Member Node Description: SEAD VA

Version 1.0 1/13/2014 B. Plale

General

Name of resource:	SEAD Virtual Archive (SEAD VA)
URL(s):	http://www.sead-data.net
Institutional affiliation(s):	Indiana University, Univ of Michigan, Univ of Illinois
Primary geographic location:	USA
Project Director & contact info:	Beth Plale, plale@indiana.edu (SEAD co-PI)
Technical Contact & contact info:	Inna Kouper, inkouper@indiana.edu
Age of resource:	6 mos
Funding support:	Currently, National Science Foundation Cooperative Agreement #OCI0940824
Proposed Unique Identifier:	urn:node:SEAD

Content

Content description/collection policy (1 paragraph, domain and spatial/temporal coverage, uniqueness of content, exclusions, as applicable):

SEAD offers several software tools and services that are designed to lower the barrier to data management for researchers in the long tail of science, where long-tail in this case refers particularly to hydrology, environmental science, social-ecological systems, social science, and earth science. The three primary tools are a user profile service that links people, data and publications; an active curation environment that combines research tools and smart data management (SEAD Active Content Repository, ACR), and a federation service that serves as a single point of ingest/submission and access to long tail data for multiple academic institutional repositories. The latter, called SEAD Virtual Archive (SEAD VA), is the Data One member node. That is, references to "SEAD member node" hereafter refer to SEAD VA.

SEAD VA is part of the larger SEAD project (Sustainable Environments Actionable Data), which is funded by a grant from the National Science Foundation, OCI0940824, and led out of the University of Michigan.

Types of data (complex objects, text, image, video, audio, other):

The SEAD member node serves published research objects; these are complex objects containing metadata, images, videos, shapefiles, text documents (Word, pdf, plain text), and spreadsheets. Initial offerings of SEAD member node are simple data collections from the National Center for Earth-Surface Dynamics (NCED).

Data and metadata availability (rights, licensing, restrictions):

The SEAD member node serves research objects that have been prepared for publication and reside in one of several Institutional Repositories: IUScholarWorks at Indiana University, IDEALS at University of Illinois, Inter-university Consortium for

Political and Social Research (ICPSR), and a default repository, the latter for which SEAD VA is responsible. Rights are as follows:

ICPSR: The open ICPSR repository uses the Creative Commons Attribution 4.0 International license.

IU ScholarWorks: Depositors can choose to license their data any way they wish (or not at all).

UIUC IDEALS: Depositors can choose to license their work however they wish. UIUC does not provide a system integrated with IDEALS for depositors to indicate that their item(s) are CC licensed, however; they would need to indicate that in a rights statement.

SEAD VA default repository: policy under development.

Option for embargo (yes/no, duration):

ICPSR: does not support embargo periods.

IU ScholarWorks: Embargo period is any period of time, up to five years from the date of deposit. For embargoed content, the metadata is open.

UIUC IDEALS: Embargo period can be any length of time that is required. For embargoed content, is the metadata can be either open or closed.

Size of holdings (number and size of datasets, mean and median granules (files) per dataset):

The current holdings of SEAD VA member node are 1.6TB in approximately 450,000 files. Due to an ongoing conversation with NCED, the files must be accessed directly from NCED. This issue of access could also be addressed by SEAD member node moving to a Tier 2 node.

Please describe recent usage statistics, if known, including information on annual data product downloads, annual number of users, annual number of data products used in publications:

Statistics are not currently available.

User interactions

How does a user contribute data? (what can be deposited, how are data prepared, are specific software required, documentation/support available)

The SEAD VA accepts research objects that are submitted to it through an interface that accepts a BagIT package.

SEAD VA has a separate BagIT service that can be adapted to work with any metadata storing service. In SEAD's case, the BagIT service interacts with SEAD ACR to extract relevant metadata that SEAD ACR has about a research object, and does so in the process of constructing the BagIT payload.

The BagIT object is expected to include a metadata file for scientific metadata, currently FGDC is supported. The research object once arrives at SEAD VA is queued for manual data curation. SEAD VA relies on tools in SEAD ACR, to automatically derive the temporal, geolocation, and non-scientific

metadata for the research object and make this available through its SPARQL interface. This automatic metadata collection reduces the manual data curation required at time of ingest.

See source repository for documentation on BagIT interface https://github.com/Data2Insight/sead-virtual-archive

How does a user acquire / access data?

Data are accessed through links provided in the metadata.

What user support services are available (both for depositing and accessing/using data)?

Discovery access is through the SEAD VA interfaces http://seadva.d2i.indiana.edu:8181/sead-access. This includes full text search through a Solr index and faceted search over the scientific and authoritative metadata.

How does the resource curate data at the time of deposit?

SEAD VA has a BagIT client that gets triggered at the time the research object is ready for publishing. The client interacts with SEAD ACR to extract relevant metadata that was accumulated during active use in ACR. This metadata is written into the BagIT payload; other metadata such as residing in files (e.g., FGDC) can be attached to the research object at this time. In SEAD v1.0 release, the publish event is triggered by a scientist within the ACR environment, and the ingest event into SEAD VA is initiated by the data curation specialist working through the SEAD VA interface. Post v1.0, research object ingest into SEAD VA will be available directly through a REST interface, allowing extensibility of the system.

Technical characteristics and policies

Software platform description, incl. data search and access API(s):

SEAD VA extends and is a standalone profile of the Data Conservancy server (www.dataconservancy.org) source code. Search and publish access for scientists and data curation librarians for research object packaging is through a portal. This includes full text search through a Solr index. SEAD VA "talks" to the partner institutional repositories using SWORD.

Service reliability (including recent uptime statistics, frequency of hardware refresh, if known):

SEAD VA service's production instance came on-line Fall 2013 with SEAD release v1.0, so its uptime statistics are not yet available. SEAD VA is a service and not a repository itself, so statistics such as frequency of hardware refresh are dependent on its partner repositories.

Preservation reliability (including replication/backup, integrity checks, format migration, disaster planning):

The preservation reliability of SEAD VA is a function of the preservation policies of the respective member Institutional Repositories.

User authentication technology (incl. level of create/modify/delete access by users):

SEAD VA authentication and authorization is through Google OAuth or through a locally managed portal login. As to access control, in SEAD VA 1.0, objects are publically accessible for download subject to the embargo period. Post SEAD v1.0, research object manipulation upon discovery is richer, allowing object subset and merge for instance.

Data identifier system and data citation policy, if available:

The SEAD member node assigns DOIs to research objects and collections using DataCite. The individual IRs may also assign a handle identifier.

Metadata standards (including provenance):

The research object that arrives at SEAD VA supports Dublin Core + temporal + geolocation + FDGC metadata. SEAD as a project is still working on its provenance representation but draws on substantial experience in provenance by the project PIs (e.g., Karma provenance tool).

Capacity/services to DataONE

At what functional tier will you initially be operating? (see http://bit.ly/MNFactSheet for definitions)

- ☒ **Tier 1: Read only, public content**
- ☐ **Tier 2: Read only with access control**
- ☐ **Tier 3: Read/write using client tools**
- ☐ **Tier 4: Able to operate as a replication target**

If you can host data from other member nodes, what storage capacity is available?

The potential exists to host data in SEAD VA's default repository, which currently resides at IU.

Can you provide computing capacity to the broader network? If so, please describe.

Not at this time.

Other Services

What other services or resources (such as expertise, software development capacity, educational/training resources, or software tools) can be provided of benefit to the broader network?

SEAD is happy to share any of the tools in the SEAD v1.0 suite, including ACR, BagIT client, OAI-ORE research object representation, and SEAD VA.