Data Management and Distribution - Overview

Large Facilities Workshop

Brian Glendenning

NRAO Data Management and Software Department

Atacama Large Millimeter/submillimeter Array

Karl G. Jansky Very Large Array

Robert C. Byrd Green Bank Telescope

Very Long Baseline Array

Why Data Management?

- No Choice – mandates ratcheting up
  - NSF Data Management Plan requirements, OSTP data access memo
  - What taxpayers pay for should be freely available
- Review survival
  - People like me will write nasty recommendations if you don't have it (*)
- Science Return = Facility self-Interest
  - Data-use multiplier
    - STScI Archive: Data comes out 5x
  - Cost effective!
    - ALMA: ~3% construction (LSST ~25%), ~8% operations

NSF: No more facilities without strong data management plans (*)

You don't want comments like this, even from your token softaware/data management panel member

Data Products

- Expand your user base, provide automatically-generated, widely usable derived data products
- Also provide raw data products (experts will demand it, some analyses will require it, someone needs to build the next generation of instruments)
- NEON: Level 0=raw, 1=simple calibration, 2=gap filled, 3=gridded, 4=derived, multi-source
- Try to define your QA parameters up front (someday someone wll do this…)
- What to do about PI generated data products (make available through facility? Leave to PI Data Management Plan?)
- Data should be transportable to other software systems, and the future
  - Define using file formats, not APIs (data outlives software)

Data Volume

- Drown in data later, get the system to work now
- Pareto's Principle works for data as well; more data = more science, but not linear
- Moore's law is your friend
  - Easier every year ($100 becomes $0.10 in 15 years); or
  - More data/processing every year for fixed budget
    - ALMA: Average data rate can increase by 100x
  - Algorithms usually get more expensive (FLOPS/IO) with time

Other

- Open Source facility software – under threat?
  - In astronomy has been a tremendous benefit in last 15+ years
  - Pressure to monetize, gain advantage over competitors
- Allow anonymous access to (non-proprietary) data, even if it gets in the way of metrics
  - EarthScope iPad application example
- Many "pure IT" issues: Reliability/availability, Disaster recovery, Backup, Security, Privacy (do you really want to be a data center?)
- Data/Computing facilities: in-house vs. center vs. cloud decisions, cost/benefits not being systematically (re)considered
  - Role for NSF to facilitate?
- End of Life – will your data outlive your facility? Vice versa?

Questions

- What should NSF Facility "Data Management" best practices be?
  - Can/should this be formalized?
- How do we keep data management systems in construction project scope?
  - Often thrown out to obtain only modest cost savings
  - Construction projects often dominated by grizzled veterans
    - Data Management = chart recorder + HP-11C
- Can the various national HPC centers/networks play more of a role?
  - Gap: big-ish data problems, hard for facility but not interesting for HPC research
- What metrics should we use?