Towards Sustainable Curation and Preservation:

The SEAD Project's Data Services Approach

James Myers[1], Margaret Hedstrom[1], Dharma Akmon[1], Sandy Payette[1],
[1]Inter-university Consortium for Political and Social Research (ICPSR), [2]School of Information
University of Michigan
Ann Arbor, MI, USA
myersjd@umich.edu

Beth A Plale[3], Inna Kouper[4], Scott McCaulay[4], Robert McDonald[4], Isuru Suriarachchi[4], Aravindh Varadharaju[4]
[3]School of Informatics and Computing, [4]Data To Insight Center
Indiana University
Bloomington, IN, USA

Praveen Kumar[5], Mostafa Elag[5], Jong Lee[6], Rob Kooper[6], Luigi Marini[6]
[5]Civil and Environmental Engineering, [6]National Center for Supercomputing Applications
University of Illinois
Urbana-Champaign, IL, USA

***Abstract*— When the effort to curate and preserve data is made at the end of a project, there is little opportunity to leverage ongoing research work to reduce curation costs or conversely, to leverage curation efforts to improve research productivity. In the Sustainable Environment Actionable Data (SEAD) project, we have envisioned a more active approach to data curation and preservation in which these processes occur in parallel with research and generate sufficient short and long-term return on researcher investments for self-interest to drive their adoption. In this paper, we describe the conceptual framework motivating the SEAD project and the suite of data services we have developed and deployed as an initial implementation of this approach. Use cases in which these services can reduce curation effort and aid ongoing research are highlighted and, based on our experience to date, we identify some key architectural features of our approach as well as open challenges to fully realizing the value of this approach in the broad ecosystem of cyberinfrastructure.**

Keywords*— *data curation, data preservation, research productivity, semantic web, content management, web services

I. Introduction

At first glance, it appears paradoxical that, as many researchers struggle to manage the increasing volume and variety of data they work with, they none-the-less consider data curation and preservation a tax, or as a duty to their communities. This is less surprising when one compares current data curation practice with popular on-line services such as those for sharing files or managing photo collections, or performing remote analyses. Most researchers only engage with curation and preservation services at the end of project lifetimes, after opportunities to leverage them in the research process itself have already passed. Often, submitting data to third party services for preservation and dissemination involves manual re-entry of information and the use of cumbersome forms that cannot be submitted until they are 100% complete.

By comparison, today's online services motivate users to submit data, add tags and comments, and provide metadata by appealing to their short-term self-interest in sharing, organizing, and drawing attention to their work. One might expect that re-organizing data curation and preservation processes to leverage researchers' self-interest could simultaneously add value and lower costs, thereby encouraging more researchers to share, publish, and preserve their data.

It was with this expectation that the Sustainable Environment: Actionable Data (SEAD) project was launched in 2011. Supported through the U.S. National Science Foundation's DataNet program, SEAD provides data curation and preservation services that make it simpler, more valuable, and less costly to preserve data. Further, SEAD is actively exploring ways to leverage curation and preservation activities to accelerate research and enable more data-intensive projects.

In addition to the direct benefits to researchers in lowering costs and increasing the value of data, SEAD also recognized that such a shift could be critical to sustaining services over decades without ongoing project funding, a goal set by NSF within the DataNet program as a whole.

As a project, SEAD involves coupled efforts to develop data management and curation services and to deploy those services for beneficial use to active research groups. While many aspects of SEAD's approach are intended to be generalizable, the project's primary focus is in supporting the 'long-tail' of smaller projects in sustainability science. Sustainability science – broadly definable as the study of coupled natural and human systems – exemplifies many trends in science and engineering as a whole and demonstrates both the value of and challenges in multidisciplinary research. It is also an area in which advances in sensing (distributed and remote), data mining, data science, and computing are making it possible for small groups to tackle complex problems, while driving a need for turn-key solutions for managing the voluminous and highly heterogeneous data such research involves.

In the following sections, we provide further information on SEAD's conceptual approach, and the technical design of its current data services. We also briefly describe SEAD's interactions with sustainability science projects and discuss how the ongoing use of SEAD's services is refining our approach and guiding our iterative deployment plans.

II. Background

While larger projects often have significant budgets for the development of databases and specialized software as well as some level of commitment for ongoing data preservation, smaller projects have little or no capability for custom development and no clear mechanism for maintaining data after the project funding is exhausted. Further, such projects, particularly inter-disciplinary ones, rely on a broad range of third-party data sources and software suppliers. They therefore have limited ability to specify data and metadata standards and instead face the tasks of extracting, translating, and merging information from different sources and handling data in the forms consumed and produced by data analysis tools.

Traditionally, smaller projects have had only a few choices for preserving their data: informal mechanisms such as websites or shared file systems, disciplinary databases that accept specific data types (e.g. gene sequences or LIDAR), or institutional repositories. Of these methods, only the informal ones provide capabilities useful during research such as iterative development of collections, sharing of early results, support for larger data sets, or integration with research processes and collaboration support tools. Conversely, such informal approaches lack features provided in the more formal methods such as assignment of persistent identifiers, versioning, metadata review, file verification, provenance capture, etc.

Technologies and standards exist that can help bridge this gap. Content management frameworks treat data as typed objects (e.g. files or database binary objects). The base layers of the semantic web add the ideas of globally unique identifiers for data and relationships and an overall graph model for metadata. And web services provide a universal means to perform read/write operations on such objects and their metadata across distributed systems. There are a wide range of standard vocabularies that can be used with such technologies to describe various aspects of data [1-4] as well as mechanisms to reify combinations of datasets and metadata as "research objects" [5] that can be serialized [6] and treated as content themselves. These types of technologies have been used widely in e-Science to produce a broad range of powerful and flexible tools including electronic notebooks, lab management systems, workflow systems, and research/data sharing and publication [7-10].

III. Active and Social Curation

Motivated by such examples and the larger concept of linked open data [11], we are exploring, through the SEAD project, how such technologies can be applied coherently throughout the lifecycle of data creation, curation, publication, preservation, and re-use. Specifically, our intent is to go beyond just considering the immediate opportunities to increase efficiency and flexibility over that of current practice and investigate how coupling of activities that are currently conducted by different people, and/or at different times, and/or in independent systems could, through a design based on interacting data services, create value more immediately, add more value, and avoid duplicate efforts.

Consider the apparent paradox raised in the Introduction: researchers face data management challenges yet curation practices that could help, such as annotating data with metadata, are used only after research work is completed (if they are used at all). If data and metadata could be added incrementally as the data are produced, the metadata could be used to help organize and filter data during research. Assuming, for the moment, that researchers found using such a service valuable in their work, it is clear that the incremental effort required to publish data at the end of the project would be greatly reduced and researchers might be much more willing to publish their data. If the system not only preserved the data but also generated citable persistent identifiers for data and dynamically updated the project's web site with new data citations, completing the publication process could be motivated by self-interest as well as altruism. Making such a system preferable to current practices such as using a local or shared file system would also require services that mimic important characteristics of file systems such as the ability to handle any format, provide scalable storage, and support access control. Beyond that, one could envision capabilities such as extracting and displaying metadata from within files, , supporting direct ingest of metadata from instruments, and self-documentation of data as it is created by scientific applications and services that would go beyond what file systems support. Such a system would also benefit curators, who, for example, might monitor data management practices, interact with the project teams to answer questions or provide advice on good data practices, and analyze trends in data volume and in the use of specific data formats and metadata vocabularies to inform their preservation planning and curation workflow decisions. It is this type of analysis that has led to the overall architecture and feature set implemented by SEAD. SEAD has pursued an iterative/agile design approach to development with developers and designers working closely with sustainability researchers and data curators (within the project team, associated with collaborating projects and initial user groups, and in the broader community) to assess current practices and brainstorm about new curation scenarios that could be enabled via designs that support flexible metadata, interactions between services, and richer human interactions across organizations.

Overall these discussions and resulting scenarios have clustered within two general areas that we've termed Active and Social Curation [12]. Active Curation scenarios focus primarily on the activities of data producers and curators working during research projects to produce published data collections. In contrast, Social Curation explores how the actions of the user community (both researchers who discover and use existing data resources as well as other data producers) can be leveraged to provide further value. Under Active Curation, we would consider how SEAD services could be incorporated into formal data management plans, how metadata-based capabilities can support research, how curators

might embed themselves in research groups and provide advice, and how the requirements of long-term repositories could be conveyed to researchers. Social Curation could involve the ability of research groups and projects to publish derived, value-added data products, the capability to notify researchers when revisions or derived products appear, or the ability for curators to monitor the mix of file formats and metadata vocabularies in use to inform recommendations and migration strategies.

IV. SEAD'S DATA SERVICES

SEAD's architecture [13] has been developed within this broad framework of Active and Social Curation. However, development in the first three years of the project has primarily concentrated on creating robust base capabilities and implementing high priority features related to Active Curation that can help drive adoption.

SEAD's initial capabilities are provided by three primary interacting components:

- Project Spaces: secure, self-managed storage with toolsthat allow research groups to assemble, annotate and work with data resources,
- Virtual Archive: a service that manages publication of data collections from Project Spaces to a range of long-term repositories, and
- Researcher Network: a service with personal and organizational profiles that can include literature and data publications.

A. Overview of Functionality:

These components interact to provide the type of end-to-end functionality for producing, publishing, and reusing data that is now being popularized by organizations such as the National Data Service Consortium in the United States [14]. Researchers can drag and drop data into their project spaces and immediately see metadata extracted from their files, The can preview zoomable images, launch playable videos, view automatically generated graphs of time series, see geospatial datasets overlaid on a zoomable map, inspect tabular views of spreadsheets, etc., all within their favorite web browser. Users can also upload data in bulk from local file systems or configure instruments and applications to send data and metadata directly to their project space via web services. Datasets can be organized within collections and sub-collections. Data can be annotated with simple keyword tags or with a configurable set of formal terms from standard and custom vocabularies. Relationships between datasets (e.g. version and provenance derivation relationships) can also be added. Bibliographic metadata relating data to people (as creators or contacts) or publications (papers in which the data is cited) results in live links to the information about the person or publication in the Researcher Network. As the name of a person is typed, a query to the Researcher Network provides type-ahead functionality and, when a matching entry is chosen, the identifier of the person is recorded along with their name, making it possible to create a browse-able web link between the data page and the person's profile in the Researcher Network. Paper references are treated in a similar manner.. Any metadata that is entered can be used to find related data. Specifying the creator of a dataset allows researchers to find all datasets created by the same person. Tags can also be used to find data or to filter data to create custom maps that only show layers and geo-located data associated with the given tag. An RSS feed allows researchers to receive notifications when new data is available.

Publication in SEAD is triggered by marking a collection as "proposed for publication". Curators, working in the Virtual Archive component, can see incoming collections and assign specific curators to work with a given research team. Curators can use the archive interface to retrieve basic metadata from the collection but may also be given direct access to project spaces to explore the overall collection. When satisfied that the collection is ready for publication, a curator can start a workflow that retrieves and packages the data and metadata, assigns the collection a digital object identifier (DOI), identifies one or more repositories willing to accept the data, submits the packaged data to the chosen repository, and registers the new collection with the DataOne project's catalog [15]. Once published, the collection becomes searchable via the Virtual Archive's faceted search and it also becomes visible in a project-space-specific web interface from which the public can discover and download collections as well as individual data sets. Using SEAD's Social Curation capabilities, project spaces can be configured to allow researchers in the broader community to 'like' data and leave comments. Completing the circle, researchers who discover data published through SEAD that is relevant to their own research can request their own project spaces and publish derived data products that retain a provenance connection to the source material.

The Researcher Network provides editable profiles for people and organizations that include information about their publications and data collections published through SEAD. The links created between people, publications, and data remain active after publication and allow researchers to navigate from data to the creator's profiles and to publications and/or related data sets. Within the Researcher Network component, visitors can also browse or view co-authorship graphs and personal or organizational timeline graphs.

B. Architecture

SEAD's components leverage a broad range of open source technologies. Server-side components have primarily been written in Java with user interfaces leveraging HTML5, JavaScript, and the Google Widget Toolkit. Services are deployed at multiple institutions using virtual machines and Apache Tomcat. While SEAD's website and services are run on separate machines at multiple institutions, all components are linked through a common top-level menu structure. When login is required, SEAD supports single sign-on via Oauth2. SEAD also supports anonymous search/browse of published data, Researcher Network profiles, and of any pre-publication data that project groups wish to make available.

1) Project Spaces: Each project space within SEAD is managed by a separate instance of SEAD's Project Space web application. This web application leverages the Tupelo 2

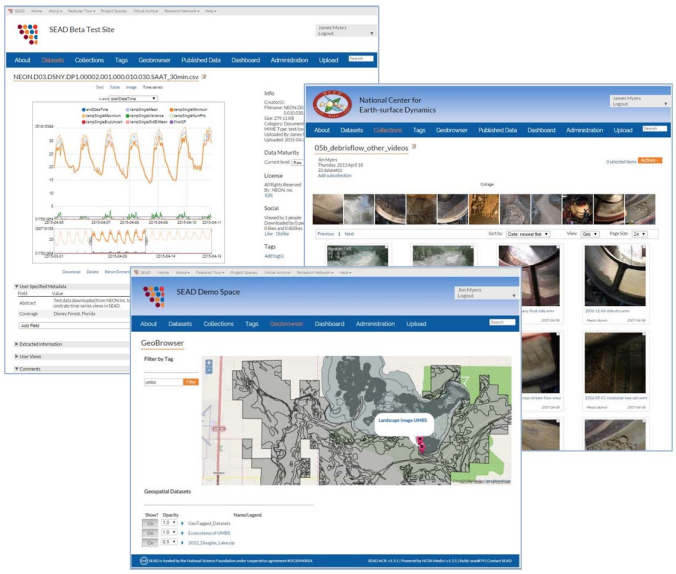

Fig. 1: Individual Dataset, Collection, and Map overlay pages from SEAD Project Spaces.

semantic content middleware [16] developed at NCSA, which provides a blob plus RDF metadata abstraction over an underlying file system and RDF store. To date, SEAD has primarily used a MySQL-based RDF store, although Tupelo 2 provides a mechanism to plug in more scalable stores.

The Project Space web application itself is a SEAD-developed extension to the Java-based Medici semantic content management web application [17] also developed at NCSA. Medici provides core functionality for managing and displaying datasets (a file plus associated metadata) and collections. Access control within a Medici instance is managed via an extensible set of roles defined over a rich set of permissions. Every dataset in Medici is given a globally unique identifier and can be displayed on a data page (Fig. 1) that displays various forms of metadata (tags, key/value pairs, relationships, comments, geolocation, and, for recognized formats, previews and or key/value metadata extracted from within the file). Metadata can be added incrementally and used immediately for browsing, sorting, and filtering data. Through the SEAD project the basic Medici functionality has been extended to include nested collections, a map interface, project-specific branding, faceted search over published collections, a 'dashboard' interface providing a quick overview of recent changes and overall holdings, Oauth2-based authentication using Google or ORCID, a capability to configure the set of metadata terms with which datasets and collections can be annotated, and a broad range of bug fixes and interface improvements. SEAD currently uses a Tomcat 6 web application server fronted by an nginx proxy to implement secure https connections.

SEAD has also added a set of restful web services providing basic create, read, update and delete capabilities for datasets, and collections. Service responses, e.g. listing datasets in a collection or the metadata associated with a dataset, are formatted using JSON-LD. These services have been used to create a stand-alone Java application for batch uploads from file systems and to create a SEAD library that can be used from within the R analysis application to read data and write data with desired provenance and metadata [18]. A SPARQL-query service, and services that report configuration information and usage statistics have been developed and are used internally to generate a dynamic list of project spaces as part of the overall SEAD project website (http:sead-data.net/) that shows live information about the number of views, contributors, datasets and collections, and total collection size for each project space.. Another service, allows researchers to transfer data directly from third-party websites into their SEAD project spaces. SEAD uses an approach called a 'bookmarklet' - a JavaScript managed as a browser bookmark - to generate a SEAD import panel within third-party web pages (Fig. 2) and to temporarily repurpose download links on those pages to trigger import of the data to a SEAD project space rather than download to the desktop. Small utilities that, for example, verify space contents or support bulk editing of metadata also exist.

The generation of previews and extraction of metadata within Project Spaces is handled through an extensible set of plugins. As with other content management systems, plugins are registered to respond to specific sets of mime-typed content. In SEAD's current implementation, extractors are created as plug-ins within a single Eclipse-RCP [19] application that implements an extraction web service. Extractors run in multiple stages that can complete asynchronously and generate extracted metadata and/or data required for a preview. Preview display is handled by type-specific classes added to the main web application. As an example, SEAD provides a zoomable image preview for common image types that relies on an image extractor that generates a pyramid of tiles and a small preview display handler that, rather than handling display directly, simply invokes Microsoft's Seadragon [20] viewer, configured to point at the supplied tiles.

The extractor for geospatial information, as part of its processing, sends the data to a Geoserver [21] instance, which makes the data available as Open Geospatial Consortium

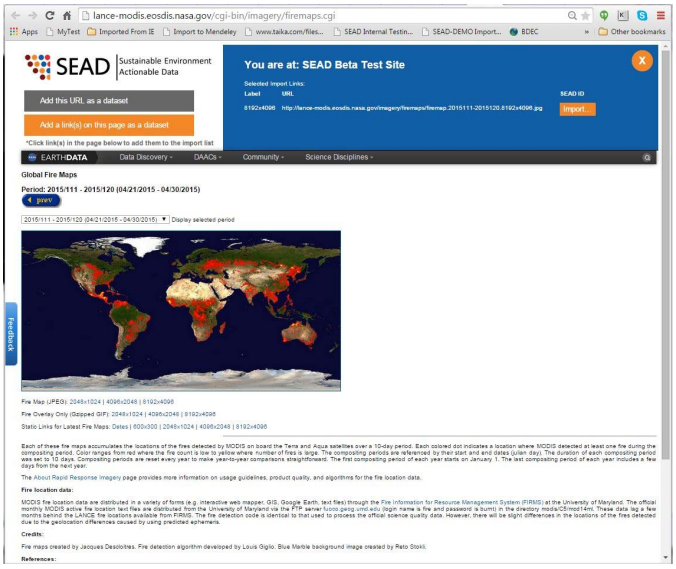

Fig. 2: SEAD bookmarklet being used to import MODIS fire map data with provenance directly from original website.

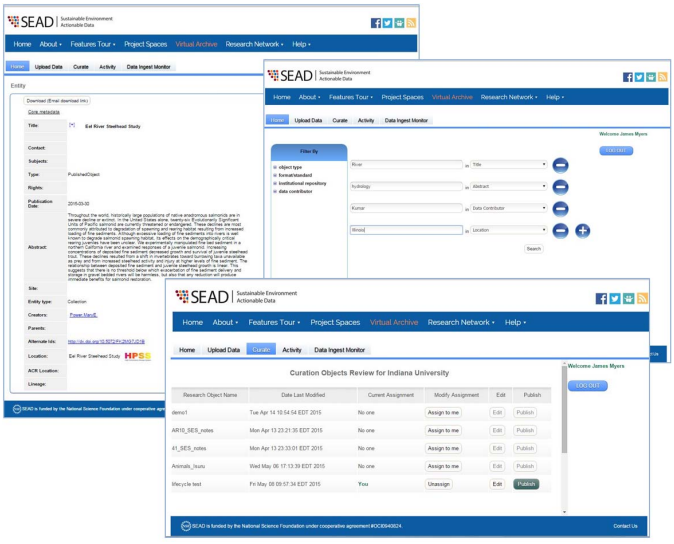

Fig. 3 Virtual Archive interfaces for viewing published data, finding data via faceted search, and publish new data collections.

(OGC) map layers. Access to these layers is proxied through the main Project Spaces application to enforce access control. Zoomable map displays in the Project Spaces application access the layers in the Geoserver and the URLs for the layers are exposed as metadata on the dataset page so it can be used with third-party applications.

2) Virtual Archive: The SEAD Virtual Archive (SEAD VA) [22] is a federation layer over multiple institutional repositories that manages an overall packaging and publication workflow and provides a global search capability across data published through SEAD (Fig. 3). It leverages archival software developed by the Data Conservancy Project [23] and the Komadu provenance service [24]. Overall, the VA is architected as a set of interacting web services accessed via a common web interface.

Processing within the VA begins when the fact that a new collection has been marked as "proposed for publication" in a project space is detected via a restful service call. The VA retrieves basic metadata about the collection and shows it in a list where it's processing can be assigned to a given curator. The basic process of retrieving data and metadata is handled via a modified version of the Data Conservancy code that makes restful service calls to the relevant Project Space and generates an Object Reuse and Exchange (ORE) formatted package as output. SEAD has added a number of extensions to this basic workflow. The processing of a collection can be assigned to a specific curator and the assignee can make modifications and additions to the collection within the VA interface (or login to the project space and review and update the collection there). Since SEAD can submit data to multiple types of repositories, the VA includes a Matchmaker service that compares the collection (using its metadata) with an extensible set of rules defined by repositories that specify what each repository will accept. Current rules address aspects such as the institutional affiliations of the data creators and the overall size of the collection. The curator can select a target repository from the list of those that will accept the package or update the collection and repeat the Matchmaking step. Once a target repository is selected, the VA invokes the submission process. Currently, the VA uses SWORD to deposit to DSpace (e.g. Illinois' IDEALS repository and Indiana's ScholarWorks). It also has connectors to the Digital Preservation Network (DPN) and to cloud and file-based storage. The latter, which can aggregate the datasets in a collection into tar files, is currently configured for use with Indiana University's Scholarly Data Archive (SDA) storage system as a default for larger collections. Given SEAD's capability to update and republish collections, and to publish collections that are derived from previously published collections, the VA also includes a registry service that captures the connections between versions and derived collections along with the detailed provenance of a collection as it is processed. The VA uses DataCite to generate DOIs for published collections. It also includes a service that implements the DataOne Member Node protocol [25] which enables SEAD to register newly published collections with DataOne's federated catalog. Collections are indexed using an Apache Solr to support full text and faceted search of the metadata. The VA's data discovery interface also includes a graphical display of the provenance relationships between published collections.

3) Researcher Network: SEAD's Researcher Network manages information about people, publications, data, and organizations. It is implemented using the open source VIVO application [26]. VIVO is a semantic web application that uses the Resource Description Framework (RDF) to represent information relevant for researchers' profiles. VIVO has primarily been deployed at an institutional level, leveraging bulk ingest of researcher names and affiliations, interests, activities, and accomplishments from existing institutional databases. SEAD's instance allows any SEAD user to create a profile, and it has been pre-populated with ~1,800 profiles for people in the co-authorship networks of one of SEAD's pilot user groups.

Unlike SEAD's Project Spaces, VIVO uses a formal ontology to define the set of objects and metadata that can be entered and displayed. To facilitate rich connections between data and people as well as powerful analytics, the SEAD VIVO has incorporated an extension to VIVO ontology developed by the Australian National Data Service (ANDS). This extension, VIVO-ANDS ontology [27], focuses on datasets and enables data to be treated as a VIVO resource such as a person or a publication. SEAD's VIVO implementation has also been extended to accept citation information, generated by the VA using the DataCite DOI API, for collections published through SEAD. SEAD also runs a Fuseki SPARQL server to expose VIVO content to other components. This is used, for example, to support type-in capabilities within Project Spaces: when a researcher starts typing a person's name to add that person as one of a dataset's creators, the list of people with matching VIVO profiles is shown. The result is that, rather than just recording a name string, Project Spaces record the persistent identifier for the person in VIVO and creates a live link from the data page to that person's profile. An analogous mechanism is used to associate data with literature publications in VIVO.

In addition to basic browse and search capabilities, VIVO can generate several types of interactive graphical displays including co-authorship network graphs and timeline views

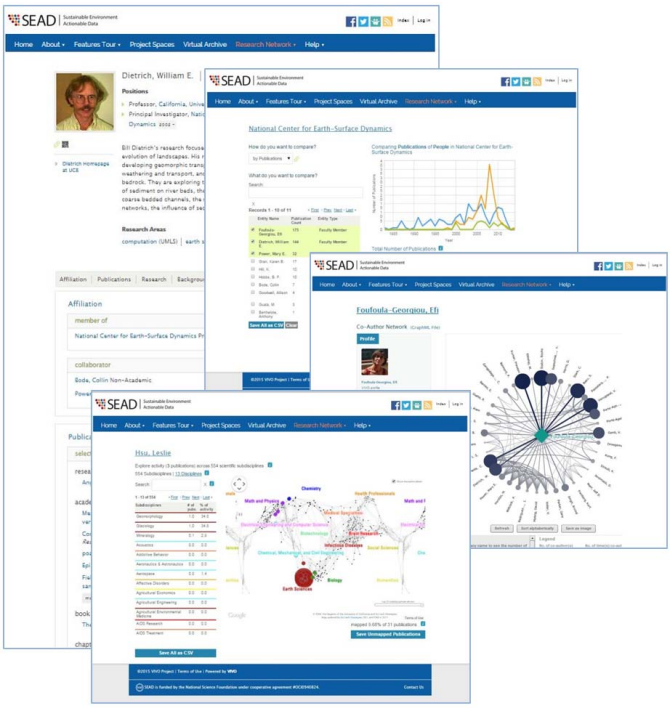

Fig. 4 The Researcher Network manages information about people, organizations, and publications, providing profile views as well as temporal, co-author, and scientific categorization graphs.

(e.g. the number of publications over time) (Fig. 4). These graphs include data publications. In combination with the live links from Project Spaces and VA to researcher's profiles, these graphs allow navigation, across SEAD's components, from DOIs to data to people to co-authors to papers to further interesting datasets.

C. *Deployment:*

The SEAD Project is based at the Inter-University Consortium for Political and Social Research (ICPSR), part of the University of Michigan and the main SEAD website is hosted at ICPSR. As shown on the "Projects" link at http://sead-data.net/, SEAD is operating more than 20 Project Spaces along with test and operational versions of the Virtual Archive and Researcher Network services. Project spaces are deployed at the National Center for Supercomputing Applications on a cluster (2 servers with dual quad-core Xeon processors, 64GB memory, and 20TB of available storage (RAID-6)) as individual virtual machines with memory and storage configured to match the needs of specific groups. The servers are monitored and maintained, and the operational cybersecurity has undergone an initial review by the NSF-supported Center for Trustworthy Scientific Cyberinfrastructure (CTSC). As of spring 2015, SEAD has also completed an internal review of Project Spaces software for the top ten vulnerabilities identified by the Open Web Application Security Project (http://owasptop10.googlecode.com/files/OWASP Top 10 - 2013.pdf) and switched to https-only access. SEAD's VA and Researcher Network services are deployed at Indiana University on servers managed by the Pervasive Technologies Institute.

All software developed by the project is available under open source licenses. Team development is managed using standard software tools including Atlassian's software development suite (source code control, bug tracking, wiki, and automated build tools) and GitHub. For developers, testers, and researchers wishing to host their own spaces, SEAD also makes the Project Spaces software stack available as a downloadable virtual machine.

V. DISCUSSION

A. *Evolving SEAD's Services to Support Sustainability Research*

Within 18 months of starting the project, the SEAD team demonstrated the core capability for researchers to import and annotate data, even for collections of data at the scale of hundreds of thousands of files and over a terabyte in total size, and for curators to review, package, validate, and route data collections to repositories for long-term storage and future use. Over the following two years, SEAD has worked to make its prototype capabilities robust, to connect with a broad and growing range of sustainability research teams interested in managing their data. SEAD has actively solicited comments and suggestions from such groups and from representatives of repository organizations and has incorporated their feedback to improve and extend SEAD's services. We also work to assess our progress in realizing the benefits of active and social curation, and, most recently, to initiate design changes that will be needed to continue to scale SEAD's services and enrich functionality while adapting to the numerous changes in the larger data services and cyberinfrastructure arenas.

The first project to publish datasets through SEAD, the National Center for Earth-surface Dynamics (NCED) is also the largest contributor of published data. They have published 20 rich collections that include over 450,000 files and approximately 1.6 TB of data in total. These data were collected over 10 years and include more than 83 file formats as identified by DROID [28] (counting versions of formats as distinct, but treating flexible formats such as XML or comma separated value (CSV) as single formats, and ignoring roughly 10,000 files with no extension/of unknown type). The collections include raw experimental data (including images, video, sonar tabular data, etc.), simulations, software, documents and presentations, and third-party reference data.

While the largest spaces have hundreds of thousands of files and hundreds to thousands of gigabytes of data, most spaces that have moved beyond simple testing include tens to thousands of files. Groups using SEAD vary in size and in purpose. Some represent individual projects and a single research aim, e.g. the analysis of the erosion that occurred during 2011 flooding on the Mississippi River due to the intentional destruction of a levee by the US Army Corp. of Engineers. Others involve larger groups with shared interest in a place, e.g. Yellowstone National Park, or topic, e.g. sediment transport. These groups seek to share primary research data, aggregations of third-party data assembled and organized for their specific purpose, data and analyses that support results in peer-reviewed publications, software project documentation, and education and outreach materials.

SEAD has solicited feedback from users (researchers and curators) through a wide range of formal and informal techniques including presentations and demonstrations at conferences such as American Geophysical Union, Ecological Society of America, and International Digital Curation Conference meetings, SEAD-specific workshops, and one-on-one interactions. From these interactions, the SEAD team has identified an ongoing set of development activities and received guidance that is influencing longer-term planning. Users and potential users have helped identify simple bugs and usability issues as well as areas where SEAD's initial conceptual model did not fit current practice and where relatively simple additions could reduce work for researchers or solve additional problems for them.

As an example in these latter areas, it became clear relatively early that, for larger collections, the relatively flat model of listing all datasets, or just all collections of datasets (providing one level of hierarchy), which SEAD shared with many existing systems, simply does not scale to thousands of datasets and beyond. While displays that show pages of datasets look useful with the tens to hundreds of entries typical in prototyping and early use, a more hierarchical model and hierarchical browsing capabilities, and more emphasis on sorting and filtering mechanisms become increasingly important as collections grow. In response, SEAD implemented the capability to have deeply hierarchical collections, added new tree-style views, added a configuration options to only show 'top-level' items in dataset and collection lists (i.e. to show only the datasets or collections that are not in a parent collection rather than all datasets or all collections), and enhanced the display of search and tag-filtered results to enable the same functionality (e.g. the ability to select multiple entries and add additional metadata or inter-dataset relationships) as when viewing all datasets.

Another early recognition was that, when researchers think about publishing data incrementally while a project continues, maintaining their branding on the data, and making it simple for researchers to leverage data to, for example, drive traffic to their project website, are critically important. This led SEAD to emphasize the ability to link data to publications, authors, and institutions and to add new capabilities that allow research teams to brand their project spaces, generate lists of the project's published data collections that could be integrated into their website, and produce custom reports on research outputs and impact. We also had to confront the array of concerns and requirements surrounding access to published and/or pre-publication data that arise from a combination of characteristics of the data, regulatory and, institutional environments, discipline-based norms and standards, and researcher- and project-based sensibilities about open sharing of data. SEAD allows researchers to and to control access to their project space through provide a range of options, including for project teams to enable anonymous access (e.g. no login required), or onymous access (login required with no additional restrictions), role-based but otherwise open to anyone) access with restrictions on uploading, viewing, editing, and downloading data, and (including hybrid options such as anonymous browsing/preview with onymous data download).

As SEAD added capabilities to display maps showing geospatial data (both geo-located items as well as datasets such as shapefiles and Geotiff images that represent geospatial map layers), researchers noted the potential for new ways of working. For example, being able to filter geospatial data based on an associated tag, which provides the ability to dynamically create a custom map, could allow graduate students to tag their latest results and send their advisor a URL to the custom map. Similarly, researchers noted that simply exposing the Open Geospatial Consortium (OGC) URLs for data layers, which SEAD was already creating to support its map interfaces, would make it easier for researchers to pull the data back into desktop geospatial analysis software. SEAD added these URLs to the metadata being generated by the geospatial extractor during upload to support such use.

At a more abstract level, SEAD's capabilities have also provoked interesting discussions about active and social curation and how technologies such as SEAD's will affect the socio-technical landscape of data management, sharing, curation, and preservation services. Simply having a scalable mechanism for assembling and publishing large collections raises questions of whether researchers should be thinking about publishing only data directly associated with publications, which is arguably the current norm, or more complete project records that could include raw and intermediate data as well as final results, and software, instrument calibration data, model validation test, etc.. Such publications are potentially much larger but would significantly enhance reproducibility and support broader re-analysis.

SEAD has been used to publish both types of collections already. Faced with collections as large as a hundred thousand files, SEAD has had to define practical mechanisms to curate and preserve them, as well as consider how cases of very large data collections could be handled better going forward. Currently, SEAD uses automated tools to check the completeness and fixity (via cryptographic hash verification) of ingested data, performs manual inspection and enhancement of basic bibliographic and geospatial metadata for the collection, and adds metadata to identify descriptive files (e.g. images representative of a collection, 'readme' files and software manuals, spreadsheets mapping individual data files to experimental parameters used in generating them) that are discovered through text search and inspection (e.g. when a sub-collection contains a few files that are not of the same type as the bulk of the datasets in the sub-collection). The project is currently looking into metrics and tools that would help curators quickly assess and identify issues, such as a few files out of thousands missing desired minimal metadata, within collections that are too large for manual inspection of each file..

SEAD has adopted several practices to handle larger collections after publication. It exposes such collections as single entities to DataOne's federated catalog (e.g. rather than as 450,000 file-level entries): the DOI for the collection references the repository preserving the data where the substructure of the collection can be explored. SEAD's default data store has also recently been switched from storing individual files to using tar to produce one aggregate file for efficiency. Going forward, SEAD is planning to create

guidelines for the granularity of data and the metadata necessary for discovery, efficient use of storage, and timely retrieval. These guidelines could be repository specific, conveyed to researchers through SEAD's Matchmaker service, and managed by curators at each of SEAD's partner repositories.

B. SEAD's Services as a Platform for Exploration

There are many other areas where the flexibility of SEAD's architecture combined with its goal to deploy operational services in support of sustainability research brings a useful urgency to finding practical solutions to complex issues. For example, while SEAD project spaces support use of formal metadata (e.g. RDF statements using external vocabularies, identifiers for people) it also allows use of tags and strings for the names of creators and contacts. Researchers find these features very convenient yet curators know that disambiguating such terms without further context is very challenging. However, SEAD's project spaces do provide context and its model allowing incremental addition of metadata and active curation makes it possible for researchers to annotate data using free form text tags and for researchers or curators to recognize when this is occurring. Researchers or curators could then use bulk operations to add valid formal terms corresponding to the meaning of the tag as used within the specific project space. Such a mechanism is not fully automated in SEAD today, but it does provide an alternative to forcing researchers to only use formal terms or expecting curators to add valid formal descriptors after publication. Going forward, we anticipate supporting this type of transformation through configurable rules and capturing the provenance of such inferred metadata in data publications.

SEAD is also trying to leverage its experience in the context of national and international 'data ecosystem' efforts such as the DataNet program, the US National Data Service Consortium and the Research Data Alliance, as well as the broad range of commercial and open source services that are emerging to manage people, publications, and data. When the SEAD project started, we selected technologies that were semantic and web-service 'aware' but otherwise independent (in terms of data model and vocabularies used) as a way to develop a complete system while architecting for a future in which an ecosystem of evolving data services would exist. In doing so, we have sought to identify the minimal level of agreement required between components to provide added value to researchers. For example, in developing and demonstrating interoperability between DataNet projects, it has become clear that there is value for researchers simply in being able to import data and metadata from external systems with a provenance link back to the source, regardless of the existence of further agreement about data format or minimal metadata standards. While such additional standardization could add more value, there is already value in being able to present metadata to users, in any vocabulary (i.e. that of the source), and in providing relatively simple mechanisms, such as the ability to perform bulk operations to map terms to preferred vocabularies (as discussed in the context of tags above) or to unzip packages from external sources when their storage granularity differs from that of a local data service. Such mechanisms allow standardization at a per-project level, which could then inform and prioritize broader community standardization efforts - as it becomes clear which data sources are being used in combination and which terms researchers are trying to map. SEAD has participated in an NDS Hackathon to develop a service providing a uniform means of retrieving data and metadata, independent of identifier scheme, data format, and metadata vocabulary that demonstrates such an approach [29].

One area where SEAD has not yet defined a full model is in the continuous flow of data from project spaces, where it is 'live' and it can be versioned, moved, and have its metadata change, to 'published', where it is fixed, to a new research environment where it may be disaggregated, combined with other data, edited, and reanalyzed, producing a new data product for publication. The expectation that published objects may still acquire further comments and provenance, or may be modified over time for preservation purposes (e.g. format migrations and vocabulary updates) further complicates this picture. For its current version (1.5), SEAD has defined a three stage model going from 'live', to 'in-curation' to 'published' [30]. Most re-use issues are handled by assuming that data brought into a new research environment or a project space is always a new object with a provenance link to the original published object, but currently with no capability to use that link to identify changes. The general issue has been recognized in other systems [31] and structured ways of addressing it have been proposed [32-33], but the challenge of implementing usable and useful capabilities that leverage this structure in a system such as SEAD that also supports access controls and annotation in different project and public contexts appears to be open.

This type of issue, along with general drivers to improve scalability, lower operating costs, and increase interoperability with external systems that manage data, people, or publications, or that provide independent services for data sharing (e.g. DropBox), format conversion (e.g. Brown Dog [34]), cataloging (e.g. DataOne), or preservation (e.g. Fedora), are at the heart of SEAD's current efforts to shift towards cloud databases and to standard messaging and preservation interfaces. In the next phase of the project, we also expect a shift from an emphasis on active curation, which relates primarily to the creation of new data resources, to a balance with social curation, which addresses data re-use and community-level contexts. We expect that techniques similar to those we have applied, which that leverage rich metadata and knowledge of context to enhance data publication, will also allow the creation of practical, '80%' solutions that better capture the value of community members' efforts in assessing data quality and creating derived data products.

VI. CONCLUSIONS

In four years, SEAD has developed an operational capability for sustainability science projects to manage, curate, and publish their data through hosted project spaces (available upon request) that represents a new option for such projects that is more powerful than simply using a shared file system yet less costly than a custom project solution. This has enabled the project to attract a broad range of collaborators and early

adopters, representing both research groups and repositories, who are interested in using SEAD's services as well as in continuing to improve usability and leveraging SEAD as a platform to explore new approaches to curating data.

At a technical level, SEAD's emphasis on interacting services exposing globally identified semantic content has proven valuable in providing rich and highly customizable capabilities, to researchers while minimizing coupling between services and providing scalable capabilities. Further, while there are many opportunities going forward, SEAD's data services are already demonstrating, with concrete functionality, how the potential of active and social curation approaches can be realized. We anticipate that the approaches used in SEAD will be broadly applicable in the growing ecosystem of linked open data sources and we see SEAD as an example of the type of schema-agnostic tools that can help realize the potential of the semantic web [35].

In terms of addressing the challenge of creating data services that are themselves sustainable, in both a technical and business sense, SEAD continues to assess how its operational costs can be further reduced and to explore, along with the larger community, the potential of different models of support. To date, SEAD has leveraged its internal resources and in-kind contributions from collaborating institutions to subsidize the costs of operating its services for the sustainability community, with development and implementation of a full long-term model for support still a future deliverable.

Solving the economic issues related to data preservation at a national and international scale will clearly require the participation by and coordination of many different players. Many conversations are taking place within organizations such as RDA and we expect that SEAD's strategy will be heavily influenced by choices made in the overall scientific community. As this discussion moves forward, we hope that our experiences in SEAD will help open the debate to include not just how to support curation and preservation as currently practiced, but how active and social curation practices, and value-added services over federated data infrastructure, which will increase value and shift and reduce costs, and broaden the options for sustaining data services.

ACKNOWLEDGMENT

This research was funded by the National Science Foundation under cooperative agreement #OCI0940824. The authors acknowledge the efforts of the broader SEAD project team (http://sead-data.net/about/sead-team/), past and present, in developing, deploying, and supporting SEAD's data services. The authors also gratefully acknowledge the efforts of SEAD's collaborators, including the National Center for Earth Surface Dynamics (NCED), repository partners including the libraries at Indiana University and the University of Illinois and the Inter-university Consortium for Political and Social Research at the University of Michigan, and SEAD's early adopters who have provided invaluable guidance and feedback.

REFERENCES

[1] Data Catalog Vocabulary (http://www.w3.org/TR/vocab-dcat/)

[2] Dublin Core Metadata Schema (http://dublincore.org/documents/dcmi-terms/)

[3] W3C recommended provenance modeling ontology (http://www.w3.org/TR/prov-o/)

[4] VIVO research network ontology (https://wiki.duraspace.org/display/VIVO/VIVO+Ontology)

[5] Bechhofer, Sean, De Roure, David, Gamble, Matthew, Goble, Carole, and Buchan, Iain. "Research Objects: Towards Exchange and Reuse of Digital Knowledge", Nature Precedings, 2010, http://dx.doi.org/10.1038/npre.2010.4626.1

[6] C. Lagoze, H. V. de Sompel, P. Johnston, M. Nelson,R. Sanderson, and S. Warner. ORE Specification -Abstract Data Model. Technical report, Open Archives Initiative, 2008.

[7] Talbott, T., Peterson, M., Schwidder, J., & Myers, J. D., "Adapting the electronic laboratory notebook for the semantic era", Proceedings of the 2005 International Symposium on Collaborative Technologies and Systems, 136-143.

[8] Frey, Jeremy, De Roure, David, Taylor, Kieron, Essex, Jonathan, Mills, Hugo and Zaluska, Ed, "CombeChem: a case study in provenance and annotation using the Semantic Web", International Provenance and Annotation Workshop, IPAW 2006 Berlin, Germany, Lecture Notes in Computer Science, 4145/2006, pp. 270-277, http://dx.doi.org/10.1007/11890850_27

[9] Fedora Commons (http://www.fedora.info/)

[10] Goble, D. R., & Goble, C.. "myExperiment - A Web 2.0 virtual research environment", presented at the International Workshop on Virtual Research Environments and Collaborative Work Environments, Edinburgh, UK., 2007

[11] Linked Open Data (LOD) (http://linkeddata.org/)

[12] Myers, J., Hedstrom, M., "Active and Social Curation: Keys to Data Service Sustainability", position paper, National Data Service (NDS) Consortium Planning Workshop, Boulder, CO, June 12-13 2014

[13] Hedstrom, M., Alter, G., Kouper, I., Kumar, P., McDonald, R.H., Myers, J., and Plale, B., "SEAD: An Integrated Infrastructure to Support Data Stewardship in Sustainability Science", NSF - Research Data Management Implementation Workshop. Arlington, VA., March 13-14, 2013

[14] National Data Service vision video (https://www.youtube.com/watch?v=BPT1FNFAvnc)

[15] DataOne OneMercury data search tool (https://cn.dataone.org/onemercury/)

[16] Futrelle, J., Gaynor, J., Plutchak, J., Myers, J. D., McGrath, R. E., Bajcsy, P., Kastner, J., Kotwani, K., Lee, J. S., Marini, L., Kooper, R., McLaren, T. and Liu, Y., "Semantic middleware for e-Science knowledge spaces" Concurrency Computat.: Pract. Exper., 23: 2107–2117, 2011, doi: 10.1002/cpe.1705

[17] Marini, L., R. Kooper, J. Futrelle, J. Plutchak, A. Craig, T. McLaren, and J. D. Myers, "Medici: A Scalable Multimedia Environment for Research", Microsoft Research eScience Workshop, Berkeley, CA, 10/11/2010

[18] rSEAD Library (https://opensource.ncsa.illinois.edu/stash/projects/MMDB/repos/rsead/browse)

[19] Eclipse Rich Client Platform (https://wiki.eclipse.org/index.php/Rich_Client_Platform)

[20] Microsoft Seadragon visualizer (http://en.wikipedia.org/wiki/Seadragon_Software)

[21] Geoserver open source geospatial data server (http://geoserver.org/)

[22] Beth Plale; Praveen Kumar; Jim Myers; Margaret Hedstrom; Robert McDonald; Stacy Konkiel; Kavitha, Chandrasekar, "SEAD Virtual Archive: Building a Federation of Institutional Repositories for Long Term Data Preservation", Infrastructure, Intelligence, Innovation: Driving the Data Science Agenda, 8th International Digital Curation Conference 2013 (IDCC13) 14-16 January 2013, Amsterdam, Netherlands

[23] Mayernik, M.S., Choudhury, G. S., DiLauro, T., Metsger, E., Pralle, B., Rippin, M., Duerr, R., "The Data Conservancy Instance: Infrastructure

And Organizational Services For Research Data Curation“, D-Lib Magazine 18, no. 9/10, 2012

[24] Suriarachchi, I, Zhou, Q and Plale, B., “Komadu: A Capture and Visualization System for Scientific Data Provenance”, Journal of Open Research Software 3(1):e4, 2015, http://dx.doi.org/10.5334/jors.bq

[25] DataOne Member Node Application Programming Interface https://mule1.dataone.org/ArchitectureDocs-current/apis/MN_APIs.html

[26] Krafft, Dean B., Cappadona, Nicholas A., Caruso, Brian, Corson-Rikert, Jon, Devare, Medha, Lowe, Brian J. and VIVO Collaboration, “VIVO: Enabling National Networking of Scientists”, Proceedings of the WebSci10: Extending the Frontiers of Society On-Line, Raleigh, NC, April 26-27th, 2010

[27] VIVO-ANDS Ontology (http://purl.org/ands/ontologies/vivo/)

[28] Digital Record Object Identification (DROID) software (http://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/file-profiling-tool-droid/)

[29] NDS@NCSA Hackathon: An experiment in community development (https://wiki.ncsa.illinois.edu/download/attachments/34768893/Hackathon.pdf)

[30] Kouper, Inna, Plale, Beth, Akmon, Dharma, Hedstrom, Margaret, “Practical and Conceptual Considerations of Research Object Preservation”, presentation, Digital Preservation 2014, Washington, DC. (http://www.slideshare.net/InnaKouper/research-objects-preservation)

[31] Jun Zhao, Carole Goble, Robert Stevens, “An Identity Crisis in the Life Sciences:, Provenance and Annotation of Data, International Provenance and Annotation Workshop, IPAW 2006, Chicago, IL, USA, May 3-5, 2006

[32] James D. Myers, “I Think Therefore I Am Someone Else: Understanding the confusion of granularity with Continuant/Occurrent and related perspective shifts”, Proceedings of the 2010 International Provenance and Annotation Workshop (IPAW 2010), Provenance and Annotation of Data and Processes, Lecture Notes in Computer Science Volume 6378, 2010, pp 292-294

[33] James P. McCusker, Timothy Lebo, Cynthia Chang, Deborah L. McGuiness, Paulo Pinheiro da Silva, "Parallel Identities for Managing Open Government Data", IEEE Intelligent Systems, vol.27, no. 3, pp. 55-62, May-June 2012, doi:10.1109/MIS.2012.5

[34] K. McHenry, J. Lee, M. Dietze, P. Kumar, B. Minsker, R. Marciano, L. Marini, R. Kooper, D. Mattson, "DIBBs Brown Dog, PaaS for SaaS for PaaS", XSEDE Reproducibility Workshop, 2014

[35] Karger, D.R., "The Semantic Web and End Users: What's Wrong and How to Fix It," Internet Computing, IEEE , vol.18, no.6, pp.64,70, Nov.-Dec. 2014, doi: 10.1109/MIC.2014.124