Sustainable Environments Actionable Data (SEAD): Three Years On
Beth Plale
Indiana University

Large Scale Data Products Meets RDA
March 8, 2015

Cooperative agreement #OCI0940824
SEAD Leadership

Clockwise from top: Margaret Hedstrom, Beth Plale, Praveen Kumar, Jim Myers, Sandy Payette
SEAD Overview

A Knowledge Network for Collaboration, Data Curation, and Discovery

SEAD enables easy management of sustainability science data and dramatically lowers the effort required to preserve data for long-term use.

TAKE OUR FEATURES TOUR

About SEAD

SEAD is an NSF-sponsored project to create data services designed to meet the needs of sustainability science research. Sustainability science requires reliable cyberinfrastructure and an enhanced ability to manage, integrate, interpret, share, curate, and preserve data across a broad range of physical and social science disciplines.

Latest SEAD News

"Information Managers, Scientists, and Land Managers Tout SEAD's Value"
SEAD Vision and Rationale

SEAD addresses data in the “long tail”

Researchers who produce and analyze heterogeneous data that is unique and at fine resolution and granularity.

Researchers who need to collaborate with others to make scientific advances.

Researchers who lack access to reliable cyberinfrastructure for managing, sharing, analyzing, publishing and archiving data.
SEAD – Researcher’s Perspective

Collection and analysis of data

SEAD

Management and sharing of data

Sharing and publication of data
Manage Data

With SEAD you can annotate and organize data as you're collecting and working with them. In your project space, YOU control who has access and you add the metadata that best supports your project. Check out SEAD's Demo Project Space if you'd like to try things out before getting your own space, or contact us to start a project space.

Link Datasets Together

SEAD supports a variety of interlinking, including via creator name, user generated tags, and collections and subcollections that you create in your project space. SEAD also makes it easy to connect multiple versions of data to one another.

Readily Drop Active Data into a Project Space

SEAD presents an information page for each dataset that displays a preview of the data and map overlays of geospatial data. SEAD also captures basic information about file formats and creation time, extracting metadata from within your files. You can specify which set of metadata fields is appropriate for your project and enter as little or as much metadata as you'd like.
**Publish & Preserve Data**

The [SEAD Virtual Archive](#) simplifies the process for getting data with long-term value into repositories where they are preserved and made available to others.

---

**Conveniently Move Data from Project Space to Repository**

With SEAD it's easy to transition active data into published and archived products. You can start this process from your project space. SEAD uses the metadata you've already entered while the data were active to make them accessible and usable to other researchers via SEAD's repository partners.

SEAD indexes your published data's metadata and registers them with [DataONE](#), so that your data collections are widely discoverable.

---

**Modify, Extend, and Even Republish Data**

After data have been published, you can keep a live copy in your project space, where you can continue working on them. Later, you can publish new versions of the data.
SEAD Technology

• Active Project Spaces
  – Collaboration / file sharing space for research projects
  – Staging area for data curation prior to publishing

• SEAD Virtual Archive
  – Push publishable data (as Research Objects) from an active Project Space to appropriate preservation and discovery environment, and back for reuse

• People, Data, Things
  – Researcher profiles, repository profiles, data profiles
  – all linked, related, dynamic
  – JSON LD
Project Spaces ... organize, describe, visualize dataset collections with metadata

geo dashboard
Virtual Archive ... publish and archive data

<table>
<thead>
<tr>
<th>Research Object Name</th>
<th>Date Last Modified</th>
<th>Current Assignment</th>
<th>Modify Assignment</th>
<th>Edit</th>
<th>Publish</th>
</tr>
</thead>
<tbody>
<tr>
<td>AravindhTest</td>
<td>Thu Oct 02 00:18:11 EDT 2014</td>
<td>Isuru Suriarachchi</td>
<td>Re-assign to me</td>
<td>Edit</td>
<td>Publish</td>
</tr>
<tr>
<td>4SES_notes</td>
<td>Thu Aug 28 13:44:44 EDT 2014</td>
<td>Isuru Suriarachchi</td>
<td>Re-assign to me</td>
<td>Edit</td>
<td>Publish</td>
</tr>
<tr>
<td>BPNM_Publication</td>
<td>Thu Oct 02 15:15:08 EDT 2014</td>
<td>You</td>
<td>Unassign</td>
<td>Edit</td>
<td></td>
</tr>
<tr>
<td>BPNM_Publication</td>
<td>Fri Oct 03 01:43:20 EDT 2014</td>
<td>Isuru Suriarachchi</td>
<td>Re-assign to me</td>
<td>Edit</td>
<td>Publish</td>
</tr>
<tr>
<td>BPNM_Publication</td>
<td>Fri Sep 19 15:16:09 EDT 2014</td>
<td>Isuru Suriarachchi</td>
<td>Re-assign to me</td>
<td>Edit</td>
<td>Publish</td>
</tr>
<tr>
<td>BPNM_Publication</td>
<td>Fri Sep 19 22:35:08 EDT 2014</td>
<td>No one</td>
<td>Assign to me</td>
<td>Edit</td>
<td>Publish</td>
</tr>
<tr>
<td>BPNM_Publication</td>
<td>Wed Oct 01 12:14:12 EDT 2014</td>
<td>No one</td>
<td>Assign to me</td>
<td>Edit</td>
<td>Publish</td>
</tr>
</tbody>
</table>
Data from: Mississippi River Flood 2011 project 2014-10-04 01:09

This collection includes data developed and used for the analysis of the Birds Point-New Madrid (BPNM) Floodway activation in 2011. The data collection includes 10 items, all of which present the processed and derived data. The processed differential LIDAR is the 2005 (pre-flood) LIDAR subtracted from the 2011 (post-flood) LIDAR and corrected for flight line errors. The original LIDAR data were obtained from US Army Corps of Engineers. There are 5 simulated maximum velocity data items from HydroSed2D at two locations (O’Bryan Ridge and Ten Mile Pond) and 2 simulation cases (vegetation and no vegetation). The maximum velocity data for the entire study area was for the vegetated case. The NASA AVIRIS dataset is classified into classes representing vegetation and bare soil. The soil dataset (K/T) is an erodibility index derived from USDA data. Additional data for this study was provided by the USGS, and is available along with this data set at the following site: http://pubs.usgs.gov/pp/1798e/. This data includes ADCP (Acoustic Doppler Current profiler) flow measurements from the inflows and outflows of the BPNM, and HOBO depth sensor measurements from various points within the floodway. This data is used to validate the HydroSed2D simulations.

Recent Additions

Collection_Aggregation_Map
Berretta, D.; Greenberg, Jonathan; Kumar, Praveen; Garcia, M.; Goodwell, Allison; Jacobson, Robert B.; Zhu, Zhenduo; Dutta, Debsunder; Holmes, R.; Parker, Gary; Rhoads, Bruce (2014-10-04)
This collection includes data developed and used for the analysis of the Birds Point-New Madrid (BPNM) Floodway activation in 2011. The data collection includes 10 items, all of which present the processed and derived data.

The 3.6 GB Research Object of data for Mississippi River Flood 2011 deposited to IDEALS repository
Cloud Storage

IR

Project Space

Match-Maker

VA Registry Services

Scientific Workflow

Curation Work Flow

Long Tail of Data Producers

Apps

Tools

Computational Environments

VIVO

ORCID

Social network

Cloud Storage

IDEALS

INDIANA UNIVERSITY

Cloud Storage

Domain Repository

SEAD

ICPSR

DPN

Projects
Matchmaker selects best destination for published data

- Research Object (OAI-ORE)
- Creator profile (LOD)
- Repository profile (LOD)

Matchmaker

- R1 Queue
  - Publish
  - Reject
- R2 Queue
  - Publish
  - Reject
- R3 Queue
  - Publish
  - Reject

R1
- UIUC institutional repository

R2
- Distributed dark archive

R3
- Premier social science repository
### Examples of SEAD’s User Base

<table>
<thead>
<tr>
<th>Project Name</th>
<th>Team Members</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Umich Biological Station</td>
<td>1</td>
<td>8 datasets; 18.76 MB</td>
</tr>
<tr>
<td>Hydroshare</td>
<td>8</td>
<td>439 datasets; 976.08 MB</td>
</tr>
<tr>
<td>Sediment Experimentalist Network</td>
<td>5</td>
<td>16 datasets; 49.59 MB</td>
</tr>
<tr>
<td>Laboratory for Sensory Ecology</td>
<td>11</td>
<td>34 datasets; 7.54 MB</td>
</tr>
<tr>
<td>UMich College Conservation Collaboration</td>
<td>9</td>
<td>18 datasets; 112.64 MB</td>
</tr>
<tr>
<td>Washtenaw County Parks &amp; Recreation Commission</td>
<td>6</td>
<td>58 datasets; 55.62 MB</td>
</tr>
<tr>
<td>Site-based Data Curation at Yellowstone</td>
<td>8</td>
<td>2624 datasets; 1.17 GB</td>
</tr>
<tr>
<td>NCED</td>
<td>25</td>
<td>481,640 datasets; 1.7 TB</td>
</tr>
<tr>
<td>Lower Mississippi Flood</td>
<td>22</td>
<td>67 datasets; 25 GB</td>
</tr>
</tbody>
</table>
• Originally... we saw SEAD as a pipeline, with end-to-end flows
  – Linear: A-active data → B-curation → C-publication → D-archiving
  – But also Circular: A ↔ C

• Now... SEAD is making new progress on:
  – Project Spaces as staging areas for data in computational environments
  – Improvements in matching data to destination repositories (e.g., ICPSR)
  – improvements in data use and reuse, especially of published data in SEAD project spaces

• Future... SEAD continues to evolve to move towards:
  – Technical architecture that is nimble and adaptable
  – Interoperability with broader emerging data infrastructure
  – Serving other communities beyond sustainability science
“Infrastructures are not systems in the sense of fully coherent, deliberately engineered, end-to-end processes. Rather, infrastructures are ecologies or complex adaptive systems; they consist of numerous systems, each with unique origins and goals, which are made to interoperate by means of standards, socket layers, social practices, and individual behaviors that smooth out the connections among them.”

http://deepblue.lib.umich.edu/handle/2027.42/97552
SEAD Long-term Vision

- SEAD as source of technologies and lessons learned for data in the long tail
- Secure SEAD’s role as a vital part of cyberinfrastructure for data intensive science, learning, and action.
- Serve all researchers producing data in the long tail
- Address full lifecycle of data
  - support data from production, to active management, to publication, to archiving and long-term preservation
SEAD and RDA

• As SEAD moves further along towards its goal of a fully realized and standalone recommendation + deposit engine, divorced from the larger OAIS environment that previously constrained it, we will seek likeminded RDA members to harmonize functionality, likely convened as separate WG evolving from Domain Repositories IG.
SEAD and RDA, cont.

• As repositories start to consolidate around protocols for data description and transfer interoperability, SEAD will be an enthusiastic contributor.

• Finally, we see SEAD and its services as a valuable contributor to the Data Fabric IG’s data fabric.