Problems in the Computerization of Chemical Information: Capture of Tabular and Graphical Data

J. H. Westbrook

Brookline Technologies, Ballston Spa, New York 12020

Received June 4, 1992

There are numerous problems in the computerization of technical information, even given a data structure appropriate to the subject content and a particular database management system. These problems arise in large part because usually the data it is desired to computerize were not initially collected by or for computer storage but rather existed in in various paper formats. Technical data appear mostly in the form of tables and graphs whose structure, typography, captions, and footnotes convey needed information, supplementing that imparted by the data values themselves. Thus, data capture for computer storage involves much more than obtaining a scanned image of the exhibit; all of the metadata (data about data), both explicit and implicit, must be thoroughly understood and incorporated in the computer record. If this is done successfully, nothing from the original source will be lost, the computer files will be fully searchable, and the user will be able to produce various presentation formats, independent of that of the original sources. Following initial capture, checks need be run for completeness of the data records (units, required variables, entity descriptors, characterization of the data values, etc.), for acceptability of the data values, and for control or harmonization of terminology. All these issues will be illustrated with examples from the field of chemistry.

INTRODUCTION

Today, with the ubiquitous presence of the computer in all phases of human life, it may seem superfluous to ask why we use it for storage of technical information. Yet consideration of the motivations provides useful background for examining some of the problems of getting this type of information into the computer in the first place. The attractions the computer offers in this context are

1. potential for storage of a large volume of data
2. possibility of random access to data
3. opportunity to facilitate the location and correction of errors
4. easy updating of information
5. potential for automated format conversion
6. direct use of data with application software
7. transparent unit conversion

These features have been elaborated in recent monographs and reports.[1-5]

Unfortunately, most data needed for such machine-readable databases are not generated automatically in digital form or structured in a way that permits easy retrieval. While some data currently emanating from our laboratories are generated by automated laboratory systems using digital sensors and while standardization of data structures for chemical and materials data has begun,[6-10] the vast majority of the information it is desired to computerize is now in *print* in journal articles, handbooks, data compilations, and encyclopedias, much of it in tabular or graphic formats. The print presentation, while familiar and with high potential for compression of information, also affords the considerable hazard of ambiguity of interpretation. In contrast, the computer representation of data demands a highly logical and completely specified data structure schema with all ambiguities and interpretational questions resolved prior to data entry. As we shall see, in tabular and graphical exhibits, information is imparted by logical structure, typography, captions, and footnotes to supplement that of the data values themselves. The information features contained in such fashion must also be incorporated in the computerized record of the print exhibits.

But, one might say, "Can't these tables and graphs be stored in the computer as images (video stills or raster scans) and hence pose no more interpretational problems than the original print forms?" The answer, of course, is "Yes, but...". As a stored image, the package of information is fixed, and all the interactive potential of the computer with this information is lost (format conversion, unit conversion, use with application software, etc.). Furthermore, even with the addition of a rich indexing scheme for such images, the possibility for full random access to *all* the information in the tabular and graphical exhibits is severely curtailed.

METADATA

Their Importance. Metadata—"data about data"—are the most important part of any database, for without them the data values it contains could not be found, interpreted, compared, or combined with other data values.[11,12] In contrast to data from other fields such as finance or demographics, technical data are complex and multifaceted, and their representation is poorly standardized. The volume of information constituting the metadata of a technical database has been estimated to be 10–100 times that of the numerical data values themselves. A representative listing of metadata parameters for the materials field is shown in Table I. An extensive discussion of metadata issues is found in ref 12. Here we will highlight only a few.

Terminology. Since data to be entered in a database typically come from diverse sources, the terminology employed, even in a narrow field, is often wildly variant. In addition, the database builder may prefer still other terms, and those familiar to an unknown coterie of future users are something else again.[13] Some simple examples will illustrate the point. Table II lists some of the commonly encountered designations for a common chemical substance, and Table III presents a similar listing for a property. The diversity in some instances is even more appalling; there are said to be more than 60

0095-2338/93/1633-0006$04.00/0

Table I. Principal Metadata Parameters for Materials Property Data

materials	characterization of data values
names	source
designations	class
equivalent designations	reliability
descriptors	derivation
properties (dependent variables)	statistical descriptors
names	data value descriptors
synonyms	descriptors of a population of data values
allowed values	data source
units	database
independent variables	exhibit id
names	data set id
synonyms	references
allowed values	footnotes
units	
units	
unit classes	
standard units	
valid units/conversions	

Table II. Synonyms for a Common Chemical

hydrogen chloride	HCl
hydrochloric acid	ClH
muriatic acid	CAS 7647-01-0

Table III. Synonyms for a Common Property

enthalpy of fusion	latent heat melting
latent heat of fusion	ΔH_f
enthalpy of melting	

names for the household drug aspirin, and dozens of equivalent designations for common industrial alloys.[14] The very use of the computer poses a terminological problem in that it is inflexible in responding fully and correctly to search commands unless it is primed with a memory store of a rich variety of synonyms and thesaurus relationships.

How can these terminological problems be solved? One action is to build, concurrently with the database, an on-line thesaurus and glossary invokable automatically during search and retrieval and supplemented by a user-callable "define" option.[15] Other beneficial steps are terminology standardization—a slow and difficult process[16,17]—and the imposition of terminological controls at the data entry stage of database building: required use of unambiguous variable names, acceptance of only allowed values for variables, inclusion of required independent variable values, appropriate units, specified default conditions, etc.

There are other types of terms beside entity names and property names, for example, material descriptors, test descriptors, variable terms, data descriptors, statistical descriptors, allowed value terms, unit terms, etc. (definitions for these term classes may be found in Appendix A). Definitions for all such terms used in the database, together with their thesaurus relationships, should be built-in in a database or system thesaurus. The diversity of synonyms (and proxies such as symbols and abbreviations) is as great for these term classes as for entity and property names. These cause terminological problems as do collective terms. Does ΔH_f mean heat of fusion or heat of formation? Is Poisson's ratio indicated by μ or ν? Only the context will tell. When the answer has been determined, an unambiguous designation must be entered into the computer. With respect to collectives, what is meant by "halogen gases", for example? The question is trivial for a chemist, but the computer is ignorant until taught that this term must be expanded to chlorine, fluorine, bromine, etc. Similarly with "natural fibers", "T6XX tempers", and so on. Unless the computer has been taught how to expand such terms and the search software designed to operate with the expanded listings, much pertinent data will be "lost" and not retrievable during search.

Characterization of a Material. Often the name of a chemical, a manufacturer's designation, or even a CAS Registry Number is not a sufficient identification for association with a particular property set in a database. The generic designation must be supplemented with other material descriptors such as grade (technical, drug, food, reagent, electronic), property (density, molecular weight, UV-resistance), or process (drawn, cast, spun, extruded).

Units. In building a database de novo, units for either properties or variables are not a particular problem. One simply adopts a certain system of units, e.g., SI units, as the standard for the database and ensures that all data entering the database are so expressed. In contrast, in building a database either from a single printed compilation or (more typically) from a variety of such sources, an enormous diversity of unit systems and forms are encountered. There are then two options: convert each unit when encountered to the standard chosen for the database or enter the units exactly as found and provide for a built-in unit conversion capability. The latter is usually the preferred course because it reduces the potential for conversion errors and allows for the fact that the ultimate user of the database may well prefer a different set of units than the database designer. Again there are problems with synonyms. For example, consider just one type of unit class, density (M/L^3). Not only are there many different related units in this class, e.g., g/mL, g/cm^3, kg/m^3, lbs/ft^3, and lbs/in^3, etc., but any one of these may be variously written—the last may also be found as pounds per cubic inch, lbs/cu·in., lbs·in^{-3}, and so on. The computer must be taught all these equivalences and conversion factors, or searches for quantitative values will be doomed to failure. Other problems arise where the author or editor has allowed the unit to stand as proxy for some frequently used property or variable. "ksi" may represent yield strength or pressure, and °C may represent test temperature or prior exposure temperature. Careful study of the context will be required to remove this kind of ambiguity.

PROCESS OF DATA CAPTURE AND CONVERSION TO CANONICAL RECORDS

Role of the Computer.[18] In addressing the problems arising in the conversion from print to computer format, the computer itself is used, insofar as possible, to identify the existence of a problem. For example:

1. Validation testing during data capture to ensure that term names and allowed values are compatible with the system vocabulary. Inconsistencies are identified, and expansion or substitution automatically made insofar as possible from system files.
2. Validation testing for completeness of the data record. Absences of required components can be detected.
3. Application of linking programs used to establish full matching or equivalency of dependent and independent variables and their values between potentially related exhibits. Incomplete links are identified.

The computer is also used to apply standards and other constraints, process exhibit records, perform linkages, implement denormalization routines, etc. to produce the desired set of final data records. Notwithstanding this machine assist,

the knowledge of a human expert in the subject field is frequently required to arrive at a database which is complete, accurate, and searchable. These matters will be more fully elaborated below.

Capture Process. In considering the capture of data from printed graphs or tables for computer entry, the first step is to recognize *all* the information elements, both explicit and implicit, which must be identified and incorporated into the computer record. Such information is critical both to ensuring the capability for decomposing the exhibit to individual data records (sometimes called canonical records) and to establishing the subsequent searchability of the database. Studies conducted by the author and his colleagues[19-21] have resulted in the construction of lists of these information elements for both tables and for graphs. Such listings can be converted, and have been, to a kind of electronic template or prompter to aid the data entry person in the capture process.

Tables. The set of elements important for tables is summarized in Table IV. Not all elements will be encountered in every instance, but consideration should be given in each case as to whether information on the element in question is necessary and where it may be found (often the required information is not within the exhibit itself but in the captions, footnotes, associated text, etc.). The different types of logic structure encountered in tables (and related lists) are summarized in Appendices B and C. Figure 1 presents examples of the three most common table logic structures.[22]

A guided table entry software program has been developed for personal computers (PCs) which is composed of a series of modules based on the classes of information element summarized in Table IV. Each module is callable by a dedicated function key, and a status control screen keeps track of the capture progression. As each module is called, the operator completes at the PC keyboard the information elements requested; these responses are automatically inserted in the electronic data record file. A generalized editor has been integrated into the program so that the same program can be used for initial capture of a table, for editing a previously captured table, or for using the file of a previously captured exhibit as a template to minimize rekeying of repeated column headers, row stubs, or field entries in similar tables. Figure 2 shows the data record resulting from this capture process for the first four rows of the row table shown in Figure 1b.

It might appear that this is a relatively cumbersome procedure for such a simple exhibit and that it results in an unduly lengthy data record. However, experience has shown that this same program can cope with even very complex tabular structures. Furthermore, when the data record is built in this fashion, several different options are now possible:

1. The table can be re-presented in exactly the format of the original exhibit (valuable for proofing purposes).
2. Every element of the table, singly or in combination with other elements, is now searchable.
3. With appropriate software programs, tabular information can be presented graphically.
4. New presentations can be created which never existed in print (e.g., column 3 of Figure 1b could be combined with a column from some other table containing similar properties for another class of resin, also in fine powder form).

Graphs. The capture of information from graphical exhibits is analogous to that described for tabular exhibits. The set of information elements important for graphs is summarized in Table V. (These elements are those appropriate for Cartesian plots; extension to polar, triangular, and other coordinate systems should be feasible but has not yet been attempted.) The software program developed for graphical capture is similar to that for tables except that when individual points or curves have to be recorded, resort is made to an electronic tablet and digitizer. Calibration must be initiated, of course, for each individual graph. Editing activities for graphs include recognition of the type of graph, distinction between dependent and independent variables and their related axes, point sets and line sets, and their symbols. As with tables, significant information elements may not be contained within the exhibit itself but are to be found in related titles, legends, keys, footnotes, and text.

The electronic data record for a graph is analogous to that shown for a table and exhibits the same sort of versatility in its application. Working from such a record, and with appropriate graphical display programs, the original exhibit can be faithfully re-presented or new exhibits with transformed axes, curves added or deleted, etc. are feasible. As with the tabular record, the electronic records for graphs captured in this way are fully searchable. Addition of mathematical representation of curves via curve-fitting routines would also make possible extrapolation or interpolation of data values.

Even such a complex graph as that of Figure 3 has been successfully captured using the program described here.[19] Note the several special features of this graph: rotated axes, auxiliary scales, three independent variables (alternating stress, mean stress, and state of notch), a family of stress ratio lines (A = alternating stress/mean stress and R = minimum stress/maximum stress), and inset notes and keys.

Completeness of Canonical Record. In describing the design of the data capture process, allusion has already been made to the concept of a "canonical data record", obtainable by decomposition of the captured exhibit records. Some further explication of this concept may be useful.[12] Canonical data records are self-contained records derived from printed exhibits or laboratory logs and designed as basic building blocks with which to construct or compile a database. For integrity of the database, each of the basic records should contain at least a minimum set of record elements as required by the database design. Inconsistencies in metadata definitions as well as failure to require completeness in the basic records can result in the effective disappearance of data records in the final database through the inability to locate and retrieve them.

For completeness the canonical record must include:

1. one material or entity; if the database record key is "material", then data records including multiple materials must be divided into separable basic records
2. at least one property or dependent variable with an "allowed value"
3. at least one independent variable with an allowed value (a default independent variable; – for example, test temperature or environment, – may be the only variable)
4. a full set of "required" independent variables with allowed values (where the definition of the property requires the accompaniment of one or more independent variables)
5. a "standard" unit or "default" unit for each property and independent variable, unless such unit and unit conversion features are treated at the system level

Table IV. Information Elements for a Table

Exhibit as a Whole
1. source (bibliographic citation)
2. ID (e.g., Table 1.015)
3. no. of tables (some exhibits consist of more than one with common headers)
4. table logic group (e.g., column table, row table,combined table, etc.)
5. title
6. material or entity

Items Relating to Table Structure
1. no. of columns
2. no. of rows
3. column/row variable type (independent, dependent, descriptor)
4. column and row labels

Items Relating to Columns/Rows
1. identification
2. variables (dependent, independent)
3. units
4. labels (column headers/row stubs)
5. format (column separators and entry placement; l, r, and c)
6. data values
7. internal relationships (internal hierarchies)

Notes/Footnotes/References
1. identification
2. text/citation
3. location

Symbols
1. identification

Access Algorithms
1. procedure

Associated Equation(s)
1. equation
2. substitution procedure(s)

Densities of Aqueous Inorganic Solutions

Table 88. Potassium Chromate
(K_2CrO_4)

%	d_4^{18}	d_4^{18}
1	1.0073	1.0066
2	1.0155	1.0147
4	1.0321	1.0311
8	1.0659	1.0647
12	1.1009	1.0999
16	.....	1.1366
20	.....	1.1748
24	.....	1.2147
28	.....	1.2566
30	.....	1.2784

a)

Table 1. Typical Mechanical Properties of PTFE Resins

Property	Granular resin	Fine powder
tensile strength at 23°C, MPa[a]	7-28	17.5-24.5
elongation at 23°C, %	100-200	300-600
flexural strength at 23°C, MPa[a]	does not break	
flexural modulus at 23°C, MPa[a]	350-630	
impact strength, J/m[b], at 21°C	106.7	
hardness durometer, D	50	
compressive stress, MPa[a] at 1% defor...		

b)

Table 159. Compressibility Factors of Methane[a]
PV = 1.0000 at 1 atm. and 0°C.

Pressure, atm.	−70°C.	−50°C.	−25°C.	0.0°C.	25°C.	50°C.	100°C.	150°C.	200°C.
1	0.7410	0.8150	0.9075	1.0000	1.0922	1.1845	1.3686	1.5525	1.7363
10	0.6985	0.7795	0.8803	0.9785	1.0733	1.1780	1.3595	1.5470	1.7348
20	0.6473	0.7402	0.8493	0.9543	1.0549	1.1590	1.3500	1.5422	1.7330
30	0.5910	0.6991	0.8183	0.9297	1.0373	1.1412	1.3411	1.5370	
40	0.5244	0.6547	0.7873	0.9061	1.0198	1.1275	1.3335	1.534	
50	0.4425	0.6069	0.7558	0.8830	1.0034	1.1152	1.3268		
60	0.3466	0.5551	0.7243	0.8607	0.9871	1.1017	1.320		
80	0.25[illegible]6	0.4604	0.6651	0.8192	0.9569	1.0799			
100	0.2808	0.4088	0.6167	0.7845	0.9319				
120	0.3175	0.4095	0.5877	0.7604					
140	0.3543	0.4304	0.5801	0					
160	0.3915	0.4601	0.58						
180	0.4288								
200									

c)

Figure 1. Examples of three table logic structures: (a) column table, (b) row table, and (c) combined table.

```
EXHIBIT_KEY = K-O, CE!t!1!513!1985;
SOURCE_FORMAT = table;
ERRORS_FOUND = none;
AUDIT_TRAIL = Thur 2 Apr 92, JHW;
MODIFIER =;
MOD.DESCR. =;

. . . . . .

PAGE_NUMBER = 513;
TITLE = Table 1. Typical Mechanical Properties of PTFE Resins;
TABLE_ID = 1;
   N-ROWS = 19;
   N_COLS = 3;
   GROUPING = table-row;
   WHOLE_TABLE_INFO;
      TABLE_MATERIAL = PTFE Resin;
      DATA_SOURCE = unstated;
      DATA_COMPILER = S.V. Gangal, EI DuPont;
      DATA_CLASS = typical values;
COL_INFO;
   COLUMN = 2;
      CINDVAR = form;
         VALUE = Granular resin;
   COLUMN = 3;
      CINDVAR = form;
         VALUE = Fine powder;
ROW_INFO;
   ROW = 0;
      RULE = under;
   ROW = 1;
      FORMAT = 1!c!c;
      ENTRY = Property!Granular resin!Fine powder;
      RULE = under;
   ROW = 2;
      FORMAT = 1!1!1;
      ENTRY = tensile strength at 23°C,MPa{note a}!7-28!17.5-24.5;
   ROW = 3;
      FORMAT = 1!1!1;
      ENTRY  = elongation at 23°C,%!100-200!300-600;
   ROW = 4;
      FORMAT = 1!1!1;
      ENTRY = flexural strength at 23°C,MPa{note a}!does not break;

. . . . . .

FOOTNOTE_INFO;
   NOTE = a;
      ENTRY = To convert MPa to psi, multiply by 145.;
   NOTE = b;
      ENTRY = To convert J/m to (lbf.ft)/in., divide by 53.38.;
```

Figure 2. Excerpt from the table record of the row table of Figure 1b (dotted lines indicate missing record segments).

Table V. Information Elements for a Graph

Exhibit as a Whole
1. source (bibliographic citation)
2. ID (e.g., Figure 2.013)
3. no. of graphs (some exhibits consist of more than one graph with a common axis)
4. no. of quadrants (i.e., do x and y axes show negative as well as positive values)
5. legend
6. inset caption (if any)

Items Relating to Axes
1. scale type (linear, logarithmic, reciprocal, etc.)
2. location (left, right, bottom, top, rotated, etc.)
3. scale labels
4. units
5. axis length
6. axis length unit
7. max and min values
8. axis breaks (number, location)
9. grid/tick lines (number, location)
10. grid/tick values
11. grid labels

Items Relating to Curves
1. no. of curve sets
2. ID of curve set
3. no. of curves in set with ID = xx
4. ID of curve
5. analytical form (linear, curvilinear, parabolic, hyperbolic, etc.)
6. line type (solid, dashed, dotted, etc.)
7. breaks (number, location)
8. curve labels
9. digitization (data tuples)

Items Relating to Points
1. no. of point sets
2. ID of point set
3. no. of points in set with ID = xx
4. point symbol (circle, square, triangle, cross, etc.)
5. point labels
6. digitization (data tuples)

Notes/References/Symbols
1. identification
2. text/citation
3. location

Access Algorithm
1. access procedure
2. interpolation/extrapolation routines

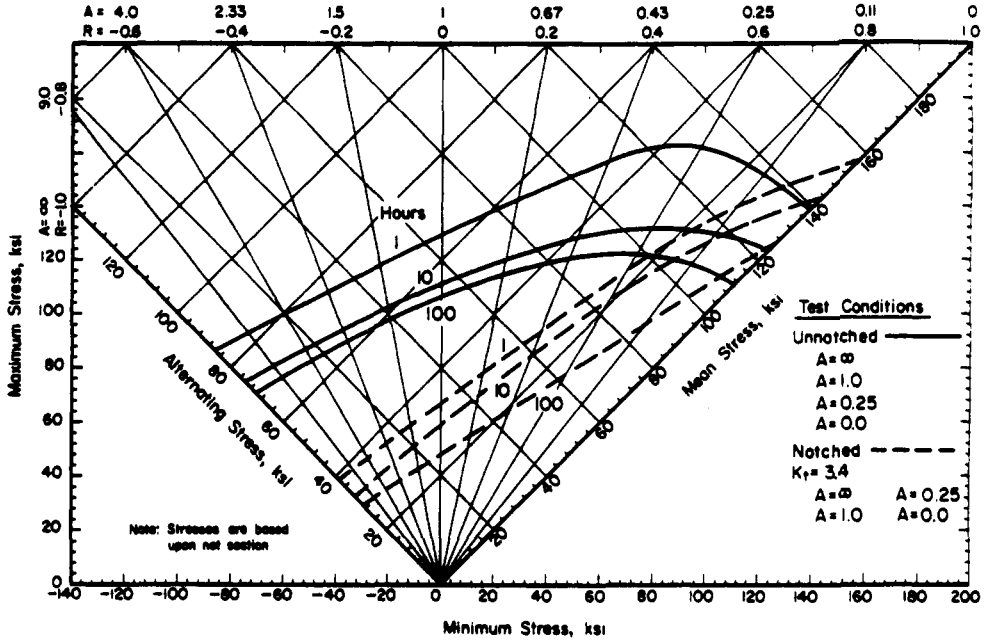


FIGURE 6.3.11.1.8(b). *Typical constant-life diagram for fatigue behavior of solution-treated and aged Udimet 500 alloy bar at 1650 F.*

Figure 3. Example of a complex graph that has been successfully captured using the techniques described.

6. a single value for each independent variable where required (Certain databases may be designed such as to disallow the occurrence of multiple values of the same independent variable in any one basic record. Where this occurs in the original exhibit, then separate basic data records are generated for each combination of values of the independent variables.)
7. where required, an indicator of the data class and quality, including values such as typical, specification min/max, statistically processed, design values, test values, etc.
8. a keyed numbering scheme for the basic record such that associated basic data records derived from the same exhibit can be retrieved together or the original exhibit reconstituted or both; e.g., for a row table, the different columns would be numbered in sequence

Thus, either as part of the validation testing associated with data entry or as part of the database compilation procedure, a completeness test of the scope just defined is desirable to ensure the integrity of the database.

Other Validation Procedures. Many existing databases are imperfect in that when searched they return incorrect values, ambiguous values, or no values at all—even though the needed information is indeed contained within them. Many of these faults can be attributed to a too simplistic digitization of the original data without the intervention of automatic validation processes, which in turn depend on the availability during the capture process of a series of stored files of acceptable terms, units, synonyms, classes, etc. Table VI lists a variety of such validation checks with examples.

Overview. A schematic of the full set of procedures required to capture data from a print source and produce a searchable database is shown in Figure 4.[18] The conversion of original document exhibits, either tables or graphs, to electronic exhibit records has already been described. These exhibit records are then decomposed by a software program to a set of canonical data records as defined by the database designer.

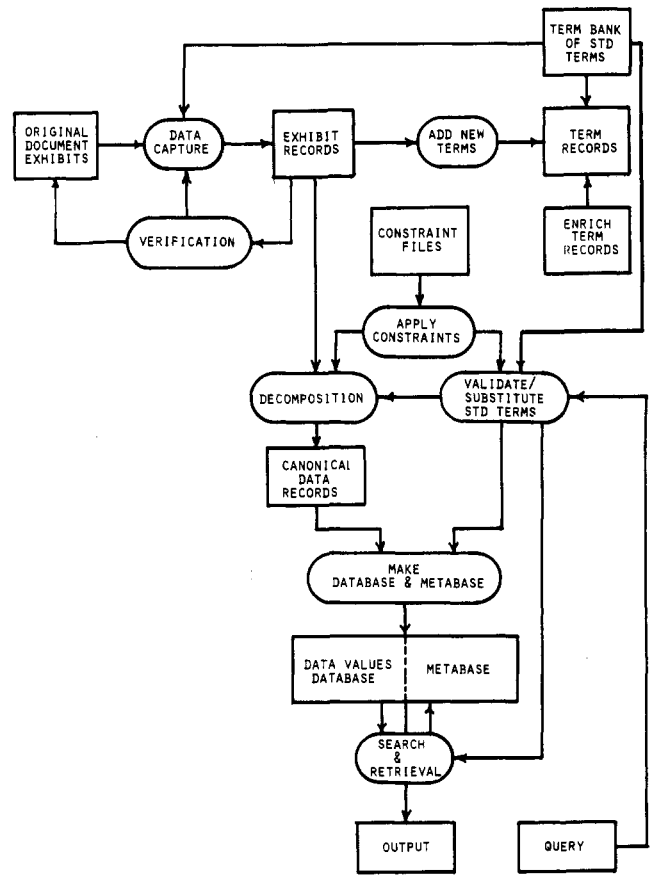


Figure 4. Schematic of the procedures required to capture data values from a print document and produce a searchable computerized database.

Typically, this results in an explosion in the *number* of individual records by 10:1 or more.

During the decomposition process, a number of auxiliary validation procedures can also be carried out which ensure that the values and combinations of variables/values in the final canonical records are consistent with local, industry, national, or international standards. Similar procedures are used for checking on the use of standard term names and for

TABLE 2.2.1.0(b). *Design Mechanical and Physical Properties of AISI 1025 Carbon Steel*

Specification	MIL-S-7952, 1025	MIL-T-5066	MIL-S-7097, Comp. 3
Form	Sheet and strip	Tubing	Bars
Condition	Cold rolled	Normalized	All ← b)
Thickness, in.			
Basis ← e)	S	S	S[a]
Mechanical properties: ← c) ← d)			
a) → F_{tu}, ksi:			
L	55	55	55
LT	55	55	55
ST			55
F_{ty}, ksi:			
L	36	36	36
f) → LT	36	36	36
ST			36
F_{cy}, ksi:			
L	36	36	36
LT	36	36	36
ST			36
F_{su}, ksi	35	35	35
F_{bru}, ksi:			
(e/D = 1.5)			
(e/D = 2.0)	90	90	90
F_{bry}, ksi:			
f) → (e/D = 1.5)			
(e/D = 2.0)			
e, percent:			
L		b	ab
LT	b		
E, 10^3 ksi		29.0	
E_c, 10^3 ksi		29.0	
G, 10^3 ksi		11.0	
μ		0.32	
Physical properties:			
a) → ω, lb/in.^{3}		0.284 ← d)	
C, Btu/(lb)(F)		0.116 (122 to 212 F)	
K, Btu/[(hr)(ft^2)(F)/ft]		30.0 (at 32 F)	
α, 10^{-6} in./in./F		See Figure 2.2.1.0	

g) → [a] Grain direction not specified.
[b] See applicable specification for variation in minimum elongation with ultimate strength.

Figure 5. Tabular exhibit showing numerous problems with implicit metadata (see text for explanation of annotations).

Table VI. Variety of Validation Checks (with Examples) for Testing Data Records

validation check	example
completeness and correctness of sequence for all metadata elements	"solubility" property requires specification of the identity of the solvent, the temperature, and perhaps the pressure for a number to be admitted as a possible value of that property
existence of the local dependent or independent variable (or its synonym or abbrevation)	variable entered from the printed source as "satunation temperature" would not be admitted because it represents a nonsensical term name which would not be in the listed known variables
existence of a valid unit drawn from the appropriate unit class for each property and quantitative variable	specific heat values in J/mol °K, cal./gm.°C, BTU/lb. °F, and many other units would be acceptable but not if given in watts
correctness of data format (character string, decimal, fraction, etc.) for various values	for a variable, "solution concentration", values of 0.1N, 3N, etc. would be accepted but not 3.2×10^7
values within allowed ranges (theoretical or arbitrarily set) for numerical quantities	melting point values in excess of 4000 °C or 7200 °F would be rejected as being out of range for that parameter
values on prescribed lists of allowed values for qualitative parameters	for a variable, "catalyst type", terms such as "metal", "oxide", "sulfide" would be accepted but "powder" would be rejected

the completeness of the records. Whenever possible, the validating files contain the terminology and data published by recognized standard bodies, such as ASTM and ANSI, or by trade associations, e.g., Chemical Manufacturers Association.

The final stage is the compilation of the canonical data records into a database and the simultaneous creation of a series of index files, the metabase, which provide access points for materials, properties, variables, and terms and which serve for identifying and categorizing links occurring between records and organizing the footnotes. Note that when accessed by the user, the system can convert the terminology, units, etc. of his query to those of the database, and that search and retrieval operates primarily on the metabase and only secondarily on values in the database itself.

OTHER PROBLEMS ENCOUNTERED

Tables. A recent paper[18] describes in some detail the problems met in computerizing one particular engineering materials handbook.[25] A single example from that source will provide an illustration. Figure 5 reproduces a tabular exhibit from *MilHandbook 5*, with appended notes highlighting the problems of implicit metadata. Among these problems we may cite the following: (a) the property names are given in a symbolic shorthand peculiar to this particular reference; (b) the collective term "all" requires definition; (c) the mechanical property values must by default be understood

Table 1. Average Volatility Data for U.S. Motor Gasolines

Property	Unleaded		Regular		Premium	
	Winter	Summer	Winter	Summer	Winter	Summer
initial boiling point, °C	28	32	28	32	28	32
10% point, °C	42	49	42	48	42	48
50% point, °C	100	104	94	99	99	102
90% point, °C	165	167	168	172	164	166
end point, °C	207	209	210	213	208	209
Reid vapor pressure, kPa[a]	86	67	85	67	85	68

[a] To convert kPa to psi, multiply by 0.145.

Figure 6. Example of a nested table showing certain complications.

to be room temperature data; (d) the mechanical data values themselves should be characterized as values recommended for "design" as opposed to test data, specification values, etc., whereas the physical property data should be characterized as "typical"; (e) the definition for "S basis" is not specified within the exhibit and must be found elsewhere in the reference; (f) the independent variable affecting certain properties needs be identified; and (g) the footnote applies only to certain values shown. In other cases of printed exhibits, structural features such as ruled lines, indentations, and horizontal or vertical spans convey other implicit metadata information, as may use of special type fonts or symbols, all of which may be lost when the original exhibit is exploded into a number of individual canonical records in the computer.

The problems described above, as well as others, can be illustrated with examples more familiar to a chemical audience. One such appeared in Figure 1a in that the symbolism of the column header is nowhere explained in the original source. It appears that the solution densities are expressed in terms of the density at 15 °C (or 18 °C) relative to water at 4 °C (point of maximum density). Figure 6 presents an interesting case of a nested table with different measures of the same conceptual property. The column headers to the right show two different independent variables, grade (unleaded, regular, premium) and season (winter, summer), while the first five rows of the body of the table present volatility data for different levels of volatilization (initial, 10%, 50%, 90%, and end point), all expressed in °C. Such representations have been referred to as "flavors" of the given property.[12] Row 6, however, presents a completely different measure of volatility, expressed in different units.

Figure 7 is an example of a column table but with some complications. Columns 3–5 represent three different physical properties, while a fourth property is shown in the three columns at the far right at three different levels of the independent variable, "test temperature". The left-hand columns 1 and 2 are actually row stubs showing different but equivalent representations of the same independent variable, "concentration". All of this must be carefully sorted out and understood at the time of data entry, or the complete record will be unusable.

Figure 8 is an example of a row table, but again with some nuances that require special attention. Row 1, despite the label of the column header, is not really a property in the sense of a dependent variable but rather a material descriptor which determines the properties to be expected within the subgroup of a given polymer family, e.g., columns 3 and 4 for epoxies. While the column headers for columns 2–6 clearly define these polymer families, columns 7 and 8 define the material in terms of form and process, with no polymer family deducible from the table; hints as to its identity must be sought elsewhere.

Graphs. Few graphs from handbooks and other reference compilations are simple cartesian plots with (0,0) origin and a single curve and/or point set. A few examples will now be presented to illustrate some of the typical complexities encountered.

Figure 9 is the analogue of a combined table in that in a single exhibit it shows the effect of two independent variables (surface temperatures of surfaces 1 and 2) on one dependent variable (radiation coefficient of heat transfer). However, it must be noted that the ordinate scale is logarithmic; that a value of a third independent variable, emissivity, appears in an inset note; and that the analytical expression for the curves is shown in the same inset. The inset further calls attention to the fact that temperature in the analytical expression is in degrees F, absolute (i.e., degrees Rankine), whereas temperature units plotted in abcissa and curve labels are in °F.

Figure 10 seems simple enough on the surface in its presentation of two curves and two point sets on a single plot. Yet study of the caption reveals that the curves are theoretical constructs rather than curves drawn through the data points, and the same symbol has been used for point sets for two different reactants.

In Figure 11, data for four curves and point sets are displayed where the abcissa scale for one curve has been displaced to prevent confusion between curves. The individual curves are labeled with codes whose explication is to be found in the figure caption.

Figure 12 is a log–log plot with multiple curves of viscosity against vapor pressure of water for a number of different liquids. The true independent variable, however, is temperature, shown above on a nonlinear scale, for which the vapor pressure of water, on the lower horizontal axis, is only a proxy.

In Figure 13, a four-quadrant plot, curves are presented for a single property of two materials under the action of two additional independent variables, values for one of which, imposed potential, appear as curve labels; the other, irradiation

Table 2. Physical Properties of Aqueous Solutions of Phosphoric Acid

Concentration, wt %		Density, 25°C,			Viscosity, mPa·s (= cP) at		
H_3PO_4	P_2O_5	g/cm³	Bp, °C	Fp, °C	20°C	60°C	100°C
0	0	0.997	100.0	0	1.0	0.48	0.30
5	3.62	1.025	100.1	−0.8	1.1	0.54	0.33
10	7.24	1.053	100.2	−2.1	1.2	0.61	0.38
20	14.49	1.113	100.8	−6.0	1.6	0.78	0.48
30	21.73	1.182	101.8	−11.8	2.2	1.0	0.62
50	36.22	1.333	108	−44.0	4.3	1.8	1.1
75	54.32	1.573	135	−17.5	15	4.8	2.4
85	61.57	1.685	158	21.1	28	8.1	3.8
100	72.43	1.864	261	42.35	140	25	9.2
105	76.10	1.925	>300	16.0	600	70	19
115	83.29	2.044	>500			1500	250

Figure 7. Example of a column table (partially nested) showing certain complications.

Table 2. Physical Properties of Commercial Rigid Foamed Plastics

Property	Cellulose acetate	Epoxy		Phenolic		Extruded plank	
density, kg/m³[a]	96–128	32–48	80–128	32–64	112–160	35	53
mechanical properties							
compressive strength, kPa[b] at 10%	862	138–172	414–620	138–620		310	862
tensile strength, kPa[b]	1,172		345–1,240	138–379		517	
compression modulus, MPa[c]	38–90	3.9	14.5–44.8			10.3	
thermal properties							
thermal conductivity, W/(m·K)[c]	0.045–0.046	0.016–0.022	0.035–0.040	0.029–0.032	0.035–0.040	0.030	
max service temperature, °C	177	205–260	205–260	132	205	74	
specific heat, kJ/(kg·K)[d]						1.1	
electrical properties							
dielectric constant,	1.12			1.19–1.20	1.19–1.20	< 1.05	< 1.05
moisture resistance							
water absorption, vol%	4.5			13–51	10–15	0.02	0.05

[a] To convert kg/m³ to lb/ft³, multiply by 0.0624.
[b] To convert kPa to psi, divide by 6.895.
[c] To convert MPa to psi, multiply by 145.
[d] To convert kJ/(kg·K) to Btu/(lb·°F), divide by 4.184.

Figure 8. Example of a row table showing some interpretational problems.

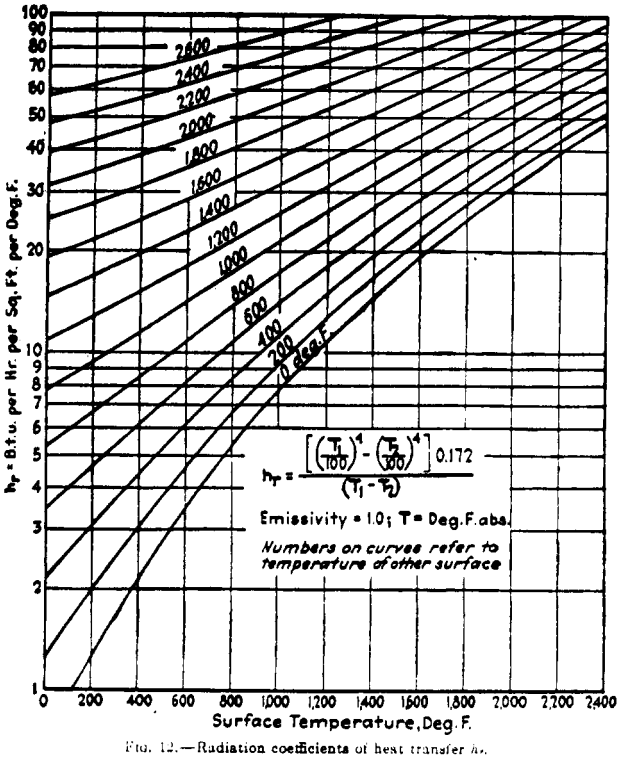


FIG. 12.—Radiation coefficients of heat transfer h_r.

Figure 9. Example of a graphical analogue of a combined table with three independent variables and other complications.

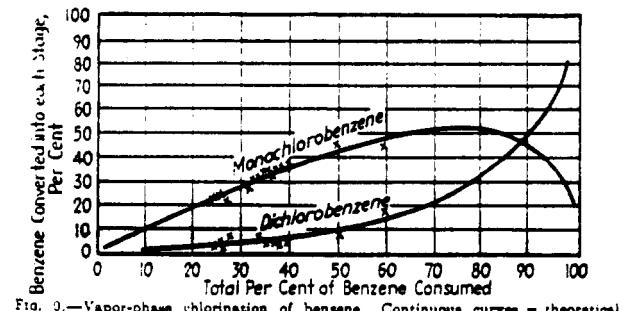


FIG. 9.—Vapor-phase chlorination of benzene. Continuous curves = theoretical curves for two-stage reaction. Points marked X = experimental values. (*From Groggins, "Unit Process in Organic Synthesis."*)

Figure 10. Example of a graph with multiple curves and multiple point sets.

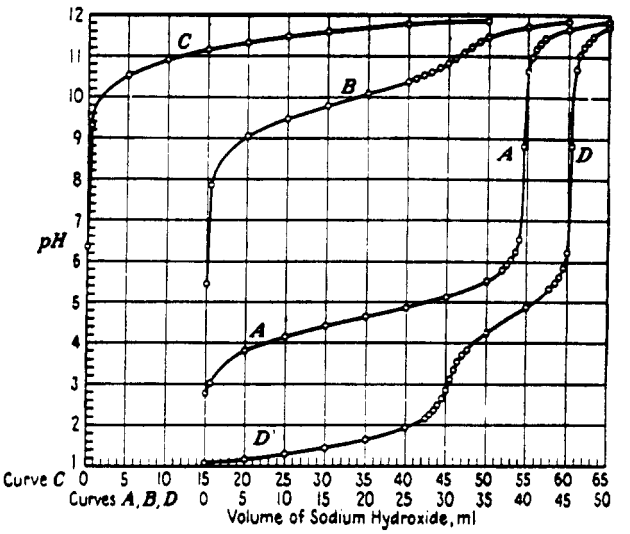


FIG. 2. Titration curves of some monoprotic acids with 0.2046*N* NaOH. Initial volume = 60 ml in all cases. (A) 8.06 millimoles of acetic acid; (b) 6.75 millimoles of phenol; (C) 8.0 millimoles of hydrogen peroxide; (D) 6.26 millimoles of hydrochloric acid plus 3.02 millimoles of acetic acid.

Figure 11. Example of a graph with displaced origin for certain curves and coded labels.

strength, appears in an inset. Further, the caption is in error because, not knowing the voltage gradient, what is plotted is a measure of relative conductance, not conductivity.

In our final example in Figure 14, data for the corrosion rate of steel in aerated water are shown in normalized fashion. Instructions for converting normalized rates to absolute rates appear in the caption. Other features of note include a reverse pH scale, two different but related ordinate scales, and separate curves showing the effects of an additional independent variable, temperature.

All of the features shown in the examples just reviewed not only can be accommodated within the graphics capture program described previously, but, as was also true for tables, the electronic prompts and validation files alert the data entry person to inconsistencies, errors, and omissions. This small

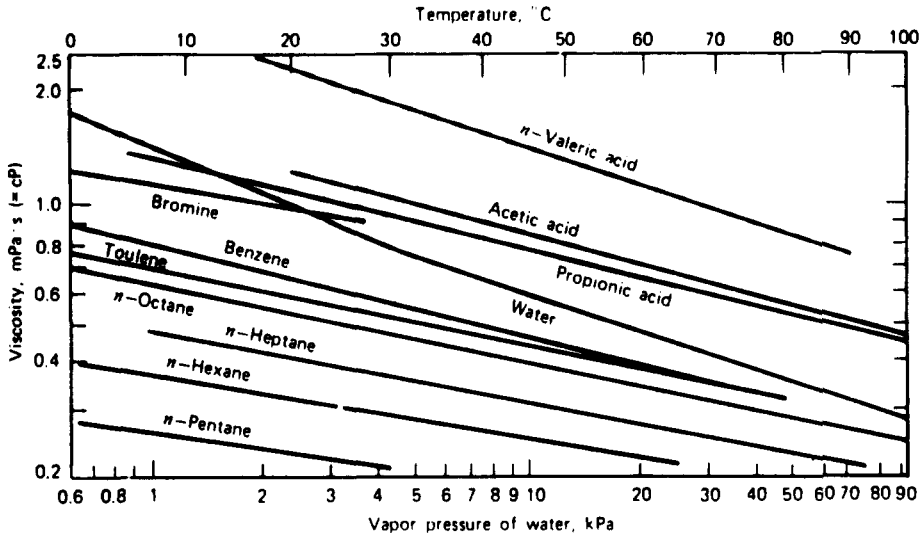


Figure 2. Plot of viscosities of eleven representative liquids against temperatures which are obtained from the corresponding vapor pressures of water. To convert kPa to mm Hg, multiply by 7.5. Courtesy of *Industrial and Engineering Chemistry*.

Figure 12. Example of a graph with a proxy independent variable.

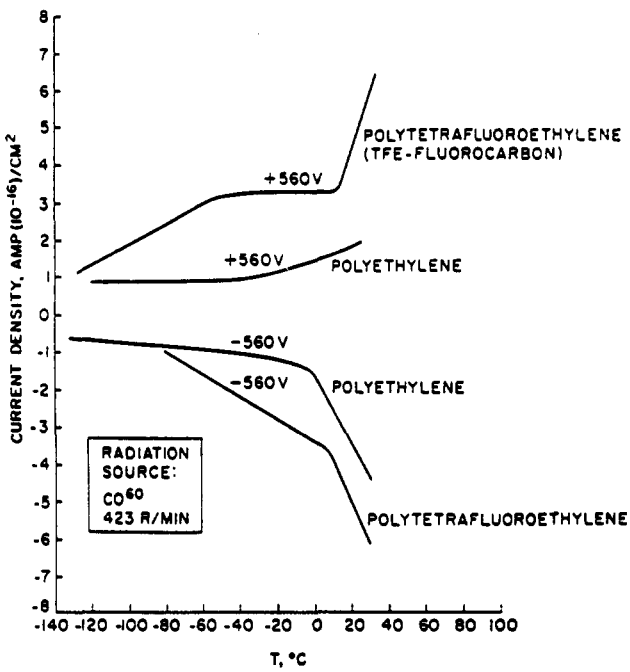
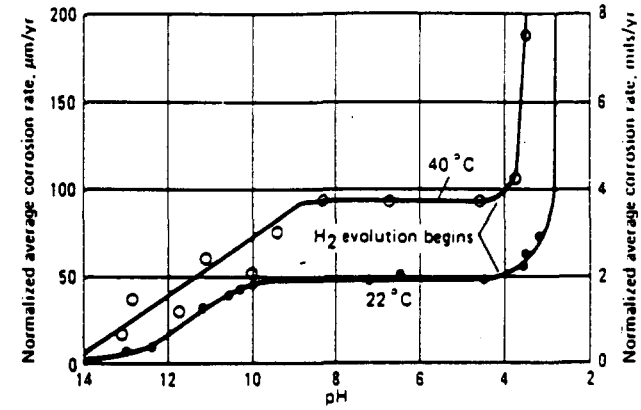


Fig. 10.3 Conductivity of irradiated polymers at various temperatures and positive and negative voltage[40]

Figure 13. Example of a 4-quadrant graph.



Corrosion rates are normalized to a solution containing 1 mL O_2 per litre of water. To estimate corrosion rates at other concentrations, multiply values derived from this graph by the oxygen concentration in mL/L.

Fig. 2. Effect of pH on corrosion of steel in aerated water

Figure 14. Example of a graph with a normalized scale and various other complicating features.

sampling demonstrates, however, the diversity and complexity of the problems faced by those attempting to convert printed graphs to searchable electronic records.

CONCLUSION

This paper has reviewed the numerous problems in the computerization of technical information, particularly that embodied in printed tables and graphs. A recently developed package of software capture programs has been shown capable of surmounting the majority of these problems, particularly when the computer itself is used to perform various completeness and validity checks. Throughout, the availability of appropriate data structures and database management systems has been assumed; this may not always be the case and such may well be the subject of further research. Additional progress in computerization of technical information will require advances in the standardization of these aspects as well as in the standardization and harmonization of terminology.

APPENDIX A: SOME TERM TYPES[29]

material term

a type of term used to identify a material or class of materials, e.g., energetic material, chemical family, CAS Registry No.

material descriptor

a term or data element that amplifies the primary identifiers of the material or sample by incorporating such characteristics as specification designation, other identifiers, compositional characterization, source, processing history, sample description, or service history, e.g., chemical structure, synthesis process, color

property term

a type of term used to identify a property of a material or entity, e.g., property class, reference property, derived property

test descriptor

a test parameter, recorded as part of the testing procedure for a designated test method, useful for validating the testing protocols or for further characterizing the testing process, e.g., test validity indicator, test environment, test identification

variable term

a type of term used to identify an independent variable affecting the properties of a material or entity, e.g., compound variable, required variable, form

data descriptor

one member of any of several sets of descriptors in different categories (e.g., source format, data class, data reliability, data treatment status) which may be used to characterize individual data values or data sets, e.g., allowable, data quality indicator, data source identifier

statistical descriptor

a term used to characterize the mathematical significance of a population of data values or one numerical value relative to a population of similar values; examples include maximum, minimum, mean, standard deviation, coefficient of variation, etc.

allowed value term

an alphanumeric string or code which is a member of the possible data values for an entity, property, variable, descriptor, or parameter; examples include raw data, certified data, or derived data as some of the permitted values of the data descriptor, "data class"

unit term

a type of term used to define or describe a quantitative standard of measurement, e.g., standard unit, unit class, valid unit

APPENDIX B: SUMMARY OF TABLE TYPES, LOGIC AND FORMAT STRUCTURES

Logic Structures

row table

a tabular array of rows and columns in which the values of each *dependent* variable are located as entries in the *same row*, with each column defining the value of one or more associated independent variables, materials, or entities or combinations thereof

column table

a tabular array of rows and columns in which the values of each of the *dependent* variables are entries in the *same column*, with each row stub defining the value of one or more associated independent variables, materials, or entities, or combinations thereof

combined table

a tabular array of rows and columns in which each data cell contains the value of one dependent variable or entity, as determined by the values of the associated independent variables, entities, or materials shown in the corresponding row stub and column header

Special Format Tables

stacked table

a table in which independent variables and their values appear in one row or column, or a set of rows and columns, but are associated with dependent variables appearing in several groups of rows or columns, adjacent to one another, either side by side or above and below

folded table

a table arbitrarily divided and rearranged side by side such that some row stubs now occupy an intermediate column, and columns to its right repeat columns to its left

nested table

a table of complex structure which incorporates superordinate column headers or row stubs or both at two or more hierarchical levels

Other Table Types

math table

in addition to a logic structure, an access algorithm must be given to direct the way the table values are to be accessed

logic table

each cell value is an operator linking the row and column entities or variables, e.g., **truth table**; a matrix that describes a logic function by listing all possible combinations of inputs, and indicates the outputs for each combination

APPENDIX C: SUMMARY OF LIST TYPES[31]

row list

a tabular format in which the values of the entry are arranged in rows; the logical structure is one-dimensional, and the list should be read only along the row

column list

a tabular format in which the values of the entry are arranged in columns; the logical structure is one-dimensional, and the list should be read only up and down the column

stacked list

a compound exhibit consisting of a group of lists arranged vertically one above the other

folded list

a list arbitrarily divided and rearranged side by side; despite the multicolumn array, such a list is still to be read in only one dimension

REFERENCES AND NOTES

(1) Date, C. J. *An Introduction to Database Systems*, 4th ed.; Addison-Wesley Publishing Co.: Reading, MA, 1987.

(2) Rumble, J. R., Jr.; Hampel, V. E. *Database Management in Science and Technology*; North Holland: Amsterdam, 1984.

(3) Shoshani, A.; Olken, F.; Wong, H. Data Management Perspective of Scientific Data. In *Role of Data in Scientific Progress*; Glaeser, P. S., ed.; North Holland: Amsterdam, 1985.

(4) Kroeckel, H.; Westbrook, J. H. Computerized Materials Information Systems. *Philos. Trans. R. Soc. London, A* **1987**, *322*, 373–391.

(5) Rumble, J. R., Jr.; Smith, F. J. *Database Systems in Science and Engineering*; Adam Hilger: Bristol, U.K., 1990.

(6) Kroeckel, H.; Reynard, K. W.; Rumble, J. R., Eds. Factual Materials Databanks, The Need for Standards. Versailles Project on Advanced Materials and Standards (VAMAS), 1987.

(7) Kaufman, J. G. Standards for Computerized Material Property Data—ASTM Committee E49. In *Computerization and Networking of Material Property Data Bases*; Glazman, J. S., Rumble, J. R., Jr., Eds., ASTM STP 1017; ASTM: Philadelphia, 1989; pp 7–22.

(8) Reynard, K. W. Standards for the Presentation and Use of Materials Data: A Review of the Activities of ASTM, CEC, CODATA, and VAMAS with Proposals for the Future. In *Scientific and Technical Data in a New Era*; Glaeser, P., Ed.; Hemisphere Publishing Corp.: New York, 1990, pp 56–60.

(9) Rumble, J. R., Jr. Standards Produced by ASTM E49. *ASTM Stand. News* **1992**, in press.

(10) Lysakowski, R. The Global Standards Architecture for Analytical Data Interchange and Storage. *ASTM Stand. News* **1992**, *20*, 44–51.

(11) Westbrook, J. H. Standards and Metadata Requirements for Computerization of Selected Mechanical Properties of Metallic Materials. *NBS Spec. Publ.* 1985; *No. 702*.

(12) Westbrook, J. H.; Grattidge, W. The Role of Metadata in the Design and Operation of a Materials Database. In *Computerization and Networking of Materials Databases II*; Kaufman, J. G., Glazman, J. S., Eds.; ASTM STP 1106; ASTM: Philadelphia, 1990.

(13) Westbrook, J. H. Terminological Problems in Computerization. *ASTM Stand. News* **1990**, *18*, 18.

(14) Westbrook, J. H. Designation, Identification, and Characterization of Metals and Alloys. In *Computerization and Networking of Materials Databases*; Glazman, J. S., Rumble, J. R., Jr., Eds.; ASTM STP 1017; ASTM: Philadelphia, 1989; pp 151–174.

(15) McCarthy, J. L. The Automated Data Thesaurus: A New Tool for Scientific Information. In *Scientific and Technical Data in a New Era*; Glaeser, P. S., Ed.; Hemisphere Publishing Corp.: New York, 1990; pp 260–264.
(16) Grattidge, W.; Westbrook, J. H.; Lund, W. B. Developing a Term Database for Use with Materials Property Data. In *Standardizing Terminology for Better Communication Practice, Applied Theory and Results*; ASTM Symposium, Cleveland, OH, June 1991; ASTM: Philadelphia, 1991.
(17) Westbrook, J. H.; Grattidge, W. Terminology Standards for Materials Databases. In *Computerization and Networking of Materials Databases III*; Barry, T. I., Reynard, K. W., Eds.; ASTM: Philadelphia, 1992.
(18) Grattidge, W.; Lund, W. B.; Westbrook, J. H. Problems of Interpretation and Representation in the Computerization of a Printed Reference Work on Materials Data. In *Computerization and Networking of Materials Databases III*; Barry, T. I., Reynard, K. W., Eds.; ASTM: Philadelphia, 1992.
(19) Grattidge, W.; Westbrook, J. H.; Brown, C.; Novinger, W. B. A Versatile Data Capture System for Archival Graphics and Text. In *Computer Handling and Dissemination of Data*; Glaeser, P., Ed., Elsevier Science Publishing, CODATA: Amsterdam, 1987.
(20) Grattidge, W.; Lund, W. B.; Westbrook, J. H. A Data Capture System for Printed Tabular Data. In *Scientific and Technical Data in a New Era*; Glaeser, P. S., Ed.; Hemisphere Publishing Corp.: New York, 1990.
(21) Grattidge, W. Capture of Published Materials Data. In *Computerization and Networking of Materials Databases II*; Glazman, J. S., Rumble, J. R., Eds.; ASTM STP 1017, ASTM: Philadelphia, 1989; pp 151–174.
(22) Sources for these and subsequent illustrative figures are as follows: Perry:[23] Figures 1a, 1c, 9, and 10; Kirk–Othmer:[24] Figures 1b, 6–8, and 12; *MilHdbk 5*;[25] Figures 3 and 5; *Encyclopedia of Chemistry*:[26] Figure 11; Charlesby;[27] Figure 13; and *Metals Handbook*, Desk Edition:[28] Figure 14.
(23) Perry, J. H., Ed. *Chemical Engineers' Handbook*, 2nd ed.; McGraw-Hill: New York, 1941.
(24) Grayson, M., Ed. *Kirk–Othmer Concise Encyclopedia of Chemical Technology*; John Wiley and Sons: New York, 1985.
(25) Metallic Materials and Elements for Aerospace Vehicle Structures. *Military Handbook MIL-HDBK-5F*, Edition F; U.S. Dept. of Defense: Washington, DC, 1987.
(26) Clark, G. L.; Hawley, G. G., Eds. *The Encyclopedia of Chemistry*; Van Nostrand-Reinhold: New York, 1966.
(27) Charlesby, A. *Atomic Radiation and Polymers*; Pergamon Press: New York, 1960.
(28) Boyer, H. E., Gall, T. L., Ed. *Metals Handbook*, Desk Edition; American Society for Metals: Metals Park, OH, 1984.
(29) Definitions for the majority of terms found in Appendices A–C were taken from Westbrook and Grattidge.[30]
(30) Westbrook, J. H.; Grattidge, W. A Glossary of Terms Relating to Data, Data Capture, Data Manipulation and Databases. *CODATA Bull.* **1991**, *23*, (1 and 2).
(31) Note, while lists may resemble tables in format, logically they are always one-dimensional rather than two-dimensional.