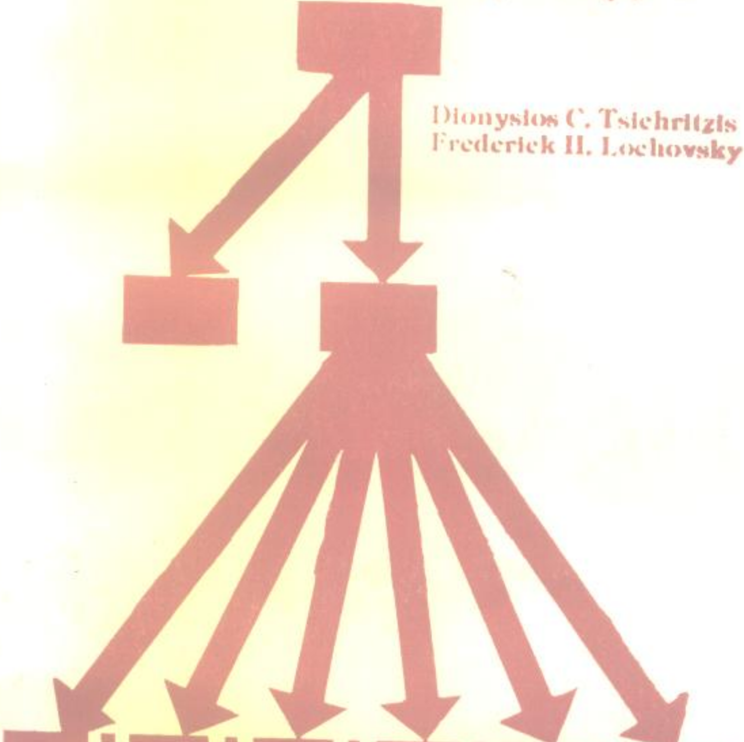
DATA MODELS



Dionysios C. Tsichritzis
Frederick H. Lochovsky

Data Models

Dionysios C. Tsichritzis
Frederick H. Lochovsky

Department of Computer Science
University of Toronto
Toronto, Canada

Prentice-Hall, Inc.
Englewood Cliffs, New Jersey 07632

Library of Congress Cataloging in Publication Data

TSICHRITZIS, DIONYSIOS C.
Data models.

(Prentice-Hall software series)
Bibliography: p. 343
Includes index.
1. Data base management. 2. Data structures (Computer science) I. Lochovsky, Frederick H. II. Title.
III. Series.
QA76.9.D3T75 001.64'2 81-8611
ISBN 0-13-196428-3 AACR2

Prentice-Hall Software Series
Brian W. Kernighan, advisor

This book was typeset by the authors, using a Graphic Systems phototypesetter driven by a PDP-11/70 running under the UNIX operating system.

UNIX is a Trademark of Bell Laboratories.

Editorial production/supervision
by **Aliza Greenblatt**

Cover design
by **George Alon Jaediker**

Manufacturing buyer
Gordon Osbourne

Printed in the United States of America

10 9 8 7 6 5 4 3 2

PRENTICE-HALL INTERNATIONAL, INC., *London*
PRENTICE-HALL OF AUSTRALIA PTY. LIMITED, *Sydney*
PRENTICE-HALL OF CANADA, LTD., *Toronto*
PRENTICE-HALL OF INDIA PRIVATE LIMITED, *New Delhi*
PRENTICE-HALL OF JAPAN, INC., *Tokyo*
PRENTICE-HALL OF SOUTHEAST ASIA PTE. LTD., *Singapore*
WHITEHALL BOOKS LIMITED, *Wellington, New Zealand*

PREFACE

The pursuit of knowledge is part of human nature. We always try to achieve some understanding of ourselves and our environment. Scientific endeavor has to do with acquiring some form of knowledge. A person may perceive or observe a phenomenon. As a result he or she acquires an incremental piece of knowledge. The increments of knowledge we will call *information* [Langefors, 1977]. The information may be valuable and may have to be recorded for purposes of communication to other people. The proper representation and communication of the information will benefit other people and will increase our collective knowledge. The representation of information is, therefore, a very important problem.

The representation of information is accomplished primarily by the use of natural language. People write papers, reports, books, etc. to communicate their ideas. Natural language, however, is not always the best tool for representing information. First, it is a general and all-encompassing tool. In certain cases, a specialized tool may be more appropriate. The use of scientific notation points to the fact that people have perceived the need for other specialized models for representing information. These models may be widely different and more appropriate for their use than natural language (e.g., maps for geographical information). Second, natural language evolved for free-style communication between human beings. As such, it did not have to be and is not a precise means for recording and transmitting information. Finally, natural language has properties that make it notoriously difficult to manipulate through computer operations.

In addition to solving complex computational problems, computers also are used to aid human intelligence by providing a tool for handling and manipulating information. This will probably continue to be the most important use of computers in the future. To utilize computers effectively for this purpose, we need to explore good ways of representing information in a manner that is amenable to computerization. As already pointed out, natural language may not be the appropriate vehicle for this role. Data models, as presented in this book, have been devised for computer-oriented representation of information. They are powerful conceptual tools for the organization and representation of information. In addition, they are translatable into structures that can be manipulated by computers. Data models provide a means for the representation and manipulation of information in a way that is amenable to computerization.

If we expect data models to fulfill such an illustrious role it is imperative that they be flexible and well understood. They should be flexible to achieve the different goals and satisfy the different tastes of their users. They should be well understood so that their properties can be rigorously defined. A general framework of data models is needed which encompasses all the different data models that have been proposed. Such a framework does not necessarily point to or propose a supermodel. It should, however, clearly identify the concepts and properties of and the similarities and differences between data models. This book is an attempt at developing such a framework, or at least recording what we know up to now in this rapidly developing area. We hope that the ideas presented will help people to organize data conceptually and to draw upon and use to advantage the information the data represent.

Social changes have created large and complex political and economic structures (e.g., large government agencies, multinational corporations, and chains of retail stores). These enterprises need to integrate the use of their data for planning and control. Data base management systems are an appropriate tool for this integration. The main issue is the effective use of the data rather than the particular system that is used for their storage. For data base technology to be introduced and used effectively, there is a need for understanding and visualizing the data and the information they represent. Data models help in achieving some understanding of the data and information needs of an organization and by extension the way in which an organization functions. They can also be used to describe any insights for both communication purposes and later use. It seems, therefore, that data modeling is an important activity for information integration regardless of the use of a data base management system. Data modeling is needed even if we use file systems.

There is a widespread misconception that data bases and data base management systems correspond to a shallow area dealing mainly with the storage and retrieval of data. Although we acknowledge that the area started out dealing with pragmatic problems, it is increasingly becoming involved with ideas that deal with the structure, organization, and effective use of data and the information they represent. This book does not deal with implementation of data models or with any properties of access paths or other access-oriented considerations. Some data models may have originated as a natural abstraction of working data base management systems. We discuss them conceptually, however, and treat only their logical properties. This is the conceptual and abstract part of data bases as opposed to the engineering part of their implementation. The richness of the logical properties should persuade readers that the area is very important, not merely fashionable.

The book is divided into four parts. The first part, consisting of the first four chapters, outlines the basic concepts used by all data models. In Chapter 1 we discuss abstractly the meaning of data and their role in an organization. The role of a data model is outlined as a conceptual tool to visualize and structure data. Data models are discussed in terms of three distinct parts: structures, constraints, and operations. Chapter 2 outlines a general framework for structuring data. Chapter 3 discusses constraints on data which provide additional semantics on the data structure. Finally, Chapter 4 discusses operations on data and relates data models to abstract data types.

The second part of the book describes the three most widely known data models in terms of the data modeling framework developed in Part 1: structures, constraints, and operations. Chapter 5 discusses relational data models, Chapter 6 network data models, and Chapter 7 hierarchical data models.

The third part of the book outlines four additional data models which, in our opinion, represent different approaches to data modeling. The order of presentation of these four data models corresponds approximately to their complexity. We start with simpler, more concrete data models and proceed to more abstract, complex, and semantically powerful data models. Entity-relationship data models are discussed in Chapter 8 as an example of simple data models which facilitates communication among users and data base designers. Binary data models are discussed in Chapter 9 as an example of data models starting out with a simple graph structure and introducing powerful operations to make them capable of modeling complex situations. Chapter 10 outlines semantic network data models as a modeling approach with a rich semantic component that can be used to move from data bases to knowledge base systems. Finally, Chapter 11 discusses infological data models which try to view information requirements in a manner most natural for people before they are mapped into data base requirements. There are many other worthwhile data models in the literature. We chose these four types not as the "best," but as representative of different approaches to data modeling. Lack of space prevents us from elaborating on other proposed models.

The last part of the book deals with the use of data models for data base design and operation. In Chapter 12 we outline the steps taken in schema design. The emphasis is mainly on capturing the application requirements in an accurate schema. Chapter 13 discusses data base theory which attempts to formalize the notion of a "good" schema. In this way a schema can be analyzed and a "better" schema can be produced. Chapter 14 concludes the book by outlining the correspondences between data models. This discussion includes schema mappings, operation mappings, and data base translation. These problems are very important to an understanding of the similarities and differences between data models. They are also relevant to important pragmatic considerations such as data base translation and data base cooperation.

This book was heavily influenced by the research conducted in the data base group at the University of Toronto over the last six years [Bernstein, 1975; Brodie, 1978; Klug, 1978; Lochovsky, 1978; Vassiliou, 1980a]. We have obviously benefited from many outside sources and had highly stimulating discussions with colleagues from many research groups. However, we have also developed our own philosophy about what data base management is and what the important problems are. In this book we try to look at data models from this perspective.

The area of data base management has attracted many researchers and ideas from widely different parts of computer science and management studies. For example, many ideas as presented in this book are drawn from other areas of computer science (e.g., artificial intelligence, operating systems, and programming languages). We hope that this intrusion into other areas and the borrowing of ideas will continue. It is not critical whether an idea started out in a different area. It is important only that it becomes useful and important in a data base context. We hope that the data base area will continue to be outward looking and serve as an umbrella for widely different ideas and people. Computer science drew from mathematics and electrical engineering. Data base management draws from computer science and management studies. Since the application of computers for management functions is one of their most important uses, we expect data base management to continue to be a focal point of much activity in the future.

The field of data base management has both practitioners who desperately need solutions and academics who would like to work on relevant problems. This book is addressed to both these widely different kinds of readers. People with practical knowledge in DBMSs may find it interesting to understand some of the more esoteric ideas of the subject. They may get some insights which may translate to useful practical techniques. In addition, they will be able to see how all the ideas in different systems fit together in a uniform conceptual framework. People with more theoretical backgrounds and little practical experience in DBMSs may also find this book useful. They hopefully will understand the different data base ideas abstractly without being hindered by all the ad hoc terminology of existing systems. Throughout the book there are many exercises, some of which can be used as a springboard for further research.

We hope that this book will provide a good framework and many ideas for research directions. Many people are becoming interested in data base problems. Their ideas are needed in the area of data models. In this way a theory can be developed. Such a theory is not only a theory of data bases. It should be the beginning of a theory of data, information, and knowledge as it can be operated on by computers. The problems associated with the development of this theory will be with us for some time even after we solve all the problems of fast data access through new hardware or software techniques.

ACKNOWLEDGMENTS

This book could not have been written without the help, encouragement, and support of many people. Foremost among these are the members of the Data Base Group at the University of Toronto, both past and present, who by their ideas, comments, and discussions contributed greatly to the nature and content of this book. The reader will discover that the research of several former members of this group — P. A. Bernstein, M. L. Brodie, A. Klug, and Y. Vassiliou — forms the basis for several of the chapters.

We would like particularly to thank Prof. A. T. Borgida, Prof. M. L. Brodie, Prof. A. Klug, Prof. Y. Vassiliou, M. H. Graham, and L. Melli who provided us with both encouragement and many useful ideas and comments. Their criticisms and suggestions were invaluable in shaping the contents of this book.

We would like to thank I. S. Weber and R. M. Reid of the Computer Systems Research Group at the University of Toronto for typing parts of the initial draft of this book.

Facilities for preparing this book were provided by the Computer Systems Research Group and the Computing Center of the University of Toronto. Most of the research of the Data Base Group at the University of Toronto that is described in this book was funded by grants provided by the Natural Sciences and Engineering Research Council of Canada.

CONTENTS

PART 1 BASIC CONCEPTS

PART 2 DATA MODELS I

PART 3 DATA MODELS II

PART 4 USING DATA MODELS

Part 1

BASIC CONCEPTS

For data to be useful in providing information, they need to be organized so that they can be processed effectively. Many different ways of organizing data exist, such as tables, lists, and forms. In data modeling we try to organize data so that they represent as closely as possible the real-world situation, yet are still amenable to representation by computers. These two requirements are often conflicting. To determine how best to organize data for a given application, we need to understand the characteristics of data that are important for capturing the essence of their meaning. These characteristics allow us to make general statements about how data are organized and processed. A consistent, formal set of such statements defines a data model. In Part 1 we examine those characteristics of data that comprise the definition of a data model. We also examine ways in which these characteristics can be represented so that they are amenable to computerization.

Chapter 1

DATA AND DATA MODELS

1.1 THE MEANING OF DATA

A perception of the world can be regarded as a series of distinct although sometimes related phenomena. From the dawn of time human beings have shown a natural inclination to try to describe these phenomena in some fashion whether they understand them completely or not. These descriptions of phenomena will be called *data.* Data correspond to discrete, recorded facts about phenomena from which we gain information about the world. *Information* is an increment of knowledge that can be inferred from data [Langefors, 1977].

The word "datum" comes from Latin and, literally interpreted, means a fact. Data, however, do not always correspond to concrete or actual facts. Sometimes, they are imprecise or they describe things that have never happened (e.g., an idea). For our purposes, data correspond to descriptions of any phenomenon or idea that a person considered worth formulating and recording. Data will be of interest to us if they are worth not only thinking ábout, but also worth recording in a somewhat precise manner.

Data are traditionally recorded using a particular communication method (e.g., pictures or language) on a particular (semi-)permanent recording medium (e.g., stone or paper). Examples of the human perpensity for recording data can be found throughout time: cave paintings of prehistoric man, ancient Greek on stone, and Egyptian on papyrus. Often, data are recorded on paper using a natural language. Usually, both the data (i.e., the facts) and their interpretation (i.e., their meaning) are recorded together, since natural languages are sufficiently flexible to do both. For example, the statement "His height is 173 cm" records both the datum "173" and its meaning "height in centimeters." In certain cases data are separated from their interpretation. For instance, an airline schedule is a table of data. Its interpretation is usually given separately at the beginning of the schedule and people are expected to know how to interpret it. However, separating data and their interpretation can lead to difficulty in using the data. Studies have shown that most people have difficulty interpreting an airline schedule. Without knowledge of its interpretation, the airline schedule is of limited use.

The use of computers for the encoding and processing of data has resulted in even more separation of the data from its interpretation. Computers deal mainly with raw data. Much of the interpretation of the data is not explicitly recorded. Consider, for instance, a numerical analysis program that solves partial differential equations. The package receives some numbers as input and produces some other numbers as output. It does not care whether the differential equation is an application in fluid mechanics or electromagnetism. The user of the package has to interpret the results in the context of the use being made of them.

There are at least two historical reasons for the separation of data and their interpretation in computers. First, computers are not very good at handling natural language, which is still the main way of encoding interpretations and meaning of data. Second, computer storage was initially rather expensive. Although there was enough storage for the actual data, their interpretation was traditionally left to the users and the manual systems outside the computer.

As the application of computers evolved, it became increasingly necessary to try to capture some of the interpretation of the data. For instance, an inventory control package interprets the data as parts and suppliers. An airline reservation package views the data as seats and flights. By interpreting the data to some extent, these systems can be useful to an inventory clerk or an airline clerk. There is no implication that the system has accurate and complete knowledge of the application. However, some understanding of the meaning of the data is present in the way that they are manipulated by the programs.

Suppose that the interpretation of data is encoded mainly in the programs that use the data. Thus, the programs are important since they interpret the world as it is portrayed by the data. The data are merely a collection of bits, on some storage device, which do not make sense unless they are first processed by a program. This approach is analogous to saying that the interpretation of the airline schedule is important, not the airline schedule itself.

In an environment where data can be shared and used by many different applications, such an approach can be followed only to a certain point. After a while, it becomes rather cumbersome to write different programs continually and provide them with similar, if not identical,