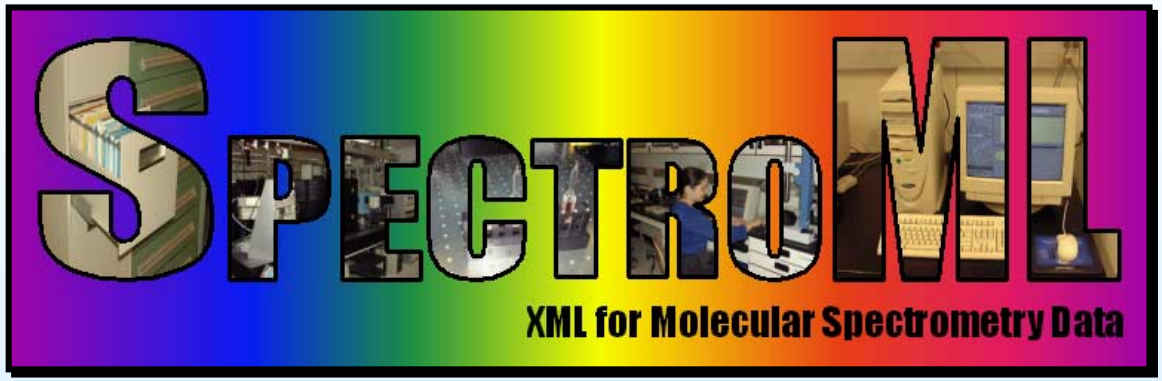
XML Standards for Analytical Chemistry



Gary W. Kramer

Analytical Chemistry Division
National Institute of Standards and Technology

NIST

In the beginning...

- Consortium on Automated Analytical Laboratory Systems (CAALS)
  - CAALS-I Communication Protocol
  - High-Level Communications Protocol
  - Common Command Set
  - Device Capability Dataset
- Laboratory Equipment Control Interface Specification (LECIS) - ASTM E1989-1998.
- System Capability Dataset
- How to Deal with Result Data?
- NIST In-House Data Interchange/Archiving Needs

NIST

Problems with Current Result Data Mechanisms

- Native Data Formats
  - Proprietary Formats
  - "Metadata" Separated from Result Data
    - Metadata & Data in Multiple Files
    - Metadata Not Available in Electronically
    - No Way to Link Metadata with Result Data
- Interchange Data Formats
  - Available for Only a Few Important Techniques
    - ANDI — GC, LC, MS
    - JCAMP-DX — IR/FTIR, NMR, UV/Vis
  - Fixed Order, Fixed Syntax, Immutable Formats
  - Content Limitations
  - Inconsistent Implementations
- Formats Incompatible with Modern Network Technologies

NIST

Goals for New Result Data Handing

- Extensible
  - Easy to Add New Elements without Breaking Existing Applications
- Flexible
  - Useful for Diverse Needs: Interchange, Interconversion, Archiving...
- Useable & Maintainable
  - Easy to Create, Use, Adapt, Maintain...
  - Readily Available Tools
- Acceptable
  - Use Standard Mechanisms Accepted by Mainstream Computing
- Network Friendly

☞ Extensible Markup Language

NIST

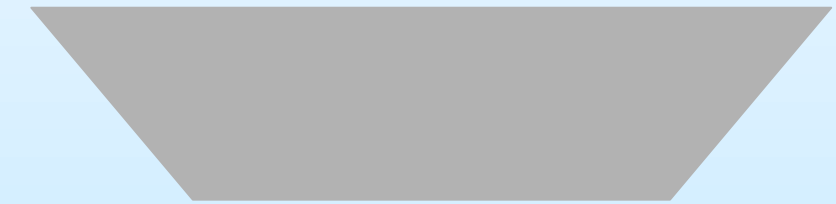
XML in 10 Points

- XML is for structuring data
- XML looks like HTML
- XML is text, but is not meant to be read
- XML is designed to be verbose
- XML is a family of technologies
- XML is new, but it has a history & a heritage
- XML turns HTML into XHTML
- XML is modular
- XML is license-free, platform-independent, and well-supported
- XML is a standard maintained by the W3C

www.w3.org/XML/1999/XML-in-10-points

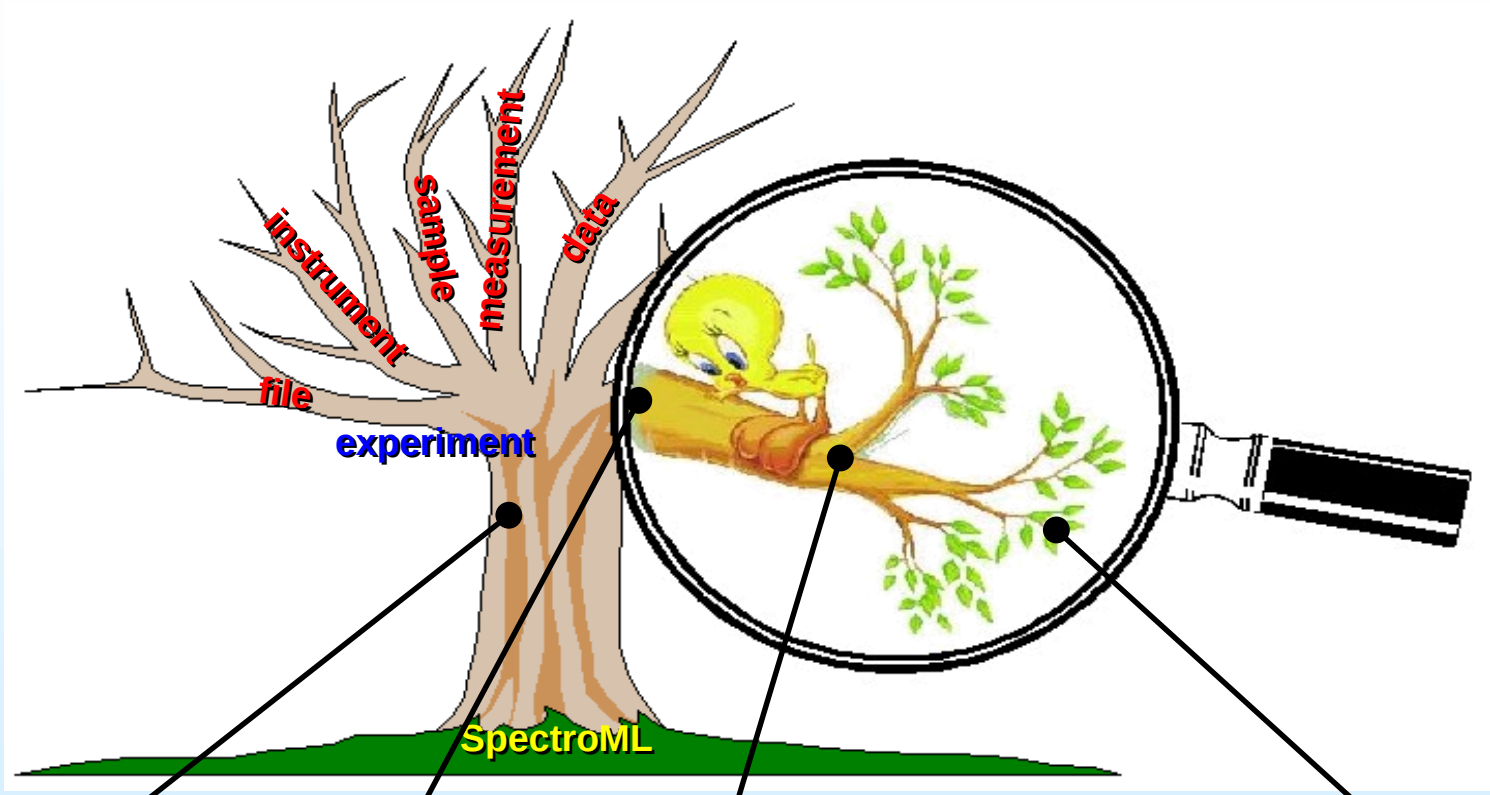
SpectroML

- SpectroML is a markup language for spectroscopy data based upon:
  - XML
  - JCAMP-DX (IUPAC)
  - ANDI/NetCDF (ASTM)
  - Thermo-Galatic GRAMS and SPC file format
  - Data definitions from instrument manufacturers
  - ASTM Definitions
- SpectroML was originally defined for UV-Visible spectral data



NIST

SpectroML Structure




experiment


groups

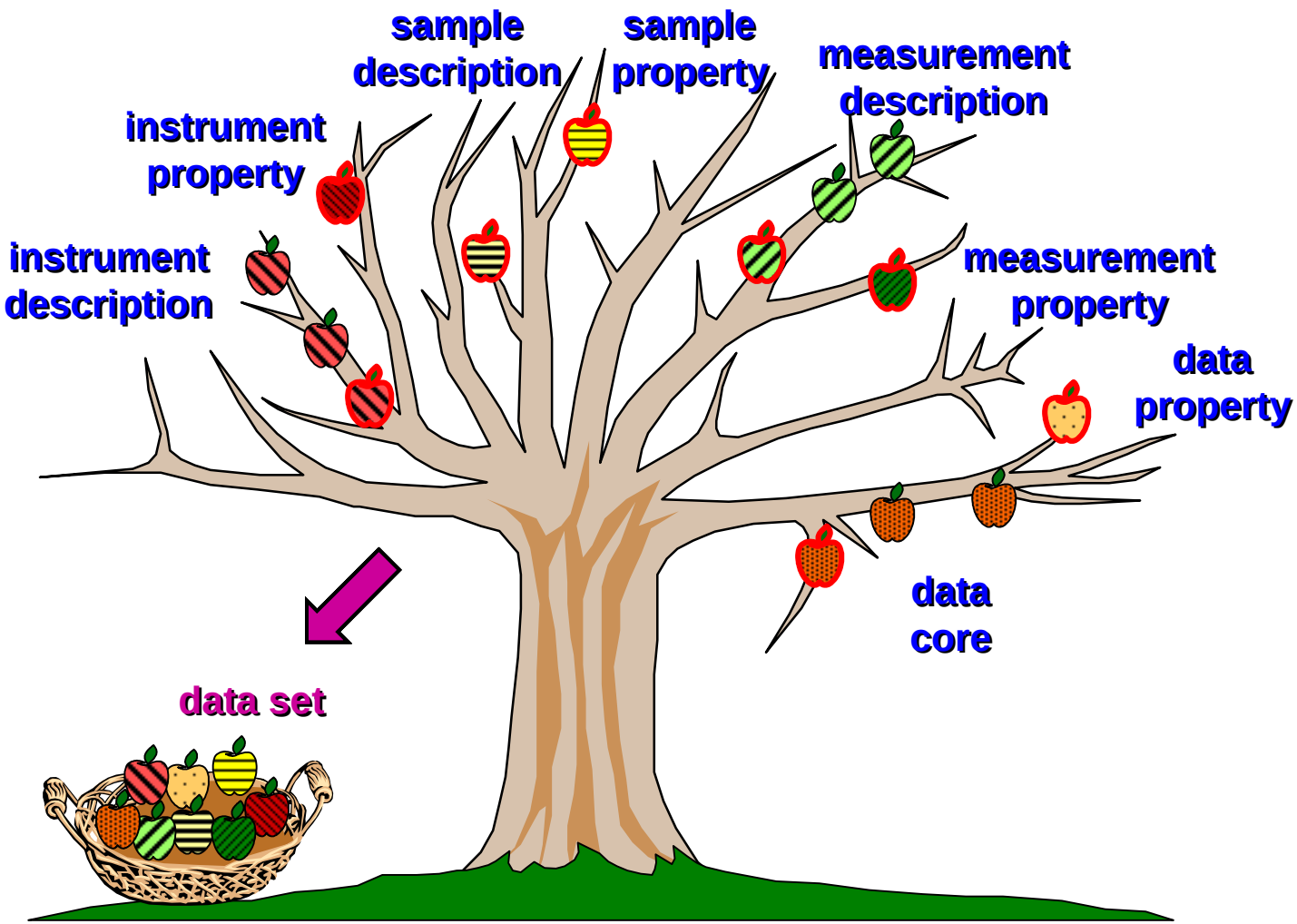

blocks


sections


elements

NIST

SpectroML Dataset



SpectroML Elements

Instrument

- instrumentDesignation
  - identifier
  - manufacturer
  - model
  - owner
  - location
- instrumentApplication
  - software
  - version
  - operatingSystem
  - firmware
  - operator
- instrumentSetting
  - resolution
  - linearDispersion
  - spectralBandWidthRange
  - wavelengthRange
  - absorbanceRange
  - detectorTypes
  - sourceTypes
- instrumentParameter
  - slitWidth
  - spectralSlitWidth
  - beamChannel
  - sampleHolder
  - samplePosition
  - scanSpeed
  - pointSeparation

Sample

- sampleDesignation
  - identifier
  - owner
  - location
  - casNumber
  - formula
  - storageMethod
  - disposalMethod
- samplePreparation
  - procedureMethod
  - timeStamp
  - operator
  - supplier
  - preparationDescription
- sampleAttribute
  - molecularWeight
  - meltingPoint
  - boilingPoint
  - density
  - refractiveIndex
- sampleParameter
  - state
  - pathLength
  - amount
  - pressure
  - temperature

Measurement

- measurementDesignation
  - identifier
  - title
  - owner
  - laboratoryReference
- measurementExecution
  - project
  - timeStamp
  - operator
- measurementParameter
  - measurementType
  - scanMode
  - referenceSample
  - filter
  - signalNoise
  - scanNumbers
  - scanDuration
- measurementCorrection
  - qualificationTimeStamp
  - qualificationReference
  - proficiencyTimeStamp
  - proficiencyReference
  - transmittanceTimeStamp
  - transmittanceReference
  - wavelengthTimeStamp
  - wavelengthReference

Data

- dataParameter
  - axisLabel
  - axisUnit
  - minimumValue
  - maximumValue
- dataCalculation
  - scaleFactor
  - numberPoints
  - pointIncrement
  - startValue

Data Values

- Single data points
- A single spectrum
- Multiple spectra
- Multi-dimensional data

NIST

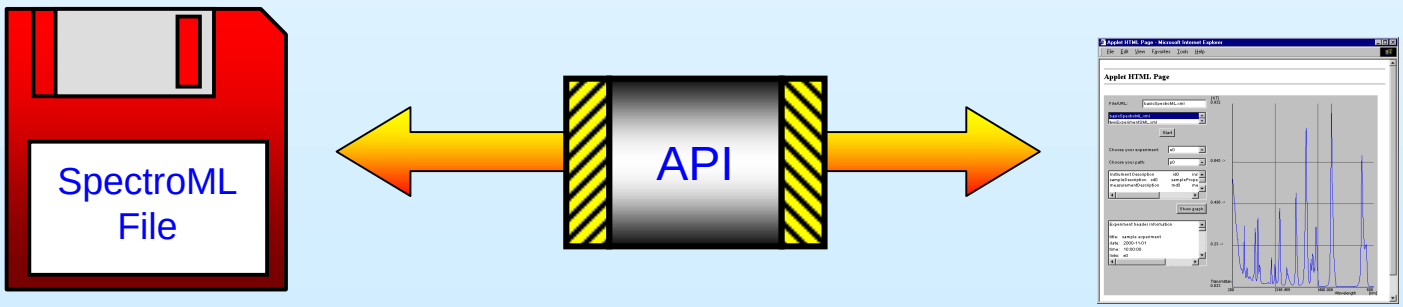
Example SpectroML File

```
<SpectroML version="1.0">
  <experiment type="UV/Vis" language="en-us">
    <file>
      <title>sample experiment</title>
      <timeStamp>
        <date>2002-11-07</date>
        <time>10:12:43</time>
      </timeStamp>
      <path pathId="p0"
        ...
        dataPropertyLink="dp0"
        dataCoreLink="dc0"/>
    </file>
    <instrument>
      ...
    </instrument>
    <sample>
      ...
      <sampleProperty samplePropertyId="sp0">
        <sampleParameter>
          <state>liquid</state>
          <amount unit="ml">5</amount>
        </sampleParameter>
      </sampleProperty>
    </sample>
    <measurement>
      ...
    </measurement>
    <data>
      <dataProperty dataPropertyId="dp0">
        <dataParameter>
          <axisLabel>
            <axis dim="x">Wavelength</axis>
            <axis dim="y">Transmittance</axis>
          </axisLabel>
          <axisUnit>
            <axis dim="x">nm</axis>
            <axis dim="y">%T</axis>
          </axisUnit>
        </dataParameter>
      </dataProperty>
      <dataCore dataCoreId="dc0">
        <values dim = "x">270 576</values>
        <values dim = "y">10 23</values>
      </dataCore>
    </data>
  </experiment>
</SpectroML>
```

NIST

Applying SpectroML

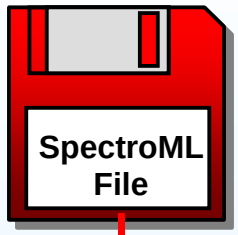
- DTD, Schema, & Sample File for SpectroML
- Stylesheet for Textual Output in Web Browser
- SpectroML API for Java and C++
- Demo Applications & Applets Using APIs

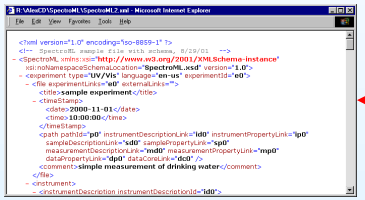


API

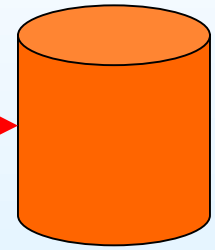
- Creating
- Selecting
- Editing
- Storing
- Searching
- Validating

Applying SpectroML





Web Browser

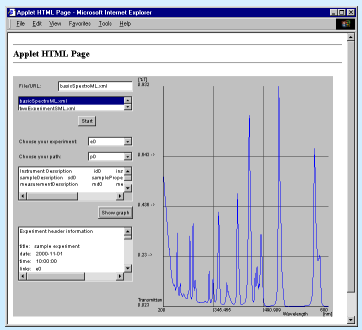


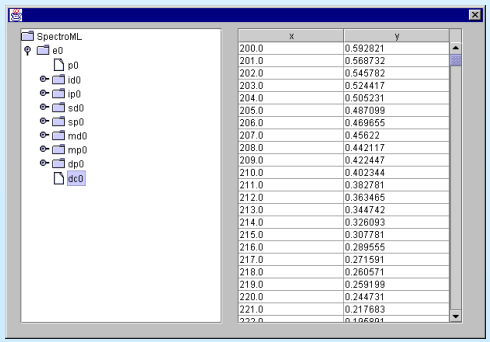
Database

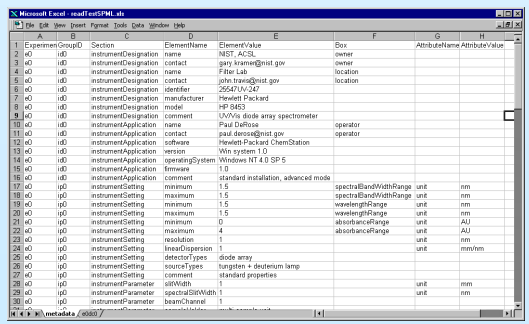
Java Applet

Java Program

Converting to
Another Format

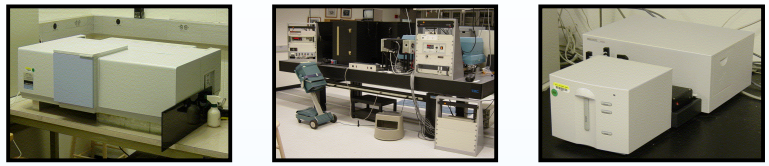






NIST

Applying SpectroML at NIST

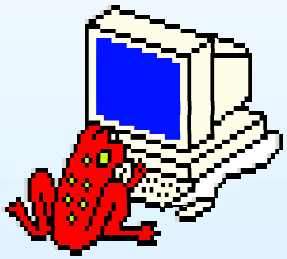


Result Data In Device-Dependent Formats

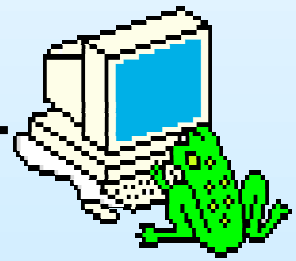
Device-Dependent Converter Software

SpectroML File

Result Data In SpectroML



[User Interface]



User Interface

Filters Access Database

Result Data In Database

NIST

Generalized Analytical Markup Language (GAML)

- Represent Analytical Data from Multiple Spectroscopy & Chromatography Techniques
- Compact, Simple Dictionary & Hierarchy (Schema)
- Use XML Datatypes & Hierarchical Structure to Mimic Relationships in Data Sources
- Avoid Parameter "Mapping" Problem
- Minimize the Need for Complex Dictionaries
- Permit Future Expansion
- Keep File Sizes Small

NIST

James Duckworth, Thermo LabSystems
www.gaml.org

GAML

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

NIST

James Duckworth, Thermo LabSystems
www.gaml.org

An XML-Based Standard for Molecular Spectrometry and Chromatography Result Data

- Possibility of XML-Based Approach Raised
  - ASTM E01.25 and E13.02 Meetings PittCon ‘00 Atlanta
- Demo of SpectroML and Applications
  - ASTM E13.01 Meeting PittCon ‘01 New Orleans
- First Organizational Meeting Held
  - ASTM E13.01 Meeting EAS ‘01 Atlantic City
- Task Group Organizational Meeting
  - ASTM E13.01.03 Meeting PittCon ‘02 New Orleans
- Task Group Working Meeting
  - ASTM E13.01.03 Meeting Shimadzu, Inc. 9/‘02
- New Subcommittee Meeting
  - ASTM E13.15 Meeting PittCon 3/‘03 Orlando
  - ANIML Working Group Meeting 4/'03 ASTM Headquarters

NIST

ASTM INTERNATIONAL
Standards Worldwide

ANIML an XML-Based Standard for Analytical Result Data

- SpectroML and GAML serve as starting points for an XML-based standard interchange format for molecular spectrometry and chromatography.

- Instrument manufacturers, data system & LIMS developers, software developers, end-users, consensus standards organizations, regulatory agencies, and other interested parties are invited to participate in this effort.

NIST

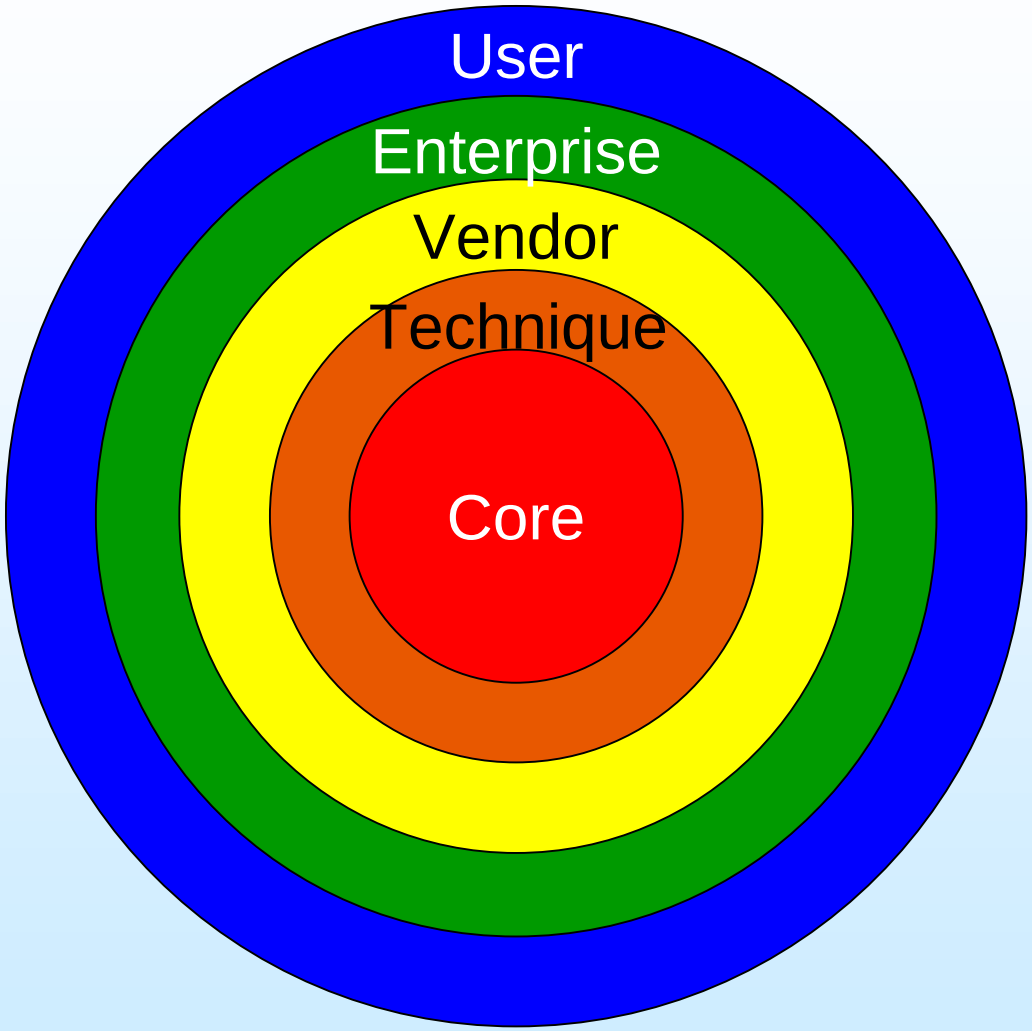
ASTM INTERNATIONAL
Standards Worldwide

Creating Analytical Information Markup Language (ANIML)

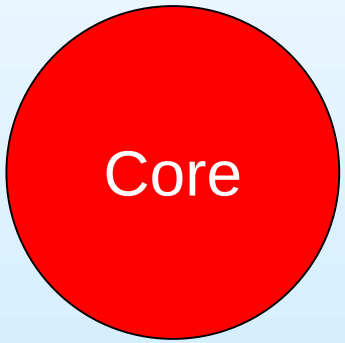
- Creating ANIML does NOT mean "starting over."
- ANIML should be built on existing ASTM, IUPAC, instrument vendor, and LIMS-developer efforts to define common data dictionaries.
- Once the schemas for ANIML are in place, straight-forward translators can be written to bridge current datasets to the new standard.
- ANIML should be developed in a way that makes it extensible to multiple techniques, yet avoids duplication of effort and dictionary entries.

NIST

Working Model

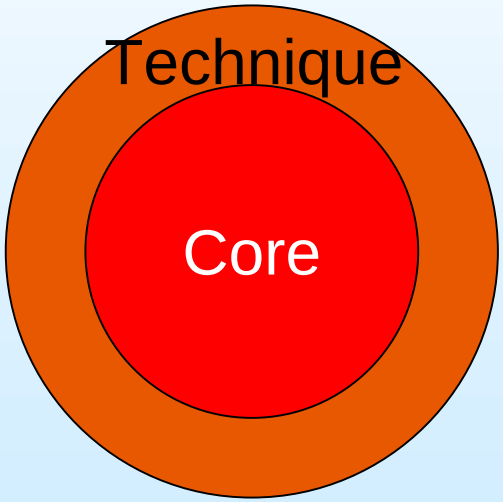


Core Schema



- Examples of Content
  - Sample Information
  - Basic System Information
  - Basic QA Information
  - Basic Units
  - Result Data
- Responsibility
  - Coalition of ASTM E-13.01.03 XML Work Group, IUPAC, ASMS,...

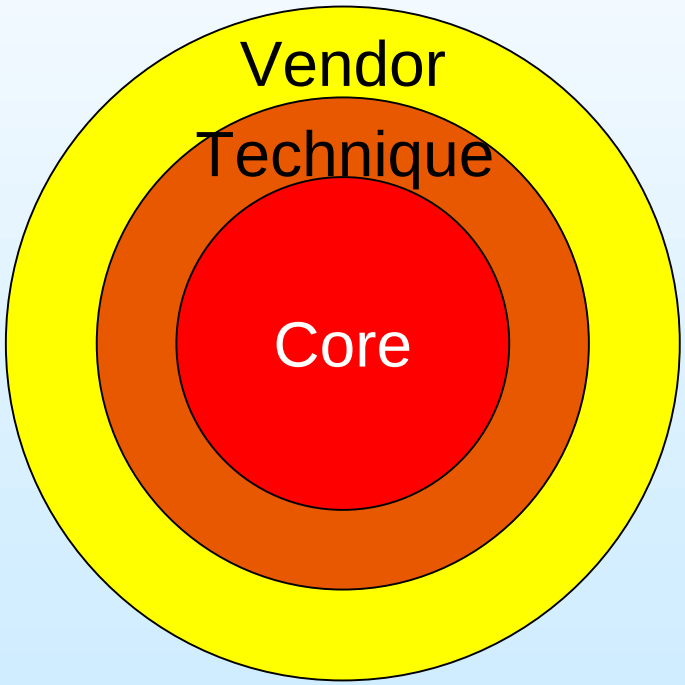
Technique-Specific Schema



- Examples of Content
  - System Information
  - Measurement Metadata
  - Method Information
  - QA Information
  - "Pre-result" Data
    - FIDs and Interferograms
  - Links for Hyphenated Measurements
  - More Units
- Responsibility
  - ASTM E-13.xx Subcommittees and Other Consensus Groups of Domain Experts

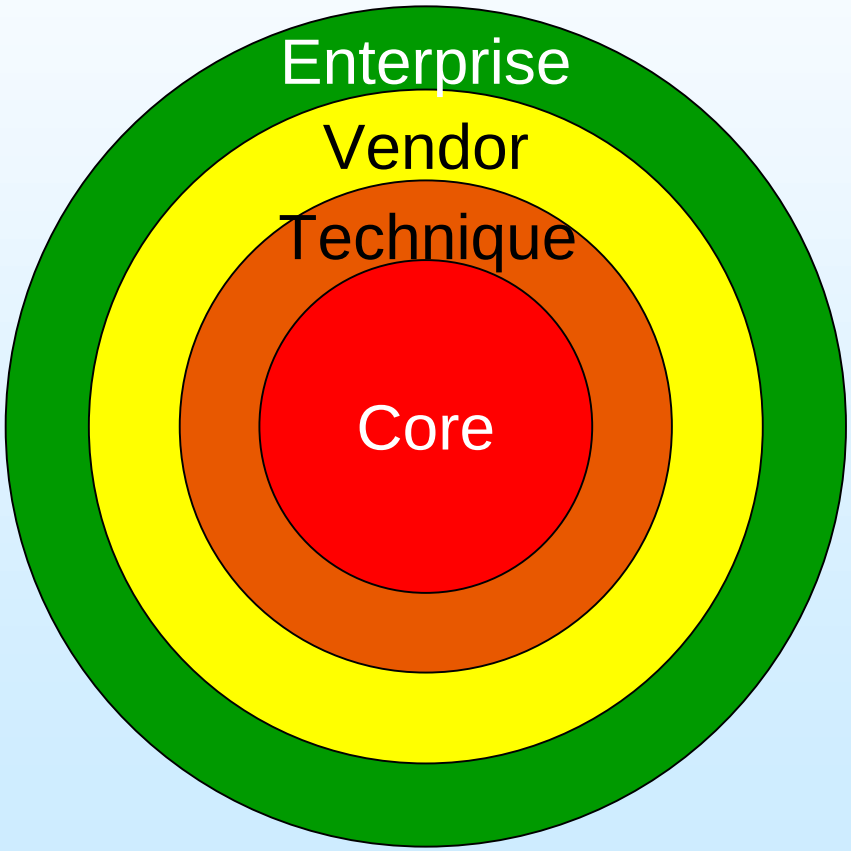
NIST

Vendor-Specific Schema



- Examples of Content
  - Method Information
  - Vendor-Specific Parameters
- Responsibility
  - Datasystem/Software/Instrument Manufacturers

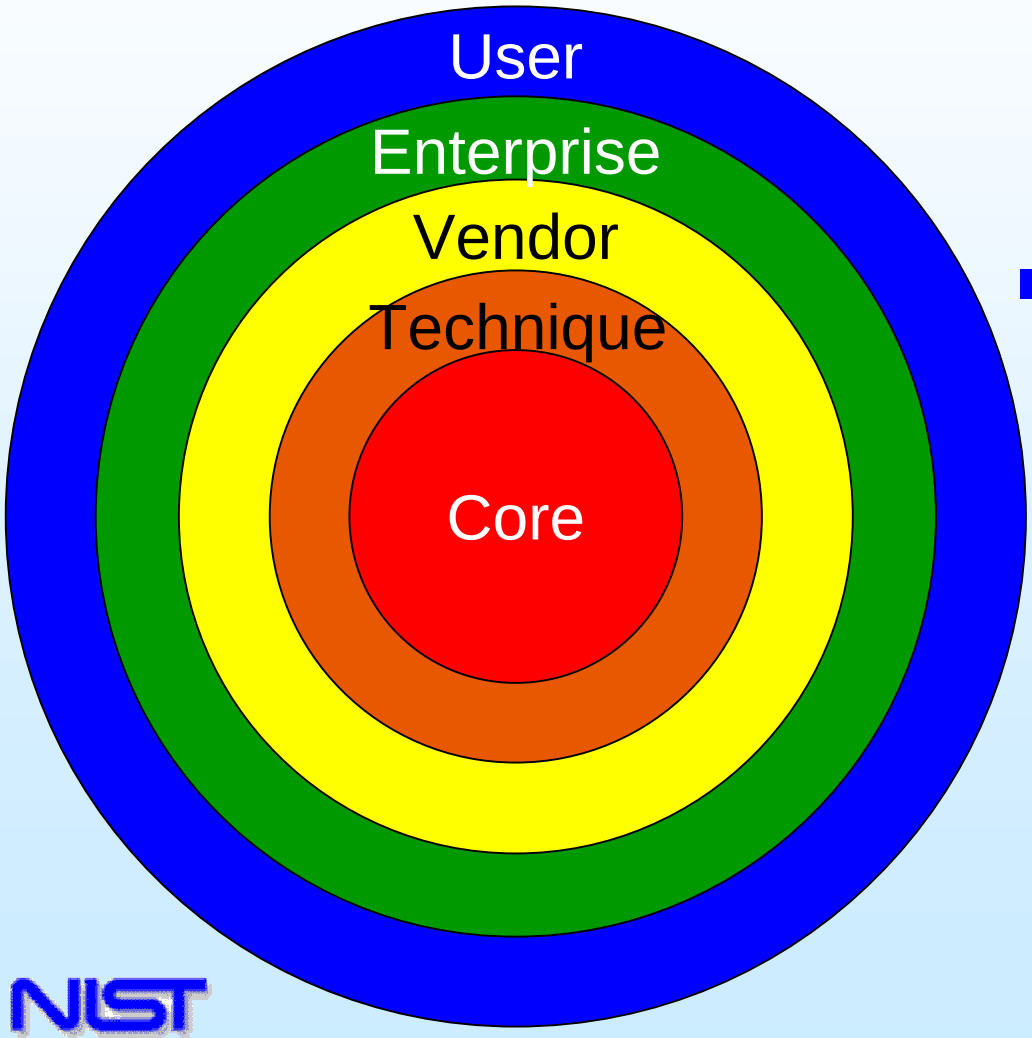
Enterprise-Specific Schema



- Examples of Content
  - Product/Project Links
  - Links to Databases
  - Enterprise-Specific QA
- Responsibility
  - Company/Organization

NIST

User-Specific Schema



- Examples of Content
  - Links to Notebooks or Other Records
  - Links to Other Result Data
- Responsibility
  - End User

NIST

Techniques

Molecular Spectroscopy and Chromatography

ASTM E13

- UV/Vis
- NMR
- Mass Spec
- IR
  - FTIR
  - NIR
  - Mid-IR
- Luminescence
  - Fluorescence
  - Phosphorescence
  - Chemiluminescence
- Raman
- Chromatography
  - GC
  - LC
  - CE
  - TLC
- Hyphenated Techniques
  - GC-MS
  - LC-MS
  - LC-Diode Array
  - LC-NMR
  - LC-Diode-Array-MS
  - ...

NIST

Considerations for Terms & Concepts in the Core

- Is the term/concept shared by most spectroscopy and chromatography techniques?
- Is the term/concept used in most datasets collected by most spectroscopy and chromatography techniques?
- Items that software needs to "understand" must be fundamental elements.
- Items that software only need for display and/or reporting can be more generically represented.
- Representations must be consistent with the requirements of FDA's US 21CFR11 and ISO 17025.

NIST

MakeUp of Core

- Sample Information
  - Name
  - Identifier
  - Solvent
- Basic System Information
  - Technique
  - System Name
- Basic QA Information
  - (ANIML Version)
  - Title/Experiment Name
  - Date
  - Time
  - Operator/User/System Name

NIST

MakeUp of Core

- Basic Units
  - Time
  - Temperature
  - Length
  - Energy
  - Mass
  - Charge
  - ...
- Result Data
  - Data Type
    - IEEE 32-Bit Floating Point little endian
    - IEEE 64-Bit Floating Point little endian
  - Data Encoding
    - Base64 Binary
  - Structure
    - n-Tuple

NIST

FAQs, Whines, What If's

- What if we don't get it right the first time?
- What happens when more "definitive" schemas appear?
- We/I don't like the definitions in the lower levels.
- What if new extensions of the technique emerge, new techniques are invented, or techniques are "hyphenated?"
- Can XML handle my variable temperature, pH-gradient, LC-MS-MS-MS/PDA/NMR data?
- Since XML is text, how can data tampering be avoided?
- What do we do with our bizillions of JCAMP (or ANDI or…) spectra?

NIST

More Information

- XML
  - http://www.w3c.org
  - http://www.xml.org
  - http://www.xml.com
  - http://www.xml101.com
- SpectroML
  - ftp://caals.nist.gov/pub/download/
  - http://www.xml.org…registry...schemas DTDs…chemistry
- GAML
  - http://www.gaml.org
  - http://www.xml.org…registry ...schemas DTDs …chemistry
- ANIML
  - http://animl.sourceforge.net

NIST

Acknowledgements

- NIST Systems Integration for Manufacturing Applications (SIMA) Program
- SpectroML
  - Professor Reinhold Schäfer - Fachhochschule Wiesbaden
  - Alexander Rühl - Definition, DTD, and Schema
  - Martin Peschke - Applications, Applets, C++ & Java APIs
  - Akyut Arslan - Instrument-to-SpectroML Applications
  - Anh Dao Nguyen - SpectroML-to-Database Application
- GAML
  - James Duckworth - Thermo Galactic
- ANIML
  - Randall Julian - Eli Lilly and Co.
  - David Martinsen - American Chemical Society

NIST