Report on Evaluation of DataONE Data Management Short Course (May 23-24, 2012) and CEE Data Management Education Lessons

Prepared by the DataONE Community Engagement and Education Working Group
Finalized October 3, 2012

Report drafted by Stacy Rebich Hespanha and reviewed by Vivian Hutchison and Heather Henkel
# Table of Contents

- **Table of Contents** .................................................................................................................. 2
- **Acknowledgements** .................................................................................................................. 2
- **Overview** ................................................................................................................................. 3
  - Brief summary of findings ............................................................................................................. 3
  - Brief summary of recommendations ............................................................................................ 4
- **Participants** .............................................................................................................................. 5
  - Research interests and experience ............................................................................................... 5
  - Expectations for Data Management Short Course ........................................................................ 7
  - Data Management and Technical Experience and Expertise ...................................................... 9
- **Relationships between Perceived Relevance of Course Curriculum and Participants’ Data** Management and Technical Experience and Expertise ................................................................. 11
  - Relationships between Ratings for Course Quality and Relevance ........................................ 12
  - Relationships between Pre-Course Data Management Practices and Perceptions of Quality and Relevance for Individual Lessons .................................................................................. 12
- **Evaluations of Course Materials** ............................................................................................... 14
  - Quantitative Results .................................................................................................................. 14
  - Discussion of Evaluations and Suggestions .................................................................................. 21
  - Reflection on Expectations ......................................................................................................... 35
- **Pre- and post-course self-reported data management practices and plans** .......................... 36
- **Obstacles to implementation of recommended Best Practices** ............................................. 49
- **Reflection and Future Plans** .................................................................................................... 70

---

# Acknowledgements

Special thanks to Rebecca Koskela, Amber Budden, Mark Schildhauer and CEE working group members Cliff Duke, Wendy Gram, Stephanie Hampton, Heather Henkel, Carly Strasser, Kristin Vanderbilt and Lynda Wayne, who provided feedback on draft versions of the evaluation surveys. We would also like to thank Nick Brand, who set up and maintains the NCEAS LimeSurvey® (http://www.limesurvey.org/) server used to administer the surveys, and to the other NCEAS staff who contributed to hosting the Short Course.
Overview
A two-day Data Management Short Course was held at NCEAS in Santa Barbara on May 23-24, 2012, led by members of the DataONE Community Education and Engagement (CEE) Working Group in cooperation with Earth Science Information Partnerships (ESIP). Vivian Hutchison (USGS), DataONE co-Lead of Community Engagement and Education was the lead workshop instructor and organizer, and Amber Budden, DataONE Director of Community Engagement and Outreach led recruitment of participants. Heather Henkel (USGS), Nancy Hoebelheinrich (NSIDC), Mark Schildhauer (NCEAS), and Carly Strasser (UCOP) presented lessons during the workshop. Three Masters students, 10 PhD students, and 2 postdoctoral researchers attended the workshop.

A primary goal of the workshop was to field test and gather feedback on the Data Management Education Lessons (http://www.dataone.org/education-modules) that have been developed by the CEE working group. The lessons are designed to target graduate students in the earth sciences; they are downloadable from the DataONE website singularly or in their entirety for teaching purposes in the classroom.

For the workshop, Stacy Rebich Hespanha developed a series of online surveys to collect information about participants and workshop feedback. Participants completed a pre-course survey and four surveys during the workshop (one at the end of each morning and each afternoon session). This document summarizes both the information gathered through the surveys and notes taken by the instructors during the workshop.

Brief summary of findings
• On average, participants found the short course information content and presentation of that content to be of high quality (‘very good’ – 5 on a scale ranging from ‘very poor’ (1) to ‘exceptional’ (7)) and approximately 80% of the content to be relevant to them personally. Ratings for quality and relevance of hands-on activities were somewhat lower (‘good’ – 4 out of 7, 70% personally relevant).
  o Short course participants already engaging in more advanced data management and sharing practices assigned significantly lower ratings to the quality and relevance of the information presented in the course than those less experienced in data management and sharing.
  o Evaluations for information quality, presentation delivery, and information relevance are positively correlated, so ratings for any individual response measure should be considered in the context of the other measures.
• Prior to attending the workshop, participants rated their own data sharing practices (or plans) as being higher than the norms for their fields. This, together with the fact that they were interested enough in learning about data management to attend a two-day course, suggests that this group of participants is not necessarily representative of the target audience for the lessons;
specifically, these participants could be expected to be more aware of/less skeptical about the importance and benefits of data management and sharing.

- After attending the workshop, participants perceive significant barriers to adopting the best practices outlined in the lessons. Lack of time was cited as the most important factor, followed by lack of training, lack of tech support, and lack of funding. Participants also perceived lack of agreed-upon metadata standards and lack of institutional support and standardized practices for data sharing and citation as important barriers that were beyond their control.

Brief summary of recommendations

- When recruiting participants for a workshop, include a pre-course survey as part of the application process. Selecting participants based upon self-reported data management and sharing practices could achieve a better fit between participants and course content.

- A decision should be made about whether the CEE Data Management Lessons are intended to be presented as a live workshop (in which case redundancy needs to be trimmed) or to function as stand-alone lessons on specific data management topics. Based on comments from participants, a single set of presentation slides will not likely be appropriate for use in both of these situations.

- Two versions of the information presented in each lesson should be prepared: one set of slides to be used for presentation that are less text-heavy than the current version, and a white paper or another set of slides with more complete text intended for reading and reference.

- Each lesson should contain multiple real-world examples or stories that illustrate the main points of the lesson. Participants reacted very positively when information was presented in this way, and suggestions to include more information of this type were more common than any other suggestion.

- Each lesson should contain information about specific tools and resources that support best practices. Inclusion of more of this type of information was the second-most common suggestion.

- In general, lack of time was seen as the most important barrier to implementation of suggested best practices. Lessons should focus on strategies and tools that minimize the amount of time required to implement good data management practices.

- ‘Jargon’ such as “metadata”, “data management”, “discovery”, and other common data management terms (see section titled ‘General feedback on lessons’ for additional terms) should be clearly defined the first time these terms are mentioned. The ‘Intro to Metadata’ and ‘Data Semantics’ lessons should be presented near the beginning of the course because they contain many such key term definitions.

- When lessons are used as the curriculum for a short course or workshop, instructors should integrate open-ended discussions and hands-on activities (e.g., critiques of data management elements such as data management plans and
metadata records, using recommended tools with own data or supplied datasets) with the presentation components.

- When advertising and recruiting participants, make sure to provide information (such as amount of time they will be expected to spend filling out feedback surveys) to help them set appropriate expectations for the evaluation aspects of the workshop or short course.

**Participants**
Fifteen early-career researchers participated in the workshop, including 3 Master’s students, 10 PhD students, and 2 postdocs.

**Research interests and experience**
Participants had a broad range of academic backgrounds and interests. Table 1 includes a brief description of each participant’s research experience and interests as submitted for the pre-course survey.

**Table 1: Participant research interests and experience**

<table>
<thead>
<tr>
<th>Responses to “Briefly describe your research. How long have you been conducting this type of research?”</th>
</tr>
</thead>
<tbody>
<tr>
<td>I am a student of information science. I do not conduct research currently. I used to be a civil engineer and did not do any research in that capacity either, but I did use data. I will answer questions as though I were still a practicing civil engineer.</td>
</tr>
<tr>
<td>I will be starting a Master’s program in Geography and Environment at Boston University in September. I will be working with hyperspectral satellite imagery, and my research will (likely) focus on applying techniques from computer science to analyze the data and address questions related to global change. Currently I work as a software engineer for NASA on the World Wind project. World Wind is an open source toolkit for 3D geovisualization. I have been building geospatial software for about four years.</td>
</tr>
<tr>
<td>Briefly, I study how land use impacts stream ecosystems. Specifically, I am interested in urbanization. I am currently exploring a series of water temperature time series data sets that stretch across about 8 sampling locations and that vary between 2 and 10 years in length. All of the sites have data gaps of varying length. I would like to collect additional temperature data and supplement it with high frequency dissolved oxygen measurements in order to assess temporal and spatial patterns of variability. Ideally, I would also collect the data appropriate to use the DO measurements in order to calculate stream metabolisms for streams of different land uses. I have only recently begun this research and I know that it will require an extensive amount of data and data analysis. I expect that over the course of the next few years I will also be working with remote sensing data and GIS layers, possibly also exploring spatial patterns of urbanization in relation to a variety of stream ecosystem measurements.</td>
</tr>
<tr>
<td>I am broadly interested in the factors that support biodiversity and in how biodiversity influences ecosystem functioning. More specifically, I am investigating diversity at multiple</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
trophic levels simultaneously and interactions with different physical regimes (e.g. water flow) and spatial scales. I have just begun this line of research in the past year.

I’m interested on the effect that the landscape has in ecological processes relevant for agriculture. For my PhD dissertation I chose to focus on pollination and I finished 1 season of field work in which I collected data regarding pollination communities associated to blueberry fields. Overall, I have been doing this research for about 1 year.

My research focuses on improving mechanistic understanding of ecohydrologic processes in the mountainous Western U.S. I have been conducting this type of research for approximately three years.

I have been studying sensors and sensor networks specific to river environments since 2006.

My current research aims to improve methods for using airborne imagery to detect, map, and quantify greenhouse gas emissions, including methane, carbon dioxide, and nitrous oxide. I have been working on this project for the last three years. I primarily work with geospatial datasets, including remote sensing imagery and mapping data, but also use thousands of ASCII files generated from the radiative transfer model MODTRAN to estimate gas concentrations.

My master’s thesis was on the ecology and management of oak woodlands at Tejon Ranch, CA. We assessed the ecological condition of oak stands, performed a historical photo analysis to determine if the oak population was increasing or decreasing and we made predictions about how the oaks would be impacted by climate change. I also have a lot of research experience related to wildlife corridors and landscape connectivity. Lastly, I’ve become involved in a landscape genetics research project where we are varying components of a landscape resistance surface (noise, range, etc.) to determine the sensitivity of such landscape elements to overall gene flow in a simulated environment.

I study the ecology of African malaria vectors (mosquitoes). Specifically, environmental controls over larval development and survivorship in their aquatic habitats.

As automated environmental sensors become more readily available to ecologists, they are more likely to be incorporated into research plans. However, what will happen to the data that are collected? I am investigating the relationship between ecology lab size and networking links with effectiveness of sharing research results. I am also working on cyberinfrastructure components to help small ecology labs using sensor networks to manage their data.

Bridging the spatial and temporal relationships between hydrologic, ecologic, and biogeochemical processes in mountain environments through the integration of in-situ and remote sensing observations. I have been doing this type of research for more than 6 years.

My research focuses on studying how root mucilages and microbial exudates aided by repeated wetting-drying cycles lead to the formation of soil aggregates. This research seeks to elucidate the processes involved in the formation of hierarchical soil structure that spans micron to centimeter length scales. Almost I finished my third year.

Monitoring of snow, soil hydrology and the response of plants to these environmental inputs. I have been doing this type of research for 3 years with also some related research prior to this current project.

I am an early stage postdoctoral researcher working in the department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign. I have assumed this research position on 24th of January, 2012 and have been entitled to carry out my research in
the direction of data management and longtime preservation of scientific data. The project I work for is titled ‘Sustainable Environment-Actionable Data (SEAD)’. This project envisions developing a cyberinfrastructure for storing, sharing, analyzing and curating environmental and social data with the intention to supporting sustainability researchers.

Expectations for Data Management Short Course
Participant expectations for the short course varied somewhat (see Table 2 for complete summary of responses), but several common themes emerged. Eleven of the 15 participants mentioned wanting to learn about better ways/best practices for managing data; six mentioned acquiring information about data management resources or tools; four mentioned interest in creating data management plans; three mentioned learning about how to create quality metadata; and three expressed interest in learning about how to better document workflows and/or data provenance. Five of the participants reported interest in the short course to help them prepare to teach others about data management.

Table 2: Participant expectations for short course

| Responses to "What are your expectations for the Data Management Short Course? What knowledge or skills are you hoping to gain by participating in the course? (e.g., Are you expecting to learn about ways that you can better manage your own data? Are you attending the short course to learn more about how to teach about data management?)"
| My formal education in data management is limited at best. To be honest, I was not 100% sure of what I might get out of this course, but I have seen the perils of poor data management. Our group runs projects in several developing countries and this poses all kinds of challenges in data management, error correction, etc... I am hoping to get new ideas for how to better plan and execute data management protocols. I DO know that data management is a critical issue, so the more know how the better, especially in my field.
| I would like to learn how to spend less time on data management (including any possible automation processes?) and more time on research. I might be interested in teaching data management in the future.
| How I can create CLEAR metadata
- How to make my data set early to be read by anyone
- Learn (for my self) and disseminate the information to my colleagues
| I hope this course will provide me with a greater understanding of the data life cycle and help me begin thinking about developing robust metadata strategies for my own research. A greater understanding of data management planning would benefit me after graduate school given I will continue working with large data sets and might join an organization with an existing data management infrastructure. I am also interested in this course given recent changes by the National Science Foundation requiring that future proposals include a well-defined data management plan.
As a relatively new PhD student who plans to participate in the modeling community for my career, I am looking for best data management practices.

I would like to understand how to enter data and use access. I am starting my dissertation and would like to be able to design a workflow and other data management planning so I start off on the right track. I would also like to learn about common techniques for good data management.

I really want to learn ways to manage my own data and to learn how to start this process from the ground up. I imagine that I will be involved in collaborations with many partners in the future and data management issues should not be a problem. Rather, I would hope that they will make everyone’s life easier.

I am specifically interested in learning how to create metadata in a formal way, how to store and share data effectively, and how to make sure that I do not lose data or the history of what I’ve done with the data.

My goal is to become more efficient and productive in my research and collaborations with other researchers. I hope to learn innovative methods for organizing, managing, and archiving my data and be introduced to some tools that I can easily use to do this.

I hope to learn about effective metadata, which I have some apprehension with as every researcher develops their own convention. I also hope to learn if there are methods to easily change workflow and directory structure due to additional layers of analysis- for instance when model analysis level is added, etc.

Both I guess. I certainly would like to learn about skills and tools for doing a better job with data curation. I also often supervise undergraduates working with data and would like to learn better ways to help them work with data.

I want to learn skills relating to systematic data management plan.

I’m hoping to learn best practices of managing scientific data. In my current work I am more involved in creating tools for data processing than involved in managing the data itself, but I expect this to change when I start school this fall. For example, I’ve never created a formal data management plan, but this seems like a useful skill.

I’m also interested in hearing about the specific challenges facing users of scientific data sets. I am often involved in the creation of data processing tools, and understanding the challenges facing users helps me build better tools.

I hope to learn detailed practices on data management so I can be a resource to researchers/students of my institution in the future. I hope to teach it one day.

I expect that this short course will provide me with an overview of all the important elements of successful data management. I also expect to learn some basic data management skills, gain some new insight into data management resources, and become aware of better standard data management practices. I expect to learn about ways that I can better manage my own data and I expect to learn more about the advantages of better data management practices.

I expect to learn methods to better manage my data and be able to store it appropriately and share it easily with other researches. In the future I might teach other people on this, especially people working with me.
Data Management and Technical Experience and Expertise

As part of the pre-course survey, participants also provided information about their experience working with metadata, different types of data, software and scripting languages, and their current data management practices or plans (related to data sharing, data management planning, data quality control/quality assurance, metadata creation and maintenance, analysis and workflows, data citation, and backups and protection). In an effort to distinguish more advanced participants from those new to data management practices, numeric ‘expertise’ score was calculated for each participant based on their responses to these survey items. Figure 1 shows mean-referenced scores on 4 components (pre-course data management practices or plans, experience with metadata, experience with a variety of data types, and software and scripting expertise) for each participant.

Overall, Figure 1 illustrates substantial variation in participant backgrounds with respect to the four measured areas of expertise. This relatively wide range in experience levels of participants, together with the participants’ greater-than-typical readiness to embrace a culture of open data sharing1, posed non-trivial challenges for the instructional team. Because the workshop curriculum consisted of a set of presentations that had been developed for data management novices who are expected to be skeptical of the idea of publicly sharing their own research data, there appeared to be some degree of mismatch between the workshop participants and some portions of the instructional materials that were to be tested. Responses to the pre-course survey were collected two weeks before the course, giving instructors had the chance to review these responses before the workshop began. With this information about participants’ prior knowledge, instructors were able to anticipate this incongruence to some degree and make some2 last-minute adjustments to the lessons and instructional format (e.g., addition of some hands-on activities to complement certain lessons). At least one of the more advanced participants was contacted in advance of the workshop in an effort to help set appropriate expectations for what the workshop would include.

---

1 All but 4 of the participants reported data sharing practices that they judged to be more open than the norm for their fields. (The remaining 4 judged their practices to be

2 While instructors had some time to make changes to the lessons in anticipation of the more advanced levels of some participants, the time frame was very short and allowed for only minor changes.
Figure 1: Component expertise scores for participants based on pre-course survey

Component Expertise Scores for Participants based on Pre-Course Survey

- pre-course DM practices score
- experience with metadata
- experience with data types
- software expertise

participant ID
Analysis of correlations between participants’ expertise scores indicates no significant relationships between participants’ levels of expertise in the four areas measured (see Table 3), suggesting that these skill areas are independent.

Table 3: Pearson correlations between reported areas of expertise for participants

<table>
<thead>
<tr>
<th></th>
<th>pre-course data management practices score</th>
<th>experience with metadata</th>
<th>experience with data types</th>
<th>software / scripting expertise</th>
</tr>
</thead>
<tbody>
<tr>
<td>pre-course data management practices score</td>
<td>0.871</td>
<td>0.449</td>
<td>0.371</td>
<td></td>
</tr>
<tr>
<td>experience with metadata</td>
<td>0.046</td>
<td>0.390</td>
<td>0.503</td>
<td></td>
</tr>
<tr>
<td>experience with data types</td>
<td>0.212</td>
<td>0.240</td>
<td>0.500</td>
<td></td>
</tr>
<tr>
<td>software / scripting expertise</td>
<td>0.249</td>
<td>-0.188</td>
<td>0.189</td>
<td></td>
</tr>
</tbody>
</table>

Relationships between Perceived Relevance of Course Curriculum and Participants’ Data Management and Technical Experience and Expertise

Correlations between participants’ prior knowledge, skills, and practices and their evaluations of slide quality, presentation delivery, and information relevance were also examined. From among the expertise variables, only pre-course data management practice scores were found to have significant relationships with any of the outcome variables. Through this analysis, pre-course data management practice scores were found to be significantly negatively correlated with perceived slide quality ($r=-0.664$, $df=13$, $p=.007$) and with evaluations of information relevance ($r=-0.681$, $df=13$, $p=.005$). A negative correlation between pre-course data management practice scores and presentation delivery ratings was nearly significant ($r=-0.507$, $df=13$, $p=.054$). These results suggest that the participants who perceived the short course to be of higher quality and more personally relevant were those who were less advanced in data management practices, and that pre-course levels of data management practices could account for approximately 44% of the variance in perceptions of information quality and 46% of the variance in perceived relevance of the information.
Relationships between Ratings for Course Quality and Relevance

Because the three of the outcome variables (average information quality rating, average presentation delivery rating, and average information relevance rating) appeared to have similar relationships with the pre-course data management practices scores, relationships between these variables was further investigated. Analysis of correlations between these three variables revealed significant relationships between the variables (see Table 5), with each variable sharing between 25% and 30% of its variance with each of the other two variables. This result indicates that it is not likely participants were able to make independent evaluations of information quality, presentation delivery, and information relevance. Furthermore, these results suggest that interpretations of the evaluations of information quality for any lesson should be made in the context of the presentation quality and relevance scores for the same lesson. For example, a lesson that was considered to have been well-presented is also likely to have received higher ratings for slide information quality.

Table 5: Pearson correlations between quality of presentation slides, quality of instructor’s presentation, and information relevance. Quality and relevance ratings were mean-normalized for each participant for this analysis to remove effects due to individual differences in willingness to give high or low ratings.

<table>
<thead>
<tr>
<th>(r values below diagonal, p values above)</th>
<th>quality of presentation slides</th>
<th>quality of instructor’s presentation</th>
<th>information relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td>quality of presentation slides</td>
<td></td>
<td></td>
<td>8.88E-16</td>
</tr>
<tr>
<td>quality of instructor’s presentation</td>
<td>0.555***</td>
<td></td>
<td>3.11E-15</td>
</tr>
<tr>
<td>information relevance</td>
<td>0.546***</td>
<td>0.502***</td>
<td></td>
</tr>
</tbody>
</table>

Relationships between Pre-Course Data Management Practices and Perceptions of Quality and Relevance for Individual Lessons

The analyses of relationships between pre-course data management practices and perceived course quality and relevance revealed a strong negative relationship: the
more advanced a participant’s data management practices, the less quality and relevance they ascribed to the short-course curriculum. However, what about the individual topics that were covered in the course? Does this relationship exist for each individual lesson, or are some lessons perceived to be of equal quality or relevance regardless of prior data management knowledge and experience? Additional correlation analyses for pre-course data management practices and ratings for individual lessons were performed, and results reveal that the previously-observed negative correlation between prior experience and evaluations of course quality and relevance is also observed for most lessons when each lesson is considered individually.

For most lessons, variance in pre-course data management practices scores could account for between 25% and 65% of one or more of the quality and relevance response variables for each individual lesson (see Table 5). Notable exceptions include the lessons on Analysis and Workflows and on Data Semantics, for which no such relationship was observed. More discussion of the lack of observed relationships for these lessons will be discussed in the sections below that focus on each lesson, but it is worth mentioning here that the Analysis and Workflows lesson was significantly modified by the instructor to include more advanced material, and the Data Semantics lesson was not one of the core lessons being evaluated through the workshop, but rather an additional session added in an effort to provide some information more relevant to the advanced workshop participants.

<table>
<thead>
<tr>
<th>LESSON</th>
<th>quality of presentation slides</th>
<th>quality of instructor’s presentation</th>
<th>information relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Why Data Management?</td>
<td>-0.607 *</td>
<td>-0.548 *</td>
<td>-0.650 **</td>
</tr>
<tr>
<td>Data Sharing</td>
<td>-0.674 **</td>
<td>-0.034</td>
<td>-0.573 *</td>
</tr>
<tr>
<td>Advertising Your Data</td>
<td>-0.577 *</td>
<td>-0.027</td>
<td>-0.622 *</td>
</tr>
<tr>
<td>Data Management Plans and Tools</td>
<td>-0.604 *</td>
<td>-0.527 *</td>
<td>-0.805 ***</td>
</tr>
<tr>
<td>Data Collection, Entry and Manipulation</td>
<td>-0.469 +</td>
<td>-0.798 ***</td>
<td>-0.502 +</td>
</tr>
<tr>
<td>Quality Assurance / Quality Control</td>
<td>-0.614 *</td>
<td>-0.524 *</td>
<td>-0.619 *</td>
</tr>
<tr>
<td>Analysis and Workflows</td>
<td>-0.188</td>
<td>0.067</td>
<td>0.307</td>
</tr>
<tr>
<td>Intro to Metadata</td>
<td>-0.645 **</td>
<td>-0.537 *</td>
<td>-0.691 **</td>
</tr>
<tr>
<td>Writing Quality Metadata</td>
<td>-0.670 **</td>
<td>-0.477 +</td>
<td>-0.629 *</td>
</tr>
<tr>
<td>Data Citation</td>
<td>-0.565 *</td>
<td>-0.213</td>
<td>-0.241</td>
</tr>
<tr>
<td>Protected Backups and Data Preservation</td>
<td>-0.543 *</td>
<td>-0.411</td>
<td>-0.391</td>
</tr>
<tr>
<td>Data Semantics</td>
<td>-0.008</td>
<td>-0.071</td>
<td>-0.164</td>
</tr>
</tbody>
</table>
Evaluations of Course Materials
Participants responded to a variety of questions intended to elicit both quantitative ratings of information and presentation quality and relevance and qualitative feedback including highlighting of memorable information and suggestions for improvement. This feedback is presented in this section, which includes an overview of quantitative results and a summary of feedback for the lessons and presenters in general as well as for each lesson individually.

Quantitative Results

Quality of Information in Presentation Slides
When considering the short course as a whole, participants most often rated the quality of information included in the presentation slides as ‘very good’, or 5 on a scale ranging from ‘very poor’ (1) to ‘exceptional’ (7). Figure 2 shows the results for this response variable. The negative correlation between pre-course data management practices scores and perceived quality of information presented is evident in this figure, with a majority of the ‘exceptional’ and ‘excellent’ ratings coming from participants less experienced in data management and sharing, and the majority of the ‘good’, ‘fair’, or ‘poor’ ratings coming from participants with higher self-reported levels of data management practice. None of the participants rated the quality of information in any of the lessons as ‘very poor’.

Most individual lessons also received a median information quality rating of ‘very good’, with the exception of the “Advertising Your Data” lesson, which received a median rating of ‘good’ (4). Further discussion of the ratings for each lesson is included in the sections below.

Quality of Instructors’ Delivery of Presentations
Ratings for presentation delivery were more varied than those for information quality, but once again participants most often rated the quality of presentation delivery as ‘very good’, or 5 on a scale ranging from ‘very poor’ (1) to ‘exceptional’ (7). Figure 3 shows the results for this response variable. The negative correlation between pre-course data management practices scores and perceived quality of presentation delivery is once again evident in this figure, with a majority of the ‘exceptional’ and ‘excellent’ ratings coming from participants less experienced in data management and sharing, and the majority of the ‘fair’ or ‘poor’ ratings coming from participants with higher self-reported levels of data management practice. None of the participants rated the quality of presentation delivery for any of the lessons as ‘very poor’.
Figure 2: Summary of ratings for quality of information included in presentation slides for each lesson. The rating scale ranged from ‘very poor’ (1) to ‘exceptional’ (7). Circle size and the number in the center of each circle indicate the total number of participants who assigned each rating, and an asterisk (*) denotes the median rating for each lesson. Colors of circle segments represent the level of pre-course data management practice reported by each participant, with green representing those with scores below the group mean, and purple representing those with scores above the group mean.

**Ratings for Presentation Slide Quality**

Please rate the quality of the information included in the "_____" presentation slides. Things to consider include (but are not limited to) completeness, accuracy, clarity, relevance, and visual appeal of the information included on the slides.
Figure 3: Summary of ratings for quality of presentation delivery for each lesson. The rating scale ranged from 'very poor' (1) to 'exceptional' (7). Circle size and the number in the center of each circle indicate the total number of participants who assigned each rating, and an asterisk (*) denotes the median rating for each lesson. Colors of circle segments represent the level of pre-course data management practice reported by each participant, with green representing those with scores below the group mean, and purple representing those with scores above the group mean.

Ratings for Presentation Delivery

Please rate the quality of the instructor's presentation of the lecture component of the “_____” lesson. Things to consider include (but are not limited to) presentation style, pace, choice of words, and interaction with workshop participants.

<table>
<thead>
<tr>
<th>presentation title</th>
<th>pre-course DM practices score</th>
<th>median rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intro to Data Management</td>
<td>below group mean</td>
<td>above group mean</td>
</tr>
<tr>
<td>Data Sharing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Advertising Your Data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Management Plans &amp; Tools</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Entry &amp; Manipulation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quality Assurance I Quality Control</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analysis &amp; Workflows</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intro to Metadata</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Writing Quality Metadata</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Citation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protected Backups &amp; Data Preservation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Semantics</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Frequencies of ratings:
- **Excellent** (5): 1, 2, 3, 4, 5
- **Very Good** (6): 1, 2, 3, 4, 5
- **Good** (7): 1, 2, 3
- **Fair** (8): 1
- **Poor** (1): 1, 1
Figure 4: Summary of ratings for relevance of the information included in each lesson. The rating scale ranged from 0% to 100% in increments of 10%. (Responses have been aggregated somewhat for visualization purposes.) Circle size and the number in the center of each circle indicate the total number of participants who assigned each rating, and an asterisk (*) denotes the median rating for each lesson. Colors of circle segments represent the level of pre-course data management practice reported by each participant, with green representing those beneath the group mean, and purple representing those above the mean for the group.

**Evaluation of Information Relevance**

In light of your own data management needs and prior experience with data management, what portion of the information in the "_____" presentation was relevant and useful for you personally?
Figure 5: Summary of ratings for quality of hands-on activities included with some lessons. The rating scale ranged from ‘very poor’ (1) to ‘exceptional’ (7). Circle size and the number in the center of each circle indicate the total number of participants who assigned each rating, and an asterisk (*) denotes the median rating for each lesson. Colors of circle segments represent the level of pre-course data management practice reported by each participant, with green representing those beneath the group mean, and purple representing those above the mean for the group.

### Ratings for Hands-on Activity Quality

Please rate the quality of the activities included in the hands-on component of the "____" lesson. Things to consider include (but are not limited to) clarity of instructions, appropriateness of the task(s), and relevance to the topic.

<table>
<thead>
<tr>
<th>Rating</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>Median Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>exceptional</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>excellent</td>
<td></td>
<td>1</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>very good</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>good</td>
<td></td>
<td>3</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fair</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>poor</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>very poor</td>
<td></td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**activity title**

**pre-course DM practices score**

<table>
<thead>
<tr>
<th>below group mean</th>
<th>above group mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>green</td>
<td>purple</td>
</tr>
</tbody>
</table>

*Please rate the (quality of the (activities included in the (hands-on (component of the ("____") (lesson. (Things to consider include (but are not limited to) (clarity of instructions, appropriateness of the task(s), and relevance to the topic.**
Figure 6: Summary of ratings for relevance of hands-on activities included with some lessons. The rating scale ranged from 0% to 100% in increments of 10%. (Responses have been aggregated somewhat for visualization purposes.) Circle size and the number in the center of each circle indicate the total number of participants who assigned each rating, and an asterisk (*) denotes the median rating for each lesson. Colors of circle segments represent the level of pre-course data management practice reported by each participant, with green representing those beneath the group mean, and purple representing those above the mean for the group.

### Evaluation of Activity Relevance

In light of your own data management needs and prior experience with data management, what portion of the hands-on activities for the "_____
\" lesson was relevant and useful for you personally?

<table>
<thead>
<tr>
<th>activity title</th>
<th>pre-course DM practices score</th>
<th>median rating</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>below group mean</td>
<td>above group mean</td>
</tr>
<tr>
<td>Data Management Plans &amp; Tools</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analysis &amp; Workflows</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Writing Quality Metadata</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Creation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Circle size and the number in the center of each circle indicate the total number of participants who assigned each rating, and an asterisk (*) denotes the median rating for each lesson.
A majority (6) of the individual lessons also received a median information quality rating of ‘very good’. Four (“Analysis and Workflows”, “Introduction to Metadata”, “Protected Backups and Data Preservation”, and “Data Semantics”) garnered median ratings of ‘excellent’, and two (“Data Sharing” and “Advertising Your Data”) received median ratings of ‘good’. Further discussion of the ratings for each lesson is included in the sections below.

Relevance of Information Included in Lessons
Ratings for information relevance were quite high, with a majority or participants reporting that 70% or 80% of the lesson content was relevant and useful to them personally. Figure 4 shows the results for this response variable. Again, the negative correlation between pre-course data management practices scores and perceived relevance of information presented is evident in this figure, with a majority of the 90% or 100% relevance ratings coming from participants less experienced in data management and sharing, and the majority of the below 50% relevance ratings coming from participants with higher self-reported levels of data management practice.

Most of the individual lessons also received median information relevance ratings of 70% or 80%, but two of the lessons (“Data Management Plans and Tools” and “Data Entry and Manipulation” received median relevance ratings of 90% or 100%. Further discussion of the ratings for each lesson is included in the sections below.

Quality and Relevance of Hands-On Activities
Four of the lessons (“Data Management Plans and Tools”, “Analysis and Workflows”, “Writing Quality Metadata”, and “Data Citation”) included hands-on components that were intended to help participants implement or practice ideas or techniques that had been introduced in the preceding lesson. Quality ratings for the hands-on activities were generally lower than the ratings for information or delivery quality, with the hands-on activities receiving a median rating of ‘good’– 4 on a scale ranging from ‘very poor’ (1) to ‘exceptional’ (7). Figure 5 shows the results for this response variable. Ratings for individual hands-on sessions varied considerably between lessons, with the “Writing Quality Metadata” activity receiving a median rating of ‘very good’ (5), the “Data Management Plans and Tools” activity receiving a median rating of ‘good’ (4), and the “Analysis and Workflows” and “Data Citation” activities receiving median ratings of ‘fair’ (3).

A summary of evaluations of hands-on activity relevance is presented in Figure 6. Although participants tended to rate the quality of the hands-on activities lower than the quality of the presentation slides and delivery, evaluations for information relevance were very similar for presentations and hands-on activities. Median relevance ratings for all hands-on sessions except the “Data Citation” session were 70% or 80%, and the median relevance rating for the “Data Citation” session was 50% or 60%. Further
Discussion of the quality and relevance ratings for each hands-on session is included in the sections below.

Discussion of Evaluations and Suggestions

General Feedback on Lessons
Several themes emerged again and again in the feedback on the lessons, and could be considered as general guidelines for continuing development of the lesson series.

• **More concrete or ‘real-world’ examples and stories that illustrate points should be added to the lessons.** When participants reported on the most memorable items from the lessons, they were most likely to mention such examples or stories, and the most common suggestion overall was to include more information of this type. Care should be taken, however, to make sure that stories/examples chosen support rather than undercut points being made. In several cases, participants felt disillusioned by stories of situations when data sharing didn't happen or data management plans weren’t followed.

• **Include more information about (and links to) tools and resources for data management and sharing.** Information about and/or links to tools or resources were pointed out as memorable information almost as often as concrete or real-world examples or stories, and were also the next most requested type of information.

• **Many of the slides are text-heavy – the amount of text on these slides should be pared down and/or spread over more slides.** Many participants noted being overwhelmed by the amount of text on some of the slides, and several suggested that two versions of the information be made available: one as a set of pared-down slides for use during the presentations, and the other a reference document that could be either a white paper or more text-heavy set of slides for use outside of a workshop/presentation.

• **Continue to use standard formatting for presentation slides.** Rather than a suggestion for improvement, this is a reaffirmation that the standardized slide formatting was appreciated by the participants.

• **Describe or define jargon.** Many participants noted the need for more attention to definitions of basic concepts when they first arise. “Metadata”, “data management”, “discovery”, “tags”, “EML”, and discussion of differences between “quality assurance” and “quality control” were pointed out as examples of terms that need more clarification when first mentioned.

• **If audience is not new to data management and sharing, minimize elements of lessons that focus on why researchers should manage/share/backup/cite data.** Several participants mentioned the lessons feeling like ‘preaching to the choir’ at different points.

• **All lessons should include information about best practices.** A number of participants mentioned that they would like to know more about best practices
for analysis and workflows and for activities such as using social media to support data sharing/advertising.

- **If audience is not new to data management, presentations should go into more depth on some topics.** More advanced participants noted that they would like more advanced information about QA/QC techniques, metadata creation, and data citation.

- **If lessons are to be used in a workshop format rather than as stand-alone lectures, redundant information should be removed.** Participants remarked on repetition of the same information in different lessons.

### General Feedback on Presentation Delivery

- **Presenters should take care to speak slowly, clearly, and loudly enough and use techniques to keep the audience engaged.** The most common critical observation regarding presentation delivery was that instructors spoke too quickly, didn’t pause long enough after presenting information that participants might wish to take note of (e.g., a URL), or spoke so quietly that it was difficult to hear. Instructors were also encouraged to make more use of gestures, movement, changes in speaking tone/pitch, and humor to keep the audience engaged. In two cases, it was also mentioned that presenters should not express cynicism about or lack of interest in the topic being presented.

- **Encourage more active participation by the audience.** Many participants remarked that the lessons would have been more engaging if the instructors had led a discussion related to the topic at hand before beginning or during the presentation, or had given participants tasks or discussion topics for them to work on in groups. Box 1 includes some suggestions for facilitating more interactive presentations.

- **Presenters should avoid reading from slides during the presentation.** Participants frequently remarked that instructors seemed to be just reading from the slides rather than using the slides as a support for an engaging presentation. Participants also provided a lot of positive feedback for speakers who avoided this presentation pitfall.

- **Presenters should be familiar with slide content and knowledgeable about all topics covered.** For several of the lessons, some participants felt that the instructor, while knowledgeable about the topic, seemed unfamiliar with the slides and/or gave a presentation that didn’t seem to match very well with the slides. Instances in which the presenter seemed not to know much about a topic (beyond what was written on the slide) were also noted.

- **Practice good time management, and take care not to try to present too much information for the time available.** Participants remarked on feeling overwhelmed when they felt that too much information was being crammed into the short time period allocated for each lesson.

- **Stay on topic. In cases where relevance of information may not be immediately apparent to participants, take care to highlight the relevance aspects.**
Especially in situations where time management was an issue, some participants were put off by what they considered to be irrelevant information or ‘too many rabbit trails’.

- **When applicable, make sure to reserve time for hands-on or interactive activities.** Participants expressed disappointment about the cases in which hands-on activities were planned but then skipped due to lack of time. At the same time, nearly all participants reported enjoyment of hands-on activities when they did occur.

- **When a hands-on task is presented, be sure that goal of the exercise/task and instructions for how to do it are clear.** A number of participants expressed confusion about what they were supposed to be doing during one of the hands-on sessions, and as a result felt that the activity was not a good use of time.

- **When a hands-on task requires participants to perform some activity using their own data, make sure to have an alternative activity available for those who don’t have an appropriate dataset.** Some participants felt they were wasting time during one of the hands-on activities because they didn’t have an appropriate dataset to work with or because the task assigned had already been done for their dataset.

- **Instructors should interact with participants during the hands-on activities.** Some participants wished that instructors had been more active in checking for questions and reviewing participant work to check for problems.

### Box 1: Suggested interactive workshop activities

**Suggested Interactive Workshop Activities**

A number of participants suggested a more interactive format for the workshop. The following are a few (of many) ways to initiate dialog between workshop instructors and participants and between participants themselves.

**Eliciting questions from participants**

*Think, pair, share*

Participants think about questions they have related to the main topic of the upcoming presentation and jot them down. After they have taken note of their questions, they discuss these questions with another participant before sharing them with the larger group via whiteboard or shared online document. The instructor(s) can use this list of questions as a gauge for prior knowledge and pace lesson material appropriately. If there are basic questions that will not be covered in the planned lesson, the instructor can discuss these questions briefly before beginning the prepared material.

*Question follow-up*

Participants write important questions they would like to have answered on small slips of paper just prior to presentation. During the presentation, a workshop assistant summarizes on a slide or on a whiteboard, etc. at the back of the room. At end of presentation, the instructor reviews this list and makes sure all questions have been answered before opening for other questions.

**Helping participants plan for new data management practices**

After a presentation, each participant writes about the ‘low-hanging fruit’ (i.e., something he/she could do that would achieve a large gain for a relatively small effort) for him/her in the data management area covered in the presentation. Participants then circulate about the room and try to find one or more other people who identified the same task. Instructor could suggest that participants put these tasks on their calendars and follow up with each other to check that fellow participants have done the tasks they planned.
• **Lesson sequencing should be carefully planned to allow participants to make the most of the information presented.** A number of participants felt that the ‘Introduction to Metadata’ and ‘Semantics’ lessons should have been given earlier in the course, and that the ‘Data Management Planning’ lesson should have come nearer to the end.

• **Let participants know in advance that the lesson slides are online and/or slides or handouts before the workshop.** Participants mentioned that they were happy to have the ability to access the slides online after the workshop, but some participants mentioned that they would have liked to review the slides before the workshop or to have a handout containing key information to refer to during the workshop.

• **When advertising and recruiting participants, make sure to set appropriate expectations for the evaluation aspects of the workshop or short course.** One participant was put off by repeated references to the ‘guinea pig’ aspect of the course, and many participants were unhappy with the amount of time they spent filling out surveys. Such dissatisfaction can be minimized by providing participants with information about how much time one should expect to spend giving feedback on the workshop before they make a commitment to attend.

**Lesson 1: Why Data Management?**
The introductory lesson entitled “Why Data Management” received median ratings of ‘very good’ (5 on a 7-point scale) and a median relevance rating of 80% (range 20% to 100%). As was the case with most of the lessons, participants with lower pre-course data management practices scores evaluated this lesson to be of higher quality and relevance than did those with higher levels of self-reported data management practice.

• **Positives:**
  - Inclusion of stories that illustrate points about the consequences of poor data management.
  - The data life cycle concept and chart.
  - The discussion of how data management and sharing benefit science and researchers.
  - Lesson was a useful model for how to communicate about the topics of data management.

• **Suggestions for improvement:**
  - Need to include more real-life/concrete examples from fields relevant to participants.
  - Jargon (e.g., “metadata”, “data management”) should be described or defined more clearly.
  - Discussion of benefits on data management should focus more on benefits to an individual researcher (e.g., organization, efficiency, career mobility, visibility, credibility).
  - There is no need for so much justification and repetition (i.e., ‘preaching to the choir’) about why researchers should manage and share data.
  - Some of the slides are text-heavy and should be pared down.
• Should include a discussion of general trends in data sharing.
• Should include longevity information about repositories and practices as a way to encourage trust to use these resources.
• Should include a broader view of data management that extends beyond the research world.
• Should include more emphasis on NSF’s and other agencies’ data management planning requirements.
• Need to update graph showing data storage not keeping up with data generation to include more recent data.

Lesson 2: Data Sharing
The lesson entitled “Data Sharing” received median ratings of ‘very good’ (5 on a 7-point scale) and a median relevance rating of 80% (range 30% to 100%). As was the case with most of the lessons, participants with lower pre-course data management practices scores evaluated this lesson to be of higher quality and relevance than did those with higher levels of self-reported data management practice.

• Positives:
  o Inclusion of useful lists of and/or links to resources.
  o Discussion of how data sharing benefits science and researchers.
  o Example of effectiveness of data sharing (NASA planet/datacasting).
  o Information about data identifiers.

• Suggestions for improvement:
  o Jargon (e.g., “discovery”, “tags”) should be described or defined more clearly.
  o Need to include more real-life/concrete examples from fields relevant to participants.
  o There is no need for so much justification and repetition (i.e., ‘preaching to the choir’) about why researchers should share data.
  o Need to include more information about best practices and best resources for data sharing and advertising, especially for using social media for these purposes.
  o Some of the slides are text-heavy and should be pared down.
  o Lesson should include more information about scientists’ concerns about data sharing (e.g., security, misuse, unwilling co-authors) and ways these concerns can be addressed.
  o Spend more time discussing data sharing portals and other tools that support data sharing, and include some examples (with screenshots) of these tools.
  o Avoid including counter-productive examples – stories about situations in which data sharing was planned or attempted, but didn’t happen.
  o Place more emphasis on the benefits of data sharing for the individual researcher (e.g., career mobility, visibility, credibility).
  o Include more discussion of data privacy issues.
  o Endnotes should be converted to footnotes.
Include more information about how to monitor data citations efficiently and effectively.
• Need more information about sequence archives and how to avoid problems when using them.

NOTE: please see additional suggestions for improvement based upon participants’ perceptions of obstacles to adoption of best practices in the section on Obstacles to adopting best practices for data sharing.

Lesson 3: Advertising Your Data
The lesson entitled “Advertising Your Data” received median ratings of ‘good’ (4 on a 7-point scale) and a median relevance rating of 70% (range 40% to 90%). As was the case with most of the lessons, participants with lower pre-course data management practices scores evaluated this lesson to be of higher quality and relevance than did those with higher levels of self-reported data management practice.

• Positives:
  o Open-ended discussion that was part of this lesson.
  o Inclusion of useful lists of and/or links to resources.
  o Discussion of how data sharing benefits science and researchers.
  o Example of effectiveness of data sharing (NASA planet/datacasting).
  o Explanation of the differences between archiving and advertising.
  o Information about advertising through social networks.
  o Inclusion of concrete, real-world examples.

• Suggestions for improvement:
  o Need to include more information about best practices and best resources for data sharing and advertising, especially for using social media for these purposes.
  o Need to include more real-life/concrete examples from fields relevant to participants.
  o Jargon (e.g., “data sharing”, “advertising”, “archiving”, “catalog”, and “registry”) should be described or defined more clearly.
  o Some of the slides are text-heavy and should be pared down.
  o There is no need for so much justification and repetition (i.e., ‘preaching to the choir’) about why researchers should share data.
  o Lesson should include more information about scientists’ concerns about data sharing (e.g., security, misuse, unwilling co-authors) and ways these concerns can be addressed.
  o Spend more time discussing data sharing portals and other tools that support data sharing and advertising, and include some examples (with screenshots) of these tools.
  o Should include longevity information about repositories and practices as a way to encourage trust to use these resources.
  o Add ResearchGate to the list of resources for data sharing and advertising.
  o Remove the word cloud that includes facebook and twitter.
Participant found it awkward to read text on agency requirements during presentation.

**Lesson 4: Data Management Planning**

The lesson entitled “Data Management Planning” received median ratings of ‘very good’ (5 on a 7-point scale) and a median relevance rating of 90% (range 50% to 100%). As was the case with most of the lessons, participants with lower pre-course data management practices scores evaluated this lesson to be of higher quality and relevance than did those with higher levels of self-reported data management practice.

- **Positives:**
  - Outline/overview of data management plans and information about resources to help create them.
  - Inclusion of concrete, real-world examples.
  - The DMPtool/website tour.
  - The opportunity to practice with the DMPtool during the hands-on section.
  - Inclusion of useful lists of and/or links to resources.
  - Information about budgeting for data management.
  - Emphasis on fact that data includes metadata.
  - Discussion of the importance of data management plans for funding agencies.

- **Suggestions for improvement:**
  - Data Management Planning lesson and hands-on activity should be presented after more topics have been covered.
  - Need more information about alternatives to DMPtool.
  - Need to include more real-life/concrete examples from fields relevant to participants.
  - Include more information about metadata (e.g., introduce EML) before introducing data management plans.
  - Avoid including counter-productive examples – stories about situations in which data management planning was completed but then not followed through with once funding had been received.
  - Include more detailed information about DMPs.
  - Some of the slides are text-heavy and should be pared down.
  - There is no need for so much justification and repetition (i.e., ‘preaching to the choir’) about why researchers should manage data.
  - Should include more discussion of the motivation to create a data management plan beyond meeting funding agency requirements.
  - Need to include more examples of good and bad data management plans. One participant suggested having participants work in groups to review and critique a variety of data management plans.
  - Two participants would have liked to have the time to actually create a DMP (or part of one) during the hands-on component of this lesson.
A couple of participants remarked that it was not interesting to go through the DMPtool website without having a particular project or example to work with.

Although most participants found this lesson very valuable, several of the participants mentioned that they weren’t convinced that the DMPtool would be useful for them.

Lesson 5: Data Collection, Entry & Manipulation

The lesson entitled “Data Collection, Entry & Manipulation” received median ratings of ‘very good’ (5 on a 7-point scale) and a median relevance rating of 90% (range 10% to 100%). Although it appears that participants with lower pre-course data management practices scores evaluated this lesson to be of higher quality and relevance than did those with higher levels of self-reported data management practice, these relationships fall short of achieving statistical significance.

- Positives:
  - Information about software and tools such as Googledocs forms for data entry
  - Inclusion of concrete, real-world examples.
  - Information about differences between a spreadsheet and a relational database.
  - Outline of best practices for data collection and entry.
  - Information about non-proprietary, commonly-used data formats.
  - Focus on tools (e.g., Excel) that researchers actually use.

- Suggestions for improvement:
  - Need to include a comparison table for data entry tools.
  - Need to elaborate the section on databases so that it will be clearer and more understandable for beginners.
  - Some of the slides are text-heavy and should be pared down.
  - Tips on data entry should be moved from the QA/QC lesson to this one.
  - Need to include more information about file types for storing data.
  - Spreadsheet vs. relational database should be a table instead of a 2-column list.
  - Content about spreadsheets, databases, and data manipulation redundant with info in Analysis and Workflows lesson.
  - Clarify ‘use’ (e.g., for data entry? manipulation? export to analytic environment?) in ‘easy to use’.
  - Need to include more information about why the situation in the Sevilleta example is a problem, and how to correct the problem.
  - Focusing on Microsoft tools conflicts with the open-source message of some of the other lessons.
  - Could mention Octave as a free alternative to Matlab.
  - Add a hands-on exercise such as creating database files from old .csv files or data validation and creation in Excel.
NOTE: please see additional suggestions for improvement based upon participants’ perceptions of obstacles to adoption of best practices in the section on Obstacles to adopting best practices for data collection, entry and manipulation.

**Lesson 6: Quality Assurance/Quality Control**

The lesson entitled “Quality Assurance/Quality Control” received median ratings of ‘very good’ (5 on a 7-point scale) and a median relevance rating of 80% (range 20% to 100%). As was the case with most of the lessons, participants with lower pre-course data management practices scores evaluated this lesson to be of higher quality and relevance than did those with higher levels of self-reported data management practice.

- **Positives:**
  - Information about differences between a spreadsheet and a relational database.
  - Idea of looking for outliers in data using basic statistics and plots.
  - Overview of basic QA/QC measures.
  - Information about software and tools such as Googledocs forms.

- **Suggestions for improvement:**
  - Need to include less obvious information and/or more advanced information such as the QA/QC topics covered in the Analysis and Workflows lesson.
  - Need to include more information about appropriate ways to handle outliers and other errors.
  - Need to include more real-life/concrete examples from fields relevant to participants.
  - Need to include more information about tools and resources.
  - Some of the slides are text-heavy and should be pared down.
  - Information seems disjointed.
  - Describe the differences between ‘quality assurance’ and ‘quality control’.
  - Tips on data entry should be moved to the Collection, Entry, and Manipulation lesson to leave more time to focus on dealing with errors in data that has already been entered.
  - Include an example of how to use Excel to do QA/QC on already-entered data.
  - Include information about if/how QA/QC steps should be included in metadata.

NOTE: please see additional suggestions for improvement based upon participants’ perceptions of obstacles to adoption of best practices in the section on Obstacles to adopting best practices for quality control/quality assurance.

**Lesson 7: Analysis and Workflows**

The lesson entitled “Analysis and Workflows” received median ratings of ‘very good’ (5 on a 7-point scale) and a median relevance rating of 80% (range 0% to 100%). In
contrast with the pattern observed for most of the lessons, there was no observed relationship between pre-course data management practices scores and evaluations of information quality and relevance for this lesson.

• Positives:
  o Engaging presentation.
  o Information about scientific workflows.
  o Inclusion of concrete/real-world examples and stories to illustrate points.
  o Information about differences between worksheet and table.
  o Inclusion of philosophical insights.
  o Information about not altering archival data.
  o Discussion of “big data”.

• Suggestions for improvement:
  o Need to include more real-life/concrete examples from fields relevant to participants.
  o Include more information about best practices/best resources for analysis and workflows.
  o Some of the slides are text-heavy and should be pared down.
  o Information too abstract and/or advanced.
  o Information seemed disjointed, and presentation delivery didn’t seem to correspond with slides.
  o Not enough information about workflows to allow participants to start building.
  o Describe or define jargon.
  o Don’t spend so much time on GIS terminology.
  o Need better transition between basic workflow graphics and slide that shows data integrated into workflow.
  o NDMS plot mistakenly labeled as PCA.
  o One participant suggested including more information about tools/resources mentioned, such as Kepler and Vistrails, while another suggested minimizing discussion of Kepler since it will not be useful to most researchers.
  o Section on data types and semantics should come earlier in the lesson.
  o Need to make a clearer distinction between formal and informal workflows.
  o Need to include information about how to deal with evolution in workflows over the course of the project.
  o Presentation included too much (possibly irrelevant) information for the time available, leaving no time for the hands-on component of the lesson.

NOTE: please see additional suggestions for improvement based upon participants’ perceptions of obstacles to adoption of best practices in the section on Obstacles to adopting best practices for analysis and workflow documentation.
Lesson 8: Introduction to Metadata
The lesson entitled “Introduction to Metadata” received median ratings of ‘very good’ (5 on a 7-point scale) and a median relevance rating of 80% (range 50% to 100%). As was the case with most of the lessons, participants with lower pre-course data management practices scores evaluated this lesson to be of higher quality and relevance than did those with higher levels of self-reported data management practice.

- **Positives:**
  - Inclusion of useful lists of and/or links to resources.
  - Inclusion of concrete/real-world examples and stories to illustrate points.
  - Information about a variety of different metadata content/formats/standards.
  - The ‘information entropy’ figure.
  - The how-to part of the presentation/the slides about ‘concerns+solutions’.
  - Good positioning of text and images.
  - Information about using updates to metadata as a method for data management.

- **Suggestions for improvement:**
  - Presentation should be given earlier in the course.
  - Remove information that is redundant with other lessons.
  - Include an example of how to search for metadata using a search engine or website.
  - Need to include more real-life/concrete examples from fields relevant to participants.
  - Too much focus on conceptual; need to include more advanced/in-depth information about metadata.
  - Describe or define jargon.
  - Include information about how metadata records can be used to track provenance.
  - Include information about the general future direction of metadata.

Lesson 9: How to Write Quality Metadata
The lesson entitled “How to Write Quality Metadata” received median ratings of ‘very good’ (5 on a 7-point scale) and a median relevance rating of 80% (range 60% to 100%). As was the case with most of the lessons, participants with lower pre-course data management practices scores evaluated this lesson to be of higher quality and relevance than did those with higher levels of self-reported data management practice.

- **Positives:**
  - Information about the USGS metadata tool.
  - Information about existing metadata standards.
  - The how-to part of the presentation/steps for creating metadata.
  - Inclusion of useful lists of and/or links to resources.
  - Inclusion of concrete/real-world examples and stories to illustrate points.
• Idea of having people not connected to a project review metadata for the project.
• Idea that writing metadata isn’t just about making sure your data can be interpreted by others, but also about checking how well you know your own data.
• Emphasis on the importance of being specific in file descriptions.
• Examples of good and bad metadata records.
• Opportunity to practice with own data.

- Suggestions for improvement:
  • Include more information about tools and resources presented such as Morpho, thesauri, and tools to crosswalk standards.
  • Include information about alternatives to the USGS metadata tool.
  • Remove information that is redundant with other lessons.
  • Include more examples of bad vs. good metadata, and discuss difficulties commonly encountered when creating metadata.
  • Provide examples of how much more can be done with thoroughly documented vs. vaguely described datasets (e.g., with respect to data discovery and reproducibility).
  • Include summary of similarities and differences between different metadata standards.
  • Include a better explanation of saving/export/publish settings for metadata.
  • Clarify differences between ‘none’ and ‘unknown’.
  • Include more information about data completeness.
  • Need more information about linking metadata for multiple datasets of a large project.
  • Include more information about how metadata is typically generated in a large government institution (e.g., is user automatically prompted to enter metadata?, are there quality control procedures?).
  • Provide information about how people choose a metadata standard to use.
  • Include more information about the next steps after creating metadata.
  • Need alternative hands-on activity for participants who don’t have a dataset to work with.
  • Suggest walking through an example form on the OME site rather than having participants write their own metadata.
  • Need to interact more with participants during hands-on portion of lesson.

NOTE: please see additional suggestions for improvement based upon participants’ perceptions of obstacles to adoption of best practices in the section on Obstacles to adopting best practices for metadata creation.
Lesson 10: Data Citation

The lesson entitled “Data Citation” received median ratings of ‘very good’ (5 on a 7-point scale) and a median relevance rating of 70% (range 30% to 100%). The observed relationships between pre-course data management practices and evaluations of information quality and relevance were somewhat less strong for this lesson than for most of the lessons – participants with lower pre-course data management practices scores evaluated this lesson to be of higher quality (but not higher relevance) than did those with higher levels of self-reported data management practice.

- Positives:
  - Explanation of what a DOI (or other identifier) is and how to create a DOI.
  - Inclusion of concrete/real-world examples and stories to illustrate points.
  - Information about how to cite a dataset.

- Suggestions for improvement:
  - Some of the slides are text-heavy and should be pared down.
  - Include more examples of accepted/standard data citation formats and information to include in a citation earlier in the presentation.
  - There is no need for so much justification and repetition (i.e., ‘preaching to the choir’) about why researchers should cite datasets.
  - Need to include more real-life/concrete examples from fields relevant to participants.
  - Remove information that is redundant with other lessons.
  - There are too many lists of attributes.
  - Need to include more advanced/in-depth information about data citation.
  - Include a clearer description of what a DOI is and more information about how to create different identifiers such as DOI and UUID.
  - Discuss rates of data citation in academic publications.
  - Show examples of data being cited in a publication.
  - Suggest that repositories should provide citation exports for datasets that can be imported into bibliographic management software.
  - Create clearer instructions for hands-on task.
  - Hands-on should just be a quick activity and/or summary slide integrated with lesson.

NOTE: please see additional suggestions for improvement based upon participants’ perceptions of obstacles to adoption of best practices in the section on Obstacles to adopting best practices for data citation.

Lesson 11: Protected Backups and Data Preservation

The lesson entitled “Protected Backups and Data Preservation” received median ratings of ‘very good’ (5 on a 7-point scale) and a median relevance rating of 80% (range 40% to 100%). The observed relationships between pre-course data management practices and evaluations of information quality and relevance were somewhat less strong for this lesson than for most of the lessons – participants with lower pre-course data management practices scores evaluated this lesson to be of higher quality (but not
higher relevance) than did those with higher levels of self-reported data management practice.

- **Positives:**
  - Engaging presentation.
  - Use of videos, stories, and real-world examples to engage audience and illustrate points.
  - Information about best practices for data backups and preservation.
  - Discussion of naming files and choosing formats.

- **Suggestions for improvement:**
  - Need to include specific recommendations for backup software/tools.
  - Suggest mentioning Dropbox and GoogleDrive as resources for data sharing and backup.
  - There is no need for so much justification and repetition (i.e., ‘preaching to the choir’) about why researchers should backup data.
  - Need more details about ways to make backup and preservation more fluid.
  - Some of the slides are text-heavy and should be pared down.

**NOTE:** please see additional suggestions for improvement based upon participants’ perceptions of obstacles to adoption of best practices in the section on Obstacles to adopting best practices for protected backups and data preservation.

**Lesson 12: Data Semantics**

The lesson entitled “Data Semantics” received median ratings of ‘very good’ (5 on a 7-point scale) and a median relevance rating of 80% (range 10% to 100%). In contrast with the pattern observed for most of the lessons, there was no observed relationship between pre-course data management practices scores and evaluations of information quality and relevance for this lesson.

- **Positives:**
  - Engaging presentation.
  - Definition of semantics and introduction to semantic web.
  - Discussion of the transition from HTML to XML to RDF to OWL.

- **Suggestions for improvement:**
  - Need to include more real-life/concrete examples from fields relevant to participants.
  - Information too abstract and/or advanced.
  - Information seems disjointed.
  - Important to spend more time explaining technical concepts for the less technical participants.
  - Need to provide examples of ontologies, RDF, etc.
  - Need more information about tools/resources mentioned such as checksum, version control, and tools for transferring files to RDF format and managing them once in that format.
  - Presentation should be given earlier in course.
Reflection on Expectations
At the end of the short course, participants were asked to reflect on their expectations for the course and whether or not these expectations were met. Table 7 shows responses to these questions.

Table 7: Reflection and feedback on expectations for the short course

<table>
<thead>
<tr>
<th>Responses to: &quot;Did the Data Management Short Course meet your expectations? What knowledge or skills did you gain by participating in the course? If the course did not meet your expectations, please describe how your experience differed from what you had expected.&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>It did, but a lot of information and not enough hands on as discussed in the afternoon. Thanks so much for all of your efforts and the opportunity.</td>
</tr>
<tr>
<td>This was a great overview. I was particularly happy about the data citation, security, and semantic modules - I think these tend to get skipped because they are considered less interesting or too difficult. But they really are important and ecologists in general need to be pushed to learn at least a small amount about these topics.</td>
</tr>
<tr>
<td>I got more than what I expect, a lot of skills I learned in this &quot;short&quot; course BUT a lot also to learn on how to use the tools that we should use</td>
</tr>
<tr>
<td>Many slides seemed redundant, and I would have appreciated individual case studies of different types of data sets and some of the challenges in dealing with and managing these data sets. I do think the presentation might be tweaked to deal with different backgrounds of the participants.</td>
</tr>
<tr>
<td>I expected it to be more technical or high-level, coming out of NCEAS I expected cutting edge. Definitely gained exposure to new tools and philosophies that I appreciated.</td>
</tr>
<tr>
<td>--no response--</td>
</tr>
<tr>
<td>I expected more hands-on experiences. I thought I would leave with a solid plan for what to do with the data I brought with me. I do have a plan from what I learned from the course but I have more to think about in terms of implementing the plan and determining the exact products I want to share. I learned an awful lot about data management, its planning and its practice. I learned a lot of cool new tools that I plan to utilize. I feel more confident in talking about data management and thinking about ways that I can more formally get novel projects off the ground. I think I wanted more, but as someone in the course put it well, it was a lot like drinking from a firehose. Funny that I want more, but this should be somewhat pleasing to the organizers because I am inspired.</td>
</tr>
<tr>
<td>yes, but I expected more of the how and less of the why.</td>
</tr>
<tr>
<td>Yes it was great. Some exercises to cement ideas were lacking, but otherwise great stuff.</td>
</tr>
<tr>
<td>Yes, very much so. I gained some exposure to tools available (which are essential for saving time) although mores hands on time would have been better. Came away with plenty of ideas on things to try to implement. Now I just need more time and more help!</td>
</tr>
<tr>
<td>Yes</td>
</tr>
</tbody>
</table>
I attended the course to get an overview of the challenges and best practices in scientific data management, and the course met this need.

I enjoyed the course overall. I was familiar with much of what was covered, but I still learned some things. I think some modules got too advanced or abstract and should have stayed more practical. The slides were too wordy. The course should have been longer for the amount of information covered, or the information covered should have stayed more high level for the short length.

The data management short course did meet my expectations, but it was definitely different than what I expected. What I wanted was at least an overview of data management issues and best practices, and I definitely got that, but I also expected quite a bit of practical and hands-on work that would force me to gain data management skills. I think that the hands-on elements of this course are the part that need the most work. I also think that the presentations, while great, would be better in the short course setting if they were broken up a little more with hands-on activities and if the hands-on activities were less like another presentation. Even without extensive hands-on activities, I think that this course really helped me understand what I need to do to be on the cutting edge of data management as a scientist. I also think that it gave me valuable resources for approaching various elements of data management.

It met my expectations regarding the information provided. I really felt that all the issues I run into when managing data were covered. However, I though that we were going to do more hands on and that these were going to be guided in a more structured way. I feel that I have been introduced to a completely new world on data management. The relevance of it is very clear (although I was aware of this before). I don’t feel that I have developed much skills yet but I feel that I can dive into this best practices procedures by myself with the support of the slides and online info that is available.

**Pre- and post-course self-reported data management practices and plans**

As part of the pre-course survey, participants provided information about their own data management practices and about perceived data management norms in their fields. At the end of the course, participants reported on their future plans for data management practices. Participants answered questions regarding data sharing, data management planning, data quality control/quality assurance, workflow documentation, creation and management of metadata, data citation, and data backup.

**Guide to interpreting figures**

Figure 8 through Figure 15 summarize responses to the questions about data management practices. Figure 7 provides an overview of how the figures should be interpreted. The vertical bars in each figure represent individual participants, and these bars are ordered by pre-course data management scores: bars representing participants
with lower scores are to the left and those representing participants with higher scores are to the right. Levels of data practice for the data management topic in focus are on the y-axis. Large gray circles indicate the levels of practice that the participants consider norms for their fields. Open semi-circles represent self-reported pre-course levels and post-course plans. Red and blue arrows highlight instances of increase or decrease between pre-and post-course surveys.

**Data sharing**
While all participants reported that researchers in their fields typically ‘share data with collaborators only’ or ‘share data with collaborators and any researcher who requests the data’, nearly all (11 of 15) indicated that their own pre-course level of practice was above what they consider to be the norm. Interestingly, only one participant reported making data available via download through his/her own website, while five reported that they already use a repository to make their data publicly available. By the end of the course, nine researchers reported plans to place their data in public repositories, while four appeared to be unconvinced of the benefits of data sharing and reported plans to share only with collaborators, or in some cases, with researchers who contact them with requests.

**Data management planning**
Prior to attending the short course, only three participants reported having created a formal data management plan (and only in cases in which they were required to do so by a funding source). While one participant reported a perception that researchers in his/her field typically create formal data management plans for each project and others reported that typical researchers never create formal DMPs, most presumed that typical researchers create formal DMPs only when required to do so by a funding source. While 12 of the 15 participants reported never having created a formal data management plan prior to attending the short course, 11 participants left the course with plans to create a formal DMP for every project.

**Data management plan communication**
Participants reported that researchers in their fields typically either don’t communicate about DMPs with members of their research teams, or communicate about DMPs with only some members of their research teams. Awareness of the benefits of communicating about plans for data management appears to have increased through participation in the short course: nine of 15 participants plan to communicate about data management plans with all members of their research teams.

---

3 As was mentioned in the section on Data Management and Technical Experience and Expertise, this indicates that this self-selected group of participants may not realistically represent typical researchers in their fields with respect to willingness to share data.
**Data quality assurance/quality control**

Over half (9) of the participants reported that typical researchers in their fields either take basic steps to ensure data quality but do not have systematic QA/QC routines, and all but one of these nine participants reported personal practices at the level of this perceived norm. The remaining participants noted that typical researchers have established QA/QC routines that they apply to all data, but only two of these participants reported having such routines themselves. By the end of the short course, twelve participants reported plans to establish QA/QC routines for all data sets that they create or use.

**Workflow documentation**

Participant perceptions of norms for workflow documentation vary considerably, with three participants reporting that researchers typically don’t do anything to document workflows, nine participants reporting that typical researchers take some basic steps to document workflows, and two reporting that researchers in their fields typically create comprehensive workflow documentation. Pre-course self-reported workflow documentation practices were similarly varied: four participants reported that they don’t do anything to document workflows, seven reported taking some basic steps to document workflows, and four reported that they already create comprehensive workflow documentation. After participating in the short course, ten participants reported plans to create comprehensive workflow documentation, either with (4 participants) or without (5 participants) the use of specialized technologies.

**Metadata creation and management**

Prior to attending the short course, a majority of the participants (10 of 15) perceived that typical researchers in their fields create some metadata but do not have systems for organizing or managing those metadata. One reported that typical researchers don’t create metadata, while four others reported that good systems for organizing and maintaining metadata (although metadata may not be comprehensive). Four participants reported that they themselves have good systems for organizing and maintaining metadata (though metadata not comprehensive), while most reported creating some metadata but not having a system for organizing or maintaining. By the end of the workshop, ten participants reported plans to have good systems for metadata organization and maintenance. On the other hand, four participants reported no change from previously-reported practices of creating some metadata but not adopting a system for organizing or maintaining those metadata.

**Data citation**

Most short course participants report that formal data citation is not the norm in their fields: typical researchers are presumed to either mention the source of datasets in the article text or acknowledgments (8 participants), or to neglect to identify or cite datasets in publications (3 participants). Three participants reported, however, that including data citations with other bibliographic references was the norm in their fields. Prior to attending the short course, three participants reported never identifying or citing datasets in publications, five reported mentioning data sources in the text or
acknowledgments, and six reported including formal data citations.\textsuperscript{4} By the end of the course, eleven participants reported plans to formally cite datasets in their publications.

\textbf{Data backups}
Most participants (10 of 15) reported that researchers in their fields typically only perform manual data backups that are not frequent enough or comprehensive enough to prevent data loss, and many participants (7 of 15) reported this as their own level of data backup as well. Approximately the same number of participants (8 of 15) had automated backup systems prior to attending the course, and all but three of these participants were responsible for setting up and maintaining their own systems. By the end of the course, 12 participants had plans for automated backup systems, but few had access to personnel who set up and maintain such systems.

\footnote{4 The high level of self-reported formal data citation prior to the course indicates that participants are likely not ‘typical’ researchers with respect to this practice.}
Figure 7: Guide to interpreting Figure 8 through Figure 15

Guide to interpreting figures illustrating pre- and post-course data management practices and/or plans

- Participant considers own level to be higher than typical in field; no change between pre- and post-course
- Participant considers own level to be typical in field; no answer provided for post-course survey
- Participant considers own level to be higher than typical in field; level planned at end of course lower than pre-course

- Participant thinks that most researchers in his/her field have adopted this level of practice.
- Participant’s self-reported levels (or planned levels) of practice.
- Change in participant’s self-reported levels (or planned levels) of data management practice.

one column per participant

participants ordered by
pre-course DM practices score

below group mean
above group mean

level of data management practice

most advanced

least advanced

missing or Other value
Figure 8: Data sharing practices and plans -- pre-course, post-course, and perceived norms

Making data publicly available by publishing in a data repository/archive

- make data publicly available and discoverable by publishing in repository
- make data available for download on website
- share data with collaborators and any researcher who requests
- share data with collaborators only
- never share data
- missing or Other value

participant ID

15 13 3 6 14 4 11 7 2 5 9 12 10 1 8

below group mean  pre-course DM practices score  above group mean
Figure 9: Data management planning practices and plans -- pre-course, post-course, and perceived norms

Creating a formal data management plan

- Create a formal DMP for every project
- Create formal DMP only when required to do so by funding source
- Never create formal DMPs
- Missing or Other value

participant ID: 15 13 3 6 14 4 11 7 2 5 9 12 10 1 8

pre-course DM practices score

below group mean  pre-course DM practices score  above group mean
Communicating about data management plans with members of research teams

- Communicate about DMPs with all members of research teams
- Communicate about DMPs with some members of research teams
- Don’t communicate about DMPs with members of research teams
- Missing or Other value

Participant ID: 15 13 3 6 14 4 11 7 2 5 9 12 10 1 8

Figure 10: Data management plan communication practices and plans -- pre-course, post-course, and perceived norms
Ensuring data quality

- Use cutting-edge practices and technologies to ensure data quality.
- Have established routines for QAQC that are applied to all data created or used.
- Take basic steps to ensure data quality, but QAQC not done in systematic way.
- Don’t do anything to ensure data quality.

Figure 11: Data quality control/quality assurance practices and plans -- pre-course, post-course, and perceived norms
Figure 12: Workflow documentation practices and plans -- pre-course, post-course, and perceived norms

**Documenting workflows**

- Use cutting-edge practices and technologies to document WFs that allow reproduction and provenance tracing.
- Create comprehensive WF documentation, but without use of specialized technologies.
- Take basic steps to document WFs, but not sufficient to exactly replicate analyses and outputs.
- Don't do anything to document workflows.
- Missing or Other value.

participant ID: 15, 13, 3, 6, 14, 4, 11, 2, 5, 9, 12, 10, 1, 8

below group mean  pre-course DM practices score  above group mean
Figure 13: Metadata creation and management practices and plans -- pre-course, post-course, and perceived norms

Creating and managing metadata

- Create comprehensive metadata, have systems for organizing & maintaining.
- Create metadata & have good systems for organizing & maintaining, but metadata not comprehensive.
- Create some metadata, but don’t have a system for organizing or maintaining those metadata.
- Don’t create metadata.
- Missing or Other value.

Typical of researchers in field:
- Pre-course
- Post-course
- Change

Participant ID:
15 13 3 6 14 4 11 7 2 5 9 12 10 1 8

Below group mean
Pre-course DM practices score
Above group mean
Figure 14: Data citation practices and plans -- pre-course, post-course, and perceived norms

Identifying or citing datasets in published studies

- Include formal citations for datasets with other bibliographic references
- Mention the source of datasets in text or acknowledgments but do not include formal citation
- Don’t identify or cite datasets in publications

Participant ID

15 13 3 6 14 4 11 7 2 5 9 12 10 1 8

Typical of researchers in field
Pre-course
Post-course
Change

Missing or Other value

Below group mean
Pre-course DM practices score
Above group mean
Figure 15: Data backup practices and plans -- pre-course, post-course, and perceived norms

**Backing up data**

- Automated backup systems, and personnel who set up & maintain these systems
- Automated backup systems, but each researcher responsible for setup & maintenance of own backup system
- Manual data backups, not frequent/comprehensive enough to prevent data loss
- Don't back up or archive data
- Missing or Other value

**participant ID**

- 15
- 13
- 3
- 6
- 14
- 4
- 11
- 7
- 2
- 5
- 9
- 12
- 10
- 1
- 8

**pre-course DM practices score**

- Below group mean
- Above group mean

**Typical of researchers in field**

- Pre-course
- Post-course
- Change
Obstacles to implementation of recommended Best Practices

In addition to collecting information about participants’ perceptions of the quality and relevance of each lesson, we also requested information about perceived barriers to implementation of the best practices advocated in that lesson. Specifically, we asked participants to report the degree of difficulty they expected to encounter in implementing each best practice as well as the relative impacts of lack of time, lack of technical support, lack of training, and lack of funding. Overall, participants perceived lack of time as the most important barrier to implementation of best practices, followed by the need for more training. Lack of technical support and funding were also identified as factors in many cases, but were perceived as less important obstacles overall.

Obstacles to adopting best practices for data sharing

Responses to the questions regarding recommended best practices to support data sharing are summarized in Figure 16. Overall, participants perceived significant barriers to adoption of these practices, and lack of time was identified as the most important of the barriers suggested. Participants pointed out a need for 1) adoption of agreed-upon standards and 2) services to monitor data use in publications and websites in order to make these practices feasible.

With respect to perceived obstacles to adoption of best practices, the lesson on Data Sharing could be improved by:

• emphasis on ways to reduce the time required to perform the suggested tasks;
• identification of agreed-upon standards that participants can adopt; and
• information about tools or services that researchers can use to help monitor use of their data in publications and websites.

A more detailed summary of these responses and participant comments in response to specific practices advocated are summarized below:

• Document and publish data using standards.
  o Most (10 of 15) of the participants reported foreseeing significant but surmountable obstacles to adopting this practice. While all of the factors (lack of time, training, tech support, and funding) were identified as important in preventing adoption of this practice, time limitations were identified as the most important factor. Several participants identified a lack of agreed-upon standards as an important barrier to implementation of the practice.

• Promote data use via presentations and meetings.
  o Many (6 of 15) of the participants reported that they already engage in this practice, and an equal number perceived only minor obstacles to doing so. Time limitations were identified as the most important factor in preventing adoption of this practice. One participant thought it was likely
that data citations included in presentations at meetings would be ignored or lost in the deluge of information, and thinks that researchers will need to be more proactive than simply including data citations. Another researcher expressed reservations about sharing links at conferences that make data available prior to publication in a peer-reviewed journal.

- **Solicit feedback from data users and address identified issues.**
  - Approximately half (7 of 15) of the participants perceived significant but surmountable obstacles to adoption of this practice, while slightly fewer (5 of 15) foresaw only minor obstacles. Time, training, and technical support were all identified as important obstacles. Several participants mentioned difficulty in finding other researchers who would be willing to contribute time and effort to review data, while several others suggested that implementation and/or documentation of revisions would pose significant problems.

- **Monitor publications and websites for data use and address misapplications.**
  - Several participants (3 of 15) perceived insurmountable obstacles to implementation of this practice, while most other participants expected either significant or minor obstacles. Lack of time was identified as a key factor, and one participant suggested that it would be necessary to have a web service to mine publication and website data and send alerts in order to make this practice feasible.

**Obstacles to adopting best practices for data collection, entry and manipulation**

Responses to the questions regarding recommended best practices to support data collection, entry and manipulation are summarized in Figure 17. Overall, participants perceived few barriers to adoption of these practices, with most participants reporting that they already engage in the practices or that they foresee only minor obstacles. Participants pointed out a need for 1) more information about setting up naming conventions and 2) more information about choosing file types and formats in order to make these practices feasible.

With respect to perceived obstacles to adoption of best practices, the lesson on Data Collection, Entry and Manipulation could be improved by:

- inclusion of best practices for those using relational databases;
- more specific information about how to establish naming conventions for variables (e.g., that contain necessary information without becoming too long) and how to update names for large numbers of files when naming conventions are revised;
- additional information about the specific file types and formats that are broadly useable and will remain so in the future.
A more detailed summary of these responses and participant comments in response to specific practices advocated are summarized below:

- **Enter all data into one table rather than multiple small tables.**
  - Most (9 of 15) of the participants reported that they already do this, and 3 others reported foreseeing only minor obstacles. One participant mentioned that this advice seems to be contrary to the rationale behind relational databases, and another suggested that training on how to design a good table that will remain applicable to a study over time would be necessary.

- **Enter one type of data per column and use consistent names, codes, and formats for the data in each column.**
  - Nearly all (13 of 15) of the participants reported that they already engage in this practice, and the remaining 2 perceived only minor obstacles to doing so. One participant pointed out the importance of learning to develop support tables that include codes and their definitions as well as the relationships between the code tables and the data tables.

- **Create descriptive column names without spaces or special characters.**
  - Most (12 of 15) of the participants reported that they already engage in this practice, and the remaining 3 perceived only minor obstacles to doing so. One participant mentioned difficulties encountered when trying to work with column names in MATLAB, and suggested it would be useful to learn about ways to manage column headers when analytic software requires tables without column headers.

- **Use descriptive file names.**
  - Most participants (10 of 15) reported already engaging in this practice, 3 perceived only minor obstacles, and one foresaw significant but surmountable obstacles. Lack of time was identified as a key factor, and one participant suggested that it would be necessary to have a web service to mine publication and website data and send alerts in order to make this practice feasible. One participant mentioned that it would have been useful to receive some more specific information about how to design file naming conventions so that they would contain enough information without becoming too long. On a related note, several pointed out that descriptive names are good for eliminating confusion, but impractical for reading into an analytical environment such as R or impossible for other programs that have strict limits on file name length. Another participant mentioned that even with good file naming conventions, generating many files with similar names can become overwhelming and necessitate a better system. Yet another mentioned that revising descriptive file names as an analytic process evolves can be very cumbersome.

- **Consistently mark missing data by leaving fields empty (NULL), using a distinct value (such as 9999 or NA) to indicate missing data, or using data flags in a separate column to qualify missing values.**
Most participants (10 of 15) reported already engaging in this practice, 3 perceived only minor obstacles, and one foresaw significant but surmountable obstacles. One participant strongly disagreed with the suggestion to use a value such as 9999 to mark missing values since these values need to be filtered before analysis and are prone to entry errors.

- **Store data in a consistent format that can be read well into the future and that can be used by any application now or in the future.**
  - Approximately half (7 of 15) reported that they already do this, while 6 foresaw only minor obstacles and 2 foresaw significant obstacles. A number of participants expressed uncertainty about which data types or formats conform with this suggested practice and suggested that more information about choosing file types/formats be added to the lesson.

**Obstacles to adopting best practices for quality control/quality assurance**

Responses to the questions regarding recommended best practices to support data quality control and quality assurance are summarized in Figure 18. Overall, participants perceived some barriers to adoption of these practices, although the significance of the perceived obstacles varied between practices. Participants pointed out a need for time-saving tools and/or approaches in order to make these practices feasible.

With respect to perceived obstacles to adoption of best practices, the lesson on Data Quality Control / Quality Assurance could be improved by:

- increased focus on time-saving tools and approaches to make QA/QC more efficient;
- including some basic guidelines for data storage design;
- including discussion of how to handle documenting changes to data when there are several people working with the data at the same time;
- providing information about how to formulate queries to check data quality when visual inspection is not feasible (e.g., too many records);
- including a discussion of what to do with outliers when they are identified.

A more detailed summary of these responses and participant comments in response to specific practices advocated are summarized below:

- **Define and enforce standards for formats, codes, measurement units, and metadata.**
  - Most (10 of 15) of the participants foresaw minor obstacles to implementing this practice, while 3 others reported that they were already doing this. Lack of time was identified as the most important obstacle to implementation. One participant noted that it would be important to know what standards are in a particular domain in order to avoid ‘reinventing the wheel’.

- **Assign responsibility for data quality.**
Participants were equally split between foreseeing significant obstacles, foreseeing minor obstacles, and already engaging in the practice. Lack of time was identified as the most important factor, and some participants mentioned that they didn’t have anyone to whom they could delegate this responsibility.

- **Use double entry, transcription from recorded reading, or read entered data back using a text-to-speech program to ensure quality during data entry.**
  - Responses to this suggested practice ranged more widely, with some perceiving it as irrelevant for their work and others foreseeing significant or even insurmountable obstacles. Lack of time was identified as the most important barrier. One participant requested a specific recommendation for a good text-to-speech program.

- **Design data storage well by minimizing the number of times an item must be entered, using consistent terminology, and atomizing data.**
  - Participants were approximately equally split between already engaging in this practice and foreseeing only minor obstacles. Lack of time was identified as the most important barrier, but several participants also mentioned needing more training and/or practice in order to be able to implement this well.

- **Document changes to data.**
  - Some (6 of 15) of the participants reported already doing this and the remainder were nearly equally split between foreseeing minor or significant but surmountable obstacles. Lack of time and training were identified as the most important barriers. Several participants commented on the difficulty of doing this well, and one participant mentioned difficulty with this when there are several people working with the data at the same time. Another mentioned that it is difficult to justify creating so much documentation when working with a data set for a couple of weeks just to determine whether or not it is ‘good’.

- **Visually inspect data table for improper column line-up and missing, impossible, or anomalous values.**
  - Nearly all participants reported already doing this, and the remaining one foresaw only minor obstacles to implementation. Two participants expressed interest in learning more efficient ways to do this, especially when dealing with data sets containing millions of records.

- **Look for outliers by performing statistical summaries and creating data visualizations.**
  - Most (10 of 15) participants reported that they already do this, and 4 foresaw only minor obstacles. One participant expressed interest in learning ways to do this more efficiently, and another remarked that manual creation takes time (and therefore funding) while creating dynamic reporting functions requires training and technical support. Another participant mentioned philosophical issues about what to do with outliers.
Obstacles to adopting best practices for analysis and workflow documentation

Responses to the questions regarding recommended best practices to support analysis and workflow documentation are summarized in Figure 19. Overall, participants perceived mostly minor obstacles to adoption of these practices, with some participants reporting that they already engage in the practices and few foreseeing significant obstacles. Participants pointed out a need for 1) time-saving tools and practices that are accessible to non-programmers and 2) additional training in how to create and manage workflow documentation using available tools in order to make these practices feasible.

With respect to perceived obstacles to adoption of best practices, the lesson on Analysis & Workflows could be improved by:

- increased focus on time-saving tools and approaches to make scientific workflow and data provenance documentation more efficient;
- inclusion of information about resources that participants can use to acquire more training related to workflow and provenance documentation;
- inclusion of practical advice on how to set up a system for updating workflow documentation as a project progresses, especially when there are multiple people working on the project;
- inclusion of more information about (more user-friendly) alternatives to Kepler that can be used to document workflows and data provenance.

A more detailed summary of these responses and participant comments in response to specific practices advocated are summarized below:

- **Create scientific workflow documentation that includes information about data provenance.**
  - Very few (2 of 15) of the participants reported that they already do this, while 6 foresaw only minor obstacles and 5 others foresaw significant but surmountable obstacles. One participant perceived insurmountable obstacles to adopting this practice, and mentioned that complicated data and model made it difficult to see where to even begin in documenting provenance. Other participants commented on the difficulty of updating workflows as the analysis progresses and the difficulty in learning to use available tools without in-depth training. Lack of time, training, and adequate technical support were all identified as barriers to adoption of this practice.

- **Create scientific workflow documentation that includes information about analyses and parameters used.**
  - Approximately half (8 of 15) of participants reported that they already engage in this practice, while the remaining participants were nearly evenly split between foreseeing minor obstacles and foreseeing significant but surmountable obstacles. Those who commented on this
practice liked the idea of commenting in code to implement this practice, but one participant felt that a more systematic approach was necessary in order to do this well. Lack of time and training were perceived as equally important barriers, and lack of technical support to a lesser degree.

- **Create scientific workflow documentation that includes information about connections between analyses via inputs and outputs.**
  - Nearly half (7 of 15) of participants reported that they already do this, while most of the remaining participants foresaw only minor obstacles to implementation. Two foresaw significant but surmountable barriers. Lack of time and training were considered the most important obstacles to implementation.

- **Create informal (e.g., flowchart) documentation for all analyses.**
  - Approximately half of the participants reported already engaging in this practice, while the remaining half foresaw only minor obstacles. Lack of time was seen as the most important obstacle, and lack of training a less important obstacle.

- **Create formal (e.g., Kepler) documentation for all analyses.**
  - A few (3 of 15) of the participants reported that they already engage in this practice, and an equal number foresaw only minor obstacles to doing so. The other participants foresaw either significant but surmountable obstacles (5 of 15) or insurmountable obstacles (4 of 15). Several participants reported being put off by the apparently steep learning curve for Kepler and for the lack of documentation/training materials that make using it accessible to non-programmers. Lack of training was considered the most important barrier to implementation, followed by lack of time. Lack of tech support and lack of funding were perceived as significant but less important obstacles.

**Obstacles to adopting best practices for metadata creation**

Responses to the questions regarding recommended best practices to support metadata creation are summarized in Figure 20 and Figure 21. Overall, most participants perceived minor obstacles to adoption of these practices, with some participants reporting that they already engage in the practices and few foreseeing significant obstacles. Participants pointed out a need for 1) agreed-upon standards and 2) more time to dedicate to metadata creation in order to make these practices feasible.

With respect to perceived obstacles to adoption of best practices, the lesson on Metadata could be improved by:

- increased focus on ways to minimize the time required to implement these practices;
• more information about the most widely-accepted metadata standards for specific fields as well as tools available for crosswalking these standards;
• inclusion of information about how to find someone who is qualified and willing to review metadata records for accuracy and completeness if such individuals are not accessible to the project or the practice is counter to the culture of the research team;
• more information about different types of data limitations that should be included as part of a metadata record and strategies that can be used to identify these limitations;
• inclusion of more practical guidelines and examples that outline and illustrate the levels of comprehensiveness and specificity that should be included as part of a metadata record; and
• inclusion of more information about 1) finding thesauri appropriate for different research fields and 2) why the use of such keywords is such an important element of metadata creation.

A more detailed summary of these responses and participant comments in response to specific practices advocated are summarized below:

• **Organize your information before beginning to create your metadata record.**
  o A few (4 of 15) of the participants reported that they already engage in this practice, while most of the others (9 of 15) perceived only minor obstacles and two perceived significant but surmountable obstacles. Several participants noted that significant time is required to gather necessary information, and one participant expressed concern about the amount of time that would be necessary to dig through a number of sources and ask the appropriate people in order to acquire all of the information necessary.

• **Write metadata using a metadata creation tool.**
  o A few (3 of 15) of the participants reported that they already engage in this practice, while most of the others (10 of 15) perceived only minor obstacles and two perceived significant but surmountable obstacles. Lack of time was identified as the most important barrier to implementation of this practice, and one participant mentioned that getting a lab to agree to consistently use the same standard is difficult and that things get even more complicated when sending data to collaborators who use different metadata standards. Again, lack of time was perceived as the most important obstacle to implementation of this practice.

• **After creating a metadata record, review your record for accuracy and completeness.**
  o A few (4 of 15) of the participants reported that they already engage in this practice, while most of the others (10 of 15) perceived only minor obstacles and one perceived significant but surmountable obstacles. Several participants commented on the essential nature of this practice.
Lack of time was often identified as the primary obstacle to implementation of this practice.

- **After creating a metadata record, gain feedback on that record from a reviewer.**
  - A few (3 of 15) of the participants reported that they already engage in this practice, while most of the others (9 of 15) perceived only minor obstacles and three perceived significant but surmountable obstacles. Lack of time was once again identified as a major factor, and several participants expressed concern about being able to find a reviewer who was both experienced with metadata creation and willing to invest time to review. One participant mentioned that this practice is not compatible with the culture of his/her work group.

- **Create metadata records that avoid jargon, acronyms, and symbology.**
  - Many (6 of 15) of the participants reported that they already engage in this practice, while all of the others (9 of 15) perceived only minor obstacles. Several participants commented on the power of habit and/or the inability on the part of researchers to recognize their own jargon as obstacles to implementing this practice.

- **Create metadata records that clearly state data limitations.**
  - A few (4 of 15) of the participants reported that they already engage in this practice, while most of the others (8 of 15) perceived only minor obstacles and three perceived significant but surmountable obstacles. One participant reported that the lesson did not make it clear how to do this. Others pointed to the large amounts of time necessary to 1) think ahead about all of the contingencies, 2) discover limitations of data they did not collect from studies they did not design, and 3) agree upon policies with collaborators. Lack of time and, to a lesser degree, training were identified as most important obstacles to implementation of this practice.

- **Use descriptive titles when creating metadata records.**
  - Many (6 of 15) of the participants reported that they already engage in this practice, and all of the others (9 of 15) perceived only minor obstacles. One participant commented that it would probably help to standardize the order of the where/what/when/etc. information in the title.

- **When writing a metadata record, include comprehensive and specific information about the characteristics of your data.**
  - A few (4 of 15) of the participants reported that they already engage in this practice, while most of the others (8 of 15) perceived only minor obstacles and three perceived significant but surmountable obstacles. A couple of participants mentioned the importance of learning more about how to do this, especially regarding use of domain ontologies, standard units, standard taxonomic nomenclature, etc. Lack of time was seen as the most important obstacle, followed by lack of training.
• **Select keywords for your metadata record using a thesaurus (e.g., the USGS Biocomplexity Thesaurus).**
  
  o Few (2 of 15) of the participants reported that they already engage in this practice, while most of the others (11 of 15) perceived only minor obstacles and two perceived significant but surmountable obstacles. Many of the participants noted the need to find a thesaurus suitable for their discipline, and expressed concern about the amount of time it would take to do this. One participant wondered about how important this practice really is. Lack of training and time were identified as the most important obstacles to adoption of this practice.

**Obstacles to adopting best practices for data citation**

Responses to the questions regarding recommended best practices to support data citation are summarized in Figure 22. Overall, most participants perceived minor or significant obstacles to adoption of these practices, with some participants reporting that they already engage in some of the practices and one foreseeing insurmountable obstacles to some practices. Participants pointed out a need for 1) agreed upon standards, 2) institutionalization of procedures for implementing these practices, and 3) funding to cover the time required in order to make these practices feasible.

With respect to perceived obstacles to adoption of best practices, the lesson on Data Citation could be improved by:

• inclusion of more information about specific metadata standards and tools that allow for generation of metadata that satisfy requirements for obtaining a DOI or other identifier;
• more focus on the details of how to go about getting a DOI or other identifier (e.g., which institution to contact and what to expect);
• addition of information about the point at which a data set is ready for a DOI and about when new DOIs should be requested for revised versions (if at all);
• more information about what to expect (e.g., sequence of events, time required) when coordinating publication of data and manuscripts with publishers and repositories, and suggestions regarding publications and repositories that help to facilitate this process; and
• discussion of cultural, institutional, and financial barriers to implementation of these practices.

A more detailed summary of these responses and participant comments in response to specific practices advocated are summarized below:

• **Use a software application that supports metadata creation for environmental data sets.**
  
  o Some (6 of 15) of the participants reported that they already do this, and an equal number reported foreseeing only minor obstacles and two
foresaw significant but surmountable obstacles. A couple of participants mentioned the difficulty in choosing an appropriate standard and tool and then learning to use them. Lack of training was identified as the most important obstacle to implementation of this practice, followed by lack of time and tech support.

- **When preparing metadata, use standardized keywords to describe your data.**
  - Some (5 of 15) of the participants reported already engaging in the practice, while approximately half (8 of 15) reported foreseeing only minor obstacles and one reported significant but surmountable obstacles. Several participants identified choosing an appropriate standard as an important obstacle. Lack of training and lack of time were identified as the most important barriers to implementation of this practice.

- **Obtain a persistent identifier such as a DOI or ARK for each dataset you create.**
  - Most of the participants foresaw minor (5 of 15) or significant but surmountable (6 of 15) obstacles to adopting this practice, and one foresaw insurmountable obstacles. Only one reported already engaging in this practice. Several participants mentioned that they didn’t feel like they really knew how to go about doing this after this lesson, and one mentioned the complication of choosing a metadata standard that will generate the metadata necessary to obtain a DOI. One participant expressed uncertainty about the point at which a dataset is mature enough to warrant obtaining a DOI and how to decide how often new versions of data sets should obtain new DOIs, if at all. Lack of training was identified as the most important obstacle, but lack of time, tech support, and funding were also considered important barriers.

- **Use a persistent identifier such as a DOI or ARK (when available) when citing datasets.**
  - Many (8 of 15) participants reported foreseeing only minor obstacles to adoption of this practice, while three foresaw significant but surmountable obstacles and two already engage in the practice. Lack of training was again considered to be the most important barrier, followed by lack of time and tech support.

- **Work with journal publishers and data repositories to archive data during the publication process.**
  - Most participants foresaw either minor obstacles (4 of 15) or significant but surmountable obstacles (5 of 15) to implementing this practice. Two reported already doing this, and one foresaw insurmountable obstacles. Participants expressed uncertainty and/or concern about 1) how to choose a publisher that supports this practice, 2) the amount of time that would be required for this level of communication, 3) lack of agreed-upon standards and systems, 4) journals that require data to be submitted to proprietary databases, and 5) the fact that this seems to be a work in progress that is not practical at this time for researchers. Lack of training,
tech support, and time, and funding were all identified as important obstacles.

- **Encourage other data authors to cite data and to make their own data available for reuse.**
  - Some (5 of 15) of the participants reported that they already do this, while many foresee minor (4 of 15) or significant (4 of 15) obstacles to implementation of the practice. One participant foresaw insurmountable obstacles. One participants mentioned cultural barriers to acceptance of this practice, while another reported working at an institution that supports the practice but has implemented it as an unfunded mandate. Participants identified lack of time as the most important barrier to this practice.

**Obstacles to adopting best practices for protected backups and data preservation**

Responses to the questions regarding recommended best practices to support backups and data preservation are summarized in Figure 23. Overall, most participants reported already engaging in these practices or perceived mostly minor obstacles to adoption of these practices.

With respect to perceived obstacles to adoption of best practices, the lesson on Protected Backups & Data Preservation could be improved by:

- increased focus on ways to minimize the time required to implement these practices;
- addition of details regarding how to design a backup policy for research projects within which each user is expected to be responsible for their own backups;
- inclusion of information about tools and/or strategies that can be used to automatically convert proprietary files into non-proprietary formats for backup purposes;
- inclusion of information about tools and/or strategies that can be used to efficiently perform manual checks for large numbers of files;
- addition of information about tools and/or strategies for automating backups to several different locations/formats so that users don’t have to rely on their memories to manually initiate backups to remote locations.

A more detailed summary of these responses and participant comments in response to specific practices advocated are summarized below:

- **Create a backup policy and review it periodically to ensure it is still valid and applicable.**
  - Some (5 of 15) of the participants reported that they already do this, and most others (8 of 15) reported foreseeing only minor obstacles. One foresaw significant but surmountable obstacles. One participant mentioned that creation of a backup policy would be easy, while enforcement would be much more of a struggle. Lack of time was
identified as the most important barrier to implementation of this practice.

• **Minimize or remove reliance on users to perform their own manual backups.**
  
  o Some (5 of 15) reported that they already do this, while 3 others foresaw only minor obstacles and four foresaw significant but surmountable obstacles. Two participants indicated this this practice was not applicable to their work. One participant mentioned that it is actually impossible to totally automate backups (e.g., because people need to make sure to keep backup drives plugged in and turned on), and another mentioned that it is very common for backup to be a user responsibility, but advocated creating a document describing these responsibilities so that practices are standardized. Lack of time and tech support were identified as the most important barriers to implementation of this practice, followed by lack of training.

• **Don’t assume backups are being performed for you.**
  
  o Most (11 of 15) participants reported that they already do this, while two foresaw only minor obstacles and one foresaw significant but surmountable obstacles. Lack of tech support and training were identified as barriers to implementation.

• **Use non-proprietary, standard formats for storing data and metadata.**
  
  o Some (5 of 15) participants reported that they already do this, and an equal number foresaw only minor obstacles. Three participants foresaw significant but surmountable obstacles, and one foresaw insurmountable obstacles. One participant wondered if there are tools available to convert common proprietary formats to non-proprietary for backup purposes. Lack of time was identified as the most important obstacle, followed by lack of training.

• **Check your backups manually.**
  
  o Some (6 of 15) participants reported that they already engage in this practice, while approximately the same number foresaw only minor obstacles to implementation. One foresaw significant but surmountable obstacles. Several participants mentioned the difficulty of doing this when backing up thousands of files on a regular basis, and expressed interest in tools to make such checking possible. Lack of time and training were identified as the most important obstacles.

• **Keep multiple versions of backups on multiple formats in multiple places.**
  
  o Most (9 of 15) participants reported that they already do this, while the remaining participants reported foreseeing either minor (3 of 15) or significant (2 of 15) obstacles. One mentioned the difficulty in remembering to update backups located in more remote places. Lack of time was identified as the most important barrier to implementation of this practice.
Figure 16: Obstacles to best practices for data sharing and contributing factors

**Obstacles to Best Practices for Data Sharing**

As part of the lesson on Data Sharing, we suggested the following as best practices. Do you foresee obstacles to implementing these best practices in your work? How difficult do you think it will be to overcome these obstacles?

<table>
<thead>
<tr>
<th>Obstacle Description</th>
<th>Below Group Mean</th>
<th>Above Group Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>I foresee insurmountable obstacles.</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>I foresee significant but surmountable obstacles.</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>I foresee only minor obstacles.</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>I am already doing this.</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>This best practice is not applicable to my work.</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

**Which of the following factors are contributing to the obstacles you foresee?**

- **Document and publish data using standards.**
- **Promote data use via presentations and meetings.**
- **Solicit feedback from data users and address identified issues.**
- **Monitor publications and websites for data use and address misapplications.**

Factors: time, training, tech support, funding.
Figure 17: Obstacles to best practices for data collection, entry and manipulation and contributing factors

**Obstacles to Best Practices for Data Collection, Entry & Manipulation**

As part of the lesson on Data Collection, Entry & Manipulation we suggested the following as best practices. Do you foresee obstacles to implementing these best practices in your work? How difficult do you think it will be to overcome these obstacles?

<table>
<thead>
<tr>
<th>Obstacle Description</th>
<th>pre-course DM practices score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>below group mean</td>
</tr>
<tr>
<td>I foresee insurmountable obstacles.</td>
<td>2</td>
</tr>
<tr>
<td>I foresee significant but surmountable obstacles.</td>
<td>1</td>
</tr>
<tr>
<td>I foresee only minor obstacles.</td>
<td>3</td>
</tr>
<tr>
<td>I am already doing this.</td>
<td>13</td>
</tr>
<tr>
<td>This best practice is not applicable to my work.</td>
<td>2</td>
</tr>
</tbody>
</table>

- Enter all data into one table rather than multiple small tables.
- Enter one type of data per column and use consistent names, codes, and formats for the data in each column.
- Create descriptive column names without spaces or special characters.
- Use descriptive file names.
- Consistently mark missing data by leaving fields empty (NULL), using a distinct value to indicate missing data, or using data flags in a separate column.
- Store data in a consistent format that can be read and used by any application now or in the future.

Which of the following factors are contributing to the obstacles you foresee?

- time
- training
- tech support
- funding

![Graph showing the contribution of factors to obstacles]
Obstacles to Best Practices for Quality Control / Quality Assurance

As part of the lesson on Quality Control / Quality Assurance, we suggested the following as best practices. Do you foresee obstacles to implementing these best practices in your work? How difficult do you think it will be to overcome these obstacles?

<table>
<thead>
<tr>
<th>I foresee insurmountable obstacles.</th>
<th>I foresee significant but surmountable obstacles.</th>
<th>I foresee only minor obstacles.</th>
<th>I am already doing this.</th>
<th>This best practice is not applicable to my work.</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="chart1.png" alt="Chart" /></td>
<td><img src="chart2.png" alt="Chart" /></td>
<td><img src="chart3.png" alt="Chart" /></td>
<td><img src="chart4.png" alt="Chart" /></td>
<td><img src="chart5.png" alt="Chart" /></td>
</tr>
</tbody>
</table>

Define and enforce standards for formats, codes, measurement units, and metadata.
Assign responsibility for data quality.
Use double entry, transcription from recorded reading, a text-to-speech program to ensure quality during data entry.
Design data storage well by minimizing the number of times an item must be entered, using consistent terminology, and atomizing data.
Document changes to data.
Visually inspect data table for improper column line-up and missing, impossible, or anomalous values.
Look for outliers by performing statistical summaries and creating data visualizations.

Which of the following factors are contributing to the obstacles you foresee?

- time
- training
- tech support
- funding
Figure 19: Obstacles to best practices for analysis and workflows and contributing factors

Obstacles to Best Practices for Analysis & Workflows

As part of the lesson on Analysis and Workflows we suggested the following as best practices. Do you foresee obstacles to implementing these best practices in your work? How difficult do you think it will be to overcome these obstacles?

Create scientific workflow documentation that includes information about data provenance.

Create scientific workflow documentation that includes information about analyses and parameters used.

Create scientific workflow documentation that includes information about connections between analyses via inputs and outputs.

Create informal (e.g., flowchart) documentation for all analyses.

Create formal (e.g., Kepler) documentation for all analyses.

Which of the following factors are contributing to the obstacles you foresee?
Figure 20: Obstacles to best practices for writing metadata and contributing factors

Obstacles to Best Practices for Writing Metadata

As part of the lesson on How to Write Quality Metadata we suggested the following as best practices. Do you foresee obstacles to implementing these best practices in your work? How difficult do you think it will be to overcome these obstacles?

<table>
<thead>
<tr>
<th>Obstacle</th>
<th>I foresee insurmountable obstacles.</th>
<th>I foresee significant but surmountable obstacles.</th>
<th>I foresee only minor obstacles.</th>
<th>I am already doing this.</th>
<th>This best practice is not applicable to my work.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organize your information before beginning to create your metadata record.</td>
<td>9</td>
<td>10</td>
<td>10</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Write metadata using a metadata creation tool.</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>After creating a metadata record, review your record for accuracy and completeness.</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>After creating a metadata record, gain feedback on that record from a reviewer.</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>Create metadata records that avoid jargon, acronyms, and symbology.</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>9</td>
<td>9</td>
</tr>
</tbody>
</table>

Which of the following factors are contributing to the obstacles you foresee?

- time
- training
- tech support
- funding
Obstacles to Best Practices for Writing Metadata (cont.)

As part of the lesson on How to Write Quality Metadata we suggested the following as best practices. Do you foresee obstacles to implementing these best practices in your work? How difficult do you think it will be to overcome these obstacles?

Which of the following factors are contributing to the obstacles you foresee?

- Create metadata records that clearly state data limitations.
- Use descriptive titles when creating metadata records.
- When writing a metadata record, include comprehensive and specific information about the characteristics of your data.
- Select keywords for your metadata record using a thesaurus (e.g., the USGS Biocomplexity Thesaurus).
Figure 22: Obstacles to best practices for data citation and contributing factors

Obstacles to Best Practices for Data Citation

As part of the lesson on Data Citation we suggested the following as best practices. Do you foresee obstacles to implementing these best practices in your work? How difficult do you think it will be to overcome these obstacles?

<table>
<thead>
<tr>
<th>Obstacle Description</th>
<th>pre-course DM practices score</th>
</tr>
</thead>
<tbody>
<tr>
<td>I foresee insurmountable obstacles.</td>
<td></td>
</tr>
<tr>
<td>I foresee significant but surmountable obstacles.</td>
<td></td>
</tr>
<tr>
<td>I foresee only minor obstacles.</td>
<td></td>
</tr>
<tr>
<td>I am already doing this.</td>
<td></td>
</tr>
<tr>
<td>This best practice is not applicable to my work.</td>
<td></td>
</tr>
<tr>
<td>N/A</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Factor Description</th>
<th>Percentage of Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use a software application that supports metadata creation for environmental data sets.</td>
<td>25%</td>
</tr>
<tr>
<td>When preparing metadata, use standardized keywords to describe your data.</td>
<td>50%</td>
</tr>
<tr>
<td>Obtain a persistent identifier such as a DOI or ARK for each dataset you create.</td>
<td>75%</td>
</tr>
<tr>
<td>Use a persistent identifier such as a DOI or ARK (when available) when citing datasets.</td>
<td>50%</td>
</tr>
<tr>
<td>Work with journal publishers and data repositories to archive data during the publication process.</td>
<td>25%</td>
</tr>
<tr>
<td>Encourage other data authors to cite data and to make their own data available for reuse.</td>
<td>10%</td>
</tr>
</tbody>
</table>

Which of the following factors are contributing to the obstacles you foresee?

- time
- training
- tech support
- funding
Figure 23: Obstacles to best practices for data backups and preservation and contributing factors

Obstacles to Best Practices for Data Backups and Preservation

As part of the lesson on Protected Backups and Data Preservation we suggested the following as best practices. Do you foresee obstacles to implementing these best practices in your work? How difficult do you think it will be to overcome these obstacles?

Which of the following factors are contributing to the obstacles you foresee?
Reflection and Future Plans

The Data Management Short Course and evaluation reported on here was successful in achieving two principal goals. First, participants in the workshop reported high levels of satisfaction with the quality of the workshop and the knowledge they gained regarding best data management practices. Second, the developers and presenters of the DataONE Data Management Lessons received valuable feedback about the curriculum and its presentation in a workshop format. The CEE Working Group plans both to incorporate this feedback into revisions to the Data Management Lessons, and to share tips and lessons learned with members of other working groups who are developing data management education projects.