A Quick Overview of the SEAD DataNet project

Jim Myers (myersjd@umich.edu), 5th National Data Service Consortium Workshop
SEAD: Sustainable Environment - Actionable Data

- An NSF DataNet project started in October, 2011
- An international resource for sustainability science
- A provider of light-weight Data Services based on novel technical and business approaches:
  - Supporting the long-tail of research
  - Enabling active and social curation
  - Providing integrated lifecycle support for data

http://sead-data.net/

Margaret Hedstrom, PI
Praveen Kumar, co-PI
Jim Myers, co-PI
Beth Plale, co-PI
Why do we need “Data Patriotism”? 

The future needs your data! 

“And so, my fellow Americans, ask not what your data can do for you, ask what you can do for your data” 

Only you can prevent data loss!
Imagine Having to Ask: “Why Acquire Data, Analyze It, and Draw Conclusions?”

- Funding agencies urging researchers to acquire data!?!?
- Publishers mandating that authors analyze their data?!!
- An outcry for scientists to put more effort into draw conclusions to help drive progress and economic growth?!?
SEAD Overarching Concepts

• Self-service
• Simple basic defaults, rich customization, branding, and automation options
• Leverage incremental, informal active use to capture data and metadata from first sources
• Provide data-related (metadata-driven) services to active producers, curators, and users of data
• Simplify and automate curation and preservation processes using captured information and context
• Leverage existing institutional repository technologies and organizations to provide long-term storage

*Increase Value, Lower Costs, Increase Immediacy*
Consider SEAD: An NSF DataNet Partner

Where your project can request a secure, branded Project Space in the cloud...
Consider SEAD: An NSF DataNet Partner

Drag-and-drop your data, or import from other websites
Consider SEAD: An NSF DataNet Partner

Add Tags, Formal Terms (configurable), comments...

Use them to search, coordinate with colleagues

[Image of a user interface with dropdown menus and search results]

Write a Comment (use @ to notify others)

Mention Someone (name or email)

Anna Ovchinnikova
anyao@umich.edu

Search Results

Ecosystem of UMBS
Anna Ovchinnikova
Consider SEAD: An NSF DataNet Partner

- Create relationships

Or let your software write them!
- Restful API
- R Library

Cited by

Cited in

Has Correction

Uses Calibration

Generated Using

Uses Procedure

http://myproj.org/wiki/p7
Find a Repository and click to publish!

- Cloud/File storage for large collections (# files or total size),
- Institutional Repositories for moderate collections
Consider SEAD: An NSF DataNet Partner

• Publication with SEAD means:
  – Persistent Data ID (i.e. DOI)
  – Multiple Repository options, including lightweight, standards-based package for long-term storage (BagIT, OAI-ORE, JSON-LD)
  – Registration with DataOne Catalog
  – Branded “Published Data” page to link in your website

http://dx.doi.org/10.5072/FK2FF3PK7W
SEAD Interacts with:

- Projects & their websites
- Authentication services (Google, ORCID, local, ...)
- Researcher Profile Services (ORCID, (VIVO), ...)
- Data Sources (TerraPop, NEON, ‘any’, ...)
- Data Processors (BrownDog, Geoserver, image/video players, ...)
- Repositories (Dspace, Fedora, Cloud, openICPSR, ...)
- Discovery Services (DataOne, DataCite, ...)
- Applications/Services (R, ECube Geosemantics, VIC/DFC, ...)

-- without deep agreement on architectural/model details
-- with mechanisms to help interoperability/synthesis
SEAD as NDS Infrastructure:

SEAD Project Spaces

Sharing, Curation, Publication, Reuse
Web App/REST API
Data/MD

BrownDog Svcs (RabbitMQ)

Earth Cube Geosemantics Services

SEAD Publishing

Matchmaking, Publishing, CI Ecosystem Integration
(Profile sources, papers, catalogs, events, provenance, ...)
REST API

Profiles/Pubs

Tools

Computational Environments

Browser

SEAD as NDS Infrastructure:

Long-term Repositories

NDS/Reference Publisher

DOI Landing Pages Webapp

Publishing Agent App

File System/Cloud Store

Third-Party Repositories

Profile Sources, SSO

Catalogs (DataOne,...)
SEAD as a Data Source:

• Initial Communities:
  – Sustainability Research (Ecological, Social)
  – Centers to grad students and county park managers
  – 2-3 M files, 2 TB+, 20+ groups

• Longer Term:
  – Open – long tail of research projects needing
    • hosted data services
    • core share/curate/publish capabilities for custom CI
  – Observational, Experimental, Modeling, Analytics Data across disciplines
  – Rich publications (e.g. for collaboration, integration, reproducibility)
• Interested in
  – Cloud Deployment (Labs?)
  – Interoperable Data Publications (ORE, BagIT, +) (Labs?)
  – Architecture and Business Model Issues (NDS? Ecube, RDA?)
  – Supporting National Data Publication Needs (NDS, Share?)
  – Integrating to implement the NDS vision across systems (Labs, Share, NDS, RDA, ?)