



**End User Meeting  
July 10, 2001**

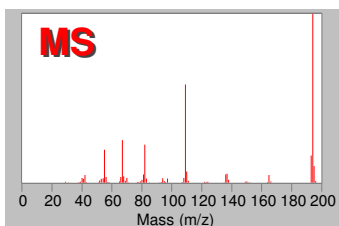
**ThermoFisher**  
S C I E N T I F I C

*The world leader in serving science*

## An XML Data Model for Analytical Instruments

James Duckworth

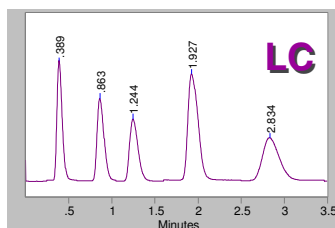
# Analytical Data: A Tower of Babel



**FOSS NIRSystems**



**Thermo Nicolet**



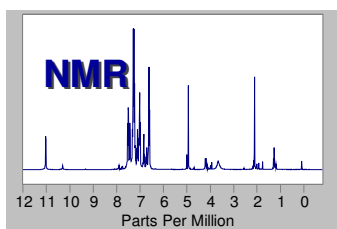
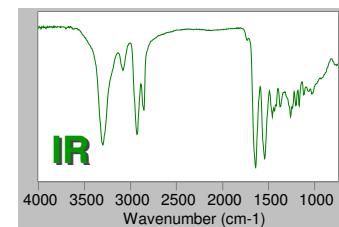
**Thermo Spectronic**

**Thermo Oriel**



**Thermo Laboratory Systems**

Waters



**HITACHI**

**JASCO**

**Thermo Finnigan**

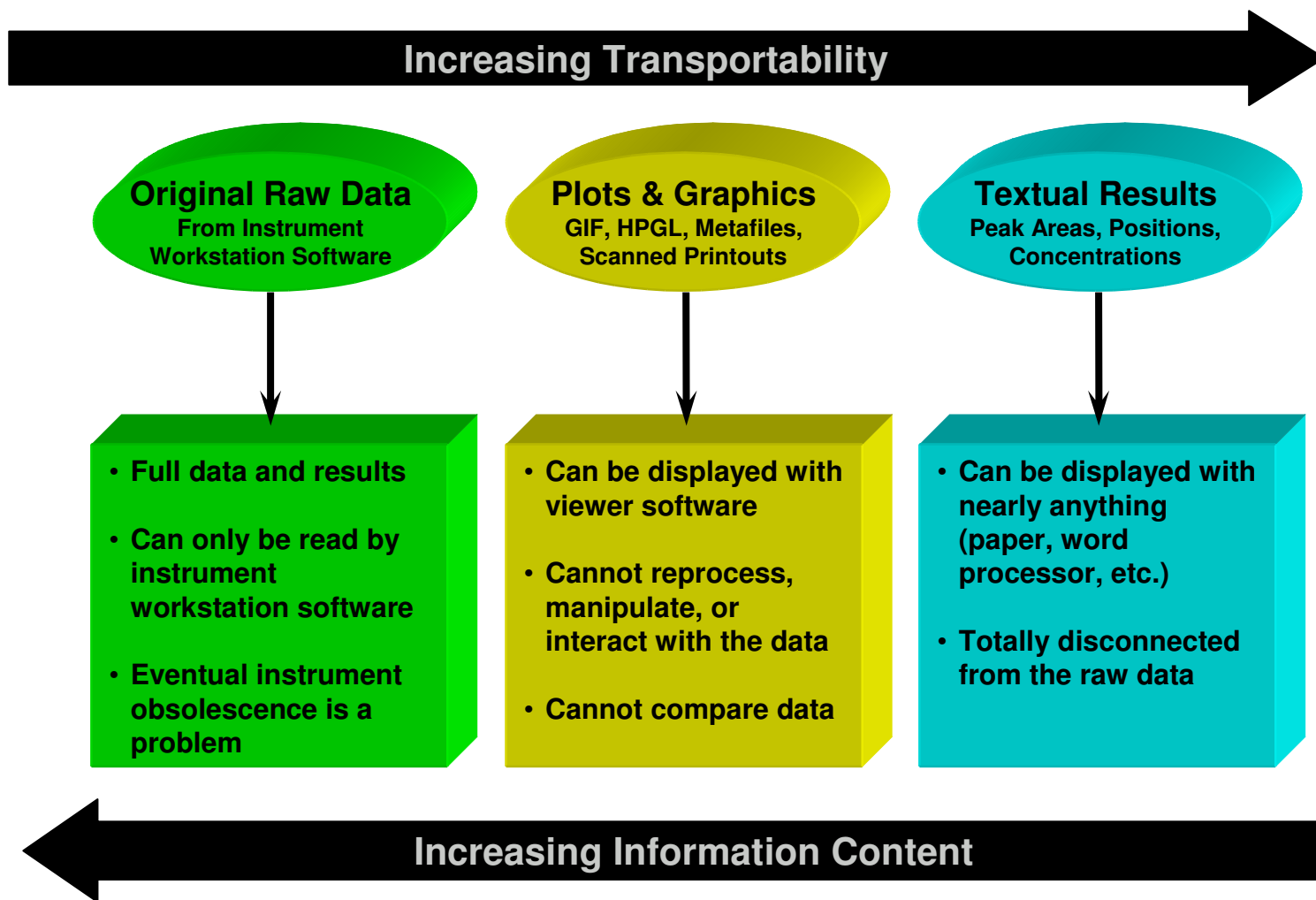


**Rigaku MSC**

## Proprietary Analytical Data Formats

- Labs are heterogeneous mix of instrumentation and vendors
- Relevant data is not always stored in one file
- Data retention periods often longer than instrument and data system lifetimes
- Potentially requires keeping outdated software operational for a long time

# Data Representations



# FDA 21 CFR 11: Data Formats

"The agency agrees that providing exact copies of electronic records in the strictest meaning of the word "true" may not always be feasible. The agency nonetheless believes it is vital that copies of electronic records provided to FDA be accurate and complete. **Accordingly, in § 11.10(b), "true" has been replaced with "accurate and complete."** The agency expects that this revision should obviate the potential problems noted in the comments. The revision should also reduce the costs of providing copies **by making clear that firms need not maintain obsolete equipment in order to make copies that are "true" with respect to format and computer system."**

# The Key To The Solution

- Translate and save in a neutral format
  - Must be both transportable and maintain information content
  - Enable data access from multiple applications
    - No need for obsolete hardware/software
  - Technology and IP from recent acquisitions of Galactic Industries Corp. and Thru-put Systems Inc.

# Technology & IP Acquisitions

- Galactic Industries Corp.
  - Founded 1988, joined Thermo 2001
    - GRAMS/32
      - Supports >150 spectroscopy file types
      - UV, FT-IR, Raman, NMR, ICP, XRD/F
- ThruPut Systems Inc.
  - Founded 1985, joined Thermo 1999
    - Target/Target DB (HP Chemserver)
      - Supports >50 chromatography file types
      - GC, LC, MS, PDA
- Now part of the Thermo Scientific Informatics Division



# Public-domain Data Formats in Use

- AnDI
  - Controlled by ASTM (E01.25)
  - MS & Chromatography only
- JCAMP
  - Controlled by IUPAC
  - Optical spectroscopy, NMR, MS
- SPC
  - Published by Galactic
  - Primarily optical spectroscopy



# AnDI Format

- Binary data format maintains data precision
  - Uses “public-domain” netCDF software maintained by Unidata
  - Source code; must be compiled for each platform
- Technique-specific data templates
  - Chromatography (ASTM E 1947-98)
  - Mass Spectrometry (ASTM E 2077-00)

# AnDI Chromatography Format

Data Element Name	Datatype	Category	Required
peak-number	dimension	C2	M2
peak-processing-results-table-name	string	C3	. . .
peak-processing-results-comments	string	C2	. . .
peak-processing-method-name	string	C2	. . .
peak-processing-date-time-stamp	string	C2	. . .
peak-retention-time	floating-point-array	C2	M2
peak-name	string-array	C3	. . .
peak-amount	floating-point-array	C2	M3
peak-amount-unit	string	C2	M3
peak-start-time	floating-point-array	C2	. . .
peak-end-time	floating-point-array	C2	. . .
peak-width	floating-point-array	C2	. . .
.			
.			

# JCAMP Format

- Completely ASCII-based
  - Simplifies transport and readability
- Fixed dictionary of tags
  - Required tags for core information
  - Custom tags allowed for private data
- Published and maintained by IUPAC

# JCAMP Format for FTIR

```
##TITLE=Polystyrene run as a film
##JCAMP-DX=4.24 $$ Nicolet v. 100
##DATATYPE=INFRARED SPECTRUM
##ORIGIN=
##OWNER=
##DATE=92/06/29
##TIME=12:57:07
##XUNITS=1/CM
##YUNITS=TRANSMITTANCE
##FIRSTX=399.241364
##LASTX=4000.128418
##FIRSTY=0.965158
##MAXX=4000.128418
##MINX=399.241364
##MAXY=0.965158
##MINY=0.000001
##XFACTOR=1.000000
##YFACTOR=1.000000E-009
##NPOINTS=1868
##DELTAX=1.928702
##XYDATA=(X++(Y..Y))
399.241 965157760 958141120 955421056 956603520 964025088 963178240
410.814 963215040 958321536 954287616 947153536 942139520 931181504
.
.
```

## Limitations of Current Formats

- Complex data description dictionaries, yet still not “complete”
- Numerical accuracy (JCAMP)
- Not “human readable” (AnDI & SPC)
- Cannot be easily validated for correct formatting and content
- Not extensible for future changes in equipment and analysis methods

# The XML Data Model

- Not a file format, but a data description language
- Can be used to represent any data structure
- Recently adopted XML Schema Definition (XSD) language provides strong data typing and syntax constraints
- Extensible by design

## Benefits of XML for Analytical Data

- Data is “human readable” ASCII text
- Public domain standard managed by W3C
- Documents can be externally validated for content and syntax (DTD or Schema)
- Hierarchical constructs for implying data relationships
- Proliferation of public domain tools
- Safe bet to be around for quite a while

## Analytical Data Model Design Goals

- Dictionary and hierarchy (Schema) must be compact and simple
- Make use of XML data types and hierarchies to mimic relationships in data sources
- Allow for future expansion
- Mind the file size, XML is all ASCII
  - It will compress nicely though...



# An XML Terminology Primer

- Element
  - Represents a fundamental piece of data or hierarchical relationship
- Attribute
  - Describes a property of an Element
- Schema (XSD)
  - Document that defines the allowed Elements, Attributes and relationships
- DTD
  - Document Type Definitions; older form of a Schema

# XML Data Representations

- Items that software need to “understand” must be fundamental elements
  - Data point values
  - Collect date/time stamp
  - Peak apex, baseline start/end
- Items that software only need for display and reporting can be generically represented
  - Peak area, height, skewness, etc.
  - Sample type, flow rate, “analyst shoe size”

# Breaking Down Analytical Data

- There are fundamental units of information that must be represented in the schema
  - Experiments (i.e. sequence lists)
  - Detectors
  - “Axes” (i.e. X, Y, Z, etc.)
  - Data points
  - Peaks (i.e. apex, baseline start/end)
  - Parameters

# Generalized Analytical Markup Language

<code>&lt;experiment&gt;</code>	data from single instrument "run"
<code>&lt;collectdate&gt;</code>	date & time of measurements
<code>&lt;parameter&gt;</code>	relevant instrument parameter
<code>&lt;trace&gt;</code>	data from a single detector
<code>&lt;coordinates&gt;</code>	coordinates for nD data (optional)
<code>&lt;values&gt;</code>	data values array
<code>&lt;Xdata&gt;</code>	X axis descriptor
<code>&lt;values&gt;</code>	data values array
<code>&lt;altXdata&gt;</code>	alternate X data descriptor (optional)
<code>&lt;Ydata&gt;</code>	Y axis descriptor
<code>&lt;values&gt;</code>	data values array
<code>&lt;peaktable&gt;</code>	peak list descriptor (optional)
<code>&lt;peak&gt;</code>	individual peak descriptor
<code>&lt;peakXvalue&gt;</code>	peak location
<code>&lt;peakYvalue&gt;</code>	peak intensity
<code>&lt;baseline&gt;</code>	baseline descriptor (optional)
<code>&lt;startXvalue&gt;</code>	baseline values
<code>&lt;endXvalue&gt;</code>	
<code>&lt;startYvalue&gt;</code>	
<code>&lt;endYvalue&gt;</code>	

# Instrumental Analysis

- Identify instrument type via "technique" attribute
  - Allows applications to know how to present/process data

```
<trace technique="CHROM" name="Chromatogram">  
.  
<trace technique="PDA" name="PDA Spectra">  
.  
<trace technique="NMR" name="13C NMR Spectrum">  
.  
<trace technique="MS" name="Mass Spectra">
```

# Curve Data Points

- Store data with no loss of information
  - Values are encoded “base64Binary” type to preserve numerical precision
  - Predefined list of "unit" attributes
  - Use "label" attribute for descriptive string

```
<Ydata label="Response" units="MILLIVOLTS">  
  <values byteorder="INTEL" format="FLOAT32" numvalues="3800">  
    8hkHQTqRBkFitAZBus8GQULjBkG6zwZBcl4GQSKVBkGiVgZB4nUGQbJ9BkG6  
    UgZBcl4GQUJmBkEyPwZBOPeGQbJ9BkECxAZBOPeGQTqRBkHidQZBaokGQcIn  
    .  
    .  
  </values>  
</Ydata>
```

# A Few Notes on "units"

- Applications must know the basis of data measurement
  - Data comparison or mining may require a transformation (i.e. "seconds" vs. "minutes")
- Similar problem exists in business applications
  - Pricing/quantity (i.e. "gallons" vs. "liters")
- Current XML standards? Not yet...
- Solution: Fixed list of units taken from IUPAC standards and past experience
  - Extend schema to adopt whatever W3C standard emerges in the future

# Parameters

- Avoid the “mapping” problem; all are stored using a single element type
  - Allowed to appear anywhere in hierarchy
  - The optional "group" attribute assigns class
  - The "name" attribute assigns identity
  - Use optional "label" attribute for descriptive string

```
<parameter group="inject" name="Inj Vol">6.00 ul</parameter>
<parameter group="inject" name="Dilution">2.5000</parameter>
<parameter group="inject" name="Position" >B124</parameter>

<parameter group="instrument" name="Flow Rate" >1.5 ml/min</parameter>
<parameter group="instrument" name="Column Temp" >27.5 C</parameter>

<parameter group="pkpick" name="Area Threshold" >27000</parameter>
<parameter group="pkpick" name="Bunch Factor">11</parameter>
```



# Peaks

- Represent key descriptors as standard elements
  - Remaining information stored in <parameter> elements

```
<peaktable name="Peaks">
  <peak name="Solvent" group="1" number="1">
    <parameter name="FIT_HGHT">178.9736</parameter>
    <parameter name="AREA">734.5404</parameter>
    <peakYvalue>187.377975463867</peakYvalue>
    <baseline>
      <startXvalue>7.3600001335144</startXvalue>
      <startYvalue>14.2526664733887</startYvalue>
      <endXvalue>27.8400001525879</endXvalue>
      <endYvalue>11.0759763717651</endYvalue>
    </baseline>
  </peak>
</peaktable>
```

## Example: Single Channel HPLC

```
<GAML>
  <experiment name="Injection 1">
    <trace technique="CHROM">
      <Xdata units="MINUTES" label="Ret. time">
        <values>
          <Ydata units="MILLIVOLTS" label="mV">
            <values>
              <peaktable>
<experiment name="Injection 2">
  <trace technique="CHROM">
    <Xdata units="MINUTES" label="Ret. time">
      <values>
        <Ydata units="MILLIVOLTS" label="mV">
          <values>
            <peaktable>
```

# Example: LC-PDA

```
<GAML>
  <experiment>
    <trace technique="CHROM">
      <Xdata units="MINUTES">
        <values>
          <Ydata units="ABSORBANCE">
            <values>
              <peaktable>
            </peaktable>
          </Ydata>
        </values>
      </Xdata>
    </trace>
    <trace technique="PDA">
      <values>
        <Xdata units="NANOMETERS">
          <values>
            <Ydata units="ABSORBANCE">
              <values>
                <Ydata units="ABSORBANCE">
                  <values>
                </Ydata>
              </values>
            </Ydata>
          </values>
        </Xdata>
      </values>
    </trace>
  </experiment>
</GAML>
```

# Example: LC-MS

```
<GAML>
  <experiment>
    <trace technique="CHROM">
      <Xdata units="MINUTES">
        <values>
          <Ydata units="UNKNOWN" label="TIC">
            <values>
              <peaktable>
            </peaktable>
          </Ydata>
        </values>
      </Xdata>
    </trace>
    <trace technique="MS">
      <values>
        <Xdata units="MASSCHARGERATIO">
          <values>
            <Ydata units="UNKNOWN" label="Abundance">
              <values>
            </Ydata>
          </values>
        </Xdata>
        <Xdata units="MASSCHARGERATIO">
          <values>
            <Ydata units="UNKNOWN" label="Abundance">
              <values>
            </Ydata>
          </values>
        </Xdata>
      </values>
    </trace>
  </experiment>
</GAML>
```

# Example: FTIR

```
<GAML>
  <experiment>
    <trace technique="FTIR" name="Interferogram">
      <Xdata units="UNKNOWN" label="Data Points">
        <values>
          <Ydata units="UNKNOWN" label="Energy">
            <values>
              <trace technique="FTIR" name="Spectrum">
                <Xdata units="WAVENUMBER">
                  <values>
                    <Ydata units="ABSORBANCE">
                      <values>
```

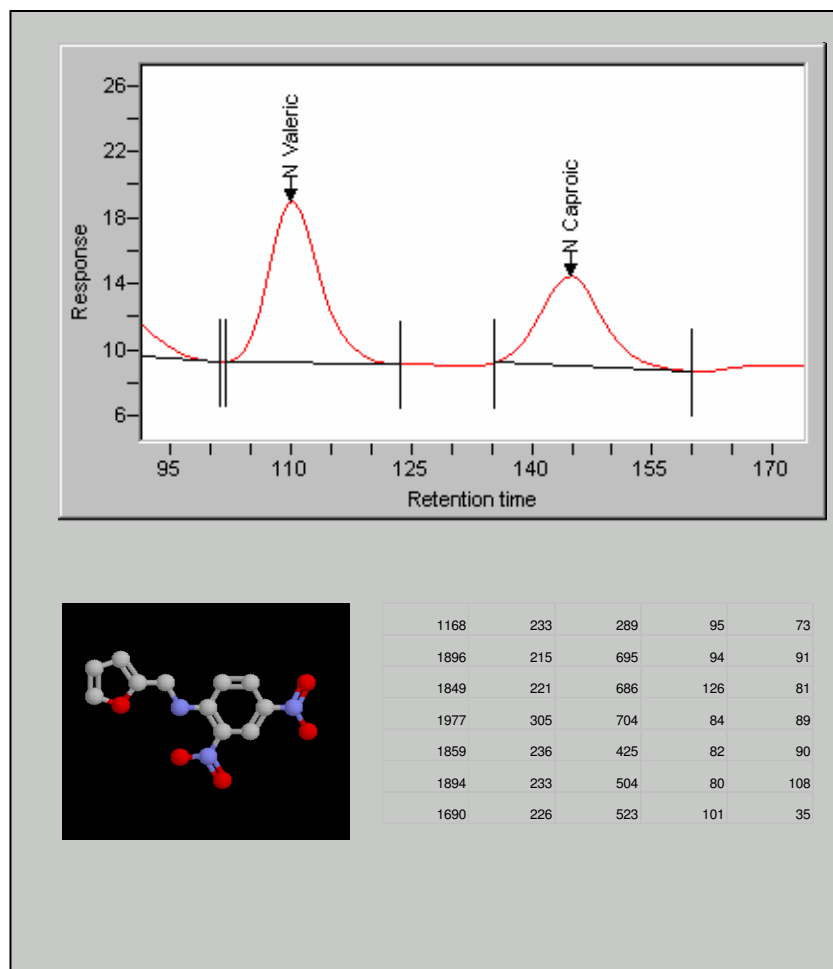
# Example: 1D NMR

```
<GAML>
  <experiment>
    <trace technique="NMR" name="FID">
      <Xdata units="SECONDS">
        <values>
          <Ydata units="UNKNOWN" label="Real">
            <values>
              <Ydata units="UNKNOWN" label="Imaginary">
                <values>
          <trace technique="NMR" name="Spectrum">
            <Xdata units="PPM">
              <values>
                <Ydata units="UNKNOWN" label="Real">
                  <values>
                    <Ydata units="UNKNOWN" label="Imaginary">
                      <values>
```

# Application Examples

- Web browser view
- Visual Basic program
  - COM controls
- XSL web pages
  - Developed independently
- XSD Schema design
  - Tools, validating document parsers

# GAML in the ELN Environment





# Where Do We Go Next?

- Schema circulated to selected instrument vendors & end users
  - Covered majority of analytical techniques
- Publish the schema
- Approach ASTM E01.25
  - Currently evaluating XML as replacement for the netCDF data model used by AnDI
- Suggestions?