Cambridge Crystallographic Data Centre. II. Structural Data File

F. H. ALLEN, OLGA KENNARD,† W. D. S. MOTHERWELL, W. G. TOWN, and D. G. WATSON*
Crystallographic Data Centre, University Chemical Laboratory, Lensfield Road, Cambridge CB2 1EW, England

Received May 22, 1973

The Cambridge Crystallographic Data Centre is concerned with the retrieval, evaluation, synthesis, and dissemination of structural data obtained by diffraction methods. This paper is Part II of a series describing the work of the Centre and deals with the organization and maintenance of a computerized file of numeric crystallographic structural data.

In Part I of this series,[1] the bibliographic file of the Cambridge Crystallographic Data Centre was described. This paper deals with the structural data file which at present contains numeric data and textual material relating to some 5000 studies by x-ray and neutron diffraction. The primary data from such studies—viz., unit-cell dimensions, symmetry, and atomic coordinates—are highly structured, and thus ideally suited to computerized organization.

The retrieval, storage, and evaluation of this information is discussed, as well as the transformation, where necessary, to a form more suitable for subsequent data synthesis. The primary data stored in the file provide the necessary information for further work in other areas of science such as theoretical chemistry, crystallography, and molecular biology.

CRITERIA FOR INCLUSION

A structural crystallographic study by x-ray or neutron diffraction is included in the file if it meets the following requirements:

> The substance studied contains organic carbon (purely inorganic carbonyls, cyanides etc. are excluded), and it is not a protein or polymer
>
> The study has been published since 1960 and has not been superseded by a later paper by the same author(s)
>
> *Either* three positional coordinates for each nonhydrogen atom in the structure have been determined, though not necessarily recorded in the publication, *or* the determination of two positional coordinates for each nonhydrogen atom has proved sufficient to resolve a chemical ambiguity.

FILE ORGANIZATION

Data relating to a specific publication on a specific compound constitute a data entry, and an example of a typical entry is shown in Figure 1. The information is stored as a series of card-image records where the last eight columns of each card hold the entry identifier or "reference code"; this code corsists of six letters and possibly two digits, and its use has been described in Part I of this series. In the example, the reference code EDCOPT was assigned during the bibliographic registration process[1] and provides the essential link between the structural data entry and the corresponding entry in the bibliographic file.

Major types of data are recorded with respect to a set of card types which are identified by the "word" in the first six columns. Thus ATOM records carry atomic coordinates, BOND records carry bond lengths, and so on. For a data entry, the various card types are ordered in a prescribed sequence, as indicated in Table I, which also lists the complete information content for each type.

Data entries are grouped in the file by chemical class according to the scheme in Table II of Part I[1] and as used in the bibliographic volumes of the "Molecular Structures and Dimensions" series.[2] In contrast to the bibliographic file, where entries are ordered by increasing carbon and hydrogen content, the entries in the structural data file are ordered alphabetically by reference code within each class.

PRIMARY ABSTRACTING AND DATA INPUT

At the initial stage, relevant information is abstracted from the publication for the following card types:

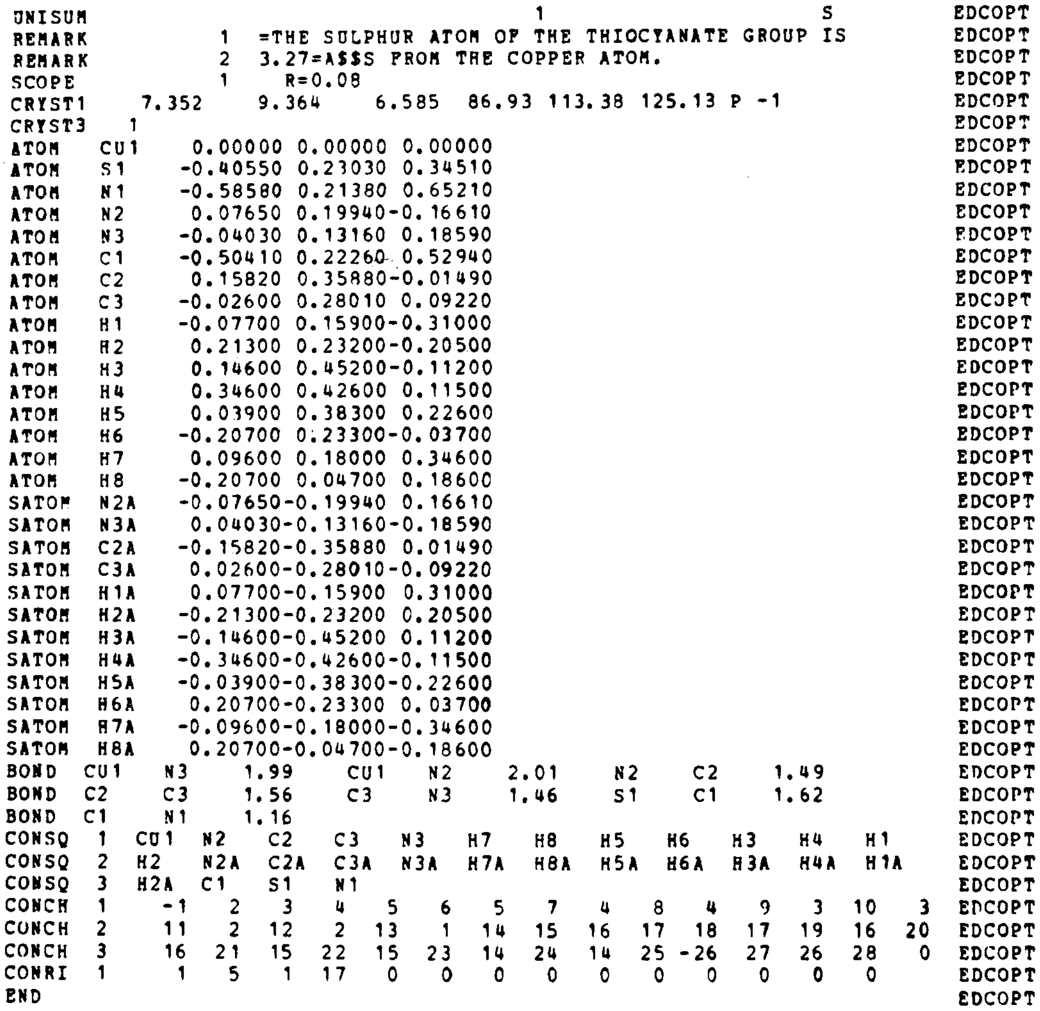
UNISUM	CRYST1	ATOM
SCOPE	CRYST2	BOND
REMARK	CRYST3	CBND
DISORD		

The information content of these cards is itemized in Table I, except for the UNISUM or summary record; the full list of information flags contained on UNISUM is shown in Table II. Only certain of these flags are handled in the primary abstracting stage and these are marked*; for cases of total disorder and for two dimensional studies, we do not input coordinates or bond lengths. All entries must have UNISUM, CRYST1, ATOM, BOND, and END records, even though these may be dummy records, as could be the case for, say, a short conference abstract where frequently no numeric data are reported.

Until fairly recently, all data were coded and then punched on cards, but we now use OCR techniques. Data for the card types UNISUM through CRYST3 in the list above are abstracted on to standard forms by a qualified chemist, while the extensive data for ATOM, BOND, and CBND are annotated in a reprint of the publication. The complete data entry for a specific compound is then typed on an IBM typewriter equipped with an OCR-B golf ball. The typed sheets are processed by a service bureau using SCANDATA OCR hardware. This operation yields a magnetic tape which forms the basic input to our system. The

†External Staff, Medical Research Council.

*To whom correspondence should be addressed.



```
UNISUM                                     1                    S        EDCOPT
REMARK          1  =THE SULPHUR ATOM OF THE THIOCYANATE GROUP IS         EDCOPT
REMARK          2  3.27=A$$S FROM THE COPPER ATOM.                       EDCOPT
SCOPE           1    R=0.08                                              EDCOPT
CRYST1    7.352     9.364    6.585  86.93 113.38 125.13 P -1             EDCOPT
CRYST3    1                                                              EDCOPT
ATOM    CU1     0.00000 0.00000 0.00000                                  EDCOPT
ATOM    S1     -0.40550 0.23030 0.34510                                  EDCOPT
ATOM    N1     -0.58580 0.21380 0.65210                                  EDCOPT
ATOM    N2      0.07650 0.19940-0.16610                                  EDCOPT
ATOM    N3     -0.04030 0.13160 0.18590                                  EDCOPT
ATOM    C1     -0.50410 0.22260 0.52940                                  EDCOPT
ATOM    C2      0.15820 0.35880-0.01490                                  EDCOPT
ATOM    C3     -0.02600 0.28010 0.09220                                  EDCOPT
ATOM    H1     -0.07700 0.15900-0.31000                                  EDCOPT
ATOM    H2      0.21300 0.23200-0.20500                                  EDCOPT
ATOM    H3      0.14600 0.45200-0.11200                                  EDCOPT
ATOM    H4      0.34600 0.42600 0.11500                                  EDCOPT
ATOM    H5      0.03900 0.38300 0.22600                                  EDCOPT
ATOM    H6     -0.20700 0.23300-0.03700                                  EDCOPT
ATOM    H7      0.09600 0.18000 0.34600                                  EDCOPT
ATOM    H8     -0.20700 0.04700 0.18600                                  EDCOPT
SATOM   N2A    -0.07650-0.19940 0.16610                                  EDCOPT
SATOM   N3A     0.04030-0.13160-0.18590                                  EDCOPT
SATOM   C2A    -0.15820-0.35880 0.01490                                  EDCOPT
SATOM   C3A     0.02600-0.28010-0.09220                                  EDCOPT
SATOM   H1A     0.07700-0.15900 0.31000                                  EDCOPT
SATOM   H2A    -0.21300-0.23200 0.20500                                  EDCOPT
SATOM   H3A    -0.14600-0.45200 0.11200                                  EDCOPT
SATOM   H4A    -0.34600-0.42600-0.11500                                  EDCOPT
SATOM   H5A    -0.03900-0.38300-0.22600                                  EDCOPT
SATOM   H6A     0.20700-0.23300 0.03700                                  EDCOPT
SATOM   H7A    -0.09600-0.18000-0.34600                                  EDCOPT
SATOM   H8A     0.20700-0.04700-0.18600                                  EDCOPT
BOND  CU1    N3      1.99    CU1    N2      2.01    N2     C2     1.49   EDCOPT
BOND  C2     C3      1.56    C3     N3      1.46    S1     C1     1.62   EDCOPT
BOND  C1     N1      1.16                                                EDCOPT
CONSQ  1  CU1  N2   C2   C3   N3   H7   H8   H5   H6   H3   H4   H1     EDCOPT
CONSQ  2  H2   N2A  C2A  C3A  N3A  H7A  H8A  H5A  H6A  H3A  H4A  H1A    EDCOPT
CONSQ  3  H2A  C1   S1   N1                                              EDCOPT
CONCH  1     -1    2    3    4    5    6    5    7    4    8    4    9    3   10    3  EDCOPT
CONCH  2     11    2   12    2   13    1   14   15   16   17   18   17   19   16   20  EDCOPT
CONCH  3     16   21   15   22   15   23   14   24   14   25  -26   27   26   28    0  EDCOPT
CONRI  1      1    5    1   17    0    0    0    0    0    0    0    0    0    0       EDCOPT
END                                                                      EDCOPT
```

Figure 1

tape is then converted to a card-image file using prescribed formatting rules. In addition to formatting, the conversion routine manipulates the upper and lower case characters and generates the necessary typesetting signals in the card-image records.

PRELIMINARY DATA CHECKING

The checking of the basic input is performed by three programs, REFMATCH, DATCHECK, and DATENT.

REFMATCH is used to check that the reference code of each new data entry has been correctly typed. The program checks each new code against a master file of all reference codes from the bibliographic file.

DATCHECK is used to check each card record individually and produces both error and warning messages. A large number of different checks are carried out which can be categorised as follows:

Tests for legality of characters

Tests for legality and constitution of fields with respect to the prescribed card formats

Tests for reasonable values for certain data elements—e.g., cell dimensions, melting points, etc.

Tests for internal consistency within a given card record—e.g., that the crystal system indicated by the unit-cell dimensions is compatible with the space group.

DATENT is used to check that the make-up of a data entry is valid. It checks that card records are in the correct sequence and that mandatory card types are present. The internal consistency of the complete entry is also checked, for example:

Atom names present on BOND records shall also be present in the ATOM list

If dummy card records are present, a test is made to see if the appropriate flags on UNISUM are set

A check is made to ensure that no information is duplicated—e.g., two atom names the same, or two sets of *x y z* coordinates the same in an ATOM list

The internal consistency of the important UNISUM record is also examined.

EVALUATION OF DATA

A check on the internal consistency of the numeric data is achieved by recomputing the bond lengths in a structure and comparing these with values reported by the author(s) and coded on BOND records. The program, SYSMOL, written to perform the evaluation is described in detail in Part III of this series,[3] but the essential features will be indicated here.

All bonded networks in the structure are established by using a set of covalent radii and the space-group symmetry to find bonded atom pairs. These networks may contain atoms which are symmetry related to those in the input asymmetric unit. For example, a molecule having a crystallographic center of symmetry as its only symmetry element has an asymmetric unit of one half-molecule whose coordinates are stored on ATOM records. The coordinates of the symmetry-related atoms are stored on SATOM records. The connectivity of the system or crystal chemical unit (ATOM + SATOM) is then logged in compact form on the card types CONSQ, CONCH, and CONRI.

The crystallographically unique bond lengths are recalculated and any bond length found to be shorter, by more than a preset tolerance (usually 0.40Å), than the sum of covalent radii, is flagged as a "short bond" (SB). Each calculated bond length is compared with the corresponding published value (on BOND), and the difference, Δ, is obtained.

The bond is flagged * if 0.02Å ≤, $\Delta <$ 0.05Å (BC1)
or ** if $\Delta \geq$ 0.05Å (BC2)

Finally, for certain common elements, a check is made to ensure that the normal valency is not violated.

A data entry is defined to be an error set if any of the following conditions are established by SYSMOL:

Existence of one or more short bonds (SB)

Existence of one or more bond length comparisons at the **-level (BC2)

Existence of one or more valency check errors (VAL)

All other entries are classed as "error-free."

Table I

Information content of the card types relevant to the structural data file: the order of card types corresponds to an entry where all types are present. Mandatory card types are indicated*.

*UNISUM Summary flags from initial abstraction, evaluation and checking: a full list is given in Table II

REMARK Any textual comment which is considered useful

DISORD Nature of disorder in cases of partial disorder (PD on UNISUM)

ERROR Details of errors in the original publication

SCOPE Data flag for visual, densitometer or diffractometer measurement of intensities

Landolt-Börnstein number (see section on scientific checking).

R-factor(s) indicating the measure of disagreement between theoretical and experimental models

TOLER Special value for bond-length tolerance (see section on evaluation)

RADIUS Special value(s) of atomic covalent radii to be used during evaluation

*CRYST1 Unit-cell dimensions $a,b,c,\alpha,\beta,\gamma$ and space-group symbol in Hermann-Mauguin notation

CRYST2 Standard deviations of unit-cell dimensions, $\sigma(a)$, $\sigma(b)$, $\sigma(c)$, $\sigma(\alpha)$, $\sigma(\beta)$, $\sigma(\gamma)$

Temperature of data collection (°C)

Melting point (°C)

CRYST3 Z-value—i.e., the number of formula units per cell

Measured density (D_m) with standard deviation and calculated density (D_x), together with the temperature (°C) of density measurement

Range of measured densities

SYMTRY Explicit list of symmetry operators, one per record, input only where the space-group origin choice is nonstandard

*ATOM Atom names and x,y,z coordinates for each atom of asymmetric unit

Standard deviations of atomic coordinates for 1960–71 data

SATOM The x,y,z coordinates of any symmetry-generated atoms which bond directly or indirectly to the asymmetric unit. The atom name is amended with an alphabetic code referring to the symmetry operator

*BOND Atom names and bond lengths grouped as three bond lengths per record

CBND Bond lengths corrected for thermal vibration, coded as for BOND

CONSQ Sequence of atom names in the connectivity table, grouped twelve per record

CONCH Chain specification in connectivity table, coded as 15 per record

CONRI Specification of bonds which form ring-closures, coded as seven pairs of atoms per record

*END Termination of data entry

ERROR-FREE SETS

The card records of the input data entry are copied to the output file, but with the following additions and amendments:

On the UNISUM record (see Table II), ERF is set to 1, and BC1 is set if applicable. The flags DA, DB, DCD, DSG may also be set if they were missed on input.

The input ATOM records are replaced by a new set in which the atomic coordinates are listed with respect to the bonded units. SATOM records are added if the set possesses crystallographic symmetry.

The connectivity records CONSQ, CONCH, CONRI are added.

Table II

Information content of the summary record UNISUM. Those flags which may be set during the primary abstracting are indicated*

*TD Structure is totally disordered (implies DA, DB)

*PD Structure is partially disordered

*PC Partial connectivity only will be input, frequently associated with PD

*TDS Two-dimensional structure (implies DA, DB)

*DA Dummy atom record, i.e. no coordinates input

*DB Dummy bond record, i.e. no bond lengths input

*DCD Dummy cell dimensions i.e. these were not published

*DSG Dummy space-group symbol, i.e. this was not published

*CBL Bond lengths corrected for thermal vibration are reported

*AC Absolute configuration determined by x-ray methods

*NT Neutron-diffraction study

*LT Low-temperature study

*MS Maximum standard deviation for C-C bonds, for 1973 data onwards

*AS Average standard deviation for C-C bonds, for 1973 data onwards

WC Wrong connectivity obtained during check process

SB Number of short bonds obtained by the evaluation program

BC1 Number of bond comparisons at the *-level obtained by the evaluation program

BC2 Number of bond comparisons at the **-level obtained by the evaluation program

BM Bond matching problem encountered during evaluation, i.e. bond given on BOND record not found in calculation

VAL Valency check(s) encountered during evaluation

LB Long bond encountered during evaluation, i.e. longer than the sum of covalent radii plus tolerance value

PE Program exit has occurred during evaluation; the reason is indicated in an output message from SYSMOL

ERF Error function, can be coded 1, 2, 3 as follows:

3 indicates that a data entry has *not* been processed by SYSMOL

2 indicates that the data entry has been classed as an *error set* by SYSMOL

1 indicates that the data entry has been classified as *error-free* by SYSMOL

CHK Check status, may be blank, C or S:

C indicates that the data entry has been clerically checked

S indicates that the data entry has been scientifically checked

RPA Refer problem to author, can be coded 1 or 2:

2 indicates a serious error; the data entry cannot be used unless the author replies and the problem is clarified

1 indicates a less serious error in which case the data entry can be used with caution pending a reply

ERROR SETS AND CLERICAL CHECKING

In these cases, the output file from SYSMOL contains the entire input data entry but with the following amendments to UNISUM (see Table II).

ERF is set to 2.
The flags SB, BC1, BC2 and VAL are set as necessary.
The flag DB may be set if omitted at the abstracting stage.

For each error set, the transcription of data from the publication to the card records CRYST1, ATOM, BOND is checked. Errors detected in this process are corrected, and the data entry is reprocessed The principal types of clerical error are transposition of digits and omission of negative signs. However, much more trouble is caused by typographic errors in the original publication, and in certain cases it is possible to rectify these at this stage. If such a correction is made, then an ERROR record is added to the data entry. Any residual errors can be corrected only by a detailed scientific check followed by referral of the problem to the author if necessary.

UPDATING OF THE MAIN FILE

For each new batch of entries, the output file from SYSMOL contains both error sets and error-free sets in a random order. They are immediately sorted into chemical classes (by reference to the bibliographic file), and then alphabetically within class using the programs DISCORD and ORDERDAT. They are then merged with the main structural data file.

The error sets for each batch are clerically checked and new decks generated by reprocessing. The UNISUM flag CHK is set to C, and the new decks are checked by DATENT/DATCHECK to safeguard against card-handling errors. These then replace the old error sets on the main file via the program REPLACER.

SCIENTIFIC CHECKING AND FILE UPGRADING

Those error sets which cannot be corrected by clerical checking should all be examined by scientific personnel and, as a last resort, the problems should be referred to the authors. Because of lack of manpower it has only been possible to attempt scientific checking on a subset of the file corresponding to the 1960–65 literature and for chemical classes 1–43 for the period 1966–68. A total of 1305 entries for the 1960–65 period were thoroughly checked in preparation for the production of a compendium of interatomic distances.[4] The use of the structural data file for the synthesis of data for this volume will be described in a later paper in this series.

During the course of production of the "Interatomic Distances" volume, the file was upgraded by providing cross-references to the recent volume in the Landolt-Börnstein series[5] covering organic crystal data. *R*-factors, which give an indication of the degree of refinement of the structure, were also added to the 1960–65 subset. In our current operation, we input *R*-factors at source, and we expect to upgrade the file uniformly to cover the period 1966–72.

USE OF STRUCTURAL DATA

The user of the file can make a very rapid assessment of the content and "status" of a data entry by checking three flags on the UNISUM record—viz:

DA = 1 : no atomic coordinates stored
ERF = 1, CHK = blank : error-free on first pass through SYSMOL
ERF = 1, CHK = C : error-free after clerical check
ERF = 1, CHK = S : error-free after scientific check
ERF = 2, CHK = C : set still in error after clerical check
ERF = 2, CHK = S : set still in error after scientific check

The other flags listed in Table II serve to amplify this information in a variety of ways.

As indicated earlier, when atomic coordinates are present in the input file then SYSMOL will establish the bonding connectivity provided the data entry is error-free. The connectivity records CONSQ, CONCH, CONRI indicate the chains of connected atoms and ring closure pairs in a compact form; no attempt has been made to assign bond orders on the basis of molecular geometry (this is virtually impossible to do in a uniform manner since hydrogen coordinates have not been reported for a large number of entries). Therefore, a simple search for benzene compounds would yield cyclohexane compounds as well; however, it would be possible to combine a simple search of the connectivity record with a molecular geometry analysis to screen out many of the unwanted compounds.

The connectivity records, in conjunction with the transformed coordinates stored on ATOM and SATOM, provide the fundamental data for further work in various areas of scientific research—e.g.:

Automatic recomputation of various elements of molecular geometry such as bond lengths, bond angles, and torsion angles
Calculation of intermolecular distances for studies of molecular packing
Generation of molecular and packing diagrams—both mono and stereo
Input to calculations in theoretical chemistry

Standard programs are generally available for utilizing the data base as outlined above.

It should be borne in mind when searching the structural data file that no coordinates are available for short communications, totally disordered structures and two-dimensional structures; therefore, no connectivity records exist for these categories. However, a project has been initiated recently to code full "*Chemical Abstracts*-type" connectivity tables for all entries in the file. Since this file of chemical connection tables will be fully retrospective, it is envisaged that it will be used for file searches rather than the crystallographic connectivity records.

FILE STATISTICS

As of 1 January 1973, the flagged (UNISUM) numeric data base comprised 4962 individual entries. Of these 1435 had no coordinates (no error check possible), 3069 were error-free with coordinates and there were 458 error sets. This corresponds to a total file size of 163,600 card images.

The rate of increase of the file is approximately 2000 entries per annum. The backlog consists of some 2000 entries which are being evaluated while a further 1500 await preliminary checking. However, with the recent implementation of direct OCR input, we expect that numeric data will be checked on a current basis by 1 January 1974.

RELEASE OF STRUCTURAL DATA

There are two methods of accessing the numeric data files—through requests for individual data sets directly to the Cambridge Centre, or through accredited data centers on a regional basis.

Accredited data centers are being established in various parts of the world and are intended to serve the needs either of a large research institution or of a scientific community on a geographic basis. An accredited center may lease the whole or a part of the data base and use it to provide a variety of services subject to certain lease restrictions.

ACKNOWLEDGMENT

We thank the Office for Scientific and Technical Information, Department of Education and Science for their support of the activities of the Crystallographic Data Centre, and individual staff members for their advice and helpful discussion. Space for the Centre was provided by the University of Cambridge and computing facilities by the Institute of Theoretical Astronomy (IBM 360/44) and the Cambridge University Computing Service (IBM 370/165). T. Scott of the Medical Research Council Computing Service took a large part in the development of the evaluation program SYSMOL. We are indebted to the following for their assistance during preliminary abstracting, data input, and checkout processes: Stella Weeds, Ann O'Brien, Eileen Doyle, Anne Town, Bernice Roberts, Adele Boskey, Maryla Chiechanowicz and D. Sales.

LITERATURE CITED

(1) Kennard, O., Watson, D. G., and Town, W. G., *J. Chem. Doc.* **12,** 14 (1972).

(2) Kennard, O., Watson, D. G., "Molecular Structures and Dimensions," Vols. 1, 2, 3, and 4 (with Town, W. G.) published for the Crystallographic Data Centre, Cambridge, and the International Union of Crystallography by N. V. A. Oosthoek's Uitgevers Mij, Utrecht, 1970, 1972, 1973

(3) Allen, F. H., Kennard, O., Motherwell, W. D. S., Town, W. G., Watson, D. G., Scott, T., and Larson, A. C., to be published.

(4) Kennard, O., Watson, D. G., Allen, F. H., Isaacs, N. W., Motherwell, W. D. S., Pettersen, R. C., and Town W. G., "Molecular Structures and Dimensions, Vol. A1, Interatomic Distances 1960-65," published for the Crystallographic Data Centre, Cambridge and the International Union of Crystallography by N. V. A. Oosthoek's Uitgevers Mij, Utrecht, 1973.

(5) "Numerical Data and Functional Relationships in Science and Technology," III, Vol. 5a, 5b, Eds. K-H and A. M. Hellwege, Berlin Springer, 1971.

A Materials Index—Its Storage, Retrieval, and Display*

CAROL Z. ROSEN
Information Services Division, American Institute of Physics, 335 East 45th St., New York, N. Y. 10017

Received April 19, 1973

An experimental procedure for indexing physical materials based on simple syntactical rules was tested by encoding the materials in the journal, *Applied Physics Letters,* to produce a materials index. The syntax and numerous examples together with an indication of the method by which retrieval can be effected are presented.

In today's research and development, there are myriad physical materials available to the scientist or engineer for his work. A materials index to the scientific journals can be of great assistance in furthering the scientist's knowledge and accelerating his achievements. Our goals have been to derive a simple scheme compact enough to minimize the number of material entries per article and general enough to encompass the full range of materials and devices discussed in the present and future literature. We have tried to make the coding of the materials easy to learn, simple to recognize, and based, wherever possible, on codes which already exist in the literature. The feasibility of computer retrieval of items on the index played a large role in the selection of these codes.

It is anticipated that the Materials Index can be produced by computer-based photocomposition directly from a computer file containing a list of say, *Applied Physics Letters* (APPLA) articles and their materials. This file in turn could be merged with a file holding the Physics and Astronomy Classification Scheme (PACS)[1] information for these APPLA articles. PACS is currently used in organizing the material appearing in *Current Physics Titles* (CPT), as well as the three monthly journals *Current Physics Advance Abstracts* (CPAA) published by the American Institute of Physics.

To illustrate this point, suppose a user is interested in Kondo-materials, in particular, Pd-Ni and Pd-Mn. The materials index encoding scheme (MIES)-PACS combination could link these two specific alloys with papers about order-disorder phenomena, localized electronic states, resistivity, electronic specific heat, and magnetic impurity interactions. From this material—subject merger, the user will get a picture of the current state of research on these two Kondo alloys. In general, this system can provide detailed retrieval and display for all subjects of physics.

MATERIALS INDEX ENCODING SCHEME (MIES)

Tables I and II display the full glossary for the encoding scheme. Table I lists the "Type" symbols, along with their physical meaning, and a material example. Table II describes each "Delimiter" along with descriptive examples. The "Type" symbol describes the material's physical state or the physical process it underwent such as implan-

*Work supported by NSF Grant GN864