Project idea: Collecting ‘real-life’ success stories and cautionary tales for data management engagement and education

Goal: Collect stories to illustrate data management concepts and best practices

Background:
A consistent theme that has emerged both through the evaluation of the CEE education modules and through the work of other working groups is the need for stories related to data management concepts and best practices. A collection of narratives of this type could serve the goals of DataONE in a variety of ways:

- used as common threads through the education products created by DataONE
- serve as the basis for blog posts to enhance DataONE’s social media presence
- function as starting points for the success stories requested by the NSF.

We propose gathering such stories from people who represent DataONE’s primary stakeholder groups through focus groups and interviews using the set of questions below.

Project management:
Under direction of CEE working group leaders Stephanie Hampton and Viv Hutchison, Stacy Rebich Hespanha will coordinate this project. Anticipated coordination tasks include:

- Incorporate feedback from CEE working group members and the DataONE leadership team into project documentation (such as interview protocol and questions)
- Prepare materials to be submitted to the Institutional Review Board for approval to conduct research with human subjects
- Serve as primary contact person for those participating in the project
- Prepare and distribute documents and email templates that interviewers will use as they recruit and interview researchers (e.g., steps 1b-d and 2a of the protocol below)
- Collect audio recordings from interviewers and coordinate with project intern to convert the audio files to a suitable text format
- Coordinate responsibility for developing raw text into polished narratives with other members of the CEE working group
- Work with interested members of the leadership team to finalize completed narratives for posting on the web (e.g., the DataONE coffeehouse blog)
- Coordinate with project intern to provide periodic project updates in the form of a research blog
- Communicate with other DataONE working groups about the available narratives once they have been posted
- Draft a manuscript summarizing key findings of the study in a format suitable for publication

What we are asking:
To make this project feasible, we will need time commitments (on the order of 3-5 hours) from

CEE group members and members of the leadership team who will lead focus groups or conduct interviews. We also anticipate need for a graduate or undergraduate intern who will help with gathering audio recordings and synthesizing text narratives from these recordings. We will need to seek IRB approval through the institution that will be responsible for preserving the data collected through these interviews.

The following pages contain an overview of the anticipated data collection protocol and a set of eliciting questions that could be used to guide the interviews.

Interview protocol overview

1. Planning for interviews
    a. Interviewer peruses the set of target themes and questions included in the list below and selects several topics and/or questions with which to work
    b. Interviewer coordinates selected areas of focus with other interviewers (e.g., through an online document) to ensure that all topics and questions are used by the interviewers as a group.
    c. Interviewer arranges a meeting with a researcher or researchers, either one-on-one or as a small group. When setting up the meeting, interviewer lets potential interviewees know that the meeting 1) is part of a project to collect stories about researchers' experiences with data management, 2) will be focused on specified topics (selected by the interviewer in Steps 1a and 1b), 3) will be recorded (audio), and 4) will involve a request for interviewees' consent to allow their stories to be used anonymously as part of a research project. (*Email template containing this basic information will be prepared by the CEE WG and distributed for use by all interviewers.*)
    d. Interviewer sends the consent form and copies of the questions to be discussed to interviewees in advance of the meeting to allow interviewees time to think about their stories before the interview starts. (*The CEE WG will assist with preparing these documents for each interviewer.*)
    e. Interviewer acquires, sets up, and tests audio recording equipment to be used during the interviews. (*Note: recordings need to be of sufficiently high quality to be transcribed; quality recording equipment is especially important when interviewing more than one person at a time.*)
2. During interviews
    a. Interviewer distributes paper copies of the consent forms and asks interviewees to read and sign them if they are willing to participate in the research project.
    b. Interviewer distributes paper copies of the questions and gives interviewees a chance to read through them again before the interview starts.
    c. If interviews will be conducted in a group setting, interviewer gives interviewees time to talk about the questions (and their answers to them) with another interviewee. *(This step is important when collecting data in a group setting because it allows participants to become more comfortable speaking in the group setting, and it also gives them a chance to practice telling their stories so that they tell them more fluently during the recorded portion of the interview.)*
    d. Interviewer regains the attention of the group and lets them know that recording will begin.
    e. Interviewer selects (or lets interviewee select) a question to start with and begins interview session.
        i. If one or more interviewees have a story to tell related to that question, interviewer uses follow-up questions (listed beneath main questions and/or generated by the interviewer) to elicit more story details.

ii. If interviewee(s) does not have a story in response to the question, interviewer moves on to another question.

iii. During the interview, the interviewer focuses on asking follow-up questions that elicit a level of detail that make the stories ‘real’ and interesting. This may include both suggested follow-up questions from the list and interviewer-generated questions that cover the who, what, where, when, and why of the story.

f. When stories and/or questions and/or time available are exhausted, interviewer thanks interviewees and lets them know that they will receive information about the data management story project once the narratives have been compiled and prepared for distribution.

3. After interviews

a. Interviewer provides the project coordinator (Stacy Rebich Hespanha) with the audio recordings from the interviews he/she conducted.

b. Project intern transcribes or summarizes audio recordings and the CEE WG uses these summaries/transcripts as the basis for creating written story narratives.

c. When project narratives are made public on the DataONE website, interviewers notify their interviewees that the project is complete and that the narratives are available online.

Question Ideas

- Data management planning
    - Have you ever written a data management plan?
        - If so, what difficulties did you encounter when writing your first plan?
        - Did you use any information resources or get help from any people during this process?
        - What do you think could have helped make the process easier?
        - Has the process of writing a data management plan influenced the way you manage your data? If so, how?
        - Are you now confident about your ability to craft a quality data management plan, or do you feel that you need to continue to improve your skills in this area?
            - In which ways could your data management planning skills be improved?
            - What prevents you from gaining the needed skills you've identified?
- QA/QC + Analysis
    - Can you remember a situation in which your original analyses were flawed because of some factor that you had not taken into account (an error in the data, one or more outliers, a latent variable)?
        - How did you go about addressing this problem?
        - How did you document the changes you made to your data inputs or analytic steps?
        - What could you have done differently to either avoid this problem or deal with it more efficiently?
        - Has the process of correcting a flawed analysis influenced the way you manage or document your data? If so, how?
        - Are you happy with the quality assurance/quality control routines you currently use, or would you like to improve them in some way?
            - In which ways could they be improved?
            - What prevents you from making the improvements you've identified?
- Protecting data
    - Have you ever lost data?
        - How did the data loss happen (e.g., stolen computer, hardware or software failure, obsolete media)?
        - What did you lose?
        - Were you able to reproduce any of the data you lost?
        - Have you changed any of your data backup or archiving practices based upon this experience? If so, how?
        - Are you happy with the data protection measures you have already taken, or would you like to improve your data protection system in some way?
            - In which ways could it be improved?

- What prevents you from making the improvements you've identified?

  - If a storm or fire were to damage your office or home today, what data would you lose?
    - Would it be possible to recover the data lost due to such an event?
    - Would you be able to replicate data you were unable to recover?
      - If so, how much time would it take you to replicate this data?
    - Are you happy with the data protection measures you currently have in place, or would you like to improve your data protection system or practices in some way?
      - In which ways could they be improved?
      - What prevents you from making the improvements you've identified?
- Data and/or workflow organization
  - Have you ever been in a situation where you had trouble finding your own data or outputs?
    - How did you resolve the problem?
    - What could you have done differently to prevent this?
    - Have you changed anything about the way you store or annotate data as a result of this incident?
    - How likely is it that you will encounter a situation like this in the future? Will you be more prepared to deal with the situation due to changes you have made in your data storage or annotation practices? If so, please describe.
    - Are you happy with the data and workflow management system you currently have in place, or would you like to improve your system or practices in some way?
      - In which ways could they be improved?
      - What prevents you from making the improvements you've identified?
  - Do you know how to access your data backups?
    - Please describe the procedure you would follow to retrieve a file from backup.
    - Would the process for retrieving an old file (e.g., several years old) be different from the process of retrieving a more recent file (e.g., from a month ago)?
      - How long would it take you to retrieve an older file?
      - How long would it take you to retrieve a more recent file?
    - Are you happy with the backup system you have now, or would you like to improve it in some way?
      - In which ways could it be improved?
      - What prevents you from making the improvements you've identified?

  - Have you ever had a journal editor or reviewer request that you provide additional information on how you conducted an analysis, or that you revise an analysis?
    - If so, given that an extended time period had ensued while the paper was reviewed, how easy was it for you to respond to their requests?
    - How much time did you spend preparing the information they had requested?
      - What tasks took up most of your time in preparing your response/re-analysis?
      - What could you have done differently to make this process more efficient?
    - Have you changed any of your data management practices based upon this experience? If so, how?
    - Are you happy with the data management system you already have, or would you like to improve your system in some way?
      - In which ways could it be improved?
      - What prevents you from making the improvements you've identified?
  - Have you ever been called into your advisor's office/called into court/contacted by a scientist questioning your results?
    - What did you do to justify your results?
    - How much time would it take you to prepare the information necessary?
    - Could you have done anything differently to make this process more efficient? If so, please describe.
    - Have you changed any of your data management practices based upon this experience? If so, how?
- Communicating about data/workflows
  - Have you ever been in a situation where you had difficulty communicating to someone else what was contained/represented in one of your own datasets?
    - What factors were responsible for the communication difficulty?
    - How did you resolve the problem?
    - What could you have done differently to facilitate communication of this sort?
    - Have you changed anything about the way you store or annotate data as a result of the communication difficulties you experienced?
    - Are you happy with the data annotation system you already have, or would you like to improve your system in some way?
      - In which ways could it be improved?
      - What prevents you from making the improvements you've identified?
  - Have you ever been in a situation where you had difficulty communicating with someone else about the steps in your analysis?
    - What factors were responsible for the communication difficulty?

    - How did you resolve the problem?
    - What could you have done differently to facilitate communication of this sort?
    - Have you changed anything about the way you document your workflow as a result of the communication difficulties you experienced?
    - Are you happy with the workflow documentation system you already have, or would you like to improve your system in some way?
      - In which ways could it be improved?
      - What prevents you from making the improvements you've identified?
  - Have you ever experienced downtime due to loss of a team member (e.g., programmer, data analyst)?
    - How much time was lost?
    - How did you resolve the issue?
    - What could you have done differently to avoid this downtime?
    - Have you changed anything about the way you manage your team or workflow due to the difficulties you experienced?
    - Are you happy with the research team management system you already have, or would you like to improve your system in some way?
      - In which ways could it be improved?
      - What prevents you from making the improvements you've identified?
  - Have you ever managed or worked on a project that involved transfer of data from one team member to another (either while team members working concurrently or when one team member passes responsibility for a project on to another)?
    - Did you experience any data loss, misunderstandings, or communication difficulties while working on or managing this project? If so, please describe.
      - Have you made any changes to your procedures as a result of these difficulties?
    - Are you happy with the data transfer/sharing system you already have, or would you like to improve your system in some way?
      - In which ways could it be improved?
      - What prevents you from making the improvements you've identified?
- Metadata creation, curation, and use
  - Have you ever created a formal metadata record to describe one of your datasets or workflows?
    - What metadata standard did you choose, and how did you decide to use that one?
    - Did you use any information resources or get help from any people during this process?

        - How much time did you spend completing this process?
        - Could you have done something differently to make this process easier and/or less time consuming?
        - Could additional tools/resources be made available to make this process easier?
    - Have you ever located a metadata record in a clearinghouse or repository?
        - What motivated you to do this?
        - How much time and effort did you spend locating the metadata record?
        - Could you have done something differently to make this process easier and/or less time consuming?
        - Could additional tools/resources be made available to make this process easier?
        - Did you go on to use the data set associated with that metadata record in a research project?
        - Did the metadata record help you to understand the data? If so, please describe.
        - Would you have been able to use the data without the metadata record? If not, why not?
    - Have you ever ‘crosswalked’ a metadata record (converted a metadata record from one standard format to another)?
        - What motivated you to do this?
        - How did you go about doing it?
        - How much time did you spend completing this process?
        - Could you have done something differently to make this process easier and/or less time consuming?
        - What benefits did you gain from having crosswalked the record?
- Data sharing
    - Have you ever deposited/archived one of you datasets in a data repository?
        - How did you go about choosing the appropriate repository for your data?
        - Did you use any information resources or get help from any people during this process?
        - How much time did you spend completing this process?
        - Could you have done something differently to make this process easier and/or less time consuming?
    - Have you ever had someone use data you have shared publicly?
        - If so, how did you find out your data set was being used?
        - Did you have any contact with the data user prior to publication of the study? after publication?
        - Have you ever been made co-author of a paper based primarily on your data contribution?
        - Have you benefited professionally in any way due to sharing data? If so, please describe.
- Acquiring data for reuse

    - Is there a data set you knew was out there, but couldn't find? Or couldn't use once you found it?
        - How did this affect the project you were working on? Were you able to find a solution that allowed you to continue with the project, or did you need to abandon the project?
        - If this data had been available and useable, how would it have benefitted your project?
    - Have you ever searched for data in a public or institutional repository?
        - Did you encounter any difficulty in finding the data you were looking for? If so, please describe.
        - How did you resolve these difficulties?
        - Did you use any information resources or get help from any people during this process?
        - What could you have done differently to make your search more efficient and/or successful?
        - What features could the repository have offered to make your search easier and/or more successful?
    - If you have used data from a repository, have you had any interaction with the creator of those data? Please describe any such interactions you have had.
- Using other people's data
    - Have you ever used data that you did not collect yourself?
        - If so, how did you go about acquiring this data?
        - How much time did you invest in getting the data you were interested in?
        - Could you have done something differently to make this process easier and/or less time consuming?
        - Could additional tools/resources be made available to make this process easier? If so, please describe the types of tools or resources that you think would be useful.
    - If you have downloaded or otherwise acquired data collected by someone else, what processing was needed to make it useable for your purpose?
        - If so, what steps were required?
        - Were you able to alter the data as needed?
        - Did you use any information resources or get help from any people during this process? If so, please describe.
        - How long did the process take from start to finish?
        - What could you have done differently to make the data clean-up process more efficient?
        - Could additional tools/resources be made available to make this process easier? If so, please describe the types of tools or resources that you think would be useful.
    - Have you ever integrated several datasets to produce a new dataset?
        - What were the challenges you encountered and how did you deal with them?

    - Were you able to integrate the data as needed?
    - Did you use any information resources or get help from any people during this process? If so, please describe.
    - How long did the integration process take from start to finish?
    - What could you have done differently to make the data integration process more efficient?
    - Could additional tools/resources be made available to make this process easier? If so, please describe the types of tools or resources that you think would be useful.
  - Have you ever reused data that you collected earlier in your career?
    - Did you encounter any problems during the process of reusing these data?
      - What factors were responsible for the difficulties you encountered?
      - What could you have done differently to avoid or minimize these problems?
    - Have you changed anything about the way you manage your data due to the difficulties you experienced? If so, please describe.
    - Would the data management system you currently use have prevented the difficulties you encountered, or would it be necessary to improve your system in some way?
      - In which ways could it be improved?
      - What prevents you from making the improvements you've identified?
- Data citation
  - Have you ever cited your own data in one of your publications?
    - Who did you interact with as a part of this process?
    - Was the process different from what you had expected? If so, please describe.
    - Did you work with journal publishers and the data repository simultaneously to ensure mutual citations between article and dataset? If so, please describe the steps in this process.
    - How much time/effort was required to complete this process?
  - Have you ever cited someone else's data in one of your publications?
    - In which journal did your citation appear? What was the required citation format for that journal? (e.g., as part of the main references list?)
    - Did the data creator know that you were planning to cite the data in your article before it was published? If so, please describe the interactions you had with the data creator.
    - Did your citation in this venue generate any interaction with the data creator that you did not have prior to publication of your paper? If so, please describe.
    - Did you contact the data creator after publication to notify that you had

        cited the dataset?
        - If the data were archived in a data repository, did you contact the repository to notify that you had cited the dataset?
    - Has anyone cited a dataset that you created in one of their publications?
        - Did they contact you directly either during the process of working with your data or after publication? If so, what type of interaction did you have?
        - Did you receive any benefits directly related to this citation of your data? If so, please describe.
        - Did you receive any benefits indirectly related to this citation of your data? If so, please describe.
- Unique identifiers
    - Have you ever acquired a DOI or other identifier for one of your datasets?
        - Please describe the process you went through to obtain the identifier.
        - Did you use any information resources or get help from any people during this process? If so, please describe.
        - How long did the process take from start to finish?

Some useful references suggested by members of the CEE WG:

USGS *DRAFT* DM website - see right-hand side content:
http://sofia.usgs.gov/datamanagement/why-dm/value.php

Baker KS, Chandler CL. Enabling long-term oceanographic research: Changing data practices, information management strategies and informatics. Deep Sea Research Part II. 2008;55(18-19):2132-2142.

Baker KS, Yarmey L. Data Stewardship: Environmental Data Curation and a Web-of-Repositories. International Journal of Digital Curation. 2009;4(2).

Baker KS, Bowker GC. Information Ecology: Open System Environment for Data, Memories, and Knowing. Journal of Intelligent Information Systems. BDEI Special Series. 2007;29:127-144.

Karasti H, Baker KS, Bowker GC. Ecological storytelling and collaborative scientific activities. SIGGROUP Bulletin. 2002;23(2):29-30.

Karasti H, Baker KS. Digital Data Practices and the Global Long Term Ecological Research Program. International Journal of Digital Curation. 2008;3(2):42-58.

Karasti H, Baker KS, Halkola E. Enriching the Notion of Data Curation in e-Science: Data Managing and Information Infrastructuring in the the Long Term Ecological Research (LTER) Network. Journal of Computer Supported Cooperative Work: The Journal of Collaborative Computing, Special Issue on Collaboration in e-Research. 2006;15(4):321-358.

Baker, K. S., Ribes, D., Millerand, F., & Bowker, G. C. (2005). Interoperability strategies for scientific cyberinfrastructure: Research and practice. Proceedings of the American Society for Information Systems and Technology Conference. Charlotte, NC, USA, October 28–November 2005, p. 3.

Baker, K. S., & Millerand, F. (2008). Scientific information infrastructure design: information environments and knowledge provinces. *Proceedings of the American Society for Information Science and Technology, 44*(1), 1–9.

Reference Type: Book Chapter
Author: Bowker, Geoffrey C.
Author: Baker, Karen
Author: Millerand, Florence
Author: Ribes, David
Editor: Hunsinger, Jeremy
Editor: Klastrup, Lisbeth
Editor: Allen, Matthew
Primary Title: Toward Information Infrastructure Studies: Ways of Knowing in a Networked Environment
Book Title: International Handbook of Internet Research
Copyright: 2010
Publisher: Springer Netherlands
Isbn: 978-1-4020-9789-8
Subject: Computer Science
Start Page: 97
End Page: 117
Url: http://dx.doi.org/10.1007/978-1-4020-9789-8_5
Doi: 10.1007/978-1-4020-9789-8_5