JCAMP-CS: A Standard Exchange Format for Chemical Structure Information in Computer-Readable Form*

J. GASTEIGER,† B. M. P. HENDRIKS, P. HOEVER, C. JOCHUM, and H. SOMBERG

Technische Universität München, Organisch-chemisches Institut Lichtenbergstraße 4, D-8046 Garching, West Germany (J.G.); Duphar B.V., NL-1381 CP Weesp, The Netherlands (B.M.P.H.); Bayer AG., D-5090 Leverkusen, West Germany (P.H.); Beilstein Institut, D-6000 Frankfurt 90, West Germany (C.J.); and Bruker Analytische Meßtechnik GmbH, D-7500 Karlsruhe 21, West Germany (H.S.)

The JCAMP-DX format provides a standard for the exchange of data on IR spectra. Extensions of this format to other spectral data are being developed. Spectral data should always be accompanied by information on the chemical structure of the investigated compounds. The JCAMP-CS format provides definitions for exchanging information on the composition and the stereochemistry, as well as on the 2D and 3D atomic coordinates, of chemical structures. Where possible, the standard was designed to adhere to the conventions of JCAMP-DX. In addition, care was taken to make as simple as possible the conversion to other formats in use for representing chemical structures.

Index Headings: **Structural data; Spectral data; Connection tables; Stereochemistry; Chemical information; Molecular structure; 2D atomic coordinates; 3D atomic coordinates.**

1. INTRODUCTION

In connection with a spectroscopic database project supported by the German government, the decision was made to use the JCAMP-DX format[1] to exchange spectral data (the extension of the format to spectroscopic techniques other than IR is under development by JCAMP-DX subcommittees).

Structures are essential for the characterization of chemical compounds and the assignments of spectral features and must therefore be exchangeable by this format. A subcommittee (membered by the authors) of the "Arbeitskreis Spektroskopie" of both Fachinformationszentrum Chemie and Fachinformationszentrum Energie, Physik, Mathematik was founded to develop a structure-exchange format based on the requirements that it be compatible with JCAMP-DX; that it be able to code all information available in the connection tables of Beilstein, CAS-Online, and Gmelin-Online; and that it be able to support the special needs for band assignments by spectroscopists. Furthermore, the exchange format should be related as closely as possible to the SMD (Standardized Molecular Data) File Format, a structure exchange format in use at a variety of European chemical companies.[2]

Care was taken to keep the information *content* equivalent to the one in the SMD format. It should be realized that an SMD format has wider objectives, being also applicable to generic structures, reactions, and sequences of reactions.

With the JCAMP-CS format, however, we restricted ourselves to representing individual structures in order to keep the specifications as simple as possible. Furthermore, we were under the restriction of using for the labeled data records (LDR) the format already given by the JCAMP-DX standard.

2. SCOPE

(2.1) The JCAMP-CS protocol is meant for public use. These specifications constitute version 3.7 and are copyrighted by the Joint Committee on Atomic and Molecular Physical Data (JCAMP) for the purpose of linking them with the name JCAMP-CS. The right to copy these specifications for scientific purposes is hereby granted. Use of the name JCAMP-CS in software or datafiles implies compliance with sections 3, 4, and 5.

(2.2) To comply with these specifications, the software for encoding must be able to generate an exchange file, and the resulting JCAMP-CS file should be decodable by the software to generate an internal connection table.

JCAMP-CS provides means for defining chemical structures in free-standing JCAMP-CS files or as separate blocks in complex JCAMP-DX files [see Ref. 1 (3.3.2)], more or less as illustrated in table IX of Ref. 1.

3. GENERAL PRINCIPLES

JCAMP-CS adopts the same conventions and label names used for JCAMP-DX. For example, labels like ##DATE=, ##CAS REGISTRY NO=, and ##NAMES=' can be assigned to a structure by including them in the JCAMP-CS block.

The features and sections of the JCAMP-DX specification referred to in the definition of the structure ex-

Received 18 October 1989; revision received 4 August 1990.

† Author to whom correspondence should be sent.

0003-7028/91/4501-0004$2.00/0

change format proposed with this paper are repeated in Appendix A. The CORE of a JCAMP-CS file comprises (see Table I) the following set of obligatory labeled data records (LDR): ##JCAMP-CS=; ##ATOMLIST=, always preceded by ##MOLFORM= and always followed by ##BONDLIST= unless the species does not contain bonds (e.g., ionic compounds or pure elements); and, if applicable, ##CHARGE=. The information on the structure of a chemical species has to be concluded by the LDR ##END=. This LDR has to be placed after the CORE and SHELL information. The (optional) SHELL of a JCAMP-CS file (see Table I) immediately follows the CORE and comprises LDRs which code items such as stereochemical information or coordinates for either graphical representation and/or modeling techniques. Although optional, these LDRs should not be omitted by the encoding software if the internal representation contains this information. The ##BLOCK_ID= preferably precedes ##MOLFORM=.

It should be emphasized that the format defined here is intended for the *exchange* of structural information on chemical species. In order for this format to be acceptable to a wide community, the rules for coding have been designed to be as simple as possible. It is even possible to manually code or decode a chemical structure, but the main intention is that this format will be primarily handled by computer programs. The format will give an unambiguous representation of the molecular structure.

On the other hand, systems that are to *manipulate* structural information (e.g., substructure search, tautomer perception, delocalized bonds) will need an *internal* representation that is more sophisticated than the format suggested here. Thus, it is our belief that a promising representation of the chemical structure has to account for all atoms and all valence electrons and has to differentiate between the different types of bonds (σ, π, or coordinative). Giving such detailed information is usually beyond the scope of the uninitiated. With the specifications of the exchange format, sufficient information is available to generate, by the appropriate software, a more detailed internal representation that is better suited for different applications. The exchange format defined here should give enough information for an unambiguous representation of the molecular structure of organic compounds, organometallic compounds, and inorganic molecular species, including electron-deficient compounds (e.g., boranes).

If more than one structure for different components in a mixture must refer to one spectrum, a separate JCAMP-CS block should be created for each individual structure.

Only molecules having some physical or chemical interaction (e.g., ionic bonds, hydrogen bridges) should be put in a single JCAMP-CS block.

JCAMP-CS software should support these forms of multistructure encoding/decoding in order to prevent repeated transmission of the same large spectral datafile referring to different structures. The label ##BLOCK_ID= is provided to allow for unambiguous reference from inside an LDR in any JCAMP-block to the JCAMP-CS block containing this LDR (of course referencing is only allowed inside the same physical JCAMP transfer file).

TABLE I. Overview of the labeled data records (LDR) of a JCAMP-CS file. The LDRs of the core should be coded in the sequence given in this table, except that ##END= is always the last record, even when SHELL information is given. $$ initiates a comment.

$$CORE of JCAMP-CS	
1. ##TITLE=	$$4.1
2. ##JCAMP-CS=	$$4.2
3. ##ORIGIN=	$$4.3
4. ##OWNER=	$$4.4
5. ##MOLFORM=	$$4.5
6. ##ATOMLIST=	$$4.6
7. ##BONDLIST=	$$4.7
8. ##CHARGE=	$$4.8
9. ##END=	$$4.9
$$SHELL of JCAMP-CS	
10. ##DATE=	$$5.1
11. ##TIME=	$$5.2
12. ##RADICAL=	$$5.3
13. ##STEREOCENTER=	$$5.4
14. ##STEREOPAIR=	$$5.5
15. ##STEREOMOLECULE=	$$5.6
16. ##MAX_RASTER=	$$5.7
17. ##XY_RASTER=	$$5.8
18. ##XYZ_SOURCE=	$$5.9
19. ##MAX_XYZ=	$$5.10
20. ##XYZ_FACTOR=	$$5.11
21. ##XYZ=	$$5.12
22. ##BLOCK_ID=	$$5.13

4. JCAMP-CS CORE-LABELS

(4.1) **##TITLE= (TEXT).** This label conforms to the JCAMP-DX format [see Appendix A (6.1.1)].

(4.2) **##JCAMP-CS= (STRING).** This label is analogous to ##JCAMP-DX= [see Appendix A (6.1.2)].

It has been realized since the introduction of the JCAMP-DX standard that the source (software) that produces a file should be identified. The practice suggested for this purpose is for the name, version, and organization responsible for the software to be indicated here in a comment immediately following the version number of the JCAMP-CS standard.

(4.3) **##ORIGIN= (TEXT).** This label is analogous to ##JCAMP-DX [see Appendix A (7.1.2)].

(4.4) **##OWNER= (TEXT).** This label is analogous to ##JCAMP-DX= [see Appendix A (7.1.3)].

(4.5) **##MOLFORM= (STRING).** This label conforms to JCAMP-DX [see Appendix A (7.2.4)]. In JCAMP-CS this record is obligatory and is used for error checking.

(4.6) **##ATOMLIST= (STRING).**

$$\begin{array}{ccc} AN_1 & AS_1 & NH_1 \\ \vdots & \vdots & \vdots \\ AN_i & AS_i & NH_i \end{array}$$

The ##ATOMLIST= contains an arbitrary number (AN) for the atoms of a molecule followed by the atomic symbol (AS) (including isotope specification, if desired) and the number of those hydrogen atoms (NH) attached to that atom that have not been coded explicitly. Hydrogen atoms can be coded explicitly like any other atom through the specifications AN and AS. They have to be

coded explicitly if hydrogen isotopes are to be specified or if they are referenced in any other LDR.

The atomlist contains one atom per line; the order of the fields is fixed, and the fields are separated by at least one blank. As suppression of trailing zeros is allowed, only AN and AS are always present.

AN: user-assigned atom number: contiguous ascending numbering starting with one.
AS: atomic symbol including isotope description conforming to the ##MOLFORM= conventions [see Appendix A (7.2.4)].
NH: total number of hydrogen atoms directly bonded to the atom specified by AN and AS and not coded explicitly.

(4.7) **##BONDLIST= (STRING).**

$$\begin{array}{ccc} AN1_1 & AN2_1 & BT_1 \\ \vdots & \vdots & \vdots \\ AN1_i & AN2_i & BT_i \end{array}$$

The ##BONDLIST= defines all bonds of a molecule by coding all bonded pairs of atoms AN1 and AN2. It is essential to give *all bonds* between *all atoms* given in ##ATOMLIST= in order to arrive at an unambiguous representation of molecular structure. An attempt has been made to provide, with the exchange format, a representation related as closely as possible to a graphical representation by a structural formula. This representation is essentially a localized valence bond structure. The single, double, triple, and quadruple bonds of such a structure are coded as such (S, D, T, or Q). Aromatic bonds have to be coded as alternating single and double bonds. Any other type of bond, such as coordinative bond, hydrogen bridge bond, electron-deficient bond (in boranes), etc., is given as an additional type of bond (A). The decoding software has to make the correct assignment by taking account of the valence configurations of all atoms. Sigma bonds to hydrogen atoms which are assigned an AN in the ##ATOMLIST= must be given explicitly. Redundancies in the bondlist should be avoided but decoding software should permit it.

The bondlist contains one bond per line; the order of the fields is fixed, and the fields are separated by at least one blank.

AN1: number of the first atom of a bond.
AN2: number of the atom connected to AN1. Preferred coding: AN1 ascending, AN2 ascending, and larger than AN1.
BT: bond type: S, for single bonds; D, for double bonds; T, for triple bonds; Q, for quadruple bonds; A, for any other additional type of bond (e.g., coordinative bond, hydrogen bridge bond, electron deficient bond).

(4.8) **##CHARGE= (STRING).**

$$\begin{array}{cccccc} CH_1 & AN1_1 & AN2_1 & \dots & ANI_1 \\ \vdots & \vdots & \vdots & & \vdots \\ CH_i & AN1_i & AN2_i & \dots & ANI_i \end{array}$$

The formal charges on the atoms of a molecule are given in this list. For each charge value (CH) the atom where the charge is located is specified through AN.

Some charged chemical structures defy a simple valence bond description with localized charges. For example, this is the case for borane hydride anions (e.g., $B_{10}H_{10}^{2-}$) or certain radical anions or cations (e.g., benzene radical cation). For those species, only the net (delocalized) charges (CH) can be given.

In certain cases, there is knowledge of the atomic centers where this charge is delocalized. These atoms can, but need not, be specified by the appropriate atom numbers (AN) as they are listed in ATOMLIST. If charge delocalization is not specified, receiving software may either try to establish the extent of delocalization or assume complete delocalization.

CH: (delocalized) formal charge.
AN: user-assigned atom number (as in ATOMLIST), describing the extent of delocalization.

(4.9) **##END=.** This label conforms to JCAMP-DX.

5. JCAMP-CS SHELL-LABELS

(5.1) **##DATE= (STRING).** The date when the structure was coded, in the form: YY/MM/DD (note order) [cf. Appendix A (7.1.4)].

(5.2) **##TIME= (STRING).** The time when the structure was coded, in the form: HH:MM:SS [cf. Appendix A (7.1.5)].

(5.3) **##RADICAL= (STRING).**

$$\begin{array}{ccccc} UE_1 & AN1_1 & AN2_1 & \dots & ANI_1 \\ \vdots & \vdots & \vdots & & \vdots \\ UE_i & AN1_i & AN2_i & \dots & ANI_i \end{array}$$

The number of unpaired electrons on the atoms of a molecule can be given in this LDR. For each value of unpaired electrons (UE) the atoms where it is located is specified by AN. For delocalized systems all atom centers for spin localization can be indicated.

For localized radicals, the contents of ##RADICAL= will be redundant in terms of the information jointly specified in the LDRs ##ATOMLIST=, ##BONDLIST=, and ##CHARGE=. Decoding software should permit the presence of these redundancies.

UE: number of unpaired electrons.
AN: user-assigned atom number (as in ATOMLIST), describing the extent of delocalization.

(5.4) **##STEREOCENTER= (STRING).**

$$\begin{array}{ccc} AN1_1 & SD_1 & SG_1 \\ \vdots & \vdots & \vdots \\ AN1_i & SD_i & SG_i \end{array}$$

Here, stereochemical information on the configuration of stereocenters (chiral atoms and centers of higher coordination numbers) can be given. As suppression of trailing zeros is allowed, only AN and SD are always coded. Details are given in Appendix B.

AN: number of the atom for which stereochemical information is given.
SD: stereodescriptor.
P, M = priority based on user-defined atom numbers (AN) [see Appendix B (1)].
SG: stereogroup[3] [see Appendix B (2)]. This field

assigns to all centers with a known relative configuration the same alphabetic code:

0 (zero) = SD is an absolute stereodescriptor.
uppercase = SD is a relative stereodescriptor for a pure isomer, but its absolute stereochemistry is unknown.
lowercase = SD is a relative stereodescriptor for a racemic compound.

The SD of one atom out of a relative stereogroup is assigned an arbitrary code. The SD of all other atoms of that stereogroup are assigned with respect to this selected atom. According to this precept, the asymmetric atom of a pure enantiomer of unknown configuration has a unique uppercase SG (single-atom group); the asymmetric atoms of a racemic diastereoisomer both carry the same lowercase SG-code.

The stereodescriptors P and M (in combination with the stereogroup) can also be used for nonchiral atom centers. Usually the receiving software will simply ignore this information. However, if prochirality or hindered rotation cause different spectral/physical behavior for the substituents of such molecules, the individual atom(s) of these substituents can in this way be assigned to their physical datum. In this case hydrogen atoms of prochiral CH_2 groups have to be coded explicitly.

(5.5) **##STEREOPAIR= (STRING).**

$$\begin{array}{cccc} AN1_1 & AN2_1 & SD_1 & SG_1 \\ \vdots & \vdots & \vdots & \vdots \\ AN1_i & AN2_i & SD_i & SG_i \end{array}$$

Information on the stereochemistry at pairs of atoms can be given in this list. The two atoms of the pair may be directly bonded to each other (as in double bonds or single bonds between double bonds). However, the two atoms may also be separated by more than one bond. Then, these two atoms define an axis for giving stereochemical information (e.g., a chirality axis) (see also Appendix B).

SD: stereodescriptor.
P, M = priority based on AN [see Appendix B (3)].
0 (zero) = one isomer, but the configuration is not known.

SG: stereogroup[3] [see Appendix B (2)]. This field assigns to all centers with a known relative configuration the same alphabetic code:
0 (zero) = SD is an absolute stereodescriptor.
uppercase = SD is a relative stereodescriptor for a pure isomer, but its absolute stereochemistry is unknown.
lowercase = SD is a relative stereodescriptor for a racemic compound, or a mixture of E/Z isomers.

(5.6) **##STEREOMOLECULE= (STRING).**

YES

If this information is given, the stereochemistry of the molecule cannot be specified through descriptors in ##STEREOCENTER= and ##STEREOPAIR=. For example, this may be the case with helical structures. In this situation the stereochemistry of the molecule has to be deduced from the three-dimensional coordinates of the atoms in the molecule (cf. ##XYZ=).

(5.7) **##MAX_RASTER= (AFFN).** This value gives the maximum of the absolute value of the X or Y coordinates which are used for graphical representation of the structure. This label is obligatory if the ##XY_RASTER= label is included and is necessary to prevent numeric overflow when the raster coordinates are being read.

(5.8) **##XY_RASTER= (AFFN).**

$$\begin{array}{cccc} AN_1 & X_1 & Y_1 & Z_1 \\ \vdots & \vdots & \vdots & \vdots \\ AN_i & X_i & Y_i & Z_i \end{array}$$

This list contains the coordinates for graphical representation of the structure as integer values. A right-handed coordinate system has to be used with equal scaling factors in the x and y direction. The coordinates for the atoms should be given in ascending order of atom numbers (as specified in ##ATOMLIST=). For each atom in the list, one line containing the values of AN and the X, Y, and optional Z coordinates, each separated by at least one blank, is required. For X and Y, only integers are allowed; positive values are recommended. For the optional Z coordinate, positive (atom before plane of projection) and negative integers (atom behind plane) or 0 (atom in plane) are allowed.

(5.9) **##XYZ_SOURCE= (TEXT).** The source of the 3D coordinates should be given in this field. It should be clearly specified whether the 3D coordinates originate from X-ray structure determination, NMR-NOE measurements, quantum mechanical calculation (which program, which basis set, whether geometry is optimized or not, etc.), molecular mechanical calculation (which program, which force field, which version, etc.), or any other experimental or theoretical method.

(5.10) **##MAX_XYZ= (AFFN).** This value gives the maximum of the absolute value of either the X, Y, or Z coordinates of the atoms listed in the LDR ##XYZ=. If ##XYZ= is used, this label is obligatory, to prevent numeric overflow during data read-in.

(5.11) **##XYZ_FACTOR= (AFFN).** This value gives a multiplication factor used to scale the integer atom coordinates into the original Angstrom units. This field is obligatory if the LDR ##XYZ= is used.

(5.12) **##XYZ= (AFFN).**

$$\begin{array}{cccc} AN_1 & X_1 & Y_1 & Z_1 \\ \vdots & \vdots & \vdots & \vdots \\ AN_i & X_i & Y_i & Z_i \end{array}$$

The values in this list are physically meaningful X, Y, and Z atom-coordinates scaled in such a way that they are represented by integers. Multiplication with the XYZ_FACTOR will rescale them to genuine Angstrom units. A right-handed coordinate system has to be used. For each atom, one line containing the values of AN and the X, Y, and Z coordinate, each separated by at least one blank, has to be given in ascending order of the atom numbers (as specified in ##ATOMLIST=).

(5.13) **##BLOCK_ID= (AFFN).** A positive integer

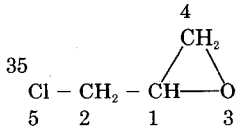
has to be given to assign a number to this JCAMP-CS block, which is unique inside the current JCAMP-transfer file.

The following LDRs of the JCAMP-DX specification[1] are candidates for referring to a chemical structure by using the contents of ##BLOCK_ID= (see Appendix A):

##CROSS REFERENCE=		(A7.1.7)
##PEAK ASSIGNMENTS=	(the ⟨A⟩-field)	(A6.4.4)
##CONCENTRATIONS=	(the ⟨N⟩-field)	(A7.2.13)

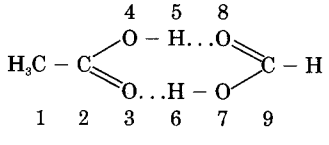
6. EXAMPLES

(6.1) *Example 1*



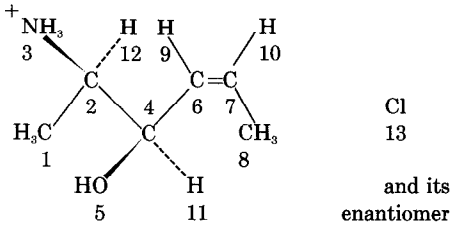

```
##TITLE= isotopically enriched epichlorohydrine
$$ a pure enantiomer of unknown configuration
##JCAMP-CS= 3.7 $$ manually coded by the authors
##ORIGIN= Prof. Dr. J. Gasteiger Tel. +89-32093750
Technical University Munich, Institute of Organic Chemistry
D-8046 Garching, West Germany
##OWNER= Public domain
##DATE= 90/04/30
##CAS NAME= 1-chloro-2,3-epoxypropane
##CAS REGISTRY NO= 13403-37-7
##MOLFORM= C/3 H/5 ^35C1 /O
##ATOMLIST=
$$     AN       AS       NH‡
       1        C        1
       2        C        2
       3        O
       4        C        2
       5       ^35C1
##BONDLIST=
$$     AN1      AN2      BT‡
       1        2        S
       1        3        S
       1        4        S
       2        5        S
       3        4        S
##STEREOCENTER=
$$ AN    SD    SG‡
   1     P     A      $$ P: arbitrarily assigned
##MAX_RASTER= 64
##XY_RASTER_FACTOR= 0.25
## XY_RASTER=
$$     AN       X        Y
       1        9        1
       2        5        1
       3       13        1
       4       11        3
       5        1        1
##END=
```

(6.2) *Example 2*




```
##TITLE= dimer of formic and acetic acid
##JCAMP-CS 3.7 $$ manually coded by the authors
##ORIGIN= Prof. Dr. J. Gasteiger Tel. +89-32093750
Technical University Munich, Institute of Organic Chemistry
D-8046 Garching, West Germany
##OWNER= Public domain
##DATE= 90/04/30
##MOLFORM= C/2 H/4 O/2 • C H/2 O/2
##ATOMLIST=
$$     AN       AS       NH‡
       1        C        3
       2        C
       3        O
       4        O
       5        H
       6        H
       7        O
       8        O
       9        C        1
##BONDLIST=
$$     AN1      AN2      BT‡
       1        2        S
       2        3        D
       2        4        S
       3        6        A
       4        5        S
       5        8        A
       6        7        S
       7        9        S
       8        9        D
##END=
```

(6.3) *Example 3*

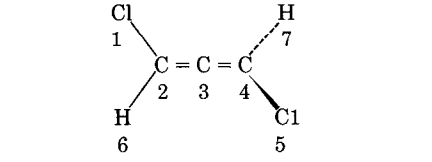


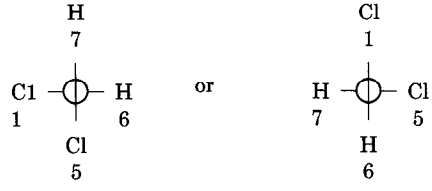

```
##TITLE= racemic diastereoisomer and a cis double bond
##JCAMP-CS= 3.7 $$ manually coded by the authors
##ORIGIN= Prof. Dr. J. Gasteiger Tel. +89-32093750
Technical University Munich, Institute of Organic Chemistry
D-8046 Garching, West Germany
##OWNER= Public domain
##DATE= 90/04/30
##MOLFORM= C/6 H/14 N O • Cl
##ATOMLIST=
$$     AN       AS       NH‡
       1        C        3
       2        C
       3        N        3
       4        C
       5        O        1
       6        C
       7        C
       8        C        3
       9        H
      10        H
      11        H
      12        H
      13        Cl
##BONDLIST=
$$     AN1      AN2      BT‡
       1        2        S
       2        3        S
       2        4        S
       2       12        S
       4        5        S
       4        6        S
       4       11        S
       6        7        D
       6        9        S
       7        8        S
       7       10        S
```

‡ This optional line is for human readability. See description of appropriate label in section 5 for definition of heading terms.

```
##CHARGE=
$$     CH          AN1‡
        +1           3
        −1          13
##STEREOCENTER=
$$     AN          SD        SG‡
        2           P         a
        4           M         a
STEREOPAIR=
$$     AN1         AN2       SD‡
        6           7         P
##END=
```

(6.4) *Example 4*






```
##TITLE= optically active 1,3-dichloroallene
##JCAMP-CS= 3.7 $$ manually coded by the authors
##ORIGIN= Prof. Dr. J. Gasteiger Tel. +89-32093750
Technical University Munich, Institute of Organic Chemistry
D-8046 Garching, West Germany
##OWNER= Public domain
##DATE= 90/04/30
##MOLFORM= C3 H2 Cl2
##ATOMLIST=
$$   AN    AS  NH‡
      1        Cl
      2        C
      3        C
      4        C
      5        Cl
      6        H $$ H-atoms 6 and 7 are present in the atomlist
      7        H $$ because of their use in the XY_RASTER
##BONDLIST=
$$   AN1   AN2   BT‡
      1     2     S
      2     3     D
      2     6     S
      3     4     D
      4     5     S
      4     7     S
##STEREOPAIR=
$$   AN1   AN2   S‡
      2     4     P
##MAX_RASTER= 64
##XY_RASTER_FACTOR= 0.5
##XY_RASTER=
$$          AN      X       Y       Z‡
             1      1       5
             2      3       3
             3      5       3
             4      7       3
             5      9       1      +1
             6      1       1
             7      9       5      −1
##END=
```

More examples can be ordered from the author identified as the author to whom correspondence should be sent, for this paper.

APPENDIX A: JCAMP-DX FORMAT

In order for the specifications of this paper to be internally consistent, the features and sections of the JCAMP-DX specifications given in Ref. 1 are repeated in this appendix, under the numerical identifications used in Ref. 1.

(4.2) **LABELED-DATA-RECORDS (LDR).** An LDR consists of a flagged data-label and an associated data-set, defined below. An LDR begins with a data-label-flag (##) and ends with the next data-label-flag.

(4.4) **DATA-LABELS.** A data-label is the name of an LDR. It is delimited by a data-label-flag (##) and a data-label-terminator (=). ##TITLE= is a DATA-LABEL with flags. A LINE contains no more than one DATA-LABEL. When LABELS are parsed, alphabetic characters are converted to uppercase, and all spaces, dashes, slashes, and underlines are discarded. (XUNITS, x units, xUNITS, and X__UNITS are equivalent.)

(4.5.1) **TEXT.** This label is comprised of data-sets that contain descriptive information for humans, not normally intended to be parsed by computer (i.e., title, comments, origin, etc.).

(4.5.2) **STRING.** A STRING is an alphanumeric field intended to be parsed by computer and read by a human. The form of each string field is described under the LABELED-DATA-RECORD in which it is used.

(4.5.3) **AFFN (ASCII FREE FORMAT NUMERIC).** This is similar to free-form input of BASIC. A field which starts with a +, −, decimal point, or digit is treated as a numeric field and converted to the internal form of the target computer. E is the only other allowed character. A numeric field is terminated by E, comma, or blank. If E is followed immediately by either + or − and a two- or three-digit integer, it gives the power of 10 by which the initial field must be multiplied.

(6.1.1) **##TITLE= (TEXT).** Required as the first LDR of all JCAMP-DX FILES and BLOCKS. Software which decodes JCAMP-DX files should check for the initial ##TITLE= to prevent accidental processing of non-JCAMP-DX files. For a BLOCK containing a spectrum, ##TITLE= should contain a concise description of the spectrum that is suitable as a title for a plotted spectrum. Software which generates a simple JCAMP-DX file should copy the internal field which most closely meets this requirement.

(6.1.2) **##JCAMP-DX= (STRING).** Required immediately after ##TITLE= of each JCAMP-DX BLOCK to show the version of JCAMP-DX; for example: ##JCAMP-DX=4.24.

(6.1.5) **##END=.** Closes a JCAMP-DX BLOCK. There must be one ##END= for each ##TITLE= in a JCAMP-DX file.

(6.4.4) **##PEAK ASSIGNMENTS= (STRING).** List of peak assignments for components, functional groups, or vibrational modes in the form:

##PEAK ASSIGNMENTS = (XYA) or (XYWA)
$(X_1[,Y_1][,W_1]\ \langle A_1\rangle)$
. . . .
$(X_i[,Y_i][,W_i]\ \langle A_i\rangle)$

Parentheses indicate the start and end of each peak entry; square brackets indicate optional elements. Each peak

entry starts on a new line, and may continue on following lines if necessary. X and Y units are specified by ##XUNITS= and ##YUNITS=, respectively. W stands for width. The width-*function* is specified by a comment on the line(s) below ##PEAK ASSIGNMENTS=. The symbol A represents a string describing the assignment (enclosed in angle brackets).

(7.1.2) **##ORIGIN= (TEXT).** Name of organization, address, telephone number, name of individual contributor, etc., as appropriate. This information is not optional; the originator of a JCAMP-DX file should always see that it is included.

(7.1.3) **##OWNER= (TEXT).** Name of owner of a propriety spectrum. The organization or person named under ##ORIGIN= is responsible for the accuracy of this field. If data are copyrighted, this line should read "COPYRIGHT (C) ⟨year⟩ by ⟨name⟩." This information is not optional; the originator of a JCAMP-DX file should always see that it is included.

If ##OWNER= contains PUBLIC DOMAIN, the implication is that the data may be copied without permission on the authority of whoever is named under ##ORIGIN=. Absence of ##OWNER= does not imply permission to copy.

(7.1.4) **##DATE= (STRING).** Date when spectrum was measured, in the form: YY/MM/DD (note order).

(7.1.5) **##TIME= (STRING).** Time when spectrum was measured, in the form: HH:MM:SS.

(7.1.7) **#CROSS REFERENCE= (TEXT).** Cross references refer to additional spectra of the same sample, i.e., different thickness, mulling agent, polarization, temperature, time, etc., or serve to link a peak table or interferogram with a spectrum.

(7.2.3) **#NAMES= (STRING).** Common, trade, or other names. Multiple names are placed on separate lines.

(7.2.4) **##MOLFORM= (STRING).** Molecular formula. Elemental symbols are arranged with carbon first, followed by hydrogen, and then remaining element symbols in alphabetic order. The first letter of each elemental symbol is capitalized. The second letter, if required, is lower case.

One-letter symbols *must* be separated from the next symbol by a blank or a digit. Sub/superscripts are indicated by the prefixes / and ^, respectively. Sub/superscripts are terminated by the next nondigit. The slash may be omitted for subscripts. For readability, each atomic symbol may be separated from its predecessor by a space. For substances which are represented by dot-disconnected formulas (hydrates, etc.), each fragment is represented in the above order, and the dot is represented by *. Isotopic mass is specified by a leading integer. In the case of isotope specification, a blank separator is obligatory. D and T may be used for deuterium and tritium, respectively. Examples:

C2H4O2 or C2 H4 O2 (Acetic acid)
C6 H9 Cr O6 * H2 O
(Chromic acetate monohydrate)
H2 ^17O (Water, mass 17 oxygen)

(7.2.5) **##CAS REGISTRY NO= (STRING).** CAS Registry Numbers for many compounds can be found in *Chemical Abstracts* indices, the Merck Index, or CAS ON-LINE.

(7.2.13) **##CONCENTRATIONS= (STRING).** List of unknown components and impurities and their concentrations in the following form, where N stands for name (enclosed in angle brackets if more than one word); C, for concentration; and U, for units of concentration:

##CONCENTRATIONS= (NCU)
(N_1, C_1, U_1)
....
(N_i, C_i, U_i)

The group for each component is enclosed in parentheses. Each group starts a new line and may continue on following lines.

(7.5.2) **$$.** Comments may be entered at any point in a line by prefixing the first word of the comment by $$. Such comments continue only to the end of the current line, and they do not terminate an LDR.

APPENDIX B: STEREOCHEMICAL DESCRIPTORS AND STEREOGROUPS

The generally accepted nomenclature systems for specifying the stereochemistry of molecules are the R,S and E,Z nomenclatures.[5,6] The assignment of the R/S or E/Z descriptors asks for an ordering of the ligands at chirality centers, chirality axes, or chirality planes, as well as at double bonds, according to the Cahn-Ingold-Prelog (CIP) rules.[5] Ordering the ligands according to the CIP rules can be delegated to appropriate software. Programs that perform this task have been written.[4,7] However, these programs are not yet generally available. Thus, we have refrained from allowing R/S and E/Z descriptors in the exchange format, as processing of this information would ask for an automatic execution of the CIP-rules.

(1) For specifying central or axial chirality through P or M descriptors the user-defined atom numbers (AN) are taken instead of the priority ordering through the Cahn-Ingold-Prelog (CIP) rules.[5] On the basis of the ordering of the atoms through the user-defined indices, the P or M descriptors are determined in the same way as the R or S descriptors. Thus, for a tetrahedral chiral center, the atom of lowest priority (highest AN) is placed away from the observer. If the sequence of the other three atoms in ascending order of AN is clockwise, the center is assigned the stereodescriptor P (otherwise, M). An implicitly coded hydrogen atom will be assigned the highest value for AN.

(2) Quite often only the relative stereochemistry, and not the absolute configuration, of a stereocenter is known. This fact has to be specified in the stereochemical information.

Furthermore, in some cases, the relative configuration of parts of a molecule is only known, and there might be several such independent parts.[3] These situations have to be differentiated by assigning these stereocenters to different stereogroups. Assigning stereogroup codes is a digital representation of Maehr's graphical approach[3] to a *complete* representation of the stereochemical information contents of a chemical structure. The relative stereodescriptor for partly unknown but pure enantiomers is indispensable for a structure exchange format.

The "racemic" case is a very compact code for a mixture with unknown or equal amounts of all implied enantiomers or (dia)stereoisomers. The procedure described in section 3 for coding mixtures should be used in all cases where more (or different) information about the content of the "racemic" mixture is available. The racemic coding option is superfluous if one is prepared to use the procedures of section 3 in any "racemic" case.

(3) For specifying the stereochemistry at an axis defined by a pair of atoms, again the user-defined atom numbers are taken [cf. Appendix B(1)].

Two situations may arise: (A) The ligands at the two atoms of the pair may all lie in one plane (e.g., when one is dealing with a double bond or with cumulenes with an odd number of double bonds); or (B), the ligands may not all lie in one plane (e.g., this is the case with allenes or cumulenes with an even number of double bonds).

In both cases, at each atom of the pair which defines the axis, that atom directly bonded to it (ligand) which has the lowest number (highest priority) is considered for the assignment of a stereochemical descriptor.

In the case of (A), the descriptor P is assigned if the two ligand atoms are on the same side of the axis; otherwise the descriptor M is given.

In the case of (B), a projection is made along the axis (it can be shown that the stereochemical descriptor is independent of the direction of projection). Then, a rotation is made to bring the ligand with lowest number at the front atom on top of the highest-priority ligand on the remote atom. If this rotation is clockwise, the descriptor P is assigned; otherwise M is given.

A stereodescriptor can also be assigned to bonds with a bond type (BT) different from D (e.g., the central C-C bond of butadiene-1,3 or the C-N bond in an amide or any other bond involved in restricted rotation). The receiving software can simply ignore this information if it does not support this extended use of stereoisomerism.

ACKNOWLEDGMENTS

Support of this work by the Bundesminister fuer Forschung und Technologie and Deutsches Forschungsnetz Verein is gratefully acknowledged.

1. R. S. McDonald and P. A. Wilks, Jr., Appl. Spectrosc. **42,** 151 (1988).
2. H. Bebak, C. Buse, W. T. Donner, P. Hoever, H. Jacob, H. Klaus, J. Pesch, J. Roemelt, P. Schilling, B. Woost, and C. Zirz, J. Chem. Inf. Comput. Sci. **29,** 1 (1989).
3. H. Maehr, J. Chem. Educ. **62,** 114 (1985).
4. L. Gann and J. Gasteiger in *Software-Entwicklung in der Chemie,* J. Gasteiger, Ed. (Springer, Berlin, 1987).
5. R. S. Cahn, C. Ingold, and V. Prelog, Angew. Chem. **78,** 413 (1966); *idem.,* Angew. Chem. Int. Ed. Engl. **5,** 385 (1966); V. Prelog and G. Helmchen, Angew. Chem. **94,** 614 (1982); *idem.,* Angew. Chem. Int. Ed. Engl. **21,** 567 (1982).
6. J. E. Blackwood, C. L. Gladys, K. L. Loening, A. E. Petrarca, and J. E. Rush, J. Amer. Chem. Soc. **90,** 509, 2203 (1968).
7. E. F. Meyer, J. Chem. Educ. **55,** 780 (1978); *idem.,* J. Comput. Chem. **1,** 229 (1980).