IJDC | *General Article*

Building Tools to Support Active Curation: Lessons Learned from SEAD

Dharma Akmon
University of Michigan

Margaret Hedstrom
University of Michigan

James D. Myers
University of Michigan

Anna Ovchinnikova
University of Michigan

Inna Kouper
Indiana University

Abstract

SEAD – a project funded by the US National Science Foundation's DataNet program – has spent the last five years designing, building, and deploying an integrated set of services to better connect scientists' research workflows to data publication and preservation activities. Throughout the project, SEAD has promoted the concept and practice of "active curation," which consists of capturing data and metadata early and refining it throughout the data life cycle. In promoting active curation, our team saw an opportunity to develop tools that would help scientists better manage data for their own use, improve team coordination around data, implement practices that would serve the data better over time, and seamlessly connect with data repositories to ease the burden of sharing and publishing.

SEAD has worked with 30 projects, dozens of researchers, and hundreds of thousands of files, providing us with ample opportunities to learn about data and metadata, integrating with researchers' workflows, and building tools and services for data. In this paper, we discuss the lessons we have learned and suggest how this might guide future data infrastructure development efforts.

Received 23 January 2017 ~ *Accepted* 4 December 2017

Correspondence should be addressed to Dharma Akmon, 2116C ISR-Perry, Ann Arbor, MI 48104-1248. Email: dharmrae@umich.edu

An earlier version of this paper was presented at the 12th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

Over the last decade, significant attention has been placed on developing infrastructure (software, tools, services, and policies) to support creation, sharing, preservation and reuse of research data. The SEAD (Sustainable Environment/Actionable Data) project[1] —sponsored as part of the U.S. National Science Foundation's DataNet program — was formed to build capabilities that serve the data management and curation needs of researchers in the long tail, which is characterized by small research teams and individual researchers without easy access to advice, training, storage, or other services for their data. The SEAD team has spent the last five years designing, building, and deploying an integrated set of services to better connect scientists' research workflows to the data publication and preservation activities that are increasingly demanded of them (Myers et al., 2015).

Throughout the project, we have promoted "active curation" by providing researchers with software tools and services to capture data and metadata from the beginning of a research project and to allow data producers and others to refine and improve the data throughout the research lifecycle (Myers and Hedstrom, 2014). We wanted to address what we and other researchers have observed (e.g. Marchionini, Lee, and Bowden, 2012) as an inefficient and cumbersome traditional approach to data preservation, which positions curation as a process separate from most other research activities, invoked primarily at the end of a project when researchers are required or otherwise compelled to make data more widely available. In promoting active curation, our team saw an opportunity to develop tools that would help scientists better manage data for their own use, improve team coordination around data, implement practices that would serve the data better over time, and seamlessly connect with data repositories to ease the burden of sharing and publishing. We also anticipated benefits in our approach for repositories, including higher deposit rates and better quality initial submissions.

SEAD is made up of several key services intended to serve data needs throughout the research lifecycle. Project Spaces[2] are secure, team-controlled workspaces where scientists can collaboratively upload, organize, annotate, preview, and download their team's research data. Like commercial cloud-based services, such as Dropbox, SEAD makes it possible for anyone to create a Project Space and, within that space, define the set of people who can upload, maintain, and share files. Unlike simple file sharing services, SEAD supports annotation of files with formal metadata, relationships, and informal tags. In addition, SEAD offers several key capabilities aimed at serving the needs of research teams and scientists. Rather than requiring users to follow specific standards for metadata, or requiring data in particular formats, SEAD users can store and share data in any format; implement the metadata schema(s) that suits their scientific needs; add custom metadata terms; and track a variety of relationships within, and between, datasets. Researchers can also create project profiles and add logos and other graphical elements to the public view of their Project Space. In combination with a role- based access control system that allows Project Space administrators to control who can view, annotate, upload, and download data, these features provide researchers significant freedom to manage their data in ways that meet individual and team needs.

1 SEAD Project website: http://sead-data.net/
2 SEAD Project Spaces: https://sead2.ncsa.illinois.edu/spaces

SEAD Project Spaces include a publication Staging Area where researchers can prepare datasets for publication, add or edit metadata, include files as they see fit, select a repository, and submit the packaged data and metadata to that repository for review, additional curation, dissemination, and preservation. A repository Matchmaker service compares traits of the proposed data publication (e.g. file formats and sizes, metadata provided, institutional affiliations of the data creators, etc.) with the acceptance criteria of SEAD's partner repositories (e.g. openICPSR[3]) to provide guidance to researchers in selecting an appropriate repository for their data. Lastly, this service also notifies repositories of new SEAD-brokered submissions so that the repository can ingest them into their own workflows and systems. SEAD provides repositories with a serialized copy of all metadata, formatted using the JSON-LD syntax of the Open Archives Initiative Object Reuse and Exchange (OAI-ORE) standard that includes links through which individual data files can be retrieved as needed for that repository's internal processing.

Through the above services, SEAD has worked with 30 projects, dozens of researchers, and hundreds of thousands of files, providing us with ample opportunities to learn about scientists' data and metadata; integrating new tools into researchers' workflows; and building tools and services for data. In this paper, we discuss what we learned and how this might guide future data infrastructure development efforts.

Data and Metadata

Our first set of project lessons concerns scientists' data and metadata. We found that to credibly serve as a platform for scientists to manage and share data while they collect and work with them, data services must be prepared to accept a wide range of both proprietary and open, non-proprietary file formats; facilitate arrangement of files into meaningful, hierarchical groupings; provide efficient navigation across such groupings; and satisfy local annotation needs while nudging researchers into best metadata practices where appropriate.

We have observed more than 140 different file formats and format variations representing raw and processed data, figures, reports, presentations, software, experimental methods, databases, and other material. The file types found in SEAD Project Spaces encompass primarily the following:

- Tabular data – .csv, .wks, .xml, .xls, and .xlsx
- Geospatial data – .tfw, .tiff,
- Digital Image data – .png, .tif,
- Documents – .doc, .pdf, .txt
- Digital Video data – .mov, .mp4
- Other software or instrument specific formats – e.g. Matlab (.mat)

While accepting such a variety of formats is a necessity for a data management service, it is not without challenges. For one, our work with scientists has shown in-browser previews of files to be a key feature of SEAD Project Spaces, because it

3 openICPSR: https://www.openicpsr.org/openicpsr/

removes download as a prerequisite to browsing file contents and annotating data. SEAD displays previews for several popular file types—including .pdf, .doc, .jpg, .tiff, .xls, and .csv—but scientists frequently upload data in formats they cannot preview through SEAD. Additionally, as data curators will note, many of the file types enumerated above are not recommended for data preservation, primarily because continued access to them depends on the availability of closed platforms that often evolve in ways that are not backward compatible. But rather than imposing format restrictions as researchers upload data into a Project Space, SEAD surfaces each repository's file format requirements when researchers begin the publication submission process. They are then given the opportunity to either choose a repository that accepts all the files in their submission — if there is such a SEAD partner repository — or to change their submission to comply with format restrictions. We think that there is room for further improvement on this model, by making such requirements apparent to users earlier in the research process and potentially nudging researchers toward preservation-friendly formats where appropriate. For example, when a researcher uploads tabular data as a Microsoft Excel file, the system could let her know that many long-term repositories require or prefer the open, non-proprietary .csv format. Furthermore, we could prioritize previewing capabilities along these lines and offer the highest-fidelity, in-browser previews to open non-proprietary formats, to encourage scientists to utilize the open format.

In terms of how SEAD's users organize their files, researchers sought to recreate the hierarchical organization of files that they commonly employ in their computer file systems. These hierarchies, which were often very deep when many files were involved, contained important contextual information about the data that would be lost, unless translated as metadata. For example, in a Project Space named for a site-based study of water movement that included several datasets (a dataset is a group of files in the current iteration of SEAD), each file name referred to the specific location and time period of data collection. Each dataset contained a flat list of files within it. Another Project Space was named for an observatory and contained datasets that were named for each specific study at the observatory along with a list of files list associated with each study.

Projects with large numbers of data files typically arranged their files into hierarchies two or more levels deep (and some had up to 12 levels). One project, for instance, contained a single dataset organized into five folders, roughly approximating steps in the research process: results and processed data; original data; log data; study design; and code and settings. Some of these folders contained subfolders to further organize the files within them. For example, the folder for results contained subfolders corresponding to type of data: topographical, photographs, movies, and so on. In another example, a dataset contained folders for code/syntax, meeting recordings, notes, output, and paper drafts and literature. Such groupings not only provided important context that indicated what files were about, they also made it easier to navigate across collections of files that could number in the hundreds, if not tens of thousands. While critical to an active data management system, this is an area where we see a significant, and possibly unnecessary, disconnect between active data management and data publication and preservation services. Many repositories only allow for single-level groupings of files (e.g. FigShare, Dryad, and institutional repositories such as DeepBlue Data at University of Michigan), thus SEAD's Matchmaker identifies appropriate repositories based, in part, on how many levels deep they allow collections to be. This approach potentially requires researcher effort to reorganize their data, document relationships between files in some other way, and make important contextual

 doi:10.2218/ijdc.v12i2.552

information that is captured in file names and hierarchies explicit through enriched metadata. In some cases, archived collections may be less navigable and more difficult to reuse than their active counterparts.

As with data file type and file organization, SEAD was designed to flexibly support researchers' metadata needs. We aimed to encourage users to add metadata throughout their projects by making it yield immediate benefits for data producers (e.g. by including metadata in text search) and reduce the effort needed to prepare data for publication later. This required a careful balance between local needs, community-specific standards, and repository requirements. SEAD captures some metadata automatically (e.g. title, file size, upload date) with most terms mapped to the Dublin Core schema. In addition, SEAD supports a configurable list of metadata terms that data producers can add values to at any time. The default set of options for user-entered metadata in SEAD Project Spaces is also largely based on Dublin Core (e.g. alternative title, creator(s), abstract, contact, spatial coverage, and temporal coverage) with some additions from other vocabularies to support common scientific metadata (e.g. funding institution and method). Researchers can also attach free-form tags to their data. The only metadata required in the system is a title. Otherwise, researchers can fill out as much, or as little, metadata as they wish and can determine the specific conventions they will use for particular fields. Project Space administrators also have the ability to change the set of metadata fields for their space. One of the groups we worked with, for example, altered the set to be more like the Ecological Metadata Language (EML) standard. Another team wanted their Project Space to be customized with metadata terms required by the Inter-university Consortium for Political and Social Research (ICPSR), where they would eventually archive the data.

An analysis of how researchers actually applied metadata to their data in SEAD reveals that many teams added metadata primarily when they were ready to publish data, sometimes over the course of several days. The most popular practice was to provide a core set of fields applied at the dataset level, as opposed to file-level. Several projects, for example, only entered a description/abstract and/or creator(s) for the dataset to complement the system-extracted metadata at the file level, such as file name, date uploaded, file type, and location. In fact, description/abstract and creator(s) were the two most commonly applied metadata fields in SEAD, a finding that is not particularly surprising, given that these two fields are required for publication and, hence, also figure prominently on the dataset information page. A few of the teams applied a much more extensive set of metadata, including spatial coverage (i.e. location) and temporal coverage (i.e. when the data were collected), relationships (e.g. "is a version of" or "is part of"), method, audience, contact, and funding institution.

The most valuable lesson we learned concerning metadata is that while scientists need some amount of flexibility in documenting their data, they also frequently want guidance in best practices and functionality that more explicitly directs their teams toward these practices. Rather than promoting metadata annotation, we found that the open-ended, highly flexible metadata functionality we implemented could be a hindrance to it. We see opportunities here to make it easier for researchers to identify the appropriate metadata standards for their data and apply them more skilfully. Some of the researchers that come to SEAD already know what repository will eventually archive their data. Where SEAD is a partner with that repository, it could suggest the set of metadata fields the researchers will eventually need, as well as provide in-line guidance on how such fields should be used. Similarly, we have considered ways to highlight metadata fields that would be required by a team's preferred archive. Without forcing researchers to adopt a standard or requiring them to enter metadata at the same

time data are uploaded, SEAD may nonetheless help researchers adopt best practices in metadata annotation, potentially yielding better-curated data and less repeated effort on the part of scientists.

Workflow Integration

The second set of lessons from SEAD concerns the benefits of and obstacles to integrating active data management into scientific research and publication workflows. SEAD was designed to support multi-disciplinary projects from the broad area of sustainability science, and we attracted teams with a range of ideas as to how they could best employ SEAD for their projects. Our initial project and product vision was predicated largely on the assumption that research teams would use SEAD Project Spaces for data management while they collected and worked with data. Many scientists, however, did not view their teams' needs or SEAD's offerings in the same way. In fact, rather than using SEAD for data management from the beginning of their projects, teams typically did not request assistance or start using the SEAD platform until they needed to identify a suitable repository and get help preparing data for publication and long-term preservation.

Researchers came to SEAD to accomplish one or a combination of three main things: to create a centralized place for their team to organize, store, and share data among team members; to more easily disseminate their data beyond their teams; and to identify appropriate long-term repositories for publishing and archiving their data. Because of the importance of incremental metadata annotation to the SEAD vision, our outreach and marketing efforts heavily emphasized SEAD as a tool for managing data during a project and then for easily transitioning data, as needed, to long-term repositories. This message seemed to resonate most with teams whose members were geographically distributed and who already knew where their data would be archived; and for data managers and curators who worked closely with researchers to get data into their repositories. SEAD Project Spaces have the capability not only to centralize the data that teams gather, but also to provide an easier way to create the documentation that publishers and repositories require of them. Data managers saw this as an opportunity to engage researchers around data management and improve the completeness and level of description of the data being submitted to their repository.

Although few teams adopted Project Spaces as an active data management tool, several teams took advantage of SEAD's support for ongoing uploads and annotation to work over the course of days—or even weeks and months—to gather and document data for publication. In this part of the overall data workflow, SEAD's enhancements for active curation, including commenting capabilities that would send email notifications to colleagues, the ability to view geospatial data from multiple files as a map overlay, and the ability to annotate a portion of a data file, were compelling reasons to adopt SEAD even if they did not make SEAD a full replacement for a project's existing active data management infrastructure.

SEAD garnered significant interest and use from teams that had already produced data and were ready to publish or otherwise share data[4], but did not necessarily know where they should go for guidance or how to disseminate them more widely. From the

4 Using SEAD, researchers can control access to their datasets, including making data public. We contrast this with publishing the data, whereby data are "fixed" and deposited in a SEAD partner repository where they receive a DOI.

 doi:10.2218/ijdc.v12i2.552

outset, SEAD was intended to serve as a broker between researchers and existing repositories to the benefit of both data producers and repositories. By supporting data management and active curation, preparing data for publication and archiving would demand less effort on the part of researchers and repositories would receive higher-quality and better-documented datasets. In spite of the rapid growth of research data management services and repositories since the SEAD project began in 2011, many of SEAD's users were not able to identify appropriate repositories for their data. Our experience working with researchers on multidisciplinary teams and producing a wide variety of data types revealed some limitations in the existing data preservation infrastructure. For multi-disciplinary teams with members from several different institutions, there was no obvious institutional repository or suitable domain repository for their data. In several cases, the diversity, quantity, and structure of the datasets taxed or exceeded the capabilities of repositories to store and disseminate the data.

Tighter coupling of research data management with publication and preservation workflows requires adjustments by both data producers and repositories. SEAD demonstrates that it is possible to offer scientists a high degree of flexibility so that they can focus on their research and still achieve consistency and adoption of good data management practices. Providing a set of interoperable web services, allowing researchers to decide which services meet their needs, and offering ways for researchers to customize metadata schema, terminology, and definitions, is preferable to a monolithic system with hard requirements and constraints. Yet achieving consistency, a steady flow of high quality data into repositories, and affordable levels of curation, will not happen automatically even with an ideal platform. Platforms like SEAD will still need to provide guidance to researchers and tools that nudge users toward best practices. At the same time, repositories will need to become more flexible and increase their capabilities to handle complex data collections, very large datasets, and many new types of data in order to meet researchers' needs and expectations.

Building Tools and Services

A final set of lessons concern what we learned about building a set of production-level data services sufficiently robust and usable for researchers to adopt them on their own merits and to integrate them into their work. To more readily accomplish this, we adopted an Agile, user-centered design approach that emphasizes short cycles of iterative development; frequent contact with users to understand their needs; and the collaborative development of user stories that both articulate and capture those needs and serve as the basis for development (Cohn, 2004; Patton, 2014).

Much of the early work in the SEAD project was typical of any project attempting to move from a prototype and capabilities that are sufficient for demonstration to a robust system that can support ongoing use by researchers, without relying too heavily on user support. In addition to identifying many software bugs, researchers made it clear that their adoption of SEAD's tools depended on how well those tools met their expectations of usability and whether or not SEAD offered features common to collaboration and file management systems. While innovative services such as repository matchmaking were well-received, they were not, on their own, enough to spur adoption. As a result, we spent significant design and development effort adding enhancements that research software developers might have characterized as "pedestrian," but that users thought of as customary in web applications. For example,

we enhanced SEAD to make it possible to parallelize file upload; upload files over 4GB in size; enter special characters in metadata; and create a Project Space without SEAD team assistance. We also added help text and documentation; rearranged and relabelled buttons to make their use more intuitive; added confirmation messages; edited jargon-laden text; and implemented "breadcrumbs" to facilitate wayfinding and navigation. While we relied heavily on the researchers using the system to identify priorities, provide examples, and evaluate potential solutions; we also worked hard to roll out updates on a frequent basis so that we could validate our enhancements and more quickly improve on them.

As we made our systems more robust, we increasingly focused on improving SEAD's active curation capabilities. In doing so, we discovered significant challenges to researchers' adoption of SEAD as an active data management platform. One of the biggest obstacles was that, while researchers could use SEAD to collaborate around gathering data together, arranging files into hierarchical groupings, and annotating them with appropriate metadata, they could not easily integrate work that required manipulation of the files' content from SEAD's Project Spaces. For example, for researchers to change a value within a file, they would need to download the file from SEAD, make the changes, and then upload the updated file. Furthermore, they would have to do this while hoping no one else on their team was also working with the file. Contrasted with making changes to data in a file that is on a shared drive, this is a meaningful difference that meant extra work and coordination on the part of data producers. Similarly, running analyses on data uploaded to SEAD required first downloading them to a local machine with the analysis tools. These limitations meant that, for many scientists, the data had to be in a relatively finished state for SEAD to be useful to them. Our team considered implementing desktop file syncing to ameliorate these issues, but quickly realized such a feature was likely to introduce significant complexity regarding the metadata attached to the files. For example, would the metadata added to the file, or to the dataset in SEAD, still be valid if the original file changed? If not, how could changes to the metadata be supported without adding significant work for the data producers? As a result, we ultimately decided not to pursue this area of development. Still, we were encouraged by researchers' requests, for improvements to features such as geospatial data display and graphed time-series data, as well as suggestions for us to add the ability to trigger email notification by mentioning a colleague's name in a comment. We interpreted these requests as indications that researchers were interested in working within SEAD to explore features of their data visually and to coordinate some of their data activities with colleagues.

Even while the Agile software development approach helped us better align SEAD with data producers' needs, it presented difficulties for our geographically dispersed, multi-institutional team. First, because SEAD's services were implemented as separate, interoperable components, we sometimes wrongly assumed a single feature would fulfil a user need, when in reality an entire set of features across the components was required. For example, we implemented parallelized file uploads thinking it would help users with large datasets. However, this was of limited usefulness until we also improved the ability to display, annotate, and publish larger collections. Our team was also challenged at times to define and prioritize use cases that were likely to make practical advances for many users versus compelling cases that were beyond our capability to build and deploy as robust services or that would serve only a subset of SEAD users. This tension came up when some of our users requested workflow-oriented tools such as document locking and the ability to use SEAD to create and assign tasks to individuals; as well as the capability to support complex data models and

 doi:10.2218/ijdc.v12i2.552

allow users to execute computation within a Project Space. Such features could have bolstered SEAD as an active data management tool and analytical platform, and we seriously considered them. However, as our team discussed these requests both internally and with users, we realized that implementation would be far more complex than it initially sounded. The ability to lock files, for instance, suggested integration with file editing, bringing up, once again, the challenge of maintaining metadata for changing files. We also recognized that task management was not one but many use cases because researchers want to address both coordination and computational workflows; group leaders need to assign and monitor specific tasks or entire workflows; and peers and collaborators frequently make ad hoc requests. With no clear consensus on the functionality required, or the extent of the challenges to connect task management with data, metadata, and computation, our team ultimately decided to concentrate on improving notification capabilities (e.g. showing the most recent uploads in a Project Space and triggering emails when users were mentioned in comments). Integration of data, metadata, workflows, computation, publication, and archiving remains a critical challenge for the next generation of scientific data platforms.

Conclusion

Perhaps the most promising lesson we learned is that researchers are highly interested in improving their team's data management practices. Furthermore, with user-friendly tools and guidance, they are very willing to provide data and metadata. The feedback that researchers have given us and what we have observed in practice, indicate that the kind of capabilities SEAD offers make it easier for researchers to assemble, annotate, and publish rich collections of data. We attribute this to the combined ability to support any data and metadata types at significant scale, with the means for research leaders to control access and branding and to define practices for their groups. SEAD is still in active development, and we anticipate ongoing improvements to better support the current workflow of uploading, annotating, and publishing data. We also recognize that significant new development will be required to fully realize our active curation vision. We are hopeful that the lessons we have learned will be applicable in those endeavours and to other data services developments.

Acknowledgements

This research was funded by the National Science Foundation under cooperative agreement #OCI0940824. The authors acknowledge the contributions of the broader SEAD project team 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 users, who have provided invaluable guidance and feedback throughout the project.

References

Cohn, M. (2004). User stories applied: For Agile software development. Addison-Wesley Professional.

Marchionini, G., Lee, C.A., Bowden, H. (2012). Curating for quality: Ensuring data quality to enable new science. Final Report for Invitational Workshop Sponsored by the National Science Foundation. Retrieved from http://openscholar.mit.edu/sites/default/files/dept/files/altman2012-mitigating_threats_to_data_quality_throughout_the_curation_lifecycle.pdf

Myers, J., Hedstrom, M., Akmon, D., Payette, S., Plale, B.A., Kouper, I., et al. (2015). Towards sustainable curation and preservation: The SEAD project's data services approach. 2015 IEEE 11th International Conference on e-Science. doi:10.1109/eScience.2015.56

Myers, J. & Hedstrom, M. (2014). Active and social curation: Keys to data service sustainability. White paper presented at National Data Service Consortium Planning Workshop, June 12-13, 2014.

Patton, J. (2014). User story mapping: Discover the whole story, build the right product. O'Reilly Media.