The AnIML Mass Spectrometry Technique Definition, an Example Including AnIML Hybrids and AnIML Evolution

Mark Bean[1]; Burkhard Schaefer[2]; Maren Fiege[3]; Gary Kramer[4]

[1]GlaxoSmithKline, Collegeville, PA; [2]BSSN Software, Mainz, Germany; [3]Waters GmbH, Frechen, Germany; [4]National Institute of Standards and Technology, Gaithersburg, MD

Introduction

The AnIML MS standard is in draft form and needs help from the MS Community.

We seek a few representatives from instrument vendors, government regulatory agencies, and the scientific community to aid in reviewing the draft. If interested, leave a card in the envelope below, add you name to the list, or send an e-mail.

The AnIML MS Technique Uses mzML Terminology

A major challenge in creating a data standard for mass spectrometry is standardizing and documenting the ever-expanding terminology. Fortunately, we can stand on the shoulders of prior efforts:

- The AnIML MS Technique Definition will shadow the proteomics standard mzML in the interest of simplification and possible future coalescence of the standards:
  - mzML[1] adheres to IUPAC terminology
  - mzML ontology is maintained via a mailing list[2]
- K.K. Murray's MS Terms Wiki is also useful[3]
- The MS Technique, like all AnIML Techniques, can be extended by instrument vendors or individuals at any time. If the extension AnIML Technique Definition Document (.atdd) is made available, it can be used to validate the MS data files.
- The AnIML MS Technique is part of the balloted ASTM standard and must be reviewed periodically, at which time new terms may be made official.
- We propose to adopt mzML terminology using their mailing list when changes are needed. The differences are that AnIML does not use accession numbers and requires title case for terms (mzML uses lower case).

Mass Spectra

Three Flavors of Data

- x and y axes are separate <Series>
- Each <Series> is recorded as either:
  - <EncodedValueSet> = base64Binary
    - Perfect fidelity with original computer data
    - Encoders and decoders available for all platforms
    - 7-bit binary encoding as text
    - 33% larger than binary

```
<Result name="Spectrum">
    <SeriesSet name="Spectrum" length="123">
        <Series name="m/z" dependency="independent" seriesType="Float32">
            <EncodedValueSet>AACIQ2ZmjEMAgIxDM7ONQ5oZjkMzM5RDZmaUQ2bmlEMA
            AJVDAICVQwAAmUMAgJpDM7O….IARRJqZEURmphFEZuYRRM2MEkQzcxNE
            AIATRJoZFkSamRdE</EncodedValueSet>
            <Unit label="m/z" quantity="Ratio"/>
        </Series>
    <SeriesSet >
<Result >
```

- <IndividualValueSet> (e.g. <Float32>234.1</Float32>)
  - Bypasses base64binary conversions
  - Maintains "scientific precision"
- <AutoincrementedValueSet>
  - <StartValue> and <Increment>
  - Useful for regularly-spaced data (e.g. chromatogram time-axis)

Methods and Experimental Metadata

- <Method> contains metadata known before the run
- <Result> contains metadata recorded during the run (useful in data-dependent experiments)
- Name-value parameter pairs are recorded as categorized <Parameter> entries, with term names defined by the MS Technique Definition Document and identical to mzML but using Title Case.
- Tabular entries encoded as <IndividualValueSet> (see above)

```
<Series name="Event Time" seriesType="Float32" >
    <IndividualValueSet>
        <Float32>0.0</Float32>
        <Float32>1.5</Float32>
        <Float32>1.9</Float32>
        <Float32>2.0</Float32>
    </IndividualValueSet>
    <Unit label="min"/>
</Series>
```

AnIML Hybrids

LC-MS

```
<ExperimentStepSet >
 <ExperimentStep name="Timeline" >
  <Technique name="Chromatography" uri=http://techniques.animl.org/current/chromatography-technique.atdd/>
  <Result name="Separation Monitoring" >
   <SeriesSet name="Separation Monitoring">
    <Series name="Time" seriesType="Float" seriesID="c01" dependency="independent">
     <Unit label="min" quantity="Time"/>
     <IndividualValueSet>
      <Float32>0.0</Float32>
     </IndividualValueSet>
    </Series>
   </SeriesSet>
  </Result>
  <ExperimentStepSet >
   <ExperimentStep name="Mass Spectrum" >
    <Technique name="Mass Spectrum" uri="http://schemas.animl.org/current/ms-technique.xsd" />
    <Infrastructure>
     <ParentDataPointReference seriesID="c01"startIndex="0" >
    </Infrastructure>
    <Result name="Spectrum" >
     <SeriesSet name=" Spectrum">
      <Series name="Mass" seriesType="Float" seriesID="ms01" dependency="independent">
       <EncodedValueSet> AABQwgA ….. wC+vUcAyL</ EncodedValueSet >
       <Unit label="m/z"/>
      </Series>
     </SeriesSet>
    </Result>
   </ExperimentStep>
  </ExperimentStepSet>
  <ExperimentStepSet >
   <ExperimentStep name="MS Time Trace" >
    <Technique name="Mass Chromatogram" uri="http://techniques.animl.org/current/mschromatogram-technique.atdd" />
    <Infrastructure>
     <ParentDataPointReference seriesID="c01"startIndex="0" >
    </Infrastructure>
    <Result name="Time Trace" >
     <SeriesSet name="TIC">
      <Series name="Time" seriesType="Float" seriesID="mt01" dependency="independent">
       <EncodedValueSet>AJB5HADwb … vCNHAGgkRoJUcAICd</ EncodedValueSet >
       <Unit label="min" quantity="Time"/>
      </Series>
      <Series name="Intensity" seriesType="Float" seriesID="mi01" dependency="dependent">
       <EncodedValueSet>MgSR … ALQ6RwB</ EncodedValueSet >
       <Unit label="counts"/>
      </Series>
     </SeriesSet>
    </Result>
   </ExperimentStep>
  </ExperimentStepSet>
 </ExperimentStep>
</ExperimentStepSet >
```

reference

reference

References lock timelines between techniques

Techniques are nested or not, depending on Sample inheritance

Mass Spectrum and Mass Chromatogram are separate Techniques at the moment

Append processing Technique information (methods or results):

- Smoothing
- Baseline Subtraction
- Peak Finding
- Spectral Summation
- Quantitation

AnIML Evolution

So how does AnIML evolve, covering ever-new MS technologies, and still force conformance to a standard? It is ever a delicate balance between flexibility and confusing variations on a standard (a problem encountered with JCAMP-DX)

- Technique extensions can be created at any time and can be used to validate AnIML data files. Extensions are useful for samples, methods and results, but do not allow extension of Allowed Values (e.g. MS source types).
- If a new MS source type is needed in Allowed Values, there are several options:
  - Enter a new Allowed Value, but recognize that there may be a validation error to be ignored.
  - Propose new Allowed Values to the E13.15 Committee at the time of its annual review.
  - Applications can simply choose to display the unexpected value or ignore the unvalidated information entirely.

The twin AnIML governing schema are intended to be immutable, requiring major review and justification before a new version is created (akin to going to XML 2.0).

- New data or post-processing techniques can be appended to an AnIML file (with a new check sum and audit trail entry).
- Selected spectra, chromatograms, or tables can be split off, together with encapsulated metadata, and saved to a database, sent in an e-mail, etc. and still be valid.

URL References

1. mzML - http://psidev.info/mzml
2. mzML e-mail list https://lists.sourceforge.net/lists/listinfo/psidev-pi-dev
3. K. K. Murray's MS Term Wiki - http://mass-spec.lsu.edu/msterms/index.php/Main_Page

Questions ?

http://www.animl.org

Mark.F.Bean@gsk.com