DOCUMENT DATABASES

Geoffrey James

VAN NOSTRAND REINHOLD COMPANY
New York

Library of Congress Catalog Card Number: 84–3685
ISBN: 0-442-28185-4

Manufactured in the United States of America

Published by Van Nostrand Reinhold Company Inc.
135 West 50th Street
New York, New York 10020

Van Nostrand Reinhold Company Limited
Molly Millars Lane
Wokingham, Berkshire RG11 2PY, England

Van Nostrand Reinhold
480 Latrobe Street
Melbourne, Victoria 3000, Australia

Macmillan of Canada
Division of Gage Publishing Limited
164 Commander Boulevard
Agincourt, Ontario M1S 3C7, Canada

15 14 13 12 11 10 9 8 7 6 5 4 3 2 1

Library of Congress Cataloging in Publication Data

James, Geoffrey, 1953–
Document databases.

Includes bibliographical references and index.
1. Data base management. 2. Technical publishing—
Data processing. I. Title.
QA76.9.D3J35 1985 001.64 84–3685
ISBN 0-442-28185-4

DOCUMENT DATABASES

For Kathie.

Special thanks to John L. Joseph, James L. Carr, and Tracy K. Tondro for their advice and inspiration, and to Richard Petkiewicz for providing photographic skill.

PREFACE

This book describes the design and implementation of document databases. Document databases are an economical and flexible means of providing multiple-format and multimedia technical information. The document database methodology exploits recent breakthroughs in publications and software technology, providing powerful capabilities that were impossible in the past.

Most books on computers and technical publications deal entirely with hard-copy printing. Little or no attention is given to using the full capability of the computer to create, update, control, and distribute documents. This book describes a revolution that is taking place even as it is being written—a revolution in publication techniques that will rival the revolution in office automation. The document database methodology represents the logical extension of the ever increasing use of computer technology in publications environments.

This book is an opportunity for documentation managers, in-plant printing managers, word processing personnel, software suppliers, and technical writers to take advantage of current technology and to plan for future developments. This book is organized to address the central issues of the new publications technology. The chapters are structured as follows:

Chapter 1 suggests that our civilization is entering an information crisis, which cannot be solved by technology alone. This chapter describes the document database theory and the basic concepts used in the remainder of the book.

Chapter 2 explains how the different text manipulation techniques either help or hinder the development and maintenance of computer-stored documents.

Chapter 3 explains the different methods of computer graphics manipulation and their application in a technical publications environment.

Chapter 4 describes a powerful application of the document database, on-line document querying. Unlike traditional printed books, on-

line querying gives the reader direct interactive access to the document database. Various querying techniques and their uses are described.

Chapter 5 illustrates how the computer assists in the preparation of indexes. Automatic and declarative indexes are described, with their respective advantages and disadvantages.

Chapter 6 describes how computers are used to distribute and maintain a large set of publications; it also describes how the publications environment can be changed to be more responsive to the reader.

Chapter 7 describes the tools that writers, editors, and managers use to create and maintain quality, which is defined as usability, accuracy, consistency, and accessibility.

Chapter 8 discusses publications systems, their hardware integration, configuration, and workflow.

Chapter 9 estimates the productivity improvements and cultural changes that result from the application of the document database methodology in the technical publications environment.

Chapter 10, the final chapter, describes how the document database concept will impact the ways that we write, read, and think about publications.

Throughout this book are sample document database printouts and reports. Unless otherwise noted, these figures were generated on Honeywell's Control Program-Six (CP-6) operating system, using its Computer Aided Publication (CAP) document database software. Certain of the examples have been altered slightly so that they are easier to explain.

GEOFFREY JAMES

CONTENTS

Chapter 1
METHODOLOGY

THE INFORMATION CRISIS

The survival of a species is dependent upon its ability to communicate. Until the advent of the human race, this communication took place through genetics and inheritance. The human race has passed through verbal and written communication, and finally to publication as the means of preserving the information that guarantees the survival and growth of civilization.

Changes are now taking place in our civilization that obsolete the communications techniques of the past. Our rapidly developing technology is outstripping the speed with which the mechanisms of publication can document it.

A single scientific development sometimes requires volumes upon volumes of technical publications. According to one engineer the author has spoken with, the manuals and specifications for an aircraft often take up more physical space than the craft itself. The problems of creating and maintaining such a large set of documents are immense. A single change in a type of bolt can be like throwing a stone into a pool, causing a ripple effect of changes throughout the entire document set. Updating manuals for a developing project often involves hundreds of clerical "editors" madly pushing paper.

As the pace of new developments quickens, libraries overflow with journal articles, textbooks, manuals, specifications. This flood of technical data, because of its sheer size, will probably never be used to its potential. Becoming an expert even in a limited technical field requires years of effort, locating and digesting technical information.

The medium for communicating technical knowledge—the published book, manual, or journal—is overburdened. A vital article buried in a stack of irrelevant paper is almost as unavailable as if it had never been written. Ironically, as the demand for technical knowledge rises, mas-

sive manpower and high-speed machines are dedicated to the task of documenting it. This produces an immense bulk of published material that remains, through its very size, unwieldy and difficult to use.

These problems are widespread. According to one industry analyst, over 83,000 companies in the United States are major producers of publications. This does not include schools and universities, nor does it include that ultimate distributor of printed paper, the U.S. Government. Magnify this vast need for technical knowledge onto a world scale, and the size of this information crisis confounds the imagination.

Many of these new scientific developments, such as genetic engineering and alternative energy sources, directly affect the ability of our civilization to survive and grow. It is even possible that lack of adequate technical knowledge might cause fatal failures in sophisticated weapons systems. The information crisis attacks the very center of our civilization and culture and in a very real sense threatens our survival.

We desperately need a communications revolution that increases our power to communicate technical information, so that we can better understand and explain to others the complexities of our fast-changing environment.

THIRD WAVE MANUFACTURING

The information crisis is not new. There has always been a need for the right information at the right time. However, three revolutions in manufacturing are taking place that magnify the inadequacy of our current methods of technical publication:

- Computer Aided Design (CAD)
- Computer Aided Manufacturing (CAM)
- Robotics

Let us examine these revolutions to gain a better appreciation of the problem.

The first revolution is taking place in design methods for advanced technology products. Computer Aided Design (CAD) allows an engineer to store, retrieve, and manipulate graphic information on a screen. More than just automated drafting, CAD systems can be directed to

create and test assemblies and even entire machine systems—in the computer's memory. This not only causes a tenfold increase in the productivity of drafters and designers, but it greatly reduces the cost of developing a product. Prototypes can be tested, rejected, and refined without ever being physically constructed. Planned CAD extensions even allow the theoretical construction of the tools that will be required to build the final product.[1]

The second revolution is taking place in the way the products are manufactured. Computer Aided Manufacturing (CAM), using computers in the design and control of the manufacturing process, results in radical speedups (as high as 50-fold) in the making of parts. Because parts can be made more quickly, the average size of production runs has been reduced. Products are easier to tool up for.

Robotics, the third revolution, begins to obsolete the entire concept of mass-manufacturing. With programmable robots driving CAM production lines and creating CAD-designed products, totally customized products become feasible. Robotics will make it economical to manufacture one-time, advanced technology products that uniquely address an individual's or individual company's needs.[2]

These revolutions in manufacturing have been fed by a fourth revolution in the field of software development. The use of design methodology and structured high-level computer languages for software development results in a far higher productivity rate among computer programmers. This creates feedback growth, as many of the new programs are used to further boost programming productivity. Potent "software engines" and "programming power tools" harness computer graphics and artificial intelligence to streamline the design and implementation of software.[3] As programming power increases, new computer applications fuel the development of new technologies.

These changes in the methodology of manufacturing are part of what futurist Alvin Toffler calls the "third wave"—a change away from the industrialism of the 19th Century toward a more diversified and decentralized economy and society. Third wave manufacturing goes beyond the techniques of mass production, beyond the classical separation of producer and consumer. Rather than a marketplace where single-use products are mass-built for a mass audience, third wave manufacturing addresses each consumer (or corporation) as an individual with unique

requirements. By storing potential products within the computer's memory, those unique needs can be met by a customized product. In many cases, the consumer contributes to the design process, or even designs and implements part of the product. The lead time for developing new products and technologies is no longer measured in years, but in months or even weeks.

This type of producer/consumer interaction can be illustrated by examples from two widely different industries.

Computer vendors used to offer hardware and software much like Ford once marketed the Model-T—"any color as long as it's black." Giant mainframes and their massive operating systems were packaged as a bundle. Typically, the newly purchased computer needed additional programming to address the data processing needs of the buyer. Today, however, computers are sold with numerous options on hardware and software, each option priced and purchased separately. Computer buyers demand turnkey systems that uniquely address their particular business or application. To fulfill this need, many small development companies have sprung up to modify or augment the problem-solving capabilities of some other company's machine. Frequently, there is considerable interweaving of the interests of two firms, who buy and sell each other's product. The distinction between consumer and producer becomes unclear.

Third wave manufacturing has even penetrated to that stronghold of mass-production, the automotive industry. Sophisticated design and manufacturing techniques have greatly reduced the time it takes for automobile manufacturers to bring new models into the showroom, creating a much wider selection than ever before. As if in direct opposition to the old Henry Ford philosophy, few cars sold today are "stock." Car buyers pick and choose the appearance and functional level of their cars, adding optional features and packages.

As a result of these three revolutions, products are both less expensive and more responsive to the consumer. When the consumers manufacture yet more products (for example, when the consumer is a company), a matrix of goods and services are set up that radically alters the way that we perceive the marketplace. Technology and innovation become self-feeding. Discovery upon discovery geometrically increase our capabilities.

IMPACT ON TECHNICAL COMMUNICATIONS

This rosy picture disintegrates under the glare of the information crisis. Current methods of creating and retrieving technical publications are already overloaded. How can they expect to cope with hundreds of variations of thousands of new technologies and products, each feeding yet further development?

We already see signs of this problem in the computer industry. The proliferation of personal computers in the home and small business has been accompanied by outcries against the inadequacy of documentation. Large commercial computer users, who have never been satisfied with the quality and completeness of their vendor's manuals, renew requests for greater depth and accuracy.

Realizing that documentation is as much a part of a product as hardware and software, computer vendors have begun to treat their technical writers as professionals rather than clerical workers. For the first time, *Datamation* magazine includes "Technical Writer" as a category in their salary survey of Data Processing professionals.

But increased writing manpower aggravates the information crisis even as it attempts to alleviate it. Many large computer systems have literally hundreds of manuals describing different elements of the system. The very size of the set almost guarantees that the user will be unable to locate the single item of information in that haystack of words. Even microprocessor vendors have similar problems. Some users demand more documentation, while others are frightened by anything larger than a pamphlet. Nobody ever seems to be satisfied with computer documentation, regardless of the time and effort expended in writing and publishing it.

The computer business is a mirror of what is happening on a larger scale as the snowball of technology gathers force. As the demand for accurate technical knowledge increases geometrically, scientific breakthroughs will be underutilized or even ignored. Vital resources will be wasted reinventing existing products and techniques. The synthesis of different disciplines, which could provide a unified plan for our future, will not take place—all because technical knowledge, the "glue" that holds the machinery of our civilization together, could not be correctly applied.