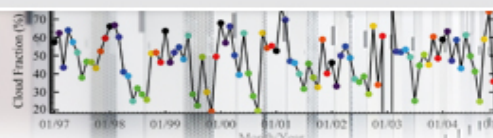


U.S. DEPARTMENT OF  
**ENERGY**



# Scientific Knowledge Management

- ▶ Scientists generate massive amounts of data
  - Experiments
  - Simulations
  - Instruments
- ▶ Processing this data produces knowledge
  - Data analysis
  - Data mining
  - Machine learning
  - Visualization



## ARM Data Archive

[Login](#)

Data collected through the routine operations and scientific field experiments of the ARM Climate Research Facility are stored at and distributed through the Archive. These data are available free of charge to the public and can be accessed through any of the interfaces below. Upon selection of an interface, a new window will ask you to sign in, or, if not already registered with the Archive, to complete the free and easy registration process.

### Get routine ARM data



#### **NEW** [Data Discovery V2](#) [ ? ]

Please try the new Unified Data Discovery Interface with built-in [Data Browser](#) capability. New interface allows filtered and faceted search logic, multi-pass data selection, filtering data based on data quality, graphical views of data quality and availability, direct access to data quality reports, and data plots. Please provide any feedback or comments to [armarchive@arm.gov](mailto:armarchive@arm.gov)



#### [Data Browser](#) [ ? ]

Note: "Data Browser" is now available within the Data Discovery.



#### [Thumbnail Browser](#) [ ? ]

View prepared plots of data to quickly find data of interest to you. The thumbnail browser uses location, measurement type and date range selections to retrieve data plot thumbnails that the user can browse. You can also download high-resolution images of the data plots, or download the data files.

### Plot previously ordered data




#### [NCVWeb](#) [ ? ]

NCVWeb is an interactive NetCDF data plotting tool users can use to plot the data they have ordered from the archive, or plot regular standing data orders, eliminating the need for separate visualization software. It has many powerful features such as producing detailed tables of NetCDF file contents, data extraction, generating statistics, and plotting one variable against another.

### Notice

This page has been replaced by <https://www.arm.gov/data>. Please change your bookmarks to use the new location. Some links on this page may no longer work.

# http://www.archive.arm.gov/discovery/#v/r results/s/fcat::sfcprop

**ARM**  
CLIMATE RESEARCH FACILITY

DATA DISCOVERY  
SEARCH RESULTS

[HOME](#) [DATA SEARCH](#) [DATASTREAM SEARCH](#)

SEARCH

Search Text:

Start Date:  End Date:   
(Start Date) (End Date)

[\[Expand All\]](#) [\[Close All\]](#) [Clear](#) [Apply](#)

CATEGORIES 5

- ☒ Surface Properties 3,552
  - ☐ Atmospheric State 797
  - ☐ Radiometric 247
  - ☐ Cloud Properties 138
  - ☐ Atmospheric Carbon 95

[Clear](#) [Apply](#)

DATA PRODUCTS 50+

SUBCATEGORIES 10

MEASUREMENTS 50+

SITES 28

FACILITIES 50+

FIELD CAMPAIGNS 34

SOURCES TYPES 10

SOURCE 50+

DATASTREAMS 50+

Home / Data Discovery

Search Results

To search for and request data, select a category, measurement, site, or source. Use the Start Date and End Date below to limit the data results timeline. Use the checkboxes below to add a data product to the Data Cart.

[Remove All](#) [Categories: Surface Properties](#)

[ROUTINE DATA](#) [PI / CAMPAIGN DATA](#)

[DATA UNRELIABLE](#) [DATA QUESTIONABLE](#) [DATA MISSING](#) [DATA NOTE](#) [LIMITED ACCESS](#)

1996-02-04  2017-01-05 Applies to this timeline view only. Sort by: Priority

Showing 1-20 of 3,552 measurements Page Size: 20

	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
1 <input type="checkbox"/> surfspecalb1mlawer a1 @ nsa C1 // Surface Spectral Albedo from first Mlawer algorithm <a href="#">[Expand]</a>	<div><div><div>★1</div><div>Surface albedo // Spectral albedo estimated from mfr albedos using surface type and Mlawer algorithm at 25m tower</div><div></div></div></div>																				
2 <input type="checkbox"/> surfspecalb1mlawer a1 @ sap C1 // Surface Spectral Albedo from first Mlawer algorithm <a href="#">[Expand]</a>	<div><div><div>★1</div><div>Surface albedo // Spectral albedo estimated from mfr albedos using surface type and Mlawer algorithm at 10m tower</div><div></div></div></div>																				
3 <input type="checkbox"/> ir10m b1 @ sap C1 // Infrared Thermometer: Ground surface temperature averaged 60-sec at 10-meter height <a href="#">[Expand]</a>	<div><div><div>★1</div><div>Surface skin temperature // Surface Infra-Red Temperature</div><div></div></div></div>																				
4 <input type="checkbox"/> swats b1 @ sap E1 // Soil Water and Temperature Profiling System (SWATS): soil temp & water profiles <a href="#">[Expand]</a>	<div><div><div>★1</div><div>Soil moisture // Volumetric Water Content, East Profile</div><div></div></div></div>																				
5 <input type="checkbox"/> swats b1 @ sap E1 // Soil Water and Temperature Profiling System (SWATS): soil temp & water profiles <a href="#">[Expand]</a>	<div><div><div>★1</div><div>Soil moisture // Volumetric Water Content, West Profile</div><div></div></div></div>																				
6 <input type="checkbox"/> swats b1 @ sap E10 // Soil Water and Temperature Profiling System (SWATS): soil temp & water profiles <a href="#">[Expand]</a>	<div><div><div>★1</div><div>Soil moisture // Volumetric Water Content, East Profile</div><div></div></div></div>																				
7 <input type="checkbox"/> swats b1 @ sap E10 // Soil Water and Temperature Profiling System (SWATS): soil temp & water profiles <a href="#">[Expand]</a>	<div><div><div>★1</div><div>Soil moisture // Volumetric Water Content, West Profile</div><div></div></div></div>																				
8 <input type="checkbox"/> swats b1 @ sap E11 // Soil Water and Temperature Profiling System (SWATS): soil temp & water profiles <a href="#">[Expand]</a>	<div><div><div>★1</div><div>Soil moisture // Volumetric Water Content, East Profile</div><div></div></div></div>																				
9 <input type="checkbox"/> swats b1 @ sap E11 // Soil Water and Temperature Profiling System (SWATS): soil temp & water profiles <a href="#">[Expand]</a>	<div><div><div>★1</div><div>Soil moisture // Volumetric Water Content, West Profile</div><div></div></div></div>																				

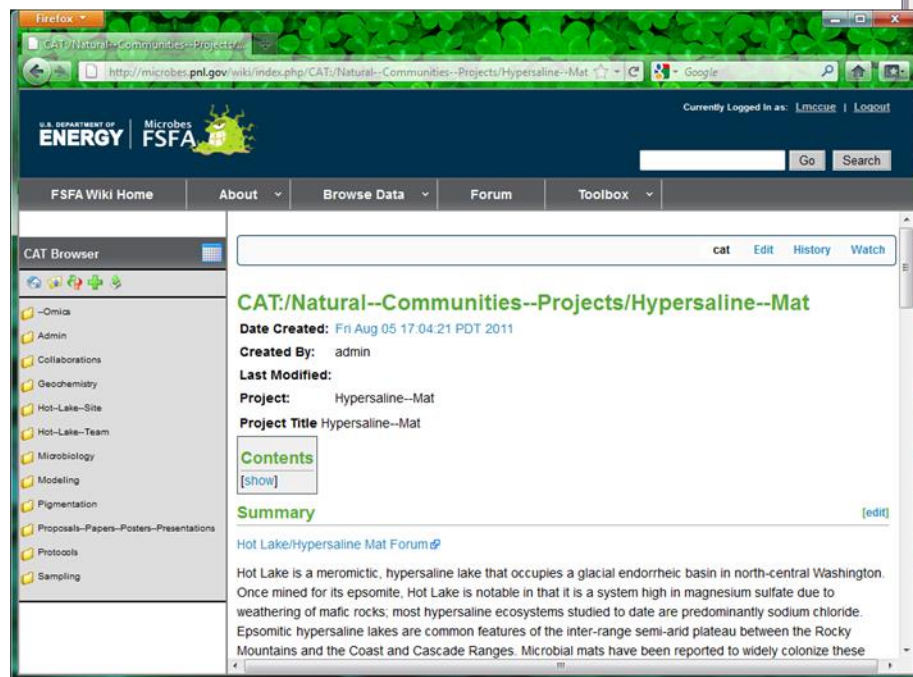
# Great Site, but ....

- ▶ Expensive to build
  - Millions of dollars
- ▶ Expensive to run/maintain
  - Own servers, disks, administration, data center
- ▶ Expensive to evolve and scale
  - More and more data
    - We're talking petabytes ...
  - More and more features needed
  - More and more to support





# Bioinformatics Example



Firefox

CAT: Natural--Communities--Projects/Hypersaline--Mat

http://microbes.pnl.gov/wiki/index.php/CAT:/Natural--Communities--Projects/Hypersaline--Mat

U.S. DEPARTMENT OF ENERGY Microbes FSFA

Currently Logged In as: [lmcucuz](#) | [Logout](#)

FSFA Wiki Home About Browse Data Forum Toolbox

CAT Browser

- Omics
- Admin
- Collaborations
- Geochemistry
- Hot-Lake--Site
- Hot-Lake--Team
- Microbiology
- Modeling
- Pigmentation
- Proposals--Papers--Posters--Presentations
- Protocols
- Sampling

CAT: Natural--Communities--Projects/Hypersaline--Mat

Date Created: Fri Aug 05 17:04:21 PDT 2011

Created By: admin

Last Modified:

Project: Hypersaline--Mat

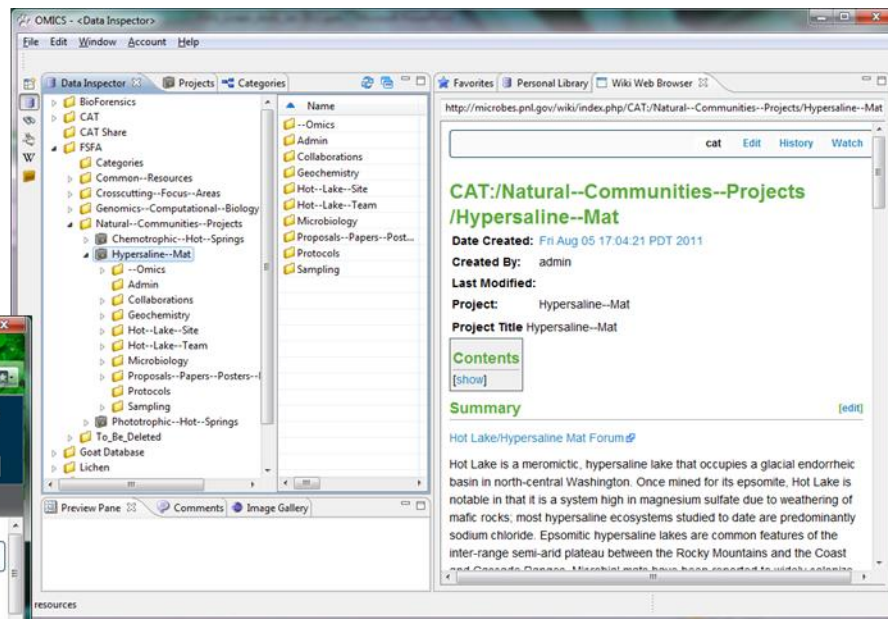
Project Title Hypersaline--Mat

Contents [show]

Summary [edit]

[Hot Lake/Hypersaline Mat Forum](#)

Hot Lake is a meromictic, hypersaline lake that occupies a glacial endorheic basin in north-central Washington. Once mined for its epsomite, Hot Lake is notable in that it is a system high in magnesium sulfate due to weathering of mafic rocks; most hypersaline ecosystems studied to date are predominantly sodium chloride. Epsomitic hypersaline lakes are common features of the inter-range semi-arid plateau between the Rocky Mountains and the Coast and Cascade Ranges. Microbial mats have been reported to widely colonize these



OMICS - Data Inspector

File Edit Window Account Help

Data Inspector Projects Categories

- BioForensics
- CAT
- CAT Share
- FSFA
  - Categories
  - Common--Resources
  - Crosscutting--Focus--Areas
  - Genomics--Computational--Biology
  - Natural--Communities--Projects
    - Chemotrophic--Hot--Springs
    - Hypersaline--Mat
      - Omics
      - Admin
      - Collaborations
      - Geochemistry
      - Hot-Lake--Site
      - Hot-Lake--Team
      - Microbiology
      - Proposals--Papers--Post...
      - Protocols
      - Sampling
    - Phototrophic--Hot--Springs
    - To\_Be\_Deleted
    - Goat Database
    - Lichen

Name

- Omics
- Admin
- Collaborations
- Geochemistry
- Hot-Lake--Site
- Hot-Lake--Team
- Microbiology
- Proposals--Papers--Post...
- Protocols
- Sampling

Preview Pane Comments Image Gallery

resources

http://microbes.pnl.gov/wiki/index.php/CAT:/Natural--Communities--Projects/Hypersaline--Mat

cat Edit History Watch

CAT: Natural--Communities--Projects/Hypersaline--Mat

Date Created: Fri Aug 05 17:04:21 PDT 2011

Created By: admin

Last Modified:

Project: Hypersaline--Mat

Project Title Hypersaline--Mat

Contents [show]

Summary [edit]

[Hot Lake/Hypersaline Mat Forum](#)

Hot Lake is a meromictic, hypersaline lake that occupies a glacial endorheic basin in north-central Washington. Once mined for its epsomite, Hot Lake is notable in that it is a system high in magnesium sulfate due to weathering of mafic rocks; most hypersaline ecosystems studied to date are predominantly sodium chloride. Epsomitic hypersaline lakes are common features of the inter-range semi-arid plateau between the Rocky Mountains and the Coast and Cascade Ranges. Microbial mats have been reported to widely colonize these

# Implications

- ▶ Thousands of different scientific research projects
- ▶ All with:
  - 'similar' data management and analysis needs
  - Limited budgets
  - Scaling and growth requirements



# So what do scientist do?

- ▶ Use wikis
- ▶ Use simple web site
- ▶ Use Sharepoint



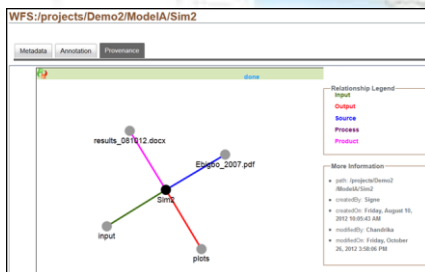


# What about Content Management Technologies?



- ▶ We built Velo 7 years ago at PNNL
- ▶ Used by many science projects
- ▶ Built on Alfresco
- ▶ Works well but has:
  - scalability challenges
  - Not easy for a new science project to deploy
    - Software/hardware configuration

# That's where you come in ...



Collaboration  
Data access  
Visualization



Security  
Upload/download  
Navigation  
Site configuration



Data storage  
Search  
Analysis  
Metadata



**Pacific Northwest**  
NATIONAL LABORATORY

*Proudly Operated by **Battelle** Since 1965*



designed by  freepik.com



# Before we go in to the details

- ▶ Expect to be confused!!!!
- ▶ Ask questions!!!
- ▶ Don't be afraid to be creative
- ▶ You won't get all the requirements 'on a plate'
  - You will have to discover some ....
  - You can create some ....
- ▶ Technology selections are pretty open
  - AWS based
  - Data storage/database?
  - Web technologies?
  - Languages?

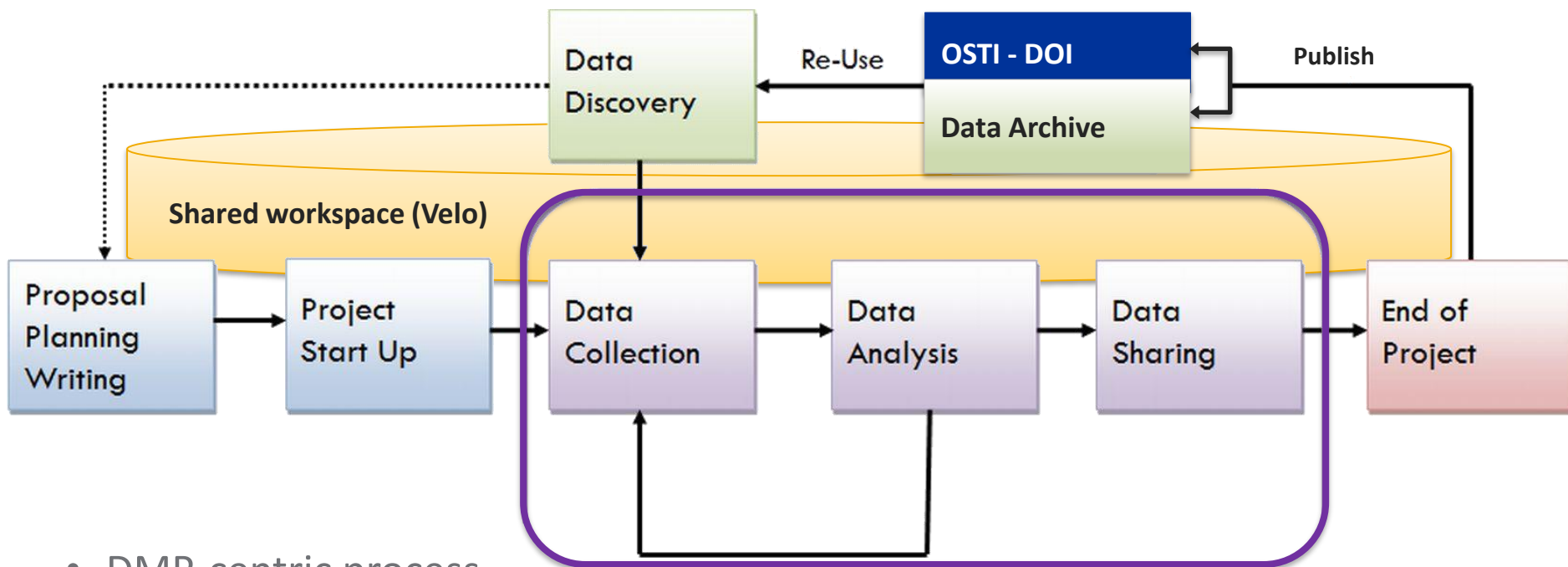




# PNNL - Introductions

- ▶ Gary Black
- ▶ Carina Lansing
- ▶ Ellen Porter
- ▶ Bibi Raju
- ▶ Zoe Gullen

## Full Lifecycle Support for Research Data



- DMP-centric process.
- Shared workspace for all project documents and research data.
- Seamless integration with PNNL systems from project proposal through publication.
- Long term: facilitate data discovery and reuse.

# Velo Scope

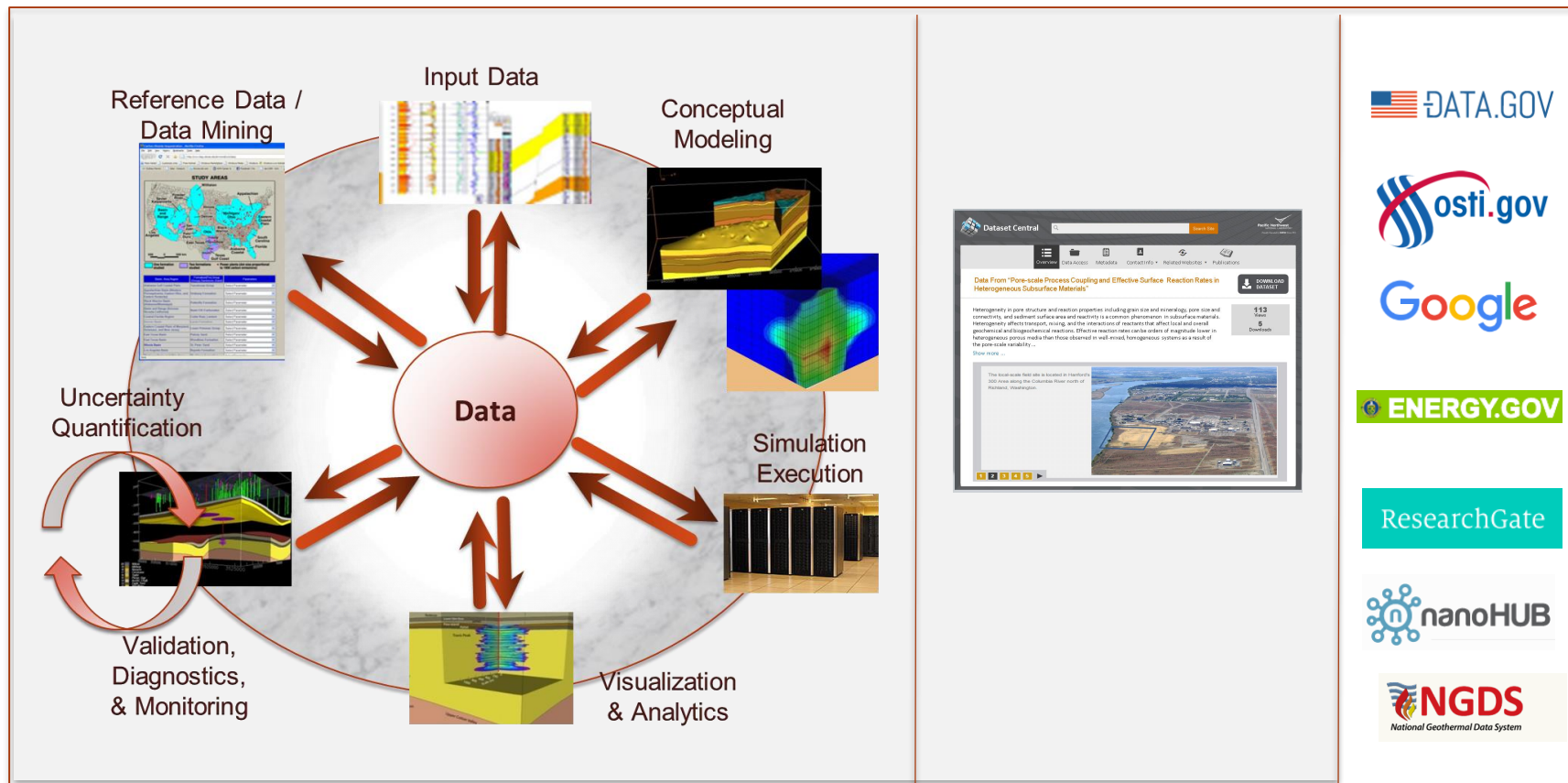
Do your work...



Publish...



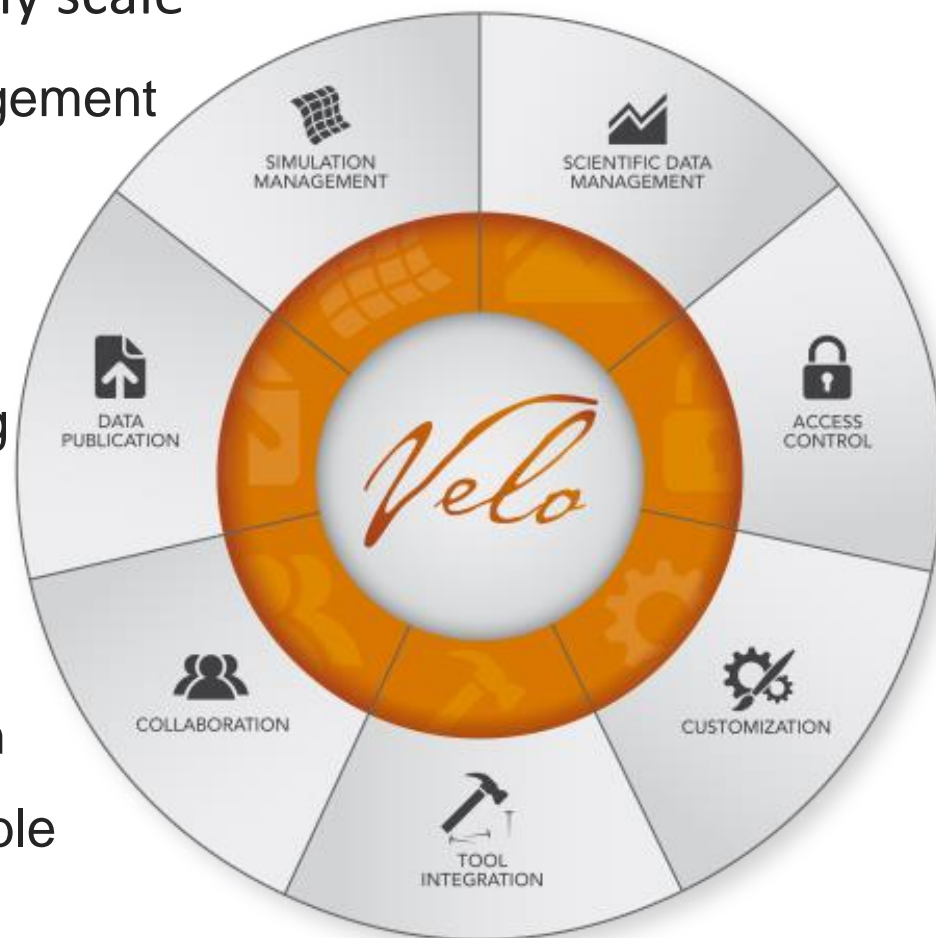
Discovery &  
Reuse



**Velo**

# Key Features

- ▶ Full life cycle, integrated support for modeling, simulation, analysis, and visualization at any scale
- ▶ Data/metadata/relationship management
- ▶ Tool Integration
  - Connect tools to data
  - Chain tools together
- ▶ Simulation management, including UQ ensembles
- ▶ Collaboration across multi-disciplinary teams
- ▶ Provenance – who did what, when
- ▶ Open, Extensible, and Customizable



- ▶ Facilitates reproducibility & discovery
  - Metadata is way harder to collect after the project is completed.
- ▶ Focused on integration – don't reinvent the wheel
- ▶ Improved organization
  - Easier for users to track where their different data parts are in a project
- ▶ Improved productivity
  - Less time spent in data orchestration tasks = more time for science
  - Low barrier of entry for HPC systems





**Pacific Northwest**  
NATIONAL LABORATORY

*Proudly Operated by **Battelle** Since 1965*

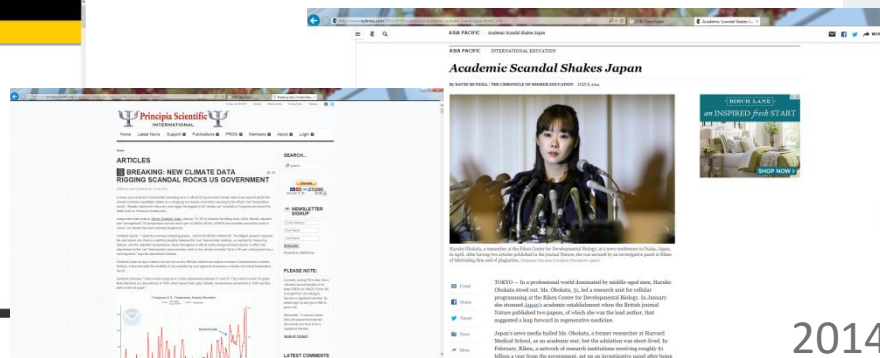
# Data Publication

# Legislative Drivers

- America Competes Reauthorization Act of 2010 & February 2013 Office of Science and Technology Policy (OSTP) Mandate – **Make the results of publicly funded research available for reuse**
- 2014 Data Management Plans (DMP) required for all new funding applications at all agencies
- Increased scrutiny of scientific results, demand for reproducibility as measure of quality and correctness
- Must comply by October 2016
- Data management & publication will soon affect lab appraisals.



2012



2014

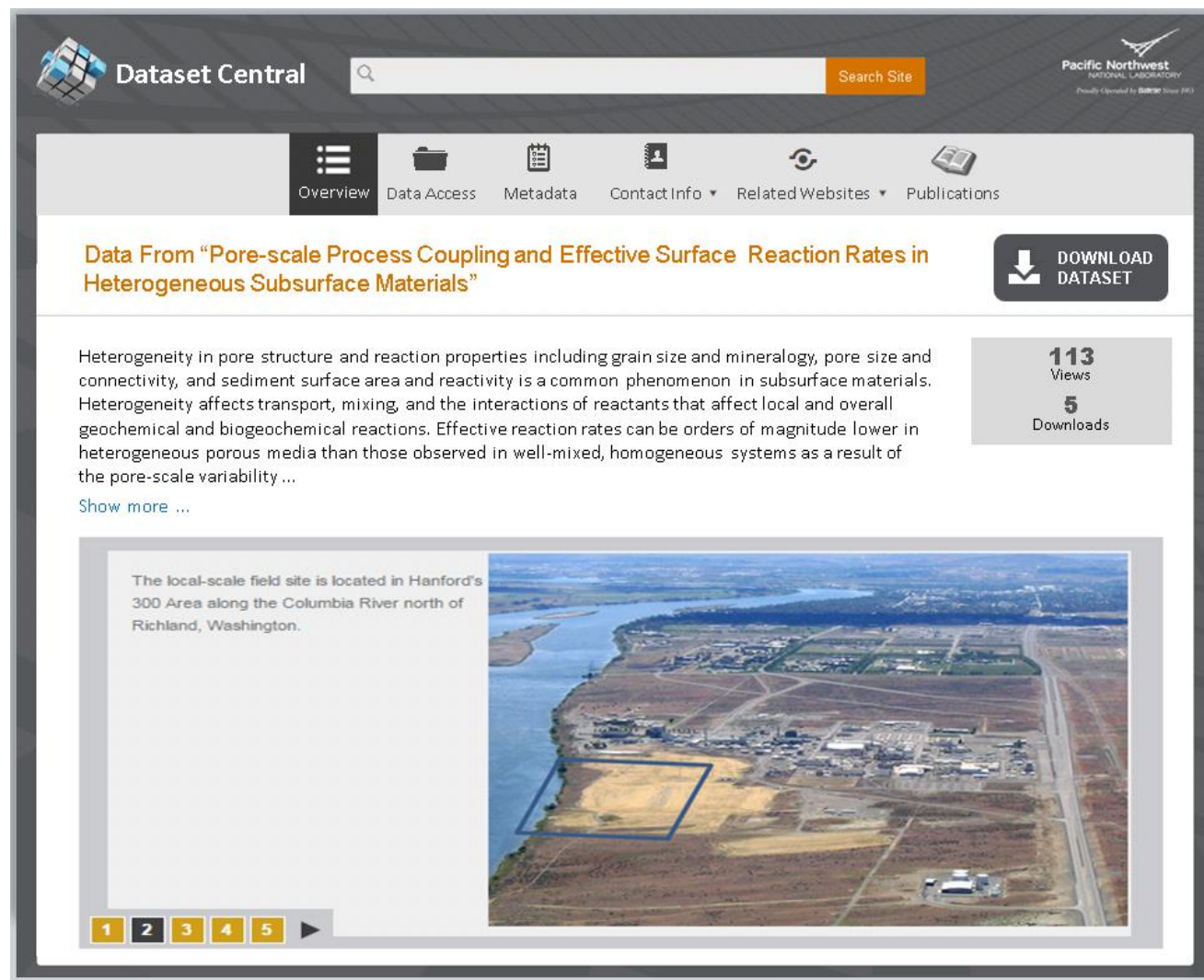


2014

2015

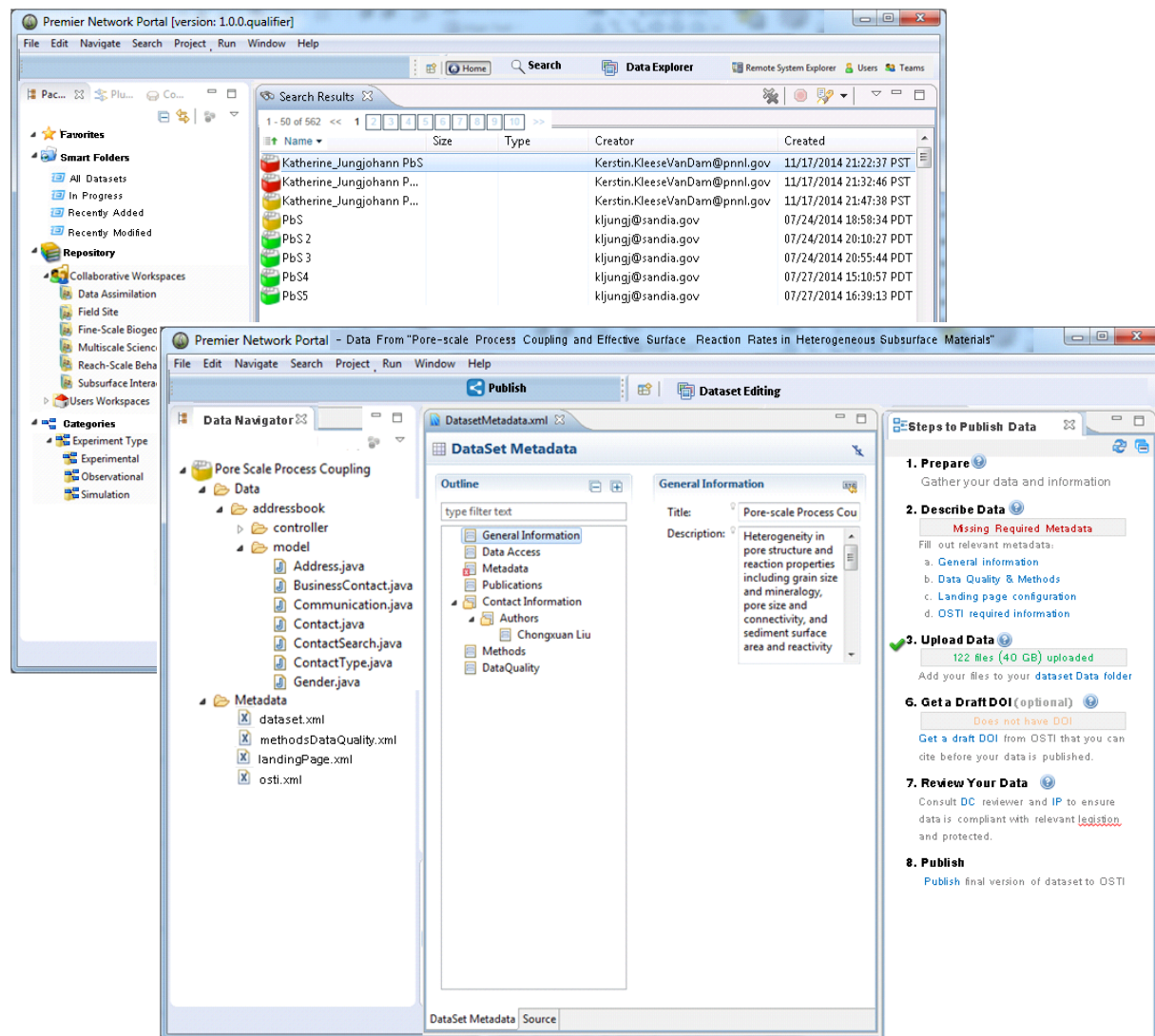
# Dataset Landing Page Design

- ▶ Designed for scientists
  - Embedded dataset visualization & analysis
  - Dataset provenance
  - Dataset browsing
- ▶ Fully compliant with DataCite and federal guidelines



# Dataset Management

- ▶ Support for dataset versioning and provenance
- ▶ Easier for scientists to use
  - More convenient metadata editing
  - More convenient file import
  - Publication compliance checklist
  - Improved navigation and search
- ▶ Custom landing page configuration





**Pacific Northwest**  
NATIONAL LABORATORY

*Proudly Operated by **Battelle** Since 1965*

# Velo Use Case: Akuna





# Velo Use Case: Akuna

## ► Akuna Architecture

