Cyberinfrastructure

David Minor
UC San Diego Libraries
San Diego Supercomputer Center

January 4, 2012

Cyberinfrastructure:

- History
- Definitions
- Examples

History

mid-1990s:

- High performance computing becoming more focused on distributed resources
- Internet boom

Foundational questions:

- How do we bring together distributed resources?
- What impact will it have on science?

National Science Foundation

2003: “Revolutionizing Science and Engineering Through Cyberinfrastructure”

- Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure
- The “Atkins Report”

http://www.nsf.gov/od/oci/reports/toc.jsp

The Panel’s overarching finding is that a new age has dawned in scientific and engineering research, pushed by continuing progress in computing, information, and communication technology, and pulled by the expanding complexity, scope, and scale of today’s challenges. The capacity of this technology has crossed thresholds that now make possible a comprehensive “cyberinfrastructure” on which to build new types of scientific and engineering knowledge environments and organizations and to pursue research in new ways and with increased efficacy.

Recommendation

NSF should establish and lead a large-scale cyberinfrastructure program that is:

- Interagency
- Internationally coordinated
- Creates, deploys, and applies cyberinfrastructure in ways that radically empower all scientific and engineering research and allied education

Office of Cyberinfrastructure (OCI)

- Funds the creation of tools and services
- Coordinates research programs
- Funds national supercomputing efforts
- Guides national discourse

http://www.nsf.gov/dir/index.jsp?org=OCI

Questions?

ARL DLF

Definitions

ARL DLF

Definition One – The Stuff

Cyberinfrastructure is comprised of the hardware and staffing needed to create and support environments for data acquisition and management, as well as the necessary tools, computing and information processing services.

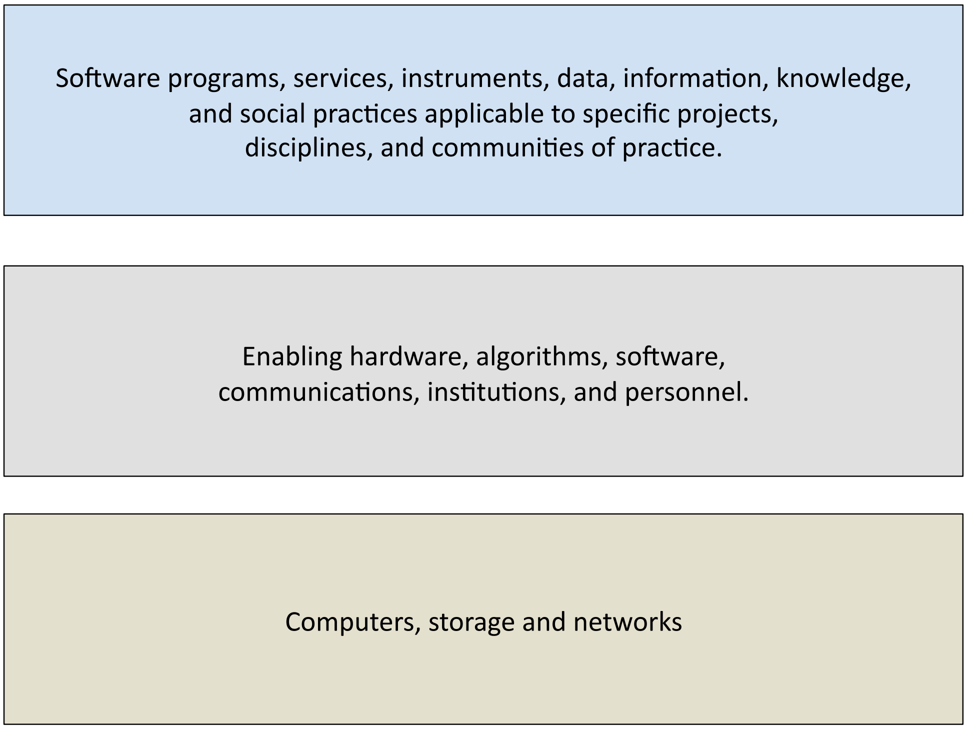
Definition Two – The Plan

For scientists and researchers, cyberinfrastructure includes the technology and programs needed to efficiently connect laboratories, data, computers, and people. This is done to create new forms of science.

Computers, storage and networks

Software programs, services, instruments, data, information, knowledge, and social practices applicable to specific projects, disciplines, and communities of practice.

Computers, storage and networks

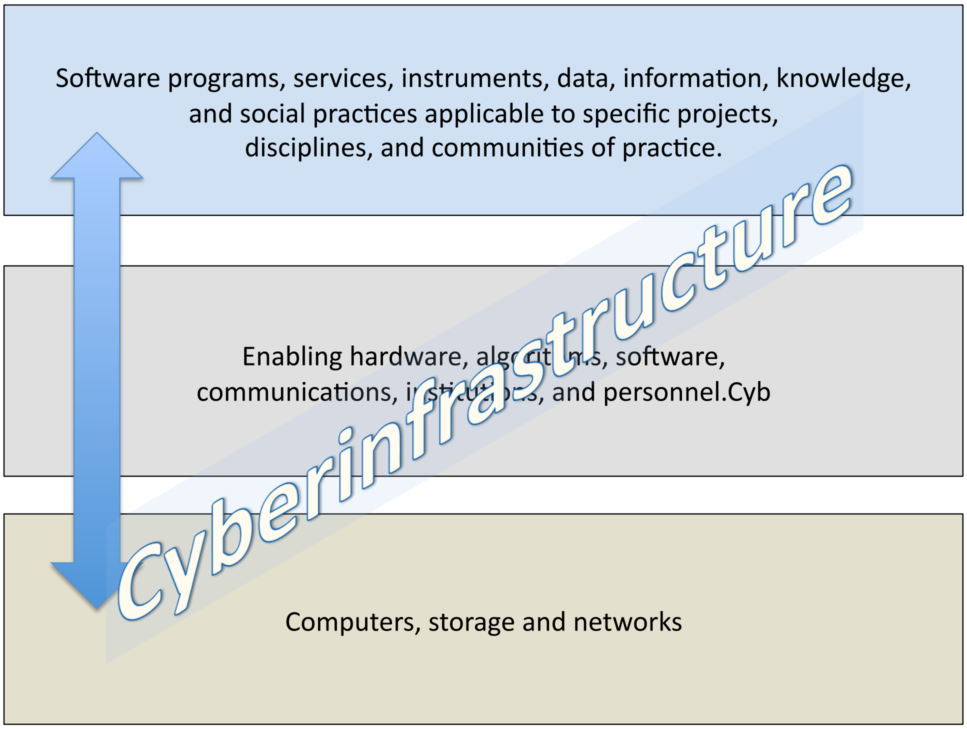


Software programs, services, instruments, data, information, knowledge, and social practices applicable to specific projects, disciplines, and communities of practice.

Enabling hardware, algorithms, software, communications, institutions, and personnel.

Computers, storage and networks

ARL DLF



Software programs, services, instruments, data, information, knowledge, and social practices applicable to specific projects, disciplines, and communities of practice.

Enabling hardware, algorithms, software, communications, institutions, and personnel.Cyb

Computers, storage and networks

ARL DLF

… uh, and E-Science

In most of Europe and much of the rest of the world, what we are talking about is called “E-Science” or “eScience.”

- Operationally and philosophically the same
- The name and concept driven by the United Kingdom in the previous decade

Questions?

Examples

TeraGrid / XSEDE

- TeraGrid 2004 – 2011
  - 2.5 petaflops of computing capability
  - More than 50 petabytes of online and archival data storage
  - More than 100 domain-specific databases

- XSEDE
  - 2011-2016
    https://www.xsede.org

nanoHUB

- Computational nanotechnology research
  - Simulation programs
  - Science gateways, networks and applications
  - Teaching and learning materials
    http://nanohub.org

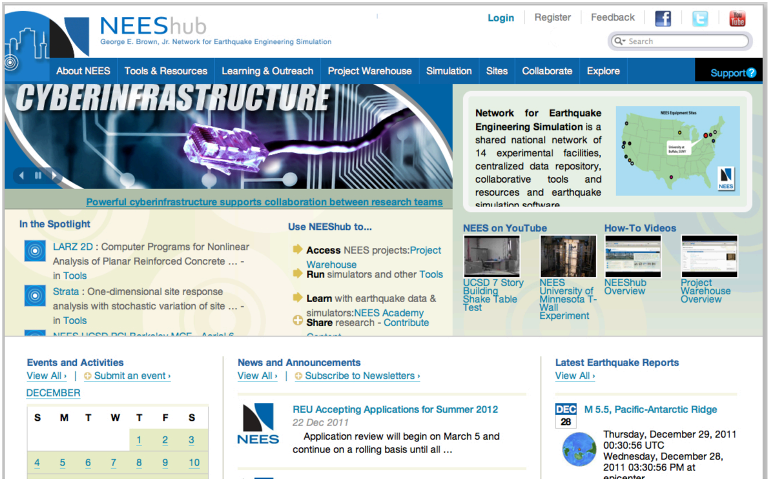
nanoHUB.org
an ncn project

DataNet program

- "Sustainable Digital Data Preservation and Access Network Partners"
  - 2008-present
  - ~$45 million dollars
  - Several large programs with many partners
    http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503141

NSF National Science Foundation
WHERE DISCOVERIES BEGIN

Network for Earthquake Engineering Simulation (NEES)



NEEShub
George E. Brown, Jr. Network for Earthquake Engineering Simulation

Login | Register | Feedback

Search

About NEES | Tools & Resources | Learning & Outreach | Project Warehouse | Simulation | Sites | Collaborate | Explore | Support

CYBERINFRASTRUCTURE

Powerful cyberinfrastructure supports collaboration between research teams

Network for Earthquake Engineering Simulation is a shared national network of 14 experimental facilities, centralized data repository, collaborative tools and resources and earthquake simulation software

In the Spotlight

LARZ 2D : Computer Programs for Nonlinear Analysis of Planar Reinforced Concrete ... - in Tools

Strata : One-dimensional site response analysis with stochastic variation of site ... - in Tools

NEES UCSD PCI Berkeley MCE - Aerial 6

Use NEEShub to...

Access NEES projects:Project Warehouse

Run simulators and other Tools

Learn with earthquake data & simulators:NEES Academy

Share research - Contribute Content

NEES on YouTube

UCSD 7 Story Building Shake Table Test

NEES University of Minnesota T-Wall Experiment

How-To Videos

NEEShub Overview

Project Warehouse Overview

Events and Activities
View All › | Submit an event ›

DECEMBER

S	M	T	W	T	F	S
				1	2	3
4	5	6	7	8	9	10

News and Announcements
View All › | Subscribe to Newsletters ›

NEES

REU Accepting Applications for Summer 2012
22 Dec 2011
Application review will begin on March 5 and continue on a rolling basis until all ...

Latest Earthquake Reports
View All ›

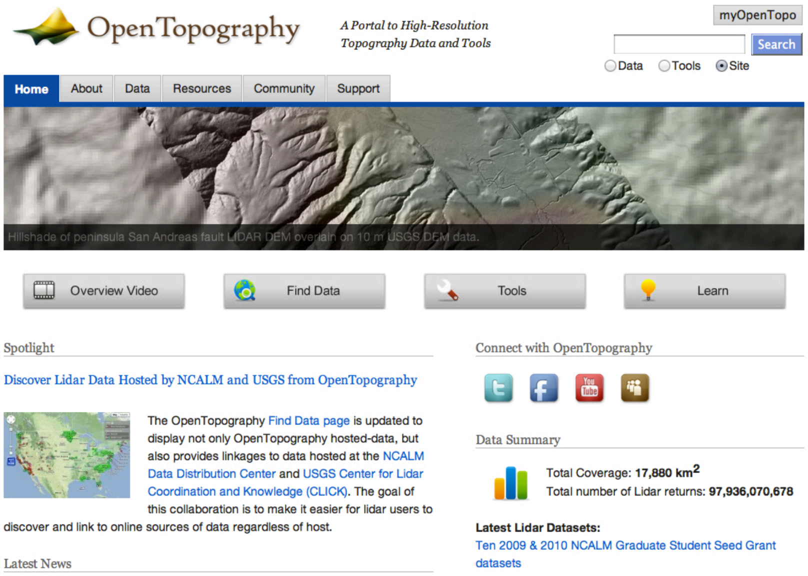
DEC 28 M 5.5, Pacific-Antarctic Ridge

Thursday, December 29, 2011 00:30:56 UTC
Wednesday, December 28, 2011 03:30:56 PM at epicenter

http://nees.org

ARL DLF

OpenTopography



OpenTopography

A Portal to High-Resolution Topography Data and Tools

myOpenTopo

Search

○ Data ○ Tools ● Site

Home | About | Data | Resources | Community | Support

Hillshade of peninsula San Andreas fault LIDAR DEM overlain on 10 m USGS DEM data.

Overview Video | Find Data | Tools | Learn

Spotlight

Discover Lidar Data Hosted by NCALM and USGS from OpenTopography

The OpenTopography Find Data page is updated to display not only OpenTopography hosted-data, but also provides linkages to data hosted at the NCALM Data Distribution Center and USGS Center for Lidar Coordination and Knowledge (CLICK). The goal of this collaboration is to make it easier for lidar users to discover and link to online sources of data regardless of host.

Latest News

Connect with OpenTopography

Data Summary

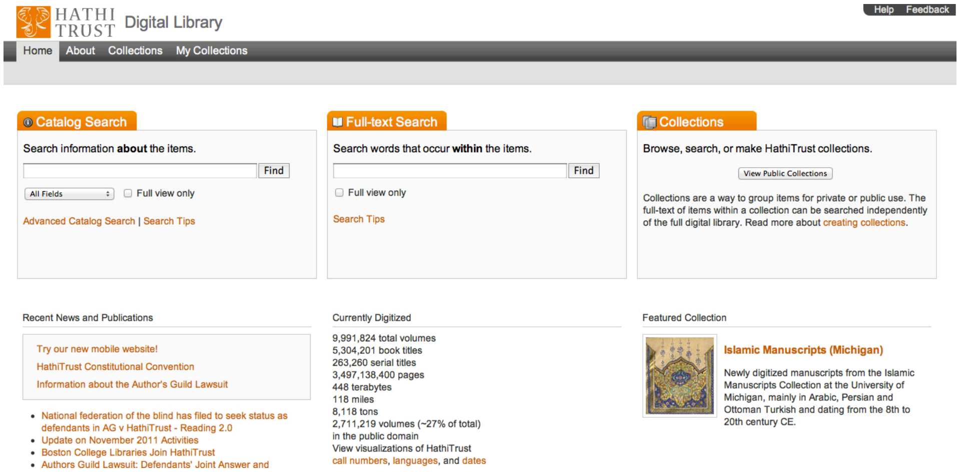
Total Coverage: **17,880 km^2**
Total number of Lidar returns: **97,936,070,678**

Latest Lidar Datasets:
Ten 2009 & 2010 NCALM Graduate Student Seed Grant datasets

http://www.opentopography.org

ARL DLF

HathiTrust



Search information **about** the items.

□ Full view only

Advanced Catalog Search | Search Tips

Search words that occur **within** the items.

□ Full view only

Search Tips

Browse, search, or make HathiTrust collections.

Collections are a way to group items for private or public use. The full-text of items within a collection can be searched independently of the full digital library. Read more about creating collections.

Recent News and Publications

Try our new mobile website!
HathiTrust Constitutional Convention
Information about the Author's Guild Lawsuit

- National federation of the blind has filed to seek status as defendants in AG v HathiTrust - Reading 2.0
- Update on November 2011 Activities
- Boston College Libraries Join HathiTrust
- Authors Guild Lawsuit: Defendants' Joint Answer and

Currently Digitized

9,991,824 total volumes
5,304,201 book titles
263,260 serial titles
3,497,138,400 pages
448 terabytes
118 miles
8,118 tons
2,711,219 volumes (~27% of total) in the public domain
View visualizations of HathiTrust call numbers, languages, and dates

Featured Collection

Islamic Manuscripts (Michigan)

Newly digitized manuscripts from the Islamic Manuscripts Collection at the University of Michigan, mainly in Arabic, Persian and Ottoman Turkish and dating from the 8th to 20th century CE.

http://www.hathitrust.org

Penn State University

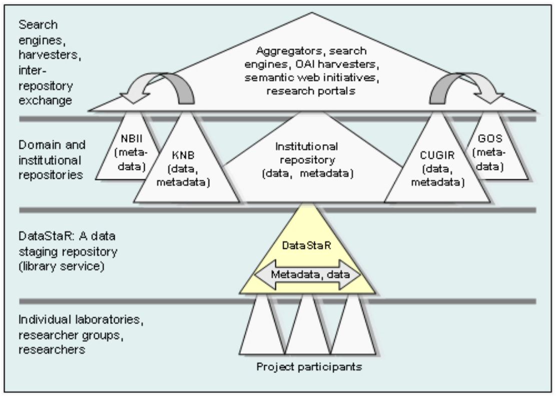
- Digital Library Technologies, a division of Information Technology Services
  - Defining service levels for archival storage and repository services for the university

- Digital Library Technologies and University Libraries
  - Publishing and curation services program in support of digital stewardship at the institution
  - Working with other organizations around the country to create flexible, shareable tools and services

  http://www.dlt.its.psu.edu/

Cornell - DataStaR

- Investigating methods for collaborating and sharing data as well as promoting high-quality metadata creation.


Search engines, harvesters, inter-repository exchange
Aggregators, search engines, OAI harvesters, semantic web initiatives, research portals
Domain and institutional repositories
NBII (meta-data)
KNB (data, metadata)
Institutional repository (data, metadata)
CUGIR (data, metadata)
GOS (meta-data)
DataStaR: A data staging repository (library service)
DataStaR
Metadata, data
Individual laboratories, researcher groups, researchers
Project participants

http://datastar.mannlib.cornell.edu/

Johns Hopkins University

Data Management Service (DMS)

Powered by Data Conservancy

- JHU Data Management Service (DMS) represents the culmination of two years of research, design, development and implementation of Data Conservancy
- Service launched in July 2011
- DC instance launched in October 2011
- Important, essential foundations in place
- There remains work to be done

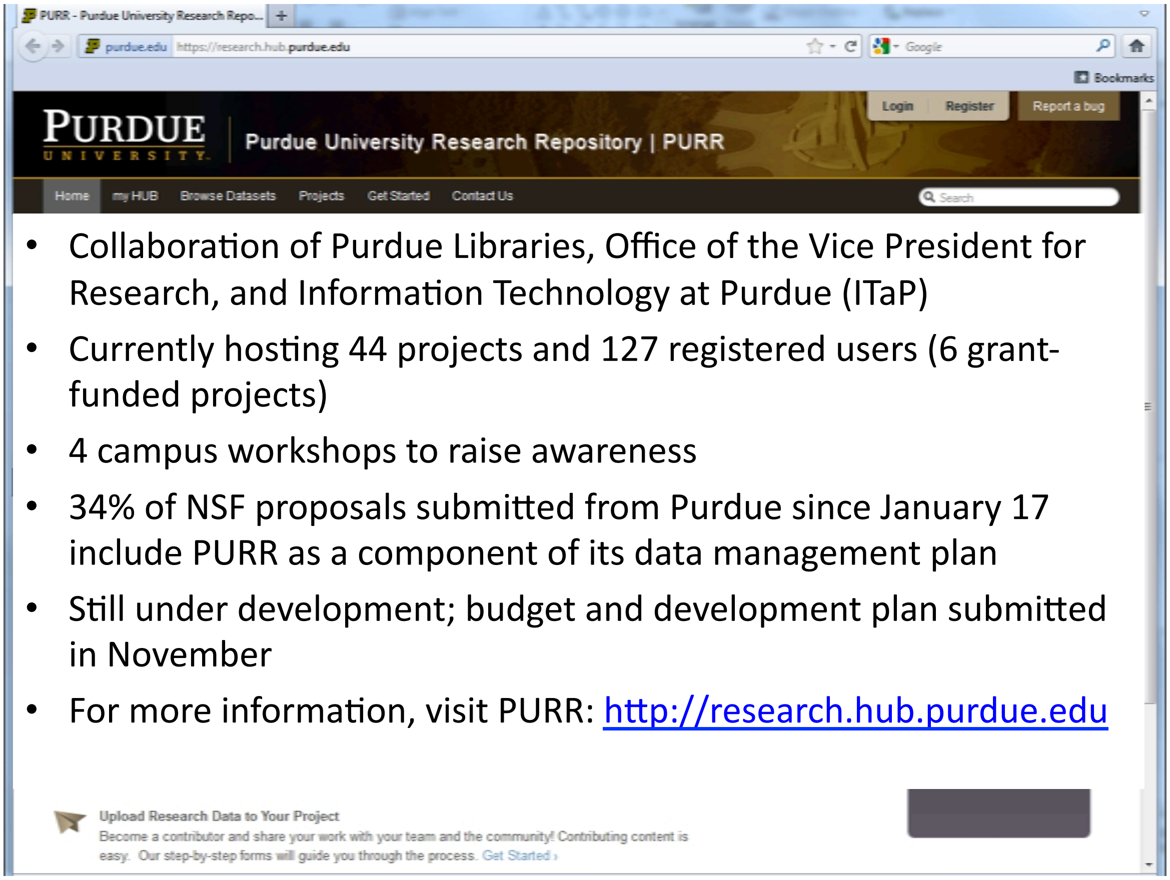
JHU Data Management Services

Two sets of services provided:

- Consultative data management planning services to support NSF proposals
  - *direct funded by schools with PIs seeking NSF funding*
- Post award data management planning and deposit in the JHU Data Archive (DC instance)
  - *funded through charge back on grant*
  - *Defined time limits, volume of data deposited per project, unencumbered data only for now*

Purdue University

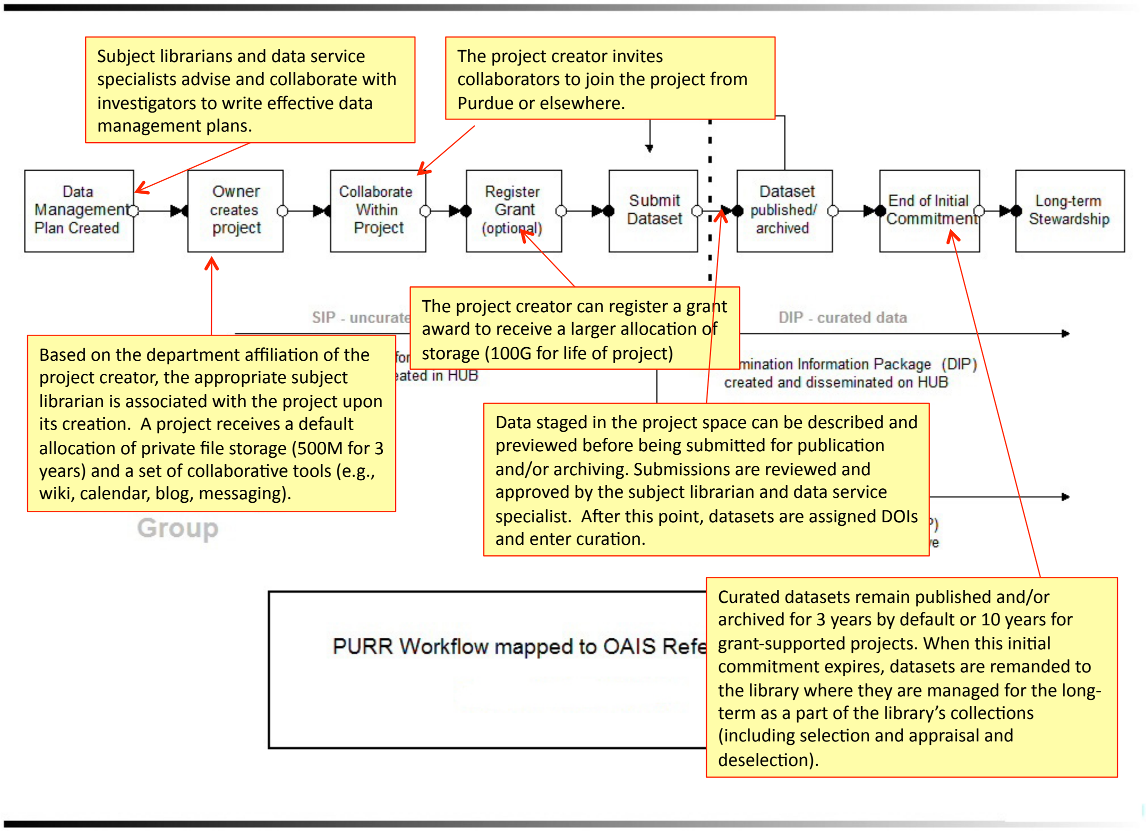
Purdue University Research Repository (PURR)


PURR - Purdue University Research Repo...
purdue.edu https://research.hub.purdue.edu
Google
Bookmarks
Login
Register
Report a bug
PURDUE UNIVERSITY
Purdue University Research Repository | PURR
Home
my HUB
Browse Datasets
Projects
Get Started
Contact Us
Search

- Collaboration of Purdue Libraries, Office of the Vice President for Research, and Information Technology at Purdue (ITaP)
- Currently hosting 44 projects and 127 registered users (6 grant-funded projects)
- 4 campus workshops to raise awareness
- 34% of NSF proposals submitted from Purdue since January 17 include PURR as a component of its data management plan
- Still under development; budget and development plan submitted in November
- For more information, visit PURR: http://research.hub.purdue.edu

Upload Research Data to Your Project

Become a contributor and share your work with your team and the community! Contributing content is easy. Our step-by-step forms will guide you through the process. Get Started ›


Data Management Plan Created
Owner creates project
Collaborate Within Project
Register Grant (optional)
Submit Dataset
Dataset published/ archived
End of Initial Commitment
Long-term Stewardship
SIP - uncurate
for
eated in HUB
DIP - curated data
mination Information Package (DIP) created and disseminated on HUB
)
e
Group
PURR Workflow mapped to OAIS Refe

Subject librarians and data service specialists advise and collaborate with investigators to write effective data management plans.

The project creator invites collaborators to join the project from Purdue or elsewhere.

The project creator can register a grant award to receive a larger allocation of storage (100G for life of project)

Based on the department affiliation of the project creator, the appropriate subject librarian is associated with the project upon its creation. A project receives a default allocation of private file storage (500M for 3 years) and a set of collaborative tools (e.g., wiki, calendar, blog, messaging).

Data staged in the project space can be described and previewed before being submitted for publication and/or archiving. Submissions are reviewed and approved by the subject librarian and data service specialist. After this point, datasets are assigned DOIs and enter curation.

Curated datasets remain published and/or archived for 3 years by default or 10 years for grant-supported projects. When this initial commitment expires, datasets are remanded to the library where they are managed for the long-term as a part of the library's collections (including selection and appraisal and deselection).

Indiana University

IU Libraries and Cyberinfrastructure

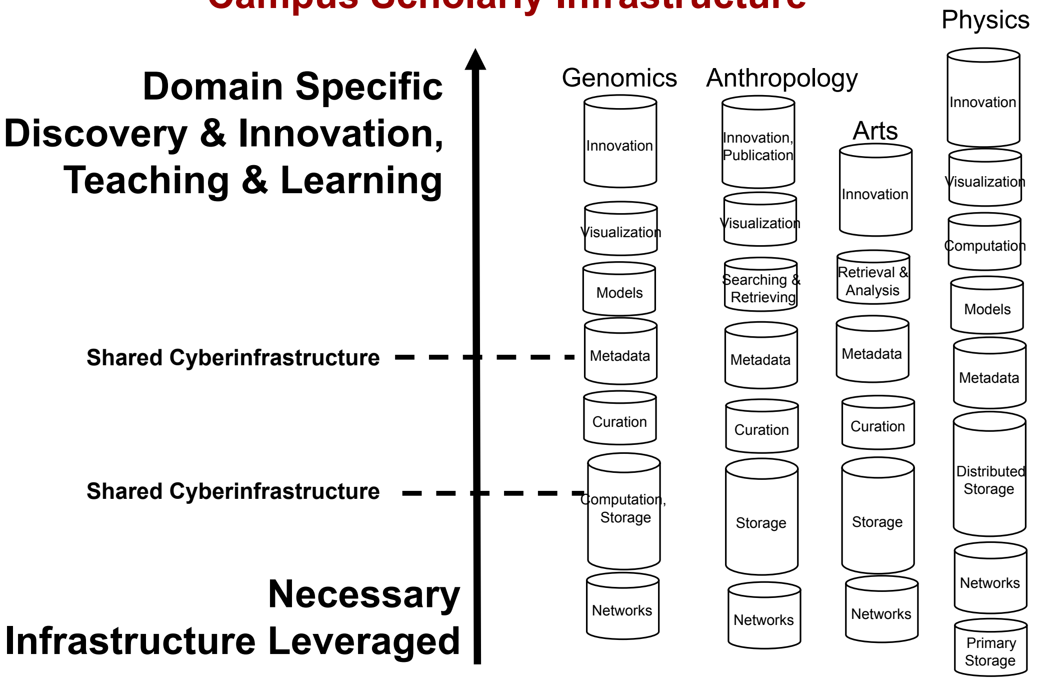
Factors driving consolidation of cyberinfrastructure

- Economies of scale in management
  - Air conditioning
  - Electricity supply
  - Staff
- Security
- VM hosting – intelligent infrastructure
- Economies of vendor partnerships and large purchases
- Redundant backups (2 locations) and our own network
- Some places are best not consolidated (Astro, some of the clusters in Chemistry)

IU Bloomington Data Center
http://dcops.iu.edu/

INDIANA UNIVERSITY MOVING FAST. THINKING BIG. SYSTEMS • NETWORKS • DATA • DISCOVERIES
PERVASIVE TECHNOLOGY INSTITUTE INDIANA UNIVERSITY
GlobalNOC Global Research Network Operations Center

Campus Scholarly Infrastructure



Domain Specific Discovery & Innovation, Teaching & Learning

Shared Cyberinfrastructure

Shared Cyberinfrastructure

Necessary Infrastructure Leveraged

Genomics: Innovation, Visualization, Models, Metadata, Curation, Computation, Storage, Networks

Anthropology: Innovation, Publication, Visualization, Searching & Retrieving, Metadata, Curation, Storage, Networks

Arts: Innovation, Retrieval & Analysis, Metadata, Curation, Storage, Networks

Physics: Innovation, Visualization, Computation, Models, Metadata, Distributed Storage, Networks, Primary Storage

Discipline Research Stacks…

IU Libraries and Cyberinfrastructure

- IU Libraries are involved in a variety of collaborative projects that utilize shared cyberinfrastructure services with a variety of partners on the IU campus. These include areas of OVPIT (Research Technologies, Learning Technologies and Enterprise Software) and areas within the Pervasive Technology Institute Research Centers, the Office of the Vice President for Research, and the College of Arts and Sciences.
  - Shared storage infrastructure for digital collections, institutional repository, publishing, data curation, and digital humanities projects.
  - Shared computational infrastructure for use with the HathiTrust Research Center, data curation, and other data transport and transformation needs.
  - Shared large-area file system for mass-data transfers and redundancy –
    - Shared data MOU with TACC
  - Shared virtual infrastructure within the IU Cloud – used for most library technologies, digital libraries and other library application services at IU.
- These resources when utilized for enterprise-wide solutions are part of the IU Empowering People Plan of abundant unmetered resources.

Indiana University
MOVING FAST.
THINKING BIG.
SYSTEMS • NETWORKS • DATA • DISCOVERIES
PERVASIVE TECHNOLOGY INSTITUTE
INDIANA UNIVERSITY
GlobalNOC
Global Research Network Operations Center

UC San Diego

Research Cyberinfrastructure (RCI)

RCI elements

- High-Performance Computing
- Data Center Colocation
- Storage
- Data curation
- Networking and other services

High-performance computing

- Triton Resource: a cost–effective and accessible high-performance computing system primarily for UC San Diego and UC researchers

- The Triton Affiliates and Partners Program (TAPP): high performance cluster computing time.
  http://www.sdsc.edu/us/tapp

Data center colocation

- Standard rack provided with ISO-Base seismic protection, aisle containment, and 2x30A power distribution
- 10+ Gb networking fabric connectivity both throughout SDSC aggregation fabric and into CENIC
- 24/7 operations staff providing facility oversight and emergency "remote hands" hardware assistance

http://rci.ucsd.edu/services/colocation.html

Storage

Storage Type	Cost per Terabyte-Year	Availability	Application Performance
Parallel File System	Free while running on an SDSC HPC machine. Medium-term parking space available by special arrangement with Project Storage purchased in an equal quantity.	99.5%	Up to 100 GB/s
Project Storage	• Standard Availability, Single-Site Durability - $600 • High Availability, Multiple-Site Durability - $900	• 99.5% • 99.95%	• Up to 1 GB/s • Up to 1 GB/s
Cloud Storage	• Single-Site Durability - $390 • Triple Copy - $650	• 99.5% • 99.5%	• Up to 100 MB/s • Up to 100 MB/s

Networking and other services

- Web & Database Hosting
- Oracle Database Hosting
- 10GigE research network throughout campus

http://rci.ucsd.edu/services/other-services.html

Data curation

- Starting with in two year pilot phase

- Using existing tools whenever possible
  - Storage at SDSC
  - Digital Asset Management System at UCSD Libraries
  - Campus high-speed networking
  - Chronopolis digital preservation network

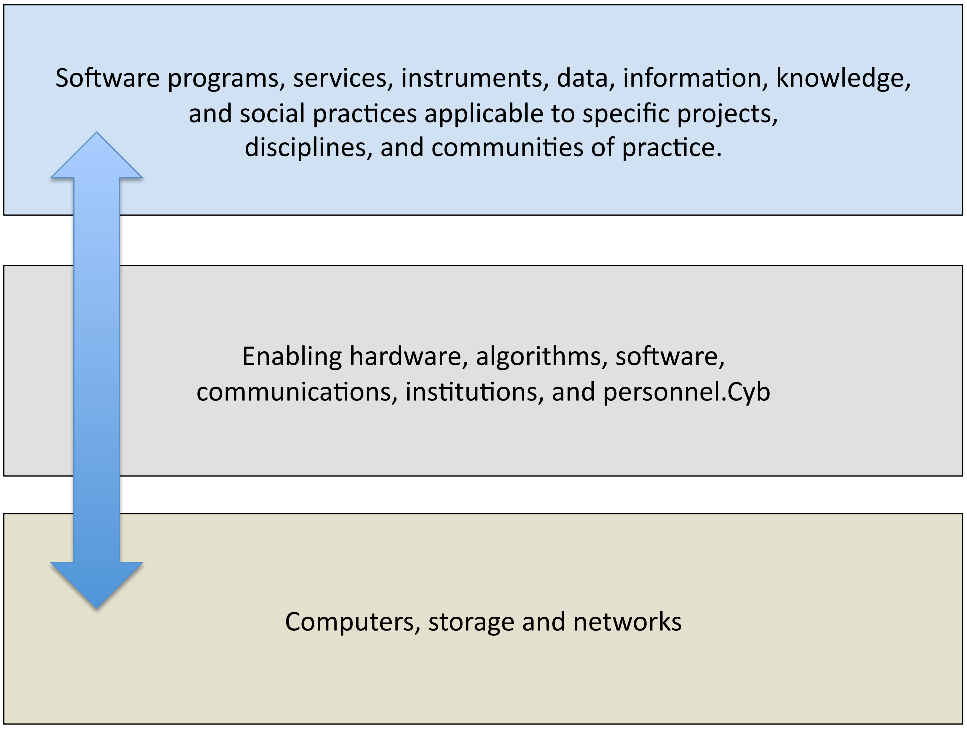
http://rci.ucsd.edu/services/data-curation.html

Data management plans

- Resources and contacts available to UCSD researchers
- Examples from submitted proposals
- Guidance, tips and recommendations for DMP preparation

http://rci.ucsd.edu/dmp/index.html

Cyberinfrastructure is …

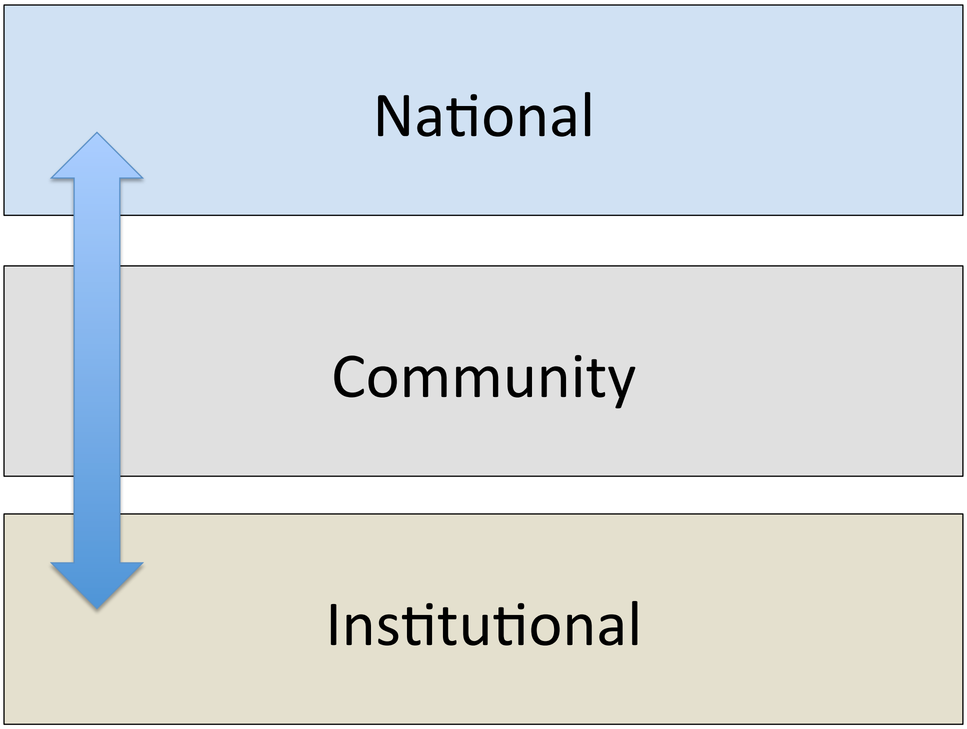


Software programs, services, instruments, data, information, knowledge, and social practices applicable to specific projects, disciplines, and communities of practice.

Enabling hardware, algorithms, software, communications, institutions, and personnel.Cyb

Computers, storage and networks

ARL DLF



National

Community

Institutional

ARL DLF

Questions!

ARL DLF

Cliff Lynch

Director of the Coalition for Networked Information

ARL DLF