DEFENSE MAPPING AGENCY LARGE SCALE DATA BASE INTEGRATION

Francis M. Mirkay
Defense Mapping Agency
U.S. Naval Observatory, Washington, DC 20305

ABSTRACT

The Defense Mapping Agency (DMA) has massive amounts of MC&G digital data holdings resulting from programs to support DoD advanced weapons systems and automated map and chart production. The digital data is used to support the DoD and other countries as well as internal DMA production. DMA requires a digital data base to maintain, store, retrieve, compare, and report on this MC&G digital data. This paper addresses current DMA efforts that will evolve into an on-line interactive, distributed, networked data base system and its associated environment for data handled by HQ DMA and the DMA Production Centers.

BACKGROUND

The Defense Mapping Agency (DMA) has massive amounts of MC&G digital data resulting from programs to support DoD advanced weapons systems and automated map and chart production. The digital data is used to support the DoD and other countries as well as internal DMA production. DMA requires a logically single automated digital data base to maintain, store, retrieve, compare, and report on this MC&G digital data.

The collection of digital data began in the 1960's when the DMA Hydrographic/Topographic Center (HTC) (formerly the U.S. Army Map Service) began collecting terrain data for use in an automated system of carving three-dimensional relief maps. A digitizing table called a Digital Graphics Recorder (DGR) recorded the position and elevation values of traced contour lines from cartographic source materials. The final output magnetic tape controlled a carving or milling machine.

In 1966, the Universal Automatic Map Compilation Equipment System (UNAMACE) was put into production at HTC. Digitized elevations are produced directly from rectified stereo imagery and are used to output orthophotographs, altitude charts, and terrain elevation matrices.

In 1972, the DMA Aerospace Center (AC) entered a new realm of production with a digital product to simulate radar displays for pilot training which today is referred to as the Digital Landmass System (DLMS). The DLMS product requires a matrix of terrain elevations called Digital Terrain Elevation Data (DTED) which, at the level of least refinement, is produced by 1-degree squares with elevation values (posts) at 3-second-of-arc intervals in both directions. DLMS also requires cultural data called Digital Feature Analysis Data (DFAD) to predict the radar reflectivity of the earth's surface. The DLMS production program covers over 18 million square nautical miles. It requires the collection of data from cartographic and photographic source materials using the Automatic Graphic Digitizing System (AGDS) (AC and HTC), stereoplotters (AC and HTC), DGR (HTC),

and other systems from both Centers to meet the production goal. This program continues today.

In 1973, DMA recognized the need for a single data base system to provide centralized management for digital data related to advanced weapons systems and automated chart production processes. The Cartographic Data Base (CDB), which became operational in 1974, provided the centralized management needed to serve DMA. The CDB initially contained the DTED and DFAD produced from photographic and cartographic source materials at AC and HTC. The CDB provided digital data accountability, graphics of area coverage, and paper copies of historical portfolios for all digital production projects. The CDB was designed to have separate file management systems for each type of digital data in a modular design to generate, maintain, sort, and retrieve data, providing an integrated digital data base to meet production requirements. HQ DMA assigned the data base management function for digital data to AC in 1975.

As computer technology grew, production methods for required products created more MC&G data in computer readable form. The Semi-Automated Cartographic System (SACARTS) went into initial production at HTC in 1973. This system of computer hardware and software allowed the cartographer to digitize map and chart manuscripts codifying features such as lakes and roads. Taking hypsographic data from the UNAMACE and elevation and cultural data from vector or line digitizers (i.e., the CALMA and BENDIX), color separation scribe coat plates were produced on a precision Concord Plotter. AC also developed chart data in digital form called Automated Chart Features (ACF) data. Conceptually, the cartographer traces, tags, and edits features on the Automated Graphic Digitizing System (AGDS), composed of a digitizing table and two CRT's for menu selection and display. The resultant output (ACF data) drives a photohead plotter for the color separation plate.

New forms of digital products evolved out of expanded user requirements. In the 1970's, the Strategic Air Command (SAC) required Terrain Contour Matching (TERCOM) data to support the cruise missile. Also identified was the need for Vertical Obstruction Data (VOD) to support SAC. AC had maintained the DMA Vertical Obstruction File (D-VOF) to support manned or other flight vehicles.

The possibility of extracting data from one set to serve another purpose became more reasonable as more digital data became available. As an example of multi-use, DMA maintains a set of data known as Minibloc. Minibloc data is used for a special chart overprint where Maximum Terrain Elevation (MTE) readings for 10-minute x 10-minute areas are extracted from DTED. This was achievable with increased availability of DTED.

Because the simulations for flight training are becoming more realistic, it is anticipated that a user requirement for a visual data base simulation, which includes color of buildings, windows and other previously unrecorded details, will evolve into a validated production program.

Of major significance to DMA are such military applications of digital data as support of the major weapons systems. For example, the cruise missile guidance system uses TERCOM data; land combat simulation and analysis use digital terrain analysis data; and navi-

gation and training simulators use Digital Landmass System (DLMS) data (DTED and DFAD).

As you can see, as a result of 15 years of generating digital cartographic data for various advanced weapon systems programs, DMA has accumulated 10^{11}-10^{12} (10^{12}-10^{13}) bytes of it. It's about time we figured out how to handle it.

DMA DATA BASE CAPABILITY EXPANSION

To manage and exploit the above-mentioned reserves of digital data effectively, DMA is currently involved in expanding its scientific and technical (S&T) computers with more memory and input/output devices as well as procuring a Data Base System (DBS) that will provide for an on-line, interactive, networked data base environment.

The tasks involved in achieving an interactive, networked data base environment with DMA are highly technical and complex. The current S&T computers will be upgraded to satisfy the needs for greater production thoughput. Concurrent with this upgrade will be the identification and implementation of the systems network. The DBS will be developed by a systems integrator (SI) that will be selected after a paid competition among two or more SI's. Each SI will provide a conceptual design of the DBS. DMA will select the best design for implementation. Each SI will be provided with a common set of DBS requirements in order for each to produce an appropriate design concept.

I will not dwell further on the hardware upgrade, software redesign, networking, or telecommunications tasks associated with the design and phased implementation of an interactive, networked data base system, but I will attempt to discuss DMA concepts of data basing in an interactive environment and data base design as it fits into the DMA MC&G arena.

A DATA BASE CONCEPT

The concept of an integrated data base needs to be substituted by that of a Leagued Data Base. Whatever we call the data base is not important. What is important is the concept behind the data base. For the sake of definition, however, the Leagued Data Base (LDB) is one with highly controlled logical redundancy in the schema which provides for improved usability of the data bases and enhances its life cycle performance.

The components of a LDB allow for collocation at a single site and distribution in a network. Most importantly, they are logically viewed as loosely coupled. Data base schemata must accurately and naturally model the application environment. It is important to recognize that a data base design that is equally good for all uses is also not particularly good for any of them.

In this regard, there is virtue in controlled redundancy. Alternate representations need to be maintained for a data item so that it may be accessed in different ways. One or several users may associate multiple meanings with a single item of information and use it in different ways. Therefore, this information should appear in several places in the schema that correspond to the several ways in which it is viewed. As an example, radio towers may appear in several

different data bases. In one, the information may be needed by a field commander to ascertain communication links. In another, it may be needed by navigation planners for obstruction avoidance. Although this produces redundancy into the schema, it provides the user with the information where he wants it. Whether or not the data is physically replicated is an issue of efficiency, turning on whether the costs of maintaining two versions of the same data item is counterbalanced by the access efficiency two copies provide.

Cost is undeniably an issue. The cost of a data base in terms of space and time are not the only ones that define its performance, and it may no longer even be its most important component. The cost of building and maintaining the applications programs that make use of the data base may be even more important in selecting designs for the data base.

The performance evaluation of the Data Base System (DBS) and the Data Base Management System (DBMS) must be in terms of excess/inefficient processing, excess device capacity, lengthy application development times, frequent data base reorganization, required reprogramming of applications programs, and the complexity of applications programs due to the DBS/DBMS design. These evaluation terms are very difficult to quantify, if it is possible at all to do so.

DATA BASE DESIGN

Logical design begins with an investigation of user requirements and ends with a logical description (schema) of a data base that can support those requirements. The logical design is logical because it does not contain details about how the data is to be represented. This information is used during the subsequent physical design phase. Four activities are involved in the logical design:

- Requirements Analysis
- Data Modeling
- View Integration
- Schema Development

Data base physical design begins with a logical schema representing user requirements and processing requirements. Four activities are involved in physical design:

- Data representation determined and documented
- Access modes selected and documented
- Allocate data to devices
- Load and reorganize data bases

Certain specific tasks are then followed: Using the data definition language (DDL) of the Data Base Management System (DBMS), assign data type and size to each data element and group in the schema. Next, the access methods for storage and retrieval are chosen for each element, record, or file. These are recorded in the DBMS

internal schema via the Device Media Control Language (DMCL). Now the actual data is loaded.

In its broadest sense, data base design encompasses activities that range from the identification of end-user requirements to the final management of data value on a physical device.

DATA BASE DESIGN TRADE-OFFS

The factors involved in data base design are numerous and interrelated. Consideration of all their relationships can enmesh, entwine, and even entrap the designer in a never-ending analysis without revealing the best data base configuration. To aid the data base design process, the design trade-offs inherent in this complex task need to be recognized, confronted, and evaluated.

Operational Trade-offs occur in both logical and physical design. During logical design, operational trade-offs are primarily related to strategies and tools selected for the development of the data base schema. While in the physical design phase, trade-offs involve alternatives for data base implementation.

General Trade-offs form the basis of a design philosophy that guide the formulation of implementation alternatives. These general trade-offs can be used to evaluate the feasibility of several approaches to implementation. There are five general trade-offs the data base designer must be cognizant of:

Specialization Vs. Generalization. Traditionally, file design is predicated on the needs of a specific application. Data is duplicated rather than shared, and storage and access decisions are made to optimize files for the primary user. In the data base environment, the emphasis must be placed on managing data as a corporate or leagued resource which changes the data base to a repository of shared information. This provides multiple views by multiple users of the same data whether it is physically replicated or not. Evaluating cost and performance in such an environment is complex. Care must be taken in this design decision-making.

Application and Configuration Requirements. The data base designer needs to make an economic trade-off between the power of the configuration and the requirements of the application by matching the structural and utilization requirements of the data base with the capabilities of the DBMS, access methods, and storage devices. The ultimate configuration should meet the requirements without providing significant unused capacity.

Future Planning. The design selected by the data base designer should be one that will be tenable for a number of years. He, therefore, must consider the life expectancy of the data base in conjunction with trends. New data structuring such as relational or set-theoretic data models should not be overlooked.

Planned Vs. Ad Hoc Processing. Decisions in design that lean toward planned processing put less emphasis on nonprocedural data base interaction. Ad hoc processing requires more storage overhead, more indexes and pointers, and is unnecessarily burdonsome if applications processes are known and repeated.

Extent of Required Analysis. Inefficient data base design and implementation causes severe and continuing penalties throughout the life of the DBS and, therefore, a front-end analysis is worthwhile and cost beneficial. Some degree of analysis is needed for most data bases. The depth of the analysis, however, must be weighed against its benefit.

REQUIREMENTS ANALYSIS

DMA has contracted for the performance and documentation of a Requirements Analysis (RA) for the DBS. In order to provide the systems integrators (SI's) with enough data to develop a conceptual design, the RA must provide the list and description of the global DMA functional areas requiring DBS support; the definition of the input, output, and transformation requirements for each functional area of the DBS; and the definition of the DBS support requirements such as the processing, communication, mass storage, and user interface requirements.

At a minimum, the RA will answer the following questions:

(1) Who defines/produces the data in the data base?

(2) Who controls/manages the data?

(3) How are security/privacy controls enforced?

(4) How are updating rights or restrictions enforced?

(5) How is the data base protected against inadvertent or intentional damage?

(6) How are recovery, backup, and restart from a software or hardware failure carried out?

(7) How are multiple, logical views of the same data described and maintained?

(8) Is logical or physical redundancy allowed or required?

(9) What are the existing and projected information needs of the Centers?

(10) What are the transaction flows, their volumes, responsiveness, and frequencies? (current and projected - FY 84-90)

(11) What are the application interrelationships in these flows?

(12) What are the current dependencies between major production systems (both manual and automated) and associated data?

(13) What kind of automated on-line indexes are required? (now and near future)

(14) What kind of queries are made? (related to items 10 and 11, above)

(15) What are interrelationships of files? (related to item 12, above)

(16) What are the relationships between data files and product lines?

(17) How can decentralized/on-line access be achieved?

(18) What is the effect of classification/security on the above?

The RA will be performed in a top-down manner. The final document will make maximum use of diagrams. The diagrams will define relationships of HQ DMA, HTC, AC, and outside organizations so far as general information flows and functions are concerned. Through functional decomposition and data decomposition techniques, lower and lower levels of detail will be described using diagrams and written definition. Decompositions will end with description of data elements and the input, processing, and output thereof.

REQUIREMENTS ANALYSIS METHODOLOGY

The DMA functions will be identified and described in the context of the overall DMA objectives and organization with special emphasis on organizational impacts. The information sources and projects required by each function will be identified. The location, geographically and organizationally, of each information source and product will be identified and described. Input, output, and transformation requirements of each information source and product will be described. This description will be done to a level of detail that will enable an SI to identify individual data items without implying specific aggregations that would constrain the system design.

From a global point of view, DMA functions are categorized as:

- Receiving
  - Images
  - Maps & Charts
  - Textual Information
- Planning
  - HQ DMA
  - Centers
  - DMA Product Preparation
- Management
  - Headquarters
  - Centers
- Requirements
- Geodesy & Control
- Image Analysis
- Production
- Crisis Support
- Distribution

- o Holdings/Inventory

Having developed a list of DMA functions, each function will be described in the context of DMA's objectives and organization. The information sources and products of each function will be identified, related to an organizational unit, and described. This process will be accomplished as follows:

- o Identify the organizational units assigned to tasks related to each function.
- o Identify the outputs required of each unit for that function:

  Who has managerial/organizational responsibility?

  Who is responsible for actual documentation?
- o Identify and collect all relevant documentation for each function and task.
- o Interview relevant personnel to define the current methodology and the future requirements.
- o Produce a draft functional description of the function. Include organizational and functional diagrams.
- o Review the draft and update it with the organizational unit and other DMA management.

As each functional area is described, the individual information sources and products will be identified. The final iteration of the top-down approach is the description of these information sources and products. Each will first be related to the responsible organizational unit. Then, for each organizational unit, the following will be described:

- o Data input (content, medium, source)
- o Processing requirements (transformations)
- o Data output (content, medium, format)
- o System and environmental requirements (e.g., frequency, time responsiveness, volume, security, and integrity requirements)

The final DMA-approved RA document will be provided to the SI's for their use in the conceptual design phase. Once a conceptual design is selected and an integrator employed, the phased detailed design and implementation process will begin.

TODAY'S DMA DATA BASE ENVIRONMENT

To form a picture of DMA's data handling challenges and its current operating environment, the following data base environmental characteristics are provided:

DMA has dozens of data bases, many of which approach a rather large size of 10^{13} bytes or larger. These data bases are product-oriented,

and, thus, some duplicate information and coverage. However, it must be noted that the data base schemata are reflective of product requirements and, therefore, physically redundant data items and names may be logically different as related to application, such as radio towers for cruise missile navigation from Vertical Obstruction Data and radio towers for scene simulation from DLMS. The current system is batch-oriented with very few interactive users, and processing resources are inadequate to meet the backlog of user requests. Most of DMA's digital collection systems are highly specialized, and, therefore, have limited data base abilities.

Based on the present operating scenario and production program requirements, a future forecast of the DMA operating environment can be made. Data base volume will increase. Similarly, the DMA workload, due to the expansion of data base volume, will undergo great expansion. There will be greater interdependency of data bases as well as a greater interdependency among digital products. There will be increased demands placed on DMA for more digital data to support more advanced weapons systems.

All of this being true, it is not unrealistic to see on the horizon significant problems arising from future growth. For DMA to be in an agressive posture and take advantage of new advancements and opportunities, an environment with the following characteristics must be developed:

- o A DBMS that supports a distributed, networked interactive system
- o On-line data bases with highly controlled logical redundancy to support multiple users and multiple product generation.
- o Advanced query capabilities
- o A DMA-tailored MC&G DBMS
- o Distributed, interactive access to data bases.

Furthermore, commercial developments in state-of-the-art computing and information processing power and storage technology must exponentially advance.

MC&G DATA BASE UNIQUENESS/FUTURE NEEDS

Conventional DBMS's provide data models such as relational, hierarchical, and network. These models do not totally satisfy the handling and manipulating of cartographic feature data. There do exist a few research tools (geoprocessing systems) such as POLYVRT, SYMAP, and the U.S. Census DIME files, but these systems, as efficient as they are, are limited-purpose data models for MC&G data. If one looks at how the DMA user uses the data, there is indication of what kinds of data manipulation are required. This is where the problem lies -- that is, MC&G data represents spatial information. Manipulation of spatial data is complex. Spatial functions include mathematical coordinate conversions and interpolation, transformations, distance, and direction calculations, area, shape, size, and orientation determination; adjacency, overlap, sameness, inclusion, and "betweenness" relationships among features. These spatial functions require analysis to determine algorithm derivation, data structure

definition, and possibly even identification of new and/or hybrid computer architectures. Truly, much research and development (R&D) is required in this area.

In this regard, DMA R&D is investigating data base systems, management systems, data base structures, data models, and spatial functions as they relate to MC&G applications and requirements. Since the commercial sector is concentrating on the general-purpose aspects of these technologies, DMA is aiming its research at the spatial and very large data base problems. The base line for some of these R&D experiments will be the Requirements Analysis for the DMA DBS.

DMA USER MODEL

A DMA user model will be one of the fallouts of the Requirements Analysis. This model must be one that is statistical, analytical, and qualitative in order to provide information revealing the varied types of systems users, load estimations by subsystem, function frequencies, products, types of user requests, and their frequencies for each product. The user model is then a forecast of future events and decisions at all levels of design.

SUMMARY

The magnitude of this effort and its highly interrelated tasks, beginning with the definition of requirements for an interactive data base system and user and data model definitions, followed by the logical and physical designs, systems architecture and management systems, detailed data base designs and implementation, required a phased development and implementation that will span at least one-half a decade. The conversion and transition from the current system will necessitate a considerable effort in software repro-gramming to achieve the required interactive capability for DMA users. The actual execution of the conversion effort is beyond the DMA staff management, and, therefore, the systems integrator approach was taken. Aside from the complexities associated with moving from a highly structured batch processing environment to a highly interactive data base environment is the natural resistance to alteration. This aspect was best expressed by N. Machiavelli's "The Prince" (1513): "There is nothing more difficult to carry out nor more doubtful of success, nor more dangerous to handle, than to initiate a new order of things. For the reformer has enemies in all who profit by the old order, and only lukewarm defenders in all those who would profit by the new order. This lukewarmness arises partly from fear of their adversaries, who have the law in their favour, and partly from the incredulity of mankind, who do not truly believe in anything new until they have had actual experience of it."