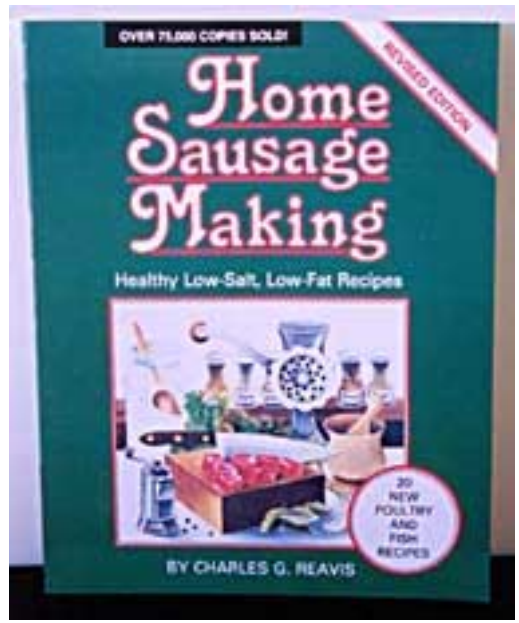
Age of exchange standards: the NeXus file format

Przemek Klosowski
NIST/NCNR

Genesis

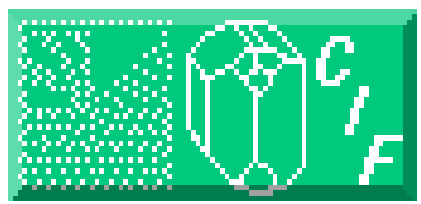
- Scientists are interested in the final result. Data is just the process, best hidden from others.
- Results are communicated by a publication. Exchange format: well-known peer-reviewed journal, latin glyphs printed on acid-free paper.



Don't knock those: Ptolemy's tables are still readable after 1800 years!

Original computer-based formats

- Ad-hoc formats, not intended for public use application-specific ASCII or binary
- Exchange formats driven by presentation of final result
  - Example: Crystallographic Information File (CIF) from the Int'l Union of Crystallography: standard method for submitting results to publication **`http://www.iucr.org/iucr-top/cif/`**



Litany of Sins

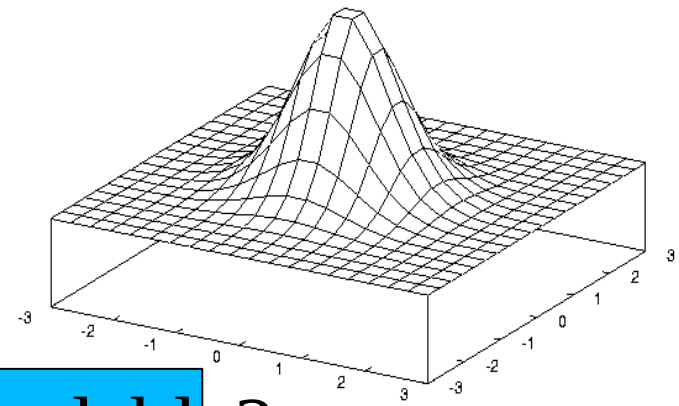
- Non-extensible
- Non-self-describing
- Incomplete
- Not codified
- No introspection
- Unsophisticated structure
- Imprecise
  - High level (spec) and low level (FP numbers)
- Infatuation with ASCII
- Or mysterious binary
- Inefficient
  - Buffering/caching
  - Compression
  - Typing (char/int/FP/double)
- NIH
- Format conversion hell

ASCII-do we need human-readable

False dichotomy between printable and binary data format:

- Thick printout syndrome: can anyone read more than, say, 1000 numbers?
- Modern data requires visualization, anyway

82,80,351,133,0,66,55,39,53,21,31,38,
99,55,14,456,767,0,74,52,85,63,290,1
86,155,372,78,86,1396,94,0,52,55,35,
96,40,31,30,39,29,157,310,521,803,0,
620,197,67,80,196,205,



Question: Is this format human-readable?

Answer: Yes, unlike most ASCII formats

Features: Important and Unimportant

- Structure/hierarchy
- Markup
- Extensibility
- Portability
- Introspection
- Self-described
- Attribute-rich

- ASCII vs. binary
- Maybe even specific encoding (method of serialization of structured data)

Age of standards :)

- Bad experiences with old data
- Success of standards (Internet, Web, even MS Office if/when it works)
- Concensus for standard formats, e.g. XML
- XML is not a panaceum:

<xml><OfficeData body=”AFBE-1039-8FCD.....

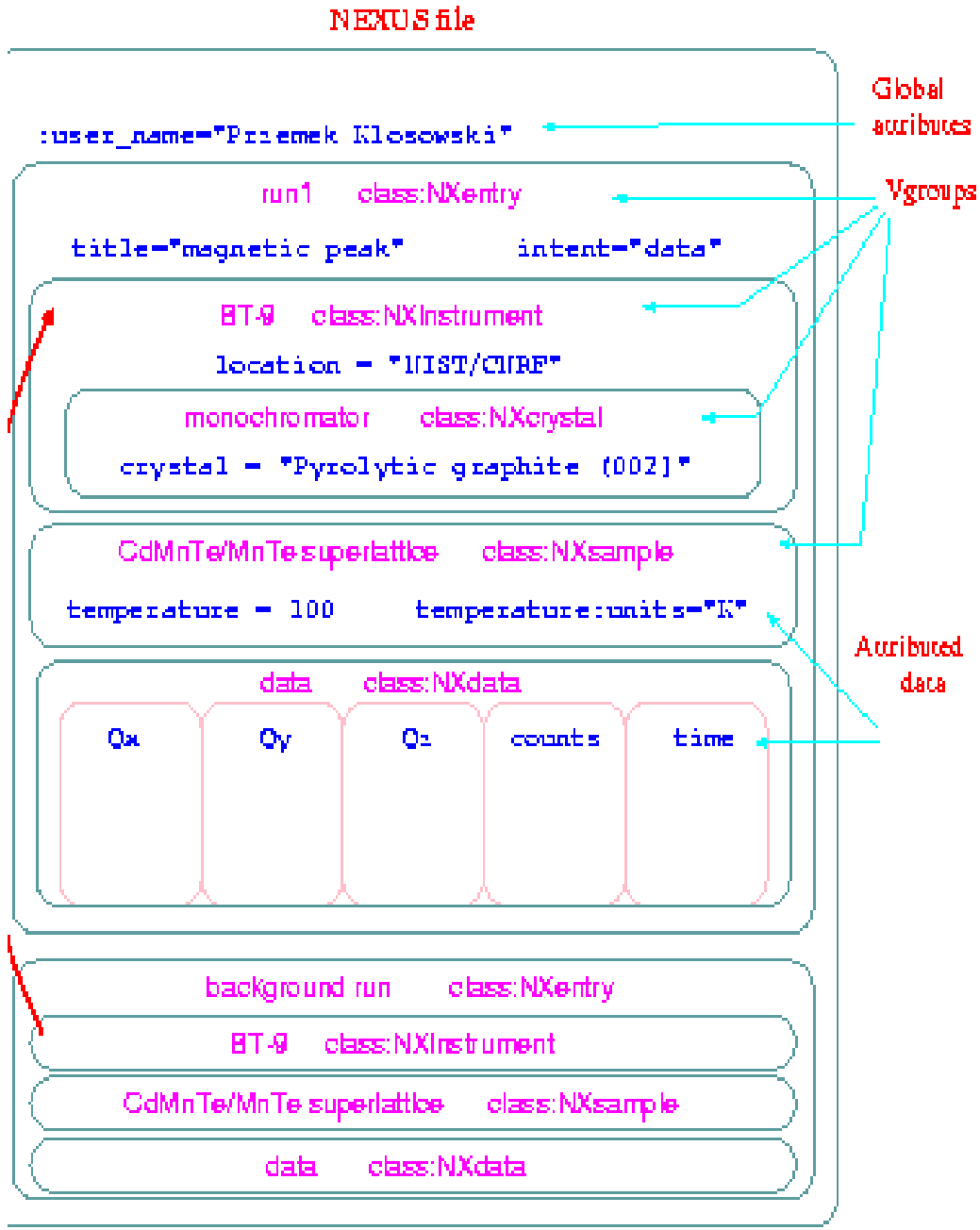
1DF5-4259-0588-E419-6614-A5FC-4177”></xml>

NeXus

- Proposed and developed since 1996 by NIST, PSI/Switzerland, Argonne and Rutherford labs
- Goal: inter-laboratory general data exchange format; initially for neutron/Xray scattering
- Uses NCSA HDF library
  - For large data sets, efficient storage, compression, random access, speed, etc
- Defines domain-specific data structuring
- Already went through a back-end change, due to the HDF4-HDF5 transition
- Possible tie-ins to XML

NeXus details

- Specification of data organization:
  - Global per-file attributes
  - Data units (NXEntry)
    - Descriptive sections (NXUser, NXInstrument)
      - Further subgroups: Nxsource
        - Data: Nxsource:intensity
        - Attribute: units=[c/s]
    - NXData section
- Fundamental data item: a multidimensional table with atrributes (units, comment)
- HDF-5 (or -4) storage+I/O



NeXus and XML

- XML might be a possible storage format
- XML could provide DTD for structure specification and conformance testing
- Many manipulation tools (parsing, transformations, debugging, etc)