Physiognomy of Numeric/Factual Chemical and Material Property Data

J. Gilbert Kaufman

National Materials Property Data Network, Inc., Chemical Abstracts Service, 2540 Olentangy River Road, Columbus, Ohio 43210

Received June 4, 1992

The rapid pace of development of new materials and new products of various kinds in recent years has led to much greater need for computer access to reliable chemical and materials property data. The growth in effort to satisfy that need has in turn focused attention on the great differences between bibliographic data, historically the basic of most computer-searchable databases, and the numeric or factual types of data. In introducing this ACS Symposium on Numeric/Factual Chemical and Materials Property Data, we will look specifically at the nature of some types of numeric data, noting the characteristics which make them different from textural data, and the added features needed in search and retrieval systems to meet users' needs and expectations about how to deal with them. Among the key characteristics are (1) the need to be searched as numbers and ranges of numbers, not strings; (2) they may range over many orders of magnitude in a single record; (3) they may have varied complex unit systems associated with them; (4) they are multivariant; and (5) they require considerable factual support. Among the software features required to deal with these characteristics are range searching, units conversion, tolerance setting, tabular formatting, and interactive thesauri. End-user searching is best augmented with a menu-driven "intelligent" interface which aids in the selection of databases, query building, and terminology.

INTRODUCTION

Activity has intensified in the past 10 years on creating scientific databases that provide direct access to quantitative information about the performance and behavior of chemical substances and of the materials made from application of the chemical sciences to such substances. This is in marked contrast to the earlier practice of concentration upon textural databases which provide bibliographic references to the literature from which such information might be determined. The shift resulted from the simple fact that searching the bibliographic literature for such quantitative information and the required supporting factual data may be a long iterative process, and in the early 1980's there was a broad challenge to improve means of making such quantitative information available.[1-6]

Despite the clarion call, and while much has been done in response, it has been and continues to be a much more complex (and therefore time-consuming and expensive) task than envisioned. In fact, even today, many of the specific complexities are still not widely understood, and so we will address them in this paper and in those which follow. We will take advantage of the great strides which have been made but will also identify the problems and gaps which remain.

Physiognomy may be an unusual, perhaps odd, term to use in this study about the characteristics of numeric properties data. Webster's New World Dictionary provides three meanings to the term, summarized as follows:

1. practice of trying to judge character and mental abilities by observation of bodily, especially facial features
2. facial features and expression, as supposedly indicative of character
3. apparent characteristics

The third definition certainly justifies the use, but there is a stronger reason for the use of the term. I choose to describe most if not all of the features in terms representing how numeric properties data look and feel to the user and will cover the specialized tools that users need to deal with them.

NUMERIC/FACTUAL DATA

For purposes of this paper, I will be using a definition of numeric/factual data as quantitative data with the supporting textural facts, and in general I will refer to it simply as numeric data. Of course, it is not strictly speaking limited to numeric data, but as the definition above implies (and we will emphasize later to a greater degree) it includes additional factual information in whatever format needed to describe the usefulness or limitations of the numeric part of the data.

In the material that follows, we will examine the following features of numeric properties data:

- Focus of numeric/factual data and how they are used
- Specific nature of numeric scientific data
- Nature of the users of numeric data and their needs
- Differences from bibliographic/textural data

FOCUS OF NUMERIC DATA AND HOW THEY ARE USED

With the substantial exceptions of financial and census data, scientific numeric data tend to focus primarily upon the characterization of materials in the broadest sense, that is materials ranging from pure chemicals and chemical compounds to highly developed finished products made from them such as pharmaceuticals or structural alloys. Numeric data are effectively used to describe almost all aspects of these materials, ranging from how they are produced to how they behave in interfacing with (possibly reacting to) other substances and how well they perform whatever tasks they were created to perform. These characteristics fit the "properties" terminology in one sense or another and, thus, form the focus of the paper.

The need for and use of extensive detailed numeric data is not limited to highly scientific chemical research; it is also important in the development of the downstream products made from the original substances and to their engineering and design into real commercial products. In fact, the case can be made that because of the dependence of the properties of most finished products on so much of the chemical and

0095-2338/93/1633-0002$04.00/0

production technology used to develop and produce them that the need for and use of scientific and technical numeric data is even greater in the downstream applications.

Specifically, then, numeric scientific data are needed for purposes as broad as research to produce them, commercial production process development, commercial process or application engineering, quality control, and, all too often, failure analysis and trouble-shooting. For all stages of the process in commercial exploitation of chemical research, detailed scientific numeric data are the key.

Therefore, in our look at the physiognomy of numeric property data, we will recognize a variety of types of data and a variety of applications of it. In describing the properties of materials (chemical substances or commercial alloys), there are several specific types of specialized quantitative numeric/factual data, most notably spectrographic and crystallographic data, that are treated separately in this symposium. While specialized, they are still bound by many of the requirements for delimiting detail and factual support described for broader classes of scientific data, and more is said about them elsewhere in this symposium.

NATURE OF NUMERIC SCIENTIFIC DATA

There are at least five major areas of characterization that need to be recognized:

Quantitative nature
Complexity
Multivariance
Significance level
Graphical impact

Quantitative Nature. The overriding characteristic is the numeric nature of the data themselves. To be of maximum use, they must be tabulated in the database as numbers and be capable of being manipulated as numbers independent of the associated text, without losing the delimiting factual support information about which more will be said later.

Implied in the requirement for treatment as numbers is the capability to (a) be recognized and queried as higher or lower than other numbers or falling into specific ranges; (b) be handled in terms of tolerances, rounding, and significant figures; (c) be used in calculations; and (d) be used in plotting of one property versus another or as a function of some variable such as temperature. More will be said about several of these characteristics and capabilities later.

Complexity. Numeric scientific properties files are more complex than textual data as well as other types of numeric (e.g., financial or census) data because they have units which must be tracked and many different units within a single data unit or record. In addition, the numeric values themselves may vary by many orders of magnitude within a single unit, often requiring the use of scientific exponential notation. Consider the following rather simple and common combination of properties for a material:

density	0.10	lb/in^3
coefficient of thermal expansion	6.1×10^{-6}	in/in/°F
thermal conductivity	48.2	BTU[(hr)ft^2 °F/ft]
modulus of elasticity	10.2×10^{-6}	psi
ultimate strength	75.6×10^{-3}	psi

Within this single simple record, 12 orders of magnitude and a great variety of units representing several unit classes are represented. These units must always be linked to the data; they are meaningless without them.

Multivariance. Most values are dependent upon a number of variables in either the way the material was processed to the point where measurements were taken, the measurement procedure(s) utilized, and/or the environmental conditions under which the measurement(s) were taken. These variables must also always be associated with the individual values because they delimit the usefulness of the numbers. Examples of each type of variable include:

Processing history	manufacturing process (casting or extrusion)
	thermal treatment (precipitation-aged or stress-relieved)
Measurement procedures	rate of application of external force
	orientation of test sample
Environmental conditions	temperature
	pressure
	humidity

Each pertinent variable must be retained with the numeric property value, and a number of them are numeric data themselves, requiring consideration of units, orders of magnitude, and other pertinent delimiters.

Significance Level. Quantitative numbers in a database must also be understood in terms of their statistical significance; without a specific statement of significance they are of limited utility beyond "ballpark" estimates.

Some numeric databases contain individual test results, as they are taken from the measurement systems; such values are important in research and also in quality assurance work, where the distributions of information from replicate measurements is useful. Other databases will contain numbers reflecting some type of statistical analysis, usually minimum values based upon a specific criteria. Still others will contain values which have resulted from some expert evaluation, by an individual or panel of experts, or have been certified by some governing authority for use in a specific design application. Knowledge of this "basis" of the numeric values in a database is essential to its use, and so that information must be closely associated with the data in the database and throughout their use.

Graphical Impact. Graphic display of numeric property data plays an important part in much of their use and interpretation. Above and beyond displays of data where graphical display is inherent, such as crystallographic and spectral data, there are many other types of graphical displays used by scientists and engineers to better understand the relationships between two properties or the dependence upon a variable such as time, temperature, or pressure. Often the relationships are complex, as in the case of analyzing the structural stability of materials under stress where properties are studied in relation to complex parameters combining a number of variables.

The net result of these conditions is the need to provide the means for the graphical interpretation of many kinds of numeric property data, either as an integral part of the data handling system or as an associated capability.

COMPARISON: NUMERIC VS TEXT FILES

Based upon the previous observations, we may summarize the major differences between numeric and textural data as follows:

numeric/properties	text/bibliographic
data are tabulated as numbers	data are embedded in text strings
data are in standardized formats	data are unformatted
units are formatted, controlled	units are part of text string
data may be treated numerically	data can not be treated numerically

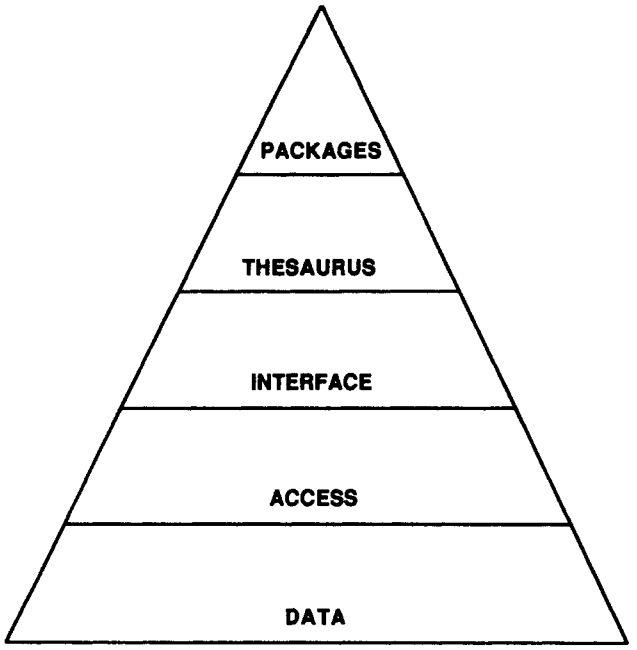

Figure 1. Numeric data searchers' hierarchy of motivation.

USER'S PERSPECTIVE ON NUMERIC DATA

Given the characteristics above, we may look at the users' perspective of numeric data and their expectations in accessing, searching, retrieving, and analyzing the data. There is literally a hierarchy of such usage, which may be considered in parallel to Maslow's hierarchy of human motivation.[7,8] The analogous levels of needs are the following:

Critical mass of data worth accessing; the greater the amount and breadth of data, the greater the driving force for access.

Reasonable mode of access; the easier access is made, the more likely users are to make the connection.

Capability for search and retrieval; the more sophisticated and yet easily understood the software is, the more likely it is to be utilized and significant numbers of data elements retrieved.

Ability to interpret meaning and significance; the greater the support features, such as helps, thesauri, and metadata, the more likely that large numbers of data records will be retrieved and analyzed.

Opportunity for analysis, interpretation, comparison, and exchange; maximum satisfaction and utilization of numeric data come with features for analytical treatment and graphical representation.

As Figure 1 illustrates, the parallel to Maslow's hierarchy is quite striking and provides a useful summary to the builders of sophisticated information systems as to how to anticipate users' needs and motivations.

ABILITY TO EXCHANGE DATA

There is one key aspect of the needs identified above which deserves particular attention: the need to exchange data. It is the key because it focuses the spotlight in turn on a series of related needs, without which all of the rest will be of limited value—standardization.

Since the need for and status of standardization in computerized data will be discussed in several papers which follow, I will address it here only in relation to what users expect of it for numeric/factual data as a class. In the simplest of terms, they expect to be able to send their data or receive it from others and compare data from several sources, while clearly understanding its content and its limitations. This is accomplished through the adoption by the industry as a whole of several types of standards:

Formats for describing materials/substances
Formats for recording specific test data
Interchange protocols
Terminology

As you will hear more about later, work is underway in all of these areas in ASTM Committee E49 on Computerization of Chemical and Materials Data.[9,10] Much remains to be done, but important starts have been made and your attention to and participation in that work is invited.

IMPLICATIONS IN REQUIREMENTS OF COMPUTERIZED SYSTEMS

Now that we have explored the nature of numeric data and noted both how it differs from textual data and what users of its expect, let us extrapolate those characteristics and describe some of the key elements required of any system intended to handled numeric data.[11-13] There are three components of this:

Specialized search software
Flexible user interface
Tabular/graphical presentation options

Specialized Search Software. Among the principal features required in handling most numeric data are the following:

Numeric range searching—looking for substances with properties in specific ranges or above or below specific values.

Tolerance handling—indicating whether a value of exactly 50 000 units is sought or of 50 000 ± some percentage of the value.

Units conversion—converting among any of five or six standard units systems for easy comparison or normalization.

Rounding—assuring that realistic indications of numbers of significant figures and properly rounded figures result from standardized calculations (such as units conversion).

Flexible User Interface. For many users, principally the nonprofessional or occasional searchers who are unfamiliar with command search procedures, as well as experienced professional searchers unfamiliar with the terminology of chemical substances or structural materials, menu-driven search and retrieval user interfaces are essential. Such interfaces permit novice and nonscientific searchers to interrogate sophisticated numeric files by either responding to straight-forward questions or selecting from preset lists of materials, properties, and variables. Among the attributes of sound interfaces are the following:

Easy to understand—user should never be in doubt as to next logical decision or choice; should never be "hung up" because system does not recognize certain responses.

Multiple search paths—user should readily recognize that there are several kinds of search paths open to them and be easily able to make the necessary decisions; never include paths or decisions that require previous knowledge or an extremely high level of technical depth in the field.

Handle multiple components—user should be able to ask about several materials at once or ask about materials having combinations of several properties; should be able to set several delimiters (variables) to limit search.

Permit easy query changes—must provide for options to change any one of several query components or add variables to narrow a search yielding too many answers; should not have to "start from scratch" again.

Tabular/Graphical Display Options. The third factor in meeting users' needs is that of logical tabular or graphical presentation of properties and characteristics. The obvious question here is "what is a logical presentation?", and we may provide some indications:

Data structure—if three properties are almost always used together and bear some logical relationship to each other in, they should be so presented. An example is the group of properties included in what are referred to as the "tensile properties" of a material; usually four specific properties are reported, in the following engineering order, with units:

tensile ultimate strength, psi
tensile yield strength, psi
elongation in gage length, percent
reduction in area, percent

Ideally these properties should always be presented in this tabular display, not in alphabetical order (the usual textual approach to listing data, making it look rather foreign to the user) nor one line under the other.

Supporting data—if there are delimiters or variables closely associated with the data whose omission results in serious loss of context or applicability, they should be present in column headings, leading descriptors, or clear footnotes.

Clutter—despite requirements for support data, perhaps including footnotes and complex headers, the tabular displays should be clear and uncluttered.

Abbreviations—use of abbreviations should be avoided, except perhaps when there are logical and widely understood industrywide standards, and even then supported with a lookup thesaurus which can be utiliized without disruption in search path or query building.

SUMMARY

In this paper, I have examined the characteristics and attributes of numeric/factual data from the perspective of the needs and expectations of the users of such data. Numeric/factual data differ significantly from textual data, principally because they are loaded and searched as discrete quantities not as text strings, and specialized software is therefore required.

Among the key features of numeric/factual property data are the dependence upon supporting factual information, including variables in the production of the materials or substances described, the methods of measurement utilized, and the conditions under which the measurements were made. Such information, much of it also quantitative in nature, must also be a part of the property data presentation, preferably in a way that will lead to continued association with the numbers.

Searches are often more specific and more direct with numeric/factual databases, and erroneous answers are more glaring and serious by their nature.

In later papers in this series, we will learn more of the peculiarities of quantitative property data and of the needs for standardization in how it is handled. And we will learn more about how specific data suppliers in several different media have dealt with the complexities described herein.

REFERENCES AND NOTES

(1) Computerized Materials Data Systems. In *Proceedings of the Fairfield Glade Conference*; Westbrook, J. H., Rumble, J. R., Eds.; National Bureau of Standards: Gaithersburg, MD, 1983.

(2) Ambler, E. Engineering Property Data-A National Priority. *ASTM Standard. News* **1985**, *Aug*, 46–50.

(3) *Materials Data Management—Approaches to a Critical National Need*; National Materials Advisory Board (NMAB) Report No. 405, National Research Council, National Academy Press: Washington, Sept 1983.

(4) *Material Property Data for Metals and Alloys: Status of Data Reporting, Collecting, Appraising, and Disseminating Numeric Data*; Advisory Board, National Academy Press: Washington, DC, 1980.

(5) Materials Data Systems for Engineering. *Proceedings of a CODATA Workshop*, Schluchsee, FRG, Sept 1985; Westbrook, J. H., et al., Eds.; Fachinformationszentrum (FIZ-Karlsruhe): Berlin, 1986.

(6) Materials Data for Engineering. *Proceedings of a CODATA Workshop*, Schluchsee, FRG, Sept 1985; Westbrook, J. H., et al., Eds.; FIZ-Karlsruhe: Berlin, 1986.

(7) Maslow, A. H. *Motivation and Personality*; Harper's Psychological Series; Harpers: New York, 1954.

(8) Kaufman, J. G. Increasing Data System Responsiveness to User Expectations. *Computerization and Networking of Materials Databases*; Kaufman, J. G., Glazman, J. S., Eds.; ASTM STP 1106, ASTM: Philadelphia, 1991; Vol. 2.

(9) Rumble, J. Standards for Materials Databases: ASTM Committee E49. *Computerization and Networking of Materials Databases*; Kaufman, J. G., Glazman, J. S., Eds.; ASTM STP 1106; ASTM; Philadelphia, 1991; Vol. 2.

(10) *Factual Materials Databanks—The Need for Standards*; Report of VAMAS Technical Working Area 10; Kroeckel, H., Reynard, K., Rumble, J., Eds.; Jun 1987.

(11) Rumble, J. R., Jr.; Smith, F. J. *Database Systems in Science and Technology*; Adam Hilger: Bristol, 1990.

(12) *Databases in Science and Technology—STN International*; American Chemical Society, Chemical Abstracts Service: Columbus, OH, 1989.

(13) Kaufman, J. G. The National Materials Property Data Network: A Cooperative National Approach to Reliable Performance Data. *Proceedings of the First International Symposium on Computerization and Networking of Materials Data Bases*; Glazman, J. S., Rumble, J. R., Eds.; ASTM STP 1017; ASTM: Philadelphia, Apr 1989; pp 7–22.