Guide for the contribution of spectra to MACE and the .mace-format

S. Schulz, August 3, 2022

Contribution to MACE is very welcome. The database is conceptually an add-on database with spectra not present in widely distributed mass spectral libraries.

The MACE database consists of EI-MS spectra obtained at 70 eV by GC/MS. Only spectra not present in the the NIST 17 databases [1] should be contributed. The contributed spectra should be preferably from compounds proven by synthesis, isolation or other methods that unequivocally confirm their structure, as well as their derivatives. The spectra will be checked for consistency before including them into the database to have a curated compound list, not filled with already known compounds.

The MACE library is provided as a simple text file using the NIST data structure [2]. The user can incorporate the data into their local databases. Unfortunately, the filename extension of NIST data files is .msp. This can lead to some difficulties in Windows that uses this filename extension for some internal files. We therefore changed the extension to .mace to make the file distinct. Nevertheless, being a text file, the extension is not important, and others can be used as well.

A sample entry is shown below. The spectrum starts always with the name and is ended by an empty line. Fields are always on one line. At the end of the text after **Num Peaks**: the actual data input follows. These data can have different formats, for details see the NIST manual [2]. It is important to exactly follow this document structure. Each field with a : has a single line. This is also true for the **Comments** line. It is one long line. Line breaks in the comment line example in this document are only here for readability.

```
Name:  Dehydrojasmone
Synonyms:  4-Methylene-5-((Z)-2-penten-1-yl)-2-cyclopenten-1-one
Synonyms:  (Z)-4-Methylene-5-(pent-2-en-1-yl)cyclopent-2-en-1-one
CAS: 2622964-68-3
InChIKey:  PRLXPOSTWLLTTJ-PLNGDYQASA-N
Formula:  C11H14O
MW: 162
ExactMass:  162.1045
RI: 1300
Comments:  Contributor=P.Stamm;_S.Schulz;_TU-Braunschweig Spectrum_id=SC-37
   Phase=DB5-MS Reference=https://doi.org/10.1021/acs.joc.1c00145
   Smiles=O=C(C=C1)C(C/C=CCC)C1=C Mode=EI-quadrupole;_Agilent_MSD License=CC_BY-SA
   Compound_class=jasmone_derivative floral scent of Araceae
   Source=Scan_1137_SD467P.D
Num Peaks:  90
37 3; 38 17; 39 176; 40 30; 41 198; 42 12; 43 7; 45 1; 49 2; 50 39; 51 113; 52
96; 53 93; 54 14; 55 136; 56 6; 57 5; 58 4; 61 3; 62 15; 63 53; 64 18; 65 125;
66 102; 67 71; 68 44; 69 75; 70 5; 72 1; 73 1; 74 8; 75 7; 76 8; 77 238; 78 99;
79 154; 80 35; 81 25; 82 10; 83 2; 86 1; 87 3; 88 1; 89 19; 90 8; 91 363; 92 86;
93 23; 94 999; 95 84; 96 6; 98 1; 101 1; 102 11; 103 72; 104 24; 105 255; 106
```

```
48; 107 113; 108 45; 109 4; 115 86; 116 16; 117 42; 118 12; 119 175; 120 65; 121
19; 122 2; 127 7; 128 20; 129 25; 130 4; 131 25; 132 20; 133 503; 134 98; 135
9; 141 1; 143 3; 144 7; 145 10; 146 4; 147 99; 148 11; 149 1; 161 16; 162 172;
163 22; 164 2;
```

The different fields will be explained now. The should be filled by the contributor if possible.

Name: Always begins the entry for a spectrum. Please use the trivial name here, if one is available for the compound. **Name** is a mandatory field.

Synonyms: Add proper chemical name here (e.g., IUPAC name), when a trivial name is used in Name. Several synonyms can be included, each one on a separate line. Please remove any stereo-descriptors that are associated with a single enantiomer because MS cannot differentiate enantiomers. The asterisk convention should be used instead if it becomes necessary to assign the relative configuration of compounds with more than one stereogenic center, according to IUPAC nomenclature, e. g. (3*R**,5*S**) [3]. This procedure will avoid claims of enantiomers being identified via mass spectral matches by inexperienced users.

InChIKey: This structure description key is used in many web and other applications. MSSearch allows direct connection to PubChem by clicking this key. PubChem provides various types of information about a compound. InChIKey keys can be generated with different chemical drawing programs, e.g. with ChemDraw using the "Copy As" tool. Additionally, free structure drawing software such as ChemSketch or online tools such as http://www.cheminfo.org/Chemistry/Cheminformatics/Generate$_I$nChI/index.html can be used to generate InChIKey codes.

CAS: CAS number if known. The number can be easily obtained by importing the Smiles code into SciFinder and performing a structure search, or simply drawing the structure in Scifinder and searching. Some double bond or other stereochemical formatting may be necessary in SciFinder before searching, to ensure that the search will bring up the CAS number for the correct stereoisomer.

Formula, MW and ExactMass: Add the respective data here. The "Molecular Weight" field must be filled in, otherwise mass spectral search programs will likely not work. Be aware that MW is the nominal mass in MS, e. g. $C_{35}H_{72}$, has an exact mass of 492.5634, but a nominal mass of 492. The exact mass is the monoisotopic molecular mass, not the molecular weight more commonly used in Chemistry.

RI: Retention index of the compound. Semi-standard non-polar retention indices are preferred, but other phases are also welcome.

Comments: **Only one line!** The Comments field includes tags. These tags can improve usability and a number of them are defined in MACE. None of them are mandatory. These tags have names that are followed by a "=" and the following term must not have spaces. The tags are separated with empty spaces. Any other text outside the tags will occur in the output under **Comments**, the rest in appropriate fields. The data are only clearly visible when the database is configured to show these tags, otherwise the data will be found in the comment line output. Spaces in tag entries are not allowed. Words in tags should be connected by a connecting character, such as an underscore, as shown in the example above. The following tags are used.

Spectrum= This entry is an internal number given by MACE. It should be left blank and will be assigned when the compound/spectrum is added to MACE.

Contributor= Submitters and laboratory contributing the spectrum

Phase= Specifies the GC column(s) used for retention index (RI) calculation(s). We usually use semi-standard non-polar phases such as HP5-MS, but if RIs are measured on more than one column, these data can also be provided.

Mode = Specifies what type of mass analyzer was used (e.g. quadrupole, sector field, ion trap or time-of-flight), and the instrument manufacturer. Spectra may differ according to the type of mass analyzer used.

Reference= The DOI reference to the publication of the original spectrum, preferably from a known authentic standard. This is helpful for users because it allows easy access to the original publications and makes citation in publications easier.

Smiles= A character code for drawing structures easily. Copying this code into drawing programs or databases such as Scifinder will automatically generate the structure. The Smiles code can be obtained as described above for the InChIKey code.

License= Should not be altered. It allows usage by individuals, but not the use of the spectrum in a commercial product without a license.

Class= Specifies the compound class. This can possibly be vague, but defining the class can be helpful in some cases, for example, terpenoids, fatty acids, steroids, etc.

Source= Local file name and scan number of the data from the submitter. This entry allows the spectrum to be traced back to its origin.

The data are better when all tags are filled. Nevertheless, one can contribute data to MACE also with some fields unfilled.

Num Peaks: Number of peaks to follow. Usually already in the data when mass spectrum is exported into the text file. Absolutely necessary for correct function of the file.

Then the actual data are following, which do not need to be on one line. They can also have slightly different format [3]. The entry ends with an empty line.

Copy all your compounds into one document. Every compound entry begins with Name, preceded by at least one blank line. Save as .mace file or other text format. Send to **mace@tu-braunschweig.de** via email. A sample file including an unfilled data template can be downloaded from our website.

All the compounds will be added to the MACE library for public use in the open access data repository of TU Braunschweig after a quality check. MACE: Mass Spectra for Chemical Ecology.

Generating .mace files via NIST MSSearch

The data for submitting a mass spectrum can be obtained via the Librarian tool of MSSearch in the NIST program. Select a peak from a GC/MS-run. A library search is then performed using the NIST program. A good spectrum should be selected, maybe performing baseline

subtraction first. A right click on the spectrum opens MSSearch. Select the `Librarian` tool. Go to the small export icon and press export. Create a file in .msp format with the name of the compound as title. Select overwrite.

Now open the .msp file in an editor that allows long lines, e. g. notepad. The file locks like this:

```
Name:  Scan 33763 (261.393 min):  AMN037.D...
DB: 15
Num Peaks:  224
38 1; 39 33; 40 6; 41 200; 42 38;
```

Replace the scan... part in the field **Name** with the name of the compound. You can save the scan data for later use in the *Source* tag. Then replace the `DB:` line with the template obtained from the sample file and fill out the respective information.

Copy all single spectrum files into one document and rename it including .mace extension.

We are sure other programs allow similar transformations as well. Send us routines, so they can be added to this document. JCAMP-DX files and other formats can be converted into NIST format by the Lib2Nist program [4]. Therefore, one can export a spectrum via JCAMP-DX and convert it via Lib2Nist. MACE is described in detail in an article by us that also describes some other resources to obtain suitable MS data from various MS systems [5].

References and Notes

[1] We opted not to use the newest library versions of NIST, because they are likely less widespread in labs than older versions. We omitted Wiley because of the many bad quality spectra and some serious errors.

[2] *NIST MS Search User Guide*, section "NIST Text Format of Individual Spectra": http://chemdata.nist.gov/mass-spc/ms-search/docs/Ver20Man_11.pdf

[3] https://www.acdlabs.com/iupac/nomenclature/93/r93$_6$35.htm according to IUPAC, Commission on Nomenclature of Organic Chemistry. A Guide to IUPAC Nomenclature of Organic Compounds (Recommendations 1993), 1993, Blackwell Scientific publications

[4] https://chemdata.nist.gov/mass-spc/ms-search/Library_conversion_tool.html

[5] S. Schulz, A. Möllerke
MACE - An Open Access Data Repository of Mass Spectra for Chemical Ecology, *J. chem. Ecol.*, online. doi: 10.1007/s10886-022-01364-4]

Stefan Schulz, Anton Möllerke August 3, 2022