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

SEAD | Sustainable Environment Actionable Data

The National DATA SERVICE Consortium

Cooperative agreement #OCI0940824

Why do we need "Data Patriotism"?

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

The future needs your data!

Only you can prevent data loss!

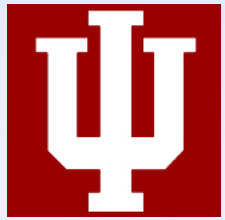
Why?

- 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

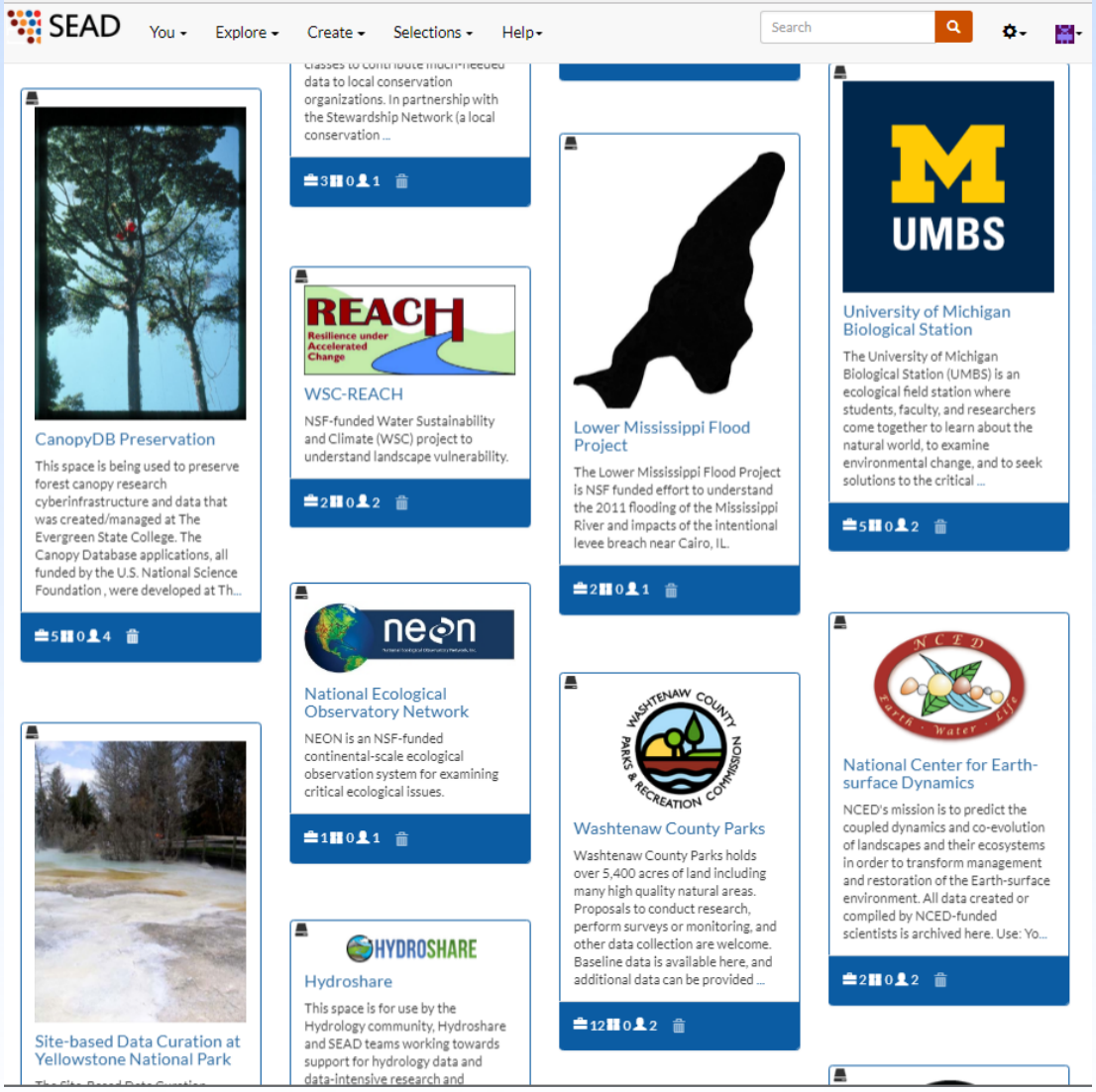
- Started in October, 2011 as part of the NSF DataNet program
- An international resource for sustainability science
- **A provider of light-weight Data Services based on novel technical and business approaches:**
  - Adopt an integrated lifecycle view to create value
  - Virtuous cycle
    - providing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

http://sead-data.net/

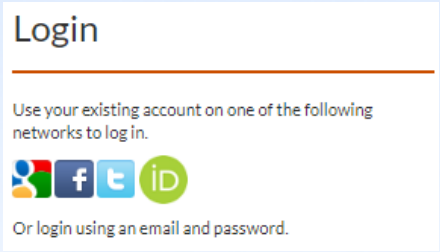


Margaret Hedstrom, PI
Praveen Kumar, co-PI
Jim Myers, co-PI
Beth Plale, co-PI

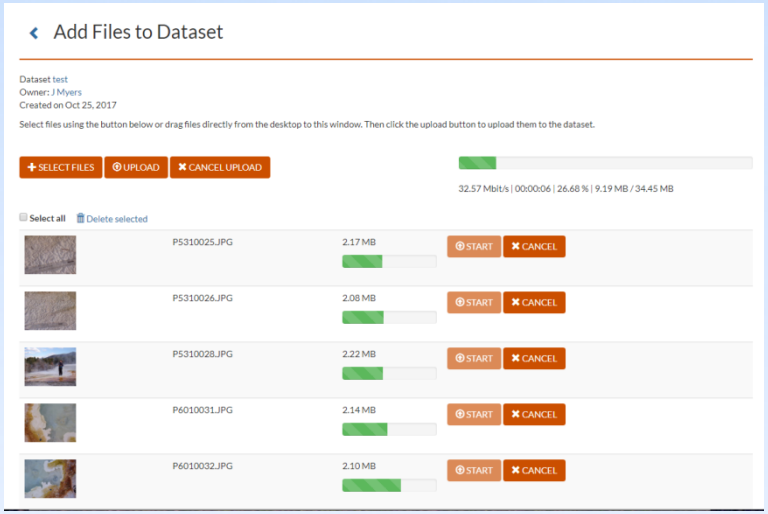
SEAD Data Services: Start today!



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

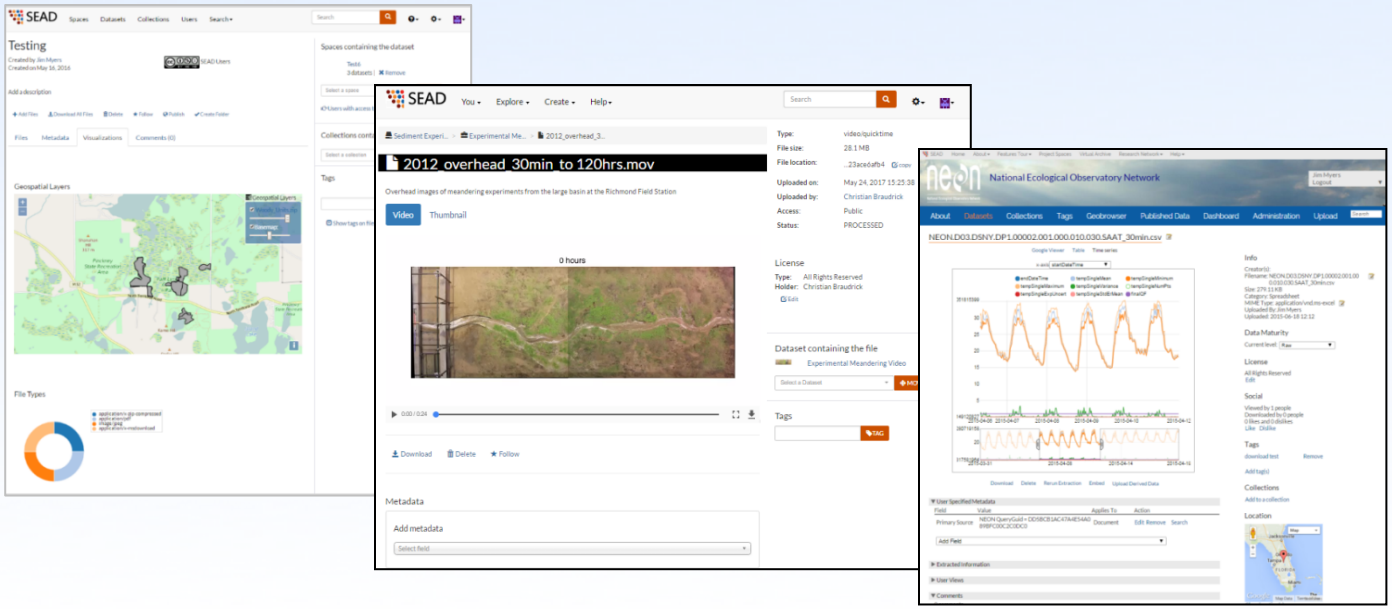


SEAD Data Services: Upload and Share



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

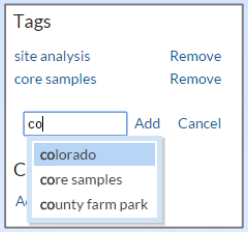
Model: SAMSUNG-SGH-I337

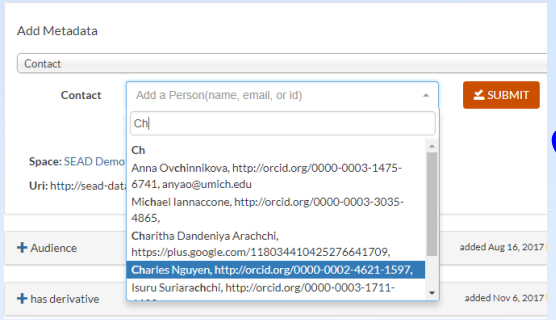
MaxApertureValue: 228/100

SEAD Data Services
Annotate, Link, Use!

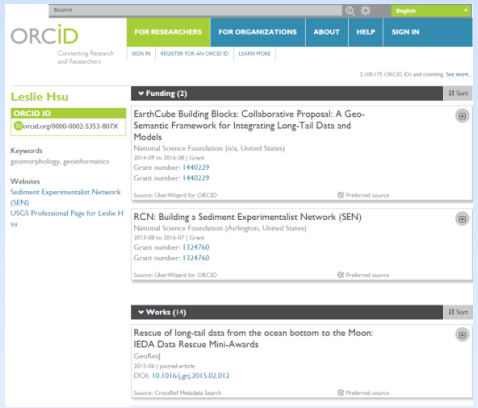
Tags, Formal Vocabularies, Full Text Indexing

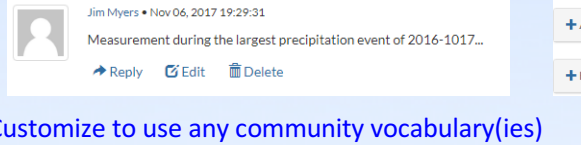
Your software can annotate for you with the Restful API



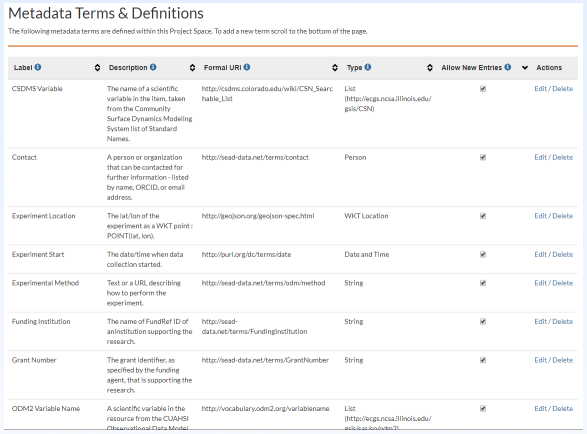


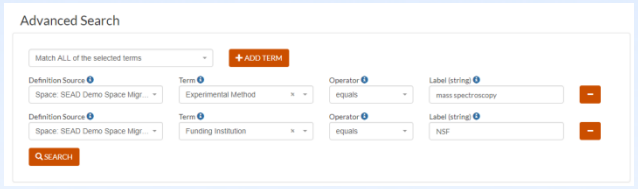
Created by

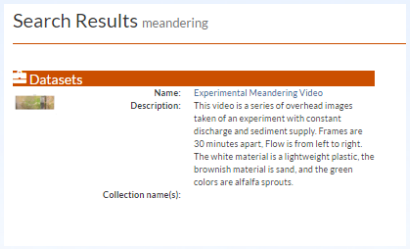




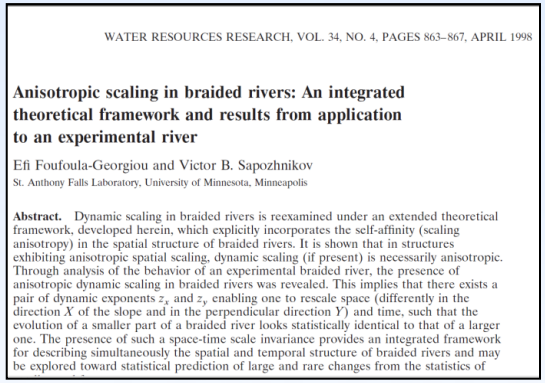
Customize to use any community vocabulary(ies)







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

GitHub

Uses Procedure

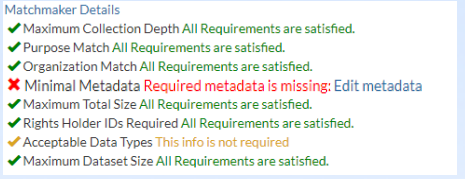
http://sedexp.net/wiki/p7

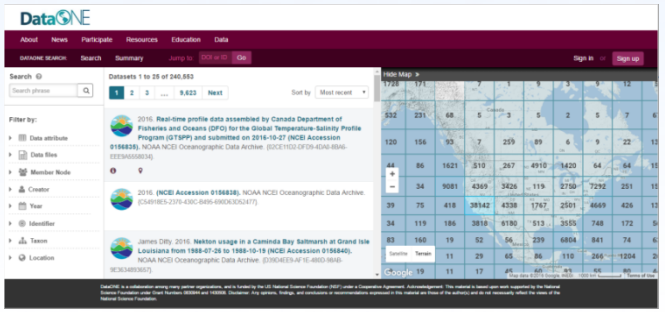
SEAD Data Services
Publish and Catalog

- **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

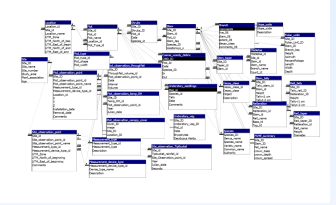




Maintaining long-term access to SEAD-published data @ NDS requires maintaining ~2000 lines of code

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.

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 …

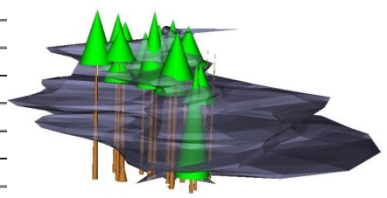
Project Website:
Containing extensive metadata, documentation, software

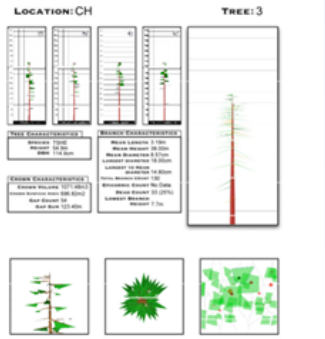
Data Rescue Project:
Plan and artifacts from the effort to organize and publish this research

"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…."

Location
Software Installers
Local Lodging
Nearest Hospital
Locked Gates
Site Characterization
Creators
Manuals
Reports
Archival Formats
Queriable DB VMs
Vizualizations
Database Schema
Methods
Collection Analysis & QA

Researcher	Database
B.VanPelt	Thousand Year Chronosequence
H. Ishii	Age-Related Development of Crown Stucture
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!

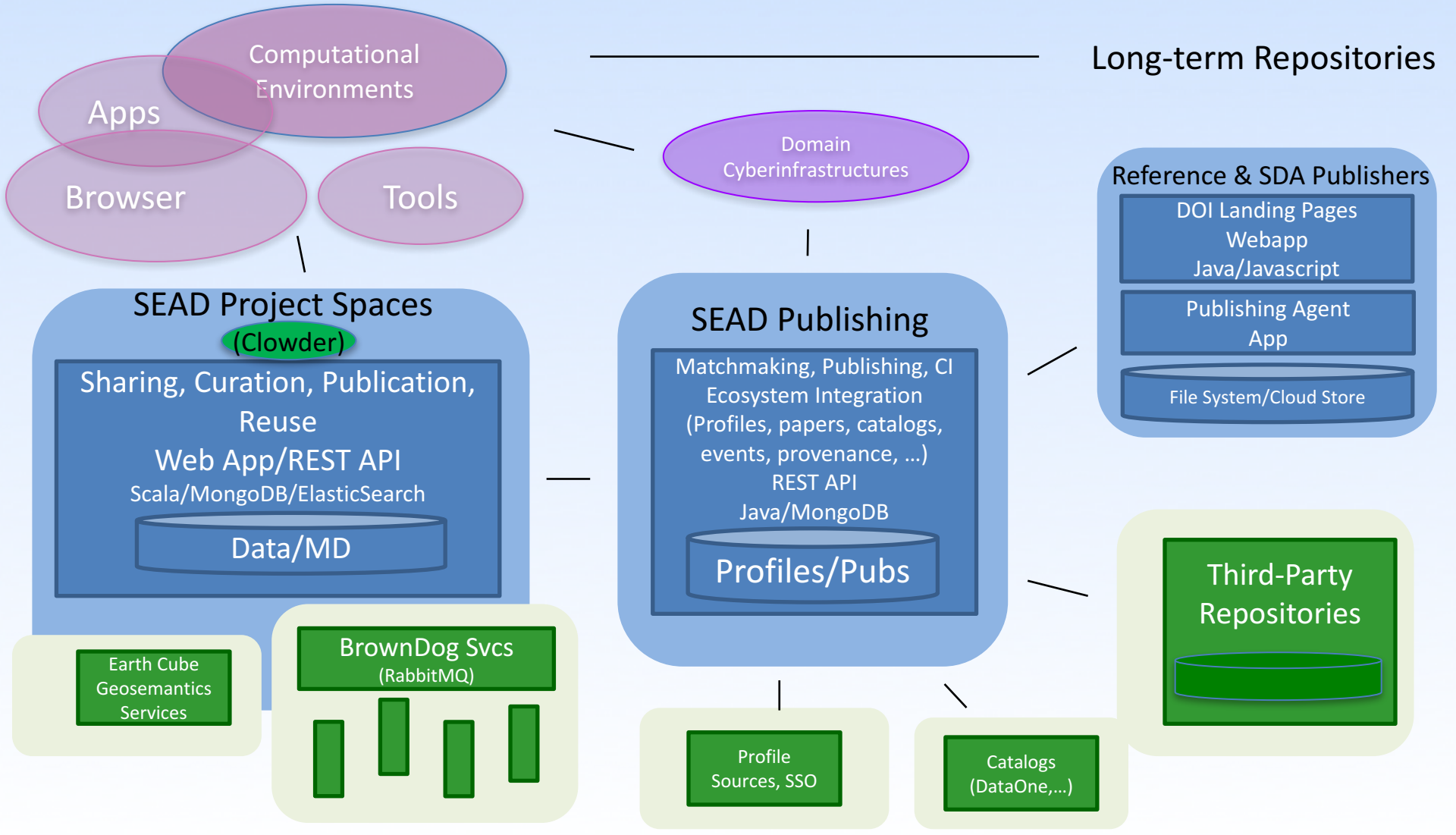
James Myers (2017). Where's My Universal Data Browser? Research Data Management Implementations Workshop, Arlington, VA, Sept. 14-15, 2017

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