

An Ecosystem Approach to Data Services and Digital Research Objects

Jim Myers(myersjd@umich.edu)

*Managing Digital Research Objects in an
Expanding Science Ecosystem, November 15, 2017*



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!



Why?

- Researchers manage data on their own during projects and then we ask them to do it again:
 - When they are busy
 - In unfamiliar software, using different terminology
 - For the potential benefit of others
 - who aren't yet ready, and
 - before the software to leverage rich data exists
 - While telling them they aren't doing a good enough job...
 - Without giving them credit
 - Without giving them any guarantee of longevity for their data or assurance that their data will be part of the ecosystem



SEAD:

Sustainable Environment – Actionable Data

- Started in October, 2011 as part of the NSF DataNet program
- An international resource for sustainability science
- **A provider of light-weight Data Services based on novel technical and business approaches:**
 - Adopt an integrated lifecycle view to create value
 - Virtuous cycle
 - providing immediate, incremental value
 - supporting integration and extension
 - Supporting the long-tail of research
 - Scaling/low operating costs



Margaret Hedstrom, PI
Praveen Kumar, co-PI
Jim Myers, co-PI
Beth Plale, co-PI

<http://sead-data.net/>



SEAD Data Services: Start today!

The screenshot shows the SEAD Data Services website interface. At the top, there is a navigation bar with the SEAD logo and links for 'You', 'Explore', 'Create', 'Selections', and 'Help'. A search bar is also present. The main content area displays a grid of project spaces, each with a thumbnail image, a title, a brief description, and a 'Create Project Space' button. The project spaces shown are:

- CanopyDB Preservation**: This space is being used to preserve forest canopy research cyberinfrastructure and data that was created/managed at The Evergreen State College. The Canopy Database applications, all funded by the U.S. National Science Foundation, were developed at Th...
- REACH**: Resilience under Accelerated Change. WSC-REACH. NSF-funded Water Sustainability and Climate (WSC) project to understand landscape vulnerability.
- Lower Mississippi Flood Project**: The Lower Mississippi Flood Project is NSF funded effort to understand the 2011 flooding of the Mississippi River and impacts of the intentional levee breach near Cairo, IL.
- University of Michigan Biological Station (UMBS)**: The University of Michigan Biological Station (UMBS) is an ecological field station where students, faculty, and researchers come together to learn about the natural world, to examine environmental change, and to seek solutions to the critical ...
- neon**: National Ecological Observatory Network. NEON is an NSF-funded continental-scale ecological observation system for examining critical ecological issues.
- Washtenaw County Parks**: Washtenaw County Parks holds over 5,400 acres of land including many high quality natural areas. Proposals to conduct research, perform surveys or monitoring, and other data collection are welcome. Baseline data is available here, and additional data can be provided ...
- National Center for Earth-surface Dynamics**: NCED's mission is to predict the coupled dynamics and co-evolution of landscapes and their ecosystems in order to transform management and restoration of the Earth-surface environment. All data created or compiled by NCED-funded scientists is archived here. Use: Yo...
- Hydroshare**: This space is for use by the Hydrology community. Hydroshare and SEAD teams working towards support for hydrology data and data-intensive research and
- Site-based Data Curation at Yellowstone National Park**: The Site-based Data Curation

Sign-up and request
a secure, branded
Project Space in the
cloud...

Login

Use your existing account on one of the following
networks to log in.



Or login using an email and password.

RESEARCHERS



Manage, Describe, and Publish Data

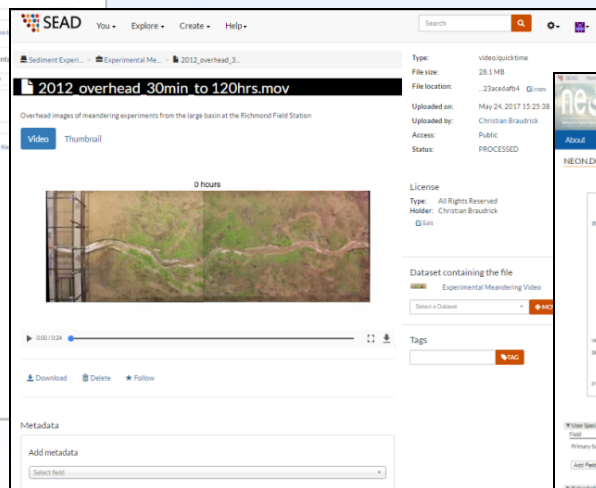
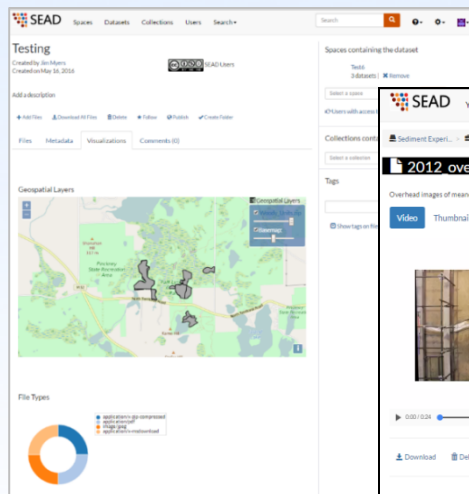
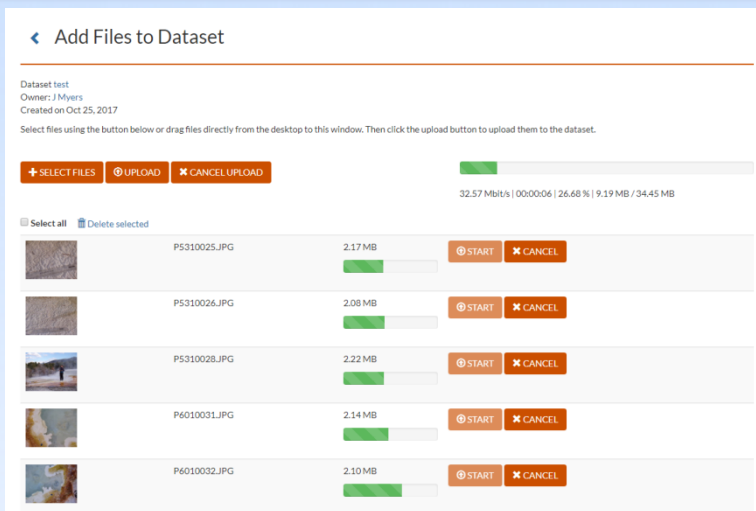
Easily organize, preview, share, publish, or simply make data public with SEAD's team-controlled, collaborative Project Spaces. Incrementally add custom metadata and collaborate with your colleagues during the entire course of your project. SEAD also helps match your data with a long-term repository and provides you with a streamlined workflow for submitting your data for publication.

[CREATE A PROJECT SPACE](#)



SEAD Data Services: Upload and Share

Drag-and-drop your data,
or upload 100K+ files from
disk
Preview, share, analyze...



Extracted Metadata

FNumber: 220/100
DateTimeOriginal: 2017:03:14 18:05:42
FocalLength: 420/100
ImageWidth: 3264
ComponentsConfiguration: 1, 2, 3, 0
ExifOffset: 232
ISOSpeedRatings: 50
Model: SAMSUNG-SGH-I337
MaxApertureValue: 228/100



SEAD Data Services Annotate, Link, Use!

Tags, Formal Vocabularies, Full Text Indexing

Tags

site analysis Remove
core samples Remove

cd Add Cancel

colorado
core samples
county farm park

Jim Myers • Nov 06, 2017 19:29:31

Measurement during the largest precipitation event of 2016-2017...

Reply Edit Delete

Add Metadata

Contact

Add a Person(name, email, or id)

Ch

Anna Ovchinnikova, <http://orcid.org/0000-0003-1475-6741>, anyao@umich.edu
Michael Iannaccone, <http://orcid.org/0000-0003-3035-4865>,
Charitha Dandenya Arachchi, <https://plus.google.com/118034410425276641709>,
Charles Nguyen, <http://orcid.org/0000-0002-4621-1597>,
Isuru Surlarachchi, <http://orcid.org/0000-0003-1711-1711>

Space: SEAD Demo
Uri: <http://sead-data.org>

+ Audience added Aug 16, 2017

+ has derivative added Nov 6, 2017

Created by

Your software can annotate for you with the Restful API

ORCID

FOR RESEARCHERS FOR ORGANIZATIONS ABOUT HELP SIGN IN

Leslie Hsu

ORCID ID: <http://orcid.org/0000-0003-5153-807X>

Keywords: geomorphology, geospatial

Publications: 1440229

Grants: 1440229

Source: Overlaid for ORCID

RCN: Building a Sediment Experimentalist Network (SEN)

Source: Overlaid for ORCID

Works: 14

Rescue of long-tail data from the ocean bottom to the Moon: IEDA Data Rescue Mini-Awards

Source: Crossref Metadata Search

Customize to use any community vocabulary(ies)

Metadata Terms & Definitions

The following metadata terms are defined within this Project Space. To add a new term scroll to the bottom of the page.

Label	Description	Formal URI	Type	Allow New Entries	Actions
CSDMS Variable	The name of a scientific variable in the item, taken from the Community Surface Dynamics Modeling System list of Standard Names.	http://cdms.colorado.edu/wiki/CSN_SearchableList	List	http://legumina330inc.edu/gis/CSN	Edit / Delete
Contact	A person or organization that can be contacted for further information - listed by name, ORCID, or email address.	http://sead-data.net/terms/contact	Person		Edit / Delete
Experiment Location	The location of the experiment as a WKT point: POINT(Lat, Lon).	http://geonames.org/geonames-spec.html	WKT Location		Edit / Delete
Experiment Start	The date/time when data collection started.	http://url.org/terms/date	Date and Time		Edit / Delete
Experimental Method	Text or a URL describing how to perform the experiment.	http://sead-data.net/terms/method	String		Edit / Delete
Funding Institution	The name of the funding ID of the institution supporting the research.	http://sead-data.net/terms/fundinginstitution	String		Edit / Delete
Grant Number	The grant identifier, as specified by the funding agent, that is supporting the research.	http://sead-data.net/terms/grantnumber	String		Edit / Delete
ODM2 Variable Name	A scientific variable in the resource from the CUAHSI Data Model.	http://vocabulary.odm2.org/variablename	List	http://legumina330inc.edu/gis/CSN	Edit / Delete

Advanced Search

Match ALL of the selected terms

Definition Source: Space: SEAD Demo Space Migr. Term: Experimental Method Operator: equals Label (string): mass spectroscopy

Definition Source: Space: SEAD Demo Space Migr. Term: Funding Institution Operator: equals Label (string): NSF

SEARCH

Search Results meandering

Datasets

Name: Experimental Meandering Video

Description: This video is a series of overhead images taken of an experiment with constant discharge and sediment supply. Frames are 30 minutes apart. Flow is from left to right. The white material is a lightweight plastic, the brownish material is sand, and the green colors are alfalfa sprouts.

Collection name(s):

Cited in

WATER RESOURCES RESEARCH, VOL. 34, NO. 4, PAGES 863-867, APRIL 1998

Anisotropic scaling in braided rivers: An integrated theoretical framework and results from application to an experimental river

Efi Foufoula-Georgiou and Victor B. Sapozhnikov
St. Anthony Falls Laboratory, University of Minnesota, Minneapolis

Abstract. Dynamic scaling in braided rivers is reexamined under an extended theoretical framework, developed herein, which explicitly incorporates the self-affinity (scaling anisotropy) in the spatial structure of braided rivers. It is shown that in structures exhibiting anisotropic spatial scaling, dynamic scaling (if present) is necessarily anisotropic. Through analysis of the behavior of an experimental braided river, the presence of anisotropic dynamic scaling in braided rivers was revealed. This implies that there exists a pair of dynamic exponents z_x and z_y , enabling one to rescale space (differently in the direction X of the slope and in the perpendicular direction Y) and time, such that the evolution of a smaller part of a braided river looks statistically identical to that of a larger one. The presence of such a space-time scale invariance provides an integrated framework for describing simultaneously the spatial and temporal structure of braided rivers and may be explored toward statistical prediction of large and rare changes from the statistics of

Has Correction



Uses Calibration



Generated Using



Uses Procedure

<http://sedexp.net/wiki/p7>



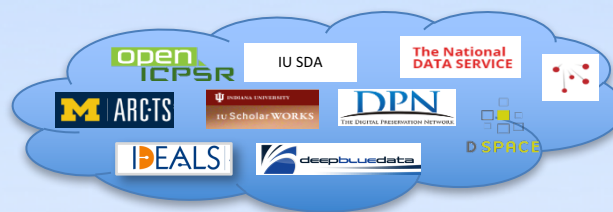
SEAD Data Services Publish and Catalog

• Push the buttons!

- Find a Repository
- Match Their Requirements
- Submit Your Data

• Publication Includes:

- Persistent Data ID (i.e. DOI) with discovery metadata
- Repository-specific storage, or
- Lightweight, standards-based package for long-term storage (BagIT, OAI-ORE, JSON-LD) @ IU, NDS
 - Web DOI landing page
 - Data, metadata, license, fixity info
- Registration with DataOne Catalog
- Branded “Published Data” page in your Space to link in your website



<http://doi.org/10.5967/M0Z0368W>

Matchmaker Details

- ✓ Maximum Collection Depth All Requirements are satisfied.
- ✓ Purpose Match All Requirements are satisfied.
- ✓ Organization Match All Requirements are satisfied.
- ✗ Minimal Metadata **Required metadata is missing:** Edit metadata
- ✓ Maximum Total Size All Requirements are satisfied.
- ✓ Rights Holder IDs Required All Requirements are satisfied.
- ✓ Acceptable Data Types **This info is not required**
- ✓ Maximum Dataset Size All Requirements are satisfied.

Maintaining long-term access to SEAD-published data @ NDS requires maintaining ~2000 lines of code

Rich Data Objects in Forest Research

Judy Cushing, Michelle Wallace, Noah Weiner,
Nalini Nadkarni, Sharon McIntee,
Anne McIntosh, Peter Lynn
SEAD: Jim Myers, Anna Ovchinnikova



Cyberinfrastructure development and 11 projects characterizing the composition, density, surface area, biomass, and spatial distribution of trees, saplings, and understory vegetation.

“Walk up from where you parked at the beginning of the plantation to around the first bend and look for a tree (~ 45 cm dbh) on the north side of the road that has a silver rectangular tag....”

CanopyDB Preservation

This space is being used to preserve forest canopy research cyberinfrastructure and data that was created/managed at The Evergreen State College. The Canopy Database applications, all funded by the U.S. National Science Foundation, were developed at Th...



Custom Databank Database Generator:

DB, Entry forms, dictionary, EML output

Populated Project Databases for 11 sites

Image Gallery:

1300+ field images and visualizations

CanopyView:

Interactive visualization tool - tree structure, canopy coverage, db fields ...

Project Website:

Containing extensive metadata, documentation, software

Data Rescue Project:

Plan and artifacts from the effort to organize and publish this research

Location

Software
Installers

Local Lodging
Nearest Hospital
Locked Gates
Site Characterization

Creators

Manuals

Reports

Archival
Formats

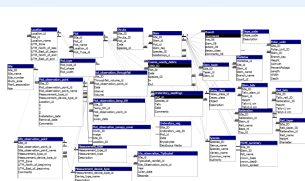
Queryable
DB VMs

Visualizations

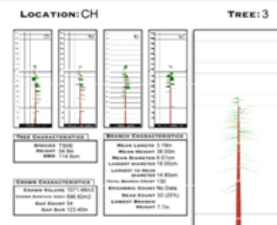
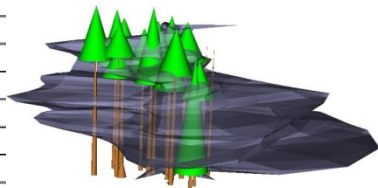
Database
Schema

Methods

Collection
Analysis & QA



Researcher	Database
B. VanPelt	Thousand Year Chronosequence
H. Ishii	Age-Related Development of Crown Structure
R. Dial	Borneo Insect Biomass and Count
B. Lyons	Epiphytes and Hemlocks
E. Menendez	Luquillo Canopy Plot Visualization
D. Shaw	Mistletoe and Hemlocks
T. Sanderson	Monteverde, Epiphyte Changes Over Time
R. Dial	Open Space in Canopy Structure
A. Sumida	Stick structure of Japanese chestnut
Y. Bar-Ness	Tasmanian Eucalyptus obliqua: Crown Structure and Arthropod Biodiversity
G. Parker	Three-Dimensional Canopy Structure



~650 metadata entries (DC, PROV, ODM, custom) describing and linking data files in the collections and reference external resources via DOI, ORCID, and URLs



Examples from related projects

- <http://terraref.org/> (LeBauer) - robotic field sensors and high-throughput phenotype analytics
- SEADTrain (Plale) – Internet-of-Things demonstration of direct publication of IOT data using RDA standards
- Sediment Experimentalists Network (SEN) Knowledge Base – equipment, method, facility information that can be linked with data published through SEAD



SEAD Data Services

Best-effort Operations

- Initial Community: Sustainability Research (Ecological, Social)
 - Large centers to grad students and county park managers
 - 3M+ files, 4 TB+, 40+ groups
 - Rescues & new data
 - 50+ Publications
 - < 1 MB – 0.6 TB
 - 1 – 135K files
 - Cited links to/from high-impact journals
 - Basic metadata to rich provenance and documentation
- Related projects - > 0.6 PB, 100's of publications
- Continuing best-effort:
 - Open to researchers in long tail of research projects needing
 - hosted data services
 - core share/curate/publish capabilities for custom CI
 - Operating through voluntary contributions, related grants, and best-effort support from **National Data Service** members



An Ecosystem Approach

- Infrastructure should provide current value and support/catalyze the creation of new value by third parties (researchers, other infrastructure efforts)
- Some types of infrastructure play the role of keystone species in ecosystems, helping define and stabilize the character of the ecosystem (but not being the most visible or populous species)
- It's time to replace FUD and Catch-22's that limit adoption with base capabilities that can support an ecosystem-wide virtuous cycle!



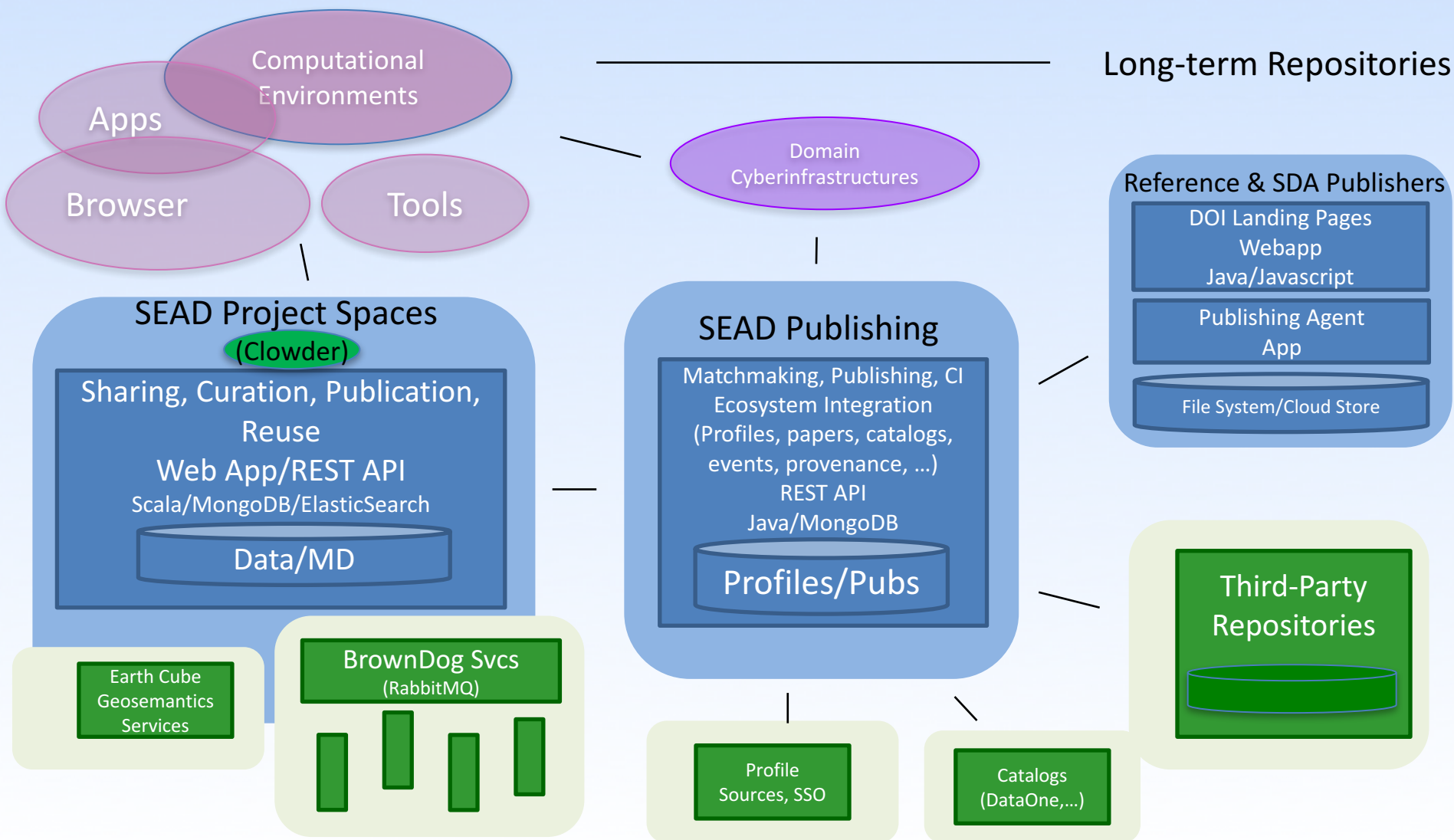
Thank you!

- Acknowledgements:
 - SEAD, NCED, SEN, NDS and other active projects that have provided guidance, feedback, and support
- For more information:
 - <http://sead-data.net/>
 - <https://sead2.ncsa.illinois.edu/>
 - <http://www.nationaldataservice.org/>





SEAD as Infrastructure:





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, ...)
- National Data Service [Universally Accessible Data Publications Pilot](#)
– sign up now!

-- without deep agreement on architectural/model details
-- with mechanisms to help interoperability/synthesis