Agency Cooperation in Processing Technical Report Literature*

PETER F. URBACH
National Technical Information Service, U. S. Department of Commerce, Washington, D. C. 20230

Received January 18, 1973

Input to the National Technical Information Service comes from over 225 Federal agencies in forms ranging from unprocessed report copies to completed bibliographic information on magnetic tape. Arrangements with contributors range from informal agreements to formal contractual relationships for bibliographic processing on a cost reimbursable basis. Four different and overlapping thesauri are used for indexing the report input. This results in inconsistencies in indexing and complicates retrieval. Solutions that have been considered include the interagency development of a single common thesaurus to meet all agencies' needs, the re-indexing of previously indexed material upon receipt by NTIS, and the development of word lists and thesauri from the existing data base for use as retrieval aids. The manipulation of the data base to develop retrieval aids currently appears to be the most promising approach.

This paper describes some of the interactions between the National Technical Information Service and the Federal agencies that generate the reports that are processed into the NTIS system. It also describes some of the problems that result from this system of interagency cooperation.

The National Technical Information Service is a component of the U. S. Department of Commerce engaged principally in processing and making available to business and industry the results of Government-funded research and development in the form of technical reports. NTIS processes bibliographic information from 70,000 reports annually from over 225 Federal agencies and sells approximately 2.5 million report copies each year.

NTIS' major products and services fall into three categories: the announcement products, a series of weekly and semimonthly abstract bulletins that announce newly-released reports; the reports themselves available in paper copy and microfiche form; and a retrospective search service providing public access to the report data base.

INPUT TO DATA BASE

The cornerstone of the NTIS products and services is the machinable bibliographic data base which is created from the report input. Some of the contributing agencies have sophisticated internal information systems which provide bibliographic information on their reports to NTIS in machinable form. Other agencies have highly decentralized information systems without any bibliographic processing or machine processing capability.

NASA, AEC, and DoD provide bibliographic data on their reports to NTIS in magnetic tape form. The DoD bibliographic processing is performed by NTIS for the Defense Documentation Center on a cost reimbursable basis. DoD reports are sent to NTIS in hard copy form and NTIS does the descriptive cataloging, indexing, and abstracting. The bibliographic information is then keyboarded into the DDC computer through terminals located at NTIS. After computer processing at DDC, a copy of the bibliographic data in magnetic tape form is sent to NTIS.

NASA and AEC produce similar bibliographic records of their reports in magnetic tape form in their own information processing facilities and send copies of these tapes to NTIS. In addition, NASA and AEC also provide NTIS with master microfiche copies of their technical reports.

Most of the agencies that do not provide magnetic tape input to NTIS do provide the basic bibliographic information in a standard hard copy format. An interagency guideline prepared under the auspices of the Federal Council's Committee on Scientific and Technical Information (COSATI) sets forth the format for the bibliographic data sheet.[1] The data sheet contains the descriptive cataloging information as well as the indexing terms and the abstract of the document. A sample data sheet is shown in Figure 1.

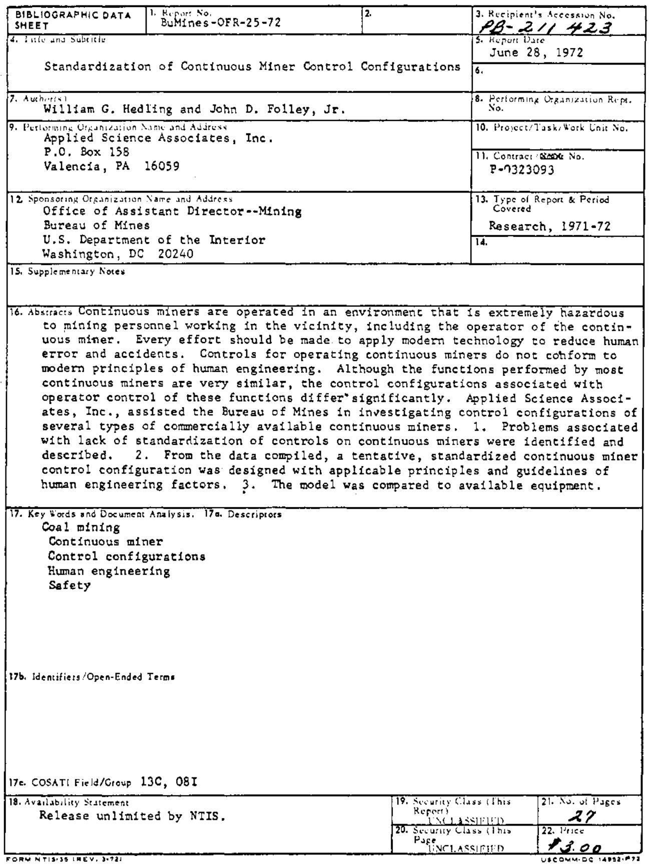
The NTIS analyst reviews the data sheet which accompanies the document to ensure that the cataloging and indexing information is correct. If necessary, the analyst edits the data sheet but generally does not have to repeat the basic cataloging and abstracting.

A third category of input comes from agencies that do not do any bibliographic processing on the reports they submit to NTIS. These agencies provide only the hard copy report. In these cases, the NTIS analysts must do the entire cataloging, indexing, and abstracting for the document.

Table I summarizes the percentage of each type of input received. Note that 79% of the input comes to NTIS in the form of other agency tapes and that only 6% of the input requires full processing by NTIS.

Clearly, a cooperative interagency system has evolved which minimizes duplicative processing. In the future, we can expect the system to develop further in the direction of more magnetic tape input and less and less central manual processing. We can expect some of the newer agencies—HUD, DoT, and EPA, for example—to develop central information systems of their own for processing

* Presented before the Division of Chemical Literature, 164th Meeting, ACS, New York, N. Y., Aug. 27, 1972.



BIBLIOGRAPHIC DATA SHEET	1. Report No. BuMines-OFR-25-72	2.	3. Recipient's Accession No. PB-211 423
4. Title and Subtitle Standardization of Continuous Miner Control Configurations			5. Report Date June 28, 1972 6.
7. Author(s) William G. Hedling and John D. Folley, Jr.			8. Performing Organization Rept. No.
9. Performing Organization Name and Address Applied Science Associates, Inc. P.O. Box 158 Valencia, PA 16059			10. Project/Task/Work Unit No. 11. Contract/[illegible] No. P-0323093
12. Sponsoring Organization Name and Address Office of Assistant Director--Mining Bureau of Mines U.S. Department of the Interior Washington, DC 20240			13. Type of Report & Period Covered Research, 1971-72 14.

15. Supplementary Notes

16. Abstracts Continuous miners are operated in an environment that is extremely hazardous to mining personnel working in the vicinity, including the operator of the continuous miner. Every effort should be made to apply modern technology to reduce human error and accidents. Controls for operating continuous miners do not conform to modern principles of human engineering. Although the functions performed by most continuous miners are very similar, the control configurations associated with operator control of these functions differ significantly. Applied Science Associates, Inc., assisted the Bureau of Mines in investigating control configurations of several types of commercially available continuous miners. 1. Problems associated with lack of standardization of controls on continuous miners were identified and described. 2. From the data compiled, a tentative, standardized continuous miner control configuration was designed with applicable principles and guidelines of human engineering factors. 3. The model was compared to available equipment.

17. Key Words and Document Analysis. 17a. Descriptors
Coal mining
Continuous miner
Control configurations
Human engineering
Safety

17b. Identifiers/Open-Ended Terms

17c. COSATI Field/Group 13C, 08I

18. Availability Statement Release unlimited by NTIS.	19. Security Class (This Report) UNCLASSIFIED	21. No. of Pages 27
	20. Security Class (This Page UNCLASSIFIED	22. Price $3.00

FORM NTIS-35 (REV. 3-72) USCOMM-DC 14952-P72

Figure 1.

Table I. 1971 NTIS Input

Agency	Type of Input	% of NTIS Input
DoD	Tape/paper copy	40
AEC	Tape/microfiche/ paper copy	23
NASA	Tape/microfiche/ paper copy	16
Other agencies	Bibliographic data sheet/ paper copy	15
Other agencies	Paper copy only	6

their technical reports and, as they do, we can expect more bibliographic data input to NTIS in magnetic tape form.

PROBLEMS

This evolving cooperative interagency information system, which minimizes duplication of effort, is not without its problems.

Much of the bibliographic input received in the form of bibliographic data sheets, for example, does not provide useful indexing information. Although the descriptive cataloging and the abstract on the data sheet are usually useful, most of the data sheets require re-indexing by NTIS analysts. In most cases, the indexing information supplied is provided by the author and has been assigned without reference to any controlled vocabulary. It is useful in guiding our indexers and is particularly helpful in supplying the free language "identifiers" but, with few exceptions, most of the bibliographic data sheets require complete re-indexing.

A different kind of problem stems from the lack of standard formats for magnetic tape interchange. Each agency providing NTIS with magnetic tape provides the tape in its own record format. The record formats differ from one another so markedly, and are so complicated, that reformatting is not a simple programming task. After over a year of programming effort, we have only recently succeeded in converting that portion of the NASA tape which contains abstracts to run in our machine system. AEC records still appear in the NTIS system without abstracts since the abstracts are not available to us in machinable form.

A related problem involves "minor" systems changes made by the agency that generates the tape which completely disable the machine systems of the tape user. Frequently, these changes are unannounced and cause established processing routines that have been running for years to hang up in the midst of a production run.

MULTIPLE THESAURI PROBLEMS

One problem of increasing seriousness that warrants special attention is caused by the use of different and frequently inconsistent indexing vocabularies.

The NTIS data base presently contains information that has been indexed in accordance with four different thesauri. Table II lists these four thesauri. DoD, NASA, and AEC each have their own controlled vocabularies which are used to index their own documents. Each agency feels that it has unique needs which are best met by its own unique indexing language. All of the other agency documents are indexed in accordance with the TEST thesaurus, the result of the DoD Project LEX and the EJC efforts.

NTIS subject analysts must index DoD documents in accordance with the DDC thesaurus and other agency documents in accordance with the TEST thesaurus. NASA and AEC input comes in on magnetic tape already indexed in accordance with their respective thesauri.

Examples of the differences between these thesauri are shown in Table III. Note that in some cases entirely different words are used to express the same concept. In other cases, pairs of words are presented in one sequence in one thesaurus and in another sequence in another thesaurus. Finally, the most common difference, the use of singular forms of words in some thesauri and the use of plural forms of the same words in others.

Since four different thesauri are used to index the documents, four different thesauri must be used for proper retrieval. This places a considerable burden upon the search analysts. Our analysts at NTIS are familiar with the four different indexing languages, but must nevertheless go from one thesaurus to another to ensure that they are using the correct terms and synonyms for retrieval. Other users of our data base, who are less familiar with the different indexing languages, have considerably more difficulty. The impact of the problem becomes greater as more and more users begin to use the NTIS data base in machine-readable form in their own information systems. The use of online terminals which provide the user with the capability of interacting intimately with the file will further highlight the inconsistencies.

Table II. Thesauri Used by NTIS

DoD	Thesaurus of DDC descriptors
NASA	NASA Thesaurus
AEC	Subject headings used by the USAEC
Other agencies	Thesaurus of Engineering and Scientific Terms (TEST)

Table III. Thesauri Differences

Concept	DDC	NASA	AEC	TEST
Teflon	Teflon	Teflon (trademark) or polytetrafluoroethylene	Ethylene/tetrafluoro-/polymers	Tetrafluoroethylene resins
Nuclear power reactors	Power reactors	Power reactors	Reactors/power	Power reactors (nuclear)
Batteries	Batteries and components	Electric batteries	Batteries	Electric batteries
Gages	Meters	Measuring instruments	Gages and meters	Measuring instruments
Diet	Diet	Diets	Diet	Diets

Furthermore, the problem promises to get considerably worse before it gets better. As new agencies begin to develop their own central information systems they also tend to develop their own indexing languages. Each agency feels a need to develop its own thesaurus with an indexing language oriented to meet its own unique needs. Thus, we can expect the number of different thesauri to increase.

As an example, the National Highway Safety Bureau of the Department of Transportation has developed its own thesaurus and is now negotiating with NTIS for the processing of highway safety literature. If these negotiations mature into an operational agreement, NTIS analysts will have to contend with a fifth thesaurus. It is not hard to envision the present system evolving into a cooperative interagency system with decentralized processing and exchange of bibliographic data in magnetic tape form but with document indexing performed in accordance with dozens of different thesauri.

POSSIBLE SOLUTIONS

There are a number of possible approaches to this problem. One approach is simply to shift the burden of rationalizing the inconsistencies in the different vocabularies to the end user who is searching the file. This is essentially the approach that NTIS is following today. The end user is required to learn to work with the various vocabularies used in the NTIS data base. As additional agencies develop their own vocabularies, the end user will simply have to learn to handle the additional vocabularies as he searches the file. This is not a solution that will enhance the use of the file. It is more likely to result in considerable frustration on the part of the user.

Another obvious approach to the problem is for all of the agencies to agree on the use of a single thesaurus in lieu of their many separate vocabularies. Though this is an obvious solution, it is not an easy one to implement. Each mission-oriented agency feels that it has special needs that require a special vocabulary to index and retrieve its own information.

Efforts to develop common vocabularies that cut across agency lines have proven to be extremely difficult and expensive. Project LEX, for example—the DoD-wide effort to develop a consolidated vocabulary for all DoD technical information—was a 1½-year effort for a full-time staff of 12 professionals and 350 participating panelists. In addition to all of the part-time contributed professional support, nearly one-half million dollars were spent on the effort.[2]

Even if a common interagency thesaurus could be developed, there is some question as to whether it would be actively adopted by the participating agencies. Quite apart from the cost of developing a new thesaurus, implementing a new thesaurus in an operating information system is an additional considerable expense. The Defense Documentation Center, for example, one of the principal participants in the Project LEX exercise, continues to use its Thesaurus of DDC Descriptors rather than the TEST thesaurus. The cost of changing the indexing vocabulary and the computer based system for controlling that vocabulary in midstream is just too great.

The goal of a single thesaurus that all of the agencies would use for indexing their documents seems somewhat beyond our present reach.

Another approach would be for NTIS to adopt, unilaterally, a single thesaurus and convert the indexing from all of the other agencies to the single common thesaurus. The magnitude of this effort might be as great as the effort required to build a common interagency thesaurus. It would be necessary to develop relationships between all of the indexing terms of all of the thesauri. Once these relationships were developed, incoming documents already indexed by the generating agencies would have to be re-indexed by NTIS. Although this approach avoids costs on the part of other agencies and the end users, the resources to accomplish the task are well beyond the reach of NTIS.

The two previous approaches, the development of a single new interagency thesaurus or the re-indexing of other agency documents by NTIS in accordance with a unilaterally developed thesaurus, both require the re-indexing of the back file of documents already in the data base. This is a massive job which could not be undertaken on a manual basis. It would require machine re-indexing of all of the old documents in accordance with the new thesaurus.

A more desirable solution to the problem would shift enough of the burden to the end user so that re-indexing of all old documents would not be required.

A fourth approach which may meet this requirement is the machine manipulation of the data base to develop retrieval aids which would assist the user in framing retrieval questions. In its simplest form such a retrieval aid would simply be an alphabetic listing of all of the index terms used in the file. In a more sophisticated form, a retrieval guide could indicate synonymy and other relationships between index terms from different vocabularies. This might be accomplished by machine manipulation of a portion of the data base to determine relationships between index terms, statistically. This type of tool would permit the user to express a search question in one vocabulary and simply translate it into all of the other vocabularies employed in the system. No re-indexing of previously indexed documents in the file would be required.

The NTIS data base is just now being put online for internal use and no decisions have been made as to how the multiple thesauri problem will be handled. We expect to learn a lot from the online manipulation of the data base which will aid in selecting an alternative from among those discussed. One of the things we have already learned from online manipulation of the data base is that there are a substantial number of errors in the records already in the data base. A significant file correction effort may be a necessary first step to improving the file.

CONCLUSION

There are some substantial cooperative efforts involved in the processing of bibliographic information into NTIS. These cooperative efforts are not without serious problems and further cooperative efforts will be required to provide solutions. The particularly difficult problem of vocabulary control and retrieval in the NTIS multiple thesauri system will require further analysis and experimentation.

LITERATURE CITED

(1) Committee on Scientific and Technical Information, "Guidelines to Format Standards for Scientific and Technical Reports Prepared By or For the Federal Government," 1968, **PB 180 600.**

(2) Heald, J. H., "The Making of TEST, Final Report of Project LEX," Office of Naval Research, November 1967, **AD 661 001.**

CHEMTRAN and the Interconversion of Chemical Substructure Systems*

CHARLES E. GRANITO**
Institute for Scientific Information, 325 Chestnut St., Philadelphia, Pa. 19106

Received January 18, 1973

The need for the interconversion of chemical substructure systems is discussed and CHEMTRAN, a new service, designed especially for creating interconversion programs, is introduced.

The organic chemist studies the three-dimensional world of chemical compounds but, as with researchers in other fields, he communicates his results in one or at best, two-dimensional media. The principle of least effort[1] guides him in most of his efforts; organic chemistry is no exception. The organic chemist has adopted the structural diagram as the means of employing "least effort" in communicating his findings to students and colleagues.

NOMENCLATURE

Unfortunately, chemists have for too long been trying to use nomenclature to serve two basically divergent needs:

1. The need for a short descriptor to replace the entire chemical structure in speech and in writing
2. The need for systematic indexing names that describe the molecular makeup of a compound

Thus, we find, for example, both adamantane and tricyclo $(3,3,1,1^{3,7})$-decane representing the same structure (Figure 1).

The problem has traditionally been that chemists usually prefer the shorter, less descriptive names where they have a choice. As the result of continual compromise, we find many "trivial" names adopted as standard, "systematic" names. For over a hundred years there has been a mixing of trivial and descriptive names. Not surprisingly, the situation has become chaotic. Organic compounds reported in the literature are not really indexed by any "systematic" nomenclature, but rather, by a nomenclatural system made up of many descriptive and many nondescriptive words. Compromise has, in this case, often reduced the system to a point of very limited utility. Furthermore, the resulting "nomenclature" has led to a dependence on whole molecules in using subject indexes, despite the fact that research chemists are usually interested in classes of compounds—i.e., compounds which have some common structural feature. Nevertheless, there is no substructure index to the 4 million compounds reportedly recorded within *Chemical Abstracts!*

SUBSTRUCTURE SYSTEMS

The failure of nomenclature for substructure searching has been recognized for some time but has not diminished the many organic chemists' desire to do substructure searching. Instead, it has led them to develop other structure representations.

Over the past 30 years, a considerable effort has gone into the development of substructure search systems. There are today, hundreds of such systems in use around the world. These systems may be generally classified as: fragment codes, line notations, and connectivity tables.

The most widely used fragment code is the Ring Code.[2] The most widely used line notation is the Wiswesser Line Notation (WLN).[3] Connectivity tables (CTs) have also received attention, but are not presently used by many organizations. However, the decision by *Chemical Abstracts* to use a connectivity table in its registry system[4] has led to increased interest in this technique in the past several years.

Figure 2 shows an example from each major system. We will not attempt to contrast the various systems in this paper.

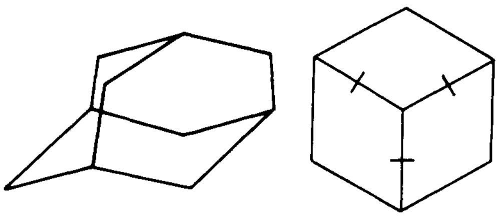


Figure 1. Adamantane

*Presented before the Division of Chemical Literature, 164th Meeting, ACS, New York, N. Y., Aug. 30, 1972.

**Present address: C G Associates, 1948 Cardinal Lake Drive, Cherry Hill, N.J. 08003