ACCESS 2007
TECHTONIC ONTOLOGIES
Victoria, British Columbia • October 10-13

DATA CYBERINFRASTRUCTURE COLLABORATION at the University of California, San Diego

Luc Declerck

AUL, Technology Services

Declan Fleming

Director, Information Technology Department

SDSC

SAN DIEGO SUPERCOMPUTER CENTER

UC SAN DIEGO LIBRARIES

UC San Diego LIBRARIES

Outline

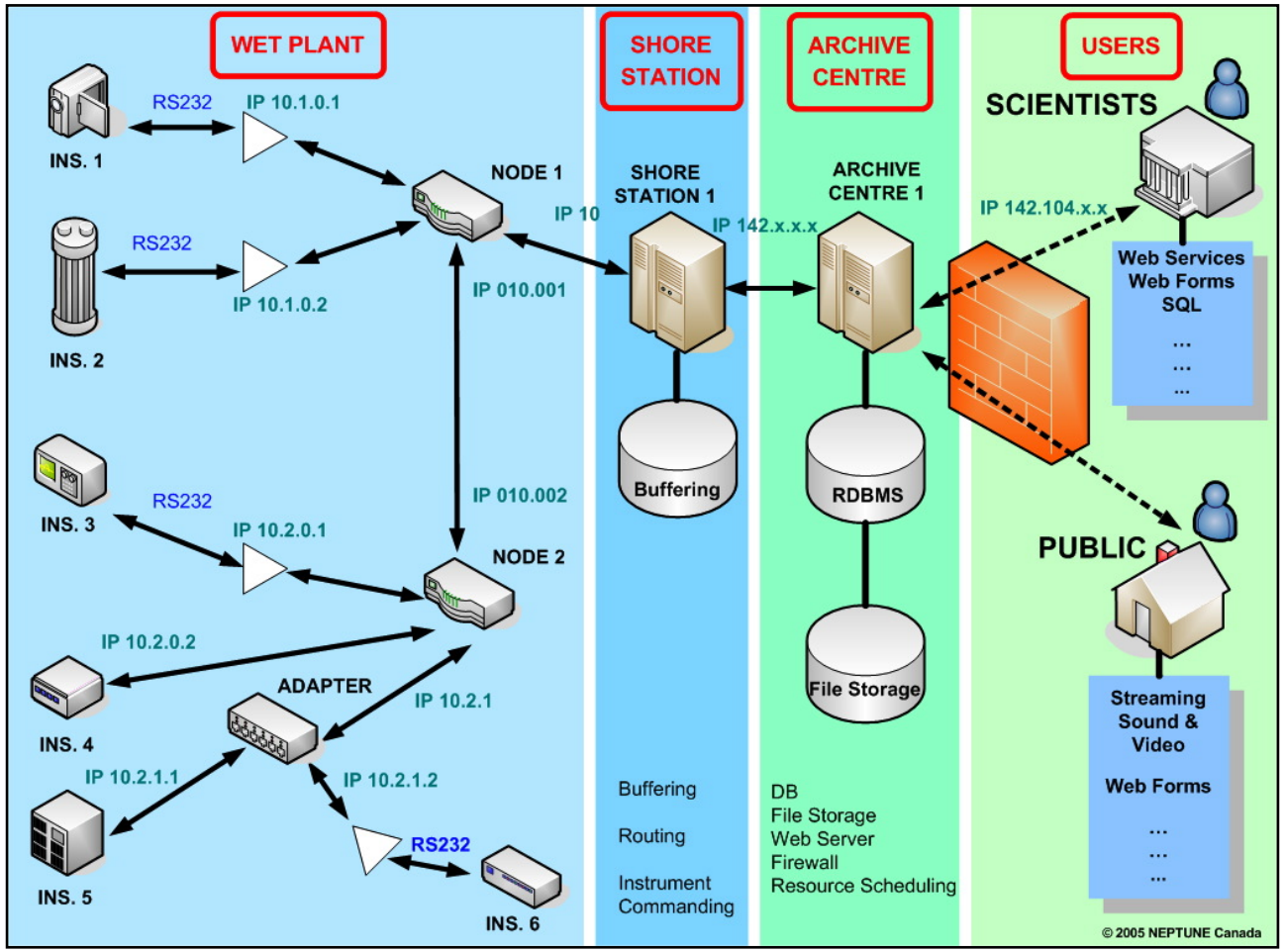
- What is cyberinfrastructure?
- Examples of cyberinfrastructure
- Why is this relevant to Libraries?
- The UC San Diego Libraries' response
- Lessons Learned
- The technology at play at national, system, and local levels
- Future plans

What is cyberinfrastructure?

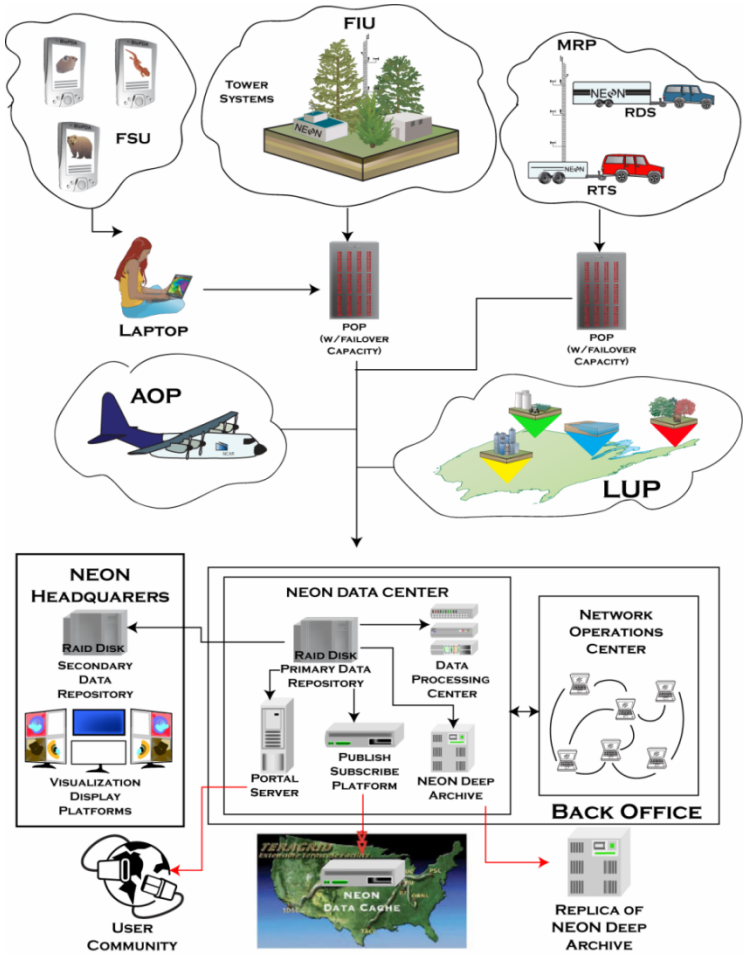
Cyberinfrastrucre is the coordinated aggregate of software, hardware and other technologies, as well as human expertise, required to support current and future discoveries in science and engineering.

Fran Berman, Director of the San Diego Supercomputer Center (SDSC)

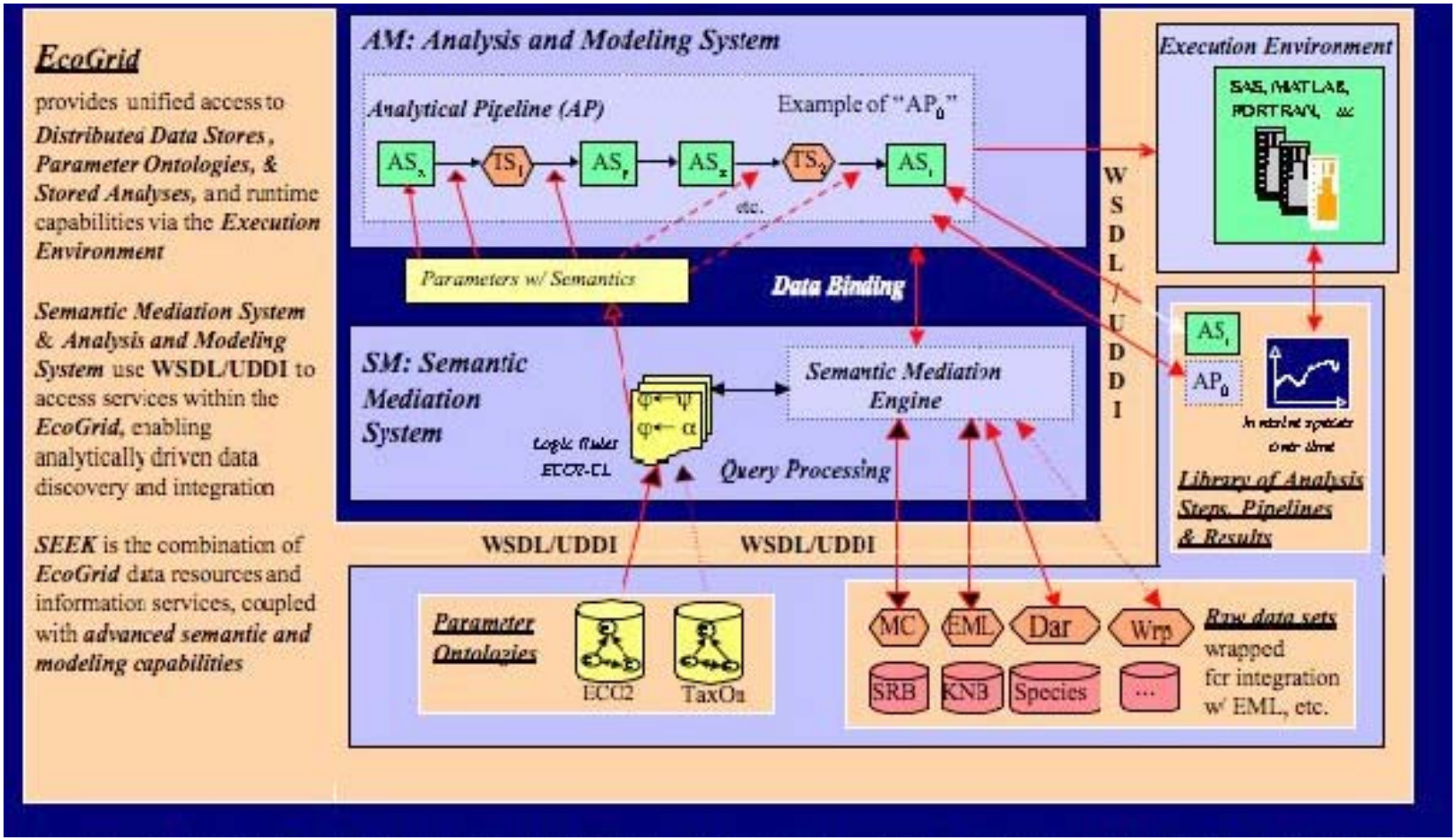
Neptune Canada



National Ecological Observatory Network (Neon)



SEEK EcoGrid (ecological, biodiversity, and geological sciences)



Common Characteristics

Data

Lots of data

Petabytes of data

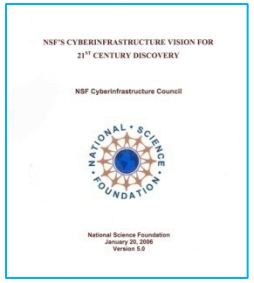
Examples

Neptune → 50 Tbytes per year

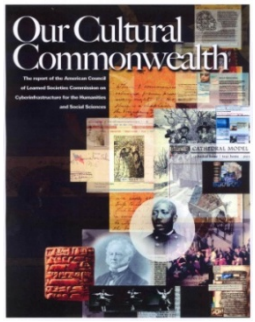
Astronomy → 40 Tbytes every 3 days

CDL → ramping up to 40 Tbytes

New Research Paradigm



NSF's Cyberinfrastructure Vision for 21st Century Discovery



Our Cultural Commonwealth: The Report of the American Council of Learned Societies' Commission (ACLS) on Cyberinfrastructure for the Humanities and Social Sciences

National Consultation on Access to Scientific Research Data (NCASRD): Final Report

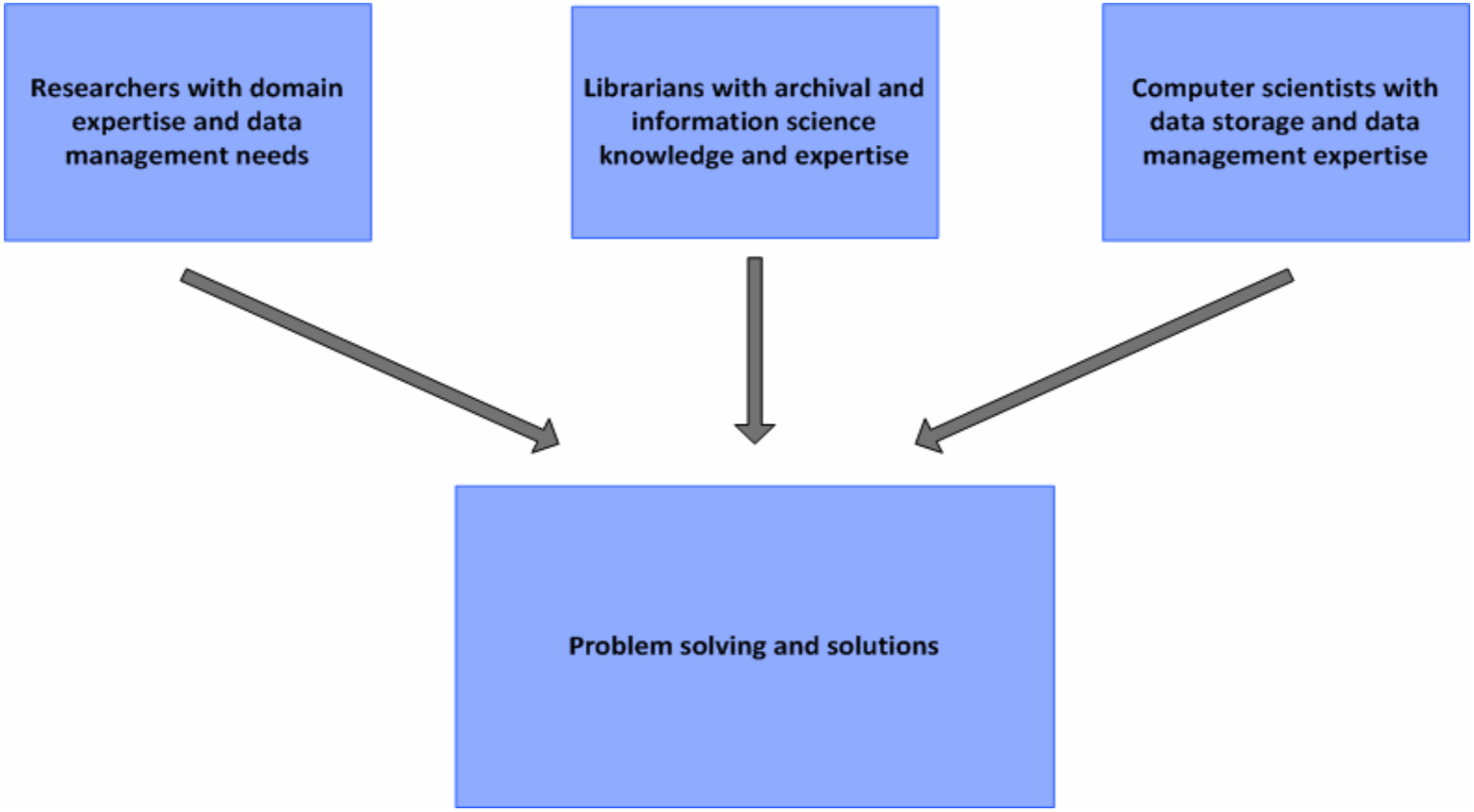
Phenomenon is Across Disciplines

- **Big science**
  - Neptune Canada, Neon, NEES, GEOSS, PDB, BIRN, HIS
- **Social science**
  - ICPSR datasets, local researcher datasets, surveys
- **Arts and Humanities**
  - Maurizio Seracini's x-ray collection of art masters, UCSD TV videos
- **Cultural institutions**
  - Library mass digitization projects (Google, MS, OCA), web crawls, local digitization activities

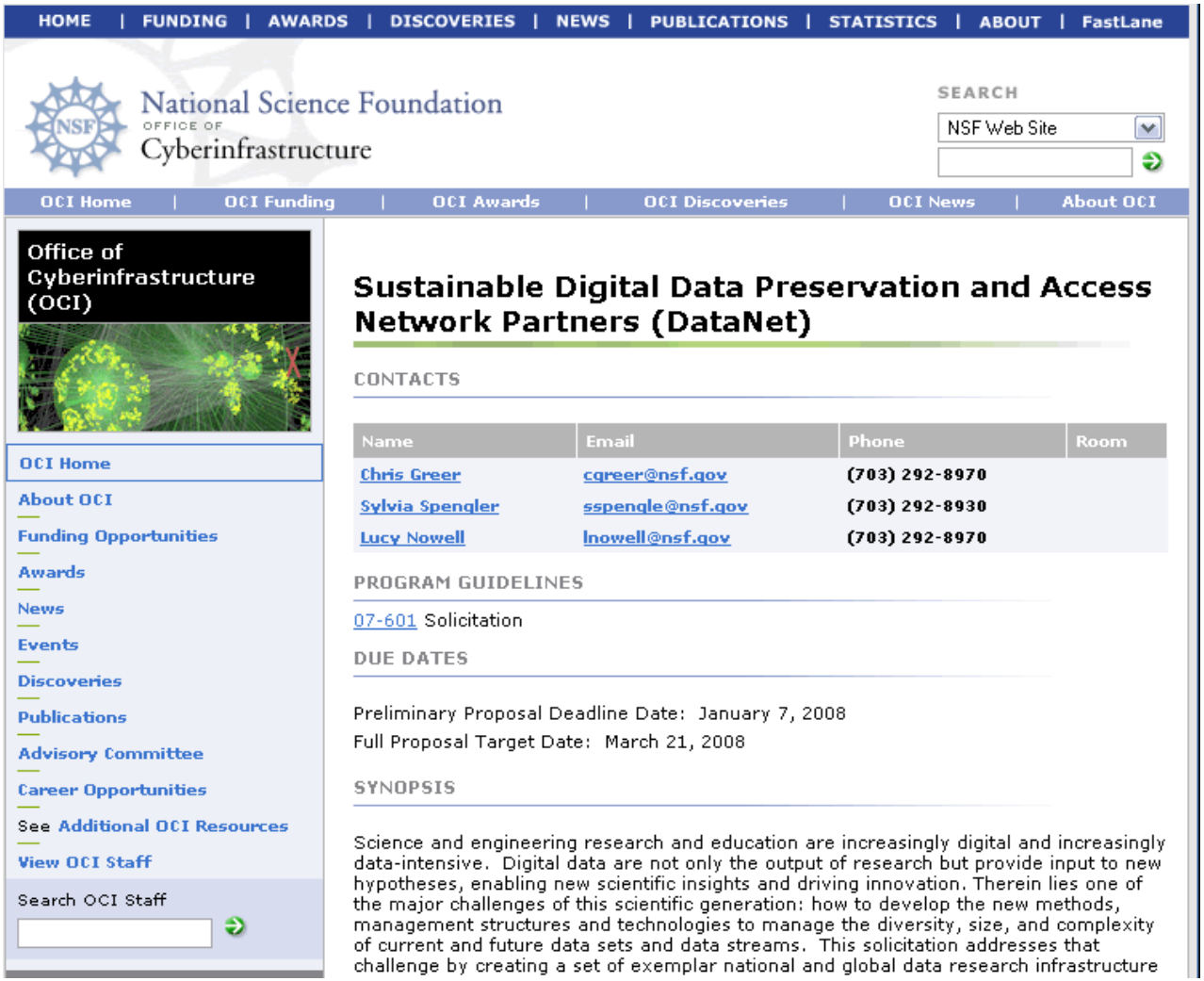
Urgent need for

- **Large-scale digital preservation infrastructure**
- **Informed (metadata/ontology-based) discovery of and access to data**
- **Links between the data and its research output**
- **Tools and services**
  - Data integration
  - Data mining
  - Data visualization

Urgent Need for Collaboration



Explicit recognition



HOME | FUNDING | AWARDS | DISCOVERIES | NEWS | PUBLICATIONS | STATISTICS | ABOUT | FastLane

National Science Foundation
OFFICE OF
Cyberinfrastructure

SEARCH
NSF Web Site

OCI Home | OCI Funding | OCI Awards | OCI Discoveries | OCI News | About OCI

Office of Cyberinfrastructure (OCI)

OCI Home
About OCI
Funding Opportunities
Awards
News
Events
Discoveries
Publications
Advisory Committee
Career Opportunities
See Additional OCI Resources
View OCI Staff

Search OCI Staff

Sustainable Digital Data Preservation and Access Network Partners (DataNet)

CONTACTS

Name	Email	Phone	Room
Chris Greer	cgreer@nsf.gov	(703) 292-8970	
Sylvia Spengler	sspengle@nsf.gov	(703) 292-8930	
Lucy Nowell	lnowell@nsf.gov	(703) 292-8970	

PROGRAM GUIDELINES

07-601 Solicitation

DUE DATES

Preliminary Proposal Deadline Date: January 7, 2008
Full Proposal Target Date: March 21, 2008

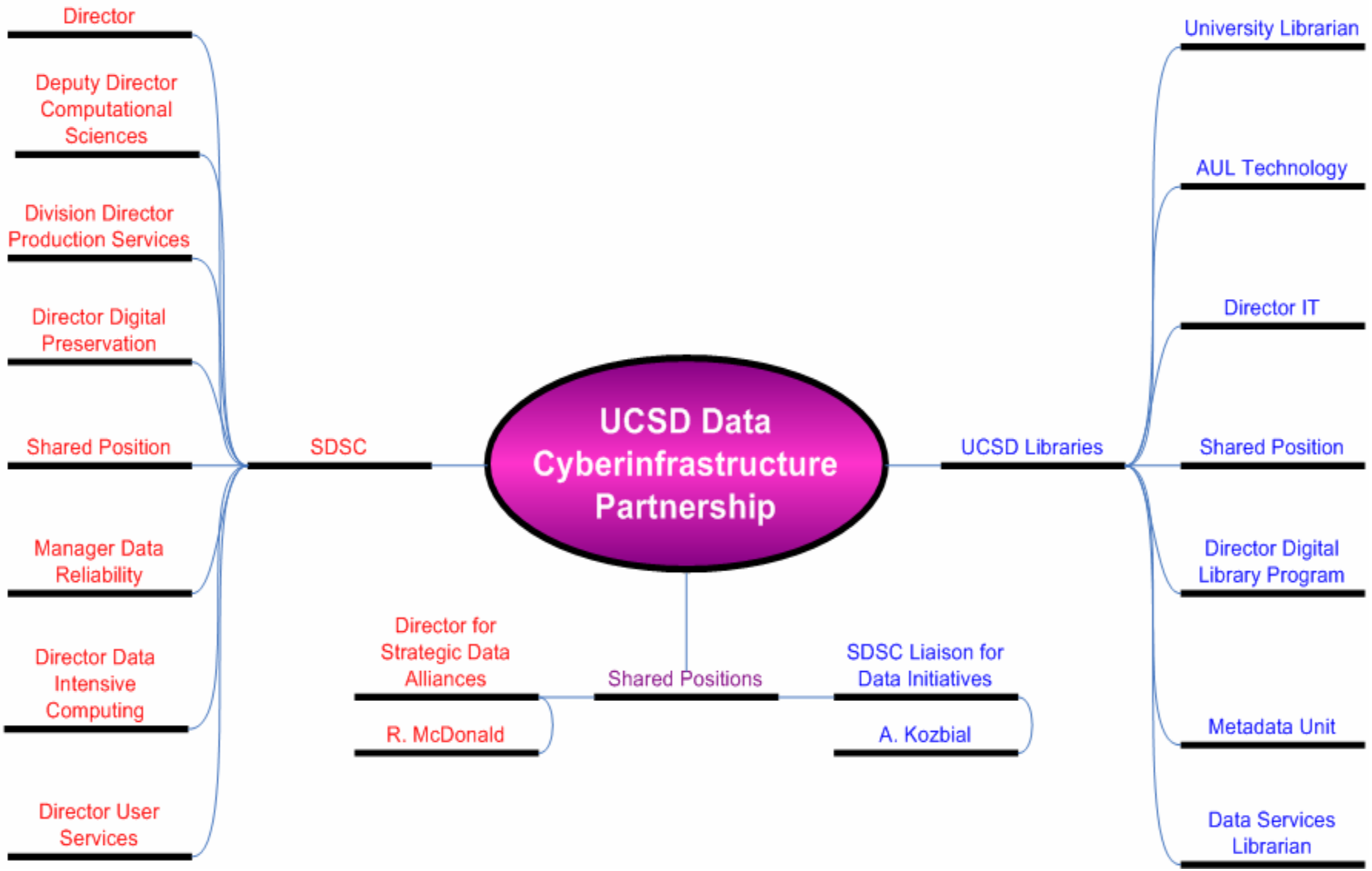
SYNOPSIS

Science and engineering research and education are increasingly digital and increasingly data-intensive. Digital data are not only the output of research but provide input to new hypotheses, enabling new scientific insights and driving innovation. Therein lies one of the major challenges of this scientific generation: how to develop the new methods, management structures and technologies to manage the diversity, size, and complexity of current and future data sets and data streams. This solicitation addresses that challenge by creating a set of exemplar national and global data research infrastructure

… new types or organizations … [that] … will integrate library and archival sciences, cyberinfrastructure, computer and information sciences, and domain science expertise …

– DataNet Program Solicitation NSF 07-601

At UCSD


UCSD Data Cyberinfrastructure Partnership
SDSC
Director
Deputy Director Computational Sciences
Division Director Production Services
Director Digital Preservation
Shared Position
Manager Data Reliability
Director Data Intensive Computing
Director User Services
UCSD Libraries
University Librarian
AUL Technology
Director IT
Shared Position
Director Digital Library Program
Metadata Unit
Data Services Librarian
Shared Positions
Director for Strategic Data Alliances
R. McDonald
SDSC Liaison for Data Initiatives
A. Kozbial

The Gretzky Rule

"Skate to where the puck will be"

- **Decided to focus on the 2nd word in "unfunded mandate," rather than on the 1st**
- **Developed an intentional relationship with SDSC (where the puck will be)**

The Gretzky Rule:
"Skate to where the puck will be"

Collaborative Projects to-date

- **Preservation infrastructure**
  - **DAMS, UC Grid, Chronopolis**
- **Collection ingest**
  - **UCSD/TV videos, LC image collections, web archives**
- **Interdisciplinary Data Integration**
  - **Neuroscience/Architecture databases**
- **Data Mining and Visualization**
  - **CalCOFI database (60 years of fish data)**

Competencies Leveraged

Faculty	Libraries	SDSC
❑ Domain expertise	❑ Archiving	❑ Grid storage
❑ Data collection	❑ Metadata management	❑ Grid services
❑ Taxonomies	❑ Discovery-tool building	❑ Data management
❑ Ontologies	❑ Culture of service	❑ Data preservation
❑ Data mining	❑ Culture of trust	❑ Format migration
❑ Data reuse	❑ Project Management	

What Have We Learned?

- We do indeed need each other
- Libraries bring a lot to the table
- Substantial organizational differences
- New organizational structure would help

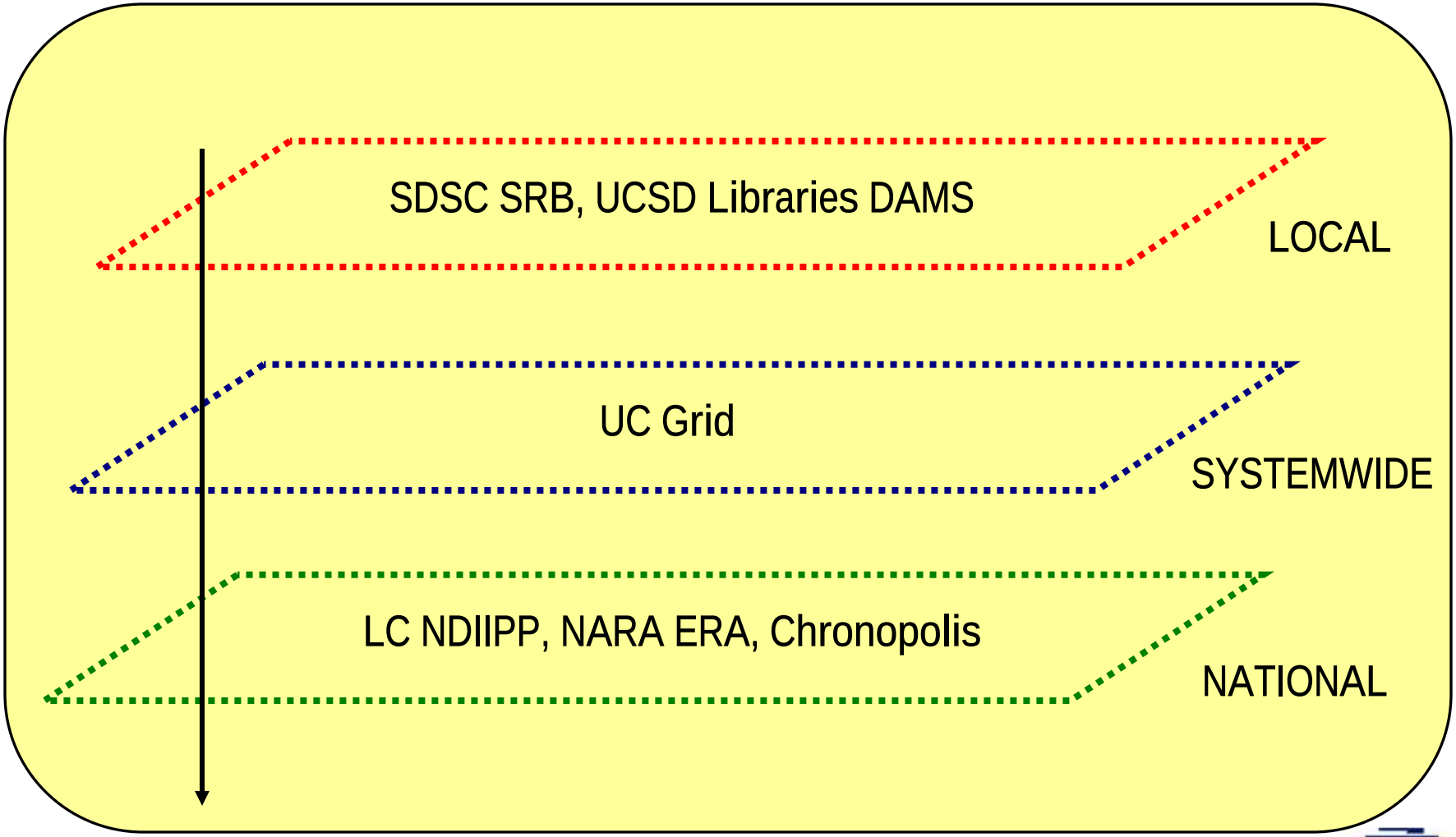
What Libraries Bring to the Table

- **Significant expertise**
  - Metadata
  - Archival management
  - Policy development
- **Organizational experience and stability**
  - Process- and Results-driven
- **Culture of trust**
  - Responsible guardians of cultural record
  - Service oriented
  - Respectful of privacy and intellectual property

What Libraries bring to the table Another view

- **Data acquisition, ingest layer**
  - Selection, taxonomy, ontology, metadata, workflow
- **Preservation layer**
  - Archival retention, format migration, QA, trust
- **Physical layer**
  - Storage, network, security, reliability standards
- **Service layer**
  - Discovery, retrieval, data mining, data visualization
- **Management layer**
  - Administration, budget, policy development

Layers of Technology Collaboration



At National Level: Chronopolis Digital Preservation Program

- **Collaborative Initiative**
  - San Diego Supercomputer Center
  - University of California, San Diego Libraries
  - National Center for Atmospheric Research
  - University of Maryland, Inst. for Adv. Computer Studies
- **Long Term Digital Management and Preservation**
  - National center
  - Latest in storage technologies
  - Grid-enabled Cyberinfrastructure
  - Operational data services
- **Research**

Chronopolis Locations

The Chronopolis demonstration data grid is composed of three geographically distributed Chronopolis provider sites.

UMIACS

NCAR

Chronopolis Digital Preservation Data Grid

Administration for Policy and Outreach

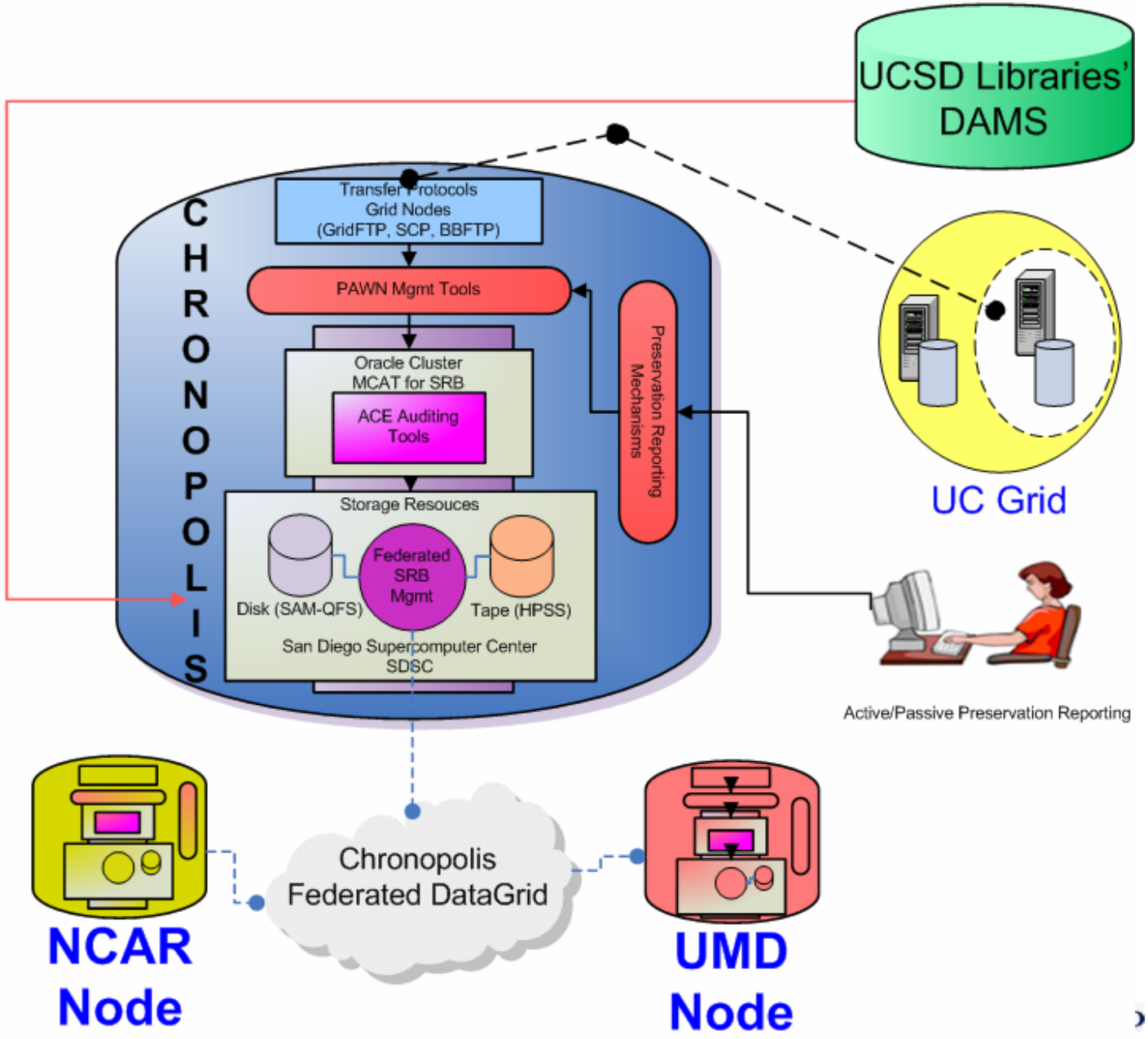
(Supports the overall partnerships and mgmt for preservation services and works as a liaison with Chroonpolis partners and other regional and national preservation programs)

Research and Development

(Research and development for rules-based preservation mgmt and technology forecasting for continual technology migration and mgmt)

Production Digital Preservation

(Long-term preservation with geographic replications and preservation services)



Chronopolis Research Areas

- **Preservation Environment**
  - Rules-Based Preservation Management
  - Content Transfer from Multiple Preservation Environments
    - Grid Federation
  - Grid-Based Storage Technologies
- **Administration, Policy, Outreach**
  - Formalized Trust Relationships
  - Sustainability Issues
  - Cost Benchmarks
  - Training
- **R&D**
  - Grid-Based Storage Technologies
    - SRB
    - iRODS
  - Rules Based Content Migration/Emulation

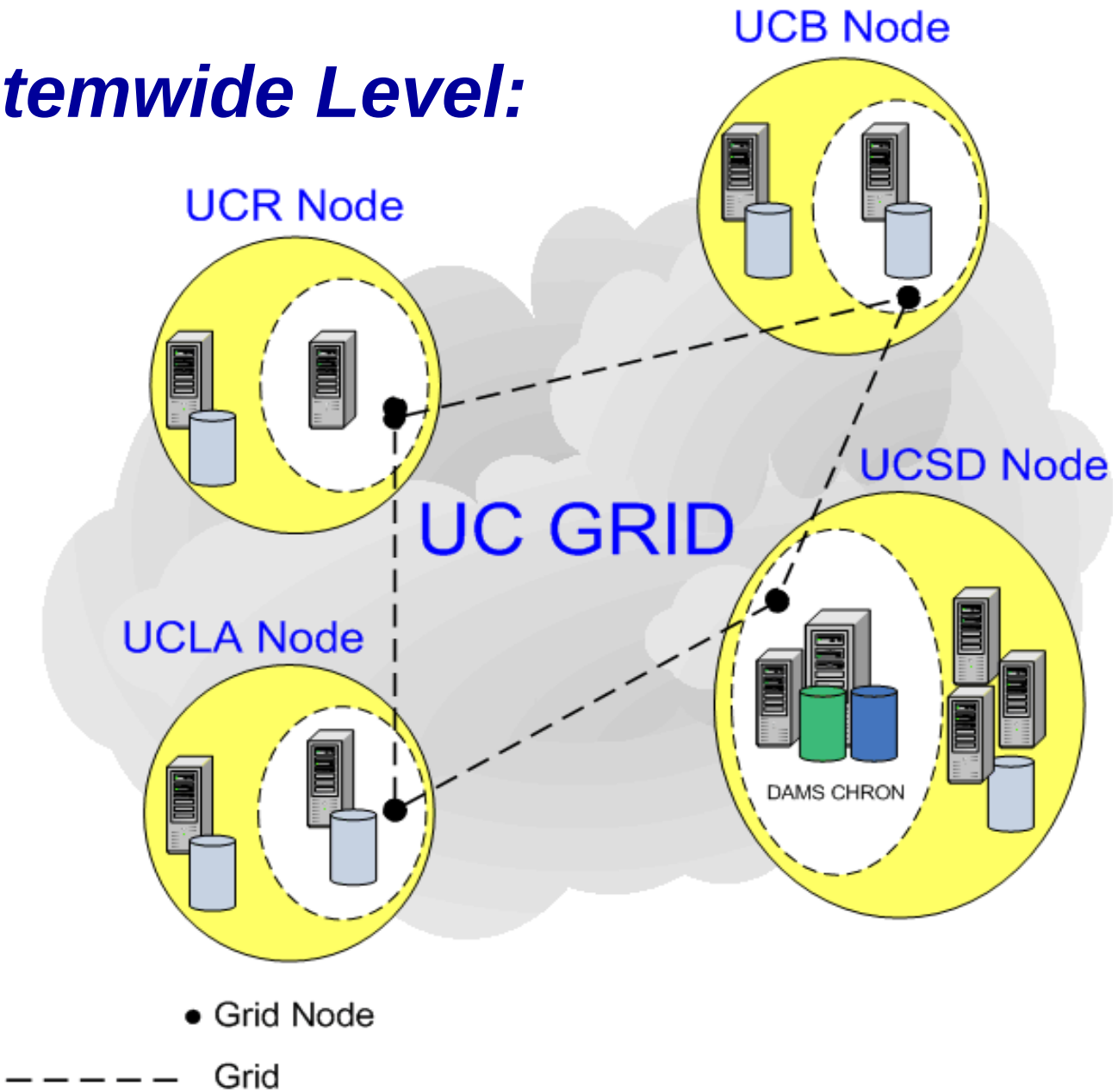
Chronopolis Collections

- **National Virtual Observatory (NVO)**
  - Currently 1 TB of Digital Palomar Observatory Sky Survey
- **Interuniversity Consortium for Political and Social Science Research (ICPSR)**
  - Currently 2 TB of Web-Based Data
  - Future plans include 10 TB of all ICPSR Data Collections
- **California Digital Library (CDL)**
  - Future Plans include 25 TB of Web-at-Risk Crawl Collections
- **Library of Congress (LC)**
  - Currently 2 TB of Prokudin-Gorskii Image Collections

At Systemwide Level: UC Grid

- **Working on Physical Connectivity**
  - 10 Gb among UC Campuses
- **UC Trust**
  - Shibboleth
  - Single Sign-On
- **Data Grid**
  - Google/OCA/Microsoft Books Project w/CDL
  - Mass Transit – data transfer between UC nodes
- **High Performance Computing**
  - Shared resources among UC campuses

At Systemwide Level:



At Local Level: SRB and DAMS

- **Collection Identified**
- **Metadata Services Unit Creates Assembly Plan**
  - Maps data to MODS, PREMIS, MIX, Local Schemas
- **Collection Ingested with JETL (Java Extraction, Transformation, and Loading) Tool**
  - Original digital object
    - Assigned a unique, permanent identifier - ARK
    - Stored in SRB
  - Technical metadata extracted with JHOVE
    - Stored in SRB in under the same ARK
  - Metadata ETL'd and stored in the SRB under the same ARK

DAMS Technical Overview

- **Front End**
  - JavaScript and HTML
  - JSON
- **Back End**
  - Clustered Tomcat Servers
  - XML
  - XSL/Style Sheets
  - Lucene
  - Oracle
- **Storage**
  - Storage Resource Broker (SRB)
  - SMB/CIFS
- **Data Model**
  - RDF
- **Supported Standards**
  - MODS
  - METS
  - MIX
  - PREMIS
  - Extendable to others

UCSD Libraries' Digital Asset Management System

Public Access

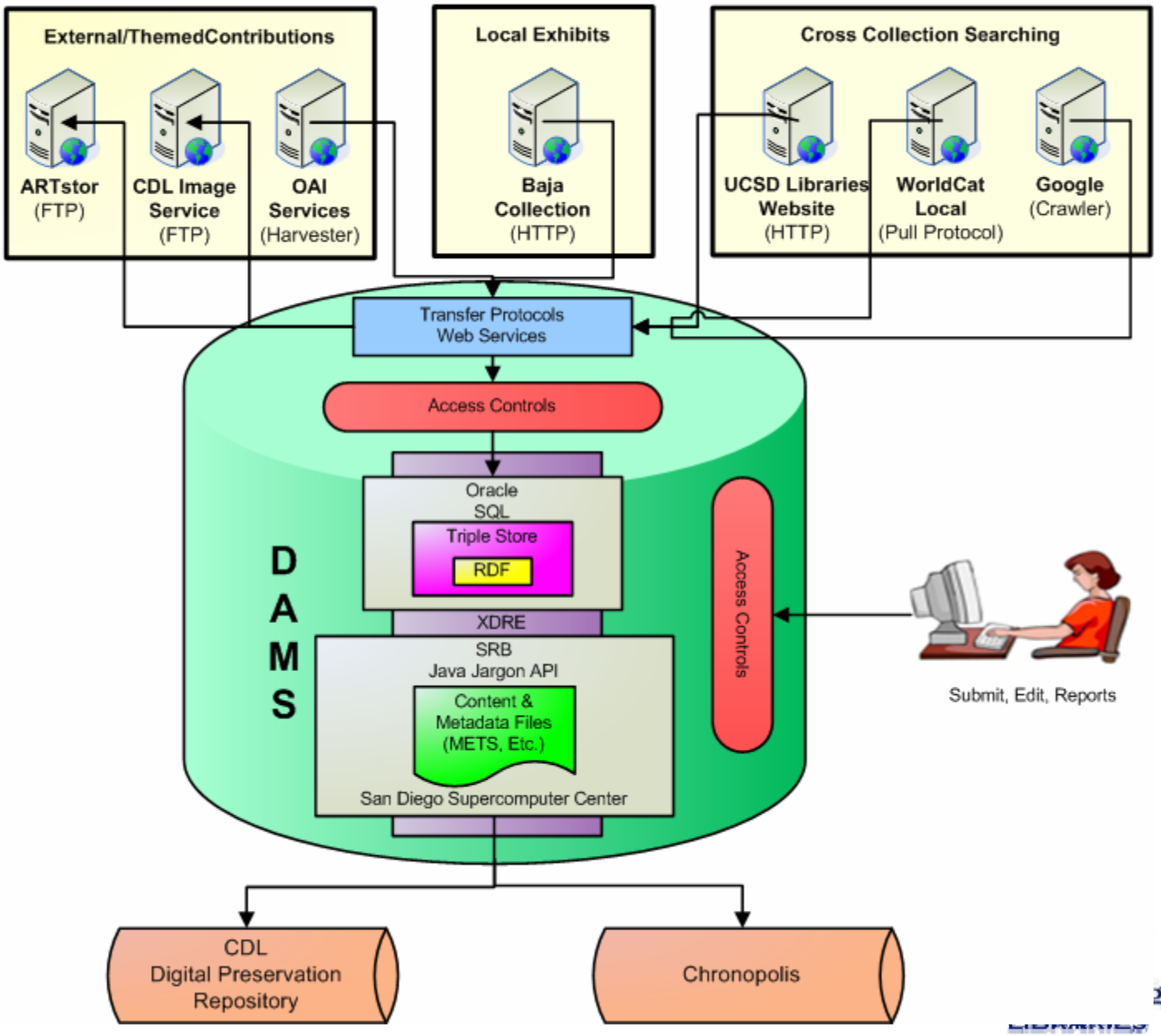
(Supports the discovery, retrieval, use, and reuse of the UCSD digital assets)

Management

(Staff interface and services for ingest, storage, management, and METS record creation)

Preservation

(Long-term preservation at remote facilities)



UCSD Libraries Digital Asset Management System (DAMS) beta

Home About Help Feedback Log out



Search

search Lucene SQL

Advanced Search

Browse by Collection

- UCSD Libraries
  - Arts Libraries
    - Film and Video Collection
    - Visual Resource Collection
  - Electronic Theses and Dissertations
  - Mandeville Special Collections Libraries
    - Baja California Collection
    - Dr. Seuss Went To War Collection
    - Southworth Spanish Civil War Collections
      - Posters of the Spanish Civil War
      - Shots Of War Collection
  - Scripps Institution of Oceanography Library
    - Archives
  - Test Collection

SDS

UC San Diego LIBRARIES

DAMS Collections

- **Current Libraries Collections (6T)**
  - Visual Resources (Art Images)
  - Spanish Civil War Posters
  - Electronic Theses and Dissertations
  - Dr. Seuss Went to War Images
- **Future Data Collections**
  - Departmental Projects
  - Research Project Datasets
- **No Collection Too Big, No Collection Too Small**
  - RDF allows extensibility into any namespace

Questions?

UCSD Geisel Library and Warren Mall
Courtesy UCSD Publications

DP_LIBG002-E