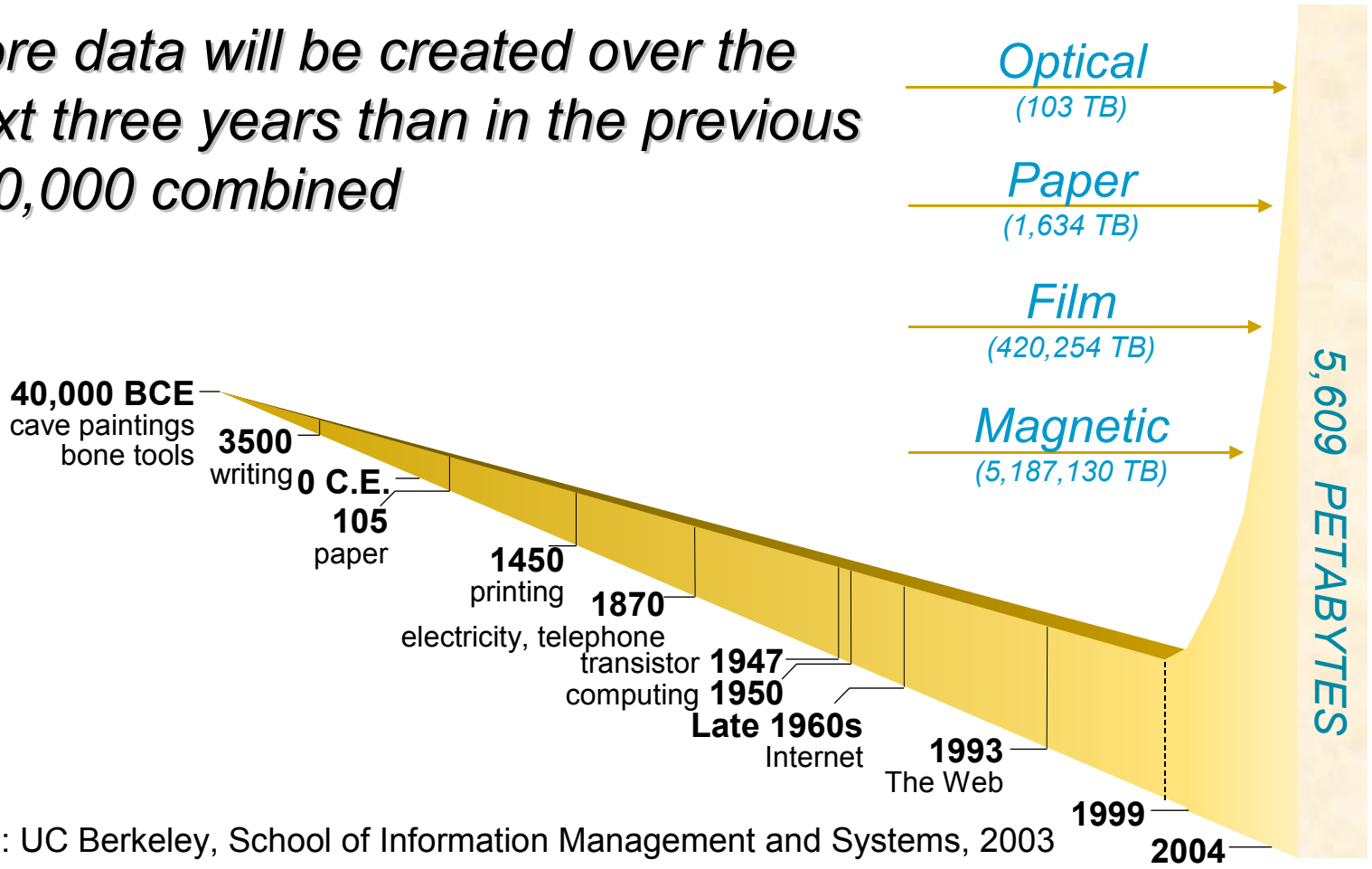
Long Term Storage of Chromatographic Data...

AnIML, TNF, Viewers, and Plenty of Challenges!

Mark Mullins
Agilent Technologies

More and more data…

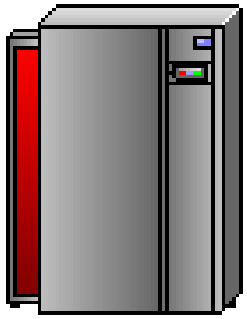
More data will be created over the next three years than in the previous 300,000 combined



Source: UC Berkeley, School of Information Management and Systems, 2003

Different sources and types of data…

Files



Databases

Structured Data

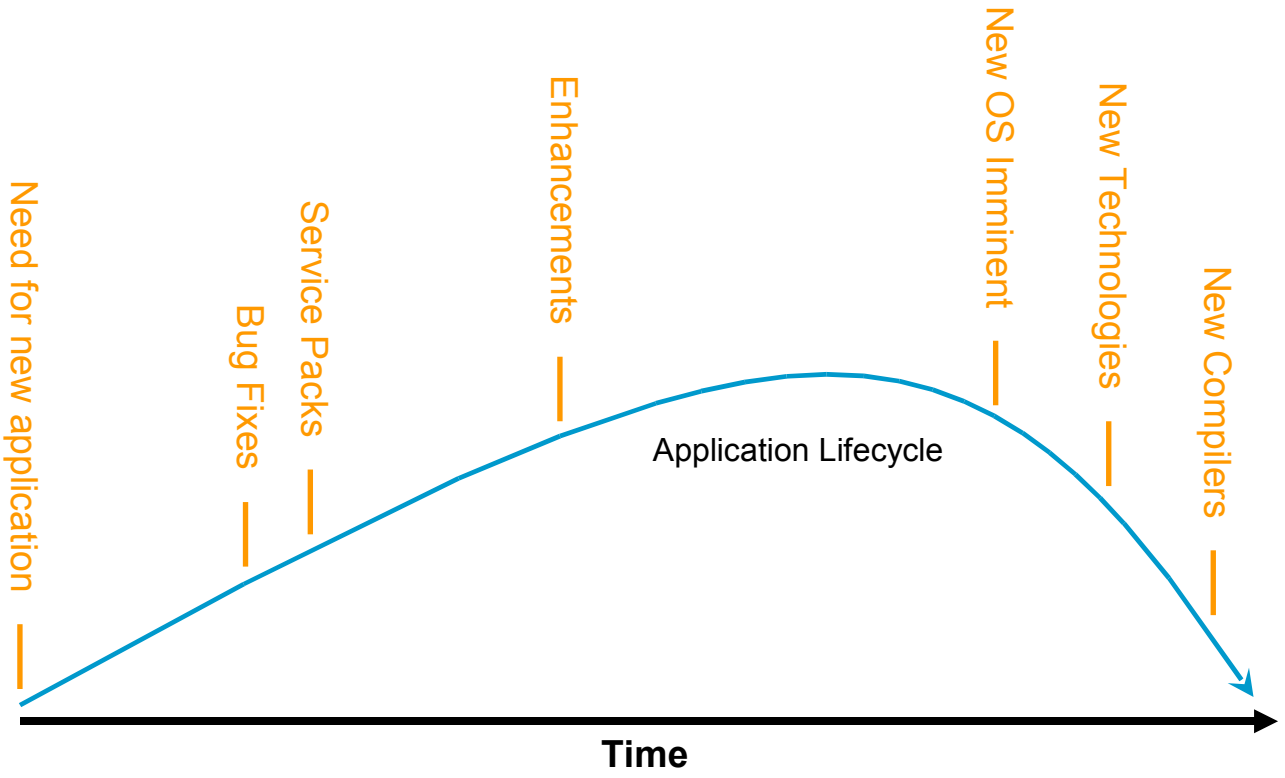
Unstructured Data

Retention periods…

- **Regulations**
  - 10, 20, 30 years
- **SOPs**
  - 40, 50… sometimes upwards of 100 years!

Retired applications

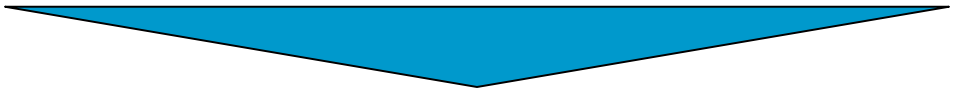
All applications progress through a natural lifecycle.



The need for Technology Neutral File (TNF) formats

Critical data must:

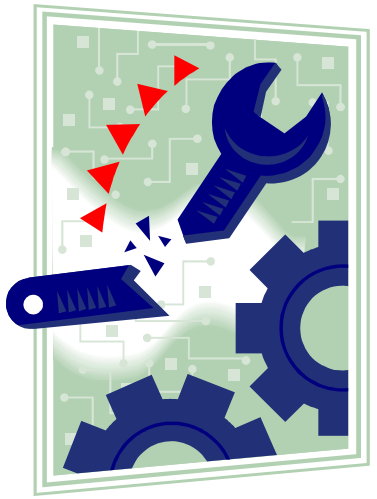
- **Be preserved in its entirety**
- **Be OS independent**
- **Outlive the creating application**
- **Must be human readable (not binary or proprietary formats)**
- **Must be usable today (viewing and analysis)**



Structured Text Files

The problems with multiple TNF formats

- Little or no interoperability
- Must create multiple viewing and analysis tools
- Proliferation of more formats
- Maintenance and versioning nightmare for developers
- New applications must support all previous formats
- “My format is best” syndrome



The advantages of a standardized format

- Easy exchange of data between applications
- Consistent and well known architecture
- Tools can be designed to work across versions
- Generic tools can be developed and shared
- Shared vendor support for standard format
- Format will be maintained and supported, even if vendors come and go

AnIML to the rescue

- AnIML is a standardized file format
- AnIML is a structured text file, using XML technology
- AnIML is generic and is not vendor specific
- AnIML is human readable
- AnIML is all-inclusive. Every bit of data from an entire experiment can be represented and stored in an AnIML file
- AnIML is flexible, while still predictable
- Data in an AnIML file can be tightly constrained for a given analytical technique

Mapping data to AnIML

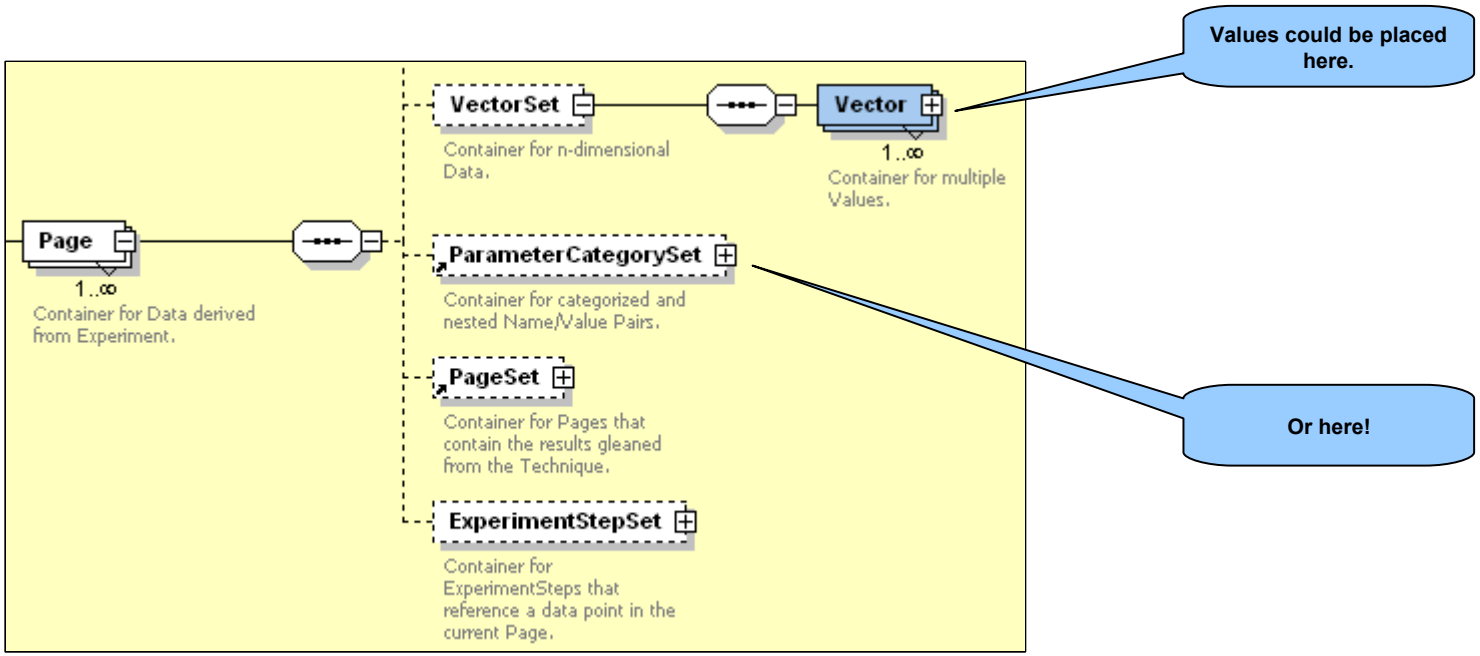
Application developers can begin to map analytical data into AnIML by educating themselves on the following topics:

- AnIML Core Schema
  - This schema is the heart of AnIML, and ultimately defines the structure for all data in AnIML XML files
- AnIML Technique Documents
  - These schemas define the rules for your structured data, given a particular analytical technique

Mapping data to AnIML

Example

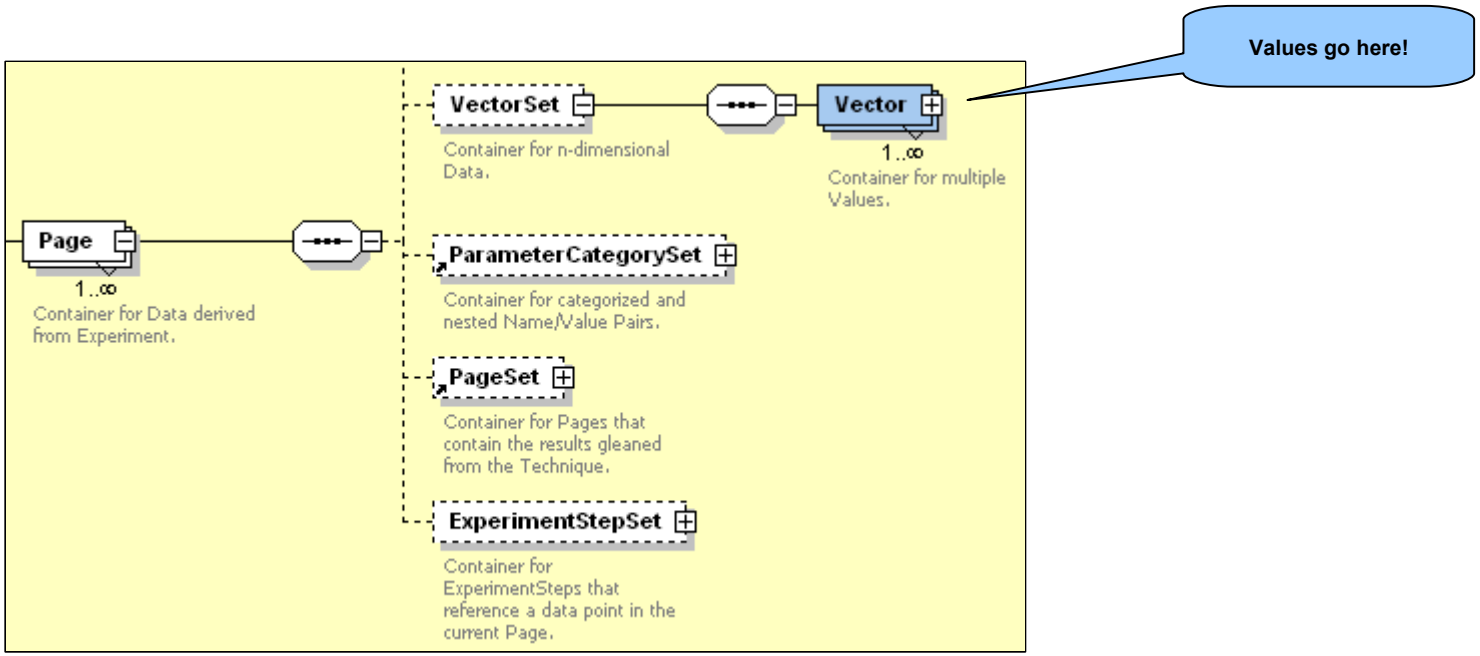
- **Mapping Position of Peak and Height of Peak into the AnIML schema**
- **Without a technique document, where do we put these items, and what are they called?**



Mapping data to AnIML

Example

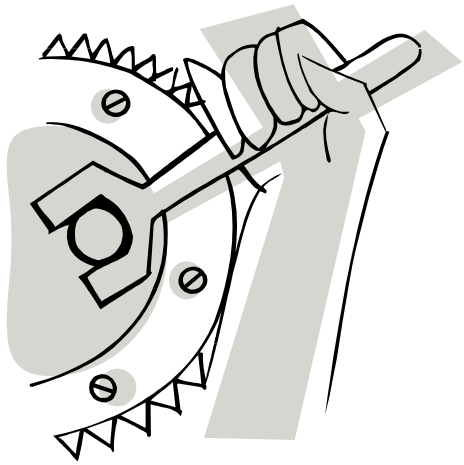
- **The technique document tells us to put these items inside of a Vector, and call them PeakPosition and PeakHeight, respectively**



Customizable Architecture

Technique documents cover common values only

Peak Number
Peak Position
Peak Height
Peak Width
Peak Area
Peak Amount
etc.



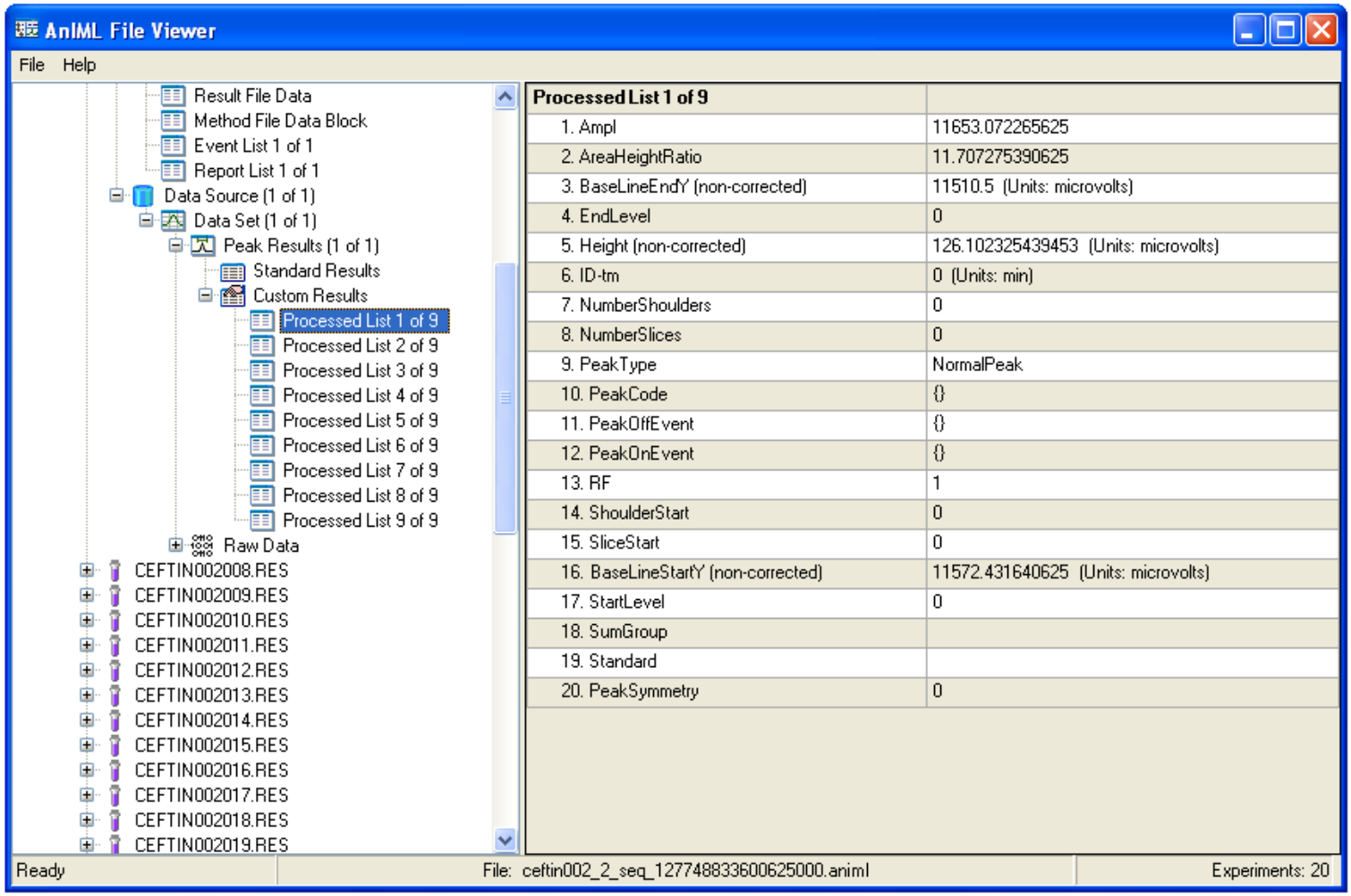
What to do with custom data system values?

Area to Height Ratio
Number of Shoulders
Peak Integration Events
etc.

Answer…

- **The core schema provides for storage of custom data through a concept called ParameterCategorySets**
- **The data is still structured, and can be discovered and viewed by generic viewers**

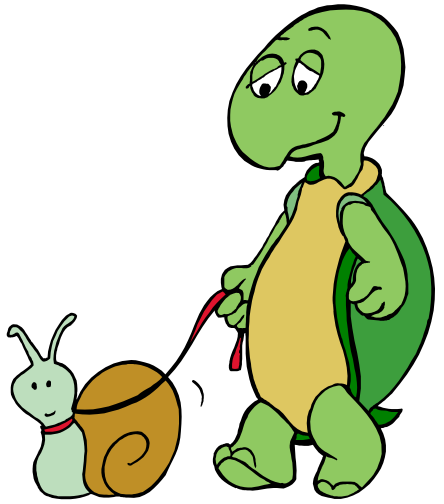
Customizable Architecture



A typical AnIML file can be quite LARGE

A typical Chromatograph AnIML file can easily be 50,000+ lines of text. Includes items such as:

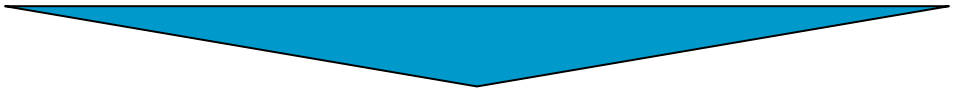
- **General file information**
- **Method configuration**
- **Instrument configuration**
- **Injector configuration**
- **Calibration information**
- **Raw data results**
- **Peak results**
- **Revision information**
- **etc.**



Developers need to be aware of the size requirements, and design viewers for speed from the ground up.

Best programming practices

- **Encapsulate logic to write sections of the AnIML file into object classes**
  - Encourages code reuse, and allows bugs to be fixed in one place
  - Enhancements and changes are easy to make
- **Maintain a level of indirection between source data and the AnIML file**
  - If new features and/or versions of the AnIML schema are released, changes are easily accommodated



- Tools, applications, viewers should operate on the indirect data
  - When changes occur upstream, the tools will continue to work unmodified, once the intermediate object classes are changed

Demo

- View real AnIML XML file
- View same AnIML file in Agilent's AnIML File Viewer

Summary

- Massive amounts of data are being generated
- Much of this data must be kept for 30+ years
- Applications retire, but the data must live on, in a TNF format
- AnIML is being created by the ASTM subcommittee E13.15, and is the standard for TNF representations of analytical data
- AnIML is a highly structured, but flexible file format
- Tools, applications, and viewers are already being generated around AnIML

Questions