How Do We (the National Research Community) Sustain Data Beyond the End of a Grant?

Pat Burns

Dean of CSU Libraries and VP for IT

March 24, 2011

Framework

- "*Education is the path from cocky ignorance to miserable uncertainty*." - by *Twain*, Mark
- Uncertainty reigns
  - What will institutions be required to provide for DM?
  - What will NSF provide for DM?
    - Incidental, one-time funding from grants?
    - Infrastructure?
      - Institutional
      - NSF infrastructure

NSF Cyberinfrastructure Model

- For HPC Infrastructure
  - Tier 1: national resource (e.g. Blue Waters at NCSA)
  - Tier 2: several regional resources, shifting toward disciplinary activities (TACC, ORNL, PSC?)
  - Tier 3: campus level (here and at other campuses)
- For DM Infrastructure
  - Uncertainty reigns
  - Likely that the three-tier model will be extended to DM
  - Partly why we established the ‘small, medium, and large’ approach (could be tiers 1, 2, and 3)

NSF DRAFT Report

- NSF Advanced Computing and Communications Task Force on Campus Bridging, Summary report draft ver. 4.1, Mar. 1, 2011

"Finding 6: New instrumentation (including that installed at the campus lab level) is producing volumes of data that cannot be supported by most current campus networking facilities."

Cont'd

- *"Strategic Recommendation to the NSF #4: The NSF should fund national facilities for at least short-term storage and management of data to support collaboration, scientific workflows, and remote visualization; management tools should include support for provenance and metadata. As a complement to these facilities and in coordination with the work in Recommendation #3, NSF should also fund the development of services for bulk movement of scientific data and for high-speed access to distributed data stores. Additionally, efforts in this area should be closely coordinated with emerging campus-level data management investments."*

Cost Model

- “Facts are stubborn things; and whatever may be our wishes, our inclinations, or the dictates of our passion, they cannot alter the state of facts and evidence.” **John Adams**, *'Argument in Defense of the Soldiers in the Boston Massacre Trials,' December 1770*
- Requirements for DM infrastructure
  - What, when, where, how, how much?
  - Cost?
  - Funding source(s)

Life Cycle Cost

- Just raw storage
  - $10,000/90 TB = $111/TB
- For storage & back-up x 2
  - $222/TB (minimum necessary)
- For storage, back-up & preservation x 5
  - $555/TB
- Life cycle for hardware (optimistic) = 5 years
  - How will the cost/TB decrease over time?

Costs vs. Term of Preservation (Costs are for back-end hardware only!!!)

	0-5 yrs.	5-10 yrs.	10-15 yrs.	15-20 yrs.
Storage and back-up only	**$222/TB**	**$444/TB**	**$666/TB**	**$888/TB**
Above plus preservation	**$555/TB**	**$1,110/TB**	**$1,665/TB**	**$2,220/TB**

Or A Blank Check?

DISCUSSION

How big are data sets, in aggregate, for a project?

For how long must data sets be accessible?

- Gazintas? I.e. Deposit

- Gazoutas? I.e. Withdrawal, annually?