, the man years required for capture are reduced. Unfortunately, technical risks are relatively high for AADES; although a number of technologies apply, none can be considered mature, nor could they be implemented without initial testing in a research environment.

Automated optical character recognition (OCR) is instrumental to improved speed. Limited OCR with a high level of analyst interaction has been implemented on graphic source materials. Known implementations of graphic-source OCR use template-matching techniques against stored font libraries. The large number of type styles contained on maps and the varied character spacing and orientation within words will challenge OCR software. Available OCR implementations do not address text with diacritics. Because many diacritics are added by hand, the success of template-matching OCR would be limited without some trainability or other upgrades.

Two major decisions must be made regarding AADES. The first is of organizational and operational concern. The second is the technical approach, which affects both the level of development risk and the possible gain in speed. The remainder of this section discusses options and recommends a course of action.

Operational Alternatives

Build as you go

One approach to AADES is to digitize names data from maps and charts in the course of product generation. Over time, data base coverage would increase. Assuming an average of 500 names/map and a capture rate of 3 minutes/name, this approach would add approximately 26 man hours to the time required to produce one map (Table 1-2). The major advantage of this approach is that it minimizes system-specific personnel and hardware demands. And the slow pace of data capture will promote orderly data base loading. A major disadvantage to this plan, however, is that the data base will neither contribute to production efficiency nor will it be a reliable source of geonames for at least 10 years. Rather, it will add to production time and exclude from the data base names whose areas are not on production schedules and names not included on DMA products. The data base, a potentially powerful reference and production tool, would be reduced to an archive. Disenchantment with the data base concept could follow.

Table 1-1. Data Entry Timing.

Names to goal 50,000,000		
Days/Year 365	Wkends & Holidays 114	Down Time 5%
Working Days 238	Hrs/Day 7	Cost/Manyear \$75,000
Mins/Name 3.0000	Man/hrs to Goal 1497.77	Human Cost \$112,332.624
1.0000	499.26	\$37,444.208
0.5000	249.63	\$18,722.104
0.2500	124.81	\$9,361.052
0.0167	8.32	\$624.070

* Minutes per name is the average number of minutes spent interactively identifying a name and its attributes. Batch portions of the task are not included.

Table 1-2. Average Map Timing.

Names on map 500		
Days/Year 365	Wkends & Holidays 114	Down Time 5%
Working Days 238	Hrs/Day 7	Cost/Manyear \$75,000
Mins/Name 4.0	Man/hrs for map 35.09	Human Cost \$1,577
3.5	30.70	\$1,380
3.0	26.32	\$1,182
2.0	17.54	\$788

Dedicated names data capture

Each government agency that has built a names data base has opted for a one-time data capture effort. Names could be captured in bulk by a small team of in-house personnel or by contractors. In either case, the effort should be physically located at DMAHTC to take advantage of hardware and reference resources. Only DMA's toponymic staff should be authorized to approve entry of data into the data base.

Dedicated names data capture would require a schedule independent of the production schedule, which it would outstrip quickly. One alternative is to capture names according to established area priorities, e.g.:

- names in areas with the highest priority are captured first at 1:50,000 scale, 100% accuracy;
- names in areas with second-order priority are captured next at 1:250,000 scale, with slightly lower accuracy levels tolerated;
- names in areas with third-order priority are captured last from 1:500,000-scale source materials.

Technical Alternatives

AADES technical alternatives offer tradeoffs between development and personnel costs. Alternatives are presented in this section according to the source media used.

Hardcopy base

Data capture direct from hardcopy source is mature technology. A map is fastened to a digitizing table, a cursor is used to digitize a coordinate location, and the name and feature attributes are keystroked. The USGS and the CIA used this method to capture their names data bases. Possible upgrades are voice entry of commands, names, or feature attributes, or selective scanning and consequent OCR of placenames. DMA has requested that voice entry not be used. And implementing OCR on scanning wands is risky, since added to standard OCR difficulties are irregularities in wand movement, pressure, and tilt.

The USGS, which has captured a U.S. names data base, documented Arkansas names digitization. Timing estimates are shown in Table 1-3. The preparation step was performed by USGS personnel, and included collecting and alphabetizing the maps to be digitized. All other work listed was performed by a contractor. Captured data were name, feature class, FIPS County Code, positional coordinates, Map Code, and elevation. Maps were re-sorted by the contractor into 10 X 10 map cells, since it was found to make digitization quicker and more accurate. Eight "annotators" worked on the source maps, underlining names and referencing them to their features. Consequently, the work of the "keyers" went more quickly, requiring only rote data entry.

Two expenses are not reflected in Table 1-3's man hour statistics. First is the cost of the source maps, which were annotated, digitized, then discarded. Second, verifying data accuracy was a major expense (the contractor was paid a fee per correct name). USGS personnel sampled 10% of all captured names (countrywide) over a period of 2.5 years, with 12-15 people involved on a steady basis. Excluding these costs, the effort averaged 3.54 minutes/name. Using Table 1-2's cost scale, digitizing the names on an average map at the rate of 3.54 minutes/name would require approximately 30 man/hours and cost in the range of \$1300.

A hardcopy-based system could be procured and delivered to DMA quickly and with minimal risk. However, speed enhancements would be limited in the near future, contributing to high labor costs throughout the names data capture effort.

Softcopy base

A softcopy-based approach scan-digitizes the entire map (unseparated) or the names overlay. After scanning, names are isolated and captured interactively and in batch.

Table 1-3. USGS data capture rate. The 27,684 Arkansas names data records (whose capture time is shown below) required approximately 3.54 minutes each to capture.

	<u>Hours</u>
Preparation	40
Data entry (1)	
Sort (2)	20
Annotate	300
Verify	108
Key	424
Edit	110
Correct keying mistakes	24
Processing (background)	20
Administration time	88
Additional corrections	8
Total data entry	1102
Quality control (3)	155
Correction	250
Editing	85
Administrative	<u>3</u>
	1635

(1) Data captured are name, feature class, FIPS County Code, positional coordinates, Map Code, and elevation.

(2) Maps, which arrived in alphabetical order by title, were sorted into 10 X 10 cells for more efficient data entry.

(3) Plots were made of captured names and compared to map sheets.

The first implementation of such a system would, by necessity, be largely interactive. Even so, rapid improvements in related technology are anticipated (automated blueprint reading has been targeted by several vendors as a profitable endeavor). OCR technology, too, has advanced considerably, a trend that is likely to continue.

Software upgrades to automate portions of the interactive process could be designed around a number of type characteristics and technologies. Type color, size, and shape help to isolate type in the raster image. Progressive or heuristic searches connect letters into words. Automated character recognition will convert a large percentage of the character images into ASCII codes, and software dictionaries would eliminate generic words. Man-machine interaction occurs between processing steps, the machine requesting approval or correction of automated deductions, and permission to proceed to the next processing step.

This option rates better on a human factors scale. Neck and back discomfort could prevent analysts from using a digitizing table for long periods, but an interactive graphics workstation's adjustable screen heights are engineered with ergonomics in mind.

The initial and ongoing equipment and development expense of this technical approach is offset by smaller personnel demands and a decrease in names digitization response time.

Recommendations

Combining production-oriented names digitization with a dedicated capture effort is the recommended strategy. It would be desirable to focus the initial data capture effort on the production schedule until production needs are answered and all known production requirements are outdistanced. Then, other prioritized names can be captured.

Because of specialized equipment and the long-term nature of the effort, it is recommended that DMA personnel be used. If data capture is contracted, it should be by area at a fee per correct name. A reasonable fee can be established once the speed of capture using AADES is known.

A softcopy-based technical approach is recommended because of upgradability and human factors. The proposed strategy is to implement low-risk software for batch processing when possible, supplemented by interactive utilities for human intervention where machine efforts fail. For example, software developed in DMA's Auto Carto Feature Identification project can automatically identify standard DMA map and chart symbols from their scanned images. Thus, on DMA source materials AADES could recognize feature type and digitize location for an estimated 50% of named features. Unrecognized features would be digitized by the analyst.

Hardware requirements for the recommended system are a large-format color scanner (which could be shared with other cartographic systems) and graphics workstations. Major software requirements include interactive graphics utilities, creative data structuring, font libraries and look-up tables, and a range of character recognition software.

OCR software must recognize an indefinite number of fonts, with mixed fonts (as many as 20) within a single map. Map type has variable spacing and orientation within words, diacritics, numbers and letters, upper and lower case. Template-matching software will be usable in some instances. A trainable system is desirable. Handwritten OCR techniques (Fay, 1985), too, should be implemented to handle characters unrecognizable through template matches and characters with handwritten diacritics.

The recommended system should be implemented in a research environment until low-risk automated procedures are installed and tested. Then, a prototype system should be evaluated at DMA for no less than 1 year. During this time, difficulties can be identified and addressed. Because of the potential for major timing improvements through software upgrades, a continuing research and development program following installation of production systems is recommended.

2.0 GEOGRAPHIC NAMES DATA BASE (GNDB)

Overview

The GNDB will handle the data banking, storage, and retrieval of geonames and their associated attributes in an all-digital environment. This single, controlled repository of geographic names will increase production throughput, facilitate toponymic queries, and standardize the names information disseminated by DMA.

The GNDB will receive batch inputs from AADES (Section 1) and will be maintained by analysts working at interactive ASP terminals (Section 3). Files compiled from the GNDB will be used to create gazetteers, names overlays, and other names-related products. The GNDB will reside at DMAHTC, and all maintenance and updates will be made by DMAHTC personnel. DMAAC and other remote users will access the GNDB via batch transmission (magnetic tape or data line).

More than any other subsystem of this design effort, the GNDB has suffered from an unclear definition of its relationship to other systems in DMA's digital future. Little serious planning can take place until several major issues are decided. These issues, in turn, can only be decided once the GNDB's interface to other DMA data bases is designed. Thus, the following discussion of design issues presents a series of tradeoffs and uncertainties, rather than clear direction.

Interface to Feature Data Bases

In the future, map and chart features will be compiled from data bases maintained at DMA. The role of names in these data bases is defined only sketchily. While names are map features, the toponymic data required to maintain names in current and correct form are not. Efficiency demands that all toponymic data be stored in digital form.

Figure 2-1 shows the relationship of toponymic and cartographic data requirements. Toponymists must know all of a name's variant spellings and aliases, as well as the sources of the toponymic information, to make qualified judgments on correct areal nomenclature. The non-Romanized version or the Romanization method are required for placenames using non-Roman alphabets and ideographs. Added to these data are requirements common to both cartography and toponymy: the placename and its feature type, location, and attributes.

The most logical way to meet toponymic needs is to design the toponymic data base as a subset of the cartographic data base. While this strategy enlarges an already formidable data handling problem, the alternative—maintaining two separate databases—results in design and quality problems.

The major design problem with maintaining separate data bases is how to cross-reference attribute information for entities stored in both data bases. A link is required to update names stored in the cartographic data base and to add detailed feature boundaries and attributes to named features in the toponymic data base. Two alternative linking methods exist. First, a software link could be designed to match name, feature type, and feature location from one data base to the other. A software link would be slow and error-prone, especially when feature positions or placename spellings do not match. Linear and areal features would be especially difficult to cross-reference, since the names data base will store only rough approximations of feature shapes.

A better alternative to a software data base link is to store a feature key in all data bases and data sets. However, given the number of features existing in a worldwide geographic data base, unique keys could be difficult to devise.

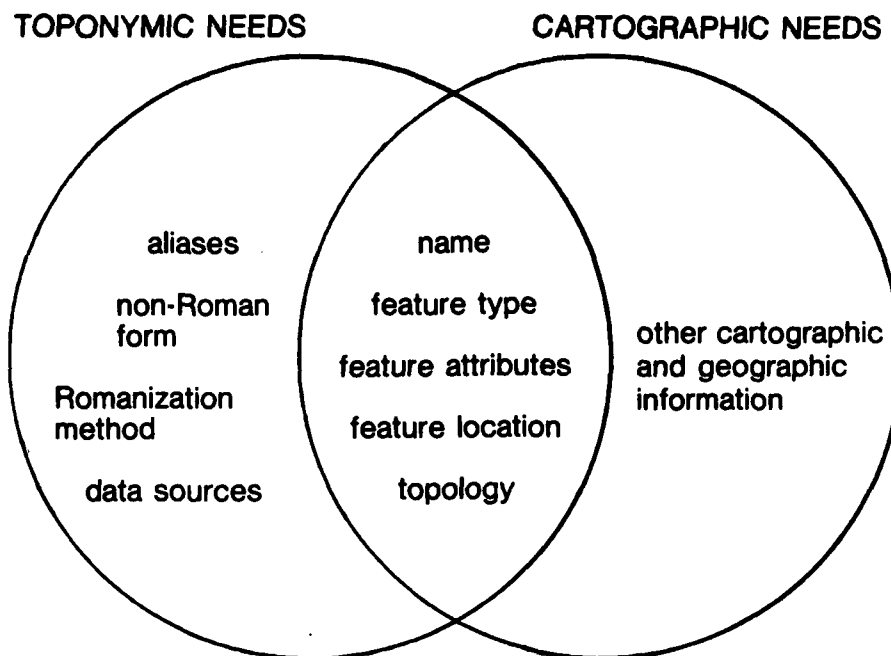


Figure 2-1. Relationship of toponymic and cartographic data requirements.

Quality control and data consistency are potentially serious problems with separate cartographic and toponymic data bases. Changes to names or to certain feature data would outdate the sister data base. Corresponding features would need constant updating so data quality is not impaired.

Despite the problems of *planning and implementing* the very large data base necessary to answer both cartographic and toponymic needs, it is the better alternative. If this tack is taken, the majority of the design issues discussed below are moot, since the needs of the cartographic data base would drive architecture and data model selection.

Size Management

To date, NORDA has planned the GNDB as a very large data base to be managed by a powerful data base management system (DBMS). Size estimates for the loaded data base range from 4 to 10 gigabytes. The GNDB's size restricts hardware and software selection. Only a limited number of commercially available DBMSs can handle system requirements—the few that can require large IBM mainframe computers.

If the GNDB is designed as a separate entity from DMA's cartographic data bases, it may be useful to consider a scaled-down approach. Lowered performance expectations (speed, range of possible queries) would reduce GNDB development to manageable proportions. Partitioning data by country, continent or by area code would result in a series of smaller data bases. More analyst participation in locating and querying data would be traded off against faster GNDB delivery, lower hardware expense, and lower technical risk.

Distributed vs. Centralized Hardware

Because of the GNDB's very large size, a centralized, single-computer approach places heavy performance requirements on both computer processing and disk drive I/O, and makes data base administration and maintenance functions technically risky. If dividing the data base into several discrete

units (as described above) is not a viable alternative, local processing capabilities would ease the main-frame processing load.

Intelligent terminals or microcomputer workstations allow the DBMS to download files for local processing. Text processing is best performed on microcomputers, since the man/machine interface is superior. Toponymic aids and quality control modules, too, could be stored on local hard disks, or downloaded from the central processor. Unfortunately, distributed processing technology is not mature, making this approach more difficult to pursue.

Data Base Structure

The relational and hierarchical data models present a direct tradeoff between performance (storage and retrieval efficiency) and flexibility. While hierarchical models have proven to perform better with large data volumes, theoretically there is no reason why a well-tuned relational model should not perform equally well. Performance requirements for this very large data base suggest that the hierarchical model should be employed. However, the newness of DMA's digital environment leads one to anticipate evolution, improvement, and restructuring after data bases are installed. Because of this, a relational model is recommended.

Data usage must be defined before data relations and access paths are configured. The toponymists and cartographers that will comprise the GNDB's most active and demanding user group should be asked for their query requirements and priorities. Sample questions are shown in Table 2-1.

Quality Control

Among the essential GNDB requirements is quality control. Bad data must not enter the data base, for once in, it will be difficult to exorcise. The ultimate quality control responsibility rests with GNDB personnel. However, to assist in detecting bad data prior to data base entry, files can be passed through a number of software modules that filter unusual values or combinations of values. Peculiarities will be flagged for inspection by a data base analyst. Quality control modules include those described next.

- Structural constraints to prevent duplicate records or duplicate paths.
- Reasonable bounds:
 - settlement population between 1 and 10 million;
 - elevation between -10 and 29,100 (for topographic or hypsographic features);
 - depth between 0 and -xxxx (for hydrographic or bathymetric features);
 - length of rivers between 1 and 4200 miles;
 - latitude between -90 and 90;
 - longitude between 0 and 360;
 - others, as applicable.
- Relational integrity:
 - an entity's position is within its country,
 - stated or implicit relations do, in fact, exist.
- Consistency assertions to avoid storing mixed attribute measures. If the extent of one feature is described in meters and another in centimeters, misleading query responses could result. Within a feature class, all like attribute classes must use the same unit measure.

Table 2-1. Defining query requirements and priorities.

Queries to the names data base can be phrased in a number of ways. Rate how useful each query type shown below would be to your application. Rate how often you might query the data base in each way. Use the scales shown below.

Usefulness: 1—not needed 2—somewhat useful 3—useful 4—essential

Frequency: 1—less than once/month
2—several times/month
3—several times/week
4—several times/day

	Usefulness	Frequency
1. "Find all names within"		
- map sheet X	_____	_____
- country X	_____	_____
- province/state X	_____	_____
- continent X	_____	_____
- a geographic area defined by minimum and maximum latitude and longitude	_____	_____
- country X, within n (mi/kms) of point (X,Y)	_____	_____
- other	_____	_____

2. "Find all names within a given area whose _____ (see list below) matches a given _____ (see list below)."

- feature code	_____	_____
- spelling	_____	_____
- feature class, e.g.,		
hypsographic	_____	_____
hydrographic	_____	_____
cultural	_____	_____
vegetation	_____	_____

3. What other types of queries might be useful?

3.0 ADVANCED SYMBOL PROCESSING (ASP)

Overview

Editing and formatting text with diacritics and foreign symbols requires specialized hardware and software. ASP is intended to provide these word processing capabilities and to be the sole read/write link to the names data base. Technology is ripe for ASP. Unfortunately, ASP delivery has been delayed because of its tie to the rest of the geonames processing system. The alternative (developing ASP outside the overall system framework) requires that flexible interfaces be designed, since it is too early to predict what hardware or software the other subsystems will use.

ASP Requirements

Of the four names processing subsystems, only ASP has a prototype. The Names Input Station (NIS) consists of a Plessey PDP-11/70 with a 2-MB disk connected to an ECD intelligent terminal. Peripherals include a tape drive, a printer, and a pen plotter. NIS hardware cannot be considered a model for future development, although much has been learned. A storage space of 2 MBs is inadequate. The Florida Data printer, selected because it has unusually high resolution for an impact dot matrix printer, is nonetheless slow, noisy, and produces low-quality output.

The most serious NIS problems were caused by the intelligent terminal. NIS diacritics word processing uses terminal commands that operate only on the text held in the terminal screen buffer (approximately 132 lines). This constrains such activities as sorting and global changes. And the terminal is programmed by a little-known macro language (that differed for a second NIS ECD terminal), making software hard to maintain.

ASP requirements are for a multistation system that is interfaced to the Multiset III. NIS keyboard layout and software work well and should be adapted to the production system. Output formats, too, should be maintained. The NIS diacritics input scheme, which divides diacritics into Regional Diacritics Sets (REDS) for ease of entry, should be implemented. However, hardware improvements must be made. Common, off-the-shelf hardware must be used. High-speed nonimpact printer/plotters have good graphic quality and are affordable.

Enhancements to the keyboard are desirable. The diacritics input scheme, as implemented on the ECD terminal, uses two outboard function keypads in addition to the standard QWERTY keyboard. Different diacritics and letters are assigned to each key, depending on the linguistic region being processed. Labels can be affixed to the keys to help the analyst remember the keyboard arrangement being used. A better alternative would be color coding or exchangeable keyboards.

ASP Configuration

The ASP configuration can be dumb terminals driven by a central processor or it can be microcomputer workstations with local storage and software. If the ASP production model predates delivery of names data base hardware and software (as seems likely), ASP designers must develop a stand-alone system that later can be integrated into a larger system.

Microcomputer workstations

Designing ASP as a series of microcomputer workstations provides a number of benefits. Microcomputer manufacturers have developed interactive devices and utilities that are uncommon in larger systems. Fixed disks holding 10 to 30 MB are becoming standard equipment on new microcom-

puters, with tape backup also available and inexpensive. Microcomputer workstations can stand alone for interim ASP capabilities. When the data base is installed, the workstations can be operated as dumb terminals while querying the data base. Once datasets are downloaded, local processing begins.

The major problem with such a configuration is the interim system's interface to other DMA systems and its ability to read archived names data from 9-track tape. A poll of microcomputer manufacturers produced no guidance on ways to read or write to standard reel-to-reel tapes without the help of a mini- or mainframe computer (microcomputer tape drives use cassettes or cartridges). The easiest solution is to interface the microcomputers (via hardware or modem) to an interim mini- or mainframe computer for tape interfacing until the data base mainframe is installed. The NIS's unused Plessey could be harnessed to this task if it is in good operating condition.

Centralized processing

A minicomputer can provide the processing power and manage centralized storage for dumb terminals. In many ways, this is the easier of the two configuration alternatives, since centralized processing technology is mature. If a centralized processing scheme were used for the interim ASP, it could be integrated into the names processing system later in one of two ways. Either the ASP software could be moved to the data base computer and the interim ASP computer retired or reassigned, or the ASP computer could be networked to the data base computer.

Display of diacritics and keyboard programmability introduce problems with using dumb terminals. Some level of intelligence may be required to provide these capabilities, whether the configuration is centralized or distributed. If intelligent terminals are used, care must be taken that programming is simple and model changes by the manufacturer will be minimal.

Recommended Software Upgrades

A number of toponymic aids can be added to the analyst aids provided by the geonames processing system. Data base or ASP software will transform coordinates from UTM to geographics and vice versa, determine the map series sheet number on which a geographic coordinate appears, and perform metric-to-English conversion. Upgrades could provide sophisticated toponymic aids. An on-line telegrapher's code helps reverse ideograph Romanization. On-line transliteration rules, as shown in gazetteer forewords, can be displayed to toponymists or added to files. An on-line dictionary defines all feature classes. These upgrades entail minimal expenditure to enter and manage simple files and look-up tables.

4.0 ADVANCED TYPE PLACEMENT (ATP)

Overview

DMA is introducing technology to produce maps and charts by interactively processing digital geographic data. Map type, too, must be generated and placed digitally to avoid production slowdowns. A digital map type placement system must display in softcopy the map symbols and the type styles and diacritics in final form. It must include interactive graphics utilities to manipulate the type on the map. There are a number of forms ATP can take that incorporate these capabilities.

Applicable Technology

The USGS interactively places digitized placenames on its provisional maps. The type placement procedure includes names capture. A map is fastened to the table of a graphics workstation. Location is digitized, the name is typed, and type style is assigned by the analyst. The name appears on the workstation display in its assigned font and size. The monochrome display also shows map linework and point symbols. The analyst employs interactive utilities to move, spread, and curve the type. Elegant use of cursor and keypad makes the process quick. The margin is laid out in batch after the analyst supplies the product-specific marginal information. Finally, the digital type separation is reproduced in hardcopy for graphic reproduction.

A number of batch methods are available to place point, line, and area labels on virtual maps relative to their stored features (Kelly, 1980; Hirsch, 1982; Greggains, 1982; Lewis, 1982; Basoglu, 1983; Ahn, 1984; Pfefferkorn, in progress). None of these methods have been used in production or undergone testing in a realistic mapping environment. Most algorithms assume that all names in the data set can be placed, and therefore focus on conflict avoidance. If a tractable data set is not passed, however, these algorithms must blindly attempt to place unplaceable names until time expires.

Unfortunately, no selection procedure exists that infallibly produces 100%-placeable names data sets. Even humans must hypothetically arrange selected names on a scratch map, since the final determiner of inclusion or exclusion is space available. Algorithmic selection criteria must consider spatial distribution as well as feature importance. In cartographic and geographic terms, relative location is a primary indicator of feature importance. Five settlement selection algorithms that meet this requirement have been developed by Langran and Poiker (in progress).

Recommended Approach

If digital names data are not data based, few upgrades to the USGS method are possible. If the mapped area names and their associated feature data can be compiled from a data base, however, a large portion of the process can be performed in batch. This strategy is recommended.

Figure 4-1 outlines the use of batch modules to improve both throughput time and the analyst interface. Following compilation, names data is verified by a toponymist (if necessary), then passed to a series of batch subroutines. Batch software filters map names based on map scale (which dictates space available), the relative location of surrounding names, and named feature importance (all excluded names are saved in a file, for later review by the analyst). Then, a batch subroutine places the names on a virtual map, continuing the edit process in areas where overcrowding prevents fit. Finally, the analyst displays the map in softcopy to review the computer's selection and placement of names.

Batch subroutines allow the analyst to work on a legible graphic representation of the map. A large percentage of the placed names need not be moved, since most point feature labels are easily

placed. Linear and areal placenames are more problematic. Their placement will depend on the link between map feature files and names files. A key element to effective automated names placement is that names selection must precede and coincide with placement. The nominal effort required to implement available selection algorithms into available placement algorithms is a worthwhile investment, improving processing times and allowing analysts to work on legible displays.

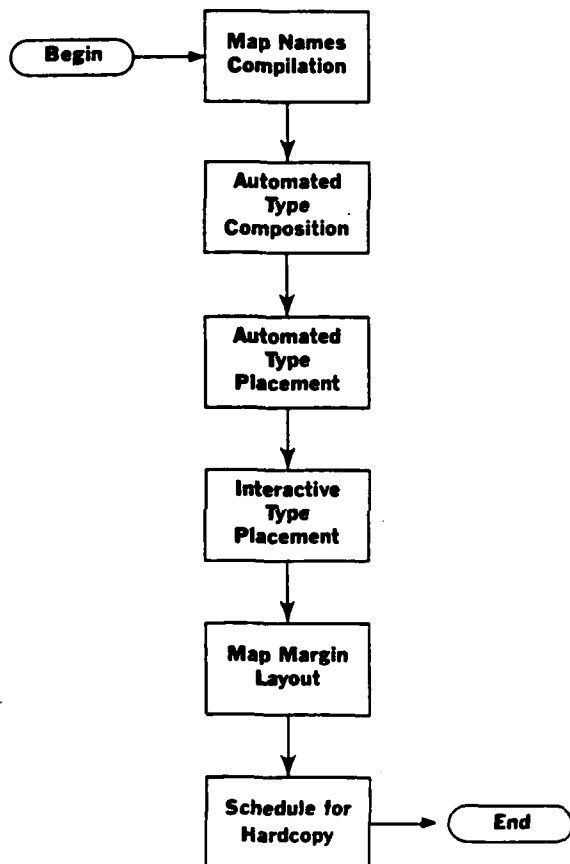


Figure 4-1. Recommended ATP process flow.

PART TWO: OVERALL SYSTEM ISSUES

Part Two discusses concerns common to all four geonames processing subsystems. Section 5 defines interfaces between the subsystems. Section 6 weighs the technical possibilities of handling non-Roman alphabets and ideographs. Personnel and space requirements for a comprehensive names processing system are estimated in Section 7.

5.0 SUBSYSTEM INTERFACES

Functional Interfaces

The functional interfaces between the subsystems are illustrated in Figure 5-1. Brief descriptions follow.

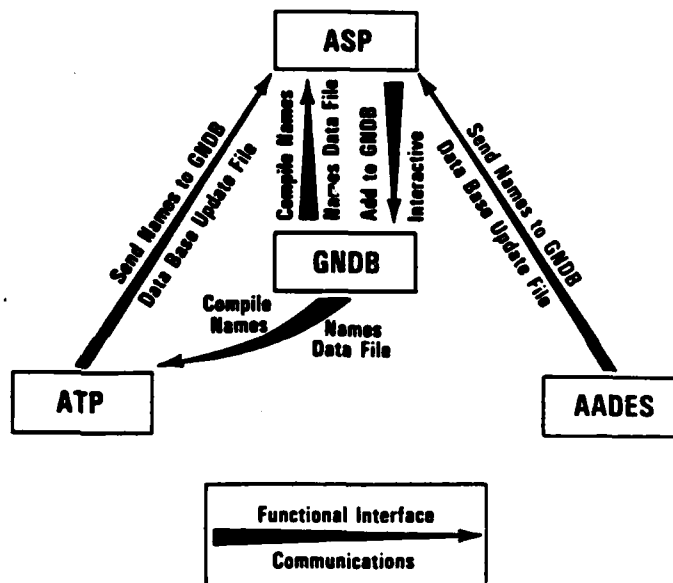


Figure 5-1. Interfaces between the subsystems

ASP is the only subsystem that can write to the GNDB. Thus, input from AADES and feed-back from ATP are entered into the data base via ASP. Files are compiled from and interactive queries are made to the data base using ASP terminals.

AADES feeds names data files to the names data base by way of ASP. AADES does not interact with ATP.

ATP compiles names data from the names data base as the first step in creating a map names overlay. Updates to compiled data are made through ASP.

The majority of system users have read-only access to the GNDB. Read/write access is granted to a limited number of ASP user accounts. Input from AADES is verified at ASP workstations, then loaded in batch.

Data Link

A basic file structure has been designed to simplify subsystem interfacing. Data Base Update files are used by ATP and AADES to send data to the GNDB via ASP. Data Base Update files are formatted to expedite routine toponymic comparisons and merging of same-area data sets gathered from different sources, including those compiled from the GNDB. The header describes the origin, ownership, and status of the file as a whole. It is a 256-byte character record that includes analyst name, file creation date, and comments. Comments describe file processing status. Data Base Update file contents are in Standard Data Transfer format. This file is comprised of Standard Data Transfer records (Table 5-1) sandwiched between 256-byte comment fields.

Table 5-1. Standard data transfer record

<u>Entity Name</u>	<u>Size (Bytes)</u>
Data Source Name	10 (1)
Number of Characters in Geoname	1
Number of Characters in Non-Anglicized Name	1
Number of Characters in Alias	1
Number of Characters in Province Name	1
Number of Characters in Country Name	1
Names (geoname, non-Anglicized name, alias, province name, country name)	140 (2)
Type of Romanization	1
Date of Data Source	3 (3)
Date of Data Capture	3
Date of Last Update	3
Position	6 (4)
Positional Accuracy	2
Feature Designator	6
Attribute	6
Administrative Code	1
Area Code	1
UTM grid	8
Selected Map Sheet	7
Approved or not Approved	1
Bounding Rectangle	13 (5)
Pointer to File Containing Feature Coordinates	8
Unused	32
	256

(1) The GNDB maintains a dictionary of legal data sources.

(2) If more than 140 characters are required, the next record is an overflow record. All names are stored in this field to substitute one large field with overflow allowances for potentially five large fields with possible overflows.

(3) Dates are numeric strings: dddmmyy.

(4) Position as currently planned is a point (the location of a point feature, the mouth of a river, or the centroid of an area feature) given as two signed numeric strings: +/- dddmmss and +/- dddmmss. Negative indicates latitude South or longitude East, positive indicates latitude North or longitude West.

(5) The bounding rectangle is high and low latitudes and longitudes, with an additional byte indicating if the bounding rectangle is incomplete due to the feature leaving the map.

6.0 NON-ROMAN SCRIPT PROCESSING

This section discusses digital ways to handle non-Roman alphabets and ideographs. Background information is provided, then two processing methods are described and one is recommended.

Background

Non-Roman alphabets include Arabic, German, Greek, Hebrew, Cyrillic, and Korean. The number of symbols in these alphabets (the sum total of lower and upper case characters) ranges from a low of 28 (Arabic and Hebrew) to a high of 62 (Cyrillic). Given appropriate equipment, processing Greek and Cyrillic is no more difficult than Latin alphabet processing, although such capabilities add to development expense. The Latin alphabet is easily substituted for the German alphabet. Korean, Hebrew, and Arabic, however, employ graphic relations as well as graphic shape in their alphabets, making them relatively more difficult to implement. Korea's 40-character "hangul" can take approximately 500 graphic forms by varying character arrangements.

ASCII codes exist for all non-Roman alphabets. DMA's Multiset III system currently supports ASCII processing of Greek, Cyrillic, and Korean. Because non-Roman alphabet processing is technically risky (although it adds to costs), the majority of this discussion focuses on technical alternatives for ideograph processing.

Chinese hanzi and Japanese kanji are fundamentally the same, although many kanji characters have been simplified. The largest Chinese scholarly dictionaries list 50,000 hanzi characters. However, China has established a 6763-character standard set (PRC Information Exchange for Chinese Character Codes, GB2312-80). In Japan, a set of 1850 selected kanji called "toyo kanji" are used in newspapers and official documents. It is not known how either of these character subsets would relate to the characters used in placenames.

Hanzi and kanji are constructed with one or more of 214 different "radicals." Radicals are the key to some kanji dictionaries. All characters with the same radical are listed together by increasing number of strokes.

Ideographs are independent of pronunciation. Many ideographs have the same meaning in China and Japan, but their spoken words vary. Even within China, different dialects pronounce characters differently. Thus, phonetic character sets have been developed. China's phonetic scheme, "pinyin," is comprised of 37 symbols and four tone marks and is taught in Chinese schools. Japan has two syllabaries ("kana") that are used for transliterated foreign words, exclamations, and grammatical inflections. "Hiragana," used for exclamations, and "katakana," used for foreign words, each have 48 characters. Some characters may be modified by diacritics (nigori) to make 25 additional characters. Japan's syllabaries are not important for names processing: most nouns, especially names of persons and places, still are written with kanji characters.

Electronic Processing of Ideographs

Digital text processing entails data entry, storage, retrieval, edit, and display. The way characters are represented to the computer is the fundamental technical decision. Characters can be computer-encoded as bit maps (images) and stored in a character field in the data base; or, characters may be ASCII-coded. The impacts of each are discussed in the next paragraphs.

ASCII coding

Two-byte ASCII codes exist for all ideographs. Thus, storage and retrieval are not serious problems. Two-byte ASCII-coded kanji is supported by Fujitsu in its AIM/RDB data base. No other non-Roman data base support is known.

Ideograph input and edit are subject to a number of difficulties. Character entry methods include keyboard (QWERTY and others), voice entry, and optical character recognition (OCR). OCR methods include page readers that constrain character forms, paper quality, document size, and real-time OCR that uses a digitizing tablet or light pen. Real-time OCR can use stroke order for more reliable character recognition, but its speed is constrained by the calligrapher's speed.

Keyboard entry systems use varied strategies. "Hunt and peck" drives the Japanese typewriter, a two-dimensional array of 2000-plus keys (one character/key, organized by radical). The Research Library Information Network (RLIN) system at Stanford University uses 214 keys (one radical to one key) to build characters by assembling their radicals.

Touch-typing strategies use encoding schemes to reduce the number of keys by increasing the keystroke-per-character ratio. A two-stroke code, memorized by rote, is popular with professional Japanese typists. For occasional use, however, rule-based coding schemes that key on character shape or pronunciation are better.

Coding based on pronunciation requires language-specific translation software. The Organizational Committee of Library Councils uses Asia Graphics Development software running on IBM microcomputers to input Chinese characters by spelling their names phonetically on a QWERTY keyboard. Ambiguities are resolved interactively. Xerox markets word processors for JACKPHY (Japanese, Arabic, Chinese, Korean, Persian, Hebrew, and Yiddish) and Thai that use a similar pronunciation-oriented strategy.

The input systems described above have one thing in common: they all require some familiarity with the input language. Voice entry and pronunciation-based keyboards require that ideograph names are known, and that they can be pronounced in a certain dialect. Real-time OCR requires knowledge of calligraphy. Operators must know radicals and stroke orders to use the Japanese typewriter and the RLIN system.

An alternative to systems requiring linguistic expertise is a 12-key device that allows character input based on appearance only. The 6763-character Chinese standard distinguishes 30 stroke types (e.g., short horizontal, long vertical). By linking keystroke length to stroke length and stroke type, a promising input method with few keys, simple rules, and minimal ambiguity among characters was developed (Clark, 1984). If ASCII-coded non-Roman script processing is required, this method is the most promising.

Bit map encoding

Bit map encoding of non-Roman script makes data entry, storage, and display considerably less complicated than ASCII-based processing. Retrieval of bit-mapped ideographs from a data base must use the Latin version of the name as an access key, a fairly minor restriction in names processing.

Capturing bit maps requires raster-scanning, isolation of text, and removal of background noise. The bit mapped image of the non-Roman script is stored in the data base as characters (character string length could preclude using certain commercial DBMSs).

Editing the raster image could be performed by rescanning (possibly using a scanning wand), or by interactive graphic edit software to correct disconnected or misplaced linework. Engineering Topographic Laboratories developed font-editing software for DMA that could be adapted to edit ideographs.

Raster graphics techniques would be used for hard- and softcopy display of bit-mapped ideographs. Laser printers and plotters, or the CRT printhead, could be configured to output the character images.

Recommendation

Non-Roman script should be processed by DMA as bit maps. The shortcoming of this approach (names must be retrieved from the data base using their Romanized counterparts) is offset by the ease of developing and using a bit map system.

Ideographs are needed for two activities. Toponymists need a placename's non-Roman form to confirm that the placename is unique and not a variant spelling or alias of an approved name. It is natural in this situation to retrieve the ideograph using its known Roman counterpart. Ideographs are also needed for bilingual maps. In this case, too, an ideograph image suffices, since such maps are likely to be printed using bit-mapped graphics (the CRT printhead or other raster device).

7.0 PERSONNEL AND SPACE

Overview

Geonames processing system users fall into the two categories of applications and systems support (Fig. 7-1). A data base administrator oversees both applications and systems support concerns. Personnel and space requirements for applications and for systems support will be estimated separately.

Data base administration may be performed by an individual or by committee. Primary responsibilities include establishing and policing standards for data size, format, and usage; administering detailed system documentation; and coordinating user needs in light of current system capabilities and data resource development. System and production statistics are directed to the data base administrator to assist in policy decisions.

The four subsystems cannot be compartmented neatly into the systems or applications categories. Most data base functions are clearly in the realm of systems support. Type placement, while requiring hardware and software maintenance, is mainly an applications function. However, data capture and the ASP subsystem are crossovers.

In this operational concept, duties are divided according to the nature of the hardware used and the type of expertise required. For example, the data processors who capture data using AADES and the toponymists who check the quality of the captured files are considered applications personnel (the data processors because they will use cartographic workstations, toponymists because of their specialized skills). However, those who oversee the batch loading of corrected files into the data base are systems personnel. Likewise, ASP hardware and personnel that maintain the data base are in the systems group, while ASP resources used for toponymic analysis and product generation are classed as applications.

Systems Support

Personnel requirements

A system manager oversees the systems support team. He/she must have expertise in computer systems and data structures. Familiarity with the data base's logical and physical design is essential for effective software maintenance and continual system upgrading.

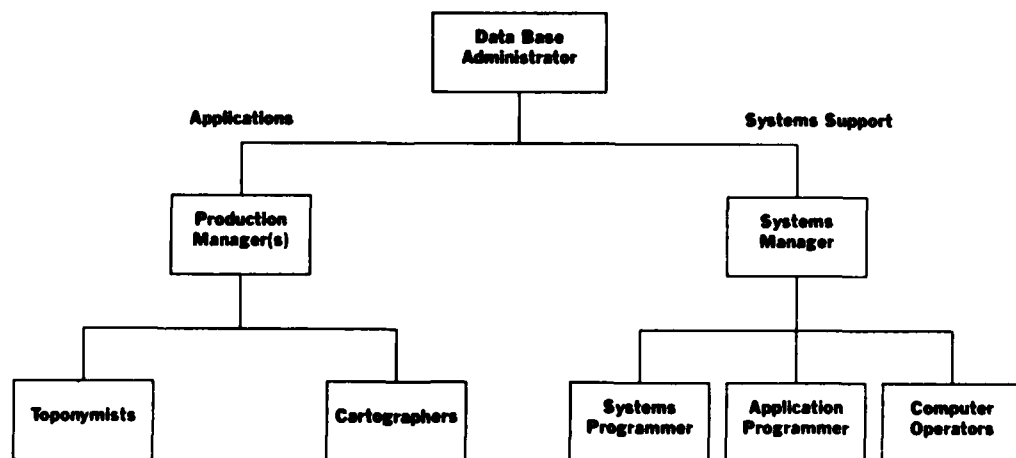


Figure 7-1. Personnel requirements.

A systems support staff performs the software maintenance and upgrades dictated by the system manager. The support staff includes a minimum of two systems programmers, two applications programmers, and one computer operator.

Space requirements

Table 7-1 is a space estimate for the systems support group. This estimate includes the space required for personnel and hardware, which includes the data base computer, mass storage, consoles, and ASP terminals for the system manager and each programmer.

Table 7-1. Systems support hardware and personnel space estimate.

Hardware system and mass storage	800
6 administrative/support staff @ 120 sq. ft.	720
5 ASP workstations @ 30 sq. ft.	150
	1670 sq. ft.

Applications

Personnel

The production manager is in charge of applications processes and personnel. Because applications includes a range of tasks, production management could be shared among several people.

Applications analysts include those toponymists and cartographers currently involved in names processing and DMA's production centers. Added to this group are data processors to perform rote clerical and data entry tasks.

Space

Table 7-2 is a space estimate for applications hardware. Space requirements depend on the number of workstations (toponymic and cartographic) to be procured, and on the magnitude of the data capture effort. Space estimates here are for 15 toponymic workstations, four cartographic workstations, and one scanner.

Table 7-2. Applications group hardware space estimate. This estimate should be added to the space currently occupied by applications personnel and materials.

15 ASP workstations @ 30 sq. ft.	450
4 cartographic workstations @ 100 sq. ft.	400
1 scanner @ 100 sq. ft.	100
	950 sq. ft.

PART THREE: SUMMARY OF RECOMMENDATIONS

Recommendations for addressing system design issues have been made throughout this report. These recommendations are summarized in Part Three, Section 8.

8.0 SUMMARY AND CONCLUSIONS

This design study has devised ways to meet all the needs expressed by DMA personnel contacted during system requirements definition. An important next step is to weigh the difficulty of responding to a requirement against the requirement's importance. A DMA working group should be formed, its membership coming from management and from the cartographers and toponymists who will be the names processing system's heaviest users. This working group must refine requirements based on the difficulty of providing certain capabilities, and establish priorities for system development. The following recommendations are offered.

Data Capture

Better data entry technology must be developed or reasonable data base coverage will not be possible for 10 or more years. A data entry system that uses raster technology, automated character recognition (OCR), and analyst interaction to capture names data from maps is the recommended configuration. The first version of such a system would rely heavily on the analyst, since OCR is an evolving technology that would be challenged by international text and map formats. Improvements in OCR, however, could be implemented in system upgrades for eventual rapid and highly automated data capture.

Hardware and development costs are higher for the recommended softcopy-based data capture strategy than for a hardcopy-based design. But, potential labor cost savings offset the technology investment (Section 1).

Data Base

Two issues make data base design problematic. First, the 50-100 million name data base will be very large by today's standards. Size restricts choice of hardware and software and makes data organization crucial. Second, the names data base must interface with a cartographic data base storing feature coordinates and attributes. Interfacing two very large data bases is a serious design problem. Possible interfacing strategies were discussed in Section 2. The recommended strategy is to design the names data base as a subset of the feature data base. Although combining the two data bases will aggravate size problems, duplicate data need not be stored if the two are combined.

Edit and Format

Technology is ripe for the diacritics word processor needed by toponymists at DMA. An immediate effort should be undertaken to develop and procure this device. The distributed and the centralized configurations described in Section 3 both have merit. The delivered system must stand alone initially, but must later be hardwired into the data base computer.

Map Names Processing

An interactive, all-digital type placement system can be procured off-the-shelf today. Algorithms that perform type layout in batch are available and are being continually improved.

Integrating automated algorithms into an interactive system requires a clear perspective of map names processing. For batch algorithms to be effective, map names selection must be integrated into type layout procedures. Without batch names selection procedures, little is gained from batch type layout. Pipeline timing suffers when an analyst must interject twice during map processing. And automated type placement results are poor if the system is not free to alleviate overcrowding by eliminating names. Automated selection and automated names placement are reviewed and, if necessary, corrected by an analyst.

Subsystem Interfaces

A standard names data transfer record is defined to simplify subsystem interfacing. A standard file format also simplifies data base loading.

To control data base content, only authorized users can write to the data base, and write operations can be initiated only from the ASP subsystem. Inputs from AADES and feedback from ATP are sent to toponymists working at ASP terminals for review and acceptance. If accepted, data are written to the data base in batch operations during off-peak hours.

Non-Roman Script Processing

Two basic methods of non-Roman script processing are described in Section 6. One method defines characters by their ASCII codes. The second method defines characters by their raster images. The second method is recommended because it simplifies capture and storage of the characters. The major drawback of bit-mapped non-Roman characters is that non-Romanized names must be retrieved from the data base by keying on their Romanized counterparts. It is felt that this concession does not constrain names processing.

Conclusions

The descriptions and recommendations stated in this report summarize findings of a 2-year design study. Of necessity, many design details have been omitted. Interested readers must refer to the 5-volume functional design specifications (Langran et al., 1985) for clarification of the information provided in this project overview.

APPENDIX A

BIBLIOGRAPHY

Ahn, John. "*Automated Map Name Placement System*." Image Processing Laboratory, Rensselaer Polytechnic Institute, Troy, New York: May 1984.

Augustine, R.F., D.R. Caldwell, and D.E. Strife. "A Prototype Geographic Names Input Station for the Defense Mapping Agency." *Auto Carto V*, Reston, Virginia: August 1982.

Basoglu, Umit. "*Automated Name Placement*." Draft of final report of findings (Phase II). CACI, Inc. Washington, D.C.: April 26, 1983.

Becker, Joseph D. "Typing Chinese, Japanese, and Korean." *Computer* 18 (January 1985): 27-34.

Brown, R., A. Zeid, A. Barnes, and E. Gough. "Advanced Type Placement and Geonames Database: Comprehensive Coordination Plan." Naval Ocean Research and Development Activity Technical Note 189. NSTL, Mississippi: January 1983.

Campbell, John, Edward Gough, and Gail Langran. "Commercial Data Base Management System for Geonames Data Base." NORDA Report 73. NSTL, Mississippi: December 1984.

Clark, W. A. "Chinese Transcription Device—Final Report." Sutherland, Sproull, and Associates, Inc. 30 November 1984.

Cui, Wei. "Evaluation of Chinese Character Keyboards." *Computer* 18 (January 1985): 54-59.

Fay, Temple. "Handprint Symbol Recognition: Alphabetic Characters (preliminary results)." NORDA Report 114, NSTL, Mississippi: 1985.

Greggains, Alan. "A Strategy for Name Placement." Unpublished report, University of Cambridge Computer Laboratory, September 1982.

Hershey, A.V. "Calligraphy for Computers." U.S. Naval Weapons Laboratory, Dahlgren, VA. August 1967, AD 662398.

Hirsch, Steven. "An Algorithm for Automatic Name Placement Around Point Data." *The American Cartographer* 9 (June 1982): 5-17.

Huang, Jack Kai-tung. "The Input and Output of Chinese and Japanese Characters." *Computer* 18 (January 1985): 18-24.

Jablinski, R. D. Strife, K. Gaar, and J. Moore. "Names Type File System." Consulting Report for the UASETL Project #P0013. April, 1983.

Kelly, Paul C. "Automated Positioning of Feature Names on Maps." M.A. Research Report. Department of Geography, State University of New York at Buffalo, 1980.

Langran, G., A. Downs, B. Glick, W. Schmidt, A. Barnes, and S. Miller. "Geonames Processing System: Functional Design Specifications Volumes 1-5." NORDA Reports 98-102, NSTL, Mississippi: 1985.

Lewis, Gail E. (now Langran). "Automated Point Labeling for Geographic Data Bases." M.S. Thesis, Department of Geography, Western Washington University, Bellingham, Washington, 1982.

Makino, Hiroshi. "Beta: an Automatic Kan-Kanji Translation System." *Computer* 18 (January 1985): 46-52.

Matsuda, Ryouichi. "Processing Information in Japanese." *Computer* 18 (January 1985): 37-45.

Opalski, William E. "Automation of Foreign Place Names at the United States Defense Mapping Agency." Unpublished report, USAETL, Ft. Belvoir: 1980.

Payne, Roger L. "Geographic Names Information System." U.S. Geological Survey Circular 895-F, Reston, Virginia: 1983.

Sheng, Jian. "A Pinyan Keyboard for Inputting Chinese Characters." *Computer* 18 (January 1985): 60-64.

Stone, Harold S. "Computer Research in Japan." *Computer* 17 (March 1984): 26-33.

Tien, H.C. "A Pinxxiee Chinese Word Processor." *Computer* 18 (January 1985): 65-66.

U.S. Army Engineering Topographic Laboratories. "Error Detection in a Toponymic Database: a Case Study Using a Database Management System." Unpublished report. Ft. Belvoir, Virginia: 1980.

U.S. Army Topographic Command. "Geographic Names Data Base Study." Corps of Engineers, Washington, D.C.: July 1969.

U.S. Geological Survey. "Geographic Names Information System." Reston, Virginia: September 1982.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

AD-A163042

REPORT DOCUMENTATION PAGE				
1a. REPORT SECURITY CLASSIFICATION Unclassified		1b. RESTRICTIVE MARKINGS None		
2a. SECURITY CLASSIFICATION AUTHORITY		3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.		
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE				
4. PERFORMING ORGANIZATION REPORT NUMBER(S) NORDA Report 125		5. MONITORING ORGANIZATION REPORT NUMBER(S) NORDA Report 125		
6. NAME OF PERFORMING ORGANIZATION Naval Ocean Research and Development Activity		7a. NAME OF MONITORING ORGANIZATION Naval Ocean Research and Development Activity		
6c. ADDRESS (City, State, and ZIP Code) Ocean Science Directorate NSTL, Mississippi 39529-5004		7b. ADDRESS (City, State, and ZIP Code) Ocean Science Directorate NSTL, Mississippi 39529-5004		
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Defense Mapping Agency	8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER		
8c. ADDRESS (City, State, and ZIP Code) HQ/STT Washington DC 20305		10. SOURCE OF FUNDING NOS.		
		PROGRAM ELEMENT NO. 64710B	PROJECT NO.	TASK NO.
11. TITLE (Include Security Classification) The Geonames Processing System Synopsis				
12. PERSONAL AUTHOR(S) Gail Langran				
13a. TYPE OF REPORT Final	13b. TIME COVERED From _____ To _____	14. DATE OF REPORT (Yr., Mo., Day) September 1985		15. PAGE COUNT 35
16. SUPPLEMENTARY NOTATION				
17. COSATI CODES		18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)		
FIELD	GROUP	SUB. GR.		
		maps, computers, software systems		
19. ABSTRACT (Continue on reverse if necessary and identify by block number) DMA has recognized a need for digital procedures to store, retrieve, and edit geographic names data and to prepare names for product generation. DMA's stated goal is a 50-100 million name digital data base with subsystems to capture names, edit and format names data, and prepare names overlays for maps. NORDA began a geonames processing system design study late in FY82. This report summarizes NORDA's study findings.				
20. DISTRIBUTION/AVAILABILITY OF ABSTRACT UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☒ DTIC USERS ☐		21. ABSTRACT SECURITY CLASSIFICATION Unclassified		
22a. NAME OF RESPONSIBLE INDIVIDUAL Gail Langran		22b. TELEPHONE NUMBER (Include Area Code) (601) 688-4449		22c. OFFICE SYMBOL Code 351

END

FILMED

24-86

DTIC