Thinking About Museum Information

PATRICIA ANN REED AND JANE SLEDGE

INTRODUCTION

IMAGINE CATALOGING *without* AACR2 (Anglo-American Cataloguing Rules, 2nd ed.), Library of Congress Name Authority, and Library of Congress Subject Headings. Welcome to the world of museum cataloging. This article discusses work in progress at the Smithsonian Institution (SI) in developing a system to understand and to articulate the information needed to support collections-related functions.

During the course of this work four major points emerged: (1) technology is not the answer to information problems; (2) a structured process of information analysis is essential to the understanding of information requirements; (3) the structured process requires allocation of scarce resources—people, money, and time; and (4) the resources expended on the structured process to analyze museum functions and data yield significant benefits that pay off in the design and implementation of systems.

In the spring of 1987, members of the "art community" began a structured process of information analysis to develop a graphic model of art data. (The "art community" is a group of seven Smithsonian museums with major art collections: The National Museum of American Art, the National Portrait Gallery, the Hirschhorn Museum and Sculpture Garden, the Freer Gallery of Art, the Arthur M. Sackler Gallery, the National Museum of African Art, and the Cooper-Hewitt Museum of Design and Decorative Arts.) What follows is a description of the use of a structured methodology and a progress report on some of the insights gained as a result of this work. Although the Smithsonian environment

Patricia Ann Reed, Office of Information Resource Management, Smithsonian Institution, Washington, DC 20560
Jane Sledge, Office of Information Resource Management, Smithsonian Institution, Washington, DC 20560
LIBRARY TRENDS, Vol. 37, No. 2, Fall 1988, pp. 220-31

may be unique because of the size and diversity of its collections, the functional analysis and data modeling methodologies provide insights applicable to other museums.

Smithsonian Background

Information processing at SI has been characterized by computer applications developed in response to individual or departmental needs. Standardization and exchange of data across departmental boundaries were not immediate goals. Each division of each department of the SI museums assumed responsibility for, and separately developed, its own data standards.

SELGEM (SELf-GEnerating Master), a computer system developed for museum collections by the Smithsonian Institution in the early seventies, served well; but it relied on magnetic tape technology, batch processing, and hard copy output for which Smithsonian staff often had to wait as long as two weeks. These constraints severely limited the ability to maintain an inventory for collections of objects numbering in the hundred millions; although when the U.S. Congress mandated and funded a complete collections' inventory, SELGEM was the only available repository for the inventory data.

Museum staff and data processing personnel agreed that there must be a better way. The better way for the Smithsonian, the Collections Information System (CIS) (an IBM 4381—a mainframe running the VM/CMS operating system—and Infodata's INQUIRE—a text-oriented database package), is formulated upon the realization that while individual solutions may be practical for the short term, they are less effective in the long term. The understanding that the Smithsonian is a community, where neighbors have common interests, precipitated a search for solutions beyond traditional hardware and software technology.

Data Administration and Information Architecture Project

In the spring of 1985, the Office of Information Resource Management, the Smithsonian's computer services department, staffed a data administration function to define and manage the Smithsonian's data as an institutional resource. Data administration proposed a comprehensive approach to systems development using an Information Architecture project to analyze and define both the functions performed at the Smithsonian and the data required to support those functions. (In April 1986, a Request for Proposal was released to acquire a methodology for building the Information Architecture. The contract was awarded to Technology Information Products (TIP), Wakefield, Massachusetts.) The project will produce a blueprint for the integration of all Smithsonian information systems. The project has two phases: Phase I will define and analyze functions and Phase II will define and analyze data.

Phase I

Phase I (functional analysis) identifies and defines the major activities of the Smithsonian and the staff doing the work. The Collections Information System will support information needed by collections management, research, and public programs. Museum collections management work requires an information system that supports the acquisition of objects, title transfer, shipping, object tracking, conservation, maintenance of collections documentation, and much more.

Figures 1 and 2 are examples of functional analysis blueprints which document workshop discussions about "Plan Collections Acquisitions," part of the broad function "Manage Collections." Figure 1 illustrates the component activities of "Plan Collections Acquisitions."

The Smithsonian collects objects both opportunistically and with predetermined intent. Not all objects offered to the Smithsonian are accepted, while other objects are purchased or solicited. Rejections and acquisitions are based on established criteria and are dependent on functions outside of the "Plan Collections Acquisitions" process. Figure 2 supports Figure 1 by showing other functions and information resources that contribute to "Plan Collections Acquisitions." Other functions, illustrated in detail on separate diagrams, that contribute information or criteria to Figure 2 are "Define Collections Policies," "Evaluate Research Possibilities," and "Identify and Select Objects." These other functions, represented in Figure 2 by three-sided boxes, send information to and receive information from "Plan Collections Acquisitions."

External factors also influence Smithsonian collecting. Collections are offered to the Smithsonian for acquisition, or government agencies are legislated to transfer collections to the Smithsonian. External factors, beyond the Smithsonian's direct control, are illustrated by a three-sided box with a bar. The type of information sent to and received from the internal and external functions is recorded on connecting lines with arrows showing the direction of the information flow.

Information is also received from and sent to "information stores." These are illustrated in Figure 2 with the name of the "information store" held between two parallel lines. The stores may be physical collections of objects, filing systems, computer systems, staff expertise and knowledge, Smithsonian policy, etc. Again, information sent from the "stores" to the function is recorded on the connecting lines with arrows showing the direction of the information flow.

Phase II

Phase II (data analysis) identifies information needed throughout the Smithsonian Institution and uncovers the relationships among sets of information. It employs a rigorous data modeling process that focuses on data rather than function. "Data modeling is the process of trying to 'uncover' the natural structure and meaning of data required

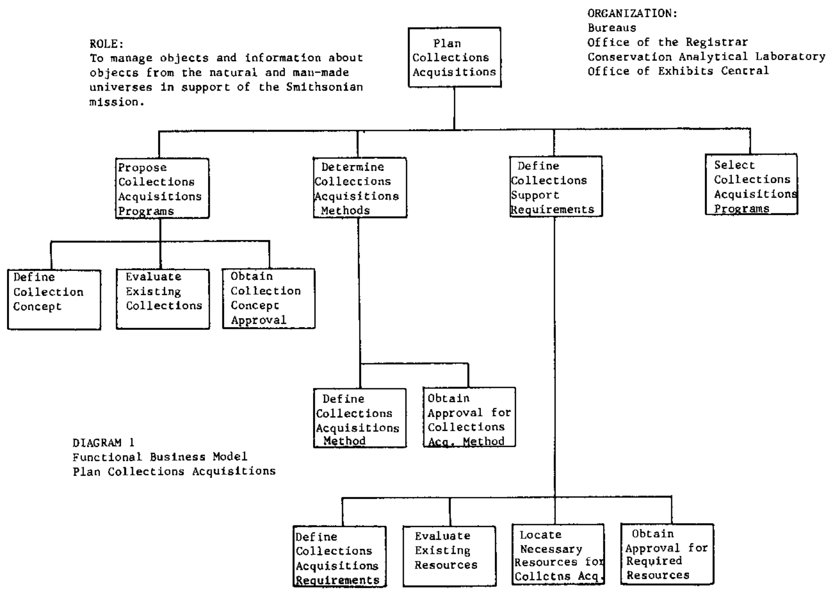


Figure 1. Functional business model plan collections acquisitions

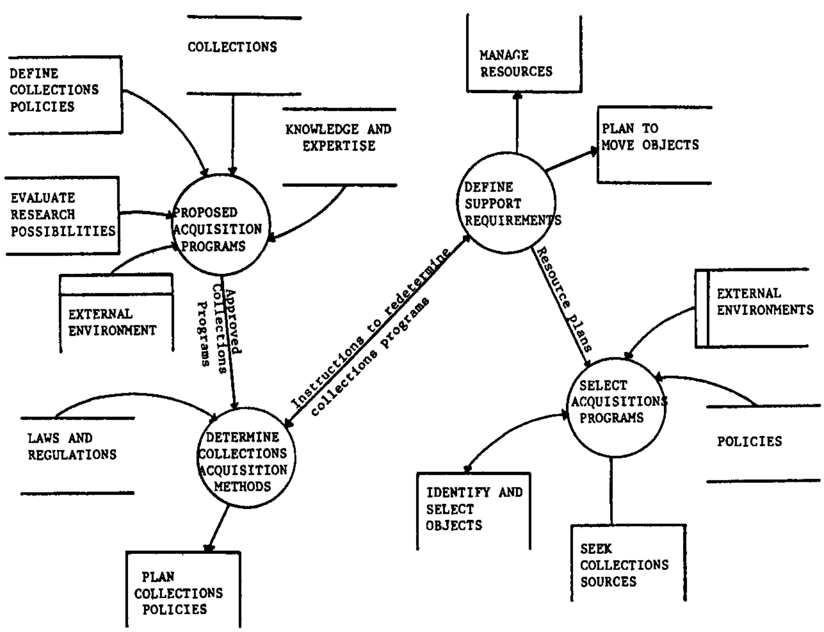


Figure 2. Simplified information usage model plan collections acquisitions

by the entire organization...a data model describes the inherent relationships of the data within a business rather than how data is currently used or will be used in the future" (Technology Information Products 1985, p. 3).

An important distinction separating the data modeling phase from the functional analysis phase of the Information Architecture project is that in the data modeling effort there is no orientation to actions. The diagrams say nothing about who does what with what information. There is no flow of information; the information is at rest. This separate study of data, without consideration of functions or automated systems, reveals the structure of the data to be used by the new Collections Information System.

Data Elements

The first step in Phase II (data analysis) began before the Information Architecture project started. Existing SELGEM data, manual records, and new data needed for the Collections Information System were defined to the data element level. This meant that each piece of data was defined and separated into component parts. For example, "Artist Name and Life Dates" was stored in SELGEM as a text field. During the process of data analysis, the data elements in "Artist Name and Life Dates" were identified as LAST NAME, FIRST NAME, MIDDLE NAME, BIRTH DATE, and DEATH DATE. Also defined was a data element called DATE QUALIFIER which holds values such as "?," "CIRCA," "BEFORE."

Teams of museum staff (curators, registrars, librarians, archivists, etc.) and data administration staff met to clarify each data element. The effort required staff to question the meaning of their data. For example, three different people defined DATE OF ACQUISITION, DATE OF ACCESSION, and DATE DONATED. Through discussion they realized that four different concepts were represented. At the Smithsonian, acquisition is different from accession as the Smithsonian may acquire objects it does not accession. Accessioning implies acceptance of additional responsibility for long-term maintenance and care required by the public trust. Donation implies an acquisition or accession by method of a gift which in turn can be a bequest. These actions occur on a particular date. After each data element is defined, staff in the Smithsonian museums understand its meaning.

The Smithsonian uses an automated data dictionary to document: the data dictionary name for each data element; the data element definition; the format of the data element; the length of the data element; the number of occurrences of the data element; and the users of the data element. The definition of data elements continues throughout the data modeling effort. The data dictionary is updated to reflect new insights on the organization and relationships of data elements.

Entities-Groupings of Data Elements. The next step in Phase II (data analysis), groups data elements into entities. An important concept is that of the primary entity. A primary entity is a person, place, thing, concept, or event which exists independently and about which the Smithsonian keeps information. The data elements that describe or define an entity are placed in a Logical Data Group (LDG) that represents the entity. Each LDG for a primary entity contains a data element, called the primary key, which uniquely identifies each occurrence of the entity.

An example of a primary entity represented in an LDG is an object. The data elements that define and describe an object are identified and grouped into an LDG called OBJECTS. Some of the data elements placed in OBJECTS describe an object's size, color, storage location, accessibility, and credit line.

Another primary entity considered by the teams seemed to be PEOPLE. In looking at the data elements which describe people, a tendency to confuse data values (the contents of fields) with data elements (the names of fields) became apparent. For example, some of the data elements originally defined included:

NAME OF ENGRAVER
NAME OF ARTIST
NAME OF CREATOR
NAME OF DONOR
NAME OF BORROWER

Analysis showed that the same data elements describe and define any individual, regardless of what he or she does, and a new concept emerged. There are two primary entities—ROLE PLAYERS and ROLES. The Logical Data Group for the primary entity called ROLES contains the names of the roles that role players can play such as engraver, artist, creator, donor, and borrower.

Data Relationships-Intersection Entities. When the primary key of a ROLE PLAYER is combined with the primary key of a ROLE, an intersection entity—ROLE PLAYERS & ROLES—is created. One primary entity is now associated with another primary entity forming a meaningful data relationship. This structure, illustrated in Figure 3, has many advantages. An important space-saving advantage is that information about a particular individual occurs only once in the ROLE PLAYERS file. The name of a role occurs only once in the ROLES file. A ROLE PLAYER entity can link to one or more roles as often as necessary by combining the primary key of a particular ROLE PLAYER with the primary key of a particular ROLE. These links appear in the intersection entity ROLE PLAYERS & ROLES. The intersection entity ROLE PLAYERS & ROLES contains additional pertinent data elements such as the begin and end dates during which the role player played the role. Another advantage is that additional roles can be added to the ROLES file at any time—data values are expandable.

The intersection entity of ROLE PLAYERS & ROLES can be combined with the primary key from the OBJECTS entity to form a more complex intersection entity called ROLE PLAYERS & ROLES & OBJECTS, illustrated in Figure 4. Role player *A*, for example, can be the creator, owner, or donor of object *A*, object *B*, or object *C*. The intersection entity documents these relationships.

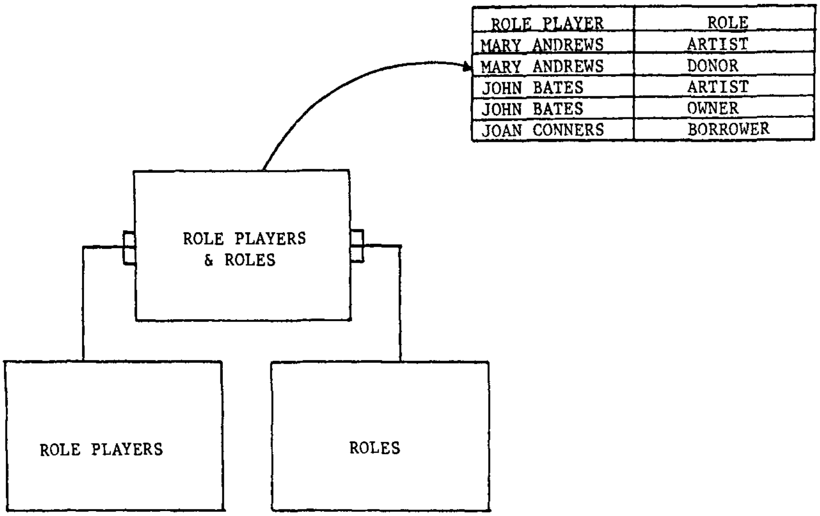
A problem arose when the realization came that there are primary entities for persons (which we have been calling ROLE PLAYERS), organizations, and culture groups. Each of these is a separate entity because the data elements kept within them are different. For example, an organization may have specific product or brand name associations, while a cultural group may have hierarchical affiliations to other cultural groups.

Each of these entities, however, can play many of the same roles. A person, an organization, and a culture group can create an object and play the role of creator.

Two solutions emerged for this problem. First, the primary entity ROLE PLAYERS became three primary entities—ROLE PLAYERS-PERSONS, ROLE PLAYERS-ORGANIZATIONS, and ROLE PLAYERS-CULTURE GROUPS. Second, a new data element—ROLE PLAYER TYPE (person, organization, or culture group)—was added for each occurrence of the ROLE PLAYERS & ROLES intersection entity to show whether the role player is a person, an organization, or a culture group.

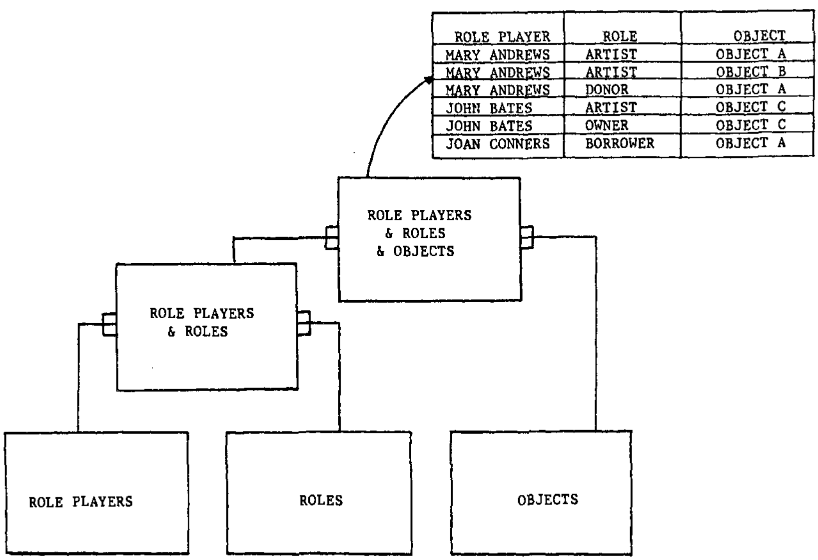
Besides primary and intersection entities, the methodology defines other kinds of entities. A type entity contains data for a primary entity that does not apply to *all* occurrences of the primary entity. Data documenting objects in the art collections differ from data documenting objects (specimens) in the natural history collections. The primary entity OBJECTS remains, but the Logical Data group representing it contains only the data elements that describe both man-made and natural history objects. OBJECTS-MAN-MADE is a type entity containing data elements used only for man-made objects, while OBJECTS-NATURAL is a type entity containing information used only for naturally occurring specimens. Within these broad types, there are subtypes. Certain data elements are needed for textiles or apparel and not for paintings, or for fish and not for minerals.

A repeating entity contains data that repeat for any given occurrence. For example, a single man-made object can have many marks. The data elements that describe a mark become a repeating entity represented by a Logical Data Group called OBJECTS-MAN-MADE-MARKS. Data elements for this entity are type of mark, material of mark, method of application, language of mark, alphabet of mark, text



ROLE PLAYER	ROLE
MARY ANDREWS	ARTIST
MARY ANDREWS	DONOR
JOHN BATES	ARTIST
JOHN BATES	OWNER
JOAN CONNERS	BORROWER

Figure 3. Data relationships—intersection entities



ROLE PLAYER	ROLE	OBJECT
MARY ANDREWS	ARTIST	OBJECT A
MARY ANDREWS	ARTIST	OBJECT B
MARY ANDREWS	DONOR	OBJECT A
JOHN BATES	ARTIST	OBJECT C
JOHN BATES	OWNER	OBJECT C
JOAN CONNERS	BORROWER	OBJECT A

Figure 4. Elements of the complex intersection entity ROLE PLAYERS & ROLES & OBJECTS

of mark, and location of mark. Another example of a repeating entity is OBJECTS-DIMENSIONS. Dimensions are often recorded in both inches and centimeters.

The Data Model

The purpose of the data model, which resembles an electrical diagram, is to represent visually the data relationships and dependencies discovered during the data analysis process.

The model is a series of named boxes, each of which represents a Logical Data Group. Each box connects to another box with a single line. The line leading into a box may have a trident that appears to "plug" into another box. The presence or absence of the trident shows whether there is a one-to-one, one-to-many, or many-to-many relationship. In Figure 3, both ROLE PLAYERS and ROLES have a one-to-many relationship with the role PLAYERS & ROLES box, meaning that one role or ONE ROLE player may occur in the intersection box many times.

When the data modeling process is complete, the data elements needed for the Collections Information System will be represented in specific entities, and entity relationships will be shown. There will be too many entities to model in an intelligible fashion on a single sheet of paper. A high-level conceptual model will show the major primary entities and related intersection entities. The complexities created by repeating and type entities will be shown in separate models. A model will be drawn for each primary entity to show its type and repeating entities. Models will also be drawn to group together all entities needed for a particular function such as "Plan Collections Acquisitions."

Finally, the Information Architecture project will merge the results of Phase I (functional analysis) and Phase II (data analysis) to create matrices that show the relationships between data and functions. The matrices will show which functions create what data, which functions use what data, and which functions send changes to the database. The matrices will help to establish priorities for automation by showing what functions produce the information needed to automate other functions.

The goal of data modeling is to develop systems that are data driven rather than process driven. Processes are subject to change while data tend to be constant. Staff change jobs, organizations reorganize, and technology advances. In contrast, the data collected at the Smithsonian will remain essentially stable. Corrections may change data content, areas of interest may expand to support new research, and data may be used in new ways to support new activities; but the base-level data are not expected to change radically. Object and role player information will remain essential to the Collections Information System.

The data model for the Collections Information System is the ideal or utopian view of the data, independent of hardware or software considerations. It is important to realize that there is a distinction between the data model (the ideal) and the system implementation (the reality). Hardware, software, time, staff, and financial resources will impact upon the system implementation. The model is the blueprint for the ideal reconfiguration and migration of SELGEM data. In translating the logical model (ideal) into the physical design (reality), the system design and development team will need to make tradeoffs and compromises to accommodate system limitations and resource constraints. The blueprint enables the system design team to preserve the data relationships and ensures that, as the system expands, all the pieces fit.

Lessons

The most astounding "whack on the head" to date is the discovery that the rigorous application of the Information Architecture methodology in Phase II (data analysis) produces a view of the data that is different in kind rather than degree. The difference in kind begins with the ability to distinguish clearly between data values (artist, donor, etc.) and data elements (i.e., NAM-ROLE) and continues with the precise definition of data elements. This creates the groundwork needed to build Logical Data Groups which relate data in new ways, providing increased flexibility and freedom to reflect the complexity of museum information.

Museum information consists of complex structures of related data groups. Data analysis shows that, while museum information may be lengthy, it is much more than paragraphs of descriptive text. Data analysis identifies and names the ideas embedded in language. The naming of concepts (such as ROLE PLAYERS, ROLES, and OBJECTS) provides the ability to associate one concept with another in a multidimensional fashion. For example, bibliographic references usually refer to the accession or catalog record as a whole rather than to specific data groups. As illustrated in Figure 5, the model attaches references by their primary keys precisely to the data groups referenced—people, places, events, concepts, and objects.

When defining data elements to build the data dictionary, museum staff exhibit a very human desire to continue cataloging traditions. Many of these traditions are implicit rather than explicit in nature. Days were spent discussing such things as object-part relationships; related objects; subject matter; classification systems; geographic naming conventions (What is a region?); multiple artist attributions; calendar schemes (Aztec, Islamic, Jewish, and Chinese); and relative time scales (eras, periods, dynasties). Data analysis provides a forum to question practices found within museum cataloging. Are labels marks? Can there be more than a single alphabet in an inscription, and how is this

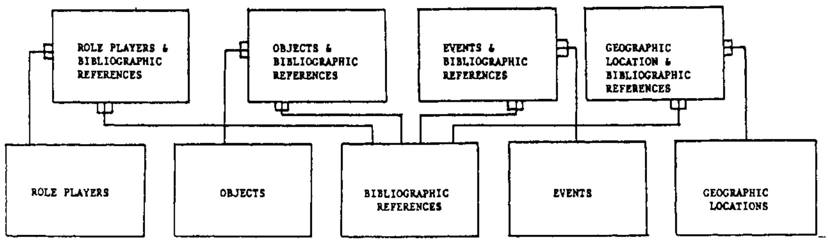


Figure 5. Model attaches references by primary keys

handled? What is the difference between decor and motif? What information is in classification systems and why? When is a photograph an object, and when is it documentary material?

Much of the success of the project is attributable to the use of graphic communication tools. The process used in the project, including the diagramming techniques and symbols, proves to be an excellent mechanism for promoting discussion. The old adage that "a picture is worth a thousand words" holds true.

One of the greatest rewards occurred during the 1987 annual meeting of the International Council of Museums' Documentation Committee. The Data Standards Subcommittee began to develop a data model by combining information produced by (among others) the Museums Documentation Association—United Kingdom, the Smithsonian Institution, the British Museum, the Victoria and Albert Museum, the National Museum of Ethnography in Sweden, and the National Museum of Civilization and the National Museum of Natural Science, Canada. Committee members compared data models developed within individual museums. The commonalities were startling. The subcommittee decided that the development of an international standard for museum information was an achievable goal and committed itself to the sharing, analysis, and integration of existing models. A representative of the International Standards Organization attending the meeting expressed interest and support for the project.

CONCLUSION

Museum staff involved with the development of the Collections Information System are pioneers of the Smithsonian's Information Architecture project. As other areas of the Smithsonian using or developing automated systems—such as personnel, finance, facilities management, libraries and archives, and security—participate in the Information Architecture project, greater benefits will be realized.

Functional analysis provides a means of establishing links and ties between different areas of the Smithsonian. Information systems are an important component in the synergy of the Smithsonian. Many areas require access to data to support work in progress and in turn generate information of interest to others.

Data modeling provides a means of standing back and examining closely-held ideas about the way museum information works while studying the realities of the data. We are beginning to relinquish our preconceptions about the way data "must be." Instead, we are on the way to understanding the reality of what information is and, equally important, how it can be structured and stored to serve our many needs.

Reference

Technology Information Projects. 1985. *TIP PLAN Reference.* Wakefield, MA: TIP.