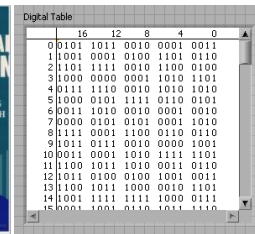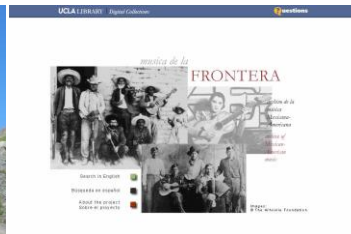# Building Services, Building Communities, Supporting Data Intensive Research

January 2012 – International Data Curation Conference

Patricia Cruse, Director
University of California Curation Center
California Digital Library

California Digital Library

Serving the University of California

- 10 campuses

- 360K students, faculty, and staff

- 100's of museums, art galleries, observatories, marine centers, botanical gardens

- 5 medical centers

- 5 law schools

- 3 National Laboratories

CDL supports the research lifecycle

- Collections

- Digital Special Collections

- Discovery & Delivery

- Publishing Group

- UC Curation Center (UC3)

Research & Scholarship Lifecycle

Research

Save — Create Knowlege — Collect

Share — Publish

UC3

University of California

CDL
California Digital Library

California Digital Library (CDL)

Our environment circa 2002-2008

Focus on preservation

For memory organizations

Content: museum & library

Infrastructure: static

Services: hosted

Sustainability: ?

UC3

The changing landscape

- Ever increasing number, size, and diversity of content

- Ever increasing diversity of partners, and stakeholders

- Decreasing resources

- Inevitability of disruptive change

  – Technology

  – Institutional mission
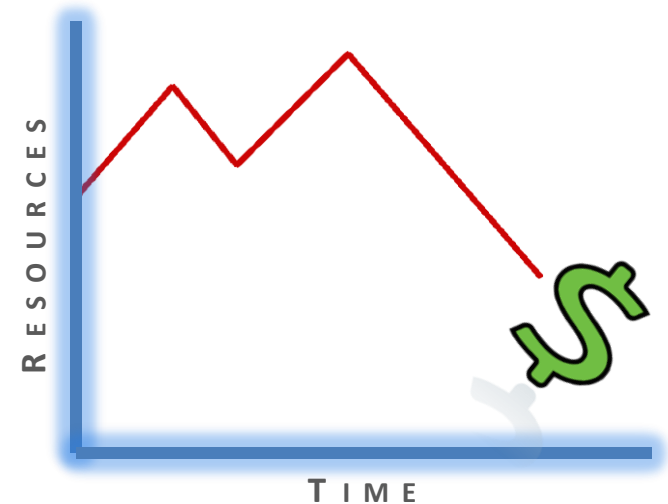
The University of California's role in research

UC share of U.S. research expenditures, Universitywide, 1999-2000 to 2009-10
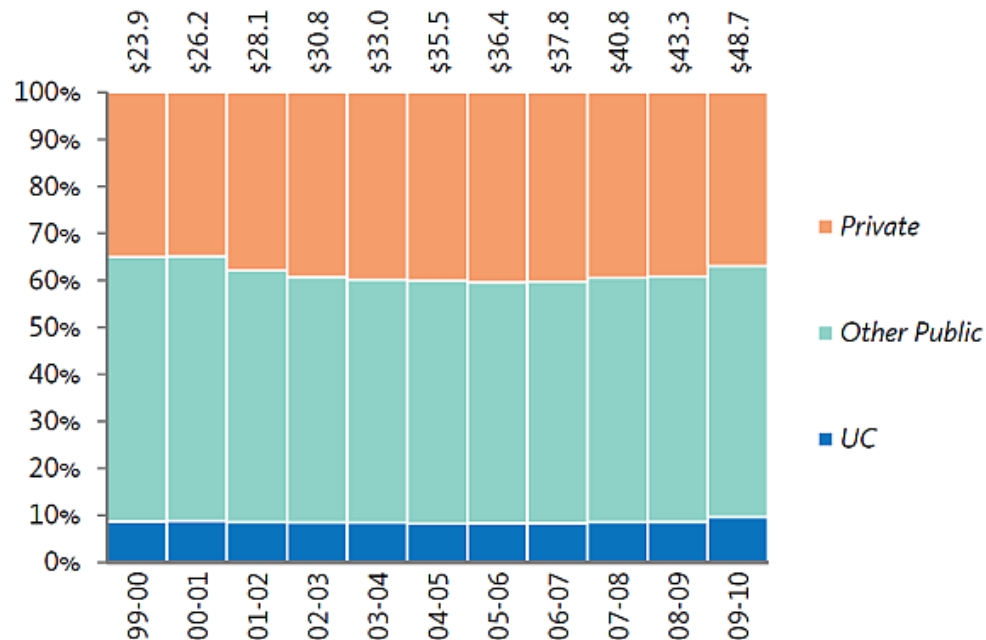
The University of California performs about one-twelfth of all the academic research and development conducted in the United States.

Source: IPEDS

What keeps users up at night?

Our environment since 2008

Focus on preservation → *curation* **(lifecycle)**

For memory organizations → **and now data producers**

Infrastructure: static → **applications enabling diverse use**

Services: hosted → **partnered, self-serve**

Content: museum & library → **research, web crawls**

Sustainability: ? → **new cost models**

UC3

University of California
CDL
California Digital Library

Pivot and work differently: build community and applications

✓ Engage with the research community early in the lifecycle

✓ Prioritize the use of scarce resources to support emerging initiatives

✓ Start small: provide simple solutions that build up to solve more complex problems

✓ Deploy simple and flexible infrastructure and services that can be used in diverse ways

✓ Develop partnerships and collaborate now more than ever!

Five services to meet diverse needs

1. Create, edit, share, and save data management plans

DMPTool
Guidance and Resources for your Data Management Plan

2. Create, share, archive, publish tabular data

❋ UC3DATA<u>UP</u>

3. Create and manage persistent identifiers

❋ UC3EZID

4. Curation repository: store, manage, preserve, and share research data

❋ UC3Merritt

5. A service to collect, manage and preserve Web-published content

❋ UC3WAS

❋ UC3

University of California
CDL
California Digital Library

Service 1: DMPTool

Meeting funding agencies data management plan requirements

- Connect researchers to resources to create a data management plan

- NSF and directorates, NIH, NEH, IMLS, foundations plus

- Customizable
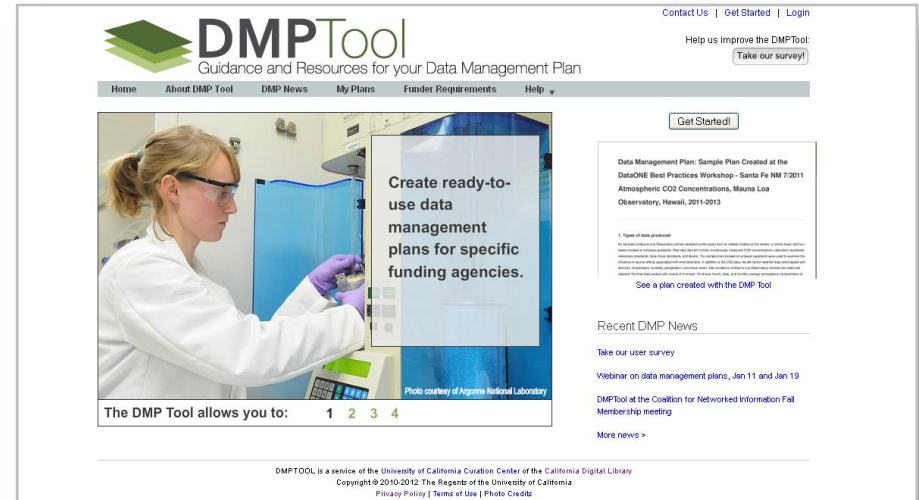
Primary Functions

1. Step-by-step "wizard"

2. Templates and examples

3. Links to institutional resources and agency information

4. Plan publication and sharing

DMTool usage
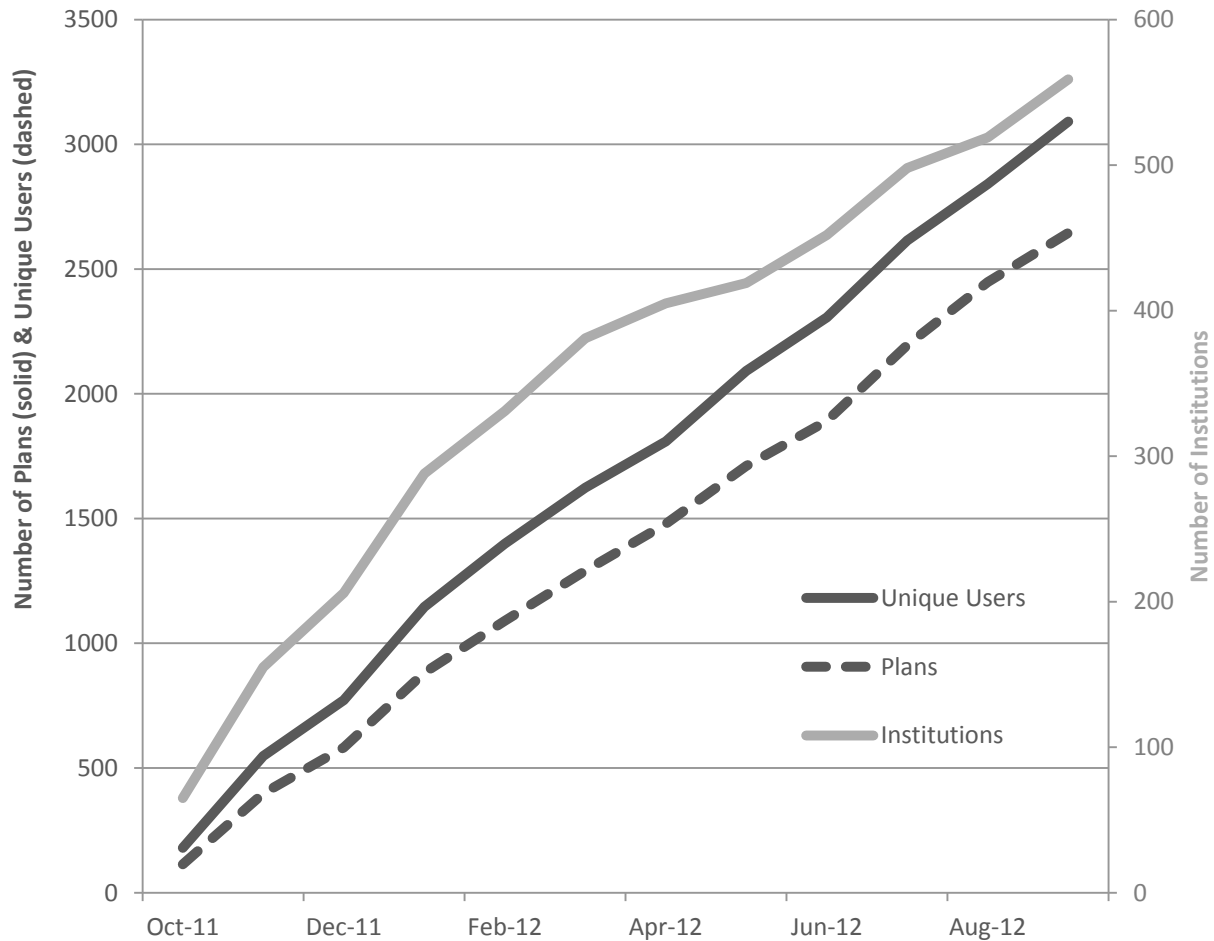
New DMPTool development

- Funding from Sloan Foundation
- Starts now
- Partners:
  - UC3
  - DataONE
  - University of Virginia Library
  - University of Illinois Urbana Champaign
  - DMPTool Partners

Goal: DMPTool can provide a centralized point for the consolidation, sharing and dissemination of expertise and advice in good data management practices and solutions
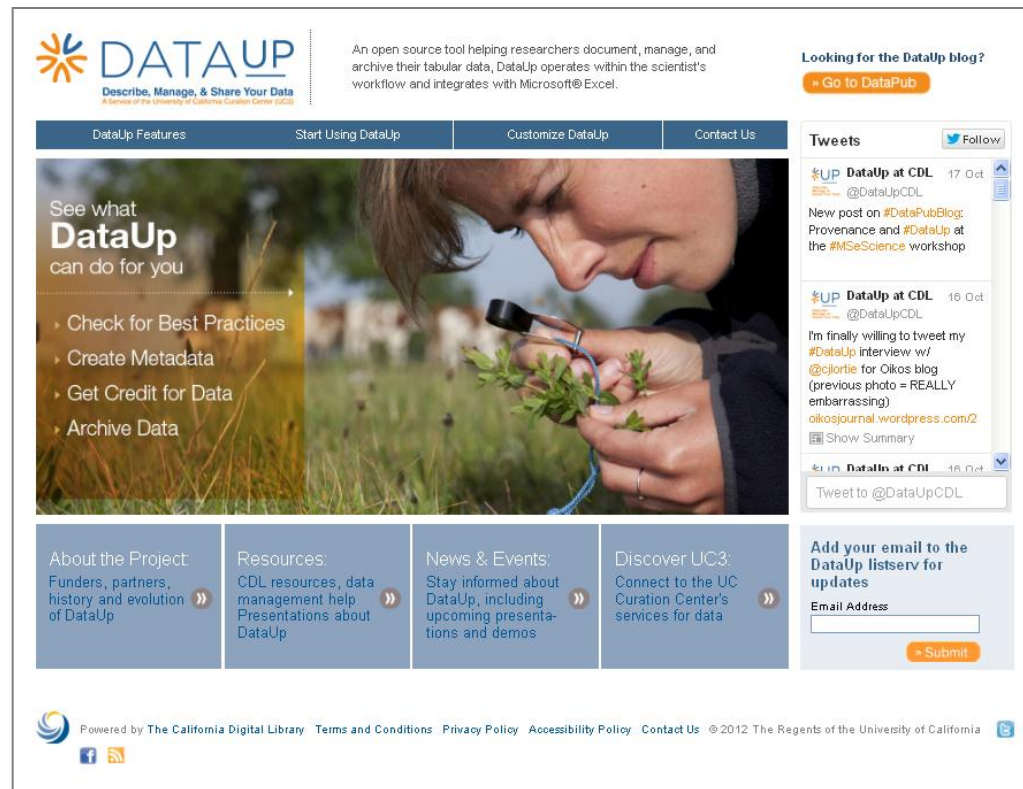
- Enable institutions to easily shape and exploit the DMPTool to meet local need
- Provide support for the full data management life cycle
- Foster the emergence of an engaged open source community of users and developers
- Enable collaboration and connect stakeholders to each other and institutional resources, including services, expertise and guidance.

UC3

University of California
CDL
California Digital Library

Service 2:

Create, share, archive, connect, publish data

Excel is the database of choice for many researcher

Primary Functions

1. An Excel 1) add-in and 2) cloud application

2. Document data

3. Check for good data practices

3. Obtain identifier and citation

4. Archive and share

Microsoft **Research**

GORDON AND BETTY **MOORE** FOUNDATION

UC3
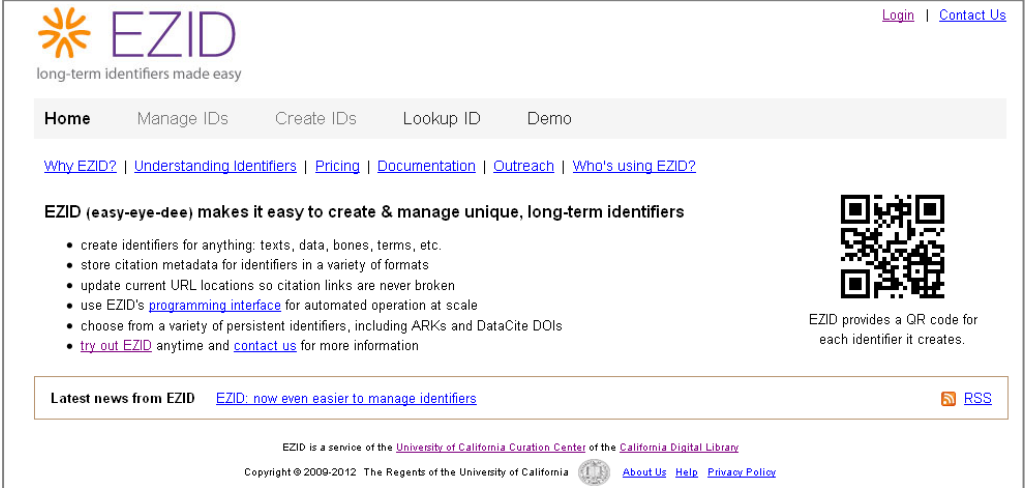
University of California
CDL California Digital Library

Service 3: ✳ EZID
Long term identifiers made easy

- Credit to data producers and data publishers

- A link from the traditional literature to the data

- Cost model in place to support infrastructure

- Range of customers

Primary Functions

1. Create identifiers (DOIs and ARKs) for anything

2. Citation metadata and persistent citation links

3. DOIs from DataCite

4. Global discovery via Primo and (soon) Web of Knowledge

Service 4: ✳Merritt

- Curation repository open to the UC community and beyond

- Discipline / content agnostic
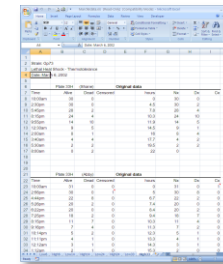
- Architecture allows for diversity

Primary Functions
1. Deposit
2. Manage (metadata, versions, etc)
3. Access (expose)
4. Share (with other researchers)
5. Preserve

✳ UC3

Merritt's diverse service offering: meeting a range of needs

- Dark archive for important digital assets

- Bright archive with direct discovery and access

- Preservation back-end for existing or new discovery and content management systems

- Integration with distributed data grids

Merritt example: UC Berkeley's Research Hub back-end preservation

Merritt example: DataShare bright archive / direct discovery

UCSF DataShare
open biomedical research data from UCSF

Catalyze widespread intra- and inter-disciplinary sharing of research data

Pilot project in neuro-degenerative disease imaging

Persistent identification and preservation through integration with EZID and Merritt

UC3

University of California
CDL
California Digital Library
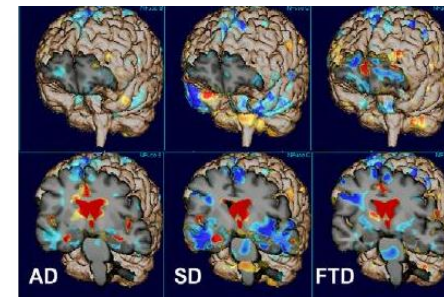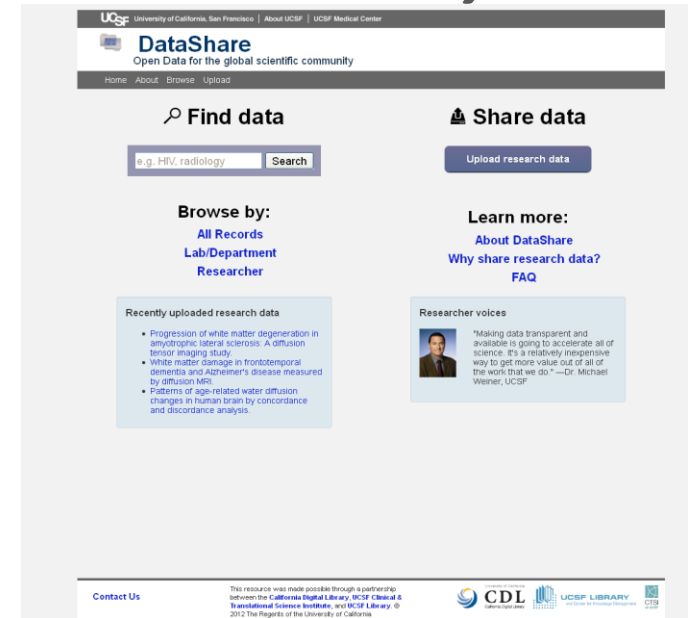
Merritt example: DataONE– part of a data grid
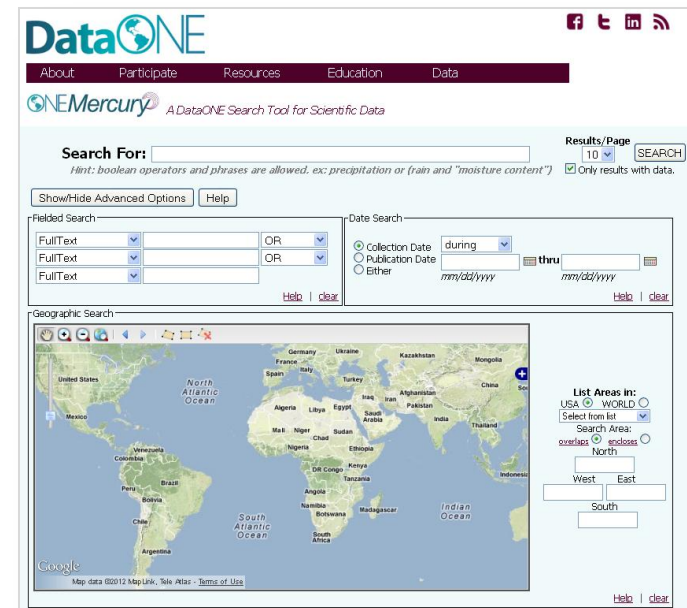

Federated framework and sustainable cyber-infrastructure for innovative, data-intensive environmental science

– Investigator's toolkit, global replication and discovery, education

A community of engaged stakeholders

– Academic
– Commercial
– Non-profit
– Governmental

Merritt participates as a DataONE member node

www.dataone.org

Data Repository for **Anyone | Anywhere**

Service 5: Web Archiving Service (WAS)

UC3 WAS

- ARCHIVE institution websites
- BUILD collections for research
- CAPTURE political and social events
- SAVE at-risk government websites

Primary Functions
1. Capture
2. Manage
3. Preserve
4. Publish

UC3

University of California
CDL
California Digital Library

Cost modeling

The long-term success of Merritt and its surrounding services is dependent on their sustainability: technically, administratively, and financially

- Goals

  - *Partial recovery of UC3 service costs*

  - *Flexibility in billing models to ensure uninterrupted service*

- Two options

  - *Pay-as-you-go*   (annual payments)

  - *Paid-up*   (one-time, up-front payment)

  http://wiki.ucop.edu/display/Curation/Cost+Modeling

Pivot and work differently

- ✓ Engage with the research community early in the lifecycle – **build community**

- ✓ Prioritize the use of scarce resources to support emerging initiatives – **be entrepreneurial**

- ✓ Start small: provide simple solutions that build up to solve more complex problems **– develop sustainability solutions**

- ✓ Deploy simple and flexible infrastructure and services that can be used in diverse ways – **enable applications to be used in diverse ways**

Develop partnerships and collaborate now more than ever!

For more information

UC Curation Center

http://www.cdlib.org/uc3

uc3@ucop.edu

Stephen Abrams   *Mark Reyes*

Patricia Cruse   *Abhishek Salve*

Scott Fisher   *Joan Starr*

Erik Hetzner   *Rosalie Lack*

Greg Janée   *Carly Strasser*

John Kunze   *Marisa Strong*

Margaret Low   *Adrian Turner*

David Loy   *Perry Willett*