Integrating Repositories for Research Data Sharing

Stephen Abrams
California Digital Library

Angela Rizk-Jackson
Julia Kochi
University of California, San Francisco

Noah Wittman
University of California, Berkeley

CC BY SA

Why is data curation important?

- Accelerating scientific progress
- Enabling appropriate scrutiny and verification of results
- Promoting integrity and debate
- Facilitating new collaborations
- Avoiding needless duplication of effort
- Increasingly, complying with institutional policies, publication requirements, and funder mandates

Cf. White and Teds (2011), "Making the case for research data management" DCC briefing paper, *www.dcc.ac.uk/resources/briefing-papers/making-case-rdm*

Merritt

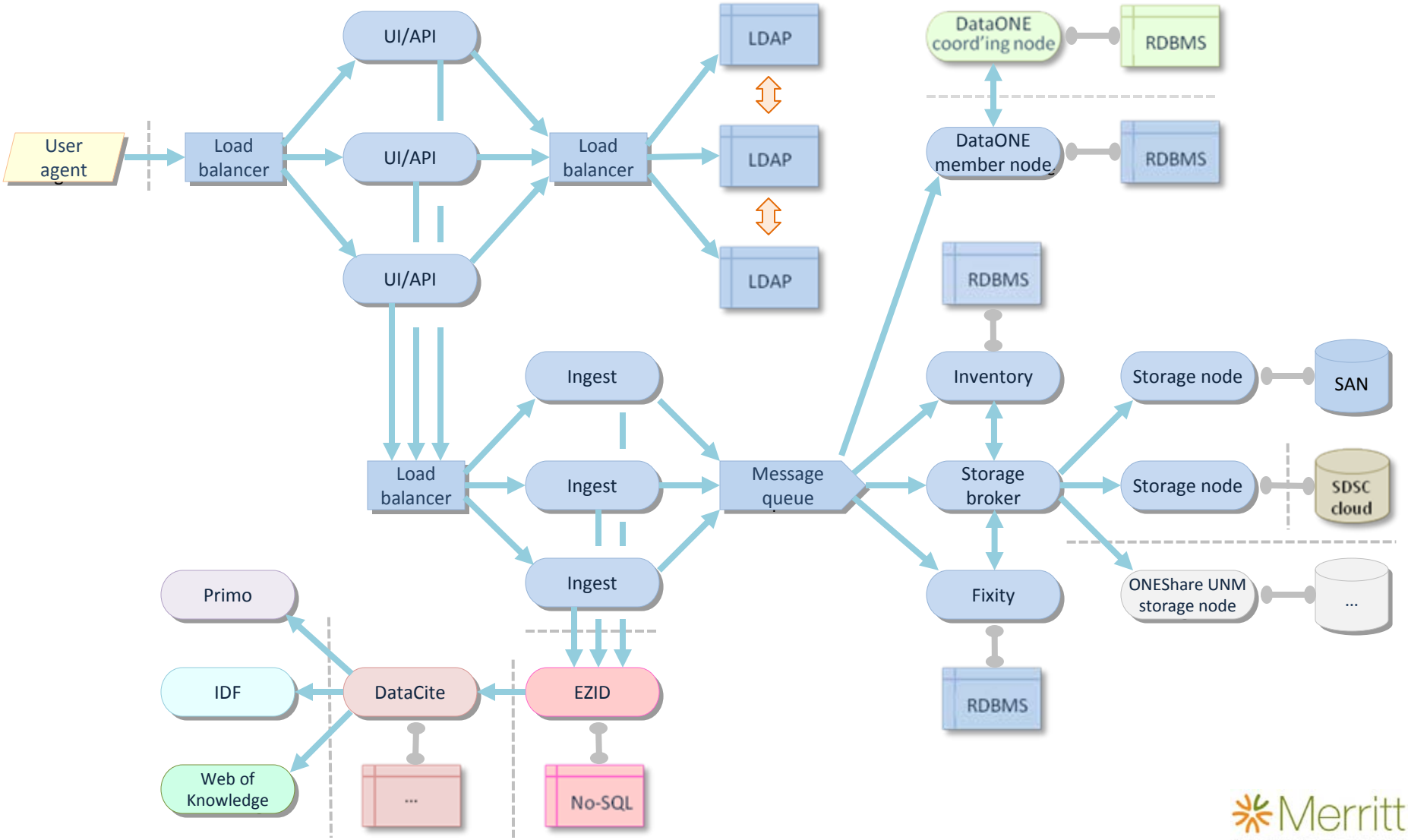
- Curation repository available to the UC community and external partners
  - *Preservation and access*
  - *Content agnostic, model free*
  - *Highly decentralized micro-services architecture*

Cf. Abrams, Cruse, Kunze, and Minor (2011), "Curation micro-services: A pipeline metaphor for repositories," *Journal of Digital Information* 12(2), *journals.tdl.org/jodi/article/view/1605*

- 26 *curatorial units*
- 271 *collections*
- 325,000 *objects*
- 450,000 *versions*
- 4,500,000 *files*
- 13 *TB*

www.cdlib.org/uc3/merritt
merritt.cdlib.org

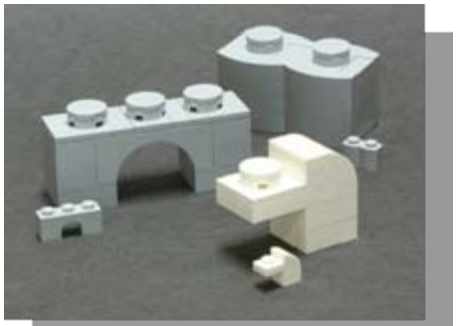
Merritt



Merritt
Share, manage & preserve your content

(*Some*) *issues to address*

- Scale
  - *Individual objects ranging from 0 to 47,000 files*
  - *Individual files ranging from 0 to 14 GB*
- Maintaining control
  - *Concern over potential loss of control over dissemination and use of data*
- User experience
  - *Switch from organizational to individual interaction*



www.flickr.com/photos/32195273@N05/51076852642

www.flickr.com/photos/traftery/4319529821

www.flickr.com/photos/vixon/116447718

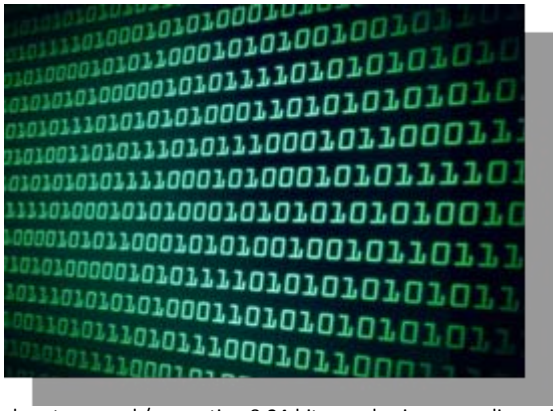
(*Some*) *issues to address*

- Scale
  - *Individual objects ranging from 0 to 47,000 files*
  - *Individual files ranging from 0 to 14 GB*
- Maintaining control
  - *Concern over potential loss of control over dissemination and use of data*
- User experience
  - *Switch from organizational to individual interaction*
- Augment repository function by composition (*when possible*) and addition (*when necessary*)
  - *Loosely-coupled integration with external community supported systems and services*

Scale

- Avoiding client timeout
  - ≤ 2 GB: *File-based → stream-based AIP-to-DIP processing*
  - > 2 GB: *Asynchronous delivery*
    - *Email notification with personalized, time-limited URL*
- Streamlined storage provisioning
  - *SDSC cloud*

cloud.sdsc.edu



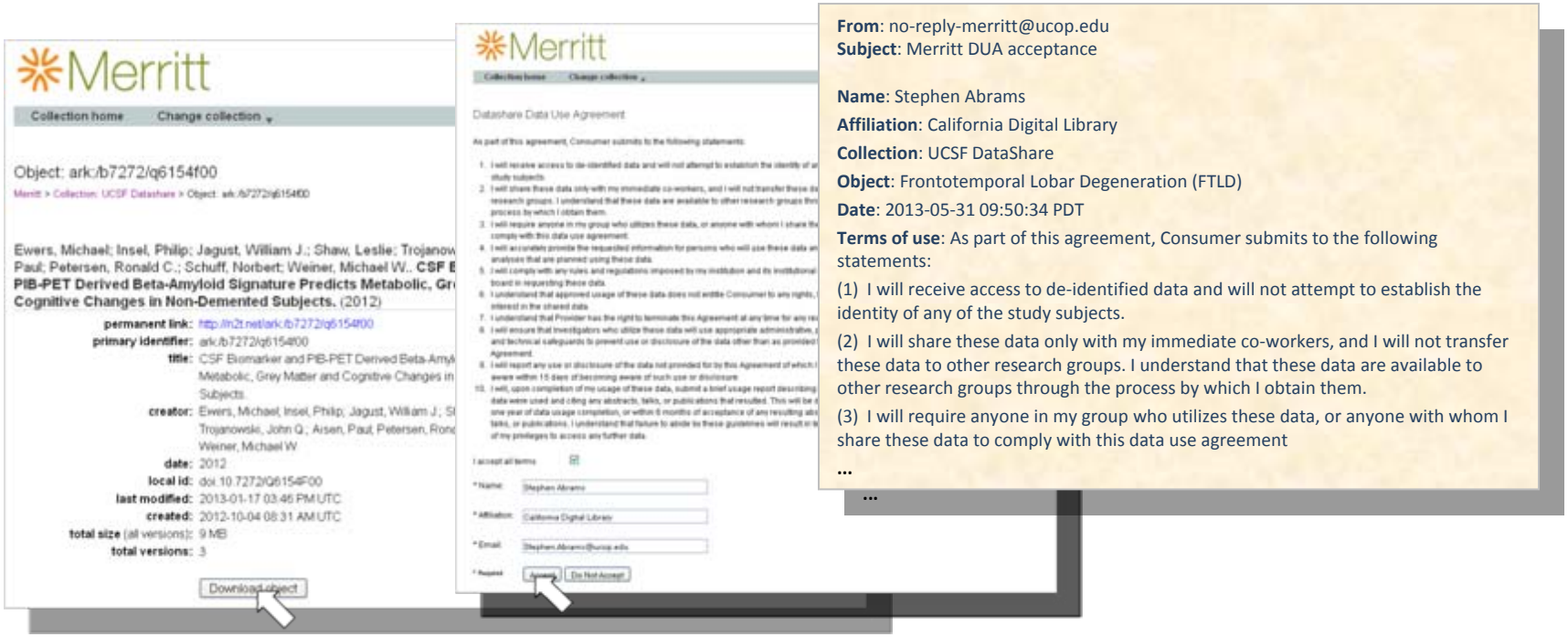
www.kevatron.co.uk/converting-8-24-bit-samples-in-coreaudio-on-ios

www.flickr.com/photos/paulbhartzog/680749585

Control

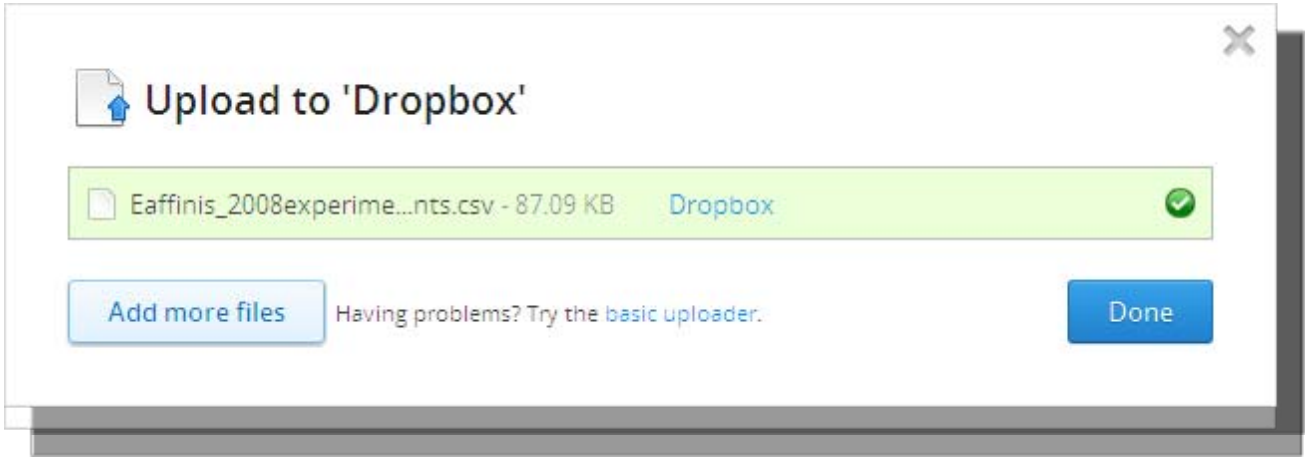
- Data use agreements (DUAs)
  - Explicit assertion of license requirements and terms of use
  - Curatorial and consumer notification of acceptance

Cf. Brazhnik and Jones (2007), "Anatomy of data integration," *Journal of Biomedical Informatics* 40(3): 252-69, *doi:10.1016/j.jbi.2006.09.001*



User experience

- Due to its open eligibility policy, Merritt will always provide a more generic UX than special-purpose or disciplinary systems
- Shifting user roles, shifting expectations
  - *Institutional → individual researcher*
  - *Behavioral expectations set by the commercial/mobile web*



User experience

- Due to its open eligibility policy, Merritt will always provide a more generic UX than special-purpose or disciplinary systems
- Shifting user roles, shifting expectations
  - *Institutional → individual researcher*
  - *Behavioral expectations set by the commercial web*
- Integration with extant services that better provide the desired UX
  - *DataShare* DataShare Open data for the global research community
  - *Research Hub* ResearchHub

DataShare

- *"The goal of the DataShare project is to catalyze widespread sharing of scientific research data"*
  datashare.ucsf.edu

  DataShare
  Open data for the global research community

  - *UCSF Clinical and Translational Science Institute*
    ctsi.ucsf.edu

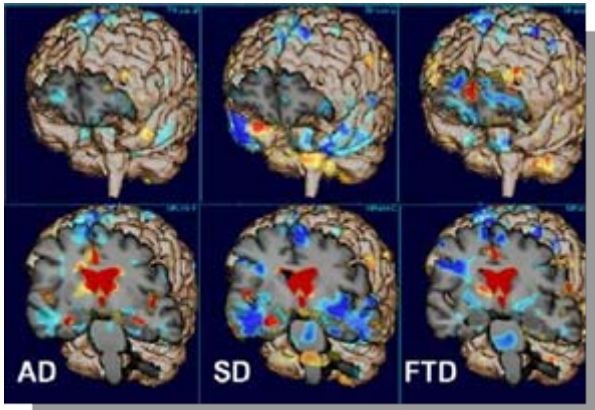
    CTSI at UCSF
    Accelerating Research to Improve Health

  - *UCSF Library*
    www.library.ucsf.edu

    UCSF LIBRARY
    and Center for Knowledge Management

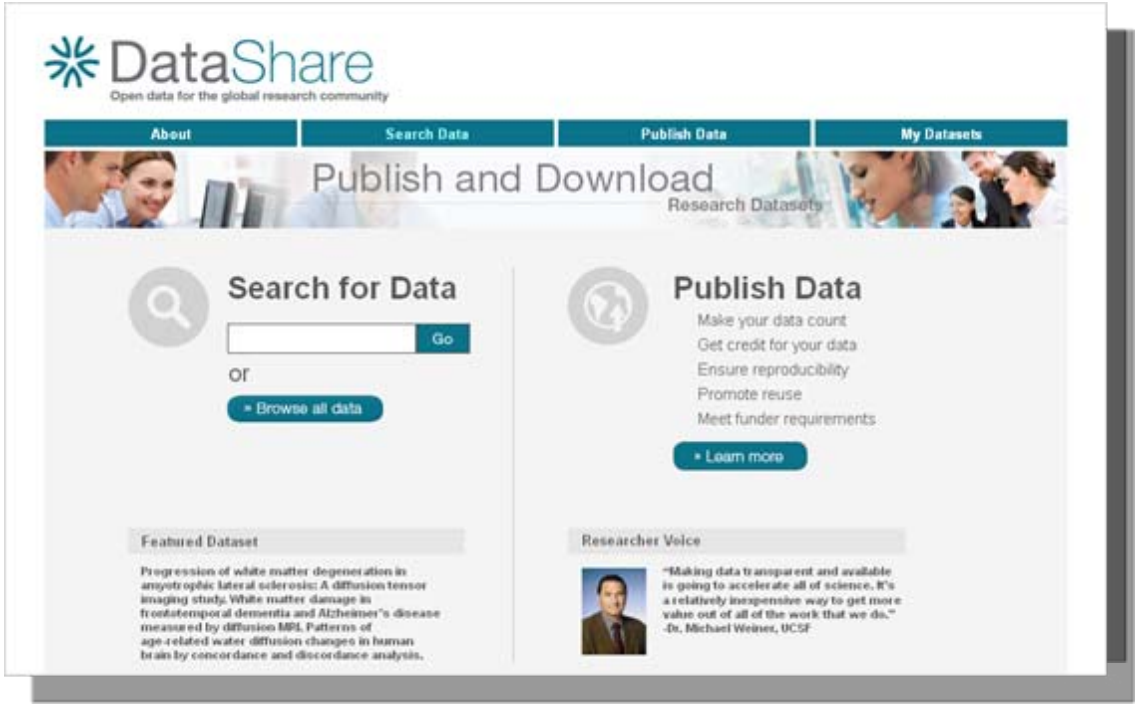
  - *UCSF Center for Imaging of Neurodegenerative Disease*
    www.radiology.ucsf.edu/cind

- Architecture
  - *DataShare submission client* (Ruby/Rails)
  - *Merritt curation repository*
  - *DataShare discovery portal* (XTF/Java)



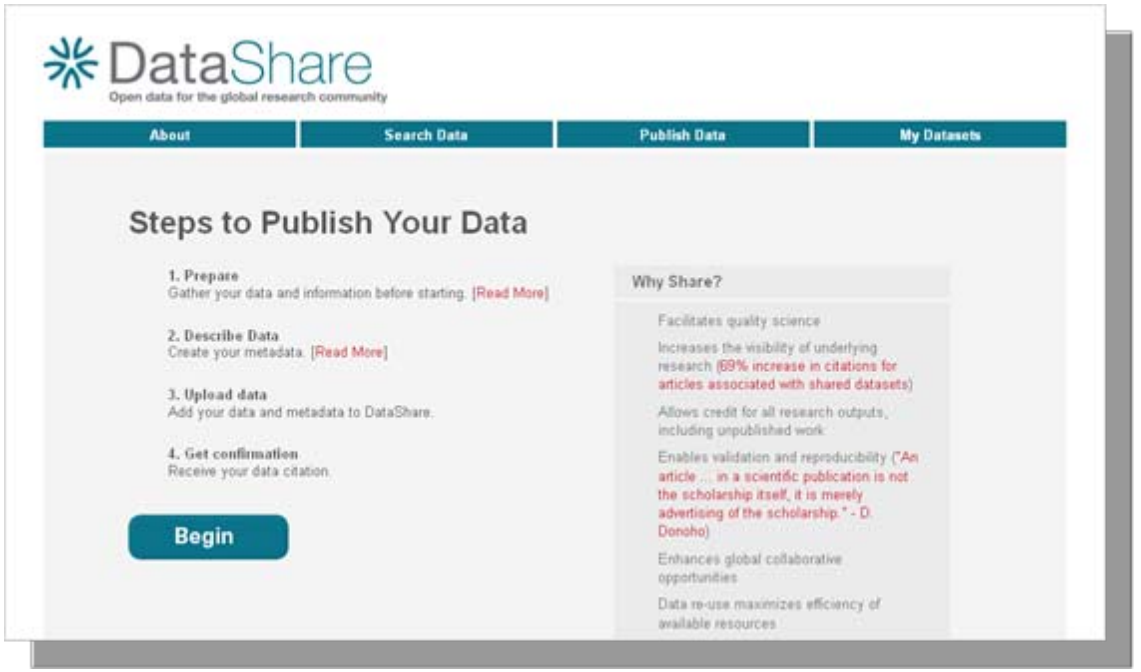
DataShare

- Prepare
- Describe
- Upload
- Curate
- Discover
- Share



DataShare

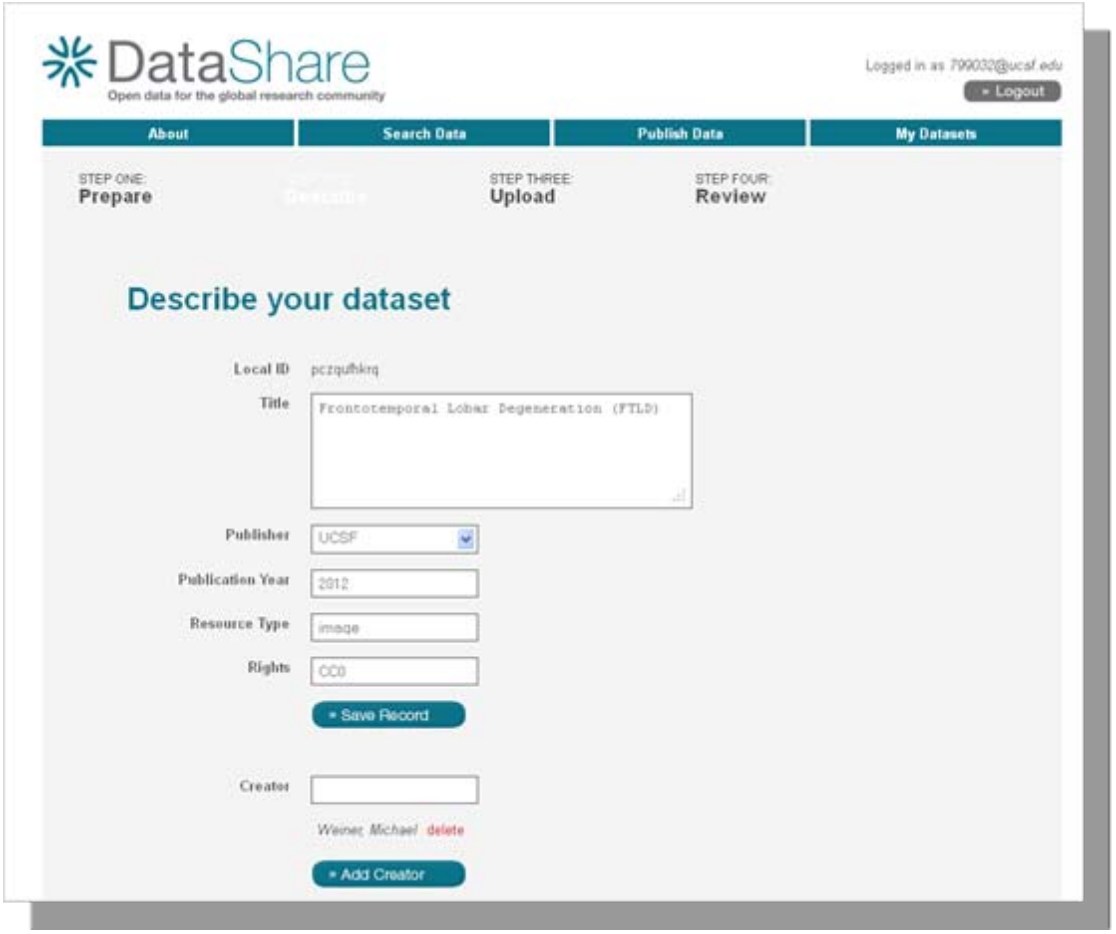
- Prepare
  - *Best practice advice*
- Describe
- Upload
- Curate
- Discover
- Share



DataShare

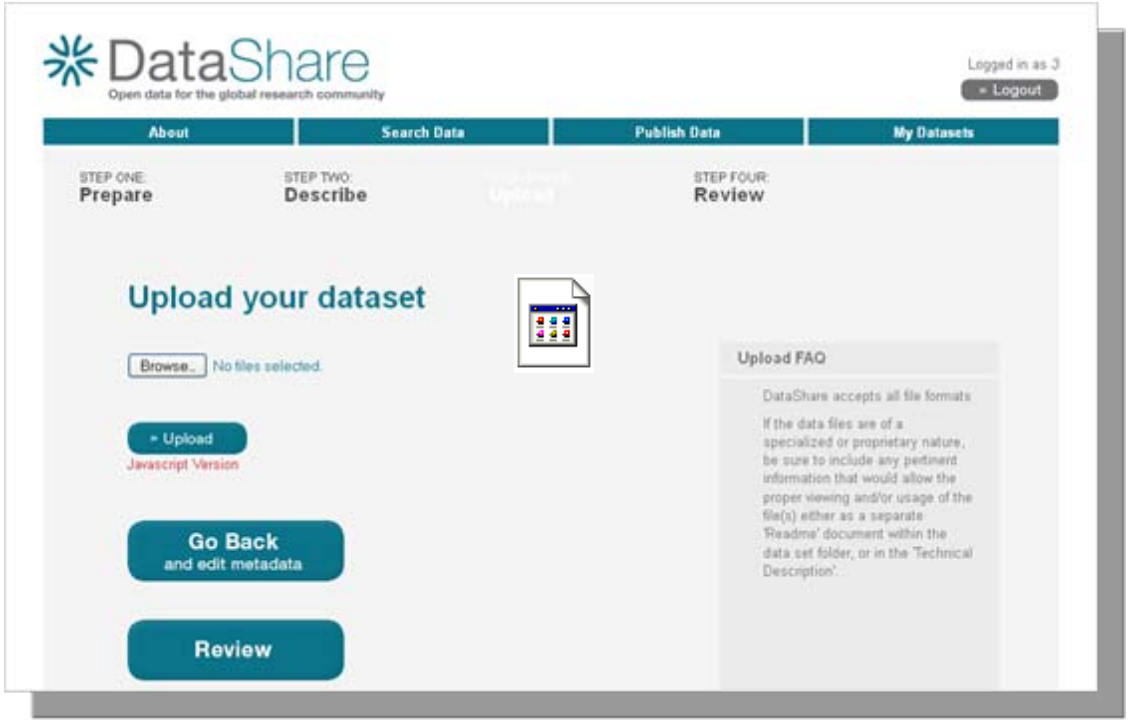
- Prepare
- Describe
  - *Schema-directed metadata editor*
  - *DataCite schema*

  schema.datacite.org
- Upload
- Curate
- Discover
- Share



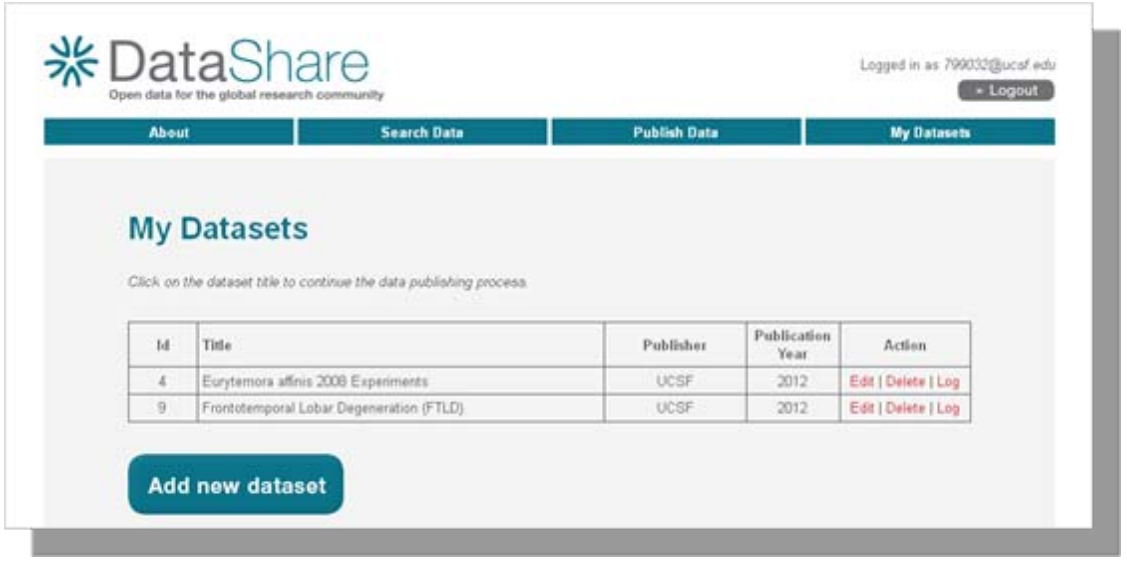
DataShare

- Prepare
- Describe
- Upload
  - *File browse or drag-n-drop*
- Curate
- Discover
- Share



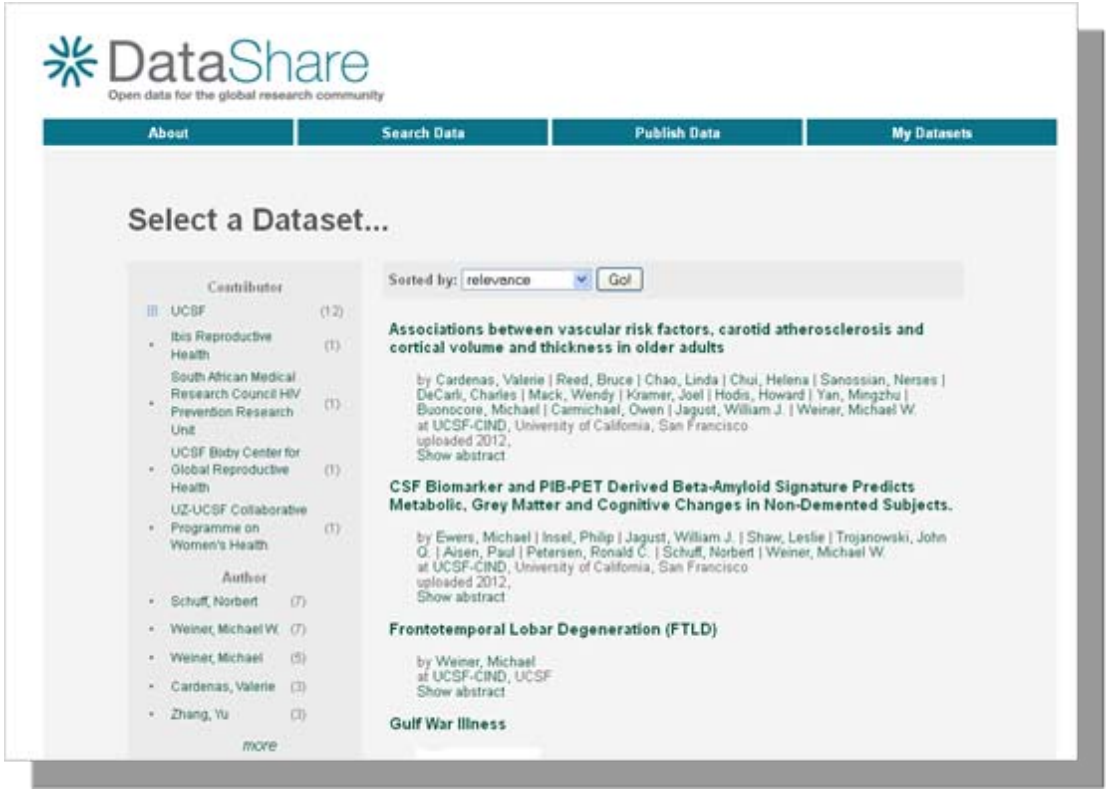
DataShare

- Prepare
- Describe
- Upload
- Curate
  - *Manage datasets*
- Discover
- Share



DataShare

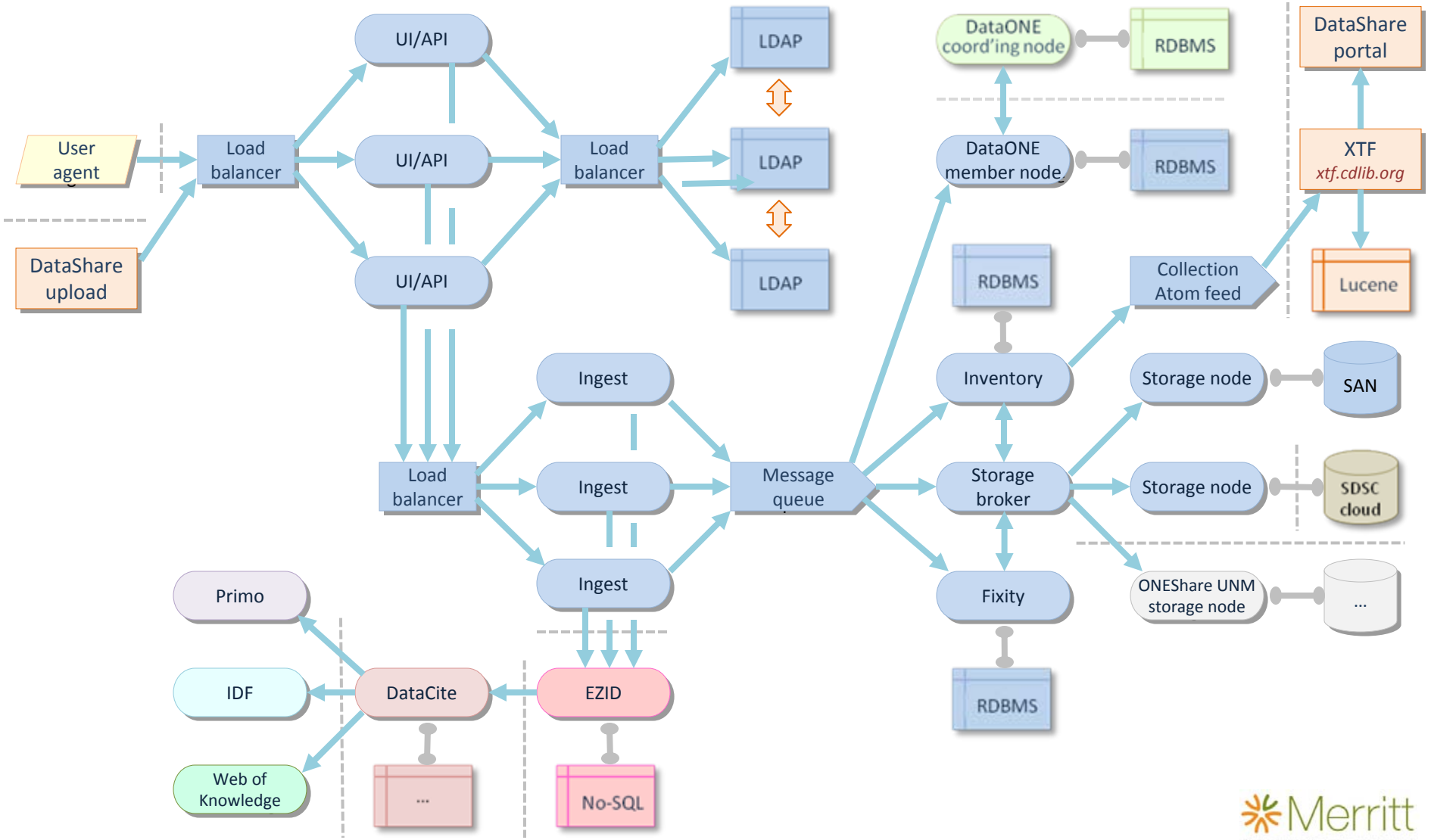
- Prepare
- Describe
- Upload
- Curate
- Discover
  - *Faceted search and browse*
- *Share*



DataShare

- Prepare
- Describe
- Upload
- Curate
- Discover
- Share
  - *DataONE*
  - *DataCite*
  - (soon) *Primo Web of Knowledge*
  - *SEO*

Merritt + DataShare



Merritt
Share, manage & preserve your content

Research Hub

- *"Research Hub provides powerful tools for content management and collaboration"*

ResearchHub

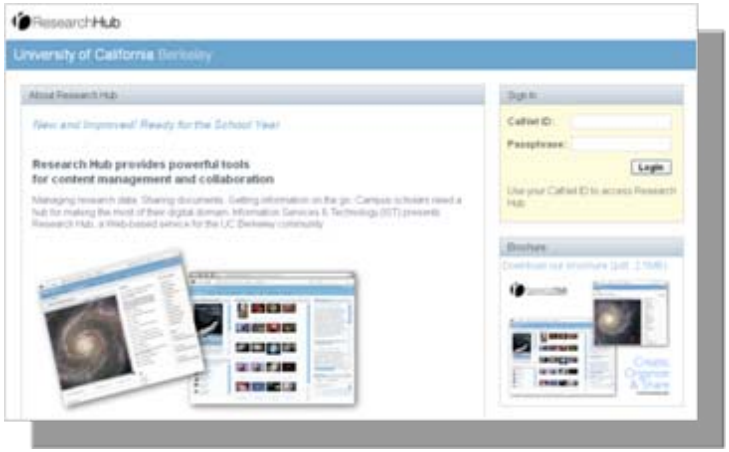
hub.berkeley.edu

  - *Alfresco CMS*

Alfresco

www.alfresco.com

  - *770 projects, 3,900 users*
    - Personal file management
    - Project collaboration
    - Departmental resource pooling
    - Research data management

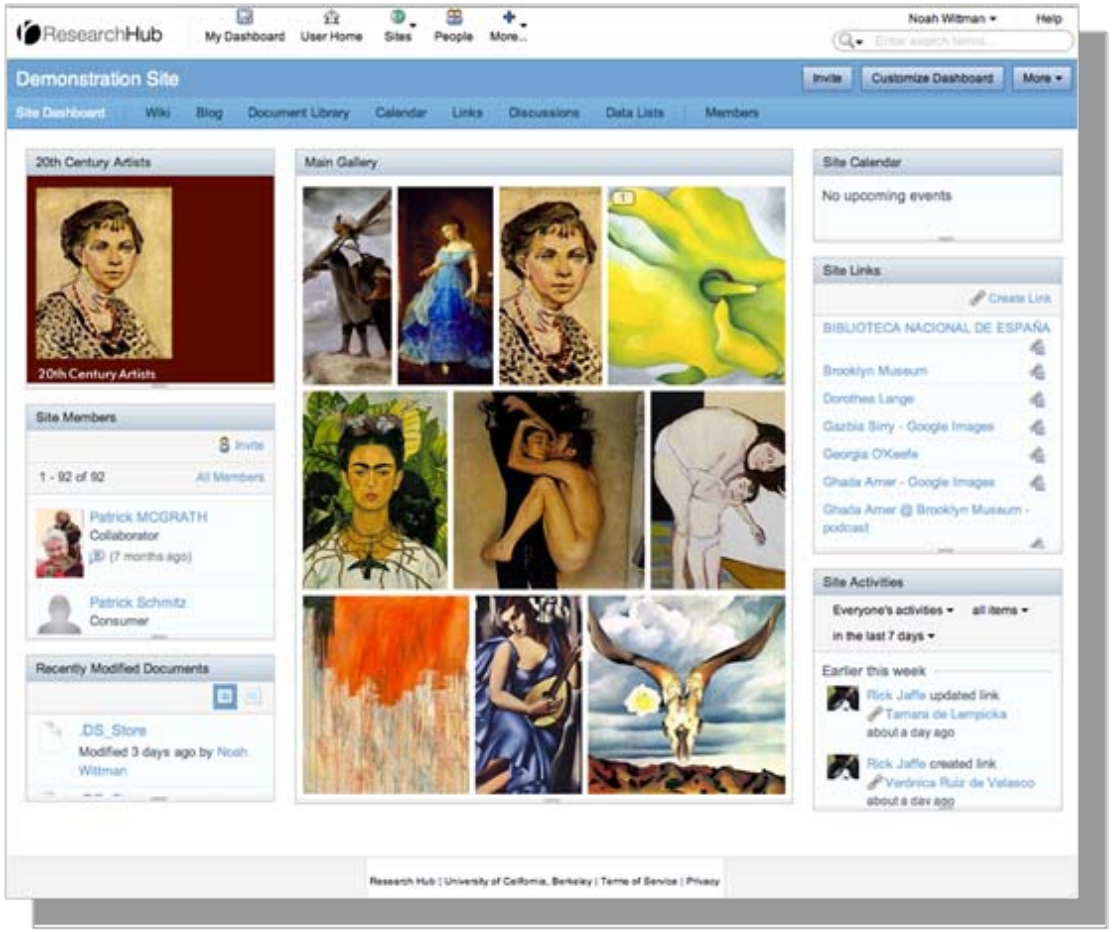


  - *Desktop sync, mobile app, Adobe Creative Suite*
  - *UC Berkeley Information Services and Technology*

ist.berkeley.edu

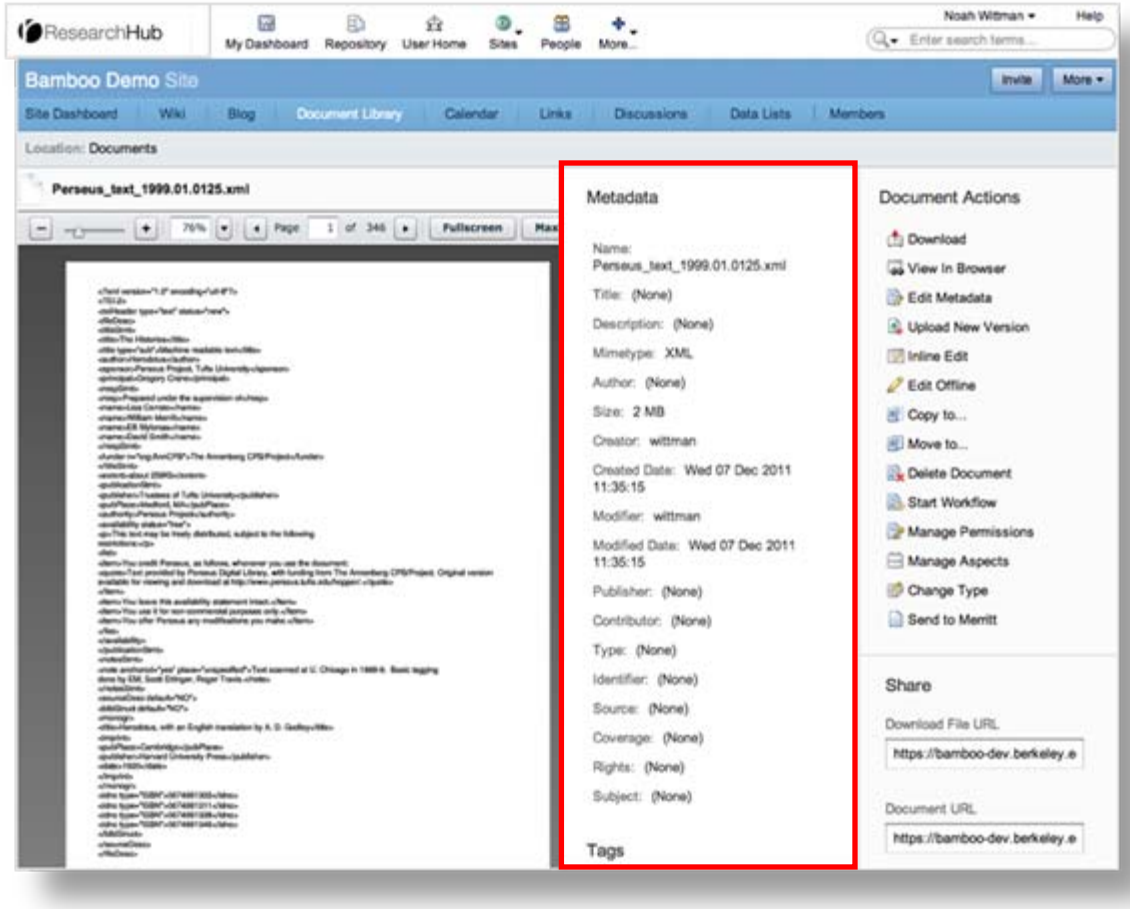
Research Hub

- Prepare
  - *Acquire and arrange*
- Describe
- Upload
- Curate
- Discover
- Share



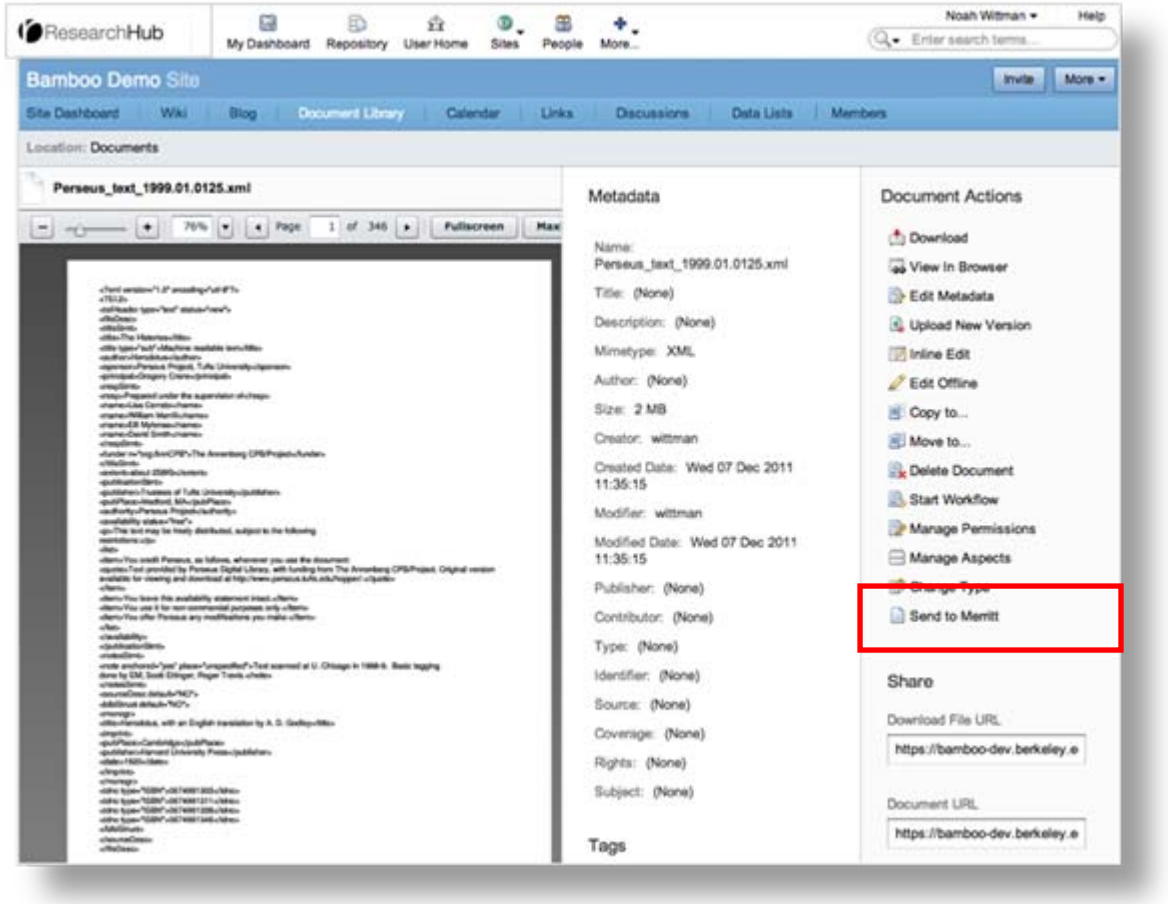
Research Hub

- Prepare
- Describe
  - *Schema-directed metadata editors*
- Upload
- Curate
- Discover
- Share



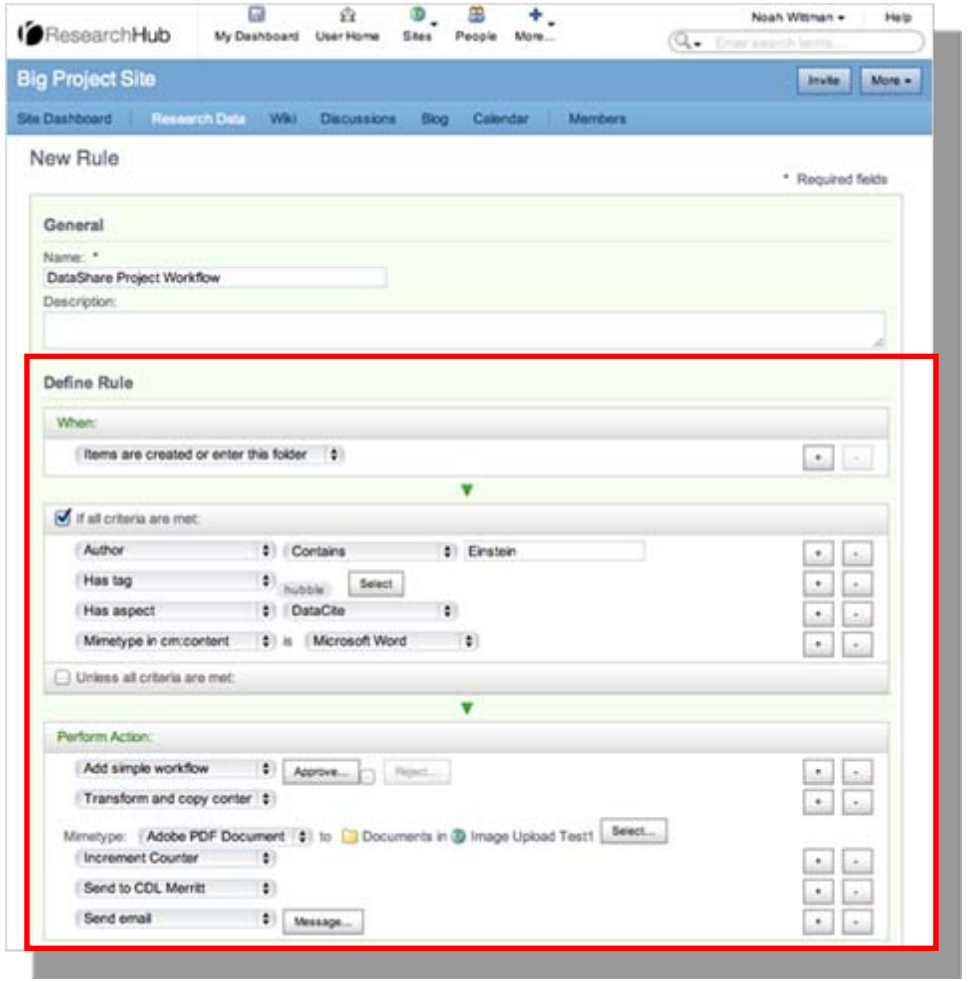
Research Hub

- Prepare
- Describe
- Upload
  - *Direct action*
- Curate
- Discover
- Share



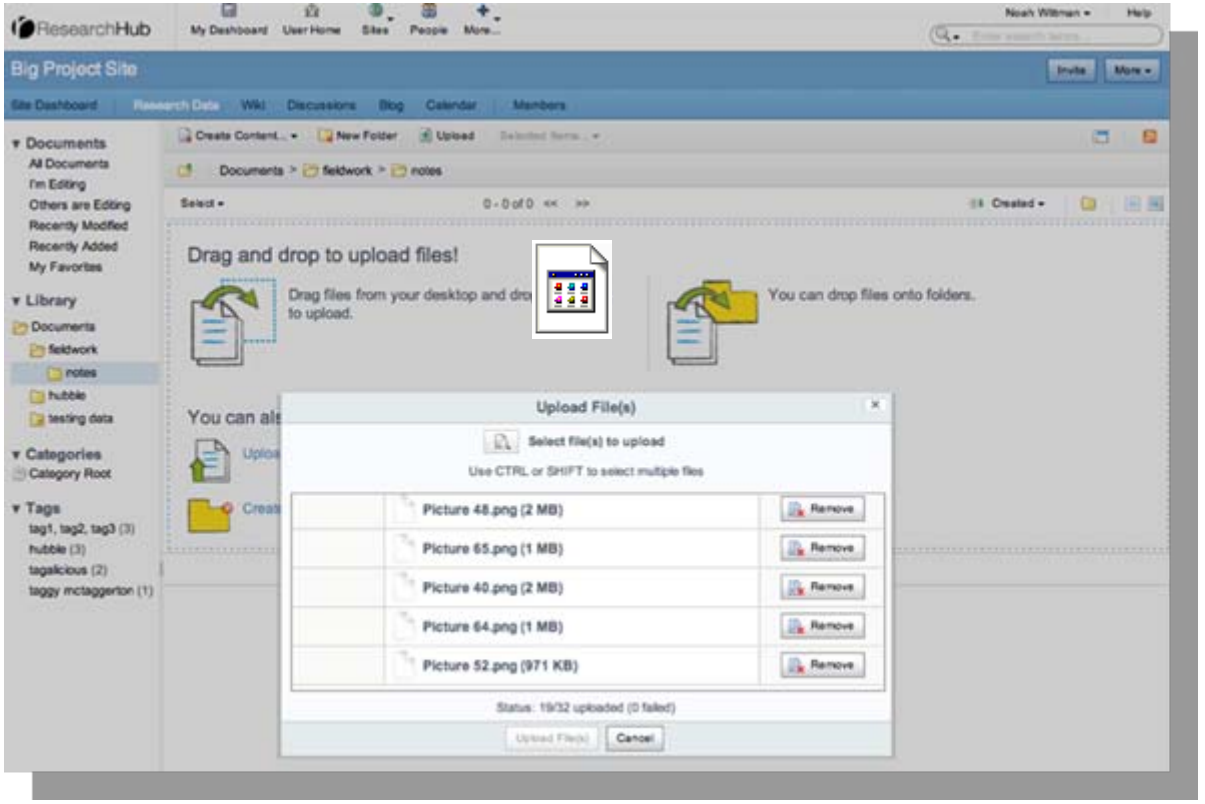
Research Hub

- Prepare
- Describe
- Upload
  - *Policy-based workflow rules*
- Curate
- Discover
- Share



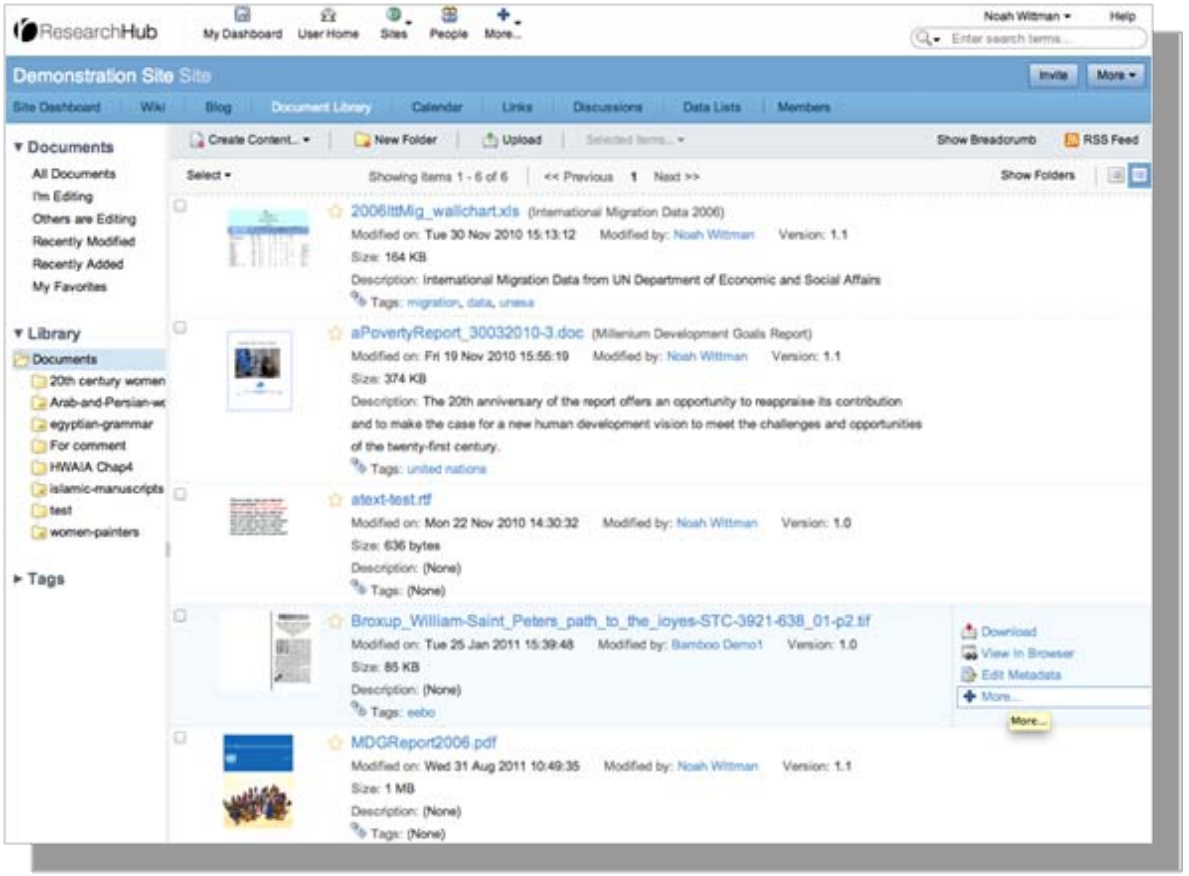
Research Hub

- Prepare
- Describe
- Upload
  - *Drag-and-drop*
- Curate
- Discover
- Share



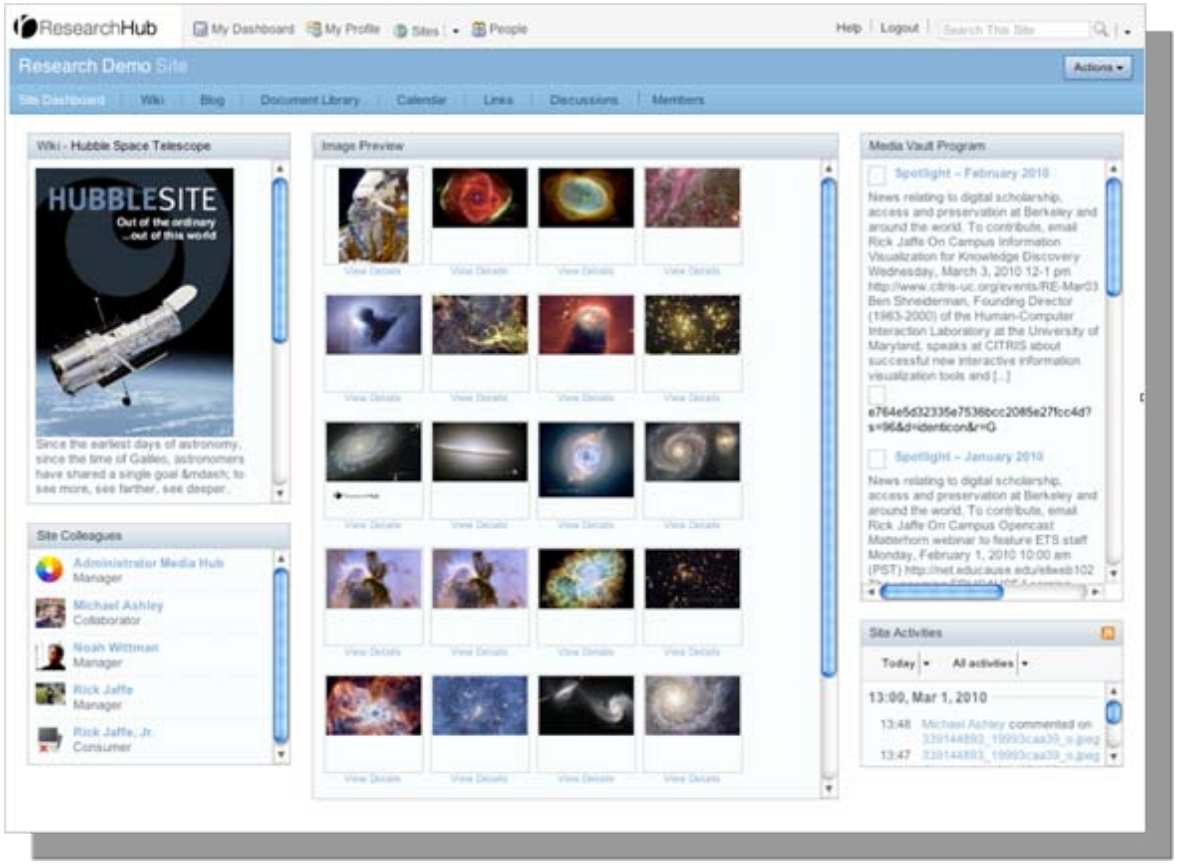
Research Hub

- Prepare
- Describe
- Upload
- Curate
  - *Manage datasets*
- Discover
- Share



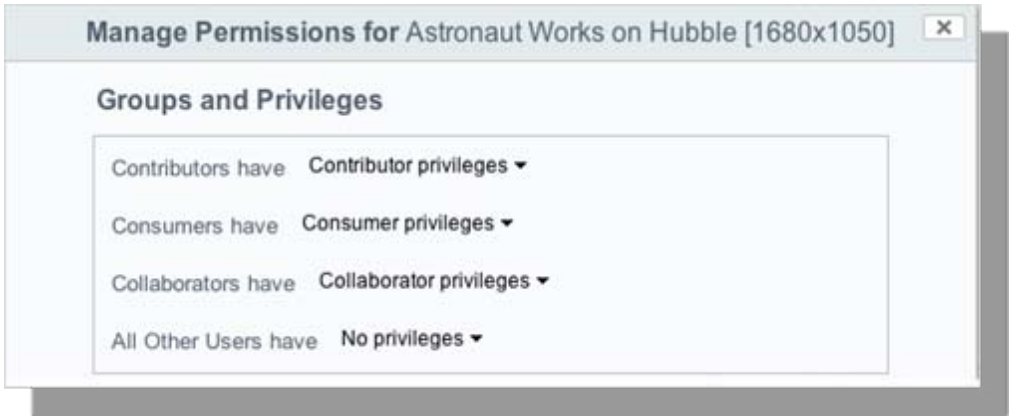
Research Hub

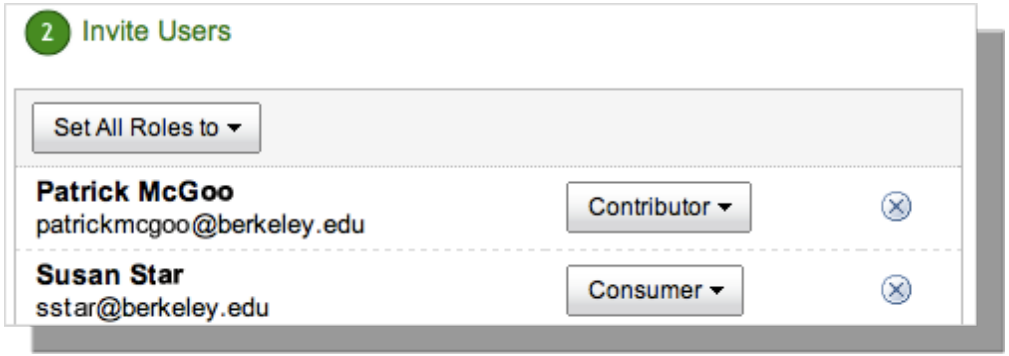
- Prepare
- Describe
- Upload
- Curate
- Discover
  - *Search / browse*
- Share



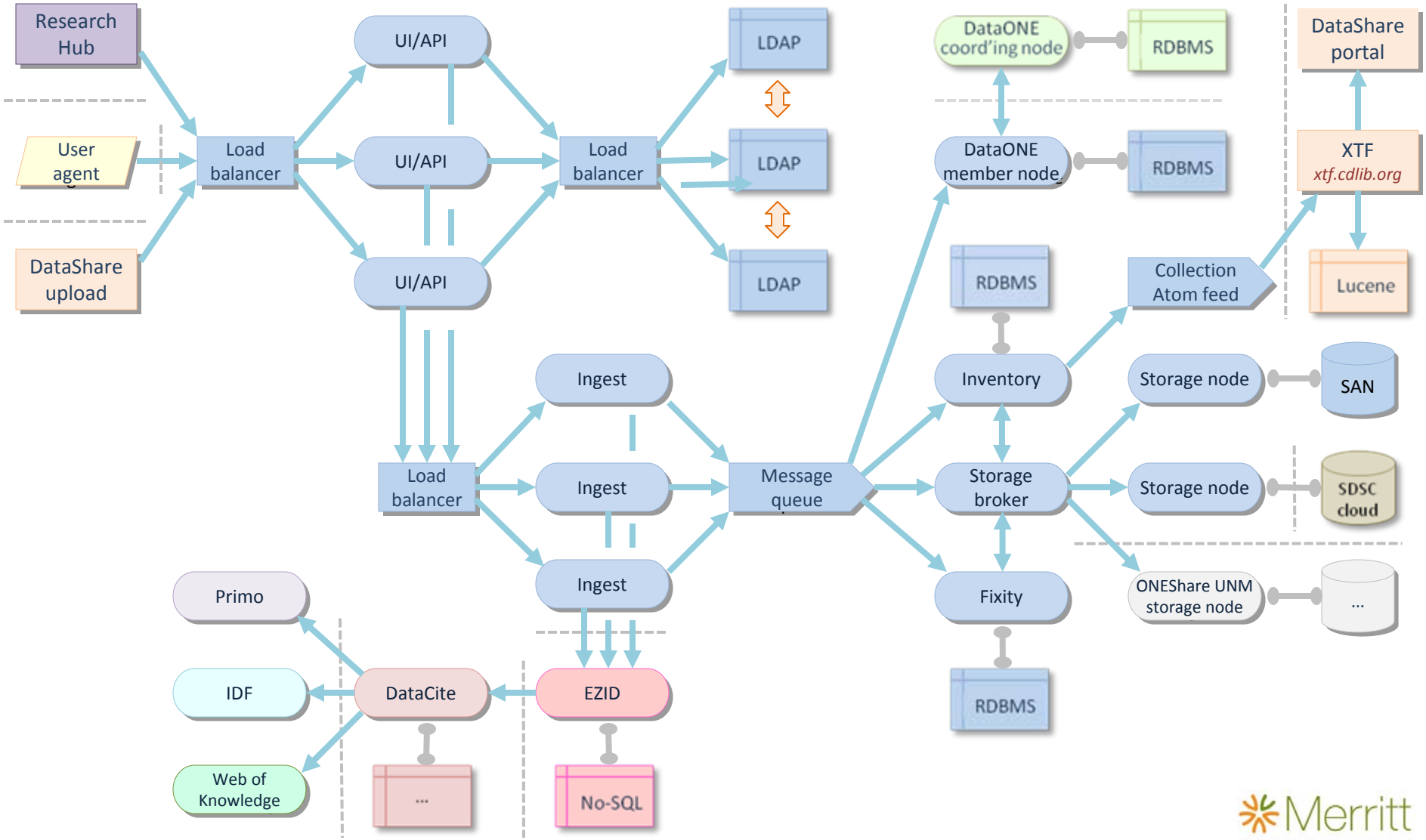
Research Hub

- Prepare
- Describe
- Upload
- Curate
- Discover
- Share
  - *Curatorial invitation*





Merritt + DataShare + Research Hub



Merritt
Share, manage & preserve your content

Future integrations

- UCTrust/InCommon federation

InCommon.

Incommon.org

- Open Context archaeological portal

OPEN CONTEXT
Web-based research data publication

opencontext.org

- Nuxeo

nuxeo / Content Management Platform For Business Applications

www.nuxeo.com

  - *UC system-wide DAMS*

- Islandora

islandora

islandora.ca

  - *Fedora → Merritt*

- DPN

DPN
THE DIGITAL PRESERVATION NETWORK

www.dpn.org

Sharing research through repositories

- Conform to institutional policy, publication requirements, and funder mandates
- Pro-active curation of valuable research outputs
- Stable citation and access
- High visibility publication and discovery
- Use metrics

Sharing research through repositories

- Conform to institutional policy, publication requirements, and funder mandates
- Pro-active curation of valuable research outputs
- Stable citation and access
- High visibility publication and discovery
- Use metrics
- Repository layering as an appropriate division of labor
  - *Exploiting existing capabilities already in local use*

For more information

Merritt

www.cdlib.org/uc3/merritt
uc3@ucop.edu

Stephen Abrams *David Loy*
Patricia Cruse *Mark Reyes*
Shirin Faenza *Joan Starr*
Scott Fisher *Carly Strasser*
Erik Hetzner *Marisa Strong*
Joshua Hubbard *Bhavitavya Vedula*
Greg Janée *Kenneth Weiss*
John Kunze *Perry Willet*
Rosalie Lack

DataShare

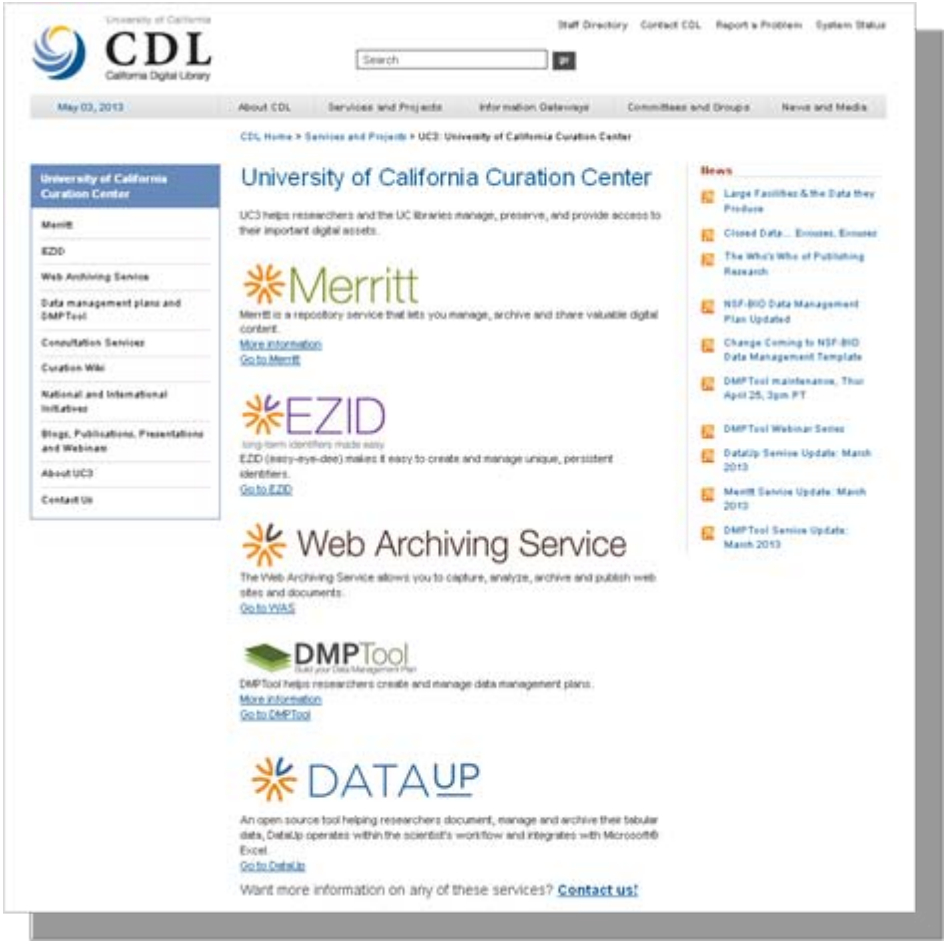
datashare.ucsf.edu

Geoffrey Boushey *Julia Kochi*
Anirvan Chatterjee *Angela Rizk-Jackson*
Maninder Kahlon *Michael Weiner*

Research Hub

hub.berkeley.edu

Ian Crew *Michael McCarthy* (Tribloom)
Noah Wittman *Patrick McGrath*



www.slideshare.net/UC3/ndsa-2013abramsintegratingrepositoriesfordatasharing