Recent ASTM Standardization Developments for Chemical Information

Charles E. Gragg

Regulatory and Scientific Information Department, Burroughs Wellcome Company, Research Triangle Park, North Carolina 27709-4498

Received June 4, 1992

Within the soope of ASTM Committee E49 on Computerization of Material and Chemical Property Data, subcommittee E49.51 on Chemical Structural Information (CSI) has developed a draft of a standard specification for computerized CSI, focused on information content. This paper reports on recent progress to reach consensus made at ASTM meetings and urges support for the proposed specification. Concepts concerning transfer of information-rich 3D chemical structural information to information-poor formats are discussed. Examples of existing and proposed formats are given.

INTRODUCTION

ASTM (The American Society for Testing and Materials) supports the process leading to voluntary consensus standards. The terms voluntary and consensus describe the work needed to write and adopt standards, as well as the perception and acceptance of standards worldwide. Committees of interested individuals volunteer time and effort to create and improve documents that are published by ASTM and made available to people everywhere. In response, producers of materials, chemicals, testing methods, systems and services are able to claim that their products meet ASTM standards.

ASTM does not certify that a product conforms to one of its standards. It is up to the producer and the user of a product to agree. When ordering a product, users can specify that it should conform to an accepted standard.

The main committees of ASTM deal with standards for ferrous metals, nonferrous metals, nonmetals, cement, building materials, miscellaneous materials, miscellaneous subjects, end-use materials, corrosion, and other subjects. The publications on steel are very popular. Steel is everywhere: I-beams, pipe, prison bars, and fencing foils.

ASTM standards are highly visible in the hardware store on pipes and other building materials. Among other things on the outside of some sections of pipe is printed the ASTM designation D-3915. That means that the manufacturer has claimed that this product conforms to the ASTM standard designated D-3915. A tube for potable water bearing the designation ASTM D-3915 is made out of polyvinyl chloride (PVC). It is important to recognize the difference between the product and the standard to which it conforms.

CHEMICAL INFORMATION

What business does ASTM have with chemical information?

One of the main committees of ASTM is E49 on Computerization of Material and Chemical Property Data, chaired by John Rumble. This committee is divided into two sections: materials and chemicals. Since all materials are made up of chemicals, we tried to find a single word that could be used for both. We looked in Roget's Thesaurus and found the word "substance", but that is associated with drug abuse. We also found the word "stuff". We could not call ourselves the committee on computerization of data about "stuff".

The chemical section of E49 is represented by three subcommittees: E49.51 on Chemical Structural Information (CSI); E49.52 on Analytical Sciences Data, chaired by Dr. Richard Lysakowski; and E49.53 on Physical Chemical Properties, chaired by Dr. Malcolm Chase.

Subcommittee E49.52 on Analytical Sciences Data has worked on drafts of standards for gas and liquid chromatography and mass spectrometry. Data dictionaries are being prepared.

Subcommittee E49.53 on Physical Chemical Properties has been preparing documents on energetic materials and heat capacity data.

Other subcommittees in E49 have interests in both the material and chemical areas, such as subcommittee E49.05 on Data and Database Quality, chaired by Gil Kaufman, and subcommittee E49.04 on Data Exchange, chaired by John Rumble. All subcommittees work in concert with E49.03 on Terminology, chaired by Jack Westbrook, and more recently by Jerry Glazman.

ASTM has standard language for use in writing documents, and some of the definitions, taken from the *ASTM Form and Style Guide*,[1] also known as the "Blue Book", are shown here.

The word "standard" means "a document that has been developed and established within the consensus principles of the Society and that meets the approval requirements of ASTM procedures and regulations".

The word "specification" means "a precise statement of a set of requirements to be satisfied by a material, product, system, or service that indicates the procedures for determining whether each of the requirements is satisfied."

In standard documents, the four verbs "shall", "should", "may", and "will" indicate mandatory, recommended, optional, and future provisions, respectively. Their use allows rapid changes to be made during the consensus process. If voters feel that an optional provision is really mandatory, then the word "should" can be changed easily to the word "shall".

CHEMICAL STRUCTURAL INFORMATION

Subcommittee E49.51 on Chemical Structural Information was formed following a meeting in Boston in April 1990, during the ACS meeting. The subcommittee met three more times that year, including during the Washington ACS meeting, during the San Antonio ASTM meeting week, and during the London Online meeting.

Standard Molecular Data (SMD) Format[2] was adopted unanimously as the basis of the proposed ASTM standard. SMD format version 5.0, as described in ref 2, is available for implementation by anyone. SMD version 4.3 has been available for a number of years.

0095-2338/93/1633-0018$04.00/0

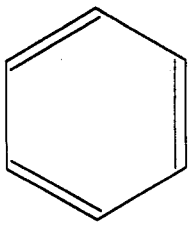

Figure 1. Benzene.

```
}GLOBAL
/DEFAULT
/END
}END
}SCOPE
@SOURCE_PROG MANUAL
@DATE 90-11-08
/MOLECULE benzene
>NODE
)ATOM
 1 C
 2 C
 3 C
 4 C
 5 C
 6 C
)END
>END
>CONVENTION SIMPLE
)BOND
 1  2 DOU
 1  6 SIN
 2  3 SIN
 3  4 DOU
 4  5 SIN
 5  6 DOU
)END
)HCOUNT
 1  1
 2  1
 3  1
 4  1
 5  1
 6  1
)END
>END
>DISPLAY_COORD
  1 20400 24200
  2 23100 23000
  3 23100 20600
  4 20400 19400
  5 17700 20600
  6 17700 23000
>END
/END
}END
```

Figure 2. SMD 5.0 format example.

Note that ASTM E49.51 is not creating a new standard format. Nor are we attempting to restate what has already been published. We are also not attempting to explain or champion an existing format.

Many formats already exist. Some of them are very robust. In general, the more robust a format, the more explicit it is. There is an amount of essential information in every format. Part of the current work of subcommittee E49.51 is to state the mandatory provisions for an acceptable chemical structural information file.

Benzene (Figure 1) can be represented using SMD version 5.0 as shown in Figure 2. This is similar to SMD 4.3. Information about atoms, bonds, hydrogen counts, and display coordinates are given in separate blocks. The display coordinates do not have any chemical significance. They are not related to bond lengths or bond angles. They are there to give a presentable picture. Additional blocks and subblocks of information about atoms, bonds, groups, and other chemical structural information can appear in a file in SMD format.

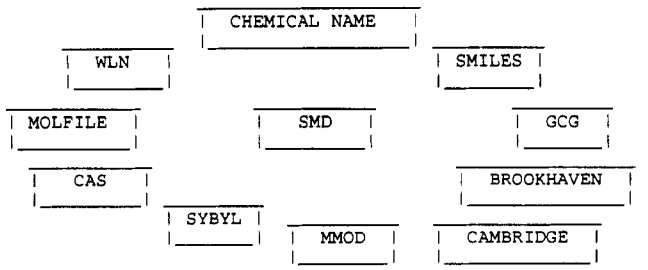

Figure 3. Some formats for chemical structural information.

SMD 4.3 is currently read and written by many chemical drawing packages on computers. SMD is suitable for representing one or more molecules, which can in turn represent pure chemicals, mixtures, formulations, reactions, transition states, patent structures, and other chemical structural information.

Other formats for chemical structural information exist (see Figure 3). The chemical name is a format for CSI that has been evolving with Chemical Abstracts and IUPAC being the major players. Wiswesser Line Notation (WLN)[3] and SMILES[4] are one-dimensional notations that serve well in computers that have limited vocabularies and limited processing power. WLN uses upper case letters, numerals, and the space, dash, ampersand, and slash characters. SMILES uses more characters from the printable ASCII set. Now that we have fast computers that use upper AND lower case, and can handle arrays of numbers and characters, we do not need to restrict ourselves to limited formats. But at the same time we do not want to neglect the valuable contribution of these formats.

Single letter codes are used to great effect by GCG,[5] which stands for the set of programs from the Genetics Computer Group at the University of Wisconsin headed by Dr. John Devereaux. Using four letters A, C, G, and T, the suite of programs in GCG can quickly search and sort out genetic and protein information that would be cumbersome if DNA and protein molecules were stored as their two-dimensional or three-dimensional representations.

There is a need to have three-dimensional information for biological macromolecules available, as shown by the Brookhaven Database.[6] However, one would not use a 3D database for sequence searching.

Three-dimensional data for chemical structures is also available from ancillary data published with journal articles and from the Cambridge Crystallography Database;[7] from molecular modeling programs such as Clark Still's Macro-Model,[8] Tripos' Sybyl;[9] and from 3D databases such as those associated with Chemical Abstracts Service[10] and Molecular Design Ltd. (MDL) MACCS-3D.[11] The MDL MOLFILE format[11] is now in the public domain, following an announcement in October 1991.

Many of these formats may contain one-dimensional, two-dimensional, and three-dimensional information about chemical structure.

The MDL MOLFILE from version 2.0 of MACCS-II is shown in Figure 4 for a molecule of benzene in two-dimensions. Cartesian coordinates in the x, y, and z axes are given, even though the z-axis values are zero. Note that hydrogen atoms may be shown explicitly using the MOLFILE format. The MOLFILE differs from SMD format in that each type of

```
BENZENE
cgMACCS-II01309213242D 1   0.00173     0.00000     0
This is a MOLFILE for Benzene.
  6  6  0  0  0  0               1 V2000
    1.3339    0.7699    0.0000 C   0  0  0  0  0  0
    0.0000    1.5398    0.0000 C   0  0  0  0  0  0
    1.3339   -0.7699    0.0000 C   0  0  0  0  0  0
   -1.3339    0.7699    0.0000 C   0  0  0  0  0  0
    0.0000   -1.5398    0.0000 C   0  0  0  0  0  0
   -1.3339   -0.7699    0.0000 C   0  0  0  0  0  0
  1  2  1  0  0  0
  1  3  2  0  0  0
  2  4  2  0  0  0
  3  5  1  0  0  0
  4  6  1  0  0  0
  5  6  2  0  0  0
M  END
```

Figure 4. MOLFILE format example.

data is not introduced by a brief descriptive phrase. Types of data are grouped together in the MOLFILE format, whereas in SMD format, different types of data are listed linearly, making for a long, narrow file.

The MOLFILE has a header that allows a place for a name for the chemical entity, the initials of the person writing the MOLFILE, the name of the software program that wrote the MOLFILE, a date and time stamp when the MOLFILE was written, whether the MOLFILE is two-dimensional or three-dimensional, other information, and a comment. Then there is a line containing the number of atoms and bonds and the version of the MOLFILE. Since this is readable by FORTRAN, the position of every number is important. The maximum number of atoms or bonds in a molecule in MACCS-II is 255. Even if this limit is raised, the FORTRAN-readability of this file limits this number to 999.

There follows a block of atom data consisting of x, y, and z coordinates, atom symbol, and numerical codes that can be used to show charge, radical, and other atom properties.

Following this is a bond block in which the columns indicate the start of a bond, end of a bond, type of bond, and other numerical codes for bond characteristics. The end of the file has an appropriate character string.

A SMILES[4] for benzene is c1ccccc1. Another is C1=CC=CC=C1. Neither contains information about atom placement in 2D or 3D space. In order to view a picture of benzene on screen, an algorithm must interpret the SMILES.

Algorithmic interpretation is also necessary when converting Wiswesser Line Notation (WLN) to a connection table format such as MOLFILE. Programs such as DARING and WLNCT/LAYOUT[11] are able to convert WLN format to MOLFILE format with value added. This conversion creates two-dimensional structure representations from one-dimensional line notations. The WLN for benzene is RH. Note that few have seen the need to write conversion programs going to WLN.

Conversion of SMILES to MOLFILE using CCT[4] or GEMINI[4] works well in the majority of cases, but fails with complex molecules. Such is the case with Buckminsterfullerene (see Figure 5). The CA index name until recently was footballene. It is now [5,6]fullerene-C60. The CAS registry number is 99685-96-8, and information can be found easily using that number. Note that a typographic error in the registry number will lead to a totally different compound.

The MOLFILE for fullerene is too long to show easily on one figure. The molecule shown in Figure 5 represents a three-dimensional MOLFILE, generated by drawing half a molecule of fullerene in MacroModel,[8] pulling the central atoms and pushing the peripheral atoms in the z direction, minimizing the resulting structure to a cup, copying the cup, rotating one cup and connecting another cup to the first, being careful to

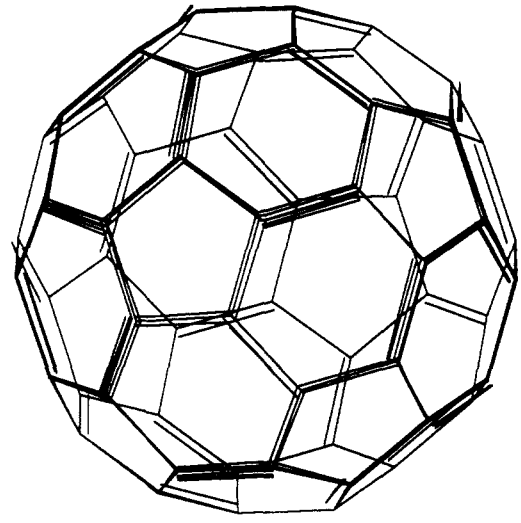

Figure 5. [5,6]Fullerene-C60.

```
c12c3c4c5c1c6c7c8c2c9c%10c3c%11c%12c4c%13c%14c
5c%15c6c%16c7c%17c%18c8c9c%19c%20c%10c%11c%21c
%22c%12c%13c%23c%24c%14c%15c%25c%16c%26c%17c%2
7c%18c%19c%28c%20c%21c%29c%22c%23c%30c%24c%25c
%26c%31c%27c%28c%29c%30%31 buckminsterfullerene c60
```

Figure 6. SMILES format example.

connect top to top, bottom to bottom, front to front, and back to back. When the resulting molecule is minimized, a molecule of fullerene is represented. This was written as a MacroModel file, which was then converted to a MOLFILE using a program written by Jim Bentley of Burroughs Wellcome Co.[12]

Three-dimensional information is valuable. Once it is obtained, a subsequent conversion to an information-poor format should be avoided, unless it is necessary. This can be done using MACCS-3D. The two-dimensional representation is important in MACCS-3D since a 3D database consists of 2D structures each registered once, and one or more 3D models associated with each structure.

Conversion of a MOLFILE for fullerene to SMILES using CCT[4] gives the result shown in Figure 6. A newer conversion system from Daylight is called GEMINI.[4] DEPICT[4] does not work well on this SMILES.

REACHING CONSENSUS

The process to reach consensus is now underway. A "standard format" does not exist. A format is the product of a producer. As stated earlier, the product may conform to a standard, but the product is not the standard. Changes and improvements to a format would be disallowed if a single format is specified in a standard. Rather, a standard specification to which a format can conform is under preparation.

It has been strongly suggested that SMD format be translated into a simpler format and retain its information content. One possible format for chemical structural information is the STAR File format.[13,14] STAR/CIF is used in the Crystallographic Information File (CIF) format which has been adopted by the International Union of Crystallography (IUCr) for use in publishing ancillary crystallographic data in Acta Crystallographica. A field for a chemical connection table is included in CIF.[14]

Shown in Figure 7 is a tentative SMD/STAR format example for benzene. Data is introduced using brief lines of descriptive text. There are five columns of atom data, and there are five lines of introduction above them. There is no

```
data_benzene

_audit_source_creation_purpose      'ASTM draft specifications'
_audit_source_creation_date         '1992-02-06 08:14EST'
_audit_source_creation_method       'manual entry'

loop_
    _atom_identity_node
    _atom_identity_symbol
    _atom_identity_hydrogens
    _atom_identity_display_coord_X
    _atom_identity_display_coord_Y

      1   C  1  315  170
      2   C  1  270  150
      3   C  1  270  110
      4   C  1  315   90
      5   C  1  360  110
      6   C  1  360  150

loop_
    _atom_bond_node_1
    _atom_bond_node_2
    _atom_bond_order_simple

     1  2  single
     1  6  double
     2  3  double
     3  4  single
     4  5  double
     5  6  single
```

Figure 7. Tentative SMD/STAR format example.

3D information in this example, but suitable 3D information could be added easily.

Another possible format (not shown) is the netCDF format[15] from the UNIDATA center in Colorado. This is also a format in which introductions precede data. Not only is the name for the data given but also the type of data and its format. This is a more robust format. It offers advantages of greater information content, widespread use, and established support. A disadvantage is that it adds complexity.

CONCLUSIONS

The work of subcommittee E49.51 has been to clarify and describe the chemical structural information content that shall be found in files conforming to the standard. Separation of content from format was necessary. Two documents are currently in development. Ballots are now being counted for the first of these documents, and the results of the ASTM balloting were available at the ASTM meeting held in Pittsburgh, May 18–20, 1992.

What are the implications? When it becomes available, an ASTM standard specification for computerized Chemical Structural Information (CSI) files or datasets will

1. be used in exchange of CSI among computer programs, databases, and systems
2. be used in electronic publishing of CSI
3. be used in electronic submission of CSI to regulatory agencies and
4. allow various formats for computerized CSI to conform

Does it conform? Having a standard specification in existence will allow the producer of a format for computerized CSI to answer the question: "Does this format for computerized CSI conform to this standard specification"? It is up to the producer and the user to agree that a format conforms.

The draft standard document cannot be shown here. Copies are available on request, and after receipt of a draft document, the following standard caveat shall be in force: "This document is part of the ASTM standards process and is for ASTM committee use only. It shall not be reproduced or circulated or quoted, in whole or in part, outside of ASTM committee activities except with the approval of the chairman of the committee having jurisdiction or the President of the Society".[1]

Table I

date	location
May 1993	Atlanta, followed by the 1st International Symposium on Chemical Data Standards
Nov 1993	Gaithersburg, MD, followed by the 4th International Symposium on Materials Data Standards
May 1994	Montreal
Nov 1994	Phoenix
May 1995	Columbus, OH, followed by the 2nd International Symposium on Chemical Data Standards
Fall 1995	Japan, followed by the 5th International Symposium on Materials Data Standards

Future meetings of ASTM E49 are shown in Table I.

Interested companies and individuals are encouraged to lend expertise to the voluntary consensus process. ASTM E49.51 gratefully acknowledges cooperation received from IUPAC and the Chemical Structure Association (CSA). Financial support given to CSA from many companies and individuals is gratefully acknowledged.

For further information, please contact the author, or contact the staff manager for ASTM E49: Ms. Teresa Cendrowska, ASTM, 1916 Race Street, Philadelphia, PA 19103-1187.

REFERENCES AND NOTES

(1) *Form and Style for ASTM Standards*, 8th ed.; ASTM: Philadelphia, PA, 1989.
(2) Barnard, J. M. Draft Specification for Revised Version of the Standard Molecular Data (SMD) Format. *J. Chem. Inf. Comput. Sci.* **1990**, *30* (1), 81–96.
(3) Smith, E. G. *The Wiswesser Line-Formula Chemical Notation*; McGraw Hill: New York, 1968.
(4) Daylight Chemical Information Systems, Inc., Irvine, CA 92715.
(5) Devereaux, J.; Haeberli, P.; Smithies, O. A Comprehensive Set of Sequence Analysis Programs for the VAX. *Nucleic Acids Res.* **1984**, *12* (1), 387–395.
(6) (a) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F., Jr.; Brice, M. D.; Rodgers, J. R.; Kennard O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank: A Computer-based Archival File for Macromolecular Structures. *J. Mol. Biol.* **1977**, *112*, 535–542. (b) Abola, E. E.; Bernstein, F. C.; Bryant, S. H.; Koetzle, T. F.; Weng, J. Protein Data Bank. In *Crystallographic Databases—Information Content, Software Systems, Scientific Applications*; Allen, F. H., Bergerhoff, G., Sievers, R., Eds.; Data Commission of the International Union of Crystallography: Bonn, Cambridge, Chester, 1987; pp 107–132.
(7) (a) Allen, F. H.; Bellard, S. A.; Brice, M. D.; Cartwright, B. A.; Doubleday, A.; Higgs, H.; Hummelink, T.; Hummelink-Peters, B. G.; Kennard, O.; Motherwell, W. D. S.; Rodgers, J. R.; Watson, D. G. *Acta Crystallogr.* **1979**, *B35*, 2331. (b) Allen, F. H.; Kennard, O.; Taylor, R. *Acc. Chem. Res.* **1983**, *16*, 146–153.
(8) Mohamadi, Fariborz; Richards, N. G. J.; Guida, W. C.; Liskamp, R.; Lipton, M.; Caufield, C.; Chang, G.; Hendrickson, T.; Still, W. C. MacroModel—An Integrated Software System for Modeling Organic and Bioorganic Molecules Using Molecular Mechanics. *J. Comput. Chem.* **1990**, *11* (4), 440–467.
(9) SYBYL Molecular Modeling Package, Version 5.4. Tripos Associates, Inc., St. Louis, MO 63144.
(10) Chemical Abstracts Service, Columbus, OH 43210.
(11) Molecular Design Ltd., San Leandro, CA 94577.
(12) Bentley, J., Burroughs Wellcome Co., Research Triangle Park, NC 27709, personal communication.
(13) Hall, S. R. The STAR File: a new format for electronic data transfer and archiving. *J. Chem. Inf. Comput. Sci.* **1991**, *30* (2), 326–333.
(14) Hall, S. R.; Allen, F. H.; Brown, I. D. The Crystallographic Information File (CIF): a New Standard Archive File for Crystallography. *Acta Crystallogr.* **1991**, *A47*, 655–685.
(15) Rew, R.; Davis, G. NetCDF: An Interface for Scientific Data Access. *IEEE Comput. Graphics Appl.* **1990**, *July*, 76–82.