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.

Historical Background

The relevant historical background is the tradition of Anglo-American descriptive and subject cataloging during the last century and a half. While some form of systematic information organization has been practiced since 2000 B.C.E.,[1] its modern history is usually regarded as beginning in the middle of the last century with Sir Anthony Panizzi's plan for organizing books in the British Library.[2] In the period following Panizzi, the groundwork was laid for the major bibliographic[3] systems in use in libraries today: the Dewey Decimal Classification (DDC), the Library of Congress Classification (LCC), the Universal Decimal Classification (UDC), the Library of Congress Subject Headings (LCSH), and the Anglo-American Cataloguing Rules (AACR). Though strong, particularly in their ideologies, these systems were jolted in the twentieth century by information explosions, the computer revolution, the proliferation of new media, and the drive toward universal bibliographic control. How they have withstood these jolts, where they have remained firm, where they have cracked, and where cracked how they have been repaired or still await repair is a dramatic — and instructive — history for those interested in organizing information intelligently.

Santayana wrote that "when experience is not retained . . . infancy is perpetual. Those who cannot remember the past are condemned to repeat it."[4] To be so condemned would not be all bad, since reinventing what has been done in different times and circumstances reinvigorates a discipline, rids it of routinized procedures and ways of thinking, and energizes it by the influx of new ideas and new terminology. Nevertheless it is instructive — especially given the recent interest and activity directed toward organizing digital information — to understand certain features of traditional bibliographic systems. Two features in particular are worth considering. One is the solutions these systems have provided to the problems that obstruct efficient access to information. While today some access problems are caused by the new technology, others — such as those that stem from the variety of information, the many faces of its users, and the anomalies that characterize the language of retrieval — have been around a long time. For instance, whether users search library shelves or the Internet, some will retrieve too much, some too little, and some will be unable to formulate adequate search requests. The thought that has gone into addressing problems like these, cumulated over a century and a half — particularly the thinking that deals with rationales for why things are done as they are — provides, independent of time and place, an informed context for systems design.

A second feature that makes traditional bibliographic systems worthy of continued interest is the vision expressed in their ideologies. A system's effectiveness in organizing information is in part a function of an ideology that states the ambitions of its creators and what they hope to achieve. The systems produced during the second half of the nineteenth century, a period regarded as a golden age of organizational activity,[5] were ambitious, full-featured systems designed to meet the needs of the most demanding users. Some would argue that they were too ambitious — that there was no need to construct elaborate Victorian edifices since jerrybuilt systems could meet the needs of most users most of the time.[6] However, good systems design begins by postulating visionary goals, if only to make users aware of the extent to which compromises are being made. The bibliographic systems of the past (in their ideologies, at least) reflect what can be achieved by intelligent information organization.

Philosophical Background

Relevant to the intellectual foundation of information organization are the points of view embraced by three philosophical movements that have permeated academic and popular thinking during the twentieth century: systems philosophy, the philosophy of science, and language philosophy.

Systems Philosophy

A philosophy of ancient origin, general systems theory was resurrected by Ludwig von Bertalanffy in the mid-twentieth century in an attempt to stopgap what he perceived to be an increasing fragmentation of knowledge.[7] General systems theory is a philosophical expression of holistic or big-picture thinking. Its credo encompasses a belief in purpose as opposed to chance processes, a way of looking at phenomena in terms of their organization and structure, and a conviction that general laws and principles underlie all phenomena. From this philosophy derives the practice of systems analysis, which in its most general form is the analysis of an object of study, based on viewing it as a system whose various parts are integrated into a coherent whole for the purpose of achieving certain objectives.

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 analysis 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 consonant 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

writing.[29] Julia Pettee in 1936 formulated the distinction explicitly, referring to a particular message content as a *literary unit* and its embodiment in a medium as a *book.*[30] In 1955 S. R. Ranganathan introduced the distinction, presenting it as the dichotomy between expressed thought and embodied thought: the former he referred to as a *work,* the latter as a *document.*[31] In the 1960s, the significance of the distinction was brought to popular attention as a result of Seymour Lubetzky's eloquent juxtaposition of the work versus the book.[32] He regarded a work as the intellectual creation of an author. A work is what in the preceding paragraphs has been characterized as (1) information, (2) the disembodied content of a message, or (3) expressed thought. It is a kind of Platonic object. A book, by contrast, is a particular physical object that embodies or manifests the work. One work can be manifested in many physical objects, and, conversely, one physical object can manifest several works.

Because of its centrality, the distinction between information and its embodiments has invited terminological confusion in the form of synonyms and near synonyms. *Literary unit, (message) content, expressed thought,* and *text* have been used either coextensively or as operationalizations of *work. Manifestation, expression, edition, version, publication,* and *carrier* have been used somewhat ambiguously to refer either to a slightly altered form of an original work, to its physical embodiment, or to both. In this book, *work* is used in the Ranganathan and Lubetzkian sense to indicate a particular disembodied information content. Ranganathan's term *document,* rather than Lubetzky's *book,* is used to indicate a material embodiment of information — at least for the most part. Exceptions are made when citing the literature and introducing further distinctions.

Purposes, Principles, and Problems

In 1674 in the Preface to the *Catalogue for the Bodleian Library,* Sir Thomas Hyde lamented the lack of understanding shown by those who never had the opportunity to make a catalog:

> "What can be more easy (those lacking understanding say), having looked at the title-pages than to write down the titles?" But these inexperienced people, who think making an index of their own few private books a pleasant task of a week or two, have no conception of the difficulties that rise or realize how carefully each book must be examined when the library numbers myriads of volumes. In the

colossal labor, which exhausts both body and soul, of making into a alphabetical catalog a multitude of books gathered from every corner of the earth there are many intricate and difficult problems that torture the mind.[33]

Three centuries and many myriads of "books" later, the problems that torture the mind when attempting to organize information have increased exponentially. It has never been easy to explain why colossal labor should be needed to organize information. If not the most successful, at least the most passionate attempt to do so was made by Panizzi when before a Royal Commission he defended his plan for organizing books in the British Library (1847–1849). Many members of the Commission did not understand the plan and, not understanding it, found it too complicated. The most celebrated of the commissioners, Thomas Carlyle, went so far as to accuse Panizzi of trying to enhance his reputation by building a catalog that was "a vanity of bibliographical display."[34] And this despite his reputation as a leading intellect of the time.

Organizing information would seem to be no different from organizing anything else. The assumption that this is the case has led to attempts to interpret it as a routine application of the database modeling techniques developed to organize entities like the employees, departments, and projects of a company. But there are important differences. One that is particularly important, because it is at the root of many of the complexities unique to organizing information, is that two distinct entities need to be organized in tandem and with respect to each other: works and the documents that embody them.

Organization can take many forms. Its prototypical form is classification. Classification brings like things together. In traditional classifications, like things are brought together with respect to one or more specified attributes. Any number of attributes can be used to form classes of documents embodying information, such as same size or color, same subject, or same author. However, the most important attribute for a system whose objective is to organize information is the attribute of "embodying the same work." No other attribute can match it in collocating power because documents that share this attribute contain essentially the same information. Organizing information if it means nothing else means bringing all the same information together.

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