DATABASE MANAGEMENT IN SCIENCE AND TECHNOLOGY

A CODATA Sourcebook on the Use of Computers in Data Activities

Sponsored by CODATA

John R. Rumble, Jr.
Viktor E. Hampel
editors

NORTH-HOLLAND

DATABASE MANAGEMENT IN SCIENCE AND TECHNOLOGY

A CODATA Sourcebook on the Use of Computers in Data Activities

Edited by

JOHN R. RUMBLE, Jr.
National Bureau of Standards, Washington, DC, U.S.A.

and

VIKTOR E. HAMPEL
Lawrence Livermore National Laboratory, Livermore, CA, U.S.A.

Sponsored by
CODATA

1984

NORTH-HOLLAND
AMSTERDAM · NEW YORK · OXFORD

ISBN: 0 444 86865 8

Published by:
Elsevier Science Publishers B.V. (North-Holland)
P.O. Box 1991
1000 BZ Amsterdam, The Netherlands

for:

Committee on Data for Science and Technology (CODATA)
51, Boulevard de Montmorency, 75016 Paris, France

Sole distributors for the U.S.A. and Canada:
Elsevier Science Publishing Company, Inc.
52 Vanderbilt Avenue
New York, N.Y. 10017
U.S.A.

Library of Congress Cataloging in Publication Data

Main entry under title:

Database management in science and technology.

Includes bibliographical references.
1. Data base management. 2. Information storage and retrieval systems--Science. 3. Information storage and retrieval systems--Technology. I. Rumble, John R. II. Hampel, Viktor E. III. CODATA.
QA76.9.D3D3586 1984 001.64 84-1490
ISBN 0-444-86865-8 (U.S.)

Printed in the Netherlands

DATABASE MANAGEMENT IN SCIENCE AND TECHNOLOGY

FOREWORD

Data in science and technology can be defined as information, usually numeric, that has been derived from some measurement, observation, or calculation. A database is an organized collection of such information on a well-defined topic. In the last 30 years, computers have changed the distribution of databases from the printed to the electronic media.

This book is concerned with these revolutionary changes and the distribution of collections of data by computers. The revolution has been a positive one, which appears to add to our scientific and technical capabilities. Ideally, computer-readable databases can be combined, sectioned, manipulated, and displayed easily and quickly in a wide variety of ways. In reality, it isn't that easy.

In this book, we, the editors and authors, will try to give scientists and engineers access to tools for using computers for databases. The emphasis is first on what must be done, then on how to find solutions. We concentrate not on specific solutions, but on the methodology of being successful.

Database management means using computers to build, change, and use databases. Database management in science and technology is a combination of:

Computer Hardware
Computer Software
A Collection of Data
Scientific and Technical Expertise

First and foremost, it must be recognized that this effort is a computer project, and the methodology of successful computer projects is indeed applicable to this subject. Second, the effort involves scientific data handling, an activity which has been discussed in some detail in a previous CODATA Sourcebook by Rossmassler and Watson. That book presents a thorough discussion of all aspects of the collection and evaluation of scientific and technical data and contains many useful references. Anyone serious about scientific data activity should consult this volume before starting.

The present volume concentrates on the use of computers to build scientific and technical databases. It is divided into three areas:

Basic Considerations - Chapters 1-4
Building the Database System - Chapters 5-8
Connecting the Database to Other Activities - Chapters 9-10

The basic underlying premise of the entire book is quite simple:

THINK THE PROJECT THROUGH FIRST

Thirty-plus years experience with computers has shown that successful projects are well-planned and well-thought-out, and database building is no exception. Many scientists and engineers have used computers routinely in their everyday technical work, performing calculations, making models, and gathering experimental data. However, most of this computer experience is only indirectly applicable to building numeric databases. In many cases, powerful computer software is already available and should be seriously considered because database programming is expensive and time-consuming. To make use of existing capabilities, the database developer must first be totally aware of the scope of the proposed project, the use of the data concerned, and the resources available. In addition, the developer must be aware of what has been done previously, how to find it, and what it can do. In our experience, many S&T data projects initially appear to have such specialized requirements that only a *home-built* system will do. However, close examination of the *true* requirements often shows that the project is not so unique as it first appears.

Let's make it clear; there will always be a need for new software. However, most S&T database projects are so resource-limited that care needs to be taken not to squander all the effort on the computer aspects of the project to the neglect of the scientific and engineering needs. The product of a S&T database project is better science and engineering, not the database itself.

So with these thoughts in mind, we invite you to share the collective experience of the authors in understanding, planning, and building scientific and technical databases.

John Rumble, Jr.
National Bureau of Standards
Washington, DC 20234

Viktor Hampel
Lawrence Livermore Laboratory
Livermore, CA 94550

Note: At points in this book, various commercial products are mentioned. No endorsement is intended either by the authors, editors or CODATA.

CODATA

The Committee on Data for Science and Technology (CODATA) was established by the International Council of Scientific Unions (ICSU) to promote the quality, reliability, and accessibility of data of importance to science and technology. CODATA works on an interdisciplinary basis through representatives of the international scientific unions and of member nations. Its activities include the preparation of key data sets and recommendations on data formats, development of better access to data sources, and general educational activities related to data handling. This Sourcebook on Database Management presents an overview of current techniques for data storage, retrieval, and dissemination which are applicable to scientific data in various disciplines.

W. W. Hutchison
President of CODATA

LIST OF CONTRIBUTORS

HELENE BESTOUGEFF is a Computer Scientist, presently Professor of Computer Science at University of Paris 7. Her research fields are databases and information systems, computational linguistics, and educational uses of computers. She had been associated with research at Stanford University, IBM, and Xerox PARC. Mailing address: University Paris 7, 2, place Jussieu 75251, Paris Cedex 05, France.

ALFRED A. BROOKS is a physical-organic chemist, presently Manager of the Information Systems organization of Union Carbide Corporation, Nuclear Division. His responsibilities have included the computer processing of research and development information, project and process information and fiscal information, as well as the computer solution of technical problems. His particular interest in data structures and information processing has been the unique requirements of R&D organizations as contrasted to the corporate requirements for fiscal and management data. He serves on the ANSI X3L5 Subcommittee and ISO TC97 SC15 WG3 Working Groups for the development of general-purpose interchange standards. Mailing address: Union Carbide Corporation, Nuclear Division, P.O. Box P, Oak Ridge, Tennessee 37830 U.S.A.

D. E. CULLEN is a nuclear physicist, presently with the Nuclear Data Section of the International Atomic Energy Agency, Vienna, Austria. His main research fields are in nuclear reactor design and associated databases. He has worked for over 20 years on computer-based information systems used in nuclear physics and engineering. Mailing address: Nuclear Data Section, IAEA, P.O. 100, A-1400 Vienna, Austria.

F. D. GAULT is a Senior Lecturer in Theoretical Physics at the University of Durham. His main field of research is elementary particle physics where he has published in several areas. He is also interested in economic modeling. Under his direction the Particle Data Group in the U.K. compiles all elementary particle scattering data and makes that and related bibliographic information easily available through computer-searchable data bases. He serves on various international committees including the Board of the Computational Physics Group of the European Physical Society and is a Fellow of the Institute of Physics. Mailing address: Particle Data Group, Department of Physics, University of Durham, Durham City, DH1 3LE, U.K.

CHARLES GOTTSCHALK is a physicist/engineer/information specialist, presently Chief of the Energy Information Program at the United Nations Educational, Scientific, and Cultural Organization (Unesco) in Paris, France. He is presently concerned with implementation of a worldwide network for energy information including numerical data, with emphasis on new and renewable sources of energy, especially in the developing countries. Mailing address: Unesco; SC/TER, F-75700 Paris, France.

VIKTOR HAMPEL is the project leader for the Technical Information System at Lawrence Livermore Laboratory. He has been very active in all aspects of database management systems and has especially concerned himself with making computerized scientific and technical data easily available to users throughout the world. Mailing address: Technical Information System, L275, Lawrence Livermore Laboratory, P. O. Box 808, Livermore, CA 94530, USA.

MICHAEL HUFFENBERGER is a research scientist in the Computer Center at Battelle Columbus Laboratories. His main interests are computer architecture configuration planning, performance evaluation, database design, and database management systems. He has published several articles on the latter topics, and has consulted on all the above on both an internal and a contract-research basis for Battelle. Mailing address: Battelle Columbus Laboratories, 505 King Avenue, Columbus, Ohio 43201.

JOHN PAGE is currently a research associate at the International Institute for Applied Systems Analysis in Austria. His research interests are the social, political, and economic impacts of new information technology, particularly electronic text transfer and satellite communications. He is a consultant for a number of organizations, including Unesco, IAEA, and the Commission of the European Communities. He is a former director at what is now the European Space Agency, where he was responsible for scientific and technical information and education programs from 1963-1973. Mailing address: IIASA, A-2361 Laxenburg, Austria.

JOHN R. RUMBLE, JR. is at the Office of Standard Reference Data of the U.S. National Bureau of Standards. He is responsible for the Materials Data Program as well as the computer activities for the entire office. He has previously been associated with the JILA Atomic Collision Data Center and the Atomic and Molecular Data Program of the IAEA. Mailing address: National Bureau of Standards, Office of Standard Reference Data, Physics Building, Room A323, Washington, DC 20234.

JOHN RUSSELL is a database administrator, working in the Computer Section of the International Atomic Energy Agency. He is responsible for many scientific and administrative databases used by the IAEA. Mailing address: International Atomic Energy Agency, P.O. Box 100, A-1400 Vienna, Austria.

CAROLE SCHERMER is currently head of Data and Word Processing at the American Chemical Society. For the last three years, she was involved in managing database administration functions and services for ACS Columbus at Chemical Abstracts Service, and prior to that was involved in many database development projects. She has served in the Software AG International Users Group as chairperson for the Data Administration and Data Base Administration Special Interest Group. Mailing address: American Chemical Society, 1155 Sixteenth Street, NW, Washington, DC 20036.

NATHAN L. SEIDENMAN works in the Data Systems Development Group of the Office of Standard Reference Data at the U.S. National Bureau of Standards. He is in charge of the computer facilities there and has extensive experience in graphics. Mailing address: National Bureau of Standards, Office of Standard Reference Data, Physics Building, Room A323, Washington, DC 20234.

F. JACK SMITH is a computer scientist, presently head of the Department of Computer Science at Queen's University, Belfast. His present main interests are scientific databases and information retrieval from text. He was previously an Atomic Physicist. He is a member of the National Board for Science and Technology in Dublin which advises the Irish Government on Science and Technology. Mailing address: Department of Computer Science, The Queen's University of Belfast, Belfast, BT7 1NN Ireland.

ACKNOWLEDGMENTS

Many persons deserve thanks for this volume, but special recognition is needed for Jeanne Bride of NBS for her excellent editorial comments and for Loretta Ellison, Terri Gladhill, Beverly Geisbert, and Terry Grimes, all of NBS, for their patient and careful preparation of the manuscript.

TABLE OF CONTENTS

		Page
	FOREWORD	v
	EDITORIAL NOTE	vii
	LIST OF CONTRIBUTORS	viii
	ACKNOWLEDGMENTS	xi
	TABLE OF CONTENTS	xii
Chapter 1	DATA, COMPUTERS, AND DATABASE MANAGEMENT SYSTEMS J. R. Rumble, Jr.	1
	1.1 A Beginning	1
	1.2 Use of Data in Science and Technology Today	2
	1.3 Advantages of Computer-based Data Operations	3
	1.4 The Definition of Databases and Database Management Systems	4
	1.5 The Importance of Communications	10
	1.6 Database Management Systems - How to Use Them	11
	1.7 References	12
Chapter 2	DATA TYPES AND STRUCTURES IN SCIENCE AND TECHNOLOGY A. A. Brooks	15
	2.1 Introduction	15
	2.2 The Theoretical Basis for Data Type and Structure	16
	2.3 Data Types	16
	2.4 Data Structures in Science and Technology	20
	2.5 References	36
Chapter 3	DATABASE MANAGEMENT SYSTEMS FOR SCIENCE & TECHNOLOGY F. D. Gault	39
	3.1 Basic Characteristics	39
	3.2 Differences Between Systems in Business and Science	46
	3.3 Flat-File, Hierarchical, Network, and Relational Systems	51
	3.4 Review of Database Management Systems in Science & Technology Project	61
	3.5 References	68

Chapter 4 ANALYSIS AND DISPLAY OF DATA IN SCIENCE AND TECHNOLOGY 75
J. R. Rumble, Jr. and N. L. Seidenman

4.1 Extraction and Preparation of Data Sets . . . 75
4.2 Statistical Analysis. 76
4.3 Editing, Composing, and Typesetting 78
4.4 Graphics. 81
4.5 Basic Graphics - Hardware 82
4.6 Basic Graphics - Software 84
4.7 Graphics - Some Examples 87
4.8 Final Remarks 89
4.9 References. 90

Chapter 5 PLANNING FOR DATABASE MANAGEMENT IN SCIENCE AND TECHNOLOGY 93
F. J. Smith

5.1 The Planning Process. 93
5.2 The Overall Plan. 101
5.3 The User Requirement. 101
5.4 Working Environment 106
5.5 Cost. 107
5.6 Alternatives and Funding. 114
5.7 Continuous Planning 116
5.8 References. 118

Chapter 6 DESIGNING THE DATABASE MANAGEMENT SYSTEM 119
H. Bestougeff

6.1 The Design Process - An Overview. 119
6.2 Database Design 126
6.3 Database Management System Design 143
6.4 Processing Requirements 153
6.5 Database Design Tools 158
6.6 References. 160

Chapter 7 SELECTING A SUITABLE DATABASE MANAGEMENT SYSTEM. . 167
M. A. Huffenberger and C. A. Schermer

7.1 Selecting a Suitable Database Management System. 167
7.2 Requirements Analysis 168
7.3 Identification of Software Options. 175
7.4 Source Determination. 177
7.5 Selection 182
7.6 Database Management System Acquisition. . . . 191
7.7 Summary 194
7.8 References. 195

Chapter 8 IMPLEMENTING AND USING THE DATABASE MANAGEMENT SYSTEM 197
J. Russell and D. E. Cullen

8.1 Roles of the Database Administrator and Manager 197
8.2 Establishing the Database 200
8.3 Updating & Integrity. 202
8.4 Measuring Performance 204
8.5 Communications & Interactive Use 205
8.6 Monthly Newsletters & User Feedbacks. 206
8.7 Marketing your Data Products. 208
8.8 References. 209

Chapter 9 INTEGRATING YOUR SYSTEM WITH DATA COMMUNICATION NETWORKS AND TO OTHER DATABASE ACTIVITIES 211
J. R. U. Page

9.1 Computer Communications - The Essentials. . . 211
9.2 The Properties and Characteristics of Data Networks. 217
9.3 Available Data Networks 220
9.4 Protocols and Standards 228
9.5 Database Operations for Remote Users. 232
9.6 References. 237

Chapter 10 DATABASE MANAGEMENT CONCERNS AT THE INTERNATIONAL LEVEL. 241
C. Gottschalk

10.1 Perspectives, Trends, and Strategies. 241
10.2 The Activities of CODATA, UNESCO, and Other International and National Groups 242
10.3 Scientific and Technical Information, Computers, and The Developing Countries . . . 248
10.4 Databases and the Developing Countries. . . . 251
10.5 References 255

Database Management in Science and Technology
J.R. Rumble, Jr. and V.E. Hampel (eds.)
Elsevier Science Publishers B.V. (North-Holland)

CHAPTER 1

DATA, COMPUTERS, AND DATABASE MANAGEMENT SYSTEMS

John Rumble, Jr.

National Bureau of Standards
Washington, DC

1.1 A BEGINNING

This book is concerned with computerizing the collection, manipulation, and distribution of numeric, scientific, and technical data. Collected data are stored in a *database* and are used by several people. Computer programs or software that create, manipulate, and use the database are called *Database Management Systems*.

In the following pages, we will guide you through the process of planning, building, and implementing database management systems for scientific and technical data. Since this is a *Sourcebook*, the aim is to present clearly the basic considerations and to point the reader to more detailed literature for further information.

The use of computers to handle data and database management systems has gained in popularity over the past few years. The technology and economics of computer hardware and software are still rapidly changing. However, the methodology of a database project remains fairly constant and is the real subject of this book. We anticipate the reader who intends to create a database and use a DBMS wants to be successful. What this book presents to such a reader are the steps to success as shown in table 1-I.

Table 1-I

Steps in a successful DBMS project:

1. Identifying and defining the data needs
2. Planning the data project
3. Designing the database
4. Obtaining a suitable DBMS
5. Implementing the DBMS
6. Linking the database to users
7. Linking the database to other databases

After the first three introductory chapters, the book follows these steps. This outline also reflects the single most important point which all of the chapter authors make. A successful database project does not begin with the selection of a computer and a database management system. Indeed, if this were done first, the data project would likely fail, perhaps spectacularly. The first steps involve identifying the needs, planning and designing possible solutions, and then, and only then, choosing the solution.

1.2 USE OF DATA IN SCIENCE AND TECHNOLOGY TODAY

The everyday work of scientists and engineers is concerned with data. They observe, measure, calculate, and predict; all these activities use and generate data. Of themselves, most data are uninteresting, although some data do demonstrate that indeed we understand a given physical phenomenon. But the real value of data is in its use by other scientists and engineers: Materials must be identified, processes designed, products manufactured, and measurements made. These and other science and technology activities rely on data generated by others.

All scientists and engineers recognize the variety and widespread use of data. The types of data are perhaps less obvious. One classification scheme for scientific and technical data, given by the Committee on Data for Science and Technology (CODATA) of the International Council of Scientific Unions, defines three broad classes of data [1].

Class A - *Repeatable measurements on well-defined systems.* These include traditional physical and chemical data resulting from measurement of well-understood properties of systems of known composition. In principle, data are subject to verification by repeating the measurements in different laboratories at different times.

Class B - *Observational data.* Here are included results of time- or space-dependent measurements that cannot, in general, be checked by remeasurement. This category includes data from biology, geosciences, and environment monitoring.

Class C - *Statistical data.* This class includes nonscientific or nontechnical data of importance in many technological problems: demographic data, chemical production records, energy consumption figures, health statistics, etc.

All three classes of data are important, and each presents different problems when data are collected and distributed on computers. Many similarities exist, and common features are important (see Chapter 2).

Since the value of data is in its use, specifically in its use in making decisions, whether as to the validity of the relativity theory or deciding the best material for an automobile tire, reliable data are important and in many cases the key issue. Quite simply, the better the data, the better the decision [2].

Many groups, both international and national, have extensive data evaluation activities in many areas of science and engineering. Gottschalk, in Chapter 10 of this volume, Rossmassler and Watson in the previous CODATA Sourcebook [3], and references in both outline the scope of these activities.

Data activities take many forms. They can be part of a coordinated national data program; they can result from interest by a technical society, either national or international. They often have come about as the need has arisen to share resources and data between different research groups. Because there often are multiple user groups for a given type of data, the past few years have seen a tremendous increase in the number of cooperative projects relating to data evaluations. The proceedings of the biennial CODATA conferences contain many examples in all areas of science [4]. For example, in the United States, in the area of materials properties data alone, three major cooperative data evaluation projects have recently been set up by the National Bureau of Standards. These include Alloy Phase Diagrams (with the American Society of Metals) [5], Ceramic Phase Diagrams (with the American Ceramic Society) [6], and Corrosion (with the National Association of Corrosion Engineers) [7]. Similar projects in other areas of science and technology are being formed.

Another aspect of the use of data and the need for data quality should be mentioned: the change that computers have brought about in the human interaction with data. Before the advent of computers, each scientist and engineer directly interacted with each piece of data used. Some one person had to select the data source, locate the data values, interpolate, if necessary, and, in short, confront the data head-on. But computers are rapidly changing this, sometimes positively, sometimes negatively. While computers reduce the potential for simple mistakes that accompany human activity, they also reduce contact between the user and the individual datum. Scientists and engineers must adjust their methodologies to take this into account. One consequence is the greater need for pedigreed data and data quality indicators. If the users cannot examine data as closely as needed, someone else should. And the result of that examination should be carried along with the distribution of the data itself.

1.3 ADVANTAGES OF COMPUTER-BASED DATA OPERATION

The above considerations serve nicely as the introduction to the second component of the subject of this book, namely, computers and their use in data activities. By now, it is obvious that computers will affect every aspect of life in the years to come, and the handling of scientific and technical data is no exception. From the very beginning, one important use of computers has been in the generation of data, either through calculations or experimental data collection. Since scientists want to share their results with other scientists, from the beginning the sharing of computerized data collections has been with us. As the size and diversity of the collections have grown, so has the development of software capability to manipulate and present the data.

In fact, computers are ideal for handling data. They can quickly gather, store, manipulate, display, and merge large amounts of data. If data are to be published, they can be transferred to paper without transcription errors. If data are to be used as input to software, they can be used without entry errors. If data need to be interpolated, this can be done easily.

The tasks listed above provide users with capabilities *equal* to those that existed in the pre-computer age. The computer should be able to perform these tasks faster and provide repeat manipulations more efficiently than by hand. But computers provide much more than this.