SEAD | Sustainable Environment Actionable Data

Report from the second SEAD User Study

May 23, 2013

sead-data.net

Final Report from Second SEAD User Study

Contents

Executive Summary

Sustainable Environment Actionable Data (SEAD) is one of five partners in the DataNet program funded by the National Science Foundation. The objective of the DataNet partnership is to build component elements of an interoperable data preservation and access network for science and engineering data. SEAD's goal is to serve researchers addressing questions related to the complex interactions between natural and social systems. More specifically, SEAD aims to support data management and sharing and provide safe, secure storage for data created and used by scientists in the emerging field of sustainability science.

The role of domain engagement in SEAD is to give stakeholders the opportunity to directly influence the design of SEAD tools, policies, and services. One way in which SEAD has engaged members of the target user community is through formal investigation of their needs and requirements. To date, two such studies have been conducted. The first was carried out in early 2012, and the second was conducted in spring 2013. This document describes the purpose, methods and findings from the second user study. This report also presents user requirements based on the results and discusses implications for future development of SEAD.

The second SEAD user study was focused on learning more about how individuals might use SEAD at the beginning of a new project and on understanding how scientists work with and share data when they are collaborating with other researchers. We interviewed seven researchers from several disciplines; most of those we spoke with were in the early stages of their careers. The interviews were conducted between March 18 and April 19, 2013 and ranged from 60 to 75 minutes in length. We asked researchers about their experiences working with and sharing data, particularly in collaborative research projects. We sought information on what gets shared and how it occurs, what works well, what limitations exist, and thoughts about what might improve the process. We also elicited feedback from interviewees on static mockups of SEAD interfaces that represented its current capabilities.

Some of the findings confirmed what was learned from the first user study; this included information regarding data heterogeneity and levels, data tasks and workflows, and categories of users. New themes also emerged. Specifically, researchers appear to want the following:

- to share a richer 'data model' than files while recognizing that additional ease-of-use and automation will be needed to enable that capability

- to share data and metadata, without having to manage keeping them in sync; thus, they want the metadata and data to be in one file/package
- version and provenance support through tools to keep track and display that metadata automatically and intuitively so that managing that extra information does not cost more than it is worth.
- to integrate their data and reference data, and integrate their data and collaboration experiences while minimizing the burdens of switching tools or adding lots of info to achieve that integration

Results from this also study indicate that while perspectives, concepts, and methods may differ across disciplines, many of the data needs and practices are similar. Thus, data management, preservation and access services do not have to be unique or designed from the bottom up for each discipline. Equally important, the findings suggest that when collaborating across disciplines, researchers face challenges that include converting formats, coordinating and communicating, recording activities and provenance, and organizing data and documentation. Some of these problems are different than those found when studying individual users in specific disciplines, and they deserve further attention in future interactions with potential users of SEAD.

SEAD is prioritizing use cases and functionality that need to be completed in order to release a version of SEAD to select groups of friendly users. The findings from the user study reported have been and will continue to be used to help determine these priorities.

1. Introduction

Sustainable Environment Actionable Data (SEAD) is one of five partners in the DataNet program funded by the National Science Foundation (NSF). The objective of the DataNet partnership is to build component elements of an interoperable data preservation and access network for science and engineering data (NSF, 2007; Wikipedia, n.d.). SEAD is a collaborative partnership among researchers at the University of Michigan (lead), Indiana University, and the University of Illinois at Urbana-Champaign. The SEAD team has expertise in core functions related to SEAD, including sustainability science, data preservation and access, and systems design and development.

SEAD's goal is to serve researchers addressing questions related to the complex interactions between natural and social systems and the ways in which those interactions affect the challenge of sustainability. More specifically, SEAD aims to support data management and sharing and provide safe, secure storage for data created and used by scientists in the emerging field of sustainability science. In addition, SEAD is focused on supporting scientists working across disciplines in "the long tail." Through prior research, reviews of reports, discussions with sustainability scientists, and findings from the first SEAD user study, SEAD has identified some of the following characteristics of the cross-disciplinary long tail.

- The focus is on solving complex problems while adding new knowledge to several different disciplines.
- Researchers require many different data types on everything that characterizes a particular system (e.g., river) and all factors that impact it.
- Research projects use a combination of observational (field) data, experiments, and models.
- Researchers extract variables and data points from large, standardized data sources, often combining these with new and unique observational and experimental data.

While data sets produced by single investigators are present in the long tail, small and medium-sized teams carrying out multi- and inter-disciplinary research are common. In fact, group research is a larger challenge to address than individual research given that teams are likely to face data integration challenges in their work and must contend with multiple data formats, metadata standards, data management practices, and disciplinary conventions.

In order to be successful in supporting scientists working in the long tail, SEAD must add value for researchers. The role of domain engagement in SEAD is to give stakeholders the opportunity to directly influence the design of SEAD tools, policies, and services. One way in which SEAD has engaged members of the target user community is through formal investigation of their needs and requirements. To date, two such studies have been conducted. The first was carried out in early 2012, and the findings were reported in Yew (2012). This document presents findings from the second study of SEAD users, which was conducted in spring 2013.

The report begins with a discussion of SEAD's development approach, a description of the SEAD prototype, and an overview of the first study of SEAD users. These topics are followed by information on the purpose, methods and findings from the second user study, including reaction to SEAD mockup interfaces. User requirements based on the results are then presented. Finally, we discuss implications of findings for future development of SEAD and next steps in domain engagement.

2. SEAD Development and Prototype

SEAD employs an agile approach to better understand users and their domain (See Fig. 1). This approach balances the need to provide clear plans and measurable outcomes – through a long-term roadmap – with the ability to respond quickly and effectively to emerging requirements. Each major phase in stakeholder engagement leads to a review of the roadmap and, as necessary, to changes in the functionality of SEAD components.

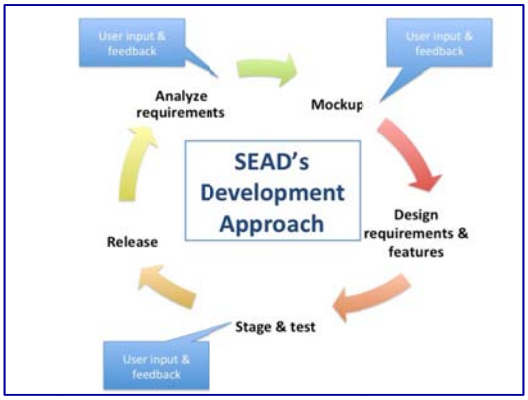


Figure 1: SEAD's agile development approach

Based on the initial vision for SEAD and lessons learned through agile development, particularly formal and informal interactions with potential users, SEAD developed a prototype to provide proof-of-concept for SEAD's ability to support the end-to-end management of data from the start of a new project to the long-term preservation of data, metadata, and publications. The current version of the prototype consists of three loosely coupled components.

- ***Active Content Repository (ACR)***: The ACR supports the active use of data. Researchers can upload, organize, annotate, store, and discover data in the ACR using simple tools. They can also auto-extract metadata and preview data. In addition, scientists have the ability to navigate their data by collection, tags and other metadata, location, and provenance or to link to author and publication information. Data owners control access to their data prior to publication or release.
- ***SEAD VIVO***: VIVO is an open-source semantic web application that enables the discovery of research and scholarship. The SEAD VIVO instance is populated with profiles of sustainability researchers, citations to their publications, and data citations. Researchers can use SEAD VIVO to find people who conduct research on topics or in geographic areas similar to their interests and to find publications and data that those researchers have published or produced.
- ***SEAD Virtual Archive (SVA)***: The SVA is a thin virtualization layer that provides the interoperability needed for researchers to publish or release their data in the ACR to a permanent preservation and access infrastructure such as a data repository or a library institutional repository. The SEAD VA also provides global search capabilities across all the member repositories.

Figure 2 shows the functions of the prototype, which include the ability to find people and data; manage and share data; curate and preserve data; and explore sustainability research.

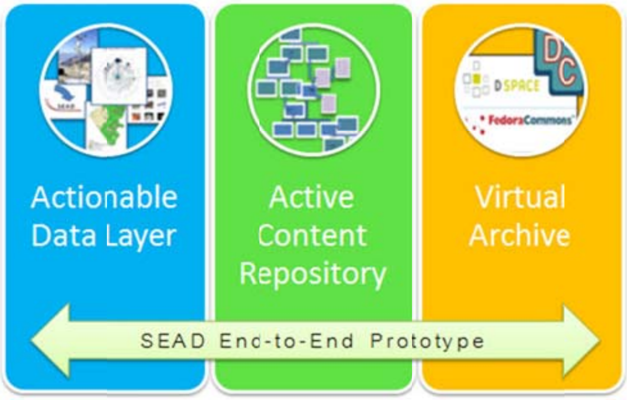


Figure 2: SEAD prototype functionality

Both the first and second studies of SEAD users have focused primarily on needs related to the Active Content Repository. The ACR has been designed to support individuals or groups for whom a simple file store is insufficient – i.e. because additional metadata and data relationships are important for data use, understanding, discovery, and reuse. SEAD also aims to serve users who need to integrate or assemble data from other projects or sources or from reference data available within SEAD. The bullet points below expand on the functionalities and perceived benefits of the current version of the ACR prototype. Figure 3 shows prototype interfaces to some of the capabilities mentioned.

- Users of the ACR see an immediate benefit from adding “metadata” such as tags, authorship, abstracts, comments, provenance relationships, and location metadata or by simply using file formats from which SEAD can extract metadata. Whenever metadata are added, the ACR makes it visible to the rest to the project team, for example, and provides new ways to navigate (find all data with a given tag, navigate to data derived from a given data set, etc.) or view the data (e.g., maps with layers filtered by tag). This is in contrast to traditional curation approaches where metadata is requested at the end of the project, and there is little incentive for the data producer to provide it.
- If individuals or groups using the ACR choose to “publish” their data when they are finished using them, the work they did (e.g., adding tags) can be made visible to the broader community of users.
- Reference data that exist within the ACR or external data or that are brought into it can be enhanced through comments, corrections, tags, etc. and linked to these primary sources and preserved for use by others. Thus, in a very

real way, SEAD enables data consumers to 'borrow' data and return them with more value.

- Users can upload derived data along with the processing and metadata that they have added to create valuable research objects.

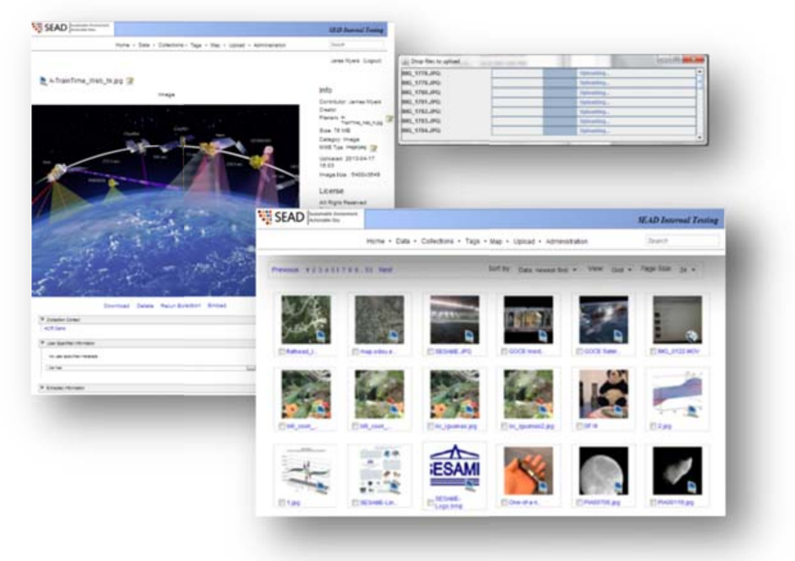


Figure 3: ACR interfaces in SEAD prototype

At the time of the report, SEAD is prioritizing use cases and functionality that need to be completed in order to release a version of SEAD to select groups of friendly users. The findings from the user study reported have been and will continue to be used to help determine these priorities.

Finally, it is important to note that interactions with additional stakeholders have informed the development of other core components of the prototype (i.e. SEAD VIVO and the SEAD VA). These "users" include science managers, who wish to track the productivity and impact of their programs through the capabilities of SEAD VIVO, and representatives from institutional and domain-specific data repositories that will provide the long-term preservation and access structure. Although the results of these engagements with stakeholders are not detailed here, they have been – and will continue to be – important to the design and development of SEAD.

3. First Study of SEAD Users

The current version of the SEAD prototype was developed primarily through engagement with researchers affiliated with the National Center for Earth-surface Dynamics (NCED), who are working on a subset of important sustainability science problems. NCED is an NSF Science and Technology Center that began operation in 2002 and is headquartered at the Saint Anthony Falls Laboratory at the University of Minnesota.[1]

NCED scientists investigate the coupled system of physical, biological, geochemical, and human processes that shape the surface of the Earth and the ways in which they respond to changes in climate, land use, and environmental management. Given this, NCED served as a good proxy for demonstrating what might be achieved in the larger more complex arena of sustainability science and for increasing SEAD's understanding of data types, file formats, data practices, and technologies.

At the time of the first user study, NCED was nearing the end of its ten-year funding cycle.[2] Although NCED has a data repository, the Center was seeking long-term options for continued access to and preservation of its data and a way to migrate additional data into its set of curated collections. Complimentarily, SEAD was looking to populate the ACR with a critical mass of scientific, publication, and people data with which to test its capabilities.

The user study that led to the prototype was carried out in early 2012 by members of the SEAD domain engagement team. The objectives of the study were to determine the needs and requirements of a targeted set of NCED researchers, develop data models and categories, and validate features for the SEAD prototype. The focus of the study was on data management tasks; data analysis and long term archiving of data were beyond the scope of the investigation.

Members of the SEAD domain engagement team visited NCED on March 13-15, 2012 where they learned about NCED research and data and interviewed individuals, including scientists, research assistants, and information technology staff. They also presented potential users with an early mockup sketch of SEAD in order to elicit feedback and gather feature requirements for developers of SEAD. Based on this feedback, new mockups were sketched and presented to interview subjects for their reactions. Members of the SEAD domain engagement team also conducted interviews via phone and Skype in cases where it was not possible to meet users

[1] http://www.nced.umn.edu/

[2] NCED has since received continued funding, at a reduced level, to continue its work.

face-to-face. Detailed findings from the first study of SEAD users are available in Yew (2012). A brief comparison of results from the first and second assessments of user needs appears in section 4.2.1. User requirements that were generated based on the NCED study can be found in Appendix B.

In the months following the user study, the SEAD development team implemented user feedback and populated the ACR with NCED data. Functionality that was added included tools to manage existing data (e.g., bulk ingest tools) and for working with data that are in active use (e.g., project summary page). In addition, NCED researcher profiles and publications were brought into SEAD VIVO.[3] Further, members of the SEAD Virtual Archive group demonstrated SEAD's capability to use the virtual archive layer to re-package data for long-term preservation and access.

4. Second Study of SEAD Users

In spring 2013, SEAD conducted a second formal assessment of user needs. The study had two primary aims. First, we were interested to learn more about how individuals might use SEAD at the beginning of a new project. We anticipated, and findings from interactions with potential users along with feedback from other stakeholders confirmed, that SEAD was most likely to be adopted at the start of a new project. A second goal of the study was to better understand how researchers work with and share data when they are collaborating with others as part of a joint research project. Again, we recognized that SEAD would be most useful at the start-up phase. However, we also needed to learn more about how researchers share data in order to identify strengths and limitations in current processes. Like the first user study, the focus of this investigation was primarily on capabilities for the ACR. A third objective was to expand the disciplinary expertise of the scientists we interacted with through this study. Given these goals, the user study was designed to address the following questions.

What can be done to enhance the sharing and discovery of data at the start of new scientific projects? In particular, what gaps exist between current and desired results when it comes to working with and sharing data among members of collaborative research teams? How can these enhancements be used to leverage the long-term dissemination and curation of data?

[3] During late summer and early fall 2012, members of the SEAD team successfully ingested 20 high-level NCED data collections. The collection consisted of 2.25 million objects, 454,000 files, and 1.6 TB of data. They also worked with NCED to add data on citations by NCED researchers and other personnel and to enhance researcher profiles in SEAD VIVO.

These questions are phrased in the form of a *needs assessment*. *Needs* are defined as gaps between current and desired results. Rather asking people what they "need", an assessment of needs helps to identify gaps in results and highlights opportunities to improve performance (Watkins, Meiers, & Visser, 2012). A needs assessment approach also recognizes that it is often difficult for people to say what they "want" because a) they do not know what is possible; b) they recognize a problem but they are unsure how to solve it, or c) they do not perceive a better solution to a particular challenge. Once gaps are recognized, however, solutions can be explored.

The remainder of this section describes the methods and findings from the second user study.

4.1. Methods

As noted above, one objective of the second SEAD user study was to talk with scientists who are working on somewhat different types of problems than those addressed by NCED. On the other hand, now that SEAD contained NCED data, we were also interested in identifying scientists who might be interested in using these data to test the current capabilities of the ACR. While we hope that a broad range of researchers will find the NCED data in SEAD useful as a source of reference data, in the short-term it seemed most likely that those who would be interested in these data would be individuals who were formerly or are currently associated with NCED. Our recruitment also focused primarily on early career stage scientists as previous interviewees and other feedback we received suggested that junior researchers are more likely to adopt SEAD.

We recruited potential interviewees through a variety of strategies, including contacting participants in EarthCube End-User workshops[4] and seeking recommendations from colleagues. We also asked those we interviewed for suggestions of others who might be willing to participate in the study.

The findings in this report are based on interviews with seven researchers. Three of these individuals had a prior affiliation with NCED (i.e. Subjects 1, 3, & 7). In addition, one of them participated in the first user study (i.e. Subject 1). We were

[4] The NSF EarthCube initiative is funding a series of domain end-user workshops. The purpose of these workshops is to allow members of earth science communities to articulate and document their cyberinfrastructure needs and what they would like to do in the future in terms of accessing data and information within and outside their disciplines. See: http://earthcube.ning.com/page/earthcube-domain-workshops

somewhat successful in our goal to expand the domain expertise of those we interviewed for this second study. Two of the four non-NCED scientists identified themselves as ecologists; a third person described herself as a geologist, and the fourth is an environmental scientist who studies soil moisture – a key variable of the climate system. Three of the interviewees were female and four were male. Table 1 provides additional information about those we interviewed.

Interview subject	Career stage	Primary fields of study	Type of Institution[5]
1	Assistant professor	Ecology & Geology	Research university/high research activity
2	PhD student	Environmental engineering	Research university/very high research activity
3	Postdoc	Geomorphology	Research university/very high research activity
4	Assistant professor	Ecology	Master's colleges & universities (larger programs)
5	Assistant professor	Geology	Research university/very high research activity
6	Associate professor	Ecology	Baccalaureate Colleges--Arts & Sciences
7[6]	PhD student	Geophysics	Research university/very high research activity

Table 1: Scientists interviewed for second SEAD user study

The interviews were conducted between March 18 and April 19, 2013 and ranged from 60 to 75 minutes in length. Five of the interviews took place using Skype and two were conducted over the telephone. All interviewees consented to have the conversation audiotaped, and transcripts were made from the recorded interviews.[7]

[5] Institution type was assigned based on the *Carnegie Classification of Institutions of Higher Education*: http://classifications.carnegiefoundation.org/

[6] This scientist defended his dissertation two weeks prior to the interview. He will begin a postdoc in summer 2013. Since he had not officially received his degree at the time of the interview, he is classified as a PhD student.

[7] Poor telephone quality prevented one of the interviews from being recorded. In this case, extensive notes were taken during and immediately following the interview.

We began each interview by asking people whether the phrases “working with data” and “sharing data” were ones that resonated with them, and if so, what meaning they held. Often people’s responses included examples from their own research which furthered our understanding of their scientific interests and the data they collect themselves or obtain elsewhere. If this was not the case, we followed up by probing for more information about their research interests and the type of data they use. From here, we asked more specifically about their experiences in terms of sharing and working with data, particularly in collaborative research projects. We probed for information on what gets shared and how it occurs, what works well, limitations that exist, and thoughts about what might improve the process.

In the latter part of each interview, we elicited feedback from interviewees on static mockups of SEAD interfaces that represented its capabilities (see Appendix A). We briefly explained each of the three figures we presented, so interviewees could get a better sense of how SEAD might be used. We also emphasized that the goal of the interview was to better understand what they needed and pointed out that SEAD would evolve based on what we learned. Where it made sense to do so, we present reactions to the mockup in the sub-sections that follow. Feedback on the mockups is also summarized and discussed in Section 5.

4.2. Findings

This section begins with a brief comparison of findings from the first user study. Specifically, we highlight results that are similar across the two studies.

4.2.1. Comparison with First User Study

Many of the findings were similar across the two studies, especially those that concern “working with” or managing data. For ease of comparison, these results are summarized below following the headings used originally in Yew (2012, pp. 4-8).

Heterogeneous and complex data: Like most NCED scientists, all of the researchers we spoke to in the most recent study use data from a variety of sources to triangulate on a particular research topic. Six of the seven interviewees routinely collect data from external sources such as state or federal agencies or from colleagues. One of these individuals relies almost exclusively on precipitation, streamflow, elevation and evapotranspiration data obtained from publicly available sources. The majority of scientists we interviewed also collect data from both the laboratory and field. Some of the many data that were mentioned include images (e.g., aerial photographs, digital photographs, aerial and ground-based lidar, movies), observational field data recorded in notebooks (e.g., wind speed, amount of vegetation, visual assessment of bluff stability), field samples that are analyzed in the laboratory (e.g., water, soil, vegetation), and experimental data.

Levels of data: Yew stated that the notion of levels of data was brought up by four of the scientists he interviewed. In this study, all interviewees referred to different levels of data at one point or another; this might be attributed to the greater emphasis on data sharing in this investigation since this is often where the subject arose. A basic typology included raw, processed, and analyzed data. Some interviewees also described sub-levels within these categories. For example, subject 5, a geologist, described the different types of analyzed data she planned to share with her collaborators.

> And then there are different levels of analysis with this. There's the level where I could be giving them just very basic data about the locations of bluffs, and show them where we determined the top of the bluff was in two different years. But I can also give them rates – and what they want is rates. So, there's some analysis that goes into that, but then there's further analysis as well that talks about the processes and that type of information as well.

Another scientist and his collaborators devised an elaborate typology for the data in their study distinguishing between raw data, proofed data, combed data, and proofed deep data.

Data tasks and workflows: Much of Yew's analyses on these topics match findings from the most recent study; we quote his most relevant conclusions below.

> …users need a way to store their data in an "active cache" where they are able to keep, organize, annotate, track versioning, and manage sharing. For Subject 1, the ideal situation would be to have things that were date stamped or versioned that had metadata kind of attached to the files so they don't get decoupled." At present, data are managed and shared using ad hoc strategies like external hard drives, USB storage keys, and shared server space. One unanimous tool that was commonly used and praised amongst all the researchers was Dropbox. The interviewees liked Dropbox because it simultaneously provides backup and version control, and more importantly, is very simple to use.

Yew also noted that the there is a high degree of interdependence between tasks such as data sharing, storage, versioning, and access control. The user requirements he developed were intended to tease apart these tasks (see Appendix B).

User categories**:** Yew developed three categories of users based on the type of research work they do and the kinds of data they work with: field researchers, experimentalists, and modelers. He also noted that the NCED researchers he interviewed carried out work that crosses all three categories. The findings from this study are similar in this regard.

The similarities in findings between the two user studies are captured and highlighted in the User Requirements table in Section 6. Because of the difference in purpose between the two investigations, not all of the requirements that Yew discerned were as apparent from the most recent needs assessment. Most notably, Yew elicited more information on needs pertaining to the management of an individual's data such as organizing, browsing and navigating, and previewing data. Finally, because this more recent investigation focused more heavily on data sharing, new requirements arose; these are also depicted in the User Requirements table.

4.2.2. Working with Data

One aim of this study was to learn more about how scientists might use SEAD at the beginning of a new project. In order to better understand what scientists would expect from a system that helps them to "work with their data," we asked those we spoke with whether this phrase resonated with them, and, if so, what meaning it held. Although this question captured information similar to what Yew obtained through questions about data management, we hoped to increase our understanding about where data "enter the picture" in a new project.

For six of the seven interviewees, "working with data" always encompassed data analysis, and for two people, this was the primary definition.[8] Of these, one person described the phrase as meaning "analysis of raw data," and the other said that to him it meant "…reading a .csv file or text file into a piece of code into a programming language, performing some operation, taking some output, and then storing that output somewhere else that can be part of the next analysis." Two scientists had a somewhat expanded conception of the phrase which included presenting data. As

[8] The seventh person had difficulty answering this question because he said that he finds data difficult to define because there are different levels of data.

one of them stated, it is that "whole thing of figuring out how to analyze, process, and present it, so you get your point across." For the remaining two researchers, what it meant to work with data spanned the entire scientific process. As one of these individuals described it:

> For me, I think of it in the sense of data collection, so going out… I'm collecting field data oftentimes, some experimental, and some other types as well. So, I think about collecting that. Also collecting data off the Internet, as well as processing and compiling that information and trying to make sense of it. And then, finally, at the end, trying to present them in a way that is useful for people outside of the project, or, in some cases, for the general public and other agencies.

Some of the most common challenges that scientists face in working with data stem from the uncertainty of the research process. The dynamic nature of the research process resulted in two frequently mentioned problems: keeping track of versions of files, scripts, etc. and having a written record of the "final" process that led to the data that is then reported in figures, publications, etc.

The challenge of documenting their work was discussed more explicitly by the postdoc and PhD students. Subject 3 summarized the problem: "I found it personally frustrating that I had trouble finding my own data and keeping track of my own data." When asked what make this hard, she said:

> It seems sort of obvious that that you should keep track of your steps exactly, but since so much of it is figuring out what works, writing down the steps is usually done in retrospect after you figure out what works. … And, as a result, it's hard to have a systematic data collection protocol from the beginning because of the way that things are constantly changing. And I can't see that changing in the future. I feel like that no matter what I do there's always going to be that nature to how the experiments are run. Frankly, that's for me the biggest challenge.

The quote above speaks to the difficulty of standardizing data collection, especially up front. Although automation is needed, recording information is not the difficult part. Documenting the same thing for different runs is hard, however, because things change. Two scientists noted that an automatic process to extract at least some of the metadata would be valuable and would also help deal with the challenge of documenting the scientific process as it changes. For subject 6, the ability to associate metadata with data is one of three things could make SEAD useful to her.[9] In general, though, this problem was perceived as a difficult one to solve. For example, subject 2, a PhD student said:

> I manage I think the way a lot of people do, which is you sort of write notes to yourself. You make files that note what you did on what date. And you're sort of relying on the discipline and responsibility of yourself and other researchers to do that kind of work. The value of something like SEAD would be if that type of documentation happened at least – at some level – more automatically. Even if you write it down on SEAD, you're still tapping into the same… Whether I put it in a Word file or put it up on SEAD, I'm still responsible for doing the right thing. I'm going to think to myself, "I just need to get this done and move onto the next task on the list." If you want to encourage better habits, it has to be easier to do the documentation on SEAD than it would be to type it into a Word file, which is already pretty easy. It has to be easier than something that it pretty easy.

So, although the problem is clear to scientists, the solution to it is not obvious. Another finding that is clear from the above is that even though most scientists define "working with data" to mean analysis, the information that they want to capture often begins prior to or with the collection of data.

The second general, and more frequent problem, that interviewees noted is keeping track of versions of scripts, files, datasets, etc. In fact, subject 6, the most senior scientist in the group, said when asked about the biggest challenge to sharing data in her projects:

> The biggest thing is version control – keeping track of the most up-to-date dataset. I would like some way to check changes that have been made. This is the biggest potential source of mistakes.

The above quote shows how the challenges of tracking data are compounded when multiple investigators are sharing data. This problem is discussed further below and in section 4.2.3.

[9] The other two functionalities she mentioned are backup capability and version control.

Subject 3 described the problem of managing different versions of scripts and their output. Her comments were prompted by looking at one of the SEAD interface mockups and are similar to what other interviewees described.

> This has been brought up with other people when I talk to them. When we write scripts to work with the data there's many versions, and scientists are not trained in versioning code or versioning scripts like developers who actually do this for a living; they might have better rules that they abide by. This isn't my own idea, but some I talked to… Well, if there were a workspace, and you had the data and you could keep track of the versions of the script that you used to work with the data, and then you make a figure, and you know which version of the Matlab script you used to make *that* figure and that were all connected together in something like this – a dashboard or whatever – that would be very useful instead of some folder that held all of these Matlab scripts when you were trying to figure things out, but you didn't want to erase any of them because you changed something, and you're not sure if the next version is going to work or not. So, I think wading through all the old scripts is one thing that is difficult in working with older data. Not just finding the right data, but finding the correct scripts.

Subject 7 said simply, "It's one of the most excruciating tasks to open up old scripts and old files and try to make sense of what you were doing a few months ago or a few years ago even." These challenges were also mentioned frequently by individuals interviewed by Yew, who listed "the capability to keep track of the latest version of a dataset" as one of the eight user requirements generated from the NCED field study. Three of the specific functionalities that Yew listed under this requirement were 1) the ability to track different versions of a file; 2) display changes committed to the file; and 3) revert changes made to a file. These are important functionalities for any version control system. What also stands out in scientists' statements is the desire to retain links between files, scripts, and the outputs generated.

Interestingly, two of the seven interviewees (subjects 3 and 5) independently referred to the potential value of a "checklist" as having a role in helping to manage these challenges, and two others (subjects 4 and 7) described a similar concept. Although individuals described different purposes for a "checklist," its overall role was as a communication and tracking mechanism.

For some, a checklist was seen as a tool to help new researchers, particularly graduate students, learn what is important to capture about the scientific process and to help them establish priorities among tasks to be completed in a particular project. As subject 5, who is now an assistant professor, said:

> I guess when I was a graduate student the hardest thing for me was I didn't always know what the most important next step was. And so I would go around doing what I thought I should do. And there was a time period when I had what seemed like four advisors, and they would all tell me different things to do – what I should be doing next. And then it would have been nice to have a list – kind of a checklist: "Well, this needs to get done; this needs to get done."

When asked what would be on the checklist, she said:

> For us – we were looking at river migration – so, getting river migration figured out. Getting bluff retreat rate figured out over decadal time periods. Then I did the terrestrial laser scanning, which is an annual time period kind of thing. So, getting those numbers worked up. And then, in some cases, it was just getting slides. So, putting together a few slides for something…a slide to show *how* we were collecting this information. … Sharing pictures and things like that. Making sure the pictures were in a usable spot. We had lots of things that we came up with that we should do at various times in the project, but in the end, not all of them got done.

She noted later that one reason "things didn't get done" is that projects take different directions, and not all ideas need to be or are followed through. Thus, a checklist is also a potential means to manage the natural dynamism in the scientific process. As someone who now an assistant professor, she also said it would be helpful for the graduate student working with her to have a checklist. It would help him prioritize and assist her to know more readily where he is in the process.

> Because I just have lots of ideas. There are certain things, and I try to emphasize which are most important and which can be pushed to the side for a little while, but I don't always think that I'm clear. … For instance, I'll know when he gets something done. And maybe I'll know a halfway point, but sometimes I'm curious, "Well, how much working is actually taking place during those in-between times?"

Subject 3, who also used the word "checklist," outlined its value at the start of a project in which data collected by two or more people would eventually be compared stating that it should make explicit, for example, the format the data will be in and the time step. She acknowledged that, "Each project will be different, so the checklist would need to be rather general." Similarly, subject 4 and his collaborators realized six months into their project that they had not all been working with the same "corrected" data. He stated that a data flow diagram – with attributes similar to what others described as a checklist – would have solved this problem.

> We sat down, and we realized we didn't have a data flow diagram. We didn't know who was downloading what and what they were doing with the data, and then how that was getting uploaded back to the Google Drive. So, we spent like half a day of our weekend meeting sorting out this data flow and how things were going to come together. That took a serious amount of effort because we didn't do it ahead of time. … We needed to sort that kind of flow out to make sure that we weren't putting the cart before the horse and weren't running things and actually starting to do some analysis before we actually were sure that those were the numbers we wanted to run.

Subject 1, who is serving as the science coordinator for a collaborative project he is involved with, created a short document to coordinate the team's work: "We're listing things that need to be done, and we kind of modify it as needed. So, we tried to keep it as simple as possible to make it useful." Based on all these comments, the concept of a checklist represents a need for something that can serve as a tool for communication, transfer of knowledge in the service of learning, and a prompt for thinking in advance about what will get shared.

The key findings that emerged from interviewee responses about "working with data" are the dynamic nature of the scientific process and the difficulty of tracking and documenting changes that occur. The latter are not necessary for their own sake, but are important for producing outputs (e.g., knowing which script and dataset versions produced the figure to be published in a scientific paper), coordinating collaborative work, or facilitating communication between students and advisors.

In addition, results from both user studies indicate that the ability to access and use analysis tools within the ACR is desirable. For a number of reasons, however, at this time, the extent to which SEAD can provide support for active analysis of data is uncertain. Comments from users suggest that this disincentive might be reduced if SEAD were to provide storage space that is as easy to obtain and use as Dropbox and has similar capabilities, but offers significantly more capacity.

Another consequence of SEAD's current functionality is that in order to keep the ACR "current" users saw that they must upload files into the ACR each time changes are made. Some interviewees indicated a willingness to do this; while others viewed the need to upload new versions of files as a significant impediment to the use of SEAD. The ability to sync up folders or files on personal hard drives to the SEAD ACR may provide a limited solution to this problem.

4.2.3. Sharing Data

After talking with scientists generally about what it means to work with data, we shifted the focus of the interview to data sharing. We began by asking interviewees what came to mind when they heard this phrase. We then sought specific examples of their experiences in sharing data in collaborative research projects, and we asked questions about topics such as what, when and how they shared.

As might be expected, the more junior scientists, especially the two PhD students, had less experience with collaboration, although they had shared their data with others. Therefore, except where noted, the remainder of this section draws primarily on the four interviewees (subjects 1 and 4-6) who had the most experience working in collaborative projects. They described several types of scenarios in terms of the way that data get shared. They also discussed different types of collaborations and various problems that arise in particular contexts. Depending upon the nature of the collaboration, "working with data" can be similar to working with one's own data, although the challenges might be compounded – or it can include an additional set of tasks and activities. It is the latter that we focus on in this section. While some of difficulties that arise might be managed through version control or annotation tools, others rely on communication and coordination. Thus, we found that interviewees in this study were more interested in SEAD functionality that facilitated the latter than those interviewed by Yew.

Collaborative research

Before talking about how data get shared it is useful to understand the nature of the collaborative work that the scientists engaged in. As we found with other aspects of this study, even a small number of interviews turns up a range of work practices. In this case, the nature of the collaborative projects that one scientist participates in are likely to vary depending on his or her role in the project. For example, subject 4, an assistant professor, currently has what he referred to as three lab-based projects, all of which have collaborators. For two of these, he and his students collect, analyze, and store most of the data, and then he shares analyzed data in the nature of figures that he and his students generate. He stated that the main reason for this is that the data being collected are "in the Catskills – like right in my back yard, so I'm the one that's doing most of the data collection." Another project that he is involved with is using data primarily from multiple external sources. In this case, the data are separated by years, and each investigator is responsible for conducting the quality control on two years of data using criteria that were established by the group in advance. Subject 6, an associate professor, said that in one of her projects "field data are being collected and the QA/QC is happening over multiple locations." The data are then collated by one person. These few examples show the variety of ways in which labor can be divided up amongst members of a collaborative team and how this influences what data are shared and how it occurs.

Most interviewees were working with collaborators who are located at other institutions, although subject 5, a geologist, was just beginning a collaborative project with anthropologists on her campus. Interviewees also noted that they and their collaborators were often working on a particular project at different times (i.e. not simultaneously). Generally, each scientist would determine his or her own work schedule, and then the group would come back together at a future date to talk, compare results, etc. This mode of working required effective communication and coordination mechanisms in order for projects to stay on course.

What is shared?

The data that get shared span the range from raw data to what one person called "synthesized sharing." This is similar to what subject 4 described for his lab-based projects, and consists of PowerPoint slides, data tables, and figures. The formats and size of the data shared are also highly variable. Scientists were more likely to note running into problems with the size of files when sharing raw data, in particular, or certain types of data such as ground or aerial lidar or elevation. In short, any data that scientists collect or data products they produce are candidates for sharing.

How sharing occurs

Interviewees described a variety of approaches for sharing data with members of their collaborative team. The tools and approaches included Dropbox, Google Drive, Google Fusion Tables, and shared server space at a supercomputing center. They also emailed files back and forth and shipped or swapped hard drives with each other. Sometimes data sharing was "planned," which we discuss further below, and at other times it was described as "ad hoc." The latter meant that when someone had a need for data, they sent an email message or phoned the owner of the data. From there, a longer discussion might occur in order to sort out what was needed and to determine the best method for sharing the data.

A couple of scientists used the words "active" and "archive" to describe differences in the nature of the sharing that can occur. Although the lines between the two are not black and white, interviewees generally spoke of "archive" as meaning the sharing of files or folders for others to access, but not necessarily to change or update and then re-post. The major challenges with sharing data in an archive sense are dealing with individual idiosyncrasies that occur at every imaginable level. For example, people have unique schemes for organizing folders and files, and this makes it hard for others to "find stuff." Scientists also use different abbreviations and headings for columns in a data spreadsheet. A quote from subject 5 summarizes what other interviewees noted, too.

> When I look at my data, it's all very easy for me to interpret in a way that makes sense to me. I think the big challenge when you're working with other people is they may have different organizational schemes or shorthand that makes it more difficult to interpret.

These idiosyncratic tendencies are widely recognized by scientists themselves and by those who have studied the data management and sharing practices of scientists. Scientists find them frustrating, but largely accept them as being difficult to avoid without a significant investment of time and effort that distracts "from doing all the things we say we're going to do."

In active sharing, the situation is dynamic because data are in a state of change. Interviewees described the need to keep track of versions of data files being worked on by multiple people or to follow the collection of a model set of values and quality assurance/quality control of data taking place over multiple locations. Subject 4's description of one of his projects is a good example of active sharing and also shows the contrast between this and the sharing of synthesized data.

> I now have a project going with Bonnie and a couple other researchers where we're looking at Lake Hampshire data from the last 5 or 6 years. This

> is where it gets a little tricky.[10] We have this data source that is externally collected, and then we're all kind of quality control/quality analysis going through the data, and then we're generating derived data using models from that data, and then we're sharing that and attempting to all do this at the same time. … The other projects, the ideas come out from the collaborators, and then I and my students collect the data, and then we share kind of end results, figures, whatever. This one has been much more challenging because it's been an active sharing of pretty large datasets of data.

One person also mentioned wanting to monitor analysis and transformations to data done by one or more members of the project. The degree of oversight that scientists wished to exert appeared to depend on their personality and/or on who is conducting the work (e.g., graduate student).

Another type of collaboration that scientists engage in is with students and others in their laboratory. Several of the seven scientists we interviewed raised the possibility of using SEAD to share data within the context of an academic laboratory. Three interviewees, including an assistant professor (subject 4), the postdoc, and one of the PhD students (subject 7) were as interested in SEAD's role in the laboratory context as in other situations. In response to looking at the project space mockup (i.e. figure 5 in Appendix A), Subject 4, who employs undergraduate students, described the benefit of SEAD as follows:

> This is really appealing – less so for the Lake Hampshire data. … I'm thinking it would be more appealing for the stuff that I'm doing with my students in the lab. What's happening in the lab is that there's a lab computer, they're updating stuff there; I do some things on my computer; they do some things on their home computer, but the lab computer is the critical piece here. I've had the lab computer break before, and that's a really scary thing because if we haven't updated it on an external hard drive recently, then it's the one place this stuff is stored. Whereas, the way we've set it up with the Google drive for the Lake Hampshire project it seems to work pretty well that everybody has access and things are going to be reasonably…I don't what the best word is… They're going to be there regardless of what happens. But I like the idea of being able to share stuff with my students and add students. I have a pretty ready flux of students both in and out of the lab just because they're only around for a couple of years… Whereas, with the grad students

[10] The name of his colleague and the lake are pseudonyms.

> or postdocs they're going to be working on it in their offices at home in the lab. With undergrads it's a little more contained, so it's easier for me to do that.

Subject 7, who would be leaving his advisor's laboratory in a few months to begin a postdoc, was planning to put together a system for sharing data and scripts for data analysis before he moved on. He was motivated to do this because when he started his PhD program, his "advisor was new, so everything was new. And I felt like I had to kind of invent everything from scratch: be it scripts to analyze the data…actually, the generation of the data itself, learning how to use particular pieces of software." He wanted to leave something behind that would make it easier for future students. He also planned to use it to "retrospectively deal with the data I've generated through my PhD, and kind of collate it and pull it together."

Planning for sharing

When we asked scientists if they discussed sharing at the start of a collaborative project, almost all interviewees indicated that planning occurred to some degree whether it consisted of a "vague notion" of what's going to be shared or was worked out in more detail.

Two scientists, in particular (i.e. subjects 1 and 4), spoke about planning for sharing at the start of collaborative projects they are currently involved in. In both cases, as they described their experiences, they talked first about the work their team did to clearly define the questions they were going to answer. Once that was done, subject 4 set up a folder on Google Drive from which all team members accessed and shared data. This was attractive to him because "every time you change a folder either online or change it on your computer, it'll sync up the file." Part way through the project, though, this group ran into unanticipated issues, which in retrospect, were a combination of the technology not working out as hoped and different work styles. The problem was both recognized and resolved at a face-to-face meeting.

Subject 1, who is serving as the science coordinator for a collaboration that includes more than six principal investigators from multiple institutions, had not yet set up the structure or process for sharing data. In this case, the data will be collected from multiple external sources as well as gathered as part of project research activities. For this multi-disciplinary group, part of the process of getting to the point where the sharing structure could be determined involved spending time understanding the language of other domains and clearly articulating what is known and unknown about the complex questions they are investigating. Subject 1 organized a cyberseminar series to facilitate these goals and also led the authorship of the short document described at the end of section 4.2.2 that is being used to help coordinate the group's work (e.g., outlining who is responsible for what). The other and related part of the delay in organizing the data sharing structure was due to "trying to think about how we can do this in most organized fashion that is going to be most useful for everyone. It's hard to know what everyone's file structures are like to go look for things." Planning for sharing in a complex, large, distributed, and multidisciplinary project such as this one is obviously a significant challenge.

The postdoc related an experience in which she and another scientist planned to collect data on certain parameters and then compare their findings. Although they determined the parameters in advance, when they got together to share the data, they realized they had done things a little differently. This meant that they had to redefine what they were going to share and specify "the actual details and the nitty gritty." Previous collaboration with other scientists can help mitigate these challenges. For example, Subject 5 described a current project that involves scientists she had worked with as a graduate student. At that time, two of them were at the same institution and another was relatively close by. This previous experience made sharing data in their latest collaboration "relatively easy," – in spite of the fact that they are now located in three different states. "Basically, this is an extension of the project we had already worked on, and so we kind of knew what everybody was going to be doing, and we knew what data we needed."

Unlike other interviewees, the most senior scientist among the interviewees (i.e. subject 6] felt it was difficult to plan for sharing in advance.

> Projects change. Collaborators are added or lost. Some projects are around long enough that the technology changes. For example, in its early days, Dropbox didn't work very well; it's much better now. Protocols are worked out but it changes. We have lidar now, whereas, we only had aerial photographs before. Radio carbon techniques have changed. Underlying software changes.

Other comments she made during the interview indicated that coordination among distributed collaborators in her project was important, however. Although her view seems at odds with the experiences of other interviewees, the challenges she described are similar to what others noted and sometimes confounded the planning they attempted to do.

Managing Data Sharing Challenges

The analysis above shows that scientists employ technological and social means to deal with data sharing challenges. The findings also indicate that communication and coordination across space and time are especially difficult to achieve with the consequence being that problems often do not surface until collaborators meet face-to-face or virtually. As the postdoc said, "Sitting down for an hour or 30 minutes would have saved a lot of time. So, the limitation is this communication thing." Communication is also important in heading off or resolving issues that arise. This may account for positive reactions to the communication capabilities of the mockup interface by two of the interviewees. Subject 5 thought that a "space" for the collaboration, along with mechanisms for communication would be useful.

> It kind of gives you a dashboard or a place to really call home for a collaborative project. It seems to me that it has the real potential to improve the collaborative nature of a project. Whereas, with a lot of collaborations, you remember that you're a collaborative when you get together for your annual meeting, or when somebody has a crazy question or an email that they need to send out to the group. Otherwise, things are relatively independent. Then all of sudden you kind of say, "Oh, yeah, I need to talk to these other people about things or ask questions." I think in that respect, it would be great to be able to really communicate.

In terms of communication, she mentioned the ability to email people from within SEAD or to have "a Facebook or Twitter-type of feed going on..." Subject 1 also thought these functions would be useful, especially if the conversations were archived within SEAD. In addition, he said:

> I guess one other potentially useful attribute of this would be for people to almost blog about the results. Kind of a place where people could say, "I just did a hydrological analysis of these watersheds here's what it's showing."

Subject 2, however, cautioned that while such features might be useful "it's one of many platforms that can sort of do this." In other words, SEAD must bring additional value, which for him would be a service to find data that other people had produced.

Finally, when time allowed and the opportunity arose, we asked interviewees about the kinds of things they would want to know about what other people were doing in the project or what they might want to be able to communicate about their activities. For example, we asked if they would be interested in receiving a message when a collaborator had reached a certain point in processing a particular data file. These were somewhat difficult questions for scientists to answer. One reason for this was people's general uncertainty about what might be monitored or shared in this regard. It was clear, however, that such sharing would need to be voluntary.

4.2.4. Accessing NCED Data in SEAD

In addition to the other aims of this study, we sought to gain an initial understanding of the value NCED scientists might derive from being able to access NCED data through SEAD. In other words, what functionality provided by the ACR might users find valuable and in what ways? As mentioned earlier, three of seven scientists we interviewed had a prior affiliation with NCED. Because of time limitations (i.e. interviews were approximately one hour) it was difficult to explore this question in detail. However, two users were interested enough following their interviews to request an account on SEAD and are contemplating use of SEAD in their new projects. Exploration of the online project materials of these users by the SEAD team identified updates to an NCED data collection that had not been incorporated back into that repository and the use of online Google spreadsheets to organize data sets via parameters. These discoveries added evidence that providing a continuing ability for researchers to update and augment existing collections is valuable and would simplify data access (i.e. because data would not be spread across multiple web sites). They also supported the idea that providing a spreadsheet-style interface to data, in which metadata associated with individual data sets are presented in tabular form and used as a way to discover and access relevant data in large collections, organize them, would mirror best practice (and eliminate the need to keep a separate spreadsheet synchronized with data creation).

4.2.5. Integrating Data

Findings from the first user study showed that the NCED scientists frequently integrated geospatial and/or temporal data. After the first user study, SEAD focused some of its development efforts on providing capabilities for working with geospatial data. Specifically, we demonstrated the ability to ingest and index geospatial information and to use that information to provide map overlays and service endpoints through which data in the ACR can be retrieved. We also developed the capability to perform a faceted search relying on FGDC metadata for published data collections. Based on feedback from scientists, analogous functionalities are planned for temporal data (i.e. ability to index temporal data as it is ingested with functionalities to show data on common timelines/temporal graphs and for temporal data in the ACR to be queried). However, we need to learn more about scientists' practices and needs relative to data integration to further guide development efforts. Two of the scientists who participated in this second user study spoke at some length about their experiences with data integration. Although their projects differed, they described the processes they used and the challenges they faced integrating data obtained from different sources, which come in different formats and are often at different resolutions and time intervals.

Subject 2, a PhD student whose research is focused upon estimates of soil moisture in real-time, relies almost exclusively on external data sources for his research. He described in depth the factors that currently impede his research.

> The limiting variable is data location, data access, and data format. So, I want lidar data, but it's sitting in some bizarre image shapefile, and that doesn't help me because I need it to be in a flat x, y ,z that I can write a code that reads it. That's a pain. And then there's the hassle of what am I going to do to get this format into this box into this code, so I can write this algorithm and then send it on its way. So, that's a limitation. There's a limitation in terms of if I want a given suite of features… We all do this as researchers; we use multiple things in tandem. For me, it's precipitation, soil-moisture, elevation, runoff – all these things – and they all come from somewhere different, and in a different format, and often, they're delineated in, you know, one of them is hourly and one of them is daily, but it's missing this day. The other one is every 15 minutes, but it's missing certain types of timings. So, you wind up writing code to do nothing other than get it altogether, line it all up, take out all the missing pieces, and then you've got a thing that is sort of a nice set of rows that you can work with. It would be nice not to have to do that. It would be nice to find a location where you had that information in a one-stop locale that has the features time-stamped the way you want and formatted in the simplest, flattest sort of format imaginable. That's what limits me. Now, it might be something different in two years. That's what tends to slow me down today.

He acknowledged that once he has processed, structured, and error corrected the data, they might be valuable to others. The codes he runs on his cleaned files, along with a description of the assumptions he made when fixing the data, are also likely to be of value to data reusers. However, in response to viewing the mockups, he said that since he is currently working solo, "it would be a lot of effort and hours to fill this framework. I would just as soon take the data, put it on my computer, and sit here to do what I have to do to publish it."

Whereas Subject 2 is working alone, Subject 4 described a collaboration in which he and his colleagues are pulling together separate data streams that are at very different resolutions and collectively integrating variables such as air and water temperatures, wind speed, humidity, barometric pressure, and dissolved oxygen.[11] They downloaded most of the data from online sources – one in particular – collected by lake buoys. They also gathered data from weather stations, requested data from other scientists, and used their own data. After the data were in hand, Subject 4 loaded them onto Google Drive for others to retrieve, and they divided up responsibilities for correcting "the extremely raw data."

> What we did is each person had one or two years where they went through and quality controlled the data. So that meant they looked for values that were anomalies; put N/As in there for bad data. We had criteria for it. And then they separated it out into the right format for some of our models, and then uploaded that data base to the Google Drive in a separate folder.

After this first pass, they referred to the data as "proofed data." Following this, Subject 4 and one of the other team members used R code to do some manipulations. They realized later – at a face-to-face meeting – that the two of them had diverged off using the same proofed data set, and thus had not been running their separate models with the same data. He explained what happened.

> So, Carrie was going through this iterative process where she was modifying the way the data *looked* as well as the content of the data based off of her QA/QC – her secondary QA/QC. Whereas, I was just pulling the original proofed data and not using the secondary QA/QC-ed data. So, we realized we needed to get on the same page where the stuff that Carrie was doing had to feed back into my model. And so to actually even figure out who was doing what…it took an hour or so and drawing up on the board.

Subject 2 thought the problem arose from trouble Carrie, a Mac user, was having with Google Drive. Everything he did to the data showed up on Google Drive in real time. Whereas, Carrie was doing some work offline, which meant that her work wasn't immediately uploaded. Earlier, we noted that Subject 2 felt that this problem would have been avoided if they had developed a data flow diagram in advance. The

[11] This particular project was also mentioned in sub sections 4.2.2 and 4.2.3 and is known as the Lake Hampshire project.

project continues, and the next step for them is to create a compiled data sheet of all of their derived and collected variables. The next challenges they face are similar to those described by Subject 2.

> The issue is going to be making sure that all the data can be meshed together in one big spreadsheet. So, making sure it's in a form, making sure we're on the same page as to how we collapse data. We're going to use things like precipitation and weather data to say why there are differences in metabolism. So, what we need to do is… So, Beth is collecting some of the weather data, and I'm responsible for some of the other drivers. So, then we're going to have to figure out ways to get data that's not necessarily collected on a 10-minute time scale to mesh with the data that we derive from sensor data to mesh with drivers that are sensor data. I think the next big challenge is going to be merging all those together in a way that's meaningful.

We asked if the derived data would be easy to view once the above steps were completed.

> The way the derived data is going to look is there's to be a column of days spanning from 2007 to 2012. So, each day we're going to have a date column, and the next column over is going to be respiration, and there's going to be a number under respiration. It's a time series of data, so you can look at it reasonably easily just by opening up a file. I do think you need to know a little… Like the column headers. We all know what they mean, but if I sent it to someone else, it might take a little bit of effort to say where that comes from.

They anticipated that the final data set would be of interest to others. So, they are planning to produce a curated data publication to share the data. In part, this was to ensure that they and the data providers would get credit for the work they had done. At the time of the interview, they had not yet determined how to share the curated data.

These two examples illustrate data integration carried out by an individual researcher and in a collaboration. For both projects, there was a significant amount of work required to correct data, resolve it, compile it, etc. Subject 4's project was able to divide up the tasks, but this also brought in the need for coordination.

4.2.6. Evolving User Expectations

We end this section with some additional observations that emerged from our analysis. Specifically, a comparison of findings between the first and second user studies shows how user expectations and inclinations change over time as technologies improve or are more widely adopted. While these factors are not specific to SEAD, they will play a role in the adoption of SEAD. These alterations in attitudes and in the propensity to use new tools emerged in the second user study largely in scientists' reactions to the mockups we presented to them.

First, it is clear from our studies and those of others' that the challenge and inconvenience of managing different logins can be a significant barrier to the use of new systems. Comments from a scientist interviewed as part of the second user study also indicated how users' tolerance for separate logins will continue to lessen as multiple services increasingly become linked through one login.

> There's just so many things to log into these days; if it can be connected to something like email. Most people have gmail accounts, so if it can be connected to something like that, or to current working things, that would really facilitate… rather than having to make my students create a new account and user name. If it's somehow connected to that then that would be a huge incentive for me to adopt something like this.

Third-party websites and applications now allow visitors to sign in using their Google user accounts. Federated Login, based on the OpenID, eliminates the need for users to set up separate login accounts for different web sites.[12] Similar capabilities are available for people through their Facebook account login. These developments will drive user expectations and can serve either as an incentive or barrier to adoption.[13]

Another increasingly available capability that is likely to affect user expectations going forward is the ability to link contents on an individuals' hard drive to an external storage space and to sync internal files and folders with these external

[12] Federated login for Google account users: https://developers.google.com/accounts/docs/OpenID

[13] In response to user feedback, SEAD is planning to implement the ability for individuals to use their external (e.g. Google) account. SEAD is also exploring the option to connect a user's ORCID ID. ORCID provides a persistent digital identifier that distinguishes each researcher from every other researcher. See: http://orcid.org/

places. Although these technologies have not been perfected, over time they will improve, and users will come to rely on and demand these features.

In addition to user expectations, we also observed changes in the interval between the two studies in scientists' general interest in or willingness to use social networking tools. Yew (2012) reported that with the exception of a couple of people, interviewees "were ambivalent to the social features of the system" (p. 10). Whereas, in the most recent study, we found that five out of the seven interviewees explicitly mentioned the general usefulness and ubiquity of social networking tools or remarked positively on these features in SEAD. One person described a key vision for the SEAD project (i.e. a place to facilitate interactions that would enhance community-level data sharing and reuse).

> That's also your critical mass problem. You get enough people to use a service like this, and then it becomes a standard from for "blank." And now this is the standard form for ildar data, and now people write code based upon that standard format … I don't know how you cross that threshold because, at the moment, you're not in a position to say, "This is how you format this data."

In response to viewing the mockups, another scientist said:

> The other thing I was thinking of is in some ways it feels like a social networking system that people **are** more familiar with. … I have friends or colleagues who do a lot of communicating on Twitter and things like that and sharing information that way. And maybe some of those social networkers are people who will really get things going.

Elsewhere, we discussed scientists' comments about the usefulness of email and blog capabilities within SEAD.

Finally, we observed a greater willingness among the interviewees in this second user study to share data more broadly. Five of the seven scientists we spoke with either had already done so or planned to share data they were currently working with. The motivations to do so included altruism and practicality (i.e. Others will request the data, so let's plan to publish them). Whether this is part of a larger trend remains to be seen. However, given NSF's requirement for data management plans, for example, the ability to "publish" data or products of data is likely to increase.

5. Reactions to Mockups

In the second portion of each interview we presented scientists with mockup interfaces to SEAD. As in the first user study, the mockups were employed as a way to generate feedback, gather further information about requirements, and function as a visual prompt for users to talk about their workflows and tasks. Interviewees' reactions to the mockups comprise an important part of the user study findings. In this section, we summarize feedback that was discussed in section 4 and discuss other reactions. We also analyze the limitations of the mockups.

Figures 4-6 in Appendix A represent the mockups that were presented to users during the interviews. Figure 4 is a simple image intended to provide an overview of the core components and functions of SEAD. The other two figures focus on subsets of SEAD functions. Figure 5 is a mockup of the project space. It shows a "landing page" that lists a scientist's projects. The summary page for each project provides various types of data summaries (e.g., maps, recent uploads, number of datasets in particular categories), a list of project members, and links to external data sources. We were also able to show interviewees a version of figure 5 by providing them with a link to an internal SEAD page. Although the content between the paper and web mockups was similar, the web version was easier to view. Figure 6 illustrates an enhanced data page. Some of the capabilities shown here include options for sharing datasets (e.g., generate an invite token via email), versioning (i.e. allow upload with option to mark as correction, version, processed (level), or derived product), integration of data from external sources, and data export. Most interviewee comments were directed to figures 5 and 6.

5.1. Positive Reactions

Feedback that appears in this section reflects positive comments made in direct response to the mockups or suggestions for improvements to ideas presented in the mockups that would make SEAD useful.

- SEAD could be valuable if it helped automate the tracking of what's been done (e.g., changes made to experimental design) or match scripts to data files.
- A backup capability would be valuable.
- A "contained," private space makes broader sharing easier when the time comes.
- Several interviewees responded favorably to the map display. Most noted that a way to filter their data would be needed, though, or the map would be difficult to read.

- Serves as a “home” for a collaborative project and a way to communicate.
- Connect the geographic location of data to a Google map as a way to generate metadata regarding location.
- It would be useful if metadata could be kept with the data.
- It would be nice be able to provide “as much metadata as I **have** rather than be told that I can’t share my data because it doesn’t have enough metadata.”

5.2. Negative Reactions

- The need to upload files each time new versions are created is a disincentive to use.
- What users wish to see in figure 5 may depend on the nature of the project or personal preference. In general, a more customizable and flexible interface was desired by some. Two interviewees commented, in particular, on the lack of utility of the bar chart showing the number and types of files.
- If the list of data shown in the middle of the page is comprised of files, it wouldn’t be useful because scientists often have a lot of files. Folders would be a more appropriate display option.
- One scientist was concerned about security and stated that the system has to be impervious because some data cannot be shared. Since SEAD is a third party system, this raises concerns.

5.3. Limitations of the Mockup

The mockup interfaces served a valuable purpose in both user studies. However, they also have limitations. The drawbacks became more apparent in the second user study because as the prototype matured, it was difficult for the interviewer to effectively convey potential SEAD functionality in the space of an hour long interview. This led to confusion or misconceptions on the part of some interviewees as to the ACR’s role in data discovery and long-term preservation.

Some of the scientists viewed the ACR as both a data discovery tool and long-term archive. While this was positive from the standpoint of finding **other people’s** data, it raised concerns for others about the security and/or ownership of **their** data. Similarly, another scientist said it would take a lot of years before SEAD would contain enough data to be useful to have a bigger role in data sharing. This comment shows that the ACR was perceived primarily as most valuable as a data repository. Some scientists, on the other hand, saw the ACR as a way to promote their science – for example, if other scientists could see their data represented on a map.

In addition, the mockups did not adequately convey the functionality of the ACR in terms of its utility in making data shareable and discoverable. It looked to one scientist, in particular, that SEAD was mainly a place to store and exchange files or to facilitate interaction among people in a project, He said:

> That's good, but it's one of many platforms that can sort of do this. It's the "everybody else's project" part that would be the real leap forward. What is valuable is that someone can look at this map and say, "Ok, somebody else is doing the same kind of work somewhere else. They've already found and done some of the legwork to get their hands on and process some of the data that I also can use, and I'm going to find out about it and use it." **That**, is suddenly interesting.

The perception of SEAD as a file store was also reflected in another interviewees' comment about the number of layers to SEAD versus Dropbox, for example, which has direct links to files and folders.

Now that SEAD has developed a prototype and is in the process of completing the steps required for an initial release to selected groups, it will be helpful to move away from – or, at least, combine – static mockups with interactive pages to help users experience current functionality, gather feedback based on their real use, and assist SEAD developers in figuring out how well things work.

6. User Requirements

Findings from the user study report here suggest the user requirements shown in Table 1. These requirements are not necessary for all users since, as the findings show, scientists' needs differ. The results also show, however, that there are many commonalities in needs that span career stage, domain, type of institution, etc.

Requirements that appeared in Yew (2012) and in the table below are noted with an asterisk. Any new functionality identified through the second user study is indicated by red text. Requirements identified by Yew that are not mentioned here are due largely to the differences between the purposes and interviewees among the two studies (see Appendix B). Many of these requirements and associated functionalities noted by Yew are needed to support requirements listed below, however (e.g., ability to organize data, ability to browse and navigate data, ability to pull in data from external sources).

Requirement	Functionality
Ability to link to other tools and systems used	--Ability to use an existing login (e.g., email) to access SEAD
Ability to link directly to folders or files stored locally and to sync automatically	--ability to sync up folders or files on hard drive to SEAD ACR --ability to update files within SEAD (e.g., add rows to an Excel spreadsheet) --ability for other programs (e.g., R, SAS) to get to the data
*Ability to keep track of the latest version of a file	--be able to upload and share latest version of a dataset --be able to track different versions of a file --be able to display changes committed to the file --be able to revert changes made to a file
Ability to provide links between research objects	--ability to retain links between files, scripts, and outputs
*Ability to describe/annotate files and data	--ability to add tags and description at the file level --ability to add tags and description at the data level (i.e. associate data and metadata) --ability to comment on file organization
Ability to coordinate collaborative work through the project space	--users can create and comment on “to-do” lists or other documents --ability to describe ‘logical/scientific’ data sets (e.g., “our agreed best precipitation estimates”); ability to match data sets to them; ability to specify logical data not yet acquired, etc. --notification system (e.g., ability to notify team members as tasks are completed or when they have a reached a certain stage)
Ability to communicate through the project space	--users can subscribe to a message or comment feed --users can send email or chat messages from the project space --users can link communications to particular objects (e.g., datasets, scripts) --communications are archived
Ability to support data management and sharing within a laboratory context	--download the ACR and SEAD VIVO as a virtual machine or provide cloud hosting --minimal technical support is required
Ability to support data integration	--index temporal data as they are ingested with functionalities to show data on common timelines/temporal graphs --ability to generate simple *x,y* time graphs

	--ability to query temporal data --ability to retain data provenance
Ability to share/publish data	--ability to package data, associated metadata, and other objects together and send via email or generate url link --ability to generate doi --ability to add information about personal data or publications via a map that is findable by other ACR users
Ability to find data produced by others	--ability to locate data on a map --ability to locate publications associated with a particular geographic area or location

Table 1: User Requirements

In summary, there are several themes that emerge in terms of user's overall requirements. Specifically, researchers appear to want the following:

- to share a richer 'data model' than files while recognizing that additional ease-of-use and automation will be needed to enable that capability
- to share data and metadata, without having to manage keeping them in sync; thus, they want the metadata and data to be in one file/package
- version and provenance support through tools to keep track and display that metadata automatically and intuitively so that managing that extra information does not cost more than it is worth.
- to integrate their data and reference data, and integrate their data and collaboration experiences while minimizing the burdens of switching tools or adding lots of info to achieve that integration

Finally, some of the requirements that scientists mentioned exist in the SEAD prototype, but they were not apparent through the mockups. Thus, as a next step, we need to learn if they will be useful to users as designed, or if they need to be revised.

7. Discussion

A needs assessment is driven by decisions that need to be made. This report documents needs and requirements for a subset of the community of sustainability scientists. The next step for SEAD is to use the information reported here along with findings from the first user study and other interactions with stakeholders to determine future steps for development and domain engagement.

In some cases, SEAD may not be the solution to particular needs identified. If SEAD perceives that it does have a role, then the team must determine what needs are priority and establish criteria to assess them. If a need meets particular criteria, then SEAD must consider the possible actions that could be taken to meet it. In addition, in some cases SEAD may be equipped to accomplish this on its own, and at other times, collaborations with other projects or researchers may be the best approach. For example, tools and approaches for helping people coordinate work across distance is an established area of scholarship in computer-supported cooperative work; some members of the SEAD team conduct research in this and related areas.

Results from this study indicate that while perspectives, concepts, and methods may differ across disciplines, many of the data needs and practices are similar. Thus, data management, preservation and access services do not have to be unique or designed from the bottom up for each discipline. Equally important, the findings suggest that when collaborating across disciplines, researchers face challenges that include converting formats, coordinating and communicating, recording activities and provenance, and organizing data and documentation. Some of these problems are different than those found when studying individual users in specific disciplines, and they deserve further attention in future interactions with potential users of SEAD.

Finally, it is clear that science is less and less conducted by the "lone scientist." All four interviewees who are assistant or associate professors, including two who are affiliated with non-research intensive institutions – are involved in collaborations that include at least one or more other scientists. While the PhD students are necessarily focused on conducting their own work, they, too, have been part of collaborative projects. Thus, SEAD users are likely to need tools and spaces for working with their own data and for sharing and managing data as part of studies they are conducting with other scientists.

References

National Science Foundation. (2007). *Sustainable Digital Data Preservation and Access Network Partners (DataNet)*. National Science Foundation, Office of Cyberinfrastructure. Retrieved April 26, 2013 from, http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503141.

Watkins, Ryan, Maurya West Meiers, & Yusra Laila Visser. (2012). *A Guide to Assessing Needs: Essential Tools for Collecting Information, Making Decisions, and Achieving Development Results*. The World Bank, Washington, DC.

Wikipedia. (n.d.). Datanet. Retrieved April 26, 2013 from, http://en.wikipedia.org/wiki/Datanet.

Yew, Jude. (2012). *User Study of the National Center for Earth-surface Dynamics: User and Domain Analysis Report*. University of Michigan, School of Information, Ann Arbor, Michigan, USA.

Acknowledgments

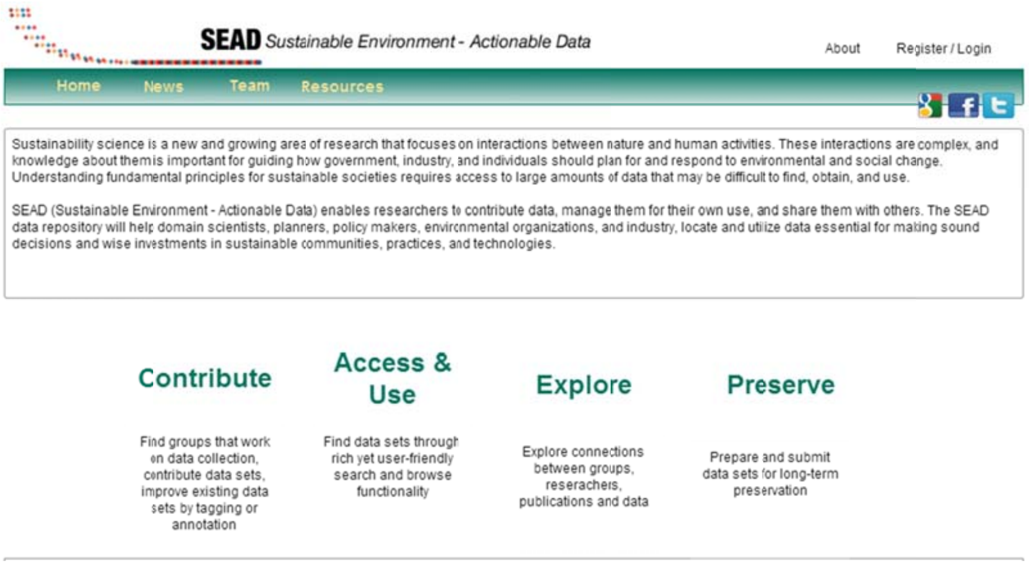
We extend our thanks to the seven scientists who shared their experiences and feedback with us. This work would not have been possible without their participation. We also acknowledge NCED, particularly Charles Nguyen and Efi Foufoula-Georgiou, for their continued collaboration and support.

This investigation was informed by the excellent work of Jude Yew and Md (Zaman) Aktaruzzaman, who conducted the NCED user study under the leadership of Praveen Kumar, SEAD Co-PI and Chair of the Domain Engagement Committee. In addition, this report and study benefited greatly from the advice and input of Margaret Hedstrom and Jim Myers. Finally, we acknowledge Beth Plale and Robert McDonald for their leadership in the design and development of the SEAD Virtual Archive and SEAD VIVO.

The user study was conducted and the report was prepared by Ann Zimmerman, an independent researcher and consultant.

This report is based upon work supported by the National Science Foundation under cooperative agreement #OCI0940824. Any opinions, findings, conclusions, or recommendations expressed in this report are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Appendix A: Mockups Presented to Interviewees



Sustainability science is a new and growing area of research that focuses on interactions between nature and human activities. These interactions are complex, and knowledge about them is important for guiding how government, industry, and individuals should plan for and respond to environmental and social change. Understanding fundamental principles for sustainable societies requires access to large amounts of data that may be difficult to find, obtain, and use.

SEAD (Sustainable Environment - Actionable Data) enables researchers to contribute data, manage them for their own use, and share them with others. The SEAD data repository will help domain scientists, planners, policy makers, environmental organizations, and industry, locate and utilize data essential for making sound decisions and wise investments in sustainable communities, practices, and technologies.

Contribute

Find groups that work on data collection, contribute data sets, improve existing data sets by tagging or annotation

Access & Use

Find data sets through rich yet user-friendly search and browse functionality

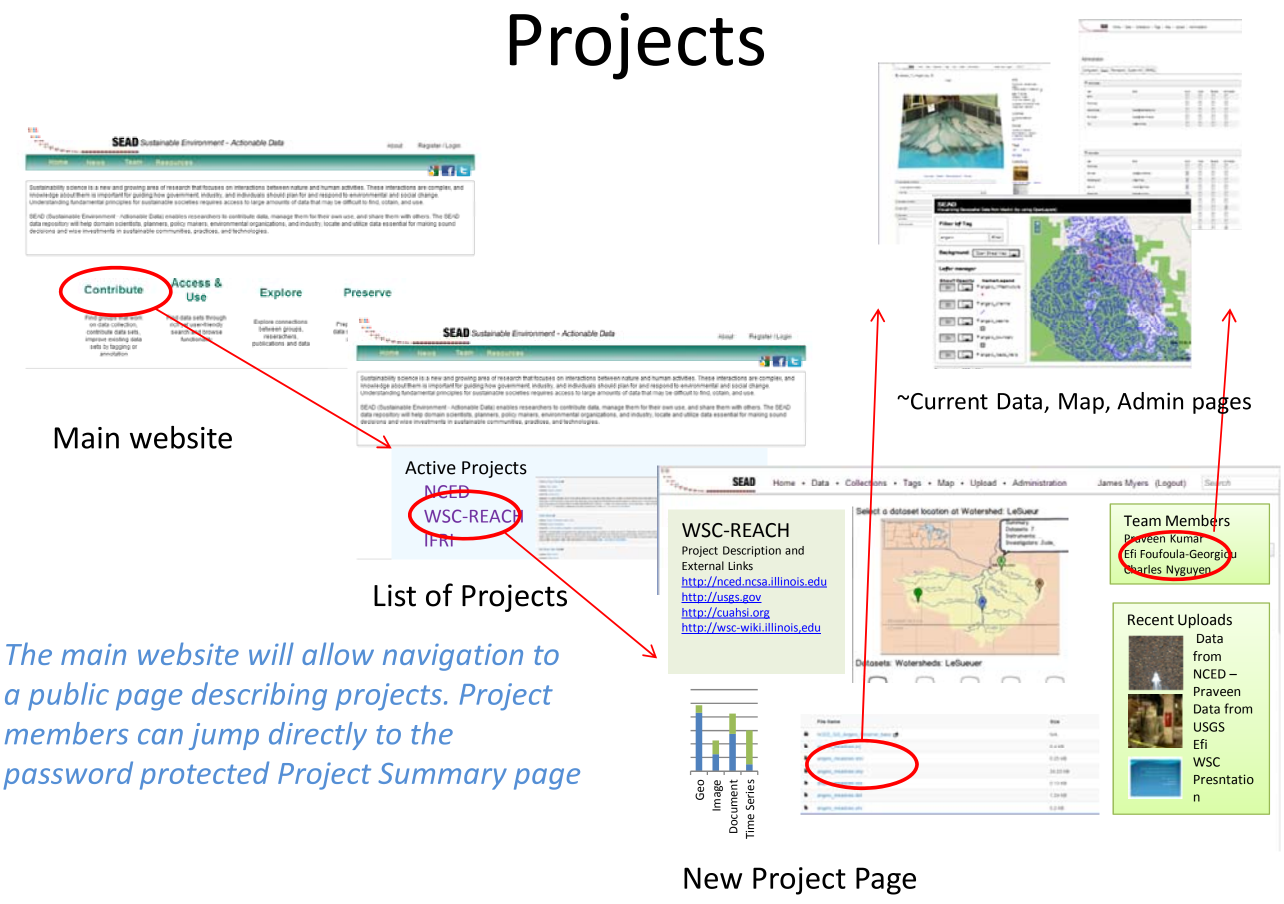
Explore

Explore connections between groups, reserachers, publications and data

Preserve

Prepare and submit data sets for long-term preservation

Figure 4: Mockup of opening SEAD screen


SEAD Sustainable Environment - Actionable Data
Home
News
Team
Resources
Contribute
Access & Use
Explore
Preserve
SEAD
Home • Data • Collections • Tags • Map • Upload • Administration
James Myers (Logout)
Search
Select a dataset location at Watershed: LeSueur
Datasets: Watersheds: LeSueur
Geo
Image
Document
Time Series

Projects

Main website

Active Projects
NCED
WSC-REACH
IFRI

List of Projects

~Current Data, Map, Admin pages

WSC-REACH
Project Description and External Links
http://nced.ncsa.illinois.edu
http://usgs.gov
http://cuahsi.org
http://wsc-wiki.illinois,edu

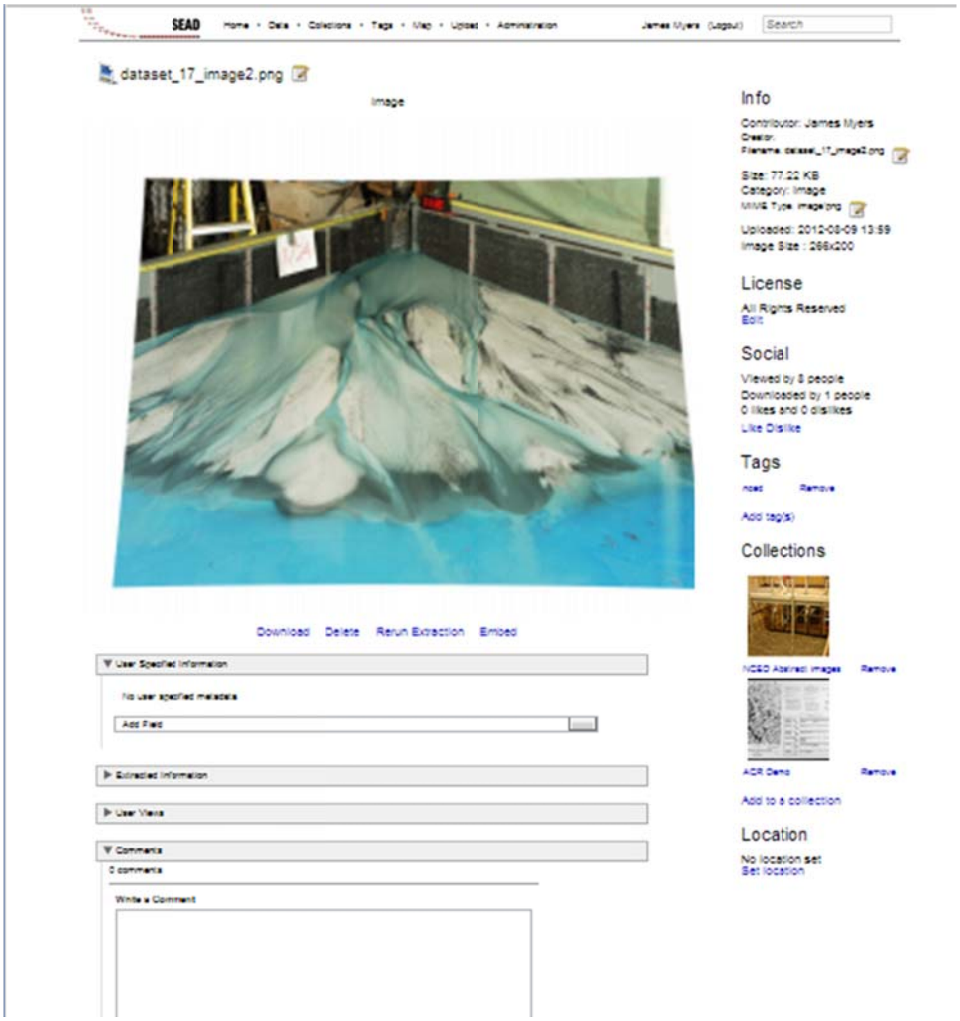
Team Members
Praveen Kumar
Efi Foufoula-Georgiou
Charles Nyguyen

Recent Uploads
Data from NCED – Praveen
Data from USGS Efi
WSC Presntation

New Project Page

The main website will allow navigation to a public page describing projects. Project members can jump directly to the password protected Project Summary page

Figure 5: Mockup of SEAD landing page and project space


SEAD
Home • Data • Collections • Tags • Map • Upload • Administration
James Myers (Logout)
Search
Download Delete Rerun Extraction Embed
Like Dislike
Edit
Remove
Add tag(s)
Add to a collection
Set location
Add Field

dataset_17_image2.png

Image

Info

Contributor: James Myers
Creator:
Filename: dataset_17_image2.png

Size: 77.22 KB
Category: Image
MIME Type: image/png

Uploaded: 2012-08-09 13:59
Image Size : 266x200

License

All Rights Reserved

Social

Viewed by 8 people
Downloaded by 1 people
0 likes and 0 dislikes

Tags

nced Remove

Collections

NCED Abstract images Remove

ACR Demo Remove

Location

No location set

▼ User Specified Information

No user specified metadata

▶ Extracted Information

▶ User Views

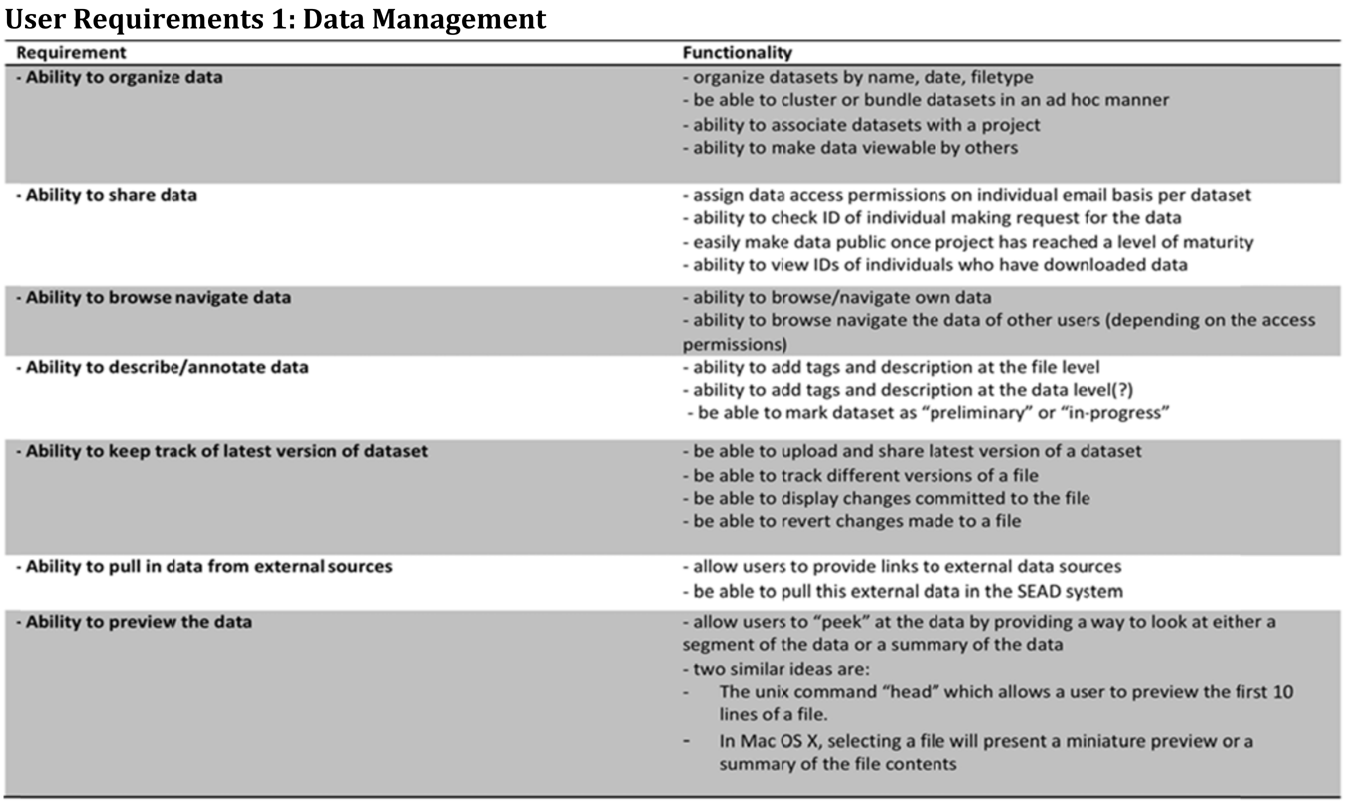
▼ Comments

0 comments

Write a Comment

Figure 6: Mockup of enhanced data page

Appendix B: User Requirements from Yew (2012)



User Requirements 1: Data Management

Requirement	Functionality
- Ability to organize data	- organize datasets by name, date, filetype - be able to cluster or bundle datasets in an ad hoc manner - ability to associate datasets with a project - ability to make data viewable by others
- Ability to share data	- assign data access permissions on individual email basis per dataset - ability to check ID of individual making request for the data - easily make data public once project has reached a level of maturity - ability to view IDs of individuals who have downloaded data
- Ability to browse navigate data	- ability to browse/navigate own data - ability to browse navigate the data of other users (depending on the access permissions)
- Ability to describe/annotate data	- ability to add tags and description at the file level - ability to add tags and description at the data level(?) - be able to mark dataset as "preliminary" or "in-progress"
- Ability to keep track of latest version of dataset	- be able to upload and share latest version of a dataset - be able to track different versions of a file - be able to display changes committed to the file - be able to revert changes made to a file
- Ability to pull in data from external sources	- allow users to provide links to external data sources - be able to pull this external data in the SEAD system
- Ability to preview the data	- allow users to "peek" at the data by providing a way to look at either a segment of the data or a summary of the data - two similar ideas are: - The unix command "head" which allows a user to preview the first 10 lines of a file. - In Mac OS X, selecting a file will present a miniature preview or a summary of the file contents