Calibrated Mass Spectrum (CMS) Format Specification

Revision 0.1

CMS format is basically an extension of the MSP and MSL formats used for GC/MS data but with additional fields defined.  Fields are defined in the same way, one field per line, with a case insensitive field identifier string at the start of the line.  In principle, all of the existing fields for MSP and MSL formats are retained.  But because I don’t have access to actual data which requires these other fields, only the fields listed in this description have been tested. The format for ion mass/signal pairs is also retained.  It should be possible for the CMS parser to read MSP and MSL files, in which case the additional fields in CMS format are either undefined and ignored or given default values.  The parser also compares the value of the "NUM PEAKS:" field with the actual count of mass/signal pairs and generates an exception if they are not equal.

A CMS file consists of an arbitrary number of either SCAN records of a mass spec measurement of an unknown or NAME records for library components all concatenated together to make a CMS data file.  SCAN records begin with the field identifier “SCAN:” and library component records begin with “NAME:”.  It is an error for a CMS file to contain both SCAN and NAME records.  Once the type of the first record is determined, all following records must be the same type.

The value for the SCAN and NAME field is a case insensitive and unique (only within this CMS file) string.  It is an error for the SCAN or NAME value to be repeated in the same CMS file.  This is to detect when a library component is duplicated, which will cause a singular matrix later.  For SCAN records, the field value might be a string representation of an integer scan number but just needs to be different for each scan record.

Other fields may be arranged in any order between the NAME/SCAN field and the “NUM PEAKS:” field.

Note: When parsing a string representation of a number, any valid string number format is acceptable including scientific.  Exceptions are thrown if the string represents an invalid number for the type of number required.  So, for example, 10.0e1 would be a valid string when an int is required (100), but any number having a non-zero fractional part would not (100.0001 for example).

Within the same CMS file, field values need to be present for each NAME or SCAN record. If not present, an illegal value may be used by default which means the value is undefined and not used for any calculations or written when the record is exported.

Fields for NAME and SCAN files:

**SOURCEP**: The sample source pressure, a Number.  A mass spectrometer is typically calibrated to measure the partial pressure of a component which is present in the spectrometer sample inlet system, often called the “sample source”. This means that the signal is proportional (hopefully linear) with the sample source pressure. Often, however, it is useful to know the relative amounts of the components present, such as the component mol fraction. To make the conversion from component partial pressure to component mol fraction, it is necessary to know the total pressure in the sample source.

In the case of a gc-ms, the sample source pressure is typically not specified. However, for mass spectrometers used for gas analysis, the source pressure is always known. For my mass spec measurements, I have a pressure transducer upstream of the molecular leak and a typical sample source pressure measurement is 1 mbar.

The reason the **SOURCEP** field is needed is that the signal intensity varies with source pressure even if the component relative composition (aka mol fraction) is held constant.  Usually instruments are set up so that this variation is linear, but at higher source pressures, the effect of source pressure on signal intensity is non-linear.  Usually, source pressure is held constant.  But, a batch analysis mass spec that is used for gas analysis will have the source pressure slowly decreasing with time as the sample is consumed.  But, so long as the effect of source pressure on signal is accounted for, the analysis can be very quantitative.

**SPUNITS**: The units used for reporting the sample source pressure, a String.  This field is used primarily to check that the library component calibration information is compatible with the measurement scan information.  If the SPUNITS are not the same, it may be possible to re-scale to make them the same.  So if SPUNITS == "torr" for a library file and SPUNITS == "mbar" for a measurement file, the SOURCEP value for the library components is multiplied internally by 1.3332237 to convert the SOURCEP value to mbar units and quantitative decomposition can proceed.

**SIGUNITS**: The units used for reporting the ion signal value in the (mass/signal) ion pairs, a String.  In my case, SIGUNITS == "amp".  In other cases, SIGUNITS might be "cps" for Counts Per Second or SIGUNITS might be "arb" for Arbitrary.  The default is "arb".

Note: When SIGUNITS == "arb" for either the library or measurement files, the decomposition results are qualitative only.  Quantitative results are possible only when the library and measurement files have the same SIGUNITS value.  In theory, it might be possible to convert signal values for different SIGUNITS, say for example convert "cps" units into "amp" units.  But even though this conversion exists, the differences between the instruments used to make the measurements are probably so great that even after converting signal units, the results would not be quantitative.

The above field additions are sufficient to produce a quantitative result.  The result is "partial pressure of this library component at the source" in SPUNITS.  These partial pressure results divided by the measurement SOURCEP give relative concentration, for example mol fraction.  Without these fields, decomposition yields qualitative results which are none the less still valuable for interpreting overlapping GC peaks.

Other new fields for NAME and SCAN files:

These fields are not necessary, but do carry useful information.

**RESCALE**: If present, this value represents the signal for the major peak in the mass spectrum. The ion signals will be rescaled so that the major peak has this signal value and the ratio between the major peak and the other peaks stays the same. This allows library spectra to be easily adjusted for changes in instrument sensitivity. Just enter the new measurement of the major peak as a **RESCALE** field and the other peaks in the cracking pattern will be adjusted. **RESCALE** is also useful when importing published cracking patterns.

**TSTAMP**: A time stamp for this scan, a String.  Date and time the measurement was made. I'm not sure of the exact format, but the TSTAMP field is there to document when the measurement was made.  Subsequent scans take the same value by default so actual date and time of subsequent scans can be computed by adding ETIMES.  For NAME files, each component would have a different TSTAMP field.

**ETIMES**: Elapsed TIME in Seconds, a number.  The elapsed time in seconds since the start of the first scan in the file.  Similar to retention time in OpenChrom CVS format, ETIMES is used to put scans into time increasing order and does not have meaning for NAME files.  For SCAN files, the ETIMES value for the first scan is 0.0.  For subsequent scans, ETIMES increases just like retention time.  For library components, ETIMES is arbitrary of left undefined.

**EENERGYV**: Electron energy in Volts, a Number.  Since the component cracking pattern can change depending on the electron energy setting in the source, this setting is recorded to aid in determining whether or not the instrument settings used to make the component calibration library measurement and the unknown measurement are compatible enough for quantitative analysis.

**IENERGYV**: Ion energy in Volts, a Number.  Component sensitivities can change depending on the ion energy setting in the source, this setting is recorded to aid in determining whether or not the instrument settings used to make the component calibration library measurement and the unknown measurement are compatible enough for quantitative analysis.

**MASSR**: MASS Resolution, a Number.  This is the instrument mass resolution expressed as (peak center)/(peak width) where both the peak and center are in m/e units so MASSR is dimensionless.  For magnetic sector mass spectrometers, resolution is often adjustable and stays roughly constant over the measured m/e range.  Instrument resolution is needed to determine whether or not library component ions have the "same" mass as an unknown spectrum ion.  I'm not sure how quadrupole mass spectrometers behave with respect to peak width vs. m/e.

**INAME**: Instrument name, a String.  The name of the instrument that made the measurement, "VG 14-80" in my case.