Standardization Requirements of a National Program for Information Transfer

RONALD L. WIGINGTON AND JAMES L. WOOD

A National program for information transfer must consider both the functions that various groups perform and the system that their numerous interrelated, but somewhat independent, activities constitute. Automation of many of these functions is widely regarded as the key for achieving significant advances over the present, predominantly manual, national system for organizing and disseminating scientific, technical, and other scholarly information. Automation of any activity requires that the function of each part and the multiple interaction between parts be brought into sharp focus. Usually, the first lesson learned in any automation effort is that even though the target activity may have been operating with some degree of success in the past, it is still not sufficiently well understood to enable the listing of complete and accurate requirements to guide the specification of the machine system. As a result, desired improvements are elusive and slow to develop.

The authors of this article represent two specializations in the spectrum of information transfer activities, that of the computer and communication system engineer and that of the librarian. These points of view are combined in examining the requirements for standardization in the national efforts to use automation in publication, library, abstracting and indexing, and information-retrieval activities. Standardization is necessary to both the representation of information and to the procedures being developed for handling it.

The details of standardization represent a major part of over-all system design. Underspecification risks operational failure, or at least it endangers over-all operational efficiency and economy. Conversely, overspecification can impede future improvements and extensions and can raise serious practical obstacles to cooperation.

Ronald L. Wigington and James L. Wood are with the Chemical Abstracts Service, Columbus, Ohio.
Standardization Requirements

The concepts of "network" and "service utility" have often been referred to in conceptualizing a national information transfer system. Frequently it is useful to consider analogies as a source for design guidance or insight. The national telephone system was one of the earliest technologically-based networks to be subjected to scientific analysis and design, and in addition, it is one form of an information-transfer system. Standardization has played a key role in its successful development and operation.

A major characteristic of the development of the national telephone network, and one that offers a valid analogy for guidance in considering the development of a national information transfer network, is that improvements are brought about by evolutionary change. Evolution in turn involves a succession of standards for each part of the system, and compatibility with the rest of the system is an essential requirement for any new part. When the investment in an existing system is very large, that system cannot be replaced or revolutionized very quickly; it must be changed gradually as the economic and human resources are available to do so. The standardization of what the system handles, and the procedures for doing it, must change with that evolution. The standardization must also be effected with great care both to gain sound current operations and to facilitate future change.

The telephone system analogy can be used to illustrate some of these points. Prior to the introduction of nationwide long-distance dialing, several different formats of telephone numbers existed, several different kinds of switching equipment were used, and correspondingly, different methods for signalling that equipment were employed in local telephone systems. Much manual intervention was necessary to establish a long-distance connection, and during periods of peak usage, performance was often poor. Before the installation of the nationwide system, a standardized method of expressing telephone numbers was established. It provided a concise, unique identification of the terminal stations in the network. The information to be transferred from one point to another in the switching network was that unique identification.

Even though the initial handling facilities did not represent the information identically, ways were worked out to translate the form of that information and pass it across the internal system boundaries. And although manual intervention was used as necessary to make the total system function, such intervention was greatly reduced compared with the former entirely manual system. As new equipment has been i-
stalled over a long period of time, greater homogeneity in standards of representation and procedure has been achieved, and performance and economy have improved. Still, some heterogeneity remains, and always will, because of continuous system change. At all times, however, the unambiguous information content of the switching control signals had to be preserved and it had to be capable of being transformed into any form needed for any system function handled by man or by machine.

Standardization is far from being a new subject in information handling. For example, that segment of information handling concerned with library science contains many instances of uniform practice established to enable the results of work done once to be used effectively by many and to improve uniformity of the interface to the library user.

A good example of this type of standardization is the descriptive cataloging on Library of Congress catalog cards. During 1968, the Library of Congress sold over 78 million cards to approximately 25,000 libraries, firms, and individuals. By accepting in part, if not in whole, the Library of Congress’ descriptive cataloging, classification, and subject headings, the library community saved millions of dollars, and at the same time provided standardized access points to its book collections. In fact, the 3 inch by 5 inch (7½ by 12½ cm.) card is itself an excellent example of library standardization.

In 1945, Wilson and Tauber in their classic work, The University Library, noted that “the progress of a profession is usually marked by the cumulation of an increasing number of generally accepted practices.” In reviewing the number of “generally accepted practices” used by the library community, it might even seem that no additional “standards” would be required for the part of the national program that involves library automation. Unfortunately, this is not the case.

First, there is a serious lack of standards formally approved for the information field by the American National Standards Institute (ANSI) and by the International Standards Organization (ISO). Furthermore, standards that have been fully approved by these organizations are applicable to only a small segment of the over-all information field. In fact, even after useful ISO or ANSI Standards have been fully formalized, there is a long delay in getting them into widespread use, usually because of economic reasons.

Second, many of the existing generally accepted practices, so often mistakenly referred to as standards, are subject to interpretation by
Standardization Requirements

different individuals functioning in a variety of environments at different times and are controlled or motivated by differing conditions. An example exists in the library field in such generally accepted practices as those outlined in the A.L.A. Cataloging Rules for Author and Title Entries, the A.L.A. Rules for Filing Catalog Cards, and in the Rules for Descriptive Cataloging in the Library of Congress. However, these do not lend themselves to simple, concise, rigid specifications such as one would find, for instance, in a standard laboratory technique for the preparation of a synthetic organic compound or in a mechanical dimension standard for machine screws. Records generated by different libraries for the same publication can, and usually do, differ considerably. Such situations cannot be automated easily.

If one accepts the hypothesis that a national program for information transfer has as its objectives the development of a coherent system for the efficient, effective, and economic transfer of information, then the need for a standardization program which is much more extensive than anything available to date becomes obvious.

As in many other fields, automation is being introduced to handle an increased load, to provide new and better services or results, and to augment scarce human skills. Automation in information handling is being achieved through modern technological methods of reprography, data processing, and communications. Their use emphasizes the need for standardization. The power of humans to interpret information in many forms and to build their own intellectual bridges between system segments will no longer be sufficient to the task.

In fact, replacing this "non-productive" human intervention is one of the improvements hoped for through automation. Information interpretation by machine, however, requires great detail of specification and uniformity of practice. Machines can transform between unambiguous alternate forms, but they cannot "understand" and re-express information or resolve ambiguity through experience—at least the machines of today and the immediate tomorrow cannot.

A national information transfer automation program will also place added emphasis on the concept of shared resources. Few, if any, modern libraries are totally self-sufficient; neither is the library community, when viewed collectively. While libraries have managed to produce for themselves the bibliographic descriptions of their book collections, they have for the past 100 years been almost completely separated from the work of providing content analysis information for their periodical and serials collections. The need to find solutions to the standardization
problems and to perform the tasks imposed by a national program presents problems that transcend the confines of the library community and confronts instead the entire information-handling community which is comprised of the efforts and activities of authors, users, publishers, dealers, and abstracting and indexing services, as well as libraries. The standardization requirements of a national program for information transfer are affected by a wide range of diverse, yet interrelated factors.

There are many levels of standardization. At the document-handling level, microforms and reprography may be used to reduce storage volume and cost by providing selective copies to information users in place of the mass distribution of documents or the removal of source material from "lending libraries." But it will not be practical to use the technology of microforms and reprography on a national scale and reap the full potential benefits until there is a widespread compatibility at prices that are comparable with the rest of the image media, until there is equipment to produce microforms and equipment to view and copy them to full document size, and until a solution to the very complicated and serious problems involving copyright and appropriate usage charges is found. A system based on medium-to-medium recopying to bridge different forms has only limited practicability because photographic recopying cannot maintain the necessary image quality through an unlimited number of copy generations. Continued technological developments may ameliorate this limitation, but today's technology is definitely limited in this respect.

Computer-readable information-transfer media are currently receiving much attention with respect to standards.\textsuperscript{11} The physical recording medium, and the character sets—both the graphic shape and coded representation—have been treated in considerable detail. Data formats and file organization are also being given increasing attention.\textsuperscript{12}

The field of information transfer must not be exclusively concerned with these machine factors in the standardization required for large scale national and international systems, however. The representation of the content of the information carried through or on the transfer media must also be designed for effective machine handling. This may well be the most important area to consider for standardization. The chief problem is the avoidance of ambiguity. A secondary problem is that of acquiring the highest practical degree of uniformity necessary to make the process as economical as possible.
Standardization Requirements

As for content representation, consideration must be given to:

1) forms derivable from original information via a set of rules, or conversely, from requirements placed on the originator of information to provide specific data items in specific forms (examples: author name formats, titles, and representation of dates);

2) codes maintained by an accepted authority and given specific meanings by definition (examples: American Society for Testing and Materials—CODEN, Standard Book Numbers, Library of Congress Catalog Card Number, and Chemical Abstracts Service Chemical Compound Registry Numbers); and

3) transformations allowable independent of machine handling (examples: transliteration rules from non-Roman to Roman alphabets, and handling of diacritical marks).

The President's Task Group for the Interchange of Scientific and Technical Information in Machine Language (ISTIM) recommended in their final report that:

USASI Committee Z39 in cooperation with concerned organizations for indexing and abstracting give high priority to the development of a standard procedure for citation and that this be disseminated to the publishers of primary and secondary literature for early consideration . . . that Z39 pursue the development of standards citations to other types of materials, such as books, patents, etc. . . . and that particular efforts be made during the immediate future within the normal framework of Z39 to enhance and enlarge the participation in Z39 by the major organizations directly involved in mechanized bibliographic data handling. In this connection, it is recommended that the scope statement and name of Z39 be enlarged to encompass the related discipline of information science.18

ANSI (formerly USASI) Committee Z39 has thus been identified as a mechanism through which these standards of primary importance can be formulated on a national basis. The recently published SAT-COM Report 14 also emphasized the broad role of Z39 in reaching agreements on bibliographic practices among libraries, documentation centers, and the abstracting and indexing services. These recommendations reflect what is actually being done by Z39, which has been supported in its efforts, since 1961, by the National Science Foundation and the Council on Library Resources.16

In a national program each of the participants becomes a node in a
network and the interactions between pairs of nodes can be significantly different in different parts of the network. For example, the transfer of information between a library as a node and the library client as a node may be quite different from that between a content-analysis center (such as the Library of Congress cataloging department or an abstracting and indexing service) and a library. The format of a machine-readable record is simply the framework which carries the defined elements of data or character strings within the machine record. In the context of a national program, then, the prime consideration with respect of formats should be focused on communication or interchange of information between nodes and not necessarily the formats used to process the information (store, search, or retrieve) at any specific organization. It may not be necessary, or even desirable, that the machine formats be uniform throughout the network, but it is mandatory that the content representations be uniform and that they be translatable into all formats used to carry that information.

To date, the most widely publicized format for English monograph bibliographic data distribution is that developed at the Library of Congress, the MARC II format.10 This format is currently pending for acceptance by the ANSI as the USA Standard for a Format for Bibliographic Information Interchange on Magnetic Tape. This format is being tested and is gaining acceptance by the library community.17 There are other formats, such as the ones being introduced in their respective fields by the Institute of Electrical and Electronics Engineers, by the American Institute of Physics, by the BioSciences Information Service of Biological Abstracts, by the Chemical Abstracts Service, and the one proposed by the Joint Agreements Group,18 for the handling of a combination of bibliographic information and other forms of information such as chemical structures, physical data and biological properties.

Regardless of their importance to the solution of current problems, none of these present forms will, in the authors' opinion, be the backbone of the eventual national interchange program with its broad range of requirements. The success of a national program depends upon having a standard interchange format, or a compatible set of interconvertible ones, but precisely what this format, or these formats, should be is still a topic to be addressed by the participants in the national program. Moreover, some solid experience with present forms is needed now to further the development of a format suitable for broad-scale national use. The successful acceptance of any format as a
Standardization Requirements

format for a national program depends on several factors. Among them are the amount of useful data made available in the format from any source, and how well it stands the test of usage.

We have strongly emphasized the necessity of being unambiguous in the representation of information content in a machine-readable record, and the necessity of being able to translate that content into any form needed for processing in any part of the information net-work. The approach that is being adopted very widely in the information community, for example, in MARC II, in the Chemical Abstracts Service Standard Distribution Format, and in the discussion of the Joint Agreements Group, is the definition of data elements and the labeling of those data elements as information transfer media by published codes. Data elements then are defined as units of information within the system. Customarily, a data element or unit of information contains a tag or identifier and the content. For example, in the MARC II system, the tag 245 or TIL has been assigned to the variable field entitled "Title Statement." In this way, the user of MARC II formatted tapes knows that a data element tag 245 or TIL contains the title of the monograph represented by the record. Similarly, in the Chemical Abstracts Service Standard File Format, data element 200 is the CODEN of the source journal.

The information community has not yet evolved to the point where these data element tags are standardized, and it is far from the point where the transfer formats themselves are identical in implementation.

In May 1967, Z39 Subcommittee 2 (Machine Input Records), Special Project on Data Elements, published a composite list of bibliographic data elements applicable to the full range of bibliographic forms. In August 1967, Subcommittee 2 under the direction of chairman Henriette Avram, Information Systems Office, Library of Congress, met and concluded that:

The most useful next step for the SC/2 would be to draft a general statement which would establish a medium of exchange between various producers and users of bibliographic information. The general principle of identification or tagging of data elements would be emphasized, but there would be no attempt at specifying all data elements to be identified in all bibliographic forms.

Since that August 1967 meeting, an increasing number of machine-readable data bases have become available to the information-handling community. In science and technology alone Cohan lists 289 such services, some of which are available on a subscription basis.
while others are available in an experimental form only. The data-element identification work done by the Information Systems Office of the Library of Congress for the MARC II format for monographs, and more recently the working document for the MARC format for serials, have been the data-element lists furnished from the library community. The abstracting and indexing community has published, or has made available to interested parties, sets of data elements including those issued by the International Atomic Energy Agency for its International Nuclear Information Service (INIS) and by the Chemical Abstracts Service.

There is ongoing work on data-element specifications that may have a direct bearing on the national plan for automated information transfer. This is the work of the Joint UNESCO/International Council of Scientific Unions UNISIST project and the National Federation of Science Abstracting and Indexing Services, Bibliographic Citations Committee (NFSAIS-BCC). The ultimate goals of these two working groups is to develop an optimal set of interchange data elements that will be acceptable to the abstracting and indexing services whose staff are members of International Council of Scientific Unions-Abstracting Board (ICSU-AB) and NFSAIS respectively.

During the past two years, ANSI Z39 Committee has addressed its efforts vigorously to the data-element identification and specification problem. Its subcommittees are currently working on the following data elements: a) country names and political subdivisions (such as countries of the United Kingdom, republics of the Soviet Union, states and counties of the United States, and provinces of Canada), b) calendar dates, c) languages, d) names and addresses of libraries, e) names and addresses of book dealers, and f) names and addresses of publishers. In addition, Z39 committees are working to develop Standard Book Numbers (SBN) and Standard Serial Numbers (SSN). Standardization of each of these data elements and of many others is required for a national program.

It is apparent that a more comprehensive and coherent attack will have to be made on the data-element standardization problems than has been waged in the past. Such an effort could merge and unify the results of various national and international group efforts, such as those of NFSAIS-BCC, MARC, UNISIST, and individual and collective abstracting and indexing services, into a coherent set of community-acceptable standards. Optimal standardization will require some sacrifice of traditional practices of individual members of the
over-all information-transfer community. However, the eventual benefits of such standardization to a national program would far outweigh any restrictions on current practices or any other disadvantages it might bring about. Getting into position to achieve this standardization is a difficult task.

There are many bibliographic data fields that can be adequately expressed for interchange purposes in the forms of abbreviations or codes rather than complete data. A name of a country, for example, can be represented by a coded form on an interchange record and can be expanded to a full name or, if necessary, reduced to an abbreviation of that name on output.

The advantages of code utilization are many. Less input transcription (keyboarding) is required, and the fewer the characters, the less the chance for error. There are fewer characters to proof, fewer to correct when errors do occur, and fewer to store and transmit. Codes are also amenable to the use of check characters for error detection. Translation of the code to the assigned value has an inherent degree of flexibility. Codes become a kind of Esperanto, or universal language. For example, a serial publication might carry titles on its cover in several languages. The standard serial code for one processor would translate to the version of the language most meaningful for his clients' requirements, while for another processor the code could translate to a library catalog entry or to the title abbreviation for still another language.

Earlier in this paper the authors identified several data elements that lend themselves to coding, including place names, languages, publishers, libraries, books, and serials. Several organizations have already developed authority files for codes for their internal use and for use by their clients. There has also been some effort to standardize codes and their coded values. The recently published Federal Information Processing Standards (FIPS) include coded values for calendar dates, for states of the United States, and for counties within the states. The International Organization for Standardization has also published a recommendation relevant to this effort. The national program for library automation or information transfer will require the identification of those bibliographic data elements whose contents can be best represented by codes and then the standardization of the codes and the data they represent.

Another area that will require standardization is transliteration. The transliteration problem includes not only those languages using ideograms...
graphs or non-Roman alphabets, but also those languages using alphabets that contain more than the twenty-six letters used in the English language. In 1969, the International Organization for Standardization published a revision of its 1955 Recommendation 9: *International System for the Transliteration of Cyrillic Characters.* This standard is worthy of serious consideration by the program. The International Organization for Standardization Technical Committee 46 Documentation Meeting held in Stockholm during October 1969 included discussions of the transliterations of Yiddish, Chinese, Japanese, Arabic, Hebrew, and non-Slavic Cyrillic languages. Eventually these discussions must culminate in community-acceptable standards.

If direct compatibility of data-element form and content of the various systems comprising a national program is to be achieved, then the full range of the transliteration problems mentioned above must be subject to rigid standardization. However, if only convertibility is required, it can be achieved with “softer” standards. Convertibility will require that each organization producing bibliographic records make known in detail the codes, their values, and the transliteration schemes that it employs. Care must be taken to insure that these local transliteration schemes do not lose information in the process.

The following statement from the Library of Congress reflects on both the national program requirements for transliterations and the need to standardize these transliteration schemes.

Until a definitive character set for nonroman alphabets and the techniques for input, storage, and output of such characters can be developed, all information will be entered into the serials system in romanized form. The romanization or transliteration schemes presently applied by the Library of Congress will be used when cataloging publications in nonroman alphabets or nonalphabetic languages.

Transliteration problems in the library and information community are analogous to the character-set problem of the computer system field. The needs of a printer, a telegrapher, an information-retrieval system, and a data-processing system, are all different. However, each of these has an impact on the national information transfer system. True standards have been formalized only for the data-processing system that handles strings of character data (e.g., ANSI 7-bit character set). However, for a national system to operate successfully, all of these interests must be brought into compatibility (transformability).
Standardization Requirements

The problem of general semantics in text is so complex as to far exceed the problems of character sets and of data elements. Within a specialized field, a thesaurus is often used to provide a standardized set of relationships among terms including synonyms and hierarchical relationships. It is fortunate that much can be accomplished in a national information-transfer program before tackling the problem of standardization of thesauri. This is because most of the attention at this time is focused on automated dissemination of, or access to, information, and not on the machine interpretation of it. There is much research on information retrieval concerned with machine interpretation of ordinary text. Unfortunately, this research has not yet matured to the point that it can be considered a major feature of a national information-transfer system. Furthermore, before it can be fully effective, the “data input” problem must be solved to get most relevant text into machine-readable form.

Because of the heterogeneity of the various links in a national information-transfer network and the immensity of the system relative to any change that can be made in it by any one effort, the authors have departed from the usual idealistic view of standardization as the achievement of total uniformity in the representation and handling of information. While no opportunity to achieve such uniformity should be wasted, progress will have to be achieved step-by-step as each part of the system develops.

In the analogy from the national telephone system, there existed a single management-planning and decision-making authority for most of the system involved. It could force the network design and standardization so that local, non-optimal actions could be taken to achieve over-all optimization. In contrast, there are many major centers of influence involved in the information problems:

a) national libraries and the general library community,
b) abstracting and indexing services,
c) major scientific and technical societies and their publications and conferences,
d) large mission-oriented activities of the U.S. government,
e) organizations that develop and sell communication, computer, and reprography technology on which a modern national network depends,
f) consumers of information, a large heterogeneous group of persons and organizations with diverse and sometimes conflicting needs,
g) commercial publishing and information-handling enterprises, and
h) generators of information in a multitude of forms.

Any attempt to present the above in any order of precedence, importance, or impact, is foolish; there is, therefore, no single entity, nor is there likely to be, with sufficient skills, prestige, resources, and authority to assume the role of monolithic management.

Thus, planners for the evolution of the national information network, and the developers of the standards which guide it, must take this difference into account as they strive to make the technological potential a reality. The network must grow together, forced only by common need and the recognition of all parties that cooperation and perseverance alone will lead to the desired result. The importance of the various standards organizations as the final mechanism for recognizing the transformation of “accepted practice” into formal standards thus becomes evident. And a necessary preliminary to creating these formal standards is the step-by-step experience of individuals and groups who work to solve the problems of ambiguity and incompatibility between nodes in the system.

Finally, we cannot consider the development of a national information transfer network without considering the international implications. Much of the information handled in scientific, technical, and scholarly publications in the national system originates outside of the United States. Any national system in the United States has an international impact. Furthermore, some of the organizations that will be instrumental in achieving standards, such as International Organization for Standardization and UNISIST, are international organizations. Thus, while we work toward achieving an improved national information-transfer system, we must be acutely aware of these international relationships.

As the expression of the major significant details in information-transfer system design, standards of representation and practice, agreed to and used by all parties, become the guiding mechanisms which replace unified management. As such they take on an importance in achieving progress in national, and international information transfer which is beyond the technical importance normally associated with standards. All parties, however, must be patient with the inevitably slow development and utilization of those standards.
Standardization Requirements

References

3. Nunn, ibid.


Standardization Requirements


