digital information technologies

THE FOUNDATION OF ORGANIZATION

INTELLECTUAL INFORMATION

Elaine Svenonius

The Intellectual Foundation of Information Organization

Elaine Svenonius

The MIT Press
Cambridge, Massachusetts
London, England

This book was set in Sabon by The MIT Press and was printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Svenonius, Elaine.
The intellectual foundation of information organization / Elaine Svenonius.
p. cm—(Digital libraries and electronic publishing)
Includes bibliographical references and index.
ISBN 0-262-19433-3 (hc.: alk. paper)
1. Information organization. 2. Bibliography—Methodology. 3. Cataloging.
I. Title. II. Series.
Z693.S92 2000
025.3—dc21 99-41301
CIP

Contents

Preface

Instant electronic access to digital information is the single most distinguishing attribute of the information age. The elaborate retrieval mechanisms that support such access are a product of technology. But technology is not enough. The effectiveness of a system for accessing information is a direct function of the intelligence put into organizing it. Just as the practical science of engineering is undergirded by theoretical physics, so too the design of systems for organizing information rests on an intellectual foundation. The topic of this book is the systematized body of knowledge that constitutes this foundation.

Much of the literature that pertains to the intellectual foundation of information organization is inaccessible to those who have not devoted considerable time to the study of the disciplines of cataloging, classification, and indexing. It uses a technical language, it mires what is of theoretical interest in a bog of detailed rules, and it is widely scattered in diverse sources such as thesaurus guidelines, codes of cataloging rules, introductions to classification schedules, monographic treatises, periodical articles, and conference proceedings. This book is an attempt to synthesize this literature and to do so in a language and at a level of generality that makes it understandable to those outside the discipline of library and information science.

A book on the intellectual foundation of information could be written in several ways. It is therefore useful to state the scope of this one, contrasting what it is not about with what it is about. First, it is not a how-to-do-it cookbook of methods used to organize information. The techne or practical skill of information organization is a function of changing technology, whereas its intellectual foundation, which encompasses theory, is relatively impervious to change. To ground the discussion of theory, however, particular devices

and stratagems used by different technologies are introduced by way of example. Thus, general statements involving abstractions are frequently followed by a detail or a graspable image.

The book does not focus primarily on how users seek information but rather on the design of organizing systems. Systems for organizing information must be designed with the user in mind, but sometimes overlooked is that the objectives and principles that undergird these systems constitute a hypostatization of users' needs. The specifications relating to user satisfaction that are embodied in these objectives and principles have been developed and refined over a period of 150 years. They are not only historically determined but also empirically warranted. Moreover, they are more stringent than can be imagined by most users or, for that matter, inferred from most studies of information seeking behavior.

This book is not primarily about how the computer is used to organize information, although the topic is discussed, since recognizing the impact of technology on information is unavoidable. The digital revolution has affected how information is embodied and what is used to organize it. It has forced a general reexamination of how the carriers of information are identified and described. Using automation to achieve the objectives of systems for organizing information has opened avenues of research and development that have significantly enriched the body of knowledge that constitutes the intellectual foundation of information organization.

The book is not written for the novice who is about to begin a job as a cataloger and wants an instant understanding of its mysteries. It is not a catechism of rules, a compendium of practice, or a training manual. Instead, the book takes a scholarly approach and looks at the rules used to describe information entities — not to spell them out but to consider their intellectual source and grounding or lack thereof. It looks at principles that have been used to guide systems design, asks why decisions were made as they were, and considers problems that were encountered and overcome. Oriented thus, the book is directed toward two groups of people: those who are interested in information organization as an object of scholarly investigation and those who are involved in the design of organizing systems.

This book does not enumerate various systems for organizing information, though meritorious features of these are referenced by way of example, but strives to express what these systems have in common — to speak

in terms of generalities rather than particulars. One of its central aims is to look at information organization holistically and thereby to raise discourse about it to a level general enough to unify the presently compartmentalized approaches for achieving it. Specifically, it endeavors to integrate the disparate disciplines of descriptive cataloging, subject cataloging, indexing, and classification. A difficulty in carrying out this aim, and indeed in writing the book, has been to reconcile different ways of referring to similar concepts, principles, and techniques. To deal with this difficulty — and to limit jargon generally — an effort has been made to eschew where possible discipline-specific terminology and to resist the temptation of inventing new terminology.

Finally, this book is not an idiosyncratic view on how to organize information effectively. Rather, it reflects practice and theory as developed within the discipline of library and information science. It adopts a particular conceptual framework that views the process of organizing information as the use of a special language of description, called a *bibliographic language*. This framework is rooted in a tradition that originated nearly a hundred years ago and has been used since then by theorists to introduce rigor, unification, and generality into theorizing about information organization.

The book is divided into two parts of five chapters each. The first part is an analytic discussion of the intellectual foundation of information organization. Chapter 1 introduces and defines what is meant by an *intellectual foundation* and the concepts of *information* and *document*. It establishes a conceptual framework that identifies the central purpose of systems for organizing information: bringing like things together and differentiating among them. It considers the function of principles in the context of systems design and concludes with an illustration of some of the problems encountered in the design of organizing systems.

The second chapter looks at one of the cornerstones of the intellectual foundation of information: the objectives of systems designed to retrieve information. It reviews their history from Antonio Panizzi (1850), through Charles Ammi Cutter (1876) and Seymour Lubetzky (1957), to the 1998 International Federation of Library Associations and Institutions (IFLA) *Functional Requirements for Bibliographic Records*. An additional objective is postulated and an argument made for it on the basis of literary and

use warrant. The degree to which the objectives can be operationalized is discussed as well as arguments pro and con their necessity.

Chapter 3 deals with ontology, the information entities mandated by the objectives, which include documents and sets of documents formed by the attributes of work, edition, author, and subject. It discusses the function of these entities in information organization and the problems that attend their definition. A distinction is made between conceptual and operational definitions. The latter, expressed in set-theoretic terms, are needed for uniformity and precision in bibliographic description and for automating aspects of organizing information.

Chapter 4 conceptualizes the organization of information as the use of a special purpose bibliographic language. This conceptualization has several advantages, two of the most important being that it unifies the traditional subject and author-title approaches to information organization and enables the development of a bibliographic-specific linguistic theory. Bibliographic languages are classified in terms of the objects they describe — whether works (information per se), documents (carriers of information), or subjects — and are categorized in terms of their components (that is, vocabulary, semantics, and syntax). The chapter concludes with a discussion of the rules governing the use of these languages and the form and function of the bibliographic descriptions created by their application.

Another foundation cornerstone consists of the principles or directives that guide the construction of bibliographic languages. This is the topic of Chapter 5. Five principles are explicated: user convenience, representation, sufficiency and necessity, standardization, and integration. These are discussed from the point of view of their origin, usefulness, internal conflict, and viability in a multimedia environment.

The second half of the book moves from generalities to particulars. It presents an overview of three bibliographic languages used to organize information — work languages, document languages, and subject languages — and looks at these languages in terms of their vocabulary, semantics, and syntax.

Chapter 6 looks at the languages used to describe works, illustrating these using the work language developed within the *Anglo-American Cataloging* tradition, which is the most sophisticated language so far developed. A distinction is drawn in the vocabulary of this language, which

Acknowledgments

This work has come about as a result of many years of association with those engaged in theoretical scholarship in the organization of information. Their influence has fanned an excitement kindled when in the Graduate Library School at the University of Chicago I learned to look at the organization of information as an object of study. I have been inspired by the interest of my students and gained insights by their questions and ideas. I feel fortunate to have been able to talk with many of the fine minds of the twentieth century who have contributed significantly to the disciplines of cataloging, classification, and indexing, both in this country and abroad, especially in India. I feel particularly fortunate for the recent conversations I have had with Seymour Lubetzky, whose principled thinking, even at age 100, is remarkable to behold.

As to the actual writing of the book, my first thanks must go to Dorothy McGarry, who willingly shouldered the burden of reading chapters as they were produced, catching me up on details, asking for clarifications, and being always and wonderfully encouraging. And then thanks go to a few of the people who reviewed the book in manuscript form for MIT Press — Barbara Tillett and Ed O'Neill. Their constructive suggestions were very welcome indeed as I was struggling to produce a better product. Thanks also to Richard Fackenthal, Helen Schmierer, and Boyd Rayward, all of whom read parts of the book and gave me reactions and points of view I found I needed. A special thanks to Bhagi Subramanyam for her bibliographic and moral support at the end of my labors. Thanks finally to the staff at MIT: copyeditors Deborah Cantor-Adams and Rosemary Winfield, who conferred elegance on my writing; Erica Schultz, who flawlessly composed the text; and acquiring editor Doug Sery, who started me on what proved to be a difficult journey and cheered me along the way.

The Intellectual Foundation of Information Organization

1

Information Organization

Introduction

A system for organizing information, if it is to be effective, must rest on an intellectual foundation. This intellectual foundation consists of several parts:

- An ideology, formulated in terms of purposes (the objectives to be achieved by a system for organizing information) and principles (the directives that guide their design);
- Formalizations of processes involved in the organization of information, such as those provided by linguistic conceptualizations and entity-attribute-relationship models;
- The knowledge gained through research, particularly that expressed in the form of high-level generalizations about the design and use of organizing systems; and
- Insofar as a discipline is defined by its research foci, the key problems that need to be solved if information is to be organized intelligently and information science is to advance.

Conceptual Framework

It is useful to begin by establishing a conceptual framework to ensure that the discussion does not become idiosyncratic and at the same time to bootstrap it to the level of theory. The conceptual framework adopted here looks at the organization of information in an historico-philosophical context. Its salient feature is that information is organized by describing it using a special-purpose language.

Systems thinking was introduced into the discipline of information organization by Charles A. Cutter in 1876.[8] Dubbed the great "library systematizer,"[9] Cutter was the first to recognize the importance of stating formal objectives for a catalog. He recognized as well the need to identify the means to achieve these objectives and principles to guide the choice of means when alternatives were available. Since Cutter's time, systems thinking has assumed a variety of different expressions, tending to become more elaborate and increasingly formalized, as, for instance, in its articulation in the form of conceptual modeling. However expressed, the ultimate aim of systems analysis is to determine and validate practice. Why certain methods, techniques, rules, or procedures are adopted to the exclusion of others in the practice of organizing information requires explanation. One way to provide this is to show that a particular element of practice can be viewed as part of a system and as such contributes to fulfilling one or more of the system's objectives.[10] An improvised practice, one that is adventitious and not rationalized with respect to the big picture, is ineffective, inefficient, and, by definition, unsystematic.

Philosophy of Science

Scientific methodology has been a central focus for philosophical inquiry for nearly a century. In the first part of the twentieth century, the dominant philosophy of science was logical positivism, whose credo was expressed by the principle of verifiability. This principle states that to be meaningful a proposition must be capable of verification. A proposition to be verified must have concepts that can be operationalized, which means (in effect) interpreted as variables and defined in a way that admits of quantification.

To the extent that problems encountered in the organization of information are definitional in nature, solutions to them can be approached by introducing constructive or operational definitions. An example of such a definition relating to information organization is the dual precision-recall measure created by Cyril Cleverdon in the mid-1950s. The measure was introduced to quantify the objectives of information retrieval. Precision measures the degree to which a retrieval system delivers relevant documents; recall measures the degree to which it delivers all relevant documents.

Defining concepts operationally enables a discipline to advance, the most frequently cited illustration of which is Einstein's use of them in his analy-

sis of simultaneity.[11] The power of operational definitions resides in their ability to provide empirical correlates for concepts in the form of variables, which, in turn allows variables to be related one to another.[12] For instance, quantifying the objectives of information retrieval in terms of the precision and recall variables makes it possible to establish propositions about the impact of various factors — such as specificity of indexing, depth of indexing, and vocabulary size — on retrieval effectiveness. Propositions that express relationships among variables are "scientific" in the sense that they represent high-level generalizations about the objects of study. This gives them an explanatory function: if verified, they assume the character of laws; if in the process of being verified, they have the status of hypotheses.

While some aspects of the philosophy of science are abstruse, its dictates are clear enough: quantify and generalize. To a greater or lesser degree all the social sciences have struggled to follow these dictates. In their striving for scientific respectability, they have pursued empirical research and undergone quantitative revolutions. Library "science" self-consciously embraced a scientific outlook in the 1930s at the Chicago Graduate Library School. This school, established for the express purpose of conducting research, had considerable influence on the field through its brand of scholarship, which encompassed theory, forced definitional clarity, and questioned assumptions.[13] Increasingly since the 1930s, understanding of the information universe and, in particular, how it is organized and navigated has been pursued through "scientific" research.

Language Philosophy

Interest in language has dominated two twentieth-century philosophies. The first was the already mentioned logical positivism, which was a linguistic form of radical empiricism. Its principle of verifiability — which states that a proposition to be meaningful must be capable of being verified — is a linguistic principle.[14] The philosophy of logical positivism was countered in the middle of the century by another language philosophy, the Wittgensteinian philosophy of linguistic analysis.[15] A major tenet of this philosophy was that the meaning of a word is its use and this use is governed by rules much like the rules that govern moves in games. As there are many different special-purpose uses of language, so there are many different language games.

The act of organizing information can be looked on as a particular kind of language use. Julius Otto Kaiser, writing in the first decade of the twentieth century, was the first to adopt this point of view.[16] Kaiser developed an index language, which he called *systematic indexing*, wherein simple terms were classed into semantic categories and compound terms were built using syntax rules defined with respect to these categories. Similar points of view have been adopted by theorists since Kaiser, mostly in the context of organizing information by subject but applicable as well to organizing by other attributes, such as author and title. The advantage to be gained by looking at the act of organizing information as the application of a special-purpose language is that linguistic constructs such as *vocabulary, semantics,* and *syntax* then can be used to generalize about, understand, and evaluate different methods of organizing information.[17] Another advantage is that these constructs enable a conceptualization that can unify the heretofore disparate methods of organizing information — cataloging, classification, and indexing.

Philosophical movements constitute the backdrop against which scholarly disciplines develop. The impact of systems philosophy on the discipline of information organization is apparent insofar as this organization is regarded as effected by a system that has purposes and whose design is guided by conceptual modeling and the postulation of principles. It is apparent as well in the discipline's increasing reliance on operational definitions, in its use of algorithms for automating aspects of organization, in frameworks it establishes for empirical research, and in generalizations that build theory.

Information and Its Embodiments

Like *meaning* and *significance,* terms with which it is allied, *information* has many senses, nuances, and overtones. This makes reaching agreement about a general definition of the term difficult. Some special-purpose definitions of the term have relatively fixed meanings. The best known of these is the one that is used in information theory, which associates the amount of information in a message with the probability of its occurrence within the ensemble of all messages of the same length derivable from a given set of symbols.[18] A definition like this, however, is too particular for use in discourse about organizing information. What is needed is one more conso-

nant with common usage, one that implies or references a person who is informed. The definition used in this book is developed in the next chapter, but as first approximation a gloss on a general dictionary meaning will do. One definition of *information* is "something received or obtained through informing."[19] Informing is done through the mechanisms of sending a message or communication; thus, *information* is "the content of a message" or "something that is communicated."

Defining *information* as the content of a message is specific enough to exclude other definitions — for instance, the definition that equates information with "a piece of fact, a factual claim about the world presented as being true."[20] This definition, which is positivistic in nature, conceptualizes *information* narrowly. Certain types of knowledge may be restricted to facts or true beliefs, but to apply such a restriction to information in general would rule out the possibility of false information or information that is neither true nor false, such as the information in a work of art or a piece of music, which when conveyed "informs" the emotions. Factual claims about the world constitute only a small subset of information broadly construed as the content of a message or communication.

Information is sometimes defined in terms of data, such as "data endowed with relevance and purpose."[21] A datum is a given; it could be a fact or, at a more elemental level, a sense perception. Either might be endowed with signatory meaning simply by focusing attention on it, as a certain smell is indicative of bread baking. While data in the form of sense perceptions and raw facts have the potentiality to inform, it cannot be rashly assumed that all information could be reduced to these. It is not possible, at least not without wincing, to refer to *The Iliad*, *The Messiah*, or the paintings in the Sistine Chapel as data, however endowed. The messages they convey represent highly refined symbolic transformations of experience,[22] different in kind from data.

While message content is probably a good approximation of what information systems organize, not all message content falls under the purview of such systems. The content contained in ephemeral messages — such as the casual "Have a nice day!" — lies outside the domain of information systems. For the most part, these domains are limited to messages whose content is (1) created by humans, (2) recorded,[23] and (3) deemed worthy of being preserved. The question of which messages fall into the latter category

is sometimes begged by equating "worthy of being preserved" with what libraries, information centers, archives, and museums in fact collect. The collective domain of all systems for organizing information — all message content created by humans, recorded, and deemed worthy of being preserved — has been likened to the "diary of the human race."[24] The purpose of these systems is to make this diary accessible to posterity.

The term *document* is easier to define and is used in this book to refer to an information-bearing message in recorded form.[25] This usage is warranted both by the information-science literature and by common usage.[26] *Webster's Third* gives as meanings of *document:*

- a piece of information
- a writing (as a book, report, or letter) conveying information
- a material having on it (as a coin or stone) a representation of the thoughts of men by means of some conventional mark or symbol.[27]

The first two of these meanings are particularly apt in that they explicate *document* with respect to *information:* "a piece of information" and "conveying information." The second is limited in that it instances "a writing," whereas in contemporary bibliographic contexts documents include not only messages using alphanumeric characters but also those expressed using sounds and images.

The third meaning of *document* introduces the concept of *material*. This underscores a distinction of great importance in the literature of information organization, one that is referenced repeatedly throughout this book: information is an abstract, but the documents that contain it are embodied in some medium, such as paper, canvas, stone, glass, floppy disks, or computer chips. Potentially any medium can serve as a carrier of information. While some media make information immediately accessible to the senses (for example, paper), others require an intermediate mechanism (such as a computer chip, a microfiche, or a compact disc). Organizing information to access it physically requires not only descriptions but also its material embodiments and the mechanisms needed for retrieval.

The distinction between information and its embodying documents is so important in the literature of information organization it warrants a brief history. It is claimed to have been recognized as early as 1674 by Thomas Hyde.[28] Certainly Panizzi in the middle of the nineteenth century acknowledged it implicitly in the design of his catalog and in certain passages of his

Normally bibliographic systems that organize information in documents do more than bring together *exactly* the same information; they aim also to bring together *almost* the same information. This introduces further complexity, particularly in trying to understand what is meant by "almost the same information." Intuitively the concept is simple to grasp. A work like *David Copperfield* may appear in a number of editions, such as one illustrated by Phiz, one translated into French, and another a condensed version. Because they are editions of the same work, they share essentially, but not exactly, the same content, differing only in incidentals such as illustrations, language, size, and so on. But the attempt to operationalize the intuitive concept in a code of rules — to draw a line between differences that are incidental and those that are not — runs into definitional barriers: What is a work? What is meant by *information*?[35]

Once editions containing almost the same information are brought together, their differences then need to be pinpointed. Panizzi insisted on this in his defense before the Royal Commission: "A reader may know the *work* he requires; he cannot be expected to know all the peculiarities of different *editions;* and this information he has a right to expect from the catalog."[36] He then went on to argue for a full and accurate catalog, one that contained all the information needed to differentiate the various editions of a work. The task of differentiation has its mind-torturing challenges and can create what to an outsider might seem like a display of bibliographic vanity. But imagine the hundreds of editions of the Bible that might be held by a library. Not only must salient differences be identified, but they must be communicated intelligibly and quickly. Intelligible communication in part is accomplished by arranging records for the different editions in a helpful order. The placing a given edition in its organizational context within the bibliographic universe is not unlike making a definition: first one states its genus (the work to which it belongs) and then, in a systematic way, its differentia.

The essential and defining objective of a system for organizing information, then, is to bring essentially like information together and to differentiate what is not exactly alike. Designing a system to achieve this purpose is subject to various constraints: it should be economical, it should maintain continuity with the past (given the existence of more than 40 million

documents already organized), and it should take full advantage of current technologies.

In addition to constraints, certain principles inform systems design. Principles are desiderata that take the form of general specifications or directives for design decisions. They differ from objectives in that objectives state what a system is to accomplish, while principles determine the nature of the means to meet these objectives. An example of a principle used to design the rules used to create a bibliographic system states that these rules collectively should be necessary and sufficient to achieve system objectives. Others are that rules should be formulated with the user in mind, they should ensure accuracy, they should conform to international standards, and they should be general enough to encompass information in any of its embodiments.

What makes the labor of constructing a bibliographic system colossal are the problems that are encountered in the process of doing so. A major source of problems is the infinite and intriguing variety of the information universe. These kinds of problems are frequently definitional in nature: defining *work*, for example, is difficult because it amounts to defining *information*. Does *The Iliad* in the original Greek consist of the same information (represent the same work) as an English translation of it? Do two different English translations represent the same work? (The answer to these questions is usually yes.) Does translation to another medium abrogate workhood? Does a film version of *Hamlet* contain the same information content as its textual counterpart? (The answer to this kind of question is usually no.) Are two recordings of a symphony, one a CD and the other a video, the same work? (Here the answer seems to be pending.) The dictum that "the medium is the message"[37] suggests that there is significant value added (or subtracted) when an original work is adapted to another medium, so that information that is to be organized is a function of its symbolic expression. The definition of *work* has become the focus of recent attention, which is hardly surprising since it is important to come to grips with the meaning of information. This is something that needs to be grasped, since how information is defined determines what is organized and how it is organized.

Another significant source of problems in organizing information stems from the need to keep pace with political and technological progress. An

example of how technological progress poses problems is the invention and proliferation of new media, which has required bibliographic systems to generalize their scope from books to any kind of media that can carry information. An example of political progress requiring adaptation is the rise of internationalism, which has required these systems to extend their reach from local to universal bibliographical control. Political problems are for the most part settled through international agreements and the establishing of standards but are addressable technically at a systems level. An example is the problem that arises from a conflict between two principles — that of universal standardization and that of user convenience. Different cultures and subcultures classify differently, use different retrieval languages, and subscribe to different naming conventions. The technical problem to be solved is how to provide for local variation without abrogating the standards that facilitate universal bibliographical control.

The most dramatic twentieth-century event to affect the organization of information is, of course, the computer revolution. It has changed the nature of the entities to be organized and the means of their organization. It has provided solutions to certain problems but spawned a host others. One of the new problems relates to the nature of digital documents. A traditional document, like a book, tends to be coincident with a discrete physical object. It has a clearly identifiable beginning and end; the information it contains — a play, novel, or dissertation — is delimited by these; it is "all of a piece."[38] By contrast, a digital document — such as a hypertext document or a connected e-mail message — can be unstable, dynamic, and without identifiable boundaries.

Documents with uncertain boundaries, which are ongoing, continually growing, or replacing parts of themselves, have identity problems. It is not possible to maintain identity through flux ("One cannot step twice into the same river").[39] A single frame is not representative of a moving picture. A snapshot cannot accurately describe information that is dynamic. This is not simply a philosophical matter, since what is difficult to identify is difficult to describe and therefore difficult to organize.

The oldest and most enduring source of problems that frustrate the work of bibliographic control is the language used in attempting to access information. In a perfectly orderly language, each thing has only one name, and one name is used to refer to each single thing. Philosophers and linguists

have idealized such languages. Leibniz, for instance, imagined a language so free from obscurities that two people involved in an argument might resolve their differences simply by saying "Let us calculate."[40] Such languages are artificial: they do not exist in nature. Natural languages are rife with ambiguities and redundancies; their robustness depends on these. But at the same time they cause problems when attempting to communicate with a retrieval system. It can happen, for instance, that a work is not found because it is known by several names and the user happens on the wrong one. Or a deluge of unwanted information may be retrieved because the user has entered a multivocal search term, one naming several different works, authors, or titles. It would seem that the most colossal labor of all involved in organizing information is that of having to construct an unambiguous language of description — a language that imposes system and method on natural language and at the same time allows users to find what they want by names they know.

2

Bibliographic Objectives

The first step in designing a bibliographic system is to state its objectives. Other design features — such as the entities, attributes, and relationships recognized by the system and the rules used to construct bibliographic descriptions — are warranted if and only if they contribute to the fulfillment of one or more of the objectives.

Traditional Objectives

Panizzi, writing in the middle of the nineteenth century, indirectly referenced bibliographic objectives when he argued in favor of the need for a catalog to bring together like items and differentiate among similar ones. It is Cutter, however, who in 1876 made the first explicit statement of the objectives of a bibliographic system.[1] According to Cutter, those objectives were

1. to enable a person to find a book of which either
   the author
   the title
   the subject
   } is known
2. to show what the library has
   by a given author
   on a given subject
   in a given kind of literature
3. to assist in the choice of a book
   as to its edition (bibliographically)
   as to its character (literary or topical).

Cutter formulated his objectives based on what the user needs and has in hand when coming to a catalog. The first objective, the *finding objective*, assumes a user has in hand author, title, or subject information and is

particularly useful for the emphasis it gives to what in the first instance is the primary act of information organization — bringing like things together. Both for its set-forming connotations and its ties to tradition it is too valuable to lose.

Also, in breaking with tradition, the first IFLA objective does not specify the sets of entities to be found but relegates this task to an accompanying entity-attribute-relationship model. This is problematic from a database design point of view. In the design of a database objectives should determine ontology and not vice versa, since for any given set of objectives, alternative models can be developed for alternative purposes. Moreover, a statement of objectives should embody a hypostatization of user needs. It should state just what it is that users need to find.

For the purposes of this book, the first IFLA objective will be amended to reintroduce the finding-collocation distinction, as follows:

1. To *locate* entities in a file or database as the result of a search using attributes or relationships of the entities:
1a. To find a singular entity — that is, a document (finding objective)
1b. To locate sets of entities representing
All documents belonging to the same work
All documents belonging to the same edition
All documents by a given author
All documents on a given subject
All documents defined by other criteria.[7]

Sufficiency of Objectives

Though objectives are postulated, they can still be evaluated insofar as they are intended to reflect user needs. They can be evaluated with respect to their sufficiency and necessity. A nontraditional position, one peculiar to this book, is that the four objectives as stated (to find, identify, select, and obtain) are in fact not sufficient. A fifth objective is needed — a *navigation objective*. Nearly half a century ago, Pierce Butler implied the existence of such an objective when he characterized *bibliography* as "the means by which civilized man navigates the bibliographic universe."[8] The metaphor is apt in its depiction of a user roaming from point A to point B and so on

to reach a destination — the desired document. The argument for explicitly recognizing a navigation objective has two parts: the first is drawn from research into users' information-seeking behavior, and the second from analyses of traditional codes for bibliographic description.

Some users come to a search for information knowing exactly what they want. But other users do not quite know or are unable to articulate the object of their search,[9] and yet they are able to recognize it immediately when they find it. Such users expect guidance. Bibliographic systems have traditionally met this expectation. An example is the guidance provided by a classification used to order books that are stored on the shelves of a library. Walking through library stacks (a microcosm of the bibliographic universe) and browsing, a user may suddenly come across just the right book and credit this luck to serendipity. But such a finding would be serendipitous only if the books were shelved in random order, whereas in fact they are ordered according to a rigorous system of semantic relationships, which like an invisible hand guides the seeker to his "lucky" find.

Another reason for postulating a navigation objective is that the bibliographic codes of rules used to organize documents assume its existence. Ideally, for each rule in a code, it should be possible to point to an objective that warrants it. Actual code construction, however, is frequently less than ideal, and rules sometimes are introduced in a Topsy-like fashion, without due regard to objectives. Many of these rules are unwarranted, but some actually have a legitimate purpose, in which case the objectives themselves can be questioned. Among rules with a legitimate purpose are those that establish bibliographic relationships. Such rules can be found in codes both for author-title description and for subject description. They include rules that specify relationships between works as well as relationships between names of work attributes, such as authors and subjects. Work-work relationships include generalization relationships *(is a subclass of)*, the aggregation relationships *(is a part of)*, and various associative relationships *(is a sequel to, is an adaptation of, is an abridgment of, is described by)*. Relationships among names of work attributes include equivalence, hierarchical, and associative relationships. The aim of the rules setting up these relationships is to map the bibliographic universe — that is, to facilitate navigation.[10]

Thus, a navigation objective has both user and code warrant. Such an objective might be formulated as follows:

• To *navigate* a bibliographic database (that is, to find works related to a given work by generalization, association, or aggregation; to find attributes related by equivalence, association, and hierarchy).

Objectives of a Full-Featured Bibliographic System

The IFLA objectives — modified to provide model independence, continuity with tradition, and a navigation objective — would read as follows:

• To *locate* entities in a file or database as the result of a search using attributes or relationships of the entities:
1a. To find a singular entity — that is, a document (finding objective)
1b. To locate sets of entities representing
All documents belonging to the same work
All documents belonging to the same edition
All documents by a given author
All documents on a given subject
All documents defined by "other" criteria;[11]

• To *identify* an entity (that is, to confirm that the entity described in a record corresponds to the entity sought or to distinguish between two or more entities with similar characteristics);

• To *select* an entity that is appropriate to the user's needs (that is, to choose an entity that meets the user's requirements with respect to content, physical format, and so on or to reject an entity as being inappropriate to the user's needs);

• To acquire or *obtain* access to the entity described (that is, to acquire an entity through purchase, loan, and so on or to access an entity electronically through an online connection to a remote computer);

• To *navigate* a bibliographic database (that is, to find works related to a given work by generalization, association, and aggregation; to find attributes related by equivalence, association, and hierarchy).

These objectives will be referred to, respectively, as the *finding*, *collocating*, *choice*, *acquisition*, and *navigation objectives*. Collectively they constitute the objectives of a full-featured bibliographic system. Though care has been taken in their formulation, they are still not without problems, as the following sections show.

Operationalization of Objectives

A subsidiary purpose of the bibliographic objectives is to specify the entities, attributes, and relationships required of a bibliographic system and to serve as instruments against which to vet system features. To achieve this purpose, they need to be operational — that is, they should be formulated in such a way that their achievement (or nonachievement) can be ascertained. The finding objective meets this requirement. Whether it is attained can be measured by ascertaining through a retrieval experiment whether the attributes used to describe the documents are sufficient to differentiate them.[12] Also measurable is attainment of the acquisition objective, which requires that data about the location and availability of a document be given. Attaining the other three objectives is more problematic, either because of the nature of their measurement or because, being open-ended, measurement cannot be completed.

The Collocating Objective

The collocating objective deserves special mention because of the composite nature of its measurement. This objective states that a bibliographic system should be capable of forming certain sets of bibliographic records. An attempt to measure it was initiated in the late 1950s in the landmark Cranfield experiment mentioned earlier. Cyril Cleverdon and his colleagues at Cranfield conducted this experiment to test the retrieval effectiveness of different methods for organizing documents. To measure effectiveness they developed a means to assess the set–forming power of a retrieval system.[13] They began with a *recall measure*, defined as the number of relevant records retrieved by the system divided by the total number of relevant records in the database. It was soon readily apparent that recall by itself was not a sufficient measure of collocating power, since even if no organizing intelligence at all were applied to structuring a database, 100 percent recall could be realized simply by sequentially examining every record in the database. Implied, but not explicitly stated in the formulation of the collocating objective, is that *only* relevant records should be brought together — that is, relevant records should not be intermixed with irrelevant ones. Collocation without discrimination is meaningless. Thus, the Cranfield team developed another measure, one that would assess the degree to which a bibliographic

system retrieves only relevant records. Called *precision*, it is defined as the percentage of retrieved records that are relevant. An ideal bibliographic system — one with full collocating and discriminating power — would retrieve all and only relevant documents. It would operate at 100 percent recall and 100 percent precision.

The precision and recall measures are not without problems. One problem is the difficulty in defining *relevance*, which is a key variable in their definition. Another is that the measure is a composite one and that in practice there is often a trade-off between collocation and discrimination. Each of these problems has generated a substantial body of thought and literature. Nevertheless, the measures have proved useful not only in evaluating how well a system achieves the collocation objective but also in testing system features (such as depth and breadth of indexing) in such a way as to generate lawlike statements about the impact of these on system effectiveness.

Originally applied to the evaluation of subject collocation, the precision and recall measures are useful as well to evaluate the set-forming power of other attributes, such as edition, author, and title. Until the early 1990s, there was little interest in applying them to this purpose, possibly because card catalogs were able to achieve reasonably good author and title collocation through the use of sophisticated filing rules. Online catalogs with their computer filing are a different matter, however. While they may retrieve records for all editions of a work and all works of an author, also retrieved is a horde of irrelevant records. Allyson Carlyle, who has studied how retrieval performance has deteriorated in the move from card to online catalogs, found that precision in the online display of records is very poor indeed. For popular works, such as More's *Utopia*, Joyce's *Ulysses*, and Shakespeare's *Sonnets*, it is less than 15 percent.[14]

Open-Ended Objectives

It is difficult to operationalize objectives that are open-ended. Take, for instance, the choice objective. As stated by Cutter, this objective specifies three ways in which a user should be assisted in choosing a book: by indicating its edition, its character, and its literary or topical nature. As stated in the IFLA document, the objective enjoins assistance in terms of "content, physical format, etc." The *etc.* is the rub. It could encompass hundreds of attributes of bibliographic entities and countless bibliographic relationships.

enduring diary of humankind, and they do not merit the bibliographic treatment reserved for documents deemed of lasting and scholarly interest, traditionally booklike objects.[18]

Indexing systems that are publicly funded are likely to confer more bibliographical control than those developed with private funds. Some systems, intentionally designed to be economical, postulate no more than a limited finding objective. No collocation is provided beyond what can be achieved by simple automatic operations applied to formal marks or character strings appearing on documents. Such systems cannot bring together a document represented as being by Mark Twain and another as being by Samuel Clemens. Bibliographic systems that rely for collocation on the automatic manipulation of character strings on documents, without attempting to interpret their meaning or to show relationships among them, are minimally featured systems. Keyword systems are of this type. While they are often useful for accessing information, they lack the retrieval power of systems in which bibliographic data are intelligently interpreted and organized through set formation and differentiation.

Most indexing systems occupy some middle ground between being minimally and fully featured. Few attempt to bring together all the editions of a work, but this is of little consequence, since the kind of documents organized by indexes tend not to be multiply manifested. Most seek to represent authors' names, as well as names of corporate bodies and places, in a uniform manner. As to subject collocation, while some rely solely on keywords, others, such as the systems created by the National Aeronautics and Space Administration (NASA) and by the National Library of Medicine and Chemical Abstracts, are very sophisticated in their set-forming capabilities — indeed, significantly more advanced than library catalogs.

In contrast to traditional indexes and catalogs are the new bibliographic systems that are being created to deal with documents on the Internet. Internet documents clearly vary in the degree of control they deserve. Many are of intellectual and lasting significance and warrant being archived for posterity, with the full bibliographic treatment this implies. Others are of an ephemeral nature, and for them the keyword access provided by low-end search engines is all that is needed. In between is a large class of documents whose bibliographical control is to be decided. This is presently the locus for innovation and experimentation.

One of the more popular systems created to deal with introducing order into the Internet is the Dublin Core. Developed at the Online Computer Library Center (OCLC) and promoted in a series of workshops, the Dublin Core provides a form of bibliographic control midway between cataloging and indexing. It differs from cataloging primarily in using many fewer metadata — thirteen as compared to several hundred. It differs also in the agents who provide bibliographic descriptions, these being not professional catalogers but frequently document authors and casual indexers. Finally, it differs in using a laxer form of vocabulary control, insofar as the use of any given metadata element, such as subject, may or may not promote collocation and differentiation depending on whether the indexer chooses to use values from a controlled vocabulary. Thus, the degree to which the Dublin Core when applied achieves the bibliographic objectives is as yet unpredictable.[19]

The rise of the Internet is affecting the actual work of organizing information by shifting it from a relatively few professional indexers and catalogers to the populace at large. In other words, this work is becoming deprofessionalized. Anyone and everyone can set up a website and organize information. The organization effected by nonprofessionals is often free and (it cannot be denied) effective. To the extent that the bibliographic universe can be organized by keyword access and beyond that by the voluntary efforts of individuals who mount information on the Web, it is self-organizing. While not consciously teleological, a self-organizing bibliographical universe nevertheless succeeds in meeting the bibliographic objectives in part, occasionally, and somewhat randomly. And for many documents and many users this is all that is needed.

An important question today is whether the bibliographic universe can be organized both intelligently (that is, to meet the traditional bibliographic objectives) and automatically. The question is important because of the ever-present danger that objectives will be sacrificed because of their cost. Can automation come to the rescue? Succeeding chapters address this question. Presently semantic barriers frustrate attempts to extend automation beyond keyword capabilities to incorporate the intellectual techniques required for collocation and differentiation. Yet the future may see the eventual creation of linguistic structures that can be used to break through these barriers, at least in part.

Implementation of Objectives in Future Systems

The question raised at the beginning of this section was whether full-featured bibliographic systems, ones that attempt to fulfill the traditional five objectives, were necessary. In an ideal world no one would question the desirability of such systems, but in the real world of economic exigency other considerations apply. In the following paragraphs, some of the traditional arguments relating to the necessity of the bibliographic objectives are presented — first those that question the necessity and then those that support it.

Of the several arguments put forward for the use of a less than full-featured system, the most frequent and the most persuasive is the cost argument. Particularly costly are the system features needed to fulfill the collocating objective. To supply these means going beyond the simple, clerical task of transcribing attributes of a document to address the not-so-simple intellectual task of ascertaining whether the document is known by more than one title or if its author has written under different names. Any task that requires an organizing intelligence to engage in research is costly. Also, as noted earlier, fulfilling the open-ended objectives is costly, requiring seemingly bottomless pockets.

Another argument favoring a less than full-featured system is user-based. A number of experimental studies have shown that often users neither need, nor are capable of exploiting, the power of a highly organized database.[20] One of the most frequently cited is a second experiment that was conducted at Cranfield, which found that an index language designed to provide only partial subject collocation satisfied users quite as well (measured in terms of precision and recall) as one that provided full collocation.[21] Similar experiments have been performed by different researchers in different environments, some with comparable and others with conflicting results.[22] Another type of experiment aiming to show that users do not require full-featured systems was performed by Alan Seal at Bath.[23] Seal attempted to assess the value of various data elements (not just subjects) used in traditional bibliographic descriptions. To this end he set up two parallel catalogs — a traditional one consisting of records with full descriptions ("full entries") and an experimental catalog consisting of short entries.[24] Observing the use of these catalogs over a two-month period, he found a failure rate for the short-entry catalog of only 8 percent, where *failure* was