Technical Features of the Chemical and Materials Property Data Network Services on STN International

V. J. Drago

Chemical Abstracts Service, 2540 Olentangy River Road, P.O. Box 02228, Columbus, Ohio 43202

J. G. Kaufman*

MPD Network Inc./Chemical Abstracts Service, 2572 Olentangy River Road, Suite 5A, P.O. Box 02224, Columbus, Ohio 43202

Received July 4, 1992

A number of technical innovations in the search and retrieval of numeric chemical and material properties data, identified in previous expert studies, have been incorporated into two new data services: the Chemical Property Data (CPD) Network and the Materials Property Data (MPD) Network on STN International. From either a PC or terminal, these services provide users easy, single-point access to hundreds of properties and over 25 000 chemicals, 20 000 materials, and 20 databases. The menu interface capability built into the CPD and MPD services enable the user to search each database cluster like a single file, with no requirement that they know which databases are most likely to contain the data. A number of software features designed especially for manipulating numeric data are incorporated into the system, including range searching, units conversion, tolerance setting, rounding and significant figures setting, calculation packages, and a wide range of data display options.

1. INTRODUCTION

At the Second International Symposium on Computerization and Networking of Materials Databases in December 1989, in Orlando, FL, some new approaches to increasing the responsiveness of computerized numeric property databases to users' needs and expectations were discussed.[1] A number of these approaches have been incorporated into the design of new data services developed by the National Materials Property Data Network (NMPDN), Inc., and Chemical Abstracts Service (CAS) of the American Chemical Society (ACS): the Chemical Property Data (CPD) Network and the Materials Property Data (MPD) Network on STN International.

This paper describes the approach and the technical features of these systems that contribute to improved responsiveness to user needs in searching for the properties of chemicals and materials.

The scope of the paper thus includes:

1. An overview of user needs
2. A description of property data
3. Technical features of the CPD and MPD Services
   a. Menu interface
   b. Cluster searching
   c. Numeric search software
   d. Logic-based presentation

A detailed search example illustrating the expertise built into the system software is provided at the end of the paper.

2. ADDRESSING USER NEEDS

In laying out the plans for the CPD and MPD services for providing easy online access to worldwide sources of reliable chemical and material property data, we carefully considered the information we had gathered on the needs of the property data users. This information came from a number of workshops that were held jointly with our associates in the Standard Reference Data Program at the National Institute for Standards and Technology (NIST) and from a pilot MPD Network operation that was run for 4 years from Stanford University.[2] From the experience gained, we not only learned much about the specific needs of scientists and engineers but also gained some insights into the broader interpretation of searchers' needs in addressing computerized systems of all types.

The primary need is for single-point access to a significant amount of quality data. Decisions on some scientific or engineering matter are needed and, indeed, will be made with the best data that can be readily obtained in the time available. The motivation is to find the highest quality values of the needed properties or performance measurements. Most technical decisions cannot be delayed too long and must be made within the limits of the available time with the best available data. The information source must have enough properties for enough chemicals or materials to make it worth the effort to conduct a search. Without this "critical mass" of data, users will not often find what they want and will lose interest in that source.

Once a substantial data resource or aggregate of resources is available, the next important need (and often a principal motivator) is ease of access. The selection of one of several candidate sources may well be based upon relative convenience and support available. Regrettably, even the quality of data seems secondary at this level. Difficulties in accessing a source (e.g., the complexity of telecommunication connection to search online sources or getting "hung up" or stranded in search and retrieval software) may be sufficient to cause the search to be abandoned and reliance placed upon readily available (though lesser quality) data.

If problems of access to either local or online sources are solved, this level of need is succeeded by one relating to the interface software for the search and retrieval system. Relative motivation to use system A or system B is likely to depend upon the relative simplicity, clarity, and flexibility of the search software for the candidate systems. An easily understood and logical menu-driven search software may well be sufficient

0095-2338/93/1633-0046$04.00/0

to get the user to choose one system over another similar source for which searching involves use of a command language requiring training and frequent use for retention.

If concerns of access and interfacing are met satisfactorily, users' motivations shift to factors of system support. A broad and valuable glossary interfacing with the search and retrieval system, for example, may provide the added motivation for a user to select one system over another. Finally, if all other needs are met, the searcher's concern shifts to what may or must be done with the data, and the selection of one data source over another may well be based upon the presence of software packages for analytical treatment, graphics, or computer-aided design and manufacturing.

The important point to take from this illustration is simply that, in order to increase the likelihood of user acceptance and regular utility of a scientific data system, we must consider all levels of the needs. In addition, it is useful to note that no matter how sophisticated the software package or expert system, it is of no value without the foundations below it, principally a strong body of reliable data.

3. NATURE OF NUMERIC PROPERTY DATA

Numeric property data are notably different from textual data in several basic respects, as detailed earlier.[3-5] First the obvious: They are quantitative and thus have implied precision, come in ranges as well as discreet values, and may vary over many orders of magnitude within a single data record containing four or five properties. Second, they have units associated with them from which they can never be separated and still retain meaning; and once again, the numbers in any one record may have multiple units, some of them extremely complex. Finally, they are usually dependent upon a number of independent variables or combinations of variables, usually called parameters, which influence their value or behavior in certain situations.

These factors all make the handling of numeric data more complex than the handling of textual data, which typically is searched for "strings" of characters without other modifiers except perhaps other strings that may be added together in various boolean modes. Numeric data must be stored and searched keeping in mind at least three major elements—name of the property, its value, and the units of that value—and often more. For example, the property may be dependent upon time of exposure at a particular combination of temperature and pressure, and so three more factors—time, temperature, and pressure—must be included as delimiters in every query about that property.

4. CPD AND MPD NETWORK SERVICE ON STN INTERNATIONAL

The National Materials Property Data Network (NMPDN), Inc., was formed in 1985 with the mission of providing easier access for engineers and material scientists to reliable materials properties. In 1987, the NMPDN began working jointly with Chemical Abstracts Service (CAS) to implement and distribute the MPD Network service on STN International. The prototype production version of MPD Network Service became available early in 1991.

In December 1991, a second service, the Chemical Property Data (CPD) Network service, was implemented on STN International as a complement to the MPD service. The CPD service provides physical and chemical property data for a wide range of chemicals.

These two services, CPD Network and MPD Network, are referred to as "networks" because, within each, the databases are networked physically by the search software and intellectually by the search strategies and the associated thesauri, which cross-link the terminology in the various databases. Thus, they provide database network services within the context of the broader framework of the scientific and technical network known as STN International.

STN International is the premier online scientific and technical information network operated jointly by CAS, a division of the American Chemical Society, in Columbus, OH; FIZ Karlsruhe, a scientific and educational organization in Karlsruhe, Federal Republic of Germany; and Japan Information Center of Science and Technology (JICST) in Tokyo, Japan.[6]

5. BASIC DATABASE NETWORK SYSTEM

Through the CPD and MPD services, STN International provides single-point access to a significant volume of property data represented in about 20 entirely separate and distinct databases. Taken as a group, the combined services provide access to hundreds of properties and over 25 000 chemicals, 20 000 materials, and 20 databases. The coverage is being extended continuously, both for chemical substances and materials as well as properties.

The CPD service includes such classes of properties as thermodynamic, electrochemical, spectral, safety, and transport. Among the specific databases currently available through the prototype version of the CPD service are the following:

DIPPR—from the AIChE—textual and numeric information on the pure component physical property data for commercially important chemical substances

HODOC—an electronic version of much of the *CRC Handbook of Data on Organic Compounds*, including the chemical and physical properties of over 25 000 organic substances

JANAF—critically evaluated thermodynamic properties from the Joint Army–Navy–Air Force Thermodynamic Tables

NISTTHERMO (formerly NBSTHERMO)—the NIST Tables of Chemical Thermodynamic Properties, containing critically evaluated chemical thermodynamic properties of over 8000 inorganic and organic substances

NISTFLUIDS (formerly NBSFLUIDS)—the NIST calculation package covering the thermophysical and transport properties of fluids as a function of temperature and pressure

TRCTHERMO—from the Thermodynamic Research Center—the evaluated thermodynamic properties of 7000 compounds.

In addition, and partially transparent to the user, the CPD service makes significance use of the NUMERIGUIDE and CAS REGISTRY databases to identify properties and chemicals, respectively, and to locate the databases that will likely contain the data the user desires. NUMERIGUIDE is property thesaurus and data directory that provides

1. Property names and hierarchies
2. Definitions/descriptions
3. Aliases
4. Abbreviations
5. Identification of databases that address each property

6. Default units
7. Variables associated with property (e.g., temperature)

The REGISTRY database provides a reliable means for searching chemicals by chemical name, molecular formula, or CAS Registry number. REGISTRY also provides identification of databases that address each chemical.

Numeric information available on the MPD service includes not only mechanical, physical, and other performance data for all structural materials, including metals, polymers, ceramics, and composites, but also the properties of connections and joints in these materials.

Among the specific databases available through the MPD service are

AAASD—from the Aluminum Association—typical and minimum tensile properties, typical mechanical and physical properties, and fabricating and application information on more than 150 commercial aluminum alloys

ALFRAC—from NIST/SRD, the Materials Properties Council, and the Aluminum Association—plane strain fracture toughness data for about 25 high-strength aluminum alloys, with validity documentation

COPPERDATA—from Copper Data Association—mechanical, electrical, thermal, and other physical properties of 166 wrought and 95 cast U.S. coppers and copper alloys

IPS (International Plastics Selector)—from D.A.-T.A. Business Communications, a division of IHS

MIL-HDBK-5—from the MIL-HDBK-5 Coordination Committee publication—design tables covering the design mechanical and physical properties of ferrous and non-ferrous alloys

MARTUF—from Materials Property Council—about 10 000 individual toughness test results for steels for marine applications

METALS DATAFILE—from Materials Information (ASM International and the Institute of Metals)—data from more than 40 000 literature citations from technical journals

MPDSEARCH—an international directory of sources for material property information that contains both databases and data centers–organizations who will research information for you

NISTCERAM—from NIST—properties of structural ceramics–silicon carbides and silicon nitrides

PDLCOM—from Plastics Design Library—chemical and environmental compatibility of plastics

PLASNEWS—a daily news file for the plastics industry

PLASPEC—from the publishers of *Plastics Technology*—typical properties from producers of about 10 000 plastics

STEELTUF—from the Electric Power Research Institute and the Materials Properties Council—results of more than 20 000 individual tests of steels for the power and petroleum industries

There are no hardware or software requirements in using these network services beyond the need for a modem and some telecommunications software. Searchers may use their PC or terminal together with a modem and a telephone connection or may obtain specially designed PC software from the supplier.

7. MENU INTERFACE

As noted earlier, the individuals most likely to have very detailed queries for specific performance data are the end-users who are not familiar with the sophisticated command languages of most online systems. NMPDN and CAS have developed and implemented interfaces for both the CPD and MPD services that are easy to use, even for someone with no previous online experience. Among the major features of these interfaces are

1. Very logical, easy-to-use, menu-driven search paths
2. A variety of search paths, recognizing that different types of users and different applications will require different queries
3. A "metadata" system in the form of an interactive thesaurus that both deals with user queries, translating them to all other acceptable nomenclature and terminology, and responds quickly to clarfy the meaning of names, terms, and abbreviations
4. The ability to define and include in the search variables associated with both the substances and properties of interest
5. A directory of data sources, including those outside the services

By nature of the detailed searching done in numeric files in contrast to the broader subject searching done in textual files, thse menus are much more detailed and sophisticated than those used to guide searchers to appropriate bibliographic files. They truly are "expert systems" of a sort, capable of taking the basic input from the user, determining the proper databases to search, and building the appropriate command-language searches required to conduct the search.

8. CLUSTER SEARCHING

The historical approach to database access and searching is based upon the assumption that each database stands alone and is accessed and searched largely independently of the others. With numeric properties databases, which each tend to be relatively small and well-focused, this approach is often not very useful for several reasons. First, a number of databases may need to be accessed to provide all the property data required for many applications. Second, the user may not know which of several databases will be most responsive to the particular needs at any given moment. Thus, to best serve the needs of searchers for properties data, it has proven desirable to link many databases together into a logical "cluster" and permit easy cross-file searching of those files.

Depending upon the nature of a query, CPD and MPD service users may approach the database cluster with different pieces of information at the heart of their query:

1. A specific database with a certain type of data, e.g., design values
2. A specific chemical or material for which a variety of types of data is sought
3. A specific property or properties for which a comparison of chemicals or materials is required, perhaps involving a specific range of values, notably those equaling or exceeding certain limiting values

The initial menu screen (Figure 1) provides the user with a versatile, easy-to-use "blank slate" on which users can build, track, or modify their search specifications.

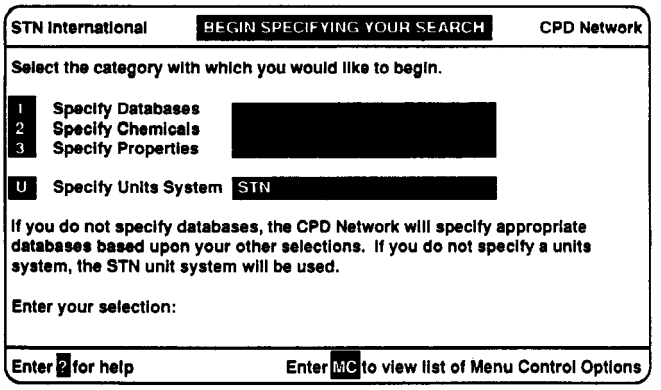

Figure 1.

The CPD and MPD services enable the user to search the database cluster simultaneously, with no requirement that the user know which databases are most likely to contain the data. Further, from such a cluster, all reasonable data sets responsive to the query will be provided, always with clear evidence of why they were provided and from where they have come along with all of the associated parametric support data.

9. NUMERIC DATA SOFTWARE CAPABILITIES

In order to provide search and retrieval software that deals with the complexities of numeric properties data, NMPDN and CAS have over the years enhanced the STN search and retrieval software to provide certain capabilities that are especially useful in meeting users' needs in handling numeric data. All the features listed below are available on the CPD and MPD services.

1. Range searching—the ability to search for chemicals or materials with combinations of properties in specific ranges or above or below certain limiting values
2. Units conversion—the ability to convert to any of the worldwide standard systems of units—International Standard (SI), meter-kilogram-second (MKS), centimeter-gram-second (CGS), engineering (ENG), and the STN user-friendly SI system
3. Tolerance setting—the ability to define ranges of search values by the tolerance on the search value (e.g., 50 000 ± 1%)
4. Table display—the ability to obtain tabular display of data that match your specific query and to predefine certain types of tabular displays
5. Calculation packages—the capability for interpolation and estimation of additional information or the application of parametric analysis of multivariant properties, valuable in providing specific answers to some complex materials questions.

10. PRESENTATION FORMATS

Property data are typically sought to support some analytic use in a specific application, perhaps ranging from developing a new, improved chemical or material to designing some specific structure. The formats in which numeric data are displayed may readily affect their usefulness for the intended application and so are of more that passing importance. To facilitate the analysis efforts of data users, the CPD and MPD service provide a wide range of data display options that take the following factors into account:

1. Types of users—engineers and scientists tend to expect data to be presented in a logical, use-oriented tabular or graphic display rather than bunched together or in a long, disorganized litany. Wherever appropriate, property data are presented in tabular matrix displays.
2. Types of queries—special display formats, known as "query-related" displays, are utilized to focus specific properties and the variables or parameters that define the applicability of the data. Searchers do not have to wade through extraneous data, yet they have all of the pertinent related information.
3. Number of answer sets—brief summary displays have been designed to allow searchers to scan large numbers of records to locate those of particular interest, which can then be displayed more fully.
4. Documentation required to define the source and applicability of data—as noted earlier, a number of variables such as time and temperature, or in the case of the materials, form and dimensions, are often required to define the range of utility of specific data. Displays are designed to accommodate reporting of these variables along with the specific data requested.

The provision of supporting documentation in the displays is a particularly important point that is often taken for granted or omitted completely in reporting data. This may go well beyond what a specific user requests. Basic facts such as the type of data (design value or individual test result) and the applicable orientation (longitudinal or short transverse) are included.

11. ILLUSTRATIVE EXAMPLE

In order to illustrate the use of the menu-driven cluster search approach employed in these services, consider the scientist needing specific numeric values for the heat of formation of phosphorus trifluoride for temperatures from 200 to 400 K. Locating such data from a bibliographic search is likely to be iterative and require several hours of ones time, in contrast to the CPD Network approach of 2 min to a direct answer.

Starting with a menu interface of the type in Figure 1, the searcher is given a choice of paths with which to start. Assuming no knowledge of which database may contain the answer, the searcher is likely to use a combination of specification of substance and property. The substance may be identified by chemical name, formula, or CAS Registry number, and the system provides an expand in Registry to enable the searcher to select from the varieties of type of substance for which data are available (Figure 2).

Upon specification of heat of formation as the property, the system searches the NUMERIGUIDE thesaurus, advises the searcher that the proper search term is enthalpy of formation (Figure 3), and builds that into the query (Figure 4). A search is requested by the searcher, and the system checks to determine which databases are likely to have the substances and properties requested, searches the appropriate files, and reports the findings (Figure 5). Of the three databases containing the substance and the property requested, answers were found in two of the three, and the searcher may elect to look at one or all of the results, leading to specific answers to the query (Figure 6). One of the answers had a different chemical name than the one searched, and the other answer (Figure 6) had no chemical name at all. With the CPD network, once a chemical name is selected, the system determines the corre-

```
STN International        CHEMICAL NAME SELECTION        CPD Network

The menu system is now consulting the Chemical Name index of the REGISTRY
file. Select names to include in your search, by numbers or ranges (e.g. 1, 3-5).

   Count  Chemical Name
   -----  -------------
1   1       PHOSPHORUS TRICYANIDE
2   1       PHOSPHORUS TRIETHYLENIMINE
3   1 —>    PHOSPHORUS TRIFLUORIDE
4   1       PHOSPHORUS TRIFLUORIDE (32PF3)
5   1       PHOSPHORUS TRIFLUORIDE DICHLORIDE

L  See continuation of name List without selecting names listed above
P  Return to CHEMICAL NAME SPECIFICATION screen

Enter your selection:3

See continuation of chemical name list? (Y/N):N

Enter ? for help              Enter MC to view list of Menu Control Options
```

Figure 2.

```
STN International        PROPERTY RECORD EXAMINATION        CPD Network

For your entry, HEAT OF FORMATION, the preferred term in the NUMERIGUIDE
thesaurus is enthalpy of formation.

This is the term that will be used in the search.

The CPD menu system is now examining the NUMERIGUIDE database record to
determine how this property may be searched.

You will be given a series of choices, depending on how your property is
reported in the databases of the CPD Network.

Enter ? for help              Enter MC to view list of Menu Control Options
```

Figure 3.

```
STN International        CURRENT SEARCH STATUS        CPD Network

Your current search terms are

1  Databases  SYSTEM WILL SELECT DATABASES.
2  Chemicals —
      PHOSPHORUS TRIFLUORIDE
3  Properties —
      enthalpy of formation (any value or range)
        Temperature 200-400 K

Select category number to insert, replace, or delete search term, or select

PV  Delete a Property Variable
U   Specify Units System  SI
S   to conduct search.

Enter your selection:S

Enter ? for help              Enter MC to view list of Menu Control Options
```

Figure 4.

```
STN International        NOW SEARCHING        CPD Network

Now searching database REGISTRY for selected chemicals.
  - Found 1 chemical(s) contained in CPD Network databases.
Now determining the CPD Network databases to search.
The following 3 database(s) will be searched.
JANAF     NISTTHERMO   TRCTHERMO

Now searching database JANAF           1 RESULT
Now searching database NISTTHERMO      1 RESULT
Now searching database TRCTHERMO       0 RESULTS

Enter GO to continue:GO

Enter ? for help              Enter MC to view list of Menu Control Options
```

Figure 5.

sponding CAS Registry number, and the search is conducted using the corresponding CAS Registry number. Therefore, even if a database uses a different chemical name than the one selected, or has no chemical name, the chemical is located.

```
STN International        *** Results 2 2 (NISTTHERMO) ***        CPD Network

L9   ANSWER 1 OF 1

RN      ***7783-55-3***   NISTTHERMO
MF      F3 P
LSF     PF3
SSTA    Ideal Gas

LSF     PF3

RE      J. Phys. Chem. Ref. Data Vol. 11, Suppl. )
Standard State Pressure = p = 0.1 MPa
Enthalpy Reference Temperature = Tr = 0 K
```

State	Enthalpy of Formation	Enthalpy of Formation	Gibbs Energy of Formation	Enthalpy	Entropy	Heat Capacity
SSTA	HFOR 0 K J/mol	HFOR 298.15 K J/mol	GFOR 298.15 K J/mol	HREF	SREF	CP
IG	-913160	-918800	-897500	12937	273.24	58.70

```
Enter GO to continue
  R to Return to the SELECT DISPLAY OPTION screen:
```

Figure 6.

If these databases would have been searched directly, without the CPD Network, using the name phosphorus trifluoride, neither answer would have been found.

Note the points of system expertise active in even this simple example:

1. The use of an expand in Registry provides numerous versions of the substance name requested for possible choice and determines the corresponding CAS Registry number for the substances.
2. The use of the thesaurus assures the use of the proper search term, even though the searcher's terminology may differ from that in the database.
3. The system prepares the correct command-mode search terms.
4. The system selects the correct databases for a search by combining lists for the substance and the property searched.
5. The system provides a list of answers available from all sources and permits the searcher to select the ones desired.

This is a substantial assist to the searcher, especially the occasional searcher unfamiliar with the structure and commands as well as the content of a comprehensive online catalog of databases.

12. SUMMARY

In developing the CPD and MPD Network services on STN International, an innovative approach was taken to addressing the needs of engineers and scientists for easy access to and flexible interaction with worldwide sources of numeric chemical and material property data. The distinct added complexities of storage, search, and retrieval of numeric property data have been addressed with special search software and a user interface to aid the scientists and engineers who are the end-users of such highly technical information. Provision has been made to ensure that a variety of types of users and their queries may be handled with ease on the system and that the confusion of names, terms, and abbreviations can be removed.

REFERENCES AND NOTES

(1) Kaufman, J. G. Increasing Data System Responsiveness to End-User Expectations. *Computerization and Networking of Materials Databases*; ASTM STP 1106; ASTM: Philadelphia, 1991; Vol. 2, pp 103–112.

(2) Grattidge, W.; et al. *Materials Information for Science and Technology*; NBS Special Technical Publication No. 726, National Bureau of Standards: Gaithersburg, MD, Nov 1986.
(3) Kaufman, J. G. The Physiognomy of Numeric/Factual Chemical and Material Property Data. Presented at the American Chemical Society Spring Meeting in San Francisco, Apr 6, 1992. *J. Chem. Inf. Comput. Sci.* **1993**, in this issue.
(4) Kaufman, J. G.; Drago, V. J. Direct Online Access to Numeric Chemical Properties Data. Online Information 91, Dec 10–12, 1991, London, England.
(5) *Database Systems in Science and Engineering*; Rumble, J. R., Jr., Smith, F. J., Eds.; Adam Hilger: Bristol, 1990.
(6) *STN International—Databases in Science and Technology*; American Chemical Society: Columbus, OH, 1987.