AnIML -The Need for Analytical Data Standards within the Pharmaceutical Industry

Mark F. Bean, Ph.D.
Molecular Discovery Research
GlaxoSmithKline, Philadelphia Area

Why Analytical Data Standards?

1. Vendor-neutral software – 10 min

    - Open-access

    - Quantitation

    - Metabolomics

    - Proteomics…

2. Report collation for NDAs – 2 min

3. Compliance (CFR 21 Part 11, GMP) – 10 min

Multiple vendors, instruments, and techniques: an Information Babel

Waiting for Unification (X**ML**)

1. Research (not regulated)

Use Case: *custom software linked to a single vendor needs to move to or include a different vendor*

- ChemStation macros
- Software to place results and metadata in a database
- Complete, custom packages

1. Open-Access LCxx

- Vendor-neutral open-access LCUV or LCMS
  - Sample login
  - Sequence-batch submission
  - Instrument and sample status reporting
  - *Data processing*
  - *Data archival*
  - *Results viewing, reviewing, correcting, reporting*

1. Open-Access LCxx

Vendor-neutral, technology-neutral formats (TNFs)

- – Official standards: JCAMP-DX, NetCDF, ANDI
- – Community standards: mzML
- – Vendor supplied: GAML, OpenLab, RPT, etc.

Companies specializing in vendor-neutral software that rely on standards where available:

- – Waters SDMS, Thermo (Galactic), Agilent ECM, Mascot, ACD, Mestrenova …

ASTM INTERNATIONAL
Standards Worldwide

NIST

IUPAC
Committee on Printed & Electronic Publication

Subcommittee on Electronic Data Standards
XML and JCAMP-DX

1. Open-Access LCxx

Binary Format Issues

- Custom numeric formats optimized for fast GCMS on early computers

- Little and Big Endean

- Octal, hexadecimal, 32- and 64-bit

1. Open-Access LCxx - CANDI

- Wrap vendor software with neutral methods
  - Sequence, batch submission
  - Status reporting
  - Data extraction
  - Data archiving
- Database
  - Atomize data to Oracle tables (samples, chromatograms, spectra, peaks, instruments, devices, etc.)
- Generic Login and Viewer
- XML datasets as common comm. format

2. New Drug Application (NDA)

Use Case 2: *collate, error-check data from varied sources and sites*

- Incoherent metadata

- Rooms full of people doing this

- Allotrope Foundation
  - http://www.allotrope.org

3. Regulated Environments

Use Case: *conform to government regulations*

- CFR21 Part11

- Cost = million$$

- Convert old data systems to TNF

- After site closures, all data must be converted and retained

- Orphan instruments – image computer and save (run in virtual environment so hardware no longer needed)

3. Electronic Data Regulations

- CFR21 part11 (electronic records and signatures)
    - Controls, validations, and audits of computer systems and software
    - Accurate, complete, human readable records
    - Signed, time-stamped, audit trail
    - Analysts to be educated, trained, and qualified
    - Are PDFs or reports suitable for LCMS?
        - *"preserve the content and meaning of the record""*
        - *FDA wants the same capabilities the user had at the time of inspection*
- Good Manufacturing Practice (GMP)
- http://www.labcompliance.com/

3. Data Retention Policies

- 65 years for:
  - Patient safety, legal, marketing, IP
  - Raw data
  - PDFs not suitable for LCMS
- 30 years for rest:
  - Pragmatically set at department level
  - Do not preserve data copies
    - Danger of invalidating patents if inconsistent data
  - LCMS PDFs and reports ok
  - Destroy most unregulated raw data after 3 years (except NMR)

3. Industry Challenges

- Difficult to maintain data and devices
  - Aging, failing, tape drives and optical
  - Outdated computers
- Instrument applications
  - Proprietary formats
  - Often not backwards compatible
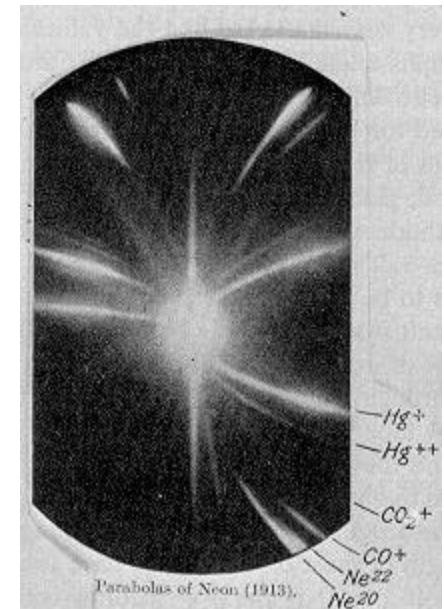 - Decommissioned Instruments
- Operating systems

30-65 Years is a Long Time - 1900s

- MS data

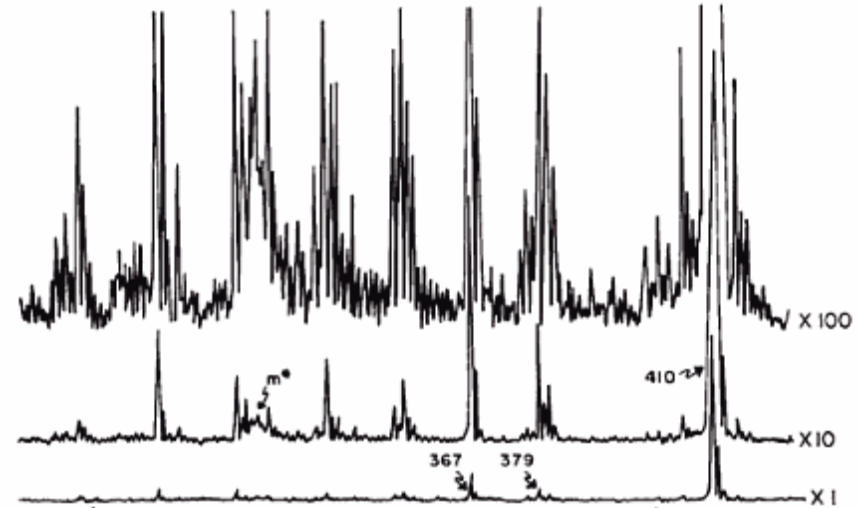

Parabolas of Neon (1913).

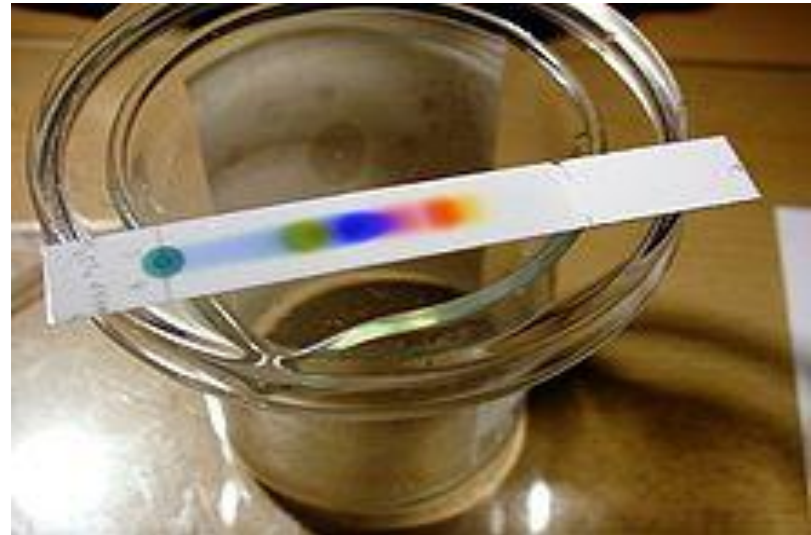- ~~Chromatography~~

- 1950s

- MS data

- Chromatography

- 1983 (30 yrs)

- MS data

- Chromatography

Encoding Binary Number Arrays

- ## JCAMP-DX uses text representations
  - "14563.09"
  - Enables scientific precision
  - Human readable

- ## AnIML uses Base64Binary

  <EncodedValueSet startOffset="0" endOffset="2846">

AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAgAAAAAAAABgOAAAAAAzM4M4AAAAAM3MLDgAAAAAmpkZOAAAAABnZmY4AAAAAAAAAYDgAAAAAMzOzOAAAAAADNzLw4AAA
AAAAACDkAAAAAmpnpOAAAAAAAAAEA4AAAAAGdm5jcAAAAAZAAAGdmNjoAAAAAmpkrOgAAAAAAADI6AAAAAAAATDoAAAAAzcx
mNjOgAAAACamWk6AAAAAJqZeToAAAAAZ2aGOgAAAAAzM5I6AAAAAM3MkjoAAAAAzcyNOgAAAADNzJA6AAAAAJqZkToAAAAAm
pmPOgAAAADNzIw6AAAAAJqZgDoAAAAAz………

……………AAAADNrEQ8AAAAAJo5QzwAAAAAAIBCPAAAAADN7EI8AAAAADOTQzwAAAAAM9NDPAAAAAAz00M8AAAAAADgQzwAAAAA
M5NDPAAAAACa2UI8AAAAAAZ6YkPAAAAADNbCM8AAAAAM1MIjwAAAAAzawhPAAAAADNrB88AAAAAM1sHTwAAAAAMxMcPAAA
AADNrBo8AAAAAGeGGTwAAAAAM1MaPAAAAAAzExs8AAAAAGfmGjwAAAAAMzMaPAAAAACaWRk8AAAAAJoZGDwAAAAAM/MXP
AAAAAAzsxg8AAAAAJpZGDwAAAAAmhkXPAAAAABnhhU8AAAAAGemEjwAAAAAmrkPPAAAAADN7Aw8AAAAAJpZCzwAAAAAzYwL
PAAAAACa2Qs8AAAAAJr5CjwAAAAAZyYJPAAAAAAAQAg8AAAAAM3MBjwAAAAAmlkFPAAAAACaWQM8AAAAAACgADwAAAAAzcz
9OwAAAAAAQPo7AAAAAM2M9TsAAAAAZ6bxOwAAAABnZu07AAAAAJrZ6zsAAAAA
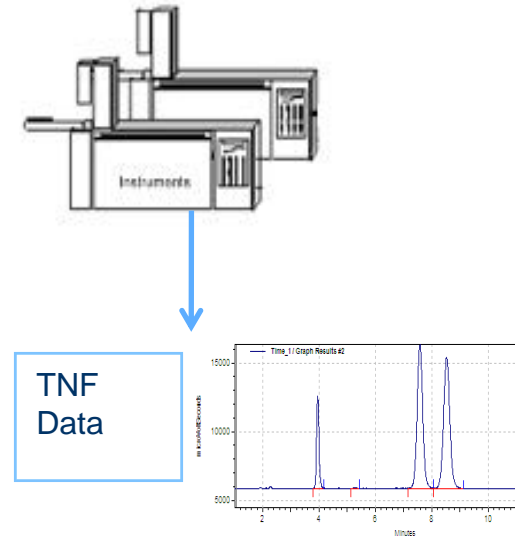
  </EncodedValueSet>

  - Preserves data integrity for regulated environments

>95% of Data Looks Like This

The remaining 5% is 95% of the work!

3. Solutions

- Short term solution:
  - Maintain instrument workstation images
  - Conversion of proprietary data to technology neutral format

TNF Data

- Long term solution:
  - Support industry standards for data
    - ASTM AnIML committee