

Sessions Proposals/Ideas for ESIP Data Stewardship Committee Summer Meeting 2015

We generally have one “reporting” session and one “planning” session at each of the two meetings in the year, and these are quite useful -- Rama.

(Information for “Time” was added based the following Google Doc and the version on May 18th -

<https://docs.google.com/a/esipfed.org/spreadsheet/ccc?key=0ArDAFB2BsbFrdFpuQnpUVzJNcHM1VmRRY0pITjFSUkE#gid=1>)

1. Data Stewardship Committee Reporting Out Session - <http://commons.esipfed.org/node/7947>
 - a. **Time: Wednesday, July 15th, 1:30pm, Room 5**
 - b. Session Lead: Justin Goldstein
 - c. Abstract/Agenda: Please join us to learn and discuss the myriad of activities that committee members have been undertaking since the Winter meeting
2. Data Stewardship Committee Planning Session - <http://commons.esipfed.org/node/7948>
 - a. **Time: Wednesday, July 15th, 1:30pm, Room 5**
 - b. Session Lead: Justin Goldstein
 - c. Abstract/Agenda: Participants in the Data Stewardship Committee will discuss the range of work to be completed over the next six (6) months.
3. Federal Agency Repository Review - <http://commons.esipfed.org/node/7925>
 - a. **Time: Tuesday, July 14th, 10:30am, Room 1**
 - b. Session Leads: Viv Hutchison and Mike Frame
 - c. Abstract/Agenda: This session will seek to understand successes and challenges facing Federal agencies in providing repositories for science data.
Some questions to explore:
 - What's been done?
 - How is it working? How is this measured?
 - What level of data curation is offered?
 - How many full-time staff are dedicated?
 - What kinds of costs were involved to develop, implement, and operate a Federal data repository?
 - What skill sets are required/sought to support the repository?
 - How are agencies addressing the issues around inter-agency data for their organization's repository?
4. Information Quality and Stewardship - <http://commons.esipfed.org/node/7918>
 - a. **Time: Friday, July 17th, 8:30am, Room 4**
 - b. Session Lead: H. K. (Rama) Ramapriyan
 - c. Abstract/Agenda: The purpose of this session is to restart discussions in the Information Quality cluster and strategize its future activities. The objective of IQ cluster is to bring together people from various disciplines to assess aspects of quality of remote sensing data. We will be learning and sharing best practices

with a goal to build a framework for consistent capture, harmonization, and presentation of data quality for the purposes of climate change studies, earth science and applications. The efforts and goals of this cluster are not predefined and are motivated by the participants of the cluster, so new ideas and participants are always welcome. The IQ cluster met during the ESIP Meeting of Summer 2014, but has been dormant since then. In the meanwhile, there have been significant activities related to this cluster's goals in NASA (through its Data Quality Working Group, one of the Earth Science Data System Working Groups – ESDSWG), and NOAA (with its Data Stewardship Maturity Matrix). We will discuss progress in these two areas and chart a course for future collaboration to meet the objectives of the IQ cluster.

5. Data Stewardship Maturity Matrix - Use Case Study Session -

<http://commons.esipfed.org/node/7956>

a. **Time: Tuesday, July 14th, 8:30am, Room 1**

b. Session Leads: Ge Peng and Ruth Duerr

c. Abstract/Agenda: Assessing the current stewardship maturity state of individual or a collection of datasets is an important part of ensuring and improving the ways that the datasets are being archived, stewarded, and disseminated to users. It is a critical step towards meeting U.S. federal regulations and user requirements exerted on datasets funded or produced by U.S government agencies. Stewardship maturity assessment models provide a uniform framework for a consistent assessment within a context of data management in organizations and portfolios, and stewardship of individual datasets, respectively. In this session, use case studies will be used to demonstrate the utility of various stewardship maturity assessment models and to provide an opportunity for people in the ESIP community to discover the capability of stewardship maturity models, to provide feedback, and to discuss a consistent way to implement them on various tiers of scientific data stewardship.

6. Planning for Next Season's Data Management Short Course Episodes! -

<http://live.commons.esipfed.bluedotapps.org/node/7992>

a. **Time: Tuesday, July 14th, Lunch (12:00pm)**

b. Session Leads: Nancy Hoebelheinrich

c. Abstract/Agenda: By hosting a lunch or dinner discussion during the ESIP 2015 Summer Meeting, the Data Stewardship Committee wishes community input on how to expand its "Data Management for Scientists Short Course" (Short Course) and seeks authors, topics, and funding for the creation and promotion of additional modules for the Short Course.

The Short Course, initiated in 2011 and published in 2012 and 2013, is currently comprised of 35 training modules that are made available on both the ESIP Commons (<http://commons.esipfed.org/datamanagementshortcourse>) and the Vimeo websites (<https://vimeo.com/album/2142831>). At the discussion session, module authors and Short Course editors will describe the module creation, peer review, and editorial

processes used for the current modules. These descriptions will serve as background information for further discussion about:

- Topics that those concerned with data management now consider high priority for training;
- Which of the current modules need updating;
- Finding additional module authors;
- Identifying prospects for future funding.

The Data Stewardship Committee would like the Short Course to evolve in order to reflect updated data management practices, thereby positioning the Short Course to further support the growing needs for training materials on data management skills and practice.

This session will be geared towards those interested in learning about and contributing to the next phase of the Short Course. All are invited.

7. Cyberinfrastructure for physical data sets - iSample collaboration -

<http://commons.esipfed.org/node/8006>

a. **Time: Tuesday, July 14th, 3:30am, Room 2**

b. Session Lead: Sarah Ramdeen

c. Abstract/Agenda: EarthCube's RCN iSamples (<http://earthcube.org/group/isamples>) is focused on developing cyberinfrastructure for physical data sets. This includes all aspects of the lifecycle for physical data – management, access, preservation, curation etc. iSamples will incorporate the perspectives of not only the field scientists who collect the data, but also the data managers, and those building the cyber systems by which access or use of the physical data is mediated. iSamples has a number of working groups that are relevant to ESIP, in particular a data citations WG, a identifiers and metadata group, and one looking at what they call “user stories” but is similar to our use cases. Thus, iSamples is interested in networking, collaborating, and coordinating with the Data Stewardship committee of ESIP in particular. Representatives from iSamples will give a brief overview of the RCN, and details about each of these working groups. Time will be dedicated to discussing overlap and future collaborations.

8. Distributed environmental data curation - piecemeal or necessity -

<http://commons.esipfed.org/node/8001>

a. **Time: Thursday, July 16th, Lunch (12:30pm)**

b. Session Leads: Corinna Gries, Don Henshaw, Margaret O'Brien, Philip Tarrant

c. Abstract/Agenda: The need for archiving environmental data is rapidly increasing due to requirements by journals and funding agencies, and the overall recognition of the value of data by different communities, including scientific researchers, policy makers, and the public. Accordingly, a great variety of data archive solutions are currently being developed. Examples of the current solutions include: 1) university wide archives and libraries taking on scientific data curation, 2) research networks employing data managers, 3) single investigator labs asking grad students to archive data, 4) consulting companies offering their

services. In this landscape of boutique approaches we would like to start the discussion of the following issues:

- i. Distributed data curation - piecemeal or natural evolution?
 1. Do we have all tools, standards, recommendations, policies to achieve the needed efficiency in a distributed model?
 2. What kinds of collaboration and governance models would be most effective?
 - ii. Is the alternative - a data "center" for the environmental sciences - a better solution?
 - iii. Role of large initiatives (NASA, NOAA, DataONE, NEON, ...)
 - iv. Possible business models
9. Tools to Wag the Long Tail: Identification of Free and Open-Source Services that Assist with Data Management Implementation - <http://live.commons.esipfed.bluedotapps.org/node/7997>
- a. **Time: Tuesday, July 14th, 10:30am, Room 3**
 - b. Session Leads: Josh Young, Karl Benedict, Christopher Lynnes, Ruth Duerr
 - c. Abstract/Agenda: Many significant efforts are underway to advance data management planning across all levels of geoscience. There is, however, a concomitant need to provide resources or guides for moving beyond data management planning to data management implementation. This session seeks to further that dialogue by identifying resources that could be included in an introductory guide for data management implementation. Specifically, presenters will conduct a 'show & tell' for free and open-source services that contribute to implementation of data management best practices. The conveners will also introduce an online resource for attendees or community members to submit additional recommended resources in this domain. The intent is that by identifying and gathering currently available services and tools a resource can be developed for researchers that might otherwise fall into the 'long tail' of data.

Session proposed by student fellow:

1. Data Management Practices for Programming - <http://commons.esipfed.org/node/7968>
 - a. **Time: Thursday, July 16th, 10:30am, Room 2**
 - b. Session Leads: Sophie Hou and Fox Peterson
 - c. Abstract/Agenda: The purpose of this session is to explore the manner in which the principles and practices of data management as discussed in ESIP "Data Management For Scientists Short Course" (Short Course) could be made applicable to software applications/scripts. The session is meant to be differentiated from software carpentry or best practices for creating scripts. In other words, the session is not intended to teach attendees how to develop software applications or how to script efficiently. Rather, based on the key concepts presented in the Short Course, the session will seek to translate the principles and practices of data management to appropriate actions for individuals who are involved in developing software applications/scripts, such as

computer scientists and scientists, for their projects. The aim is for the resulting software applications/scripts to be more “curation-ready”.

- i. Sample of principles and practices to be explored include:
 1. What does it mean to "curate" a software applications/scripts?
 2. How can curating software program/scripts enhance my reputation?
 3. What are the basic curation elements to consider in order to facilitate a deployable open-source software applications/scripts?
 4. How to choose a program/file format to enhance interoperability?
 5. What are the metadata content that should be created for the software applications/scripts?

Other sessions that are also tagged with “Data Preservation,” “Preservation and Stewardship,” and/or “Data management Training”:

1. Collaboration areas - Highlights and Coordination - <http://commons.esipfed.org/node/7957>
 - a. Session Lead: Emily Law
 - b. Abstract/Agenda: This session will focus on coordination between the ESIP collaboration areas. We are inviting each collaborative group (committees, working groups and clusters) to give a lightning talk/presentation to highlight their current activities and future plan of actions. This session provides an opportunity for discussion of cross-group coordination, exploration of connecting related interests, as well as promotion of joint future planning and activities.
2. Get Git! - <http://commons.esipfed.org/node/7965>
 - a. **Time: Thursday, July 16th, 8:30am, Room 2**
 - b. Session Lead: Fox Peterson
 - c. Abstract/Agenda: The 2015 ESIP fellows are hosting a workshop to present the fundamentals of the online integrated version control tool, "GitHub". Being able to communicate about the GitHub tool is important! With the recent switch to Github by key technological players such as Google, it is imperative that earth science professionals can direct their teams to take advantage of its powerful distributed archiving and quality control framework. In this session, we will introduce you to the "git" command line tool and to git-specific vocabulary you need to know. We'll explore how the git tool, which allows you to save versioned changes on a personal computer, scales up to GitHub, for integrated versioning amongst teams on multiple machines or at different sites. We'll then set up a sample repository and deploy a simple program to the online GitHub site. Finally, we'll demonstrate how to share repositories using methods such as "forking", "branching", or "submitting a pull request" that allow modifications with permission-based securities. This workshop is designed to be beginner friendly, but we're happy to work with you on more complex integration tasks that you might want to share, and to explore the future of GitHub as an application management platform. The student fellows hope that you will come "git" into "Git!"

3. Citing a Black Box and its Output - <http://commons.esipfed.org/node/7980>
 - a. **Time: Tuesday, July 14th, 1:30pm, Room 2**
 - b. Session Leads: Jessica Hausman, Daine Wright, Robert R. Downs
 - c. Abstract/Agenda: Citing data is becoming more significant as science journals are requiring authors to provide their data sources, and data provenance is recognized as important for its stewardship and reproducibility for experiments. But, what happens when you obtain your data through a tool that changes the data or metadata from the original file? This session will look at how to cite data that is produced from a tool that is essentially a black box to the user. This includes tools and services that subset, aggregate, calculates anomalies or climatologies, etc. Identifiers that can assist in making the citation more exact will also be investigated.

4. A Life Beyond: Leveraging Earth Science Interoperability Projects - <http://commons.esipfed.org/alifebeyond>
 - a. **Time: Thursday, July 16th, 3:30pm, Room 6**
 - b. Session Lead: John Graybeal
 - c. Abstract/Agenda: Many earth science interoperability projects have been funded; few survive to conduct stable (or growing) operations. And many projects with long, successful histories and products are passed by in the excitement to create the Next Great Thing. What can be done to improve the uptake from earth science interoperability projects, especially as they decline? This session aims to explore, ideally by presentation and open discussion, how we can:
 - i. make projects better known, connected, and used while they are operational;
 - ii. strengthen collaborative and synergistic project approaches to addressing earth science interoperability;
 - iii. promote appropriate operational funding for successful projects;
 - iv. identify new operational and funding models for projects whose grants will run out;
 - v. provide archival advice, best practices, and opportunities to projects that will otherwise go dark; and
 - vi. discover projects that may have already disappeared (or be about to) and recover key value.

Case histories, lessons learned, visions of better ways, and brainstorming, addressing any of these categories, are welcome. While we are particularly note the challenges with leveraging community and social projects that target improved interoperability, projects promoting technical interoperability (through common interfaces, standards, software, data models, or other approaches) also will contribute critically to the session's objectives.