2nd International Electronic Conference on Synthetic Organic Chemistry (ECSOC-2), http://www.mdpi.org/ecsoc/, September 1-30, 1998

[e0002]

Recommendation on Electronic Publication of Spectroscopic Data

Jun Xu

Boehringer Ingelheim Pharmceuticals, Inc. R&D Center 900 Ridgebury Road Ridgefield, CT 06877-0368, USA

Abstract: The article recommends JCAMP-DX format for spectroscopic data publication, if the standard for a particular type of spectrum is defined in JCAMP-DX protocol. For those spectra the formats are not available from JCAMP-DX, Galactic's SPC format is recommended. The MDL's MOL file or SD file format is suggested for the electronic publication of the spectra-related structures. The NMR assignment formats are also defined within the recommended MOL file format.

Modern web based electronic publication enables chemists to share their research results and data easily, fast and conveniently. Because HTML is a standard language for two-dimensional information representation, the documents composed by HTML are platform-independent. Graphic or image information can be easily posted and shared on the Internet as long as such information is in an acceptable format, such as GIF format.

Spectroscopic data are special types of graphic data. Spectroscopic data can be prepared in GIF format, but chemists really want to interpret and investigate the meanings contained in the graphics interactively. Therefore, standards for spectroscopic data have been developing by IUPAC (International Union of Pure and Applied Chemistry).

The main considerations of developing the standards for the spectroscopic data are:

1. There are many different types spectroscopic experiments, such as NMR, IR, MS, UV, GC/HPLC, EPR etc.

2. For each type of spectroscopic experiment, there are many different native formats from instrument vendors and spectral processing software vendors. Even each hardware or software vendor can produce multiple formats for one type of spectroscopic experiment, because the hardware and software are upgraded from time to time.

3. Most of spectroscopic data are related to chemical structures. And the assignments from spectral peaks to the structures or the atoms/bonds of the structures are very important.

4. The data should be stored in ASCII format. This will guarantee the data file to be platform-independent. Although, IUPAC does release binary format standards.

The above features of spectroscopic data make the data format standardization more difficult than the chemical structure format standardization. It has been noted that many chemical data formats have been developed in the past twenty years. Usually, the dominating formats are the one having larger user population, not necessarily the best defined.

IUPAC has a special committee working on printed and electronic publications for spectroscopic data standards, these standards are known as JCAMP-DX. Currently, the reliable version of JCAMP-DX is 4.24. JCAMP-DX formats are accepted by major spectral processing software.

So far, JCAMP consists of following protocols:

1. JCAMP-DX for Mass Spectrometry[1]
2. JCAMP-DX for NMR[2]
3. JCAMP-DX for IR[3]
4. JCAMP-CS: A Standard Exchange Format for Chemical Structure[4]

The PDF-formatted documents of these protocols can be found from:

http://www.isas-dortmund.de/projects/jcamp/protocol.html

The main spectroscopic instrument or software vendors accept most of these protocols, the spectral processing software installed the newer instruments should have an option to export their spectral data in JCAMP-DX formats. Even if an instrument cannot export spectral data in JCAMP-DX format, from Internet you may get programs to capture the data files via the serial port (RS232), then display or export to JCAMP-DX formats. The detail information about these programs can be found from web page:

http://wwwchem.uwimona.edu.jm:1104/software/jcampdx.html

Therefore, we recommend JCAMP-DX ASCII format for the electronic publication of spectroscopic data. For those spectra the formats are not available from JCAMP-DX, such as UV-VIS, NIR, LC, GC, HPLC, DAS, CE, Fluorescence and Raman, we recommend Galactic's SPC format (http://www.galactic.com/galactic/_gala/prods.htm) if the vendors' software provides such data exporting option.

Chemical structures related to spectra are not recommended to be published in JCAMP-CS format. The main reason is that this format is not very popular. Most of commercial or academic chemical structure information system accept MDL's SD file format or MOL file format (basically, SD file may contain more than one structure, MOL file contains one structure). The detail information about SD file or MOL file format can be found from http://www.mdli.com/prod/helper.html.

Peak assignments are considered as the parts of structural information. MDL's MOL file provides special format to store customer defined information. Recommended format for H^1 NMR assignments are described with an example as follows:

```
Unsaturated Alicyclics
-ISIS- 04249808562D
H-NMR
7 7 0 0 0 0 0 0 0 0999 V2000
0.9541 -0.3583 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
1.8495 -0.3583 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
2.3752 0.1694 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
2.2375 1.0292 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
1.4250 1.4458 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
0.5833 1.0583 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
0.4208 0.2000 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
1 2 2 0 0 0 0
2 3 1 0 0 0 0
3 4 2 0 0 0 0
```

```
4 5 1 0 0 0 0
5 6 1 0 0 0 0
7 1 1 0 0 0 0
6 7 2 0 0 0 0
M END
> <HNMR>
SHIFT 4
1,2 6.50
3,7 6.09
4,6 5.26
5 2.22
COUPLING 8
5-5 2 -13.0
6-7 3 8.9
1-7 3 5.5
1-2 3 11.2
5-6 3 6.7
1-6,2-6 4,5 1.5
2-7 4 0.8
3-7 5 -0.6
$$$$
```

Figure 1. Example for H^1 NMR assignment data format

In the above example, the first three lines are optional. The lines before " `M END` " are standard MDL's MOL file format for the structure shown in Figure 2.

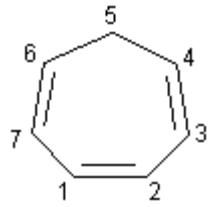

Figure 2. The structure represented in Figure 1

"`> <HNMR>`" indicates the beginning of H^1 NMR assignment data. If the assignment data set is for C^{13} NMR, the line should be "`> <C13NMR>`". This convention applies to other isotopes.

" `SHIFT 4`" starts chemical shift data, the number represents the lines of shift assignments. In this case, it is four. Be aware, the number of the assignment lines should agree with the actual lines.

" `COUPLING 8`" begins the spin coupling data assignment. Each spin coupling assignment line contains three sections: coupling path, coupling topological distance, and coupling constant. The path is represented in "from-atom" and "to-atom", the actual path is the shortest topological path between "from-atom" and "to-atom". Occasionally, one line may contains multiple paths, paths are delimited by ",". When a line contains more than one path, it contains more than coupling topological distance too. This situation happens when

the couplings are weak and overlapped, which cannot be distinguished. For example:" `1-6,2-6 4,5 1.5`".

References

1. Lampen, P., Hillig, H., Davies, A. N., and Linscheid, M., Appl. Spec. v48 (12), 1545-1552, 1994.
2. Davies, A.N. and Lampen, P., Appl. Spec. v47 (8), 1093-1099, 1993.
3. McDonald, R.S. and Wilks, P.A., Jr., Appl. Spec. v42 (1), 151-162, 1988.
4. Gasteiger, J., Hendriks, B.M.P, Hoever, P., Jochium, C. and Somberg, H., Appl. Spec. v45 (1), 4-11, 1991.

Comments

During 1-30 September 1998, all comments on this poster should be sent by e-mail to ecsoc@listserv.arizona.edu with e0002 as the message subject of your e-mail. After the conference, please send all the comments and reprints requests to the author.
