Lecture Notes in Computer Science

81

Data Base Techniques for Pictorial Applications

Florence, June 20–22, 1979

Edited by A. Blaser

Lecture Notes in Computer Science

Edited by G. Goos and J. Hartmanis

81

Data Base Techniques for Pictorial Applications

Florence, June 20–22, 1979

Edited by A. Blaser

Springer-Verlag
Berlin Heidelberg New York 1980

Herausgeber

A. Blaser
IBM Deutschland GmbH
Wissenschaftliches Zentrum Heidelberg
Tiergartenstr. 15
6900 Heidelberg

Sponsorship: IBM Europe
IBM Italy
Local Organization: Scientific Centers of IBM Italy

Programme Committee:

A. Blaser (Chairman), Heidelberg Scientific Center, IBM Germany
M. Felix, Paris Scientific Center, IBM France
F. Greco (Local Organizer), Pisa Scientific Center, IBM Italy
R. Hartwig, Heidelberg Scientific Center, IBM Germany
J. Jimenez, Madrid Scientific Center, IBM Spain
M. E. Senko, Heidelberg Scientific Center, IBM Germany
B. Smedley, Winchester Scientific Center, IBM United Kingdom

American Co-ordinator:
P. E. Mantey, IBM Research San Jose, CA

Asian Co-ordinator:
T. Takeda, Tokyo Scientific Center, IBM Japan

AMS Subject Classifications (1970): 68A45, 68A50
CR Subject Classifications (1974): 3.14, 3.63, 3.7

ISBN 3-540-09763-5 Springer-Verlag Berlin Heidelberg New York
ISBN 0-387-09763-5 Springer-Verlag New York Heidelberg Berlin

Library of Congress Cataloging in Publication Data. Main entry under title: Data base techniques for pictorial applications, Florence, June 20–22, 1979. (Lecture notes in computer science ; 81) Proceedings of a conference sponsored by IBM Europe and IBM Italy. Includes bibliographies and index. 1. Image processing–Congresses. 2. Optical data processing–Congresses. 3. Computer graphics–Congresses. 4. Cartography–Data processing–Congresses. I. Blaser, Albrecht. II. IBM Europe. III. IBM Italia. IV. Series.
TA1632.D36 621.3815'48 80-11561

Printed in Germany

Printing and binding: Beltz Offsetdruck, Hemsbach/Bergstr.
2145/3140-543210

WELCOME

The theme to be explored during this conference is that of pictorial data, in itself a difficult matter to define. Very likely, there are many conflicting definitions. The conference memorandum says, that pictorial data is that which relates to the shape of an object or its location in physical space. In planning the conference, the organizing committee was only too aware of the multitude of applications which encompass data meeting such a definition. It is obvious that these applications are very important in many areas of today's world. It was suggested that this conference works towards a first understanding of the various pictorial application areas and their potential and then tries to identify the underlying commonalities, especially as far as data is concerned, its representation, structuring, and storage.

Why are the Scientific Centers of IBM Europe interested in that subject and why are they sponsoring a conference on pictorial data? An outline of their role might provide the basis of an answer to these questions.

As technology advances and as scientific achievements become greater and technical development more rapid, we feel that the establishment of Scientific Centers is increasingly important in order to conduct long and medium range research, and to develop applications which meet the needs of modern society. The projects undertaken by IBM's Scientific Centers represent a wide range of topics in computer applications research, in advanced studies in system/user interfaces as well as system design and programming, and in many problems of information science encountered by computer users. Among these topics there has been work in pictorial applications and data base issues for many years. Some of the relevant work of the Scientific Centers might be mentioned:

- The Scientific Center at Heidelberg has worked in the past on the processing of scintigraphic images and is now dealing with man-machine interaction and with related information management.

- The Italian Scientific Centers are engaged in a wide range of problems in areas such as natural sciences, econometrics, and computer networking.

- The Scientific Center in Haifa deals with the design of computer systems for agriculture, for medical research, and for aquifer management.

- The newly established Scientific Center in Winchester is initiating work on image processing in medicine and on social sciences.

- The Madrid Scientific Center works in image processing of remotely sensed data, e.g. for the identification and assessment of earth resources.

- The Scientific Center in Paris is also engaged in this field as well as in speech signal processing with the objective to teach deaf children through visual feedback how to speak.

The result of this work is published in the scientific community. Very often, research is done in collaboration with academic and scientific institutions. Examples are the studies with the University of Pisa, the Italian National Council of Research, the University of Vienna, the Institute of Molecular Biology in Madrid, and the Universities in Berlin and Heidelberg. We consider these joint studies an ideal vehicle for the exchange of expertise and for cross fertilization.

This conference, too, is directed towards an open exchange of knowledge and results of research. It is with such an open exchange of expertise that the growth of new ideas is encouraged and science thrives. And it is in this spirit, that I would like to express my welcome and my gratitude to the participants for coming and sharing with us today's knowledge about data base techniques for pictorial applications.

I want to thank IBM Italy for the offer to host this conference in wonderful Florence and for the organizational and administrative effort undertaken to make it a pleasant and beneficial experience for all the participants.

Paris, June 1979

R. Aguilar
IBM Europe
Director of Scientific Centers

INTRODUCTION

Over the years, technological trends in hardware and software have significantly improved the processing power of computers, their primary and secondary storage capacity and accessing speed, their telecommunications facilities, as well as their user interfaces via general and special purpose terminals. This has facilitated, among others, the advancements of integrated data bases and of their administration on the one hand, and of graphical and image processing (in brief: pictorial) applications on the other.

Developments in these two fields have traditionally been unrelated. Integrated data bases have been and are still being nearly exclusively used for commercial and administrative applications of batch or transactional type. Conversely, pictorial applications have been pursued mainly in technical, scientific, and planning disciplines as diversified as architectural and engineering design, biochemistry, air traffic control, robotics, utility and geographical mapping, urban and regional planning, meteorology, medicine, and in the analysis of remotely sensed data e.g. for earth resources and agricultural inventory development and environmental protection.

There are, however, several strong arguments for an integration of data base techniques with pictorial applications. To mention just a few of them:

(1) The processing of pictorial data poses quite severe computational problems and the volumes of data to be manipulated and analyzed grow larger and larger. Therefore, much attention must be given to choosing between the various methods known for structuring and retrieving this data.

(2) Pictorial (as conventional) information is an expensive resource which should profitably be used for as many applications at as many places as possible. As an example, the same remotely sensed data can and should be used e.g. for earth resources identification, environmental protection, agriculture, and meteorology. This calls for an organization of this data on which to build a variety of different applications.

(3) Graphical and image applications enter into commercial and administrative environments. Business (or data presentation) graphics and facsimile are just 2 subjects to be mentioned. This is to some extent due to the fact that advanced problems in business and administration call for similar modes of operation (e.g. interactive problem solving) as known in technical and scientific disciplines, where graphical presentation of information is an indispensable tool.

(4) Many applications require pictorial data in combination with conventional data (computer aided design can be

mentioned as an example). This calls for the management of both types of information in the same system.

IBM Scientific Centers have traditionally been active in pictorial applications research (e.g. in engineering, medicine, remote sensing applications), as well as in data base research (e.g. in relational data base management systems and interactive enduser interfaces). Recognizing the needs mentioned above, some projects began to investigate scientific problem solving on the basis of prototyped data base management systems accommodating pictorial information.

To assess the work which has been going on for several years in application and data base fields and to discuss data base features necessary or desirable for pictorial applications, the Scientific Centers of IBM Europe organized this international conference. More precisely, its objectives are to seek an understanding of the various application areas which involve the use of pictorial data, especially to consider their requirements for the structuring, storage, and analysis of such data, and to identify commonalities and differences as well as to learn of relevant data base research. The emphasis is placed on the interaction between data base and application experts to exchange their views on the data base needs of, or the facilities for, pictorial applications.

To meet these objectives and to create a basis for mutually beneficial discussion, application oriented contributions were invited to particularly address the needs of their respective application for data regarding its volume, structure, storage, search (the algorithms used and the types of searches), and the extent to which data access can be pre-defined or needs to be dynamic and flexible. Data base contributions were encouraged to emphasize those features of their work which are relevant to pictorial data and applications.

The agenda has been structured according to our assessment of the conference subject and of the expected commonalities and differences in data related aspects of the applications. The first two sessions deal with geographic applications, e.g. in geographic and utility mapping and in urban and regional planning. The third session addresses the administration and accessing of data representing two- or three-dimensional geometric objects as they occur for instance in architecture, engineering, and biochemistry. The fourth session is devoted to remote sensing and image processing applications, among others in water resources management, agriculture, astronomy, meteorology, and biomedicine. The fifth session, eventually, covers some related data base research.

This is one of the first conferences totally devoted to the subject. It is supposed to draw its success from pointing the applications oriented audience to the common problem of organizing and accessing pictorial information and to the solutions already visible, and in making the data base researchers aware of unresolved problems deserving their attention. There will be conferences on that subject in the years to come. We would be pleased if this one could play a kind of pioneering role for the others.

The conference was instigated by the Chief Scientist of the IBM Corporation. It was sponsored by IBM Europe and IBM Italy, and locally organized by the Scientific Centers of IBM Italy. Nevertheless, the contents of this introduction and of the conference contributions express the authors' own personal opinions and not IBM's.

I would like to express my gratitude and appreciation to all the lecturers and session chairmen, to the sponsors, to hosting IBM Italy, to the members of the programme committee, to the local organizers, and to the many contributors within and outside of IBM, who gave advice and assistance in preparing, organizing, and running this conference.

Heidelberg, June 1979 A. Blaser

TABLE OF CONTENTS:

DATA BASE REQUIREMENTS FOR GEOGRAPHICAL MAPPING

by

RICHARD L. VITEK

The rapid and significant improvements in computer technology have caused a virtual explosion in the use of digital processing and digital products within the mapping community. The availability of large volumes of mapping data in digital form presents both an opportunity and a challenge. The opportunity to provide more cost effective and responsive systems and the challenge to do it well and as soon as possible. This paper defines a concept of operation which uses digital source data bases to support the production of multiple products; an approach and the status in establishing data base requirements in terms of content, structure and format; and the data collection and software problems that inhibit a simple solution. The use of this source data by other national as well as international users is also addressed in terms of the need for flexible formats and coding structures.

Defense Mapping Agency
Building 56, U.S. Naval Observatory
Washington, D.C. 20305

Introduction

The primary mission of the Defense Mapping Agency (DMA) is to provide map, chart and geodetic data to the U.S. Military Services. The primary purpose is to provide the user with geographic information on a global basis. This information has historically been presented in the form of a map or chart, that is, an abstraction of the real world in graphic form. The increasing use of the computer throughout the military is showing up as a demand for products in digital form. The content remains essentially the same as the graphic product but the form and format are changing. We now distribute digital maps and terrain elevation matrices in addition to the standard map and chart products. We are also turning more and more to the computer and automation to trim costs for our in-house production which is still very labor intensive. Digitizing is used in many cases as part of the production process even though the final product is a graphic. Over 50 percent of our manpower is devoted to digital products and digital processing.

We see the large volume of digital data as presenting both an opportunity and a challenge; an opportunity to be more cost effective and responsive, and a challenge to do the job well using data base techniques and state-of-the-art technology. Since the cost of collecting digital data is high, we must insure that the data which is collected is utilized to the maximum extent possible through efficient exchange procedures. The ideal situation is to collect data once and use it for as many products as possible. This idea leads one to consider the building of multiproduct data bases. By collecting and maintaining the data in digital form, we can, of course, provide quicker response. Besides the data that DMA collects there are also other sources of digital data. DMA is only one organization in a community of national and international producers of geographic data and products, all with very similar digital problems. Therefore we are working with the other producers and collectively developing exchange standards to take full advantage of all the digital data that will be collected.

A major consideration in the development of data base requirements is to provide a viable interface between the current and future data bases and files which will allow for phased, well-controlled change. Such an interface will provide for maximum use of new techniques and technology with minimum impact on current production. The data base requirements for DMA are not for a single data base but rather a set of management and production data bases, storage and exchange formats, data elements, and the associated environment which shapes their development and specifications. The exchange format plays a key role in providing the desired interface. This paper covers a concept of operation of that set of data bases to support multiproduct applications and an approach to establish the data base requirements in terms of the content, format and structure. There are collection and software problems which will be discussed later that inhibit a simple near-term solution; however, we feel a well-planned phased approach will allow us to meet the challenge of the future and take full advantage of the opportunities associated with the availability of digital data.

Concept and Model

The basic concept is to provide a unified set of management and production data bases which will comprise the DMA data bank. The unified set of data bases will consist of data in graphic, analog and digital form. A blend of the old and the new, existing simultaneously, which may be physically separated but will be centrally managed and logically integrated. The logical integration would insure minimum overlap or gaps in the total set of data bases by building interfaces and/or physical integration. Physical integration will only be attempted where the integration can be shown to be cost effective and still meet or exceed all current user requirements.

In the idealized model of a data bank (see Figure 1) all data bases would contain evaluated and organized data. All data would be edited and evaluated on the way into the data bank and all products would be generated from data accessed from the data bank. The intent is to gradually, over time, develop such an idealized data bank which would consist of analog and digital production data and the required associated management indices and algorithms. New indices and algorithms would be developed as components of on-line management data base systems to allow a rapid determination of what source material is on hand, how good it is, where it is located, and the products that can be made from it. The data would then be retrieved and processed with automated equipment to furnish completed products to the user. The ideal model would contain standard procedures, data elements, data bases and data base systems. The standardization would evolve through development and procurement of new systems rather than retrofitting into existing systems. New systems would be developed and tested in parallel with the older systems to insure proper operation. After the new system is demonstrated, it would be placed in operation and the old system phased out with minimum disruption of the production process.

Standardization

Additional cost savings can be achieved by minimizing data and system redundancy and their associated high maintenance costs. This approach involves developing specifications to define existing data and procedures as well as developing standards where data and procedures are essentially the same or close enough to be made the same. Standardization then becomes a means to achieve cost reduction.

In the standardization process there are many elements to be considered from total systems to specific data elements. In analyzing systems, one may consider software, hardware, procedures and the individual data bases. In the storage and exchange of data one must consider the determination of specific media, forms, format or structure, and content of files. As the content of particular data bases is explored in detail, then one must also determine the data elements, data sets, area or cell sizes, resolution, feature types and reference systems. In any specific system or data base one must determine which of these many elements should be standardized and to what degree. The elements and the degree of standardization as they relate to specific management and production data bases will be discussed in the following paragraphs.

Types of Data Bases

It is helpful to define the various types of data bases to aid in a discussion of current and future operations. The two broad categories of data bases are management and production, these are shown schematically in Figure 2. The management data base contains facts about requirements, resources and production status. This type of data is frequently described as metadata or data about data. Typical management data bases in DMA are contained in the DMA Management Information System (DMIS). The DMIS includes a family of data bases such as, Production (DMIS/P), Equipment (DMIS/E), Area Requirements and Production Status (ARAPS), and Product Maintenance System (PMS) to name a few. Indexes, automated and manual, are also considered a type of management data base. There are many automated indices for maps, charts, photography, names, and other items used in the production process.

The production data bases contain the physical data which is used as input to (source) or results from (product) the production process. Source data bases are either as collected (raw) data or preprocessed to bring the data up to a specified level of readiness for further processing. The intent is that the processed digital source data should be stored as a single class, used for the detection and recording of new or changed data, and designed to service multiple products. The product data base is a synthesized set of data (features) from several classes of source data used to support the production of a specific product or family of products. DMA currently produces terrain elevation data in a standard format for multiple applications, this data is considered as a source data base. DMA also stores data such as reproduction material which is ready for production, as well as copies of digital products such as the Digital Land Mass Simulator (DLMS) data. The graphic reproduction material and the DLMS data are considered product data bases. The data has been collected into specific sets for a specific product. Hard copy graphics are also stored as an inventory of available products and listed in our military map catalog.

Data Base Systems and Functions

A brief description of a typical data base system is needed to aid in subsequent discussions. A data base system is defined as that total set of procedures, software and hardware for the utilization of a predefined set of data. The components may be manual or automated. A functional flow of a typical data base system is shown in Figure 3. The data bases are shown separated from the functions to be performed. It is important to look at this separation of functions and data to obtain a better perspective on standardization and integration of data bases and software. For example, each data base must be considered as part of a system to insure that all input and output requirements and functions of the system are defined. However the overall set of data bases must be reviewed as a group to determine redundancies and to plan for integration of similar sets of data.

The input functions of collection, editing, and validation generate a qualified transaction for entry into the data base. The storage and retrieval functions allow for update, query, and products. The production of products is through the output application programs. Products are predefined data sets for one or more users on a regular basis. The queries are for ad hoc or non-predefined data sets. The schema is the logical structure of the data which is different than the internal physical structure of the data base. The user view of the data base is through its products and schema.

The distinction between the logical and physical views of the data is important. It is this separation which allows the internal storage structure of the data to be independent of the applications using the data. Reorganization, consolidation or integration of the internal data is transparent to the user. The schema or logical view of the data can also be applied to magnetic tape systems and is discussed in more detail in the following sections.

Exchange Format Compared to Schema

Exchange formats for digital data on magnetic tape are very common. In fact, there are national standards for labels and codes to be used in exchange formats. Discussion of schemas to be used in data bases are also very common in the literature. However a comparison of the characteristics of schemas and exchange formats is not very common, but it does lead to some very interesting conclusions.

The schema, i.e., the logical structure of the data shows the data elements and their relation to one another. These relationships are generally classified as hierarchical, network and relational. The schema shows the data elements in sets, groups, or records; the different groups that exist; and whether or not they will occur more than once as repeating groups (see Figure 4). The schema itself does not contain any values, the data values are in the data base.

Magnetic tape exchange formats are very similar to the schema. The tape format is the order in which data is to be placed on the tape. The sequential nature is not important in this comparison. The format does not limit the quantity or values of the data but rather only the logical order in which the data is to be placed (see Figure 5). The tape format includes record, file and tape levels. These levels would show the order of data elements within the record, the order of records in a file, and the order of files within the tape. The exchange format can be thought of as the external users' view of the data set (see Figure 6) and satisfies the same query function for non-predefined data sets. The exchange format also provides the same isolation as the schema and allows the physical internal storage to be independent of and transparent to the users. When considered in this way the exchange format can then become the key link between departments within the same organization, between agencies, and between older files and new data base management systems.

Exchange Versus Storage Formats

A look at exchange formats and storage formats provides a different perspective but leads to a similar conclusion about the value of the exchange format. The following example also shows a case for the correct degree of standardization. Assume a group of six users whose primary concern is to be able to exchange data (see Figure 7). If there is no standard format and they each wish to exchange with all other producers, then each one of the six users will have to develop five conversion programs for a total of 30 programs. If any one user changes his internal format, five others are affected. If on the other hand they can agree on and develop a standard exchange format, then each user only has to develop two conversion programs: one to and one from the standard exchange format for a total of 12 programs. If any user changes his internal storage format, the change is transparent to other users and no one else has to change their conversion programs. One could go one step further and try to use the exchange format for internal storage. This would eliminate reformatting the data for exchange purposes but does not tell the whole story. Other internal operations for update and production may require reformatting because of the use of the exchange format. If this is the case then one must select the format which causes the least amount of reformatting. The conclusion to be drawn from the analysis is that if a group of users wishes to exchange data, they should develop a standard exchange format but leave its use as an internal standard a free choice of the individual users. The exchange standard should facilitate the maximum utilization of the total set of data without constraining the internal operations of any of the producers, and the storage standard should be oriented toward optimizing internal operations.

Exchange and Storage Specifications

Both the exchange and the storage specifications will contain a format, data elements, data items or values for the data elements and codes for the data items. A brief definition of these terms follows. An example of a data element is "month of the year." The data items are the values, that is, the names of the months, January through December. The data items can also have abbreviations or codes to minimize the number of characters needed for storage of the data item. For example one could use the numbers 1 to 12 to represent the months. For the more complex data elements such as geographic (map) features there are a large number of data items, and the codes and coding system for the codes then become very important. The coding system may be used to define categories, types and so on. The storage specifications should also define standard data sets to be stored.

The exchange format must be very flexible to accept a wide range of data types and values because it must meet the needs of the whole community of users. For geographic data the format should allow for variable area or cell size, resolution, and reference systems. The data elements, data items and codes must be a standard set published in a catalog or dictionary for access by all users. The coding system used to establish the specific codes must be comprehensive, expandable, and the categories and types of data must be mutually exclusive. The exchange specifications must be able to handle the data for all current users and be expandable for additional data types from current and new users.

The storage specifications should be oriented to internal needs. The format should be designed to minimize the reformatting required for update, production and exchange operations. The coding structure can be less complex than the one for exchange codes because the internal set of features is only a subset of the total exchange set. The only requirement on internal coding structure is that it should allow for some expansion to collect additional data elements as changes or needs arise. The internal codes must also have a one-to-one unambiguous relation to the exchange set. This can normally be handled by generating the proper code in the exchange set and is not a constraint on the internal codes.

Standard Data Elements

The first step in developing an exchange specification is to develop a set of standard data elements. Developing such a set of standard data elements (DE) for geographic (feature) data is a formidable task. Each DE has its own name, abbreviation, data items, data item codes, and most important a unique and unambiguous definition. The definition is essential so that all users not only know what data is available but also so they can identify synonyms. To establish a standard set of data elements one must first define the set of DE for each user (see Figure 8). The initial set of users would be internal to the agency but would be expanded to include external users who wished to participate. By using the definitions for each user's subset one must then identify common elements. These are synonyms; that is, elements which are meant to be the same but may have different names or formats. The next step is to determine if a standard DE already exists in a Federal or National Standard Set and if not, then one must be established. The task of identifying synonyms and standards is aided greatly by using a DE Dictionary. Many DE dictionaries, software packages, exist which provide for on-line storage and query operations. These dictionaries facilitate analysis and selection of synonyms and the subsequent printing of the synonyms and selected standard DE's. The resources to identify the user's subset of data elements as well as the final exchange set can be significant. However it should be recognized that the cost of identifying the DE is small in comparison to the data collection cost and the benefits of developing standard DE's will far outweigh the costs.

The above process is further complicated by the fact that many geographic features have attributes that must also be identified. The attributes are also