Chronopolis™:

Federated Digital Preservation Environment Using Data Grid Technologies

2006 PRDLA/PNC/ECAI Meeting
Seoul, Korea

Brian E. C. Schottlaender
University Librarian, UCSD

PREVIEW

- Pyramids
- “Slow Rot”
- Choices
- Decisions
- Chronopolis™

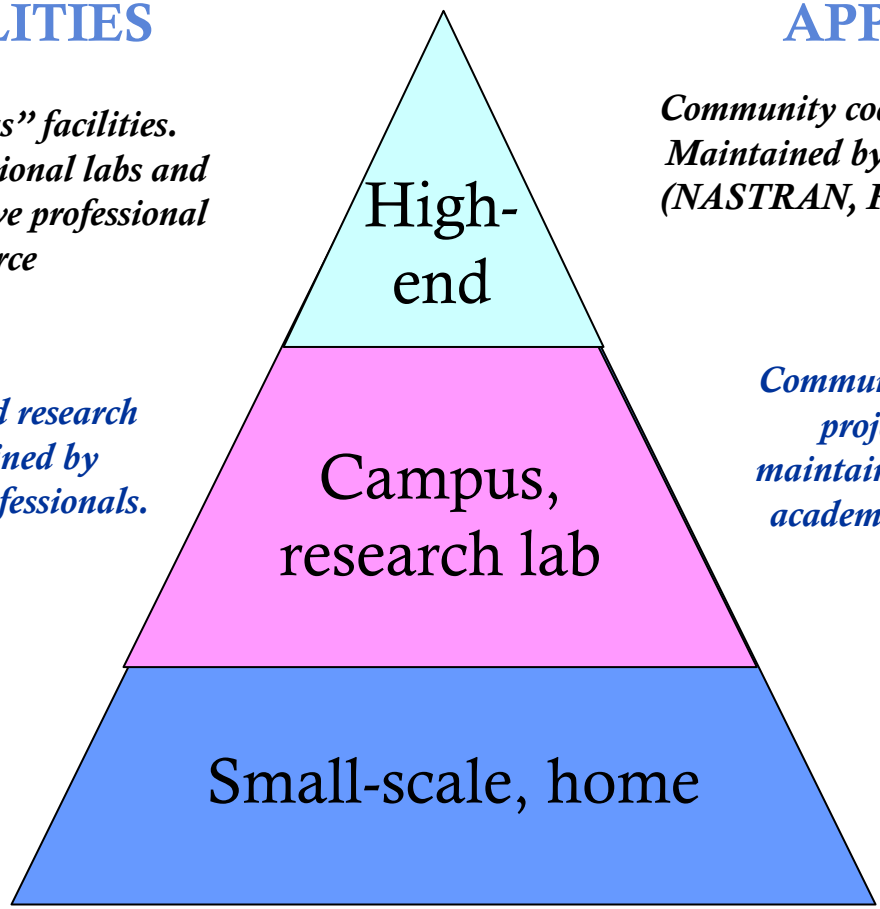
The Branscomb* Pyramid for Computing

FACILITIES

"Leadership-class" facilities. Maintained by national labs and centers. Substantive professional workforce

Mid-range university and research lab facilities. Maintained by professionals and non-professionals.

Private, home, and personal facilities. Supported by users or their proxies.



APPLICATIONS

Community codes and professional software. Maintained by large groups of professionals (NASTRAN, Powerpoint, WRF, Everquest)

Community software and highly-used project codes. Developed and maintained by some professionals and academics (CHARMM, GAMESS, etc.)

Research and individual codes. Supported by developers or their proxies.

*Chairman, NSF Blue-Ribbon Panel on High Performance Computing (1993)

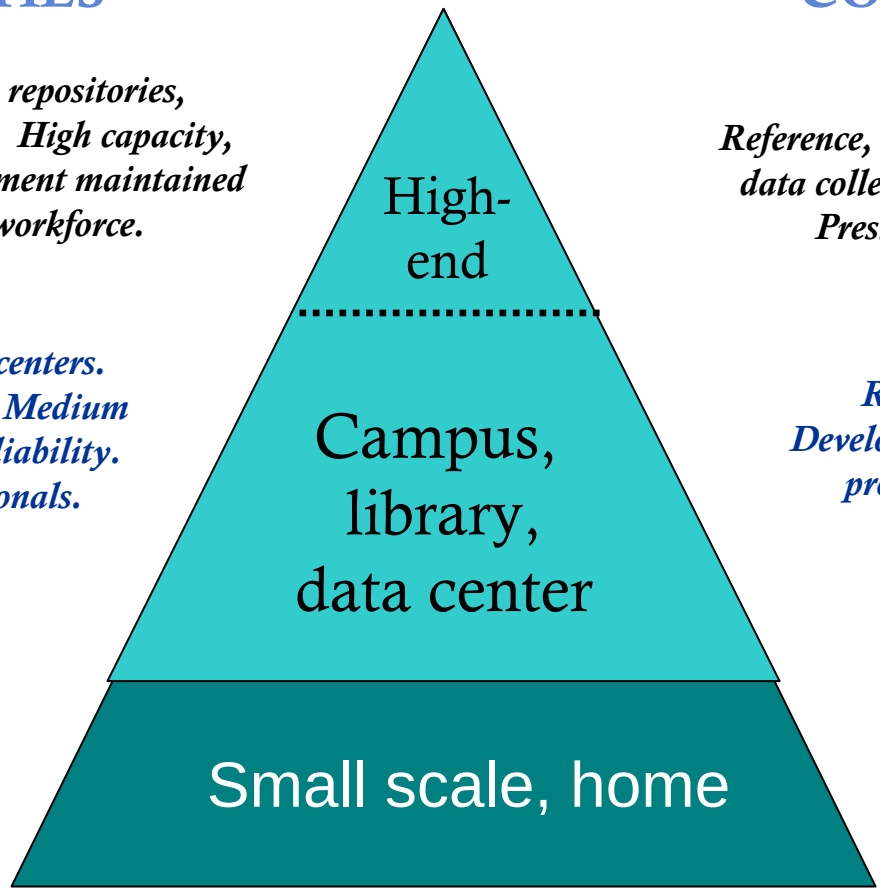
The Berman* Pyramid for Data

FACILITIES

National-scale data repositories, archives, and libraries. High capacity, high reliability environment maintained by professional workforce.

Local libraries and data centers. Commercial data storage. Medium capacity, medium-high reliability. Maintained by professionals.

Private repository. Supported by users or their proxies. Low-medium reliability, low capacity.



COLLECTIONS

Reference, important, and irreplaceable data collections PDB, PSID, Shoah, Presidential Libraries, etc.

Research data collections. Developed and maintained by some professionals and academics

Personal data collections. Supported by developers or their proxies.

*Director, San Diego Supercomputer Center

Tick … Tick … Tick … Tick … Tick

- There is a pressing need to preserve digital assets that represent the intellectual capital of scientific disciplines, educational communities, and government and cultural agencies.
- Many of these assets are increasingly at risk, whether as a consequence of:
  - lack of financial support;
  - technology evolution of storage and delivery systems, access mechanisms, or encoding formats; or,
  - calamity
  - neglect.

ISSUE: Frailty

- Dynamic:
  - May be revised or updated → instances, versions, editions
  - May change cumulatively or interactively → e.g., contributions to a listserv
  - May be available in various "views"
- More easily altered [without recognition]
- More easily corrupted
- Storage media have shorter life spans

ISSUE: Complexity

- Linkages between and amongst them may change
- Increasingly data and associated metadata cannot, or should not, be separated
- Some resources, like multimedia, are so closely linked to the software and hardware technologies that they cannot be used outside those proprietary environments
- Need to be “renderable” on a variety of delivery devices
- Require access technologies that are changing at an ever-increasing pace

ISSUE: Selection

- **Intellectual** question → What is "worth" archiving?
  - Scientific content (e.g., PDB)
  - Scholarly content (e.g., *Electronic Cultural Atlas*)
  - Cultural content (e.g., Shoah)
  - "Official" content (e.g., Govt. docs.)
- **Physical** question → What is the 'archival unit?'
  - What is its extent?
  - What are its boundaries?
    - Links?
    - Content of links?
- Intellectual and physical selection dimensions are not separate, but interrelated. E.g., determination of extent of digital object is necessary before harvest-based selection can take place.
- Selection criteria cannot be generalized because they are dependent on the goals and policies of the particular stakeholder.

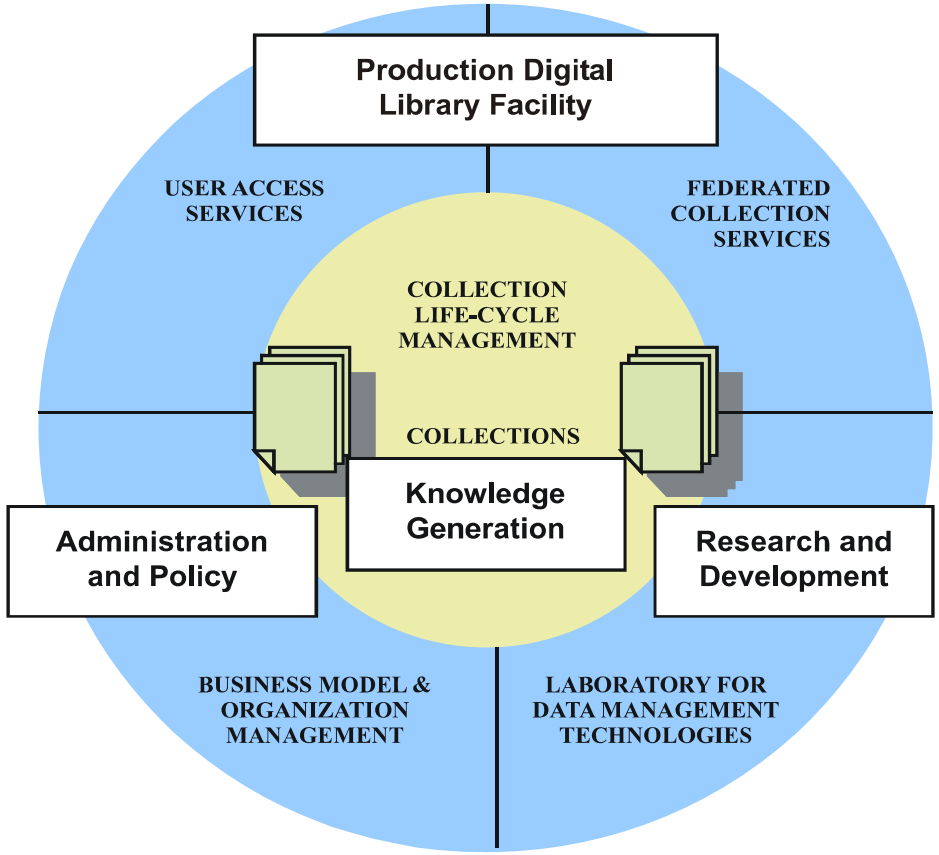
Questions, Questions, Questions

- Who gets to decide what's worth preserving?
- Who's responsible for preserving it?
  - Where?
  - How?
  - For how long?
- Who gets access?
  - Why?
  - When?
- Who pays?
  - Content creators?
  - Content users?
  - The government?

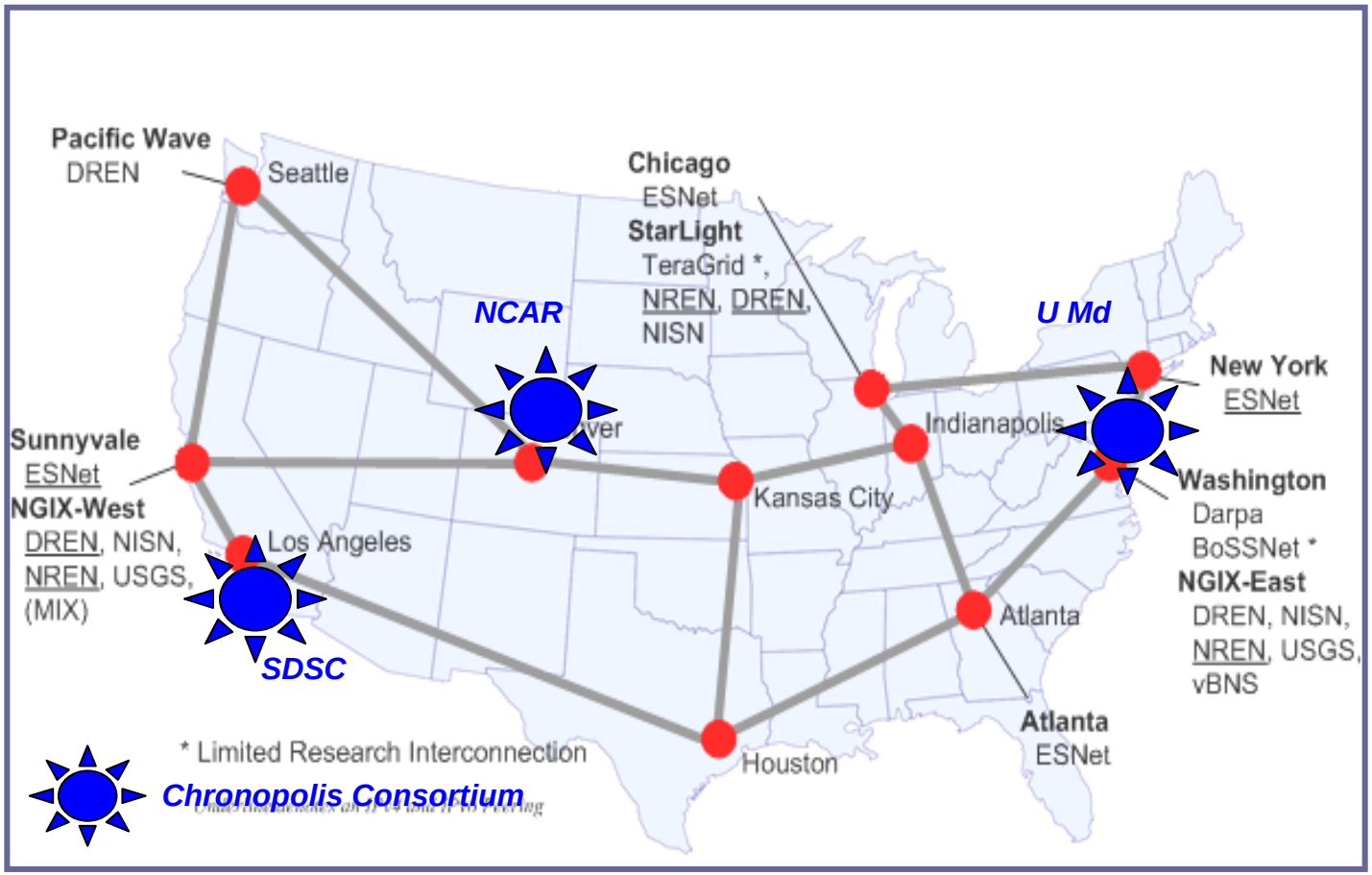
CHRONOPOLIS™

- National **center** for the management, long-term preservation, and promulgation of digital assets.
- Model **facility** for long-term support of collections, ensuring that:
  - Standard reference datasets remain available;
  - Collections can expand and evolve over time, as well as weather evolution in the underlying technologies; and
  - Preservation of "last resort" is available for critical "at risk" resources.
- **Tools**, **software**, and **services** needed to manage data, information, and knowledge at the scales required for national digital holdings.
- Distributed national **data backbone** that federates data and information (preservation across space) and that provides operational data services for maintaining key digital collections for the long term (preservation across time).

CHRONOPOLIS™: Conceptual Architecture



CHRONOPOLIS™: Federation Architecture



CHRONOPOLIS™: Replication and Distribution

- 3 replicas of valuable collections considered reasonable mitigation for risk of data loss.
- Chronopolis Consortium will store 3 copies of preservation collections:
  - "Bright copy": Chronopolis site supports ingest, collection management, user access.
  - "Dim copy": Chronopolis site supports remote replica of bright copy and user access.
  - "Dark copy": Chronopolis site supports reference copy that may be used for disaster recovery, but no user access.
- Each site may play different roles for different collections.

CHRONOPOLIS™: Users, Partners, Providers

- Chronopolis "**Users**:" will utilize the Chronopolis environment and services for data management and preservation of their collections.
- Chronopolis "**Partners**:" will support the installation of servers (e.g. SRB, DSpace, or Fedora) at their sites, register their collections into Chronopolis, and use the Chronopolis environment to replicate their collections.
- Chronopolis "**Providers**:" will constitute the federated Chronopolis environment, including deploying distributed storage infrastructure at their sites and working as a team to develop and support preservation tools and services.