Implementing AnIML 1.0

Crosswalking AnIML from Legacy Data Formats

Stuart Chalk
University of North Florida

Overview

- A place for everything…
- Metadata mapping
- Required data
- Audit trail
- Best practices
- Keep the original

AnIML Format

- Sample
  - Sample Data
  - Substance Data
- ExperimentStep
  - Method Parameters
  - Instrument Parameters
  - Result Data
  - Peak Table (if appropriate)

- Framework for information
- Small amount of required information
- Many “pigeonholes” for information if you have it

JCAMP Format

- Header
- Data

```
##TITLE= fixinc1.jdx
##JCAMP-DX= 4.24 $$ DIGILAB
##DATA TYPE= INFRARED SPECTRUM
##ORIGIN= UWS, Nepean Campus, Australia
##OWNER= public domain
##$URL= http://wwwchem.uwimona.edu.jm:1104/spectra/testdata/index.html
##DATE= 93/03/08
##TIME= 11:01:12
##CREATED= Mon Mar 08 11:01:12 1993
##SPECTROMETER/DATA SYSTEM= Digilab Data system=3207
##RESOLUTION= 2.
##XUNITS= 1/CM
##YUNITS= TRANSMITTANCE
##XFACTOR= 9.64405731e-01
##YFACTOR= 4.768371582e-07
##FIRSTX= 3.99263973e+02
##LASTX= 4.00131938e+03
##NPOINTS= 3736
##FIRSTY= 1.128905654e+02
##XYDATA= (X++(Y..Y))
414 236748675 223618025 205188125 191890525 182832000 176101675 170927175
421 167160375 164879175 162388125 159973600 158139850 156466525 155060050
428 153930675 152825250 151700050 150190125 149084200 149805825 152784900
435 155833825 155405850 154709625 154273700 154565950 155179175 155376400
...
4125 154868500 154563925 154276375 154006900 153665250 153294400 152911400
4132 152539775 152144575 151716675 151235100 150666025 150180525 149697775
4139 149173400 148704800 148234400 147790475 147365825 147016075 146719525
4146 146462625 146290125 146170425 146072575
##END=
```

AnDI Format

- Header

TABLE 2 Sample-Description Information Class

Date Element Name	Datatype	Category	Required
sample-ID-comments	string	C5	. . .
sample-ID	string	C1	. . .
sample-name	string	C1	. . .
sample-type	string	C1	. . .
sample-injection-volume	floating-point	C3	. . .
sample-amount	floating-point	C3	. . .

TABLE 3 Detection-Method Information Class

Data Element Name	Datatype	Category	Required
detection-method-table-name	string	C1	. . .
detection-method-comments	string	C1	. . .
detection-method-name	string	C1	. . .
detector-name	string	C1	. . .
detector-maximum-value	floating-point	C1	M1
detector-minimum-value	floating-point	C1	M1
detector-unit	string	C1	M1

- Data

TABLE 4 Raw-Data Information Class

Data Element Name	Datatype	Category	Required
point-number	dimension	C1	M1
raw-data-table-name	string	C1	. . .
retention-unit	string	C1	M12
actual-run-length	floating-point	C1	M12
actual-sampling-interval	floating-point	C1	M12
actual-delay-time	floating-point	C1	M12
ordinate-values	float-array	C1	M1
uniform-sampling-flags	boolean	C1	M1
raw-data-retention	float-array	C1	M1
autosampler-position	string	C1	. . .

http://www.astm.org/Standards/E1947.htm

GAML Format

- Header
- Data

```
<?xml version="1.0" encoding="UTF-8"?>
<GAML version="1.20" name="3DLIVE">
    <integrity algorithm="SHA1">33bca32b653a2dcce6dbb794ecb224ecd6232aa3</integrity>
    <parameter name="component_name" label="Component name" group="GAML Generation">GAMLIO</parameter>
    <parameter name="component_version" label="Component version" group="GAML Generation">9.1.3.6</parameter>
    <parameter name="converter_name" label="Converter name" group="Data Conversion">SLM_BWM2</parameter>
    <parameter name="converter_description" label="Converter description" group="Data Conversion">Thermo Spectronic SLM-
    <parameter name="converter_version" label="Converter version" group="Data Conversion">6.1.0.1</parameter>
    <parameter name="conversion_date" label="Conversion date" group="Data Conversion">2010-09-16 11:10:50</parameter>
    <parameter name="converter_input_source" label="Converter input source" group="Data Conversion">D:\temp\GAMLORG\Ther
    <parameter name="converter_output_file" label="Converter output file name" group="Data Conversion">C:\Temp\GCnvtstep
    <experiment name="3D Ex/Em Matrix, scan rate 50nm/sec (4sec/trace), 40 traces/200sec ">
        <collectdate>1993-12-16T22:10:18Z</collectdate>
        <parameter name="Title" label="Data record title" group="Header">3D Ex/Em Matrix, scan rate 50nm/sec (4sec/trace
        <parameter name="DataFileSignature" label="Data file signature" group="Header">SLM Data File, Format 1.5 </param
        <parameter name="Format" label="Data record format" group="Header">6</parameter>
        <parameter name="Axes" label="Number of X-axes in this data record" group="Header">41</parameter>
        <parameter name="Comment" label="Data record comments" group="Header">NULL</parameter>
        <trace name="3D Ex/Em Matrix, scan rate 50nm/sec (4sec/trace), 40 traces/200sec " technique="FLUOR">
            <parameter name="ZaxisType" label="Z-axis type" group="ZAxis">0x3 = Excitation Monochromator</parameter>
            <parameter name="ZaxisUnits" label="Z-axis units" group="ZAxis">0x2 = Nanometers (nm)</parameter>
            <parameter name="ZaxisLL" label="Z-axis lower limit" group="ZAxis">200.000000</parameter>
            <parameter name="ZaxisUL" label="Z-axis upper limit" group="ZAxis">400.000000</parameter>
            <parameter name="LabelZ" label="Z-axis Label" group="ZAxis">Excitation wavelength</parameter>
            <parameter name="PlotInfo" label="Z-axis plotting annotation" group="ZAxis">NULL</parameter>
            <coordinates label="Excitation wavelength" units="NANOMETERS" valueorder="EVEN"> [2 lines]
            <Xdata label="Emission wavelength" units="NANOMETERS" valueorder="EVEN">
                <parameter name="DataType" label="Type of data stored" group="XAxis">0x01 = Normal Data</parameter>
                <parameter name="Incs" label="# of different X-axis increments" group="XAxis">1</parameter>
                <parameter name="Channels" label="# of Y-axis associated with this X-axis" group="XAxis">1</parameter>
                <parameter name="Points" label="# of data points along the X-axis" group="XAxis">201</parameter>
                <parameter name="XFlag" label="X-axis values included flag" group="XAxis">0x0</parameter>
                <parameter name="Zaxis" label="Z-axis values for this X-axis" group="XAxis">200.000000</parameter>
                <parameter name="LabelX" label="X-axis Label" group="XAxis">Emission wavelength</parameter>
                <parameter name="XaxisType" label="X-axis type" group="XAxis">0x2 = Time-Dependent Emission Monochromato
                <parameter name="XaxisUnits" label="X-axis units" group="XAxis">0x2 = Nanometers (nm)</parameter>
                <parameter name="Inc" label="Data point increment" group="XAxis">1.000000</parameter>
                <parameter name="XaxisLL" label="X-axis lower limit" group="XAxis">300.000000</parameter>
                <parameter name="XaxisUL" label="X-axis upper limit" group="XAxis">500.000000</parameter>
                <parameter name="Header" label="Data type specific header info" group="XAxis">0</parameter>
                <parameter name="PlotInfo" label="X-axis plotting annotation" group="XAxis">NULL</parameter>
                <values byteorder="INTEL" format="FLOAT32" numvalues="201">AACWQwCAlkMAAJdDAICXQwAAmEMAgJhDAACZQwCAmUMAA
                <Ydata label="Fluorescence" units="UNKNOWN"> [86 lines]
                ...
            </Xdata>
        </trace>
    </experiment>
</GAML>
```

Data Transfer

- Keep all the metadata you have!
  - Find the most appropriate places for data you have
  - Keep the precision of data
- Verify the conversion was done correctly
  - Dates
  - Data values
  - Strings use correct characters (UTF-8)
- Metadata in different formats (i.e. date)
  - European v's American v's UTC
  - Time zones!

Fitting in the Data

- Required metadata not available

```
<AnIML xmlns="urn:org:astm:animl:schema:core:draft:0.37"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="urn:org:astm:animl:schema:core:draft:0.37 http://animl.cvs.sourceforge.net/viewvc/animl/schema/animl-core.xsd"
    version="0.37">
    <SampleSet>
        <Sample name="Test Sample" sampleID="sample1">
            <CategorySet>
                <Category name="Description">
                    <ParameterSet>
                        <Parameter name="Descriptive Name" parameterType="String">
                            <String>Pond water sample from retention pond by the arena parking garage</String>
                        </Parameter>
                    </ParameterSet>
                </Category>
            </CategorySet>
        </Sample>
    </SampleSet>
    <ExperimentStepSet>
        <ExperimentStep name="Analysis" experimentStepID="step1">
            <Method> [14 lines]
            <Result name="Spectrum"> [17 lines]
        </ExperimentStep>
    </ExperimentStepSet>
</AnIML>
            <Result name="Spectrum"> [17 lines]
        </ExperimentStep>
    </ExperimentStepSet>
</AnIML>
        </ExperimentStep>
    </ExperimentStepSet>
</AnIML>
```

Fitting in the Data

- Required metadata not available
  - Inclusion of a metadata element requires addition of other metadata

```
<Category name="Instrument Settings">
    <ParameterSet>
        <Parameter name="Slit Function" parameterType="String">
            <!-- REQUIRED -->
            <String>Constant resolution</String>
        </Parameter>
        <Parameter name="Spectral Slit Width" parameterType="Float32"> [4 lines]
        <Parameter name="Slit Width" parameterType="Float32">
            <!-- REQUIRED -->
            <Float32>2.0</Float32>
            &nm;
        </Parameter>
        <Parameter name="Integration Period" parameterType="Float32"> [4 lines]
        <Parameter name="Degree of Derivatization" parameterType="Int32">
            <!-- REQUIRED -->
            <Int32>0</Int32>
        </Parameter>
        <Parameter name="Derivatization Algorithm Description" [4 lines]
        <Parameter name="Scan Speed" parameterType="Float32"> [4 lines]
        <Parameter name="Scan Response Filter" parameterType="Float32"> [4 lines]
        <Parameter name="Spectral Bandwidth" parameterType="Float32"> [4 lines]
    </ParameterSet>
</Category>
```

Fitting the Data

- Many to one
  - Formatting
  - Consolidation
- One to many
  - Redundancy
  - Where is the best place?
- Incomplete Data
- Inaccurate Data

```
##TITLE=o-Phenanthroline
##JCAMP-DX=4.24
##DATA TYPE=MASS SPECTRUM
##ORIGIN=Japan AIST/NIMC Database- Spectrum MS-IW-3470
##OWNER=NIST Mass Spectrometry Data Center
Collection (C) 2008 copyright by the U.S. Secretary of Commerce
on behalf of the United States of America. All rights reserved.
##CAS REGISTRY NO=66-71-7
##$NIST MASS SPEC NO=230013
##MOLFORM=C12H8N2
##MW=180
##$NIST SOURCE=MSDC
##XUNITS=M/Z
##YUNITS=RELATIVE ABUNDANCE
##XFACTOR=1
##YFACTOR=1
##FIRSTX=26
##LASTX=182
##FIRSTY=10
##MAXX=182
##MINX=26
##MAXY=9999
##MINY=10
##NPOINTS=65
##PEAK TABLE=(XY..XY)
26,10 27,20 28,40 37,20
38,40 39,99 49,20 50,249
51,209 52,70 60,10 61,60
62,139 63,559 64,50 65,10
73,20 74,219 75,299 76,529
77,239 78,30 79,10 85,10
86,40 87,70 88,50 89,90
90,899 91,10 97,10 98,80
99,179 100,209 101,169 102,109
103,50 104,10 111,10 112,10
113,20 114,20 122,10 123,20
124,40 125,379 126,369 127,489
128,119 129,229 130,20 140,10
150,20 151,169 152,599 153,989
154,1279 155,139 156,10 177,40
178,10 179,3039 180,9999 181,1389
182,90
##END=
```

Fitting the Data

- Controlled vocabularies
  - Take free format data and put in an enum field

```
<ParameterBlueprint name="State" parameterType="String" modality="optional" maxOccurs="1">
    <Documentation literatureReferenceID="JCAMP-DX IR">Phase of matter of the sample.</Documentation>
    <AllowedValue>
        <String>solid</String>
    </AllowedValue>
    <AllowedValue>
        <String>amorphous solid</String>
    </AllowedValue>
    <AllowedValue>
        <String>crystalline solid</String>
    </AllowedValue>
    <AllowedValue>
        <String>liquid</String>
    </AllowedValue>
    <AllowedValue>
        <String>liquid crystal</String>
    </AllowedValue>
    <AllowedValue>
        <String>gas</String>
    </AllowedValue>
    <AllowedValue>
        <String>supercritical fluid</String>
    </AllowedValue>
    <AllowedValue>
        <String>colloid</String>
    </AllowedValue>
    <AllowedValue>
        <String>plasma</String>
    </AllowedValue>
    <AllowedValue>
        <String>crystal</String>
    </AllowedValue>
</ParameterBlueprint>
```

Defining a Crosswalk

- A crosswalk is a table that shows equivalent elements (or "fields") in more than one database schema. It maps the elements in one schema to the equivalent elements in another schema.

JCAMP-DX LDR	AnIML Element (s)	Path
TITLE	@name (Serieset)	/AnIML/ExperimentStepSet[1]/ExperimentStep[#]/Result['Spectrum']/SeriesSet[1]/@Name
JCAMP-DX	<add to comments in audittrail>	
DATATYPE	@name	/AnIML/ExperimentStepSet[1]/ExperimentStep[#]/Technique[1]@name
ORIGIN	Name, Affiliation, Phone, Locatior	Multiple
OWNER	<add to dsig>	
XUNITS	Unit	/AnIML/ExperimentStepSet[1]/ExperimentStep[#]/Result['Spectrum']/SeriesSet[1]/Series['Wavenumber']/AutoIncrer
YUNITS	Unit	/AnIML/ExperimentStepSet[1]/ExperimentStep[#]/Result['Spectrum']/SeriesSet[1]/Series['Intensity']/AutoIncrement
XFACTOR	<used to convert values>	
YFACTOR	<used to convert values>	
FIRSTX	StartValue	/AnIML/ExperimentStepSet[1]/ExperimentStep[#]/Result['Spectrum']/SeriesSet[1]/Series['Wavenumber']/AutoIncrer
LASTX	<used as check only>	
NPOINTS	@length	/AnIML/ExperimentStepSet[1]/ExperimentStep[#]/Result['Spectrum']/SeriesSet[1]/Series['Wavenumber']/@length
FIRSTY	<used to generate Y values>	
XYDATA	SeriesSet	/AnIML/ExperimentStepSet[1]/ExperimentStep[#]/Result['Spectrum']/SeriesSet
END	<not used>	
CLASS	?	
DATE	TimeStamp	/AnIML/ExperimentStepSet[1]/ExperimentStep[#]/Infrastructure[1]/TimeStamp[1]
TIME	TimeStamp	/AnIML/ExperimentStepSet[1]/ExperimentStep[#]/Infrastructure[1]/TimeStamp[1]
SAMPLEDESCRIPTION	Descriptive Name	/AnIML/SampleSet[1]/Sample[#]/CategorySet[1]/Category['Description']/ParameterSet[1]/Parameter['Descriptive Na
CASNAME	CAS Name	/AnIML/SampleSet[1]/Sample[#]/CategorySet[1]/Category['Substance Description'][#]/ParameterSet[1]/CategorySe
MOLFORM	Molecular Formula	/AnIML/SampleSet[1]/Sample[#]/CategorySet[1]/Category['Substance Description'][#]/ParameterSet[1]/CategorySe
CASREGISTRYNO	CAS Registry Number	/AnIML/SampleSet[1]/Sample[#]/CategorySet[1]/Category['Substance Description'][#]/ParameterSet[1]/CategorySe
WISWESSER	Wiswesser	/AnIML/SampleSet[1]/Sample[#]/CategorySet[1]/Category['Substance Description'][#]/ParameterSet[1]/CategorySe
MP	Minimum Temperature	/AnIML/SampleSet[1]/Sample[#]/CategorySet[1]/Category['Description']/ParameterSet[1]/CategorySet[1]/Category
BP	Minimum Temperature	/AnIML/SampleSet[1]/Sample[#]/CategorySet[1]/Category['Description']/ParameterSet[1]/CategorySet[1]/Category
SOURCEREFERENCE	Spectrum Data Source	/AnIML/ExperimentStepSet[1]/ExperimentStep[#]/Result['Spectrum']/CategorySet[1]/Category['Spectrum Descriptic
SPECTROMETERDATASYSTEM	DeviceIdentifier, Manufacturer, Na	Multiple
INSTRUMENTALPARAMETERS		
SAMPLINGPROCEDURE	Measurement Type, Sample Holder Position, Optical Path Environment, Optical Path Pressure	
DATAPROCESSING	Spectral Post-Processing	/AnIML/ExperimentStepSet[1]/ExperimentStep[#]/Method[1]/CategorySet[1]/Category['Spectral Post-Processing']
RESOLUTION	Spectral Resolution	/AnIML/ExperimentStepSet[1]/ExperimentStep[#]/Method[1]/CategorySet[1]/Category['Interferometric Method']/Ca
DELTAX	Increment	/AnIML/ExperimentStepSet[1]/ExperimentStep[#]/Result['Spectrum']/SeriesSet[1]/Series['Wavenumber']/AutoIncrer
MINX	<used as check only>	
MAXX	<used as check only>	
MINY	<used as check only>	
MAXY	<used as check only>	
BLOCKS	ExperimentStep	/AnIML/ExperimentStepSet[1]/ExperimentStep[#]@experimentStepID
CROSSREFERENCE	SampleReferenceSet	/AnIML/ExperimentStepSet[1]/ExperimentStep[#]/Infrastructure[1]/SampleReferenceSet[1]/SampleReference[#]@s
PEAKTABLE	<uvvis-peaktable.atdd>	
PEAKASSIGNMENTS	<uvvis-peaktable.atdd>	
XYPOINTS	SeriesSet	/AnIML/ExperimentStepSet[1]/ExperimentStep[#]/Result['Spectrum']/SeriesSet
NAMES	Name, Descriptive Name	Multiple
BEILSTEINLAWSONNO	Beilstein Lawson Number	/AnIML/SampleSet[1]/Sample[#]/CategorySet[1]/Category['Substance Description'][#]/ParameterSet[1]/CategorySe
REFRACTIVEINDEX	Refractive Index	/AnIML/SampleSet[1]/Sample[#]/CategorySet[1]/Category['Description']/ParameterSet[1]/CategorySet[1]/Category
DENSITY	Density	/AnIML/SampleSet[1]/Sample[#]/CategorySet[1]/Category['Description']/ParameterSet[1]/Parameter['Density']
MW	Molar Mass	/AnIML/SampleSet[1]/Sample[#]/CategorySet[1]/Category['Substance Description'][#]/ParameterSet[1]/Parameter[
CONCENTRATIONS	Concentration	/AnIML/SampleSet[1]/Sample[#]/CategorySet[1]/Category['Substance Description'][#]/ParameterSet[1]/Parameter[
STATE	State	/AnIML/SampleSet[1]/Sample[#]/CategorySet[1]/Category['Description']/ParameterSet[1]/Parameter['State']
PATHLENGTH	Sample Path Length	/AnIML/ExperimentStepSet[1]/ExperimentStep[#]/Method[1]/CategorySet[1]/Category['Common Method']/Category
PRESSURE	Optical Path Pressure	/AnIML/ExperimentStepSet[1]/ExperimentStep[#]/Result['Spectrum']/CategorySet[1]/Category['Ambient Conditions'
TEMPERATURE	Temperature	/AnIML/ExperimentStepSet[1]/ExperimentStep[#]/Result['Spectrum']/CategorySet[1]/Category['Ambient Conditions'
##= or $$ (Comments)	AuditTrail	
XLABEL	Series['name']	/AnIML/ExperimentStepSet[1]/ExperimentStep[#]/Result['Spectrum']/SeriesSet[1]/Series['Wavenumber']
YLABEL	Series['name']	/AnIML/ExperimentStepSet[1]/ExperimentStep[#]/Result['Spectrum']/SeriesSet[1]/Series['Intensity']

Best Practices

- Define guidelines for dealing with required elements
- Define/publish a crosswalk for each file format type
  - Indicate location/version used in AnIML file
- Start from template XML file OR
- Use XSL Transform (XML -> XML)
- Use audit trail to record how conversion was done
- Keep original file
  - Include in AnIML document
  - Create archive with unique identifier

Best Practices

- Include in AnIML file the converter used
- Use audit trail to record how conversion was done
- Keep original file
  - Include in AnIML document
  - Create digital archive with unique identifier