Data-Tagging Experiment for Energy-Related Documents at Chemical Abstracts Service

DAVID F. ZAYE

Chemical Abstracts Service, Columbus, Ohio 43210

Received October 16, 1980

An experiment is described that evaluated the usefulness and economic feasibility of incorporating data tags into secondary information files. Tags are brief codes that indicate the presence in a primary document of specific types of numerical data. Tags were selected in cooperation with users of the U.S. Department of Energy's (DOE) RECON system and added to a special version of the computer-readable ENERGY file. RECON users assisted in an evaluation of the tagging process. It was found that data tags do increase access to numerical data in the primary literature, but a file spanning several years is necessary to produce useful results. Evaluation of processing costs revealed that after the initial training period, costs were not significantly greater for the number and types of data tags employed in this study. The feasibility of including data tags in nonenergy sections of *Chemical Abstracts* was examined with primary emphasis on economic and editorial factors. It was shown that subject content governs the selection of data types to be tagged for a section or group of related sections with the overriding consideration of user requirements.

INTRODUCTION

Two primary energy sources that are the focus of intense investigation today are fossil fuels and nuclear fuels. In-depth research is being conducted in locating additional geographical areas and determining geological reserves of these fuels, analyzing conversion methods that allow for energy storage and transport, and characterizing patterns of energy usage. In each of these areas, large amounts of numerical data are being produced and evaluated, and policy decisions ensue based on these data. Some of the problems associated with energy data generation, storage, accessibility, and retrieval are found in the report "Data Needs for Energy" produced by the Committee on Data for Science and Technology.[1] Cost factors have also been explored for searching out data values that had been derived and/or evaluated vs. conducting the experimental procedures to establish these values initially.[2]

The problems of conserving and accessing data have burgeoned over the past three decades with the rapid improvement and extension in the power and the versatility of data-creating and -recording equipment and methodology. Also contributing to these problems is steady growth in the range of science and technology to which search capabilities are applied. This trend seems likely to continue. A special facet of the data management problem is the relatively recent emphasis on critical review and/or evaluation of data. Various national and international organizations have been established to improve the accessibility and the certification of data. Among these are the National Standard Reference Data System, operating out of the U.S. National Bureau of Standards, the Committee on Data for Science and Technology of the International Council of Scientific Unions, the Analytical Chemical Division of the International Union of Pure and Applied Chemistry, and the Numerical Data Advisory Board, which is under the aegis of the Division of Chemistry and Chemical Technology of the U.S. National Academy of Sciences/National Research Council. There are also some mission-oriented data-gathering activities, such as the International Cancer Research Data Bank. The National Data Advisory Board emphasizes the importance of compiling and providing access for data in the following statement:

> In the last analysis, progress in research and development is based on quantitative data. Therefore, the extraction of such data from the primary literature, and analysis and the evaluation of these data, and their dissemination in compilations or in critical views has been recognized as a most efficient and important intellectual condensation of the literature.[3]

The interest in data values in relation to energy processes and the emphasis on increased accessibility are reflected in an experiment conducted by Chemical Abstracts Service (CAS). The experiment focused on the subset of total coverage of *Chemical Abstracts* (CA) contained in the computer-readable file ENERGY. Special issues of ENERGY were created which contained, in addition to the regular abstracts, index entries, and bibliographic information, pointers for locating numerical data found in the corresponding source documents referenced in the file. The pointers consisted of a set of unique, computer-verifiable codes called data tags.

The tags selected for the experiment consisted of alphanumeric codes varying in length from two to four characters. Attempts were made to devise codes that would be recognizable; examples were BP for boiling point and CD for current density. It was anticipated this would minimize "look-up" time for both the user of the tags and those staff members at CAS who analyzed the source documents for the numerical data content. It was also decided at the outset that this level of tagging would be preferable to alternative schemes.[4,5] These schemes employ tags ranging from a single character, indicating either the presence or the absence of data in a source document, to those procedures where a multidigit, hierarchical code is used, denoting very specific kinds of numerical data and the conditions under which measurements were made. In accordance with CAS policies, the tagging of a data type was "neither critical nor evaluative".[6]

For provision of as complete a record as possible of the data content of a source document, a tag was assigned whenever a corresponding data type was located, regardless of whether that data type would be included in the abstract title, etc. The user, searching only this tag field, could therefore obtain the data description. All parts of the source documents were examined for data: text, tables, charts, figures, and graphs. Correlations were made, where necessary, between the text and corresponding figures when the latter were not completely defined, yet contained data to be tagged. Curves and charts were also examined to determine if graph coordinates resulted in a type of numerical data of interest to users and therefore a candidate for tagging.

The primary objective of this experiment was to explore and evaluate the trade-offs between the usefulness of tagging in a CAS computer-readable service and the expense of gener-

0095-2338/81/1621-0073$01.25/0

Table I. Energy File Content

CA section no.	CA section title
50	Propellants and Explosives
51	Fossil Fuels, Derivatives, and Related Products
52	Electrochemical, Radiational, and Thermal Energy Technology
69	Thermodynamics, Thermochemistry, and Thermal Properties
70	Nuclear Phenomena
71	Nuclear Technology
72	Electrochemistry

ating the tags. An approach to the development of the tagging experiment was selected that permitted examination of different levels of tagging content and format without altering the content of the regular ENERGY file. Also, because all content of CA is computer processed, the experimental tags were added to the file record without separate, follow-on procedures.

ENERGY FILE CONTENT

CAS provides a variety of information services in computer-readable form. These files are designed to transfer information to computer searching systems which search and manipulate large quantities of information for rapid retrieval.

ENERGY is a computer-readable file of abstracts, bibliographic information, and keyword and volume index entries for documents abstracted in certain sections of CA (Table I). Issued every two weeks, ENERGY covers literature which reports on the chemical and chemical engineering aspects of energy sources, production, and technology.

This file consists of abstracts, document titles, names and affiliations (and/or work locations) of authors, patentees, and patent assignees, source document bibliographic citations, CA section and subsection numbers, keyword index entries, and volume index entries including CA index names, molecular formulas, and CAS Registry Numbers for chemical substances and CA General Subject Index entries for concepts. The evaluation packages of ENERGY used in this experiment also contained data tags in addition to the usual ENERGY file items.

PROCESSING OPERATIONS FOR DATA TAGGING

CAS Procedures for Data Tagging. Data tagging was carried out as an integral part of the CAS document analysis, information recording, and editorial operations in routine production. This provided a controlled experiment, whereby the identification of differences in efficiencies and costs introduced by the addition of data tagging to the processing efforts could be measured.

Guidelines and procedures were distributed to CAS document analysts working in the several component ENERGY sections, and the generation of data tags continued for slightly over one year. A new data element was created for the tag, and the encoded types of numerical data were directly input by the analysts. In addition to the usual manual editing for correct assignment of tags, various computer edits were also applied. Appropriate diagnostic messages were automatically displayed on terminal screens to aid the document analysts in resolving any errors detected by these edits.

Data Tags for the ENERGY File. Types of numerical data to be tagged in this experiment were selected by a set of users of the U.S. Department of Energy's (DOE) RECON information retrieval system. RECON participants were requested to assign priority rankings to data types found in the source documents referenced in the ENERGY file and to indicate what additional data types would be of value to them if tagged.

Table II. Examples of Types of Numerical Data for Tagging

activation energy	octane number
corrosion rate	radiation dose
current efficiency	radioelement half-life
enthalpy	solar cell related data
flash point	vapor pressure
neutron flux or fluence	viscosity
nuclear moment	zero charge potential

On the basis of the priority rankings and further recommendations from participating RECON users, 81 data types were selected for the initial tagging program. Examples are given in Table II. For provision of the most effective coverage possible with a limited list, "explicit" tags were assigned to more frequently appearing data types in the ENERGY file and "broader range" tags to groups of related types of data.

A follow-on effort to this initial tagging was conducted, and this included the tagging not only of the energy-specific types of data of primary concern to the participating RECON user group, but also 58 other types of numerical data recorded in the corresponding primary documents. As in the primary procedure, these additional data types were identified and coded in the routine document analysis process.

The experimental issues of ENERGY were delivered to the RECON operations center at Oak Ridge National Laboratory where they were translated to the generalized format. The CAS Standard Distribution File data elements were equated to the predefined set of data components and edited to a standard style so the original data were kept intact. The translated tapes were then processed for the RECON system. Indexes were created for the inverted and linear files. The related terms file consisted of a list of common names for chemical substances to provide fast and easy access to CAS Registry Numbers. For assistance in searching the ENERGY file as implemented on RECON, a guideline was prepared that described file format and content.

Searchable Fields for the ENERGY File on RECON. Two primary constraints guided the formatting and structuring of the ENERGY file on RECON. The first was the number of available descriptor fields and the length of these fields.

The second was the selection of searchable data elements from ENERGY with emphasis on data tag exploitation. The 10 available searchable fields were the following:

(1) *Data Tags.* Individual data tags were searchable so that all references containing a given type of data could be isolated and retrieved.

(2) *Information Tags.* Six types of information to be tagged (e.g., nuclear reactor safeguards and security, impacts of coal gasification on society, the economy, and the environment) were selected by RECON users. These concepts were restricted to avoid duplication of the content found in other parts of the ENERGY file.

(3) *Tagged Descriptors.* Chemical substance names, CAS Registry Numbers, or energy types were linked to the appropriate data tags. Thus, the user could define a descriptor/tag combination within a single search parameter. Retrieval indicated that the corresponding source document contained the required numerical data for the substance or energy type of interest.

(4) *Chemical Substance Index and General Subject Index Terms.* Names from the CA Chemical Substance Index and from the CA General Subject Index were derived from a controlled vocabulary. Controlled vocabulary was used when formulating searches. For constructing search profiles, one source of index terms was the CA Index Guide which provides cross-references, where necessary, from terminology used in the primary literature to the controlled indexing vocabulary. Synonyms and indexing notes are also included to assist the user in selecting search terms. An alternative procedure in

obtaining subject terms was to browse through the RECON inverted index that contained the list of descriptors for the ENERGY file. Often the desired term from the controlled vocabulary is displayed in alphabetic proximity to the input term. A name in the CA Chemical Substance Index is usually systematic and constructed from segments which correspond to the structural entities of that substance. A file was available to aid the user in deriving chemical substance search terms.

(5) *CAS Registry Numbers.* The CAS Registry Number is a unique code that identifies a chemical substance. Registry Numbers are independent of nomenclature, and their values have no chemical significance. Users were encouraged to use CAS Registry Numbers for searching—either alone or in conjunction with data tags—since systematic names tend to be long and complex. Registry Numbers were obtained from the CA Index Guide and from the online file described below.

(6) *Author Names.* Each author's name was searchable.

(7) *Journal CODEN.* The CODEN is a five-character code, unique and unambiguous, assigned to the titles of serial and nonserial publications.

(8) *Publication Year.* This field contained the year in which the document was published.

(9) *CA Section/Subsection/Section Cross-References.* The ENERGY file consists of seven sections of CA, each of which is divided into subsections. The contents of these sections and subsections were provided to the searchers. Since an abstract appears in only one CA section, cross-references are used to identify other sections in which the abstract is of probable interest. Sections, subsections, and cross-references were searchable.

(10) *Common Names for Chemical Substances.* For assistance in locating CAS Registry Numbers, a set of common names for substances and corresponding CAS Registry Numbers was prepared. This set was incorporated into the related terms file for ENERGY. The input of a common name for a substance resulted in the display of the number of descriptors (i.e., CAS Registry Numbers) related to that common name. The CAS Registry Number would then be used as a search term.

Data-Tag Distributions. A total of 139 data tags (81 tags for the first phase of the experiment, 58 for the second) were considered for assignment during the experiment. Ten were never used. Dropping these 10 tags from the list was considered advantageous from the viewpoint of the data base producer since it would decrease editorial analysis time. However, because numerical data are difficult to locate, the tagging of one of these very infrequently occurring data types—even if only once during a period spanning several years—may provide the user with information that would otherwise be lost.

The most frequently occurring data tags in one CA section differed from those of other sections, except, of course, if the sections were very closely related in types of studies referenced. Indeed, the most frequently occurring tag for each ENERGY section correlated very closely with the respective section title (Table III).

Data tags were assigned with the maximum degree of specificity. Thus a general or broad tag would not necessarily have a greater number of postings than a member of its hierarchical subclass. Automatic up-posting appears not to be useful, since tag relationships may be mathematically correlated rather than hierarchically associated.

EVALUATION OF DATA TAGGING AND DATA TAGS

Evaluation of the operational experiment consisted of two parts: an appraisal of the tagging procedure by CAS from economic and processing viewpoints and an assessment of data tag utility in searches of the ENERGY file.

Table III. Most Frequently Occurring Data Tags

CA section no.	CA section title	most frequently occurring data tag
50	Propellents and Explosives	combustion velocity (CV)
51	Fossil Fuels, Derivatives, and Related Products	sulfur content of fuels (SC)
52	Electrochemical, Radiational, and Thermal Energy Technology	battery-related data (BATT)
69	Thermodynamics, Thermochemistry, and Thermal Properties	enthalpy (ET)
70	Nuclear Phenomena	nuclear cross section (NX)
71	Nuclear Technology	radiation dose (RD)
72	Electrochemistry	electric potential (ELPT)

Evaluation of Data-Tagging Procedures. For more precise identification of the expenses incurred by CAS during the course of producing the ENERGY file with data tags, charges were separated into two categories: direct charges, including training, special problems, etc., and production charges, including routine identification, recording, and correction for incorporation of data tags into the data base. The total direct charges by the departments which carried out the tagging process were allocated to three functions: administration/training, language problems, and technical problems. One analyst from each department was designated to handle technical inquiries. Situations handled by these analysts involved (1) responding to questions pertaining to the tagging procedure in general, (2) explaining the properties encompassed by particular data tags, (3) resolving problems that resulted from the machine and manual editing of tags, and (4) performing periodic quality control checks to ensure the proper assignment of tags. As with any new procedure added to editorial operations, time was expended in the beginning for interpreting policies. This was especially true for data tagging, since in some instances the scope of coverage of a particular tag could be ambiguous. Also, the same types of data are not always expressed the same way. Synonymous terms had to be harmonized and variant units identified so that tags could be assigned correctly.

The effects of data tagging on CAS production rates were monitored throughout the experiment. Initially, the time necessary to analyze a source document increased approximately 30%. A significant increase was expected, since analysts were acquainting themselves with this added procedure and look-up time was substantial. However, as the process became more routine, the time necessary to identify the data types decreased. Within a 5-month period, the increased effort was reduced from the initial 30% to about 10% for the 81 data tags. A plot of increased production effort vs. time resulted in a hyperbolic curve. When additional data types to be tagged were added to the production stream, a new but similar curve would be derived. The time required for tagging depends on two main factors: the number of data types to be tagged and the frequency of occurrence of individual data types. While the former might be well defined, the latter would be more indeterminate and would impact significantly on any proposed tagging technique. Frequency of occurrence would also vary as different subject areas were evaluated for data content.

Utility of Data Tags. The second part of the evaluation focused on the utility of data tags in searches of the ENERGY file. This assessment was particularly important, since user feedback was needed to "tune" the level of tagging detail in the experiment as well as to provide guidelines to CAS in formulating plans for a potential routine tagging operation.

To aid in determining the utility of data tags in searches, RECON users participating in the experiment were sent a

questionnaire to record their evaluation of file structure, format, and accessibility. In addition they evaluated the potential of data tags for locating numerical values in the source literature. Analysis of the completed questionnaires revealed the following:

(1) *Usefulness of the Search Guidelines for Using the ENERGY File.* The 33-page guideline contained (a) reference to the standard RECON User's Manual, (b) a brief description of the file content, (c) a detailed description of the searchable fields and their access modes, (d) an annotated sample printout from the linear or display file, (e) a listing of the available data tags and information tags and their corresponding meanings, and (f) the subject coverage and arrangement of abstracts by sections for those seven sections included in the ENERGY file. All respondents said the guidelines were satisfactory and found the information on search access points and the annotated printout particularly useful.

(2) *Usefulness of Data Tags for Locating Numerical Values in the Primary Literature.* Responses were mixed on this point depending on whether particular tags were present in the data base for the substance(s) under investigation. This, obviously, was to be expected. Users indicated that the number of file records was too limited to conduct extensive searches and that the file did not go back far enough in time to contain frequently used data types for substances which were commonly encountered. A few users, who did not participate in the initial selection of data types, reported that either the data type or the substance was not on file.

(3) *Specificity of Tagging Detail.* When the list of candidate data types for tagging was initially formulated, it was evident that for the ENERGY file not all the properties which could have numerical data associated with them could be highlighted with an individual tag. The level of specificity conveyed by a tag was therefore dictated by the content of studies referenced in the file. Thus, for example, explicit tags were selected for thermal conductivity, thermal diffusivity, and thermal efficiency rather than one general tag for thermal properties. However, all data pertaining to the mechanical properties of materials were collected under one broad tag and thereby included such items as tensile strength, coefficient of friction, hardness, etc. It is reasonable to assume that tags selected for this file would be different from those for one of the other computer-readable abstract text files and also different from a choice of tags that would encompass the whole of CA content. Because of the method for determining items to be tagged (i.e., both specific and general candidate tags were submitted to RECON users who then selected those of primary interest for the initial set), all users were satisfied with the specificity of tagging. It was pointed out by the users that too many specific tags would overload the file and the formulation of search profiles would become too exacting. It was also indicated, however, that for the ENERGY file some queries dealt with specific data types but had to be answered in broader terms at this point.

(4) *ENERGY File Access Points on RECON.* No two data base processors will load an information file in an identical fashion. Each is constrained primarily by hardware limitations and, in some cases, by software capabilities. DOE/RECON was no exception. Storage capacity precluded the incorporation of keywords from the ENERGY file, and RECON software did not exist for extracting substantive words from titles for the inverted index. However, users stated that keywords and title terms would be an asset in searching and that molecular formulas would be a useful adjunct to the related terms file.

(5) *Additional Search Aids.* RECON users had little to offer in the way of additional search aids. It was suggested that an online dictionary of data types corresponding to each of the tags be made available. This could be implemented in a routine operation. All users stated that the CA Index Guide and its most recent supplement were readily available in the area where searches were performed. For most searching of index headings, this aid is the most valuable.

Because of the subjectivity involved in a file appraisal, CAS attempted to obtain as many user assessments as possible. However, this was not too successful. Users responded that the ENERGY file was competing with several others on RECON containing energy information that focused on aspects not included in the CAS file and that the other files were more established.

ENERGY-RELATED DATA TAGS IN OTHER CHEMICAL ABSTRACTS SECTIONS

In conjunction with the operational data-tagging procedure described above, related studies investigated the effects of tagging in areas of CA coverage not included in the seven sections encompassing the ENERGY file. These studies focused on the tagging of energy-related numerical data in the nonenergy sections of CA, the types of numerical data found across the whole of CA subject coverage, data types specific to a given CA section or group of sections, and the feasibility of incorporating energy-related data tags into nonenergy portions of CA.

In selected portions of CA coverage the frequencies of occurrence of certain data types were examined to identify those sections not included in the ENERGY file for which energy-related tags may be useful. Statistical samplings were designed to provide a quantitative measure of energy-specific data types as well as other kinds of data found in these sections. Data types which are common to specific CA sections were identified to determine if such sets could be used, for example, in building program-based edits for a routine data-tagging operation.

On the basis of these factors, the feasibility of incorporating data tags into these sections was evaluated. The economic and production aspects of adding data tags to CAS computer-readable files and printed publications were investigated. Specific factors included the cost of staff training, the impact on timeliness of coverage, and the potential usefulness of data tags in searches.

Energy Data Tags in Other CA Sections. CAS document analysts are specialists in the various subject areas covered by CA and familiar with the general content of studies in their fields. Hence these staff members were selected to identify various types of numerical data occurring in the 73 nonenergy CA sections. Energy-related data types submitted by the analysts were compared with those recommended by RECON users participating in the first part of the experiment. Those data types common to both are shown in Table IV along with the number of CA sections associated with each.

Lists of data types submitted by the document analysts were not exhaustive. In order for these lists to have been comprehensive, document analysts would have had to note hundreds of quantitative properties, some of which occur only rarely. The utility of establishing a data tag for infrequently occurring data types and identifying them in source documents in a routine tagging operation would be questionable. Such tagging would be extremely time consuming, since the primary literature would have to be examined even more closely. As the number of infrequently occurring types increased, the total number of data tags would become cumbersome and undesirable, especially since users have indicated that a large number of data tags would be unwieldy in searching.

As shown in Table IV, only 4 data types appeared in 10 or more nonenergy CA sections, and three of these (melting point, boiling point, viscosity) may be viewed as relevant to data users

Table IV. Energy-Related Data Types Included in the Nonenergy Sections of CA

data type	no. of nonenergy CA sections containing data type	data type	no. of nonenergy CA sections containing data type
melting point	37	refractive index	4
thermodynamic energy	28	heat capacity	3
boiling point	26	radiation dose	3
viscosity	10	electric potential	2
electric resistivity	9	compressive strength	1
electric conductivity	9	electric capacitance	1
diffusion coefficient	7	flame temperature	1
activation energy	7	nuclear cross section	1
enthalpy	6	nuclear moment	1
entropy	5	radioelement half-life	1
vapor pressure	5		

in other areas as well as to those concerned with energy data. Additionally, these tags are distributed over a wide range of CA sections. The most frequently posted section, Synthetic High Polymers (section 35), contained only 9 of the data types, while 43 of the 48 sections contained 5 or less. These results, however, were not unexpected since data types selected for tagging by RECON users were limited to a narrow subject area from the whole of CA coverage.

Identification of Data Types in Nonenergy Sections. The types of nonenergy numerical data imbedded in studies referenced in the 73 nonenergy CA sections were identified by document analysts. Only the more commonly occurring data types were recorded, since the total number of unique types would be extremely large.

Certain data types were common to a group of CA sections due to the similarity of subject matter. This is best illustrated by 12 of the CA organic chemistry sections since all contain the same types of studies but focus on different classes of compounds. It is expected, therefore, that some of the same physical properties would be reported throughout this grouping.

A more detailed analysis of similarities of data types within allied sections was conducted in three of the organic chemistry sections: Non-Condensed Aromatic Compounds (section 25), Heterocyclic Compounds (More Than One Hetero Atom) (section 28), and Alkaloids (section 31). Data sheets were prepared that listed types of numerical data found in these three sections. Analysts were requested to indicate the number of chemical substance names associated with each data type for each primary document when they generated the abstract and index entries for that document. This procedure avoided reanalysis of a document. Information from each data sheet was then encoded, keyboarded, and processed by machine.

Data-type frequencies were classified according to CA section and subsection with a secondary classification on publication type: journal article, review or state-of-the-art study, or patent. A total of 2377 documents were evaluated over a 4-month period. For the 3 sections analyzed, there were an average of approximately 25 occurrences of data associated with chemical substance names per document. This number increased to 32 when only those documents containing data were considered. The overall ratio of data types to chemical substance names was 1.25. Extrapolating these results to encompass only the 12 organic chemistry sections (sections 23–34) indicated that in a 1-year period about 65 000 occurrences of data would result.

Frequencies of occurrence may serve as a gauge to estimate roughly how long it may take to locate a given type of data in a source document. For example, analysts would have little or no trouble locating a data type which occurs very frequently since they would be expecting it and know what form these data would take. (A case in point would be melting points or boiling points in organic syntheses papers. Such data most likely would be found in the experimental section of the paper.) However, locating a data type which occurs very rarely would take more time, particularly in foreign-language articles. The units attached to this "obscure" data type would not be familiar to the analysts; hence more time would be required for reading and analyzing documents which may contain such a data type.

Editorial Data-Tagging Procedures. The cost of data tagging is directly related to the total number of data types to be tagged and their frequency of occurrence. However, there are editorial factors that are related to cost and must be considered if a tagging procedure were to be implemented in a routine production operation.

All document analysts must be aware of all data types to be tagged, even if such types are routinely confined to one given CA section or group of sections. For instance, enthalpy data are very likely to occur in studies referenced in CA section 69 (Thermodynamics, Thermochemistry, and Thermal Properties) but may also be located in others and mentioned only incidentally. Thus, analysts not involved in thermodynamic studies would still have to be aware of the enthalpy data tag. This situation is somewhat different from a routine indexing procedure. Usual entries for a subject index are derived from those aspects of a study considered by the author or indexer to be of significant interest. In most instances, this is not the case with numerical data, because such data are produced to substantiate theories or experimental methods. Specific types of numerical data thus range from the very obscure and narrowly defined to those which are general and well-known.

Analysts must also be able to recognize and equate variations in units with data types to be tagged. Similarly, it may be that a type of data to be tagged is indicated in a document—neither in the text nor unambiguously listed in a table—as an inflection on a curve. Analysts must recognize and evaluate the coordination of the plotted variables that produce this point.

It follows, therefore, that data types to be tagged must be selected and defined so that ambiguities in application and use are avoided. A type of data may have one connotation or meaning when applied to one section of CA and a different connotation in the context of another section, since CA encompasses the entire field of chemistry and chemical engineering.

The experiment has shown also that the scope of coverage of given data tags may have to be modified after a routine tagging operation has been underway for a period of time, if it is shown that such tags have only limited utility. Tags encompassing broad or general data types may have to be subdivided if too large a number of postings occur and hamper search retrieval precision. Infrequently used data tags would slow down the assignment process, since an analyst would have to check each source document to ascertain their presence.

Finally, experimental results indicate that if a tagging operation were implemented across the full coverage of CA, it would have to be done incrementally. An overall coordination of such an effort would be essential to ensure that tags were applied consistently and uniformly and that no disruption of routine document analysis would occur.

CONCLUSIONS

The experiment in data tagging conducted by CAS evalu-

ated the operational procedures needed for a data indexing scheme. It was established that a tagging mechanism could be incorporated into the routine production operations of document analysis. Important elements to be considered in such a data-tagging technique are a combination of the number of types of numerical data to be tagged as well as their frequency of occurrence in the source documents.

ACKNOWLEDGMENT

This work was supported in part by the National Science Foundation under Grant DSI75-03491 and was conducted in cooperation with the Department of Energy/Oak Ridge National Laboratory.

REFERENCES AND NOTES

(1) "Data Needs for Energy", CODATA Bulletin No. 31, International Council of Scientific Unions, March 1979.
(2) Committee on Data Needs, Numerical Data Advisory Board, Assembly of Mathematical and Physical Sciences, "National Needs for Critically Evaluated Physical and Chemical Data", National Academy of Sciences, National Research Council, May 1978.
(3) "Evaluating Quantitative Data Information for Research and Development", Numerical Data Advisory Board (NDAB) pamphlet (undated).
(4) "Flagging and Tagging Data", CODATA Bulletin No. 19, International Council of Scientific Unions, June 1976.
(5) Murdock, John W. "Current Knowledge on Numerical Data Indexing and Possible Future Developments", Informatics, Inc., Rockville, MD; report 1978; order No. BP279924, 58 pp. (English) Available from NTIS.
(6) "Introduction", *Chem. Abstr.* **1980**, 93(1), ix.

Chemical Information Resources Directory: An Integrating Component of the Chemical Substances Information Network[1]

RONALD A. RADER* and SHASHI P. SOOD

Environmental, Information, and Safety Systems Department, MITRE Corporation, Metrek Division, McLean, Virginia 22102

Received November 3, 1980

The Chemical Information Resources Directory (CIRD) serves as an integrating mechanism for the diverse information resources to be included in the Chemical Substances Information Network. The present CIRD, available as a published handbook with its two parts, the Subject Catalog and the Descriptive Catalog, uniformly catalogs and indexes the functions, content, access, and other attributes of 53 chemical-related information resources. The CIRD will soon be available as part of the prototype Chemical Substances Information Network (CSIN).

Implementation of the Toxic Substances Control Act (TSCA) requires that the most reliable information and data be available to regulate the risks associated with the manufacture and use of chemical substances. Section 10b of TSCA mandates the design and establishment of an effective system for managing and disseminating data submitted under the act and other information which could be useful in its implementation.

Subsequent to the passage of TSCA, the MITRE corporation was contracted jointly by the Council on Environmental Quality, the Environmental Protection Agency, and the National Library of Medicine to conduct a study to identify the information needs of the various institutions implementing and affected by TSCA and investigate approaches required to store, integrate, and provide access to information as mandated by the act.[2] Development of a Chemical Substances Information Network (CSIN) was recommended.

The CSIN concept includes a collection of core component information resources integrated in a computer network, so that they would be accessed automatically and appear as a single resource to the user, and a set of noncore (noncomputer network) component information resources which would be referenced through the network, but not directly accessible through it. CSIN, it is expected, would provide users with convenient means to identify and access diverse chemical-related information and data systems. To accomplish this, two major components of CSIN were envisioned—the Chemical Structure/Nomenclature System (CSNS) and the Chemical Data Bases Directory, subsequently termed the Chemical Information Resources Directory (CIRD).

The CIRD and CSNS would function as integrating mechanisms for the diverse information resources to be incorporated in the CSIN. These data bases would serve as locaters and support such advanced computer network capabilities as automatic user query analysis, automatic selection and querying of relevant data bases, and the integration and coordination of data retrieved from multiple computerized data bases. These locater systems provide two different tools to access CSIN information resources. The CSNS would help the user identify relevant information resources in CSIN on the basis of chemical structure, substructure, and nomenclature of substances referenced by component information resources. The CSNS would incorporate many of the present capabilities of CHEMLINE of the National Library of Medicine's online network and the Structure and Nomenclature Search System (SANSS) of the NIH–EPA Chemical Information System.

The CIRD, on the other hand, would provide the user with a tool which would identify information resources on the basis of their characteristics and subject content and provide descriptions of each resource. With the help of the CIRD, the user could identify and prioritize information resources to be accessed for information and data relevant to his needs. Presently, both locater systems are being developed independently for incorporation in a prototype CSIN computer network. The discussion below will confine itself to CIRD development.

The CIRD in its current published handbook version presents detailed characterizations of 53 information resources (listed in Table I).[3] An information resource is defined from the user's point of view—it is the interface between the user

0095-2338/81/1621-0078$01.25/0