Data sharing in collaborative scientific efforts: why is it so hard and what can we do to improve it?

Carol J. Volk[1*] and Yasmin Lucero[2]

[1]South Fork Research, Inc.
44842 SE 145th St
North Bend, WA 98045
carol@southforkresearh.org
ph: 206-240-0301
fax: 206-860-3400

[2]Northwest Fisheries Science Center
NOAA-Fisheries
2725 Montlake Blvd. E
Seattle, WA 98112
Yasmin.lucero@noaa.gov

*Corresponding author

Abstract

Increasingly, research and management in natural resource science relies on very large datasets compiled from multiple sources. While it is generally good to have more data, utilizing large, complex datasets has introduced challenges in data-sharing, especially for collaborating researchers in disparate locations. Scientific collaborations are frequently plagued by data handling inefficiencies. To demonstrate this, we surveyed natural resource scientists about common data-sharing problems. The major issues identified by our survey respondents were insufficient metadata documentation and insufficient communication about data availability, quality and management procedures. We advocate alleviating these problems with the use of data dictionaries, data catalogues, read-me files, data flow diagrams, data request forms and organizational charts to facilitate communication. In particular, we argue that distributed research teams magnify data-sharing challenges; if natural resource scientists fail to overcome communication and metadata documentation issues, then negative data-sharing experiences will likely continue to undermine the success of many large-scale, collaborative projects.

Keywords: metadata, data sharing, data flow diagrams, distributed teams, data transfer

Introduction

Technological advances for collecting, storing, and analyzing data have facilitated the collection of more data now than ever before in history—a phenomenon known as data proliferation (Borgman et al. 2007; Quinn and Alexander 2008). The era of data proliferation has brought new opportunities and new challenges in arenas as diverse as marketing, homeland security and molecular biology (Spengler 2000; Shaw et al. 2001; Seifert 2004); natural resource management and monitoring programs are no exception. The needs of natural resource management and monitoring programs frequently dictate large spatial and temporal scales for research (Pikitch et al. 2004), generating extensive and complex datasets. Regional and national scientific data warehouses such as GenBank (www.ncbi.nlm.nih.gov/genbank/) and data catalogs such as Data.Gov and the Knowledge Network for Biocomplexity (http://knb.ecoinformatics.org/index.jsp) have exemplified the benefits and utility of data repositories, yet it remains difficult for individual research programs to prepare data for such repositories or even to share data within their own programs (Michener et al. 2008, Parr and Cummigs 2005). Here, we discuss the challenges that data proliferation bring—particularly problems around data sharing within teams—and suggest tools to help face these challenges.

We can classify the challenges of data proliferation into three categories: (1) Technical challenges: We measure an exceptionally large variety of physical and biological attributes that can be difficult to concurrently store and search (Jones et al. 2006). (2) Training challenges: Traditional natural resource training is focused on ecological principles, minimally preparing us for the sophisticated and rapidly evolving database and software tools essential to our work. (3) Data sharing challenges: Assembling very large data sets almost always requires extensive data sharing and often involves multi-institutional collaborations (Robertson 2008; Schmidt 2009). These challenges can notably inhibit data sharing and project performance.In this paper, we focus on the third set of challenges: challenges with data sharing and the loss of productivity, confusion and miscommunication that are the frequent companions of data sharing.

Monitoring projects are particularly prone to data proliferation issues because they have broad missions which require them to capture a large quantity and high diversity of data types. Small experimental projects generally have a well-defined suite of questions, and thus a clearly defined path between data collection and use, which limits the scope of data collection requirements. In contrast, large-scale monitoring programs often must support a wide range of research questions, many of which cannot be defined at the start of the monitoring program. Monitoring programs prepare for contingencies by collecting a wide diversity of metrics and thoroughly documenting metadata so that data may be repurposed however needed. Additionally, monitoring programs often have a diverse and disparate user community. As a result, monitoring programs frequently require a large staff to conduct data collection, to manage databases and content, and to support data analysis. Often this staff is distributed across locations, i.e. data collection staff may be located at field sites and analysts may be located at outside research institutions. This diverse and dynamic staff face several challenges with data access and handling.

Large scale natural resource projects and monitoring programs typically involve a dispersed set of experts that function as a "distributed team," a group of people who need to work closely but are separated by space, institutional status and/or training (Hinds and Kiesler 2002). Several known problems with the effectiveness of distributed teams are outlined by sociological literature (Cordery and Soo 2008); distributed teams have been associated with information flow failures, poor decision making and overall degraded team performance (Caldwell et al. 2008; Cummings and Kiesler 2007). Distributed data-sharing teams are plagued by the natural pitfalls of a distributed organizational structure and, despite positive intentions and high stakes, data sharing is often subject to inefficiencies with time, money, coordination and effort, as well as lost productivity and accuracy (Wallis et al. 2008; Brunt and Michener 2009; Mager and Pipe 1997).

As scientists working collaboratively with state, federal, and independent agencies for 10 years, our experiences have led us to conclude that these problems are common in natural resources research and that data sharing obstacles may stem in part from the distributed nature of our research teams. To bolster our personal assessments of data sharing inefficiencies, we conducted a survey of scientists actively engaged in large scale data management projects. We use the survey responses to describe data sharing problems. We then discuss a series of tools—developed in the fields of computer science, data management, and people management—that can help diagnose and solve these data sharing issues.

Understanding the challenges with data sharing: the survey

We conducted a survey of natural resource scientists actively engaged in large-scale data management. The survey consisted of 21 questions and included multiple choice, ranking, and fill-in-the-blank questions. The survey was created to identify the relative importance of issues related to data handling within or among programs with similar missions. The multiple choice and ranking questions were used for diagnostic purposes while the fill-in-the-blank and short answer questions were included to draw out examples of data handling scenarios that may be different from our own experiences.

The survey was distributed to about 175 natural resource scientists and data managers. We assumed that the response population represents individuals that are relatively engaged in the management of ecological data as this criterion was part of the solicitation email for survey responses. The initial 175 survey recipients were encouraged to pass the survey to colleagues, a technique called ‘snowball sampling,’ and resulted in 131 respondents. About 47% of these respondents were associated with federal or state government offices, 20% with academic institutions, 13% other organizations (non-governmental offices or consulting firms) and 20% did not identify their affiliation.

To understand our response population, the first part of our survey centered on data-handling responsibilities and educational backgrounds. We presented four distinct roles and responsibilities related to data handling—data collection, content management, data management and data analysis—and asked respondents to report the proportion of their work spent occupying each of these roles (Figure 1). Most respondents reported spending their time in at least two of these roles, although no respondents identified with all four roles.

Our respondents were predominantly educated in ecology and quantitative sciences; 85% and 90% of our survey response group, respectively, reported some formal training in these fields. But even among our survey response population, which is highly engaged in data management, we found limited training in data management despite our response population’s active involvement in data transfer and management. Of the the survey respondents who are data stewards and content mangers, only half reported formal training in database management or computer science, and only 7% of this subgroup have formal training in informatics.

The second part of the survey asked respondents to review their familiarity with vocabulary commonly used among data handlers. We looked for gaps in understanding of vocabulary and key data management concepts, such as "metadata" and "relational database." Of the eleven terms that we presented, all were familiar to most of our survey respondents. While our survey group has limited formal training to prepare them for their current work, they appear to be closing the gap with their own efforts.

Finally, we presented survey respondents with a list of potential obstacles to sharing data. We asked them to indicate how often they experienced these problems and to rank the problems according to frequency of occurence (never, rarely, sometimes, or often). Of the listed survey problems, all were ranked among the top three by at least some survey respondents (Figure 2). Nearly all of these issues were reported as occuring at least sometimes. There was generally good correspondence between how often a problem was reported to occur and how highly it was ranked. Nonetheless, several survey respondents did select the lowest ranked issues to be among their top concerns. This suggests that while these issues may be rare, when they do occur they are very serious problems.

When we examined the highest priority obstacles of receiving data, we found a recurring theme of problems with inadequate metadata, or data about the data. Our survey respondents articulated problems with unexplained missing data, undefined column headers (a.k.a. a lack of a data dictionary), lack of essential protocol information, and data summarization without methodological descriptions:

"...Most common problem is that people haven't provided metadata, so no idea [regarding] the reliability or accuracy of the data..."

"...Receive spatial data without georeference information (e.g., projection, coordinate system, and datum)"

"...Surveys have changed for some reason, but am left to guess..."

"...Oftentimes the data are aggregated or summarized but it isn't clear how..."

"...Hard to know sometimes what the primary data are and what the derived or calculated data are..."

"...Wasted time before realizing data was missing without explanation."

"...Spend a considerable amount of time getting definitive answers to the meaning of zero, blank, and null values."

A recurring theme around metadata issues was that they were often never fully resolved. Some respondents described situations where data were ultimately deemed unusable because of poor documentation. Many respondents reported spending large amounts of time communicating with data providers over email, telephone and in person in order to acquire metadata. This time was frequently described as excessive, wasteful or inefficient. For example,

"...it wastes the time of the data collector when the data can't be used,” or “[I] email back and forth with the providor to get clarity, which wastes time."

Our survey group readily agreed on the primary challenge to providing data: lack of clarity in the data request (Figure 2). However, the written comments failed to elaborate on the nature of this ambiguity; instead, the comments emphasize issues that we associated with inadequate metadata. From this, we concluded that these two concerns are two sides of the same coin: because metadata is not routinely provided with datasets, data receivers cannot make clear, specific requests for data.

Describing metadata: exactly what is it?

Because metadata emerged as a strong theme in the survey research, we take a moment here to expand on the basic definition of the term. Metadata is any information that is needed in order to interpret and utilize data. Ellison et al.

(2006) usefully suggests dividing metadata into two categories: descriptive and process metadata. Descriptive metadata is "data about the structure, content, producer, and location of a dataset." It is the who, what, when, where, and how of data. Process metadata is information about summarization, aggregation or data processing and should be documented as a written document, data processing script, spreadsheet with embedded calculations or in a machine-readable file.

And why is metadata so important? Metadata supports the fundamental scientific goals of repeatability and transparency (Ellison 2009). Furthermore, metadata supports synthesis: a scientist must understand the origin of data in order to understand the significance of data (Ellison et al. 2006); data has little value in the absence of sufficient metadata. In data sharing, metadata is potentially invaluable, as it can provide data transparency that allows others to utilize it freely, without relying on the institutional knowledge of the provider. Therefore, optimal metadata will 'release' the provider from responsibilities associated with interpretation and data requests.

Why do we struggle to share data?

Data sharing is essential to science and increases data value (Parr and Cummings 2005). A variety of sophisticated metadata documentation tools and standards exist—e.g., Ecological Metadata Language (http://www.nceas.ucsb.edu/ecoinfo/tools) and Nongeospatial Metadata Standards (Michener et al. 1997). In fact an entire discipline is emerging, called ecoinformatics, which develops concepts and vocabulary for ecological data management. Note that the informatics discipline works to store and retrieve information across any storage platform, and should not be confused with the more specific database management arena, which focuses on storage and retrieval of information utilizing databases. Despite these progressive movements, there is still a lack of metadata documentation and data-sharing success stories; we suggest that the distributed nature of scientific teams is an overlooked hindrance.

In natural resources, the distributed team organization structure has emerged as projects, especially large-scale monitoring programs, demand a network of experts that are based in many places. Working in distributed teams limits face-to-face communication, which undermines the development of social bonds, social contracts and group identity (Kiesler and Cummings 2002). This is important because strong social ties are the easiest route to coordination, cooperation, and trust development (Nardi and Whittaker 2002). The research literature has unambiguously shown that the more distributed a team is, the more difficult it is to resolve differences of opinion, perspective and expectations (Hinds et al. 2002). When team members are separated by distance, it becomes more difficult to reach common understanding of practices and goals, weakening motivation for attaining shared goals. Madin et al. (2008) suggests that data management practices are still based on the needs of small groups with limited practice of making data widely available to a broader audience. In our experience, the natural resource community has not yet adapted to the reality of data sharing in large, distributed groups. This has created problems for researchers who depend on large-scale data sharing, and it has prevented us from satisfying federal mandates to share data with the public (Ellison et al. 2006). If we are to be effective in this work context, we must adopt strategies that compensate for the limitations of the distributed team environment, rather than continuing to rely only on informal, extensive face-to-face communication.

Documentation tools and practices that can improve team communication

In distributed teams, communication does not occur naturally and subsequently many data-sharing difficulties arise due to inadequate communication both among team members and with collaborators outside of the team. Here, we describe several tools that can help alleviate this problem by formalizing communication . Note, while we advocate the use of advanced metadata documentation tools developed within the ecoinformatics community, such as Kepler and Morphos
(http://knb.ecoinformatics.org/index.jsp), we recognize that some communication challenges can be better resolved with less formal solutions. In this section we describe several simple and long-trusted practices for documentation and communication practices that can help improve communication in distributed teams. We have assembled this group of tools to address the particular communication gaps that we have observed to be common in natural resource monitoring projects. Our goal is to apply these tools in order to increase efficiency in data handling practices.

Tools to improve communication

Tool 1: Define roles and responsibilities: use organizational charts

Although most people understand their own responsibilities within research programs, the roles and responsibilities of others may be ambiguous, especially in large-scale programs and distributed teams. Relationships between supervisor and supervised may be well defined, but relationships among team members are often worked out on an ad hoc basis without a priori guidelines. Taking the time to develop a formal organizational chart for a project can help team members understand their role and define their responsibilities and relationships to others on the team. Clarifying roles provides the infrastructure for data handling (e.g., who do you request data from? who needs data from you?) and can help diagnose areas where communication breakdowns occur (e.g., who is responsible for resolving meta-data issues).

For example, we are familiar with one team of 9 scientists involved in starting a large-scale monitoring program spanning four states. Work for the project was to be distributed among multiple contracts and resources, leading to fears of replicating efforts and/or missing elements within the program. Despite their shared mission, this team was unable to make progress until the scientists spent time identifying roles, responsibilities, and creating an organizational chart. Since then, this distributed group has found that roles must be continuously defined and maintained to maximize efficiency, and their organizational chart is a good tool for articulating these roles.

Tool 2: Define metadata: use data dictionaries, read-me files, protocols and other metadata tools to describe the data

Metadata is a broad topic covering several types of information, including descriptive and process metadata. Although descriptive metadata requirements are well established (e.g. Federal Geographic Data Committee 1999; Ecological Metadata Language (EML) http://www.nceas.ucsb.edu/ecoinfo/tools), these formal requirements are often too formal for typical data transactions among scientists. We do recommend that programs store metadata in formats consistent with long-term program objectives (CITE metadata decision tree). However in practice, a large portion of data transactions among scientists are small-scale and relatively informal. The following descriptions of metadata tools are suggested as informal documentation tools that can be used to improve these commonplace casual data transactions.

A read-me file is a simple text file that should accompany most datasets. The read-me file contains basic descriptive metadata: the who, what, where, when and why of a dataset. This file acts as a 'bread crumb trail' between the current data file and its origins. It should reference written documents that describe data collection protocols or process metadata and include contact information for everyone who has handled the data since its inception. The concept is simple, yet it is difficult to find large, monitoring programs with well-defined descriptive metadata that covers the entire breadth of the program. For example, a program may document the name of the coordinating individual with a dataset, but detailed information about who collected the data in the field or the source GIS datasets may be difficult to track down, despite being key to understanding the origin of the data.

The data dictionary is the most basic piece of descriptive metadata; it is a table or file that succinctly defines the column headers and terms used within a dataset. Data dictionaries may include the calculation, term definition, or possible values of the field. For example, the column header "species" may refer to the latin or common name of a species, and a data dictionary would identify the preferred definition. Survey respondents clearly identified this as an issue: "I've often requested data from other biologists where spreadsheet is not labeled and no protocol is given with the data." Incidentally, we have found data dictionaries useful for tracking column names and formats, which can help both the data provider and receiver pass data in a consistent format.

Protocols are a tool for documenting descriptive metadata. While the data dictionary gives a succinct description of the origin of the data, the protocol provides details. The principle of documenting the data collection process is well understood by natural resource scientists and are commonly referred to as Standard Operation Procedures (SOPs). Data collection methods, contextual information, such as the study design and training requirements, and study objectives can be bundled into a document called a protocol and we suggest following recommendations by Oakley (2003). Additionally field personnel frequently make changes to the methodology to accommodate logistical concerns and changing objectives. These changes are usually irrecoverable unless documented at the time of data collection; there should be a process in place for updating protocols to accommodate this practical reality. The Pacific Northwest Aquatic Monitoring Program (PNAMP) has made considerable progress in providing tools to document protocols and facilitate communication about protocols amongst research scientists through an online tool (www.monitoringmethods.org). By documenting and sharing protocols amongst colleagues, protocol differences are more easily identified and provide opportunities for growth and learning by all.

Process metadata is usually recorded as a written document in the form of a methods section in a scientific paper where data summary and calculation techniques are described. Process metadata may include a data dictionary for derived variables as well as helpful executable files, such as scripts for statistical analysis, database queries, or spreadsheet files with calculations. Sufficient documentation should include a detailed description of calculations, executable files (such as scripts) and any other information necessary to understand how data has changed from its raw format. Although documenting summary and analysis procedures is often second nature to statisticians, we have witnessed many data handlers that are more likely to summarize their summary and analysis procedures post-hoc, rather than keeping a detailed log and file of executables.

Tool 3: Define null values: zero, NA, NaN, placeholders and blank cells

Defining null values within datasets is really a subtopic of defining metadata, but this was such an important recurring theme in our survey that we thought it should be discussed separately. Here, we consider null values to be those that are recorded as the numeral zero, NA, NaN, a numerical placeholder (e.g. 999) or as empty cells within a dataset.

The key thing to recognize is that null values encountered in datasets could be interpreted to have several meanings. They could mean that: (1) a measurement was made and a value of zero was found, but the zeros were not recorded or have been removed during processing; (2) a measurement was made, but the value was not recorded; (3) a measurement was supposed to be made, but was not; (4) a measurement was not made because it was not supposed to be made; or (5) a measurement was not made because it was somehow inappropriate, impossible or undefined (for example, flow rate in a dry creek). These distinctions are very important to interpreting the significance of null values in data and may trigger different procedures for different data handlers—a statistician will change their analysis if the null values are actually zeros, and a field manager may change their collection protocol if they find a large abundance of missing data.

There are no firm rules for handling null values, but there are some conventions. Generally, the numeral zero should be reserved for the case where a measurement was made and the value found to be zero. Additionally, these true zeroes should only be recorded as the numeral zero. In computing, the character strings "NA," "Na" or "na" are to be read as "not applicable." Similarly, the character string "NaN" or "nan" should be read as "not a number." Usually, "NaN" is reserved for derived values where a calculation is undefined (e.g., divides by zero) or impossible because of dependencies (e.g., a non-numerical dependent variate in a sum calculation). In some cases, data storage tools will not accept character strings. In these cases, a numerical placeholder may be used in lieu of "NA," most commonly "999." Above all, empty or blank cells are to be avoided because many computing tools can not store empty cells and may automatically convert them to zero or "NA."

Tool 4: Define the overall team process: using data flow graphs

Generating, processing, and summarizing data using multiple personnel can be similar to the childhood game of 'telephone,' in which a person creates a story and tells a second person, and the second person tells a third person, and the third tells a fourth, and so on, until the last person in the chain tells the story back to the entire group. Typically, each person along the chain makes slight alterations to the story and the ultimate result bears little resemblance to the original. Data are at risk of being processed in a similar fashion. To avoid a game of telephone, it is important for every individual to know what has previously been done to a dataset. We suggest creating data flow graphs to outline the processes (and individuals) a dataset passes through in the course of a project.

A data flow graph is a schematic diagram of the data processing life cycle. The concept comes from computer science where machine process-based data flow graphs document software development (Booch et al. 2005). They may be very detailed blueprint type diagrams—which include complete details of personnel, resources, standards, tools and formats—to generalized conceptual diagrams—which are focused on communicating procedures. In Figure 3, we show a data flow diagram that outlines the life cycle of a dataset within the Integrated Status and Effectiveness Monitoring Program (ISEMP), a program that designs and develops monitoring programs for ESA-listed salmonids (www.nwfsc.noaa.gov/isemp) in Oregon, Washington and Idaho. ISEMP has found these diagrams to be helpful for defining where and when to do quality assessments during the collection process as well as who should be responsible for these assessments.

Although diagramming data flow for a project may initially feel like busywork, we have found that diagramming data flow improves communication (Ludäscher et al. 2009) because data flow graphs can surface divergent project goals or assumptions of individuals involved in a common project. A data flow graph establishes a common vocabulary and project scope, clarifying individual roles and responsibilities and providing a common ground for discussing project needs. These graphs expand to meet the changing needs of a project and are a concrete step towards diagnosing inefficiencies in the data handling and transfer process.

Tool 5: Define the rules of data transfer: make a data catalog and a data request form

Our survey respondents report the process of data transfer to be ill-defined, inefficient and sometimes ineffective. Data collectors noted that data requests are frequently unclear; a request such as "I need all of your data" may require significant refining before it can be executed. There are two elements to improving this process: data requesters need to be better informed about the state of data prior to the request, and data providers need requests to be made in an unambiguous and reasonable format.

To clarify which data is readily available for analysis, we suggest that datasets be catalogued. A data catalog could include a list of variates, their extent in time or space, the existence and location of metadata, availability, and the data quality. From the catalog, a data requester can explicitly describe the data of interest and data givers can accurately assess what information the data provider needs and the effort required to complete the request.

A data request form is a strong complement to a data catalog and may simply be a list of information that should be included within the request itself, such as: 1) a list of requested data fields, 2) the spatial and temporal extent of the data, 3) any summarization or aggregation that is required (or undesired), 4) quality requirements, 5) desired file format, 6) expected timelines, 7) priority, and 8) the purpose of the data request. Finally, frustration can be minimized if expectations for the data transfer process are well defined, such as the how the data will be transferred, who will be informed of the request completion and a timeline for completion.

In practice, it’s imperative both data requester and provider need to identify the information outlined in the previous paragraph; whether this information is solicitied via web request form, an email, or a phone call is irrelevant. The faster all of these aspects of a dataset have been identified, the smoother and faster the data transaction is likely to transpire.

Tool 6: Define expectations: timelines and data update cycles

Establishing formal timelines is a significant tool or strategy for managing expectations on every aspect of the data transfer process, especially within distributed teams. Establishing detailed timelines is an opportunity for each party to communicate all of the anticipated steps of the data transfer process and voice concerns about what may prevent them from meeting their timelines. Taking timelines seriously, by meeting deadlines or clearly communicating modifications to timelines, builds trust and creates stonger working relationships (Armstrong and Cole 2002).

Our survey respondents indicated that defining timelines clearly is usually more important than how long requests take (Figure 2). One survey respondent stated:

"It's better to know if data is or isn't available immediately, or to know that it will be a while in coming, rather than [waiting] and thinking you might 'one day' get what you had requested."

Another respondent communicated the value of clarity for the data giver as well:

"...It would be much easier to get data in a reasonable format if folks had a set date that data needs to be sent in."

Another important aspect of expectation management is communicating the existence of update cycles. When data will be updated for reasons of quality control or additions of data for new regions or time periods, it is important this be made clear in the data transfer process as data changes may influence the time investment that data collectors, managers, programmers and analysts put into a dataset.

Discussion

Data proliferation has brought many new challenges to the natural resource sciences, not least among them the now common practice of working in distributed teams. In this transitional era, data-sharing experiences are too often confusing, difficult and inefficient. In particular, the new conditions of this era have greatly increased the importance of metadata documentation. To achieve more functional research programs within the context of distributed teams, we must overcome the limitations of current metadata documentation practices. The communication and documentation practices we’ve suggested for general use, within teams and for data sharing outside of the core team are tools that can help negotiate the burden of data sharing that is overwhelming in an era of data proliferation.

We acknowledge that there are many obstacles to good data sharing practices. Several survey respondents pointed to 'elephants in the room'—such as lack of resources, agency politics, and team interpersonal dynamics—that are the 'true' obstacles to success. While these problems are real and present, it is not necessary to use them as excuses to neglect team organization. In this paper, we suggest tools that may be utilized to prevent, diagnose, or solve inefficiencies in data handling. Many of the problems that we discuss here are simply easier and cheaper to solve than these larger systemic issues. And we truly believe that adopting more formal strategies for communication and data sharing in distributed teams can improve the data-sharing experiences, resulting in less stress and inefficiency for everyone. However, there must be a balance between adopting formal data sharing and communication practices and the utility of these practices for increasing data handling efficiencies. Negotiating this balance depends on the team and project goals—if goals are met, improvements and additional investments are irrelevant. The key is having a clear project or program definition of success and efficiency, and from there, inefficiencies can be diagnosed, a solution implemented, and the system improved. We hope the suggested tools are good stepping stones for individuals or programs to untangle, understand and diagnose inefficiencies one step at a time.

Acknowledgements
This work was supported by the Integrated Status and Effectiveness Monitoring Program (funded by Bonneville Power Administration (2003-017-00) and the National Science Foundation (XXXX)

References

Booch G, Rumbaugh J, Jacobson I. 2005. Unified Modeling Language User Guide, The Addison-Wesley Object Technology Series. Addison-Wesley Professional.

Borgman C, Wallis J, Enyedy N. 2007. Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries. International Journal on Digital Libraries 7(1):17–30.

Brunt J, Michener W. 2009. The resource discovery initiative for field stations: enhancing data management at North American biological field stations. BioScience 59(6):482–487.

Caldwell B, Palmer R, Cuevas H. 2008. Information alignment and task coordination in organizations: an'information clutch' metaphor. Information Systems Management 25(1):33–44.

Cordery J, Soo C. 2008. Overcoming impediments to virtual team effectiveness. Human Factors in Ergonomics and Manufacturing 18(5):487–500.

Cummings J, Kiesler S. 2007. Coordination costs and project outcomes in multi-university collaborations. Research Policy 36(10):1620–1634.

Ellison A. 2009. Repeatability and transparency in ecological research. Ecology, *In Press.*

Ellison A, et al. 2006. Analytic webs support the synthesis of ecological data sets. Ecology 87(6):1345–1358.

Federal Geographic Data Committee. 1999. Content Standard for Digital Geospatial Data, Part 1, Biological Data Profile. Federal Geographic Data Committee and USGS Biological Resources Division. Report no. FGDC-STD-001.1-1999

Hinds P, Kiesler S. 2002. Distributed work. The MIT Press.

Jones M, Schildhauer M, Reichman O, Bowers S. 2006. The new bioinformatics: integrating ecological data from the gene to the biosphere. Annual Review of Ecology, Evolution and Systematics 37: 519-544.

Kiesler S, Cummings J. 2002. What do we know about proximity and distance in work groups? Pages 57-80 in Hinds PJ and Kiesler S, eds. Distributed Work. MIT press.

Ludäscher B et al. 2009. Scientific Process Automation and Workflow Management. Chapter 13 in A Shoshani and Rotem, Eds. Scientific Data Management: Challenges, Existing Technology, and Deployment, Computational Science Series. Chapman & Hall/CRC.

Madin J, Bowers S, Schildhauer M, Jones M. 2008. Advancing ecological research with ontologies. Trends in Ecology and Evolution 23:159-168.

McLaughlin RL, Carl LM, Middel T, Ross M, Noakes DLG, Hayes DB, Baylis JR. 2001. Potentials and pitfalls of integrating data from diverse sources: lessons from a historical database for Great Lakes stream fishes. Fisheries 26: 14-23.

Michener WK, Brunt JW, Helly JJ, Kirchner TB, Stafford SG. 1997. Nongeospatial metadata for the ecological sciences. Ecological Applications 7:330-342.

Nardi BA, Whittaker S. 2002. The place of face-to-face communication in distributed work. Pages 83–112 in Hinds PJ and Kiesler S, eds. Distributed Work. MIT press.

Parr CS, Cummings MP. 2005. Data sharing in ecology and evolution. Trends in Ecology and Evolution 20:362-363.

Pikitch E, et al. 2004. Ecology: ecosystem-based fishery management. Science 305(5682):346.

Quinn M, Alexander S. 2008. Information technology and the protection of biodiversity in protected areas. Pages 62-84 in Hanna KS, Clark DA, Slowcombe S, eds. Transforming Parks and Protected Areas: Policy and Governance in a Changing World. Routledge.

Robertson G. 2008. Long-term ecological research: re-inventing network science. Frontiers in Ecology and the Environment 6(5):281–281.

Schmidt B. 2009. Considerations for regional data collection, sharing and exchange. StreamNet. Pp. 27.

Seifert J. 2004. Data mining and the search for security: challenges for connecting the dots and databases. Government Information Quarterly 21(4):461–480.

Shaw M, Subramaniam C, Tan G, Welge M. 2001. Knowledge management and data mining for marketing. Decision Support Systems 31(1):127–137.

Spengler S. 2000. Bioinformatics in the information age. Science 287(5456):1221–1223.

Wallis J, Mayernik M, Pepe A, Borgman C. 2008. An Exploration of the Life Cycle of eScience Collaboratory Data. Center for Embedded Network Sensing, Pp. 2232-2238.

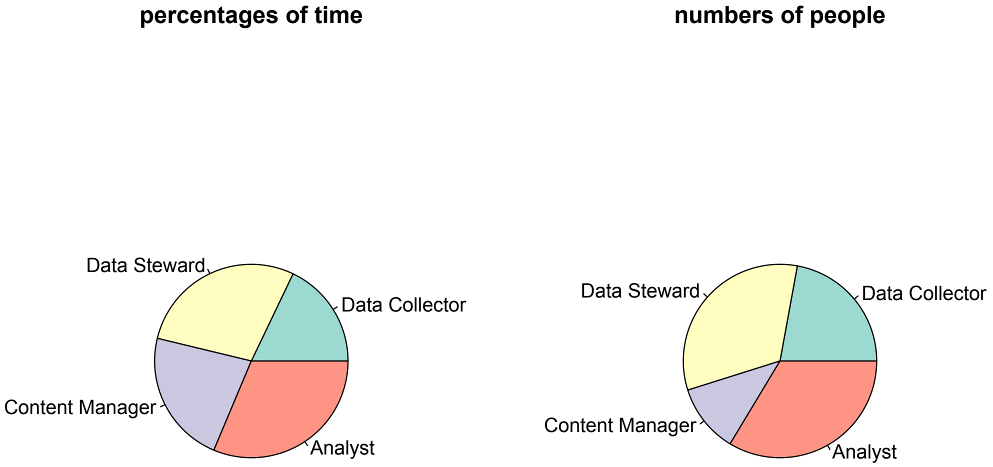
Weisbord, M.R. 1978. Organizational Diagnosis: A Workbook of Theory and Practice. Perseus Books Group.

Figure Legends

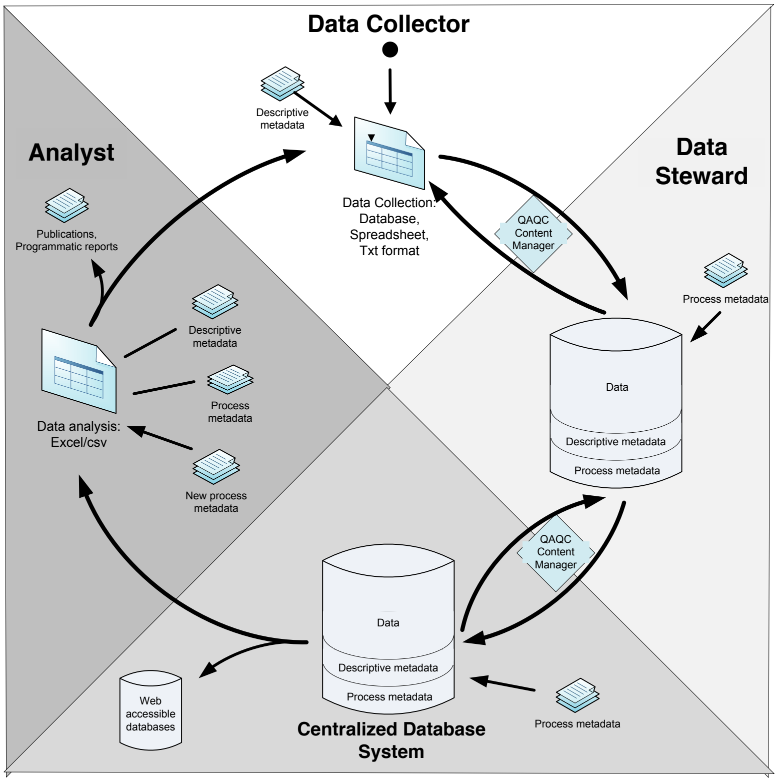
Figure 1. Composition of survey response population. Survey respondents were asked to report what proportion of their work was spent in each of four roles. For this figure, each respondent is categorized into a single role based on where they report spending the largest portion of their time. The numbers in the pie chart indicate the number of respondents categorized into each role, out of 131 total respondents.

Figure 2. List of obstacles to data sharing as ranked by our survey respondents. The rankings are determined by the number of people to identify the problem as among their top three concerns of those listed. Survey respondents were given an option of "other" and a space to list alternate concerns, but only a few used this option. The text shown here to describe the problems is the same text that was provided in the survey.

Figure 3. Data flow graph describing the life cycle of data collected in the Integrated Status and Effectiveness Monitoring Program. Data originate with data collectors, who are the first to put raw data and metadata into a digital format. Data are then shared with the data stewards, who are responsible for validating that data are in a structure that conforms to ISEMP data quality standards. The data are then shared with a content manager, who reviews data and metadata in accordance with program content standards. After review, data are exported to a centralized database from which data are exported to analysts, who develop monitoring metrics and other calculations that may inform decision-making or study designs.



	rank	the problem	number of respondents to rank this problem among their top 3
Problems with giving data	1	lack of clarity in data request	75
	2	format of data request not specified	39
	3	unreasonable timeline	39
	4	unreasonable expectations	32
	5	lack of credit share after data was provided	21
	6	data not used after request fulfilled	20
	7	no timeline associated with request	19
	8	data request changes frequently	18
	9	data receiver doesn't respond to emails	11
	10	lack of job performance credit	8
	11	data request is misdirected (went to wrong person)	7
	12	data receiver dissatisfaction not communicated	6
	13	too many data requests due to misdirected requests	5
Problems with receiving data	1	no data collection description/no protocol	52
	2	data aggregated or summarized without explanation	36
	3	failure to define zeros, empty cells, NA, or placeholders	32
	4	column headers are absent or undefined	29
	5	missing data without explanation (incomplete dataset)	29
	6	data manager is overextended and unavailable when needed	23
	7	protocols exist but are difficult to locate	18
	8	unreliable timelines for when data will be available	17
	9	data file in inaccessible format	15
	10	data inaccessible (e.g. in wrong format for your use)	14
	11	data received does not correspond with data request	11
	12	data is delivered in pieces because of data scope issues	9
	13	data manager doesn't respond to emails	9
	14	data incorrectly labeled/not what it is claimed to be	7


Data Collector
Descriptive metadata
Analyst
Data Steward
Publications, Programmatic reports
Data Collection: Database, Spreadsheet, Txt format
QAQC Content Manager
Process metadata
Descriptive metadata
Process metadata
Data analysis: Excel/csv
Data
Descriptive metadata
Process metadata
New process metadata
QAQC Content Manager
Data
Descriptive metadata
Process metadata
Web accessible databases
Process metadata
Centralized Database System