The AnIML Data Model

Peter J. Linstrom
Physical and Chemical Properties Division
National Institute of Standards and Technology

Summary

- What is a “data model”?
- What is the basic structure of the model?
- More details
  - Vectors and parameters
  - Atomic data types
  - Technique definitions
  - Containers
- Flexibility versus ease of use

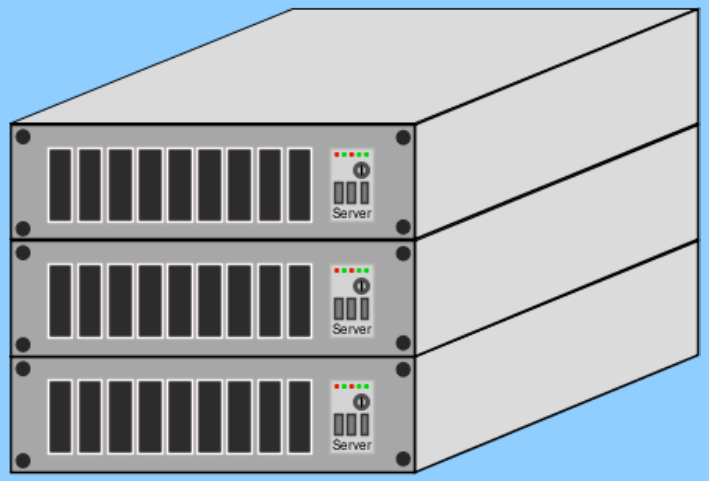
What is a data model?

- The logical structure of a data in a system.
  - Determines the type(s) of data that can be stored
  - May provide information on relationship(s) between data
  - May be optimized for a given problem or type of data

Some well known data models

	A	B
1	**Item**	**Budget**
2	Apples	$4.00
3	Oranges	$8.00
4	Bananas	$50.00
5	Star fruit	$12.50

Spreadsheet / table



Relational DB schema

```
##TITLE=2-Butanone
##JCAMP-DX=4.24
##DATA TYPE=INFRARED SPECTRUM
##STATE=gas
##XUNITS=1/CM
##YUNITS=ABSORBANCE
##XFACTOR=1.0
##YFACTOR=0.000079898
##DELTAX=4.0
##FIRSTX=450.0
##LASTX=3966.0
##FIRSTY=0.073586
##MAXX=3966
##MINX=450
##MAXY=0.79898
##MINY=0
##NPOINTS=880
##XYDATA=(X++(Y..Y))
450.0 921 395 57 0 101 66 263 239 222 433
490.0 295 264 386 382 535 614 618 683 606 571
```

JCAMP-DX

JCAMP-DX

```
##TITLE=2-Butanone
##JCAMP-DX=4.24
##DATA TYPE=INFRARED SPECTRUM
##STATE=gas
##XUNITS=1/CM
##YUNITS=ABSORBANCE
##XFACTOR=1.0
##YFACTOR=0.000079898
##DELTAX=4.0
##FIRSTX=450.0
##LASTX=3966.0
##FIRSTY=0.073586
##MAXX=3966
##MINX=450
##MAXY=0.79898
##MINY=0
##NPOINTS=880
##XYDATA=(X++(Y..Y))
450.0 921 395 57 0 101 66 263 239 222 433
490.0 295 264 386 382 535 614 618 683 606 571
```

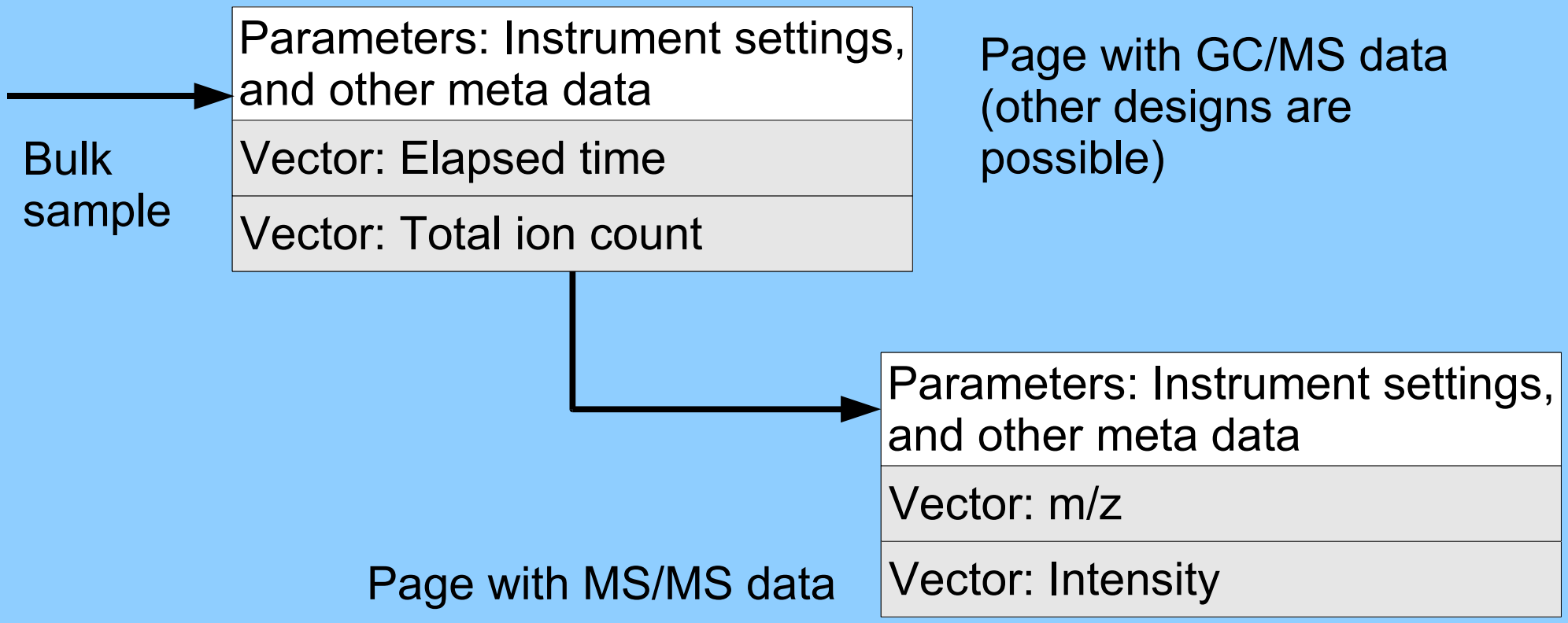
Name-value pair
(single value)

Sequence of X,Y points
X = fixed increment sequence
Y = sequence of values

What is the basic structure of the model?

- A collection of data for a related group of measurements is called a *page*
- Each page contains
  - Sequences of data called *vectors*
  - Single values called *parameters*
- Vector and page types are defined in technique definitions

Example – GC/MS/MS



- Only a simple example of AnIML's capabilities
- This talk will only covers the contents of a page

What are the benefits to this approach?

- Can represent a wide variety of data types
- Technique definitions allow semantic details of vectors and parameters to be specified
- “Hyphenated” experiments represented with multiple pages
- Multivariate experiment steps represented with pages with multiple vectors (not limited to just X, Y data)

Vectors and parameters

- Vectors typically hold the main experimental output
  - Intensity and wavelength values in spectroscopy
  - X,Y, intensity data for raw FTIR
- Parameters typically hold information about the experiment or procedure
  - Instrument settings
  - Time / location of run
  - Tracking information

Supported atomic data types

- Data types for vectors and parameters are set in the corresponding technique definition.
- Supported types include the following:
  - Floating point number
  - Integer number
  - Boolean value
  - String (label)
  - ISO date / time string
  - SVG (graphics)

Floating point types

- AnIML supports three different floating point types
  - 32 bit IEEE floating point values as text
  - 64 bit IEEE floating point values as text
  - 64 bit IEEE floating point values as base-64 encoded little-endian binary values (base-64 encoding converts binary to text)
- All types can be used for “floating point” data
- Units can be specified

Technique definitions

- Define vector and parameter types for a type of experiment
- Definitions include
  - Name
  - Atomic data type
  - Valid range for data
  - Indicator if the item is required
  - Text description of the item
  - Additional items for vectors

Additional items in vector definitions

- Continuity (discrete versus continuous)
- Dependency (independent versus dependent)
- Information to help viewer programs plot data
  - Should the item be used for plotting data
  - Preferred scaling (log versus linear)
  - Displayed by default (yes or no)

Example – centroided mass spec

- Vectors
  - m/z is an independent variable
  - Intensity is a discrete dependent variable – data plotted as peaks
- Parameters
  - Instrument / inlet type
  - Instrument settings
  - Local tracking information (time, location, etc.)

Containers

- Vectors and parameters are containers
- A parameter may contain a single value of one of the supported atomic data types
- A vector may contain one or more of the following
  - A sequence of discrete items
  - A compact representation for sequences which change by a constant increment (auto increment set)
  - A base-64 encoded set of binary floating point data

Vector example – GC temperature

- Consider a heat program where temperature is constant, ramped for a period of time and then held constant:

Auto increment set – initial temperature
Auto increment set – temperature ramp
Auto increment set – final temperature

Vector with temperature program with increment sets used for each portion of the program. The first and last sets have zero increment.

Vector example – spectroscopy

- Consider spectroscopic data where wavelength is regularly incremented and intensity is recorded as binary values:

Auto increment set – incrementing wavelength

Wavelength vector

Base-64 encoded set of 64 bit IEEE little endian floating point numbers – final temperature

Intensity vector

Viewing spectroscopy data

- From the technique definition (or data file)
  - Wavelength is an independent variable (x axis)
  - Intensity is a dependent, continuous variable (y axis, points should be connected)
- From the data file
  - Wavelength values calculated from increment set specification
  - Intensity values (binary) obtained by decoding base-64 encoded data
  - Units specified in data file

Flexibility versus ease of use

- Does all this flexibility come at a price?
  - Yes – increased complexity over JCAMP-DX
  - No – well designed software tools and pre-defined technique definitions should mean that complexity is not visible to the end user

Conclusion

- The AnIML data model is designed to support a wide range of experimental data.
- More information can be found on the AnIML web site:

http://animl.sourceforge.net/