Data Stewardship Committee Planning Session [1]

Submitted by rduerr on Thu, 2014-04-03 13:06   Friday, July 11, 2014 - 11:00 to 12:30
Event: Summer Meeting 2014 [2]
Session Type: Breakout [3]
Collaboration Area: Data Preservation [4]
Preservation and Stewardship [5]
Abstract/Agenda:
During this session we will review the status of the activities of the committee since the winter meeting and make plans for the remainder of the year. Current activities to be reported on include:

- How should this committee cope with the ever growing numbers of partially overlapping ESIP clusters/working groups?
- Planning next steps on existing activities
- A couple of potential new activities to consider

Notes:
Agenda:

During this session we will review the status of the activities of the committee since the winter meeting and make plans for the remainder of the year. Current activities to be reported on include:

How should this committee cope with the ever growing numbers of partially overlapping ESIP clusters/working groups?

Planning next steps on existing activities

A couple of potential new activities to consider

Notes:

Ruth - Planning for the next 6 months instead of reporting on activities.

Review of agenda.

Rama suggested we address the first topic last.

Vicky asked what is the list of new/existing activities that we need to evaluate? The group discussed the order of topics…

Closeout activities: (unless further action is needed in this area, consider it closed).

Use cases, data collection structure, physical object stewardship - Denise has looked a the principles and does not think many changes are needed. During her report she also talked about the PCCS, and these findings should fall under a separate activity and this one should be closed out.

Other dormant activities?

Identifiers, data citations. Dormant (not in the world at large). RDA is working on some of these things, but this group has not moved forward on identifiers. And for data citations, the effort is to push out the knowledge but not necessarily new projects for us to do internally. Opinions?

Matt said looking at the wiki on data citations, there are three areas. A shell area for publishers and a blank area for users. There are recurring questions coming up. The data user piece, which is linked but not filled out would be useful. He is working on recommendations for the meteorological society which will have to address some of these things and there might be an initial draft he would share as a suggestion to this group. It is a gap that could and should be filled.

Vicky said that one of the new activities is software citations and that is relevant to this discussion.

Ruth - the way the work gets done in this group - people volunteer. So unless someone volunteers there is no activity? She asked Matt if he would cover this, he said yes. Working on this for AMS and the publishers are all scientists and these questions keep coming up.

Ruth said now the guidelines that AGU came out, and that is an easy way forward for publishers, but for users, she agrees. Like the educational modules, they would like some practical examples. If you have streaming data - do it this way. Matt - what should go on our page? What should this look like? Are the questions they have been asked in this area.

Looked at the activities list
https://docs.google.com/document/d/1lPFHUXmAXA4eOtRAD4-aUqVgPA0EOvkhbvFjwMMdVY4/edit?pli=1 [6]

The leader for the citations - the phase two, Matt was made leader. Mark suggested that we not just talk about data specific, but Rama said that was a major concern. The need to distinguish citations for data from software. Ruth said what we want to do about software is a new issue. Mark said there is not a clear line between the two. Vicky said this is a topic that should be part of the discussion on citations. Often the model or software gets mentioned in the text of the paper, so there might be more traction for shifting the behavior on that process as opposed to data.

Sky - there is a push back from publishers as well, for including DOIs and GitHub stuff. And maybe the editor roundtable can help or else it will get worst as we thorugh more into it. Bob - Datacite is going to get in on this too, and how can a publisher say no?

Ruth - there are issues with software - there are some communities that already have norms on how they do it, and they are not the same. Mo - wouldnt the doi resolution be different for that as well? Ruth - yes, it will show up and impact your reputation. Mark - all the data increases your citations. open data gets you more academic credit, but that doesn't mean your data gets cited. Mark - Data papers with open data behind them get cited more than those who don’t. Matt should focus on his topic, and we should start a discussion on software citations. The publishers Ruth will crib from AGU, and she and Matt will work on the AMS stuff and getting the user perspective.

Matt provided more background of the process for AMS, and what ever comes out of that can be shared with ESIP. OR should ESIP make recommendations to AMS and then have a feedback from them?

Mastpha (sp?) spoke about some of the international views, there is a group which involves publishers - trying to create a service to cross reference publications and data sets. And they are engaging the publishers in this process. To include the formal citations. Do not think these groups would come up with recommendations on citations as this is already being dealt with by Force 11, but we want to feed in any recommendations from that group into the bibliometrics working group and this cross referencing group. His group has endorsed the data citations principles, and that this group will come up with recommendations for citations and we will endorse them. In the context of the working group, they are looking at implementation and need to have an agreement on citation formats. And where is the best group to have make those ....

Ruth - The problem with that is that some of those communities have gone in a very different direction than the earth science, and she is not sure how quickly a set of suggestions that will work for everyone will come out. Even questions about what is an identifier - this concept is different in biology. And how they cite that is different than the way we do. Not sure how it will play out either, their completion process.

We can scratch both of our planned activities from last time off the list. The executive committee has agreed to send out the endorsement for force11. And the AGU data policy, they have done a very good job there.

Ongoing activities

Data stewardship training - will have a slide on this.

Prov-ES - no one is sure what is going on… Rama mentioned there is another working group focusing on this. Sky and Anne said they have not seen any emails or documents on this topic. Like the matrix spreadsheet etc. Someone asked if we talked to Sara Graves, and that they were talking about PROV-ES as if it was a real thing. It is a real thing, but it is moving. And what can this group do to make this visible? This community needs to be able to see this! Sky asked if it was PROV-NASA. Rama said Hook had a session at the last meeting and hopes to share that with the community. Anne - Hook might benefit with some help in championing this effort. Ruth suggested that why don't we invite Hook to a monthly meeting to provide an overview and get their results visible to the people on our mailing list - ACTION ITEM FOR RUTH. Get Hook to attend a meeting.

Data study - report in review. And waiting to figure out what to do next. The EOS article is about to be published, got revisions and made them. And waiting for a workshop report. Will send out an editorial review to everyone later this week. Not sure what the next steps would be. Someone asked for a summary of this project. Anne said - January 2013 she raised a question of a NRC data decadal survey, formed a cluster and working group. Lots of discussions on it, a workshop in January on it and just produced a report, which the EOS article will refer to. It will also have a press release with some pictures. Bob asked if the url will have a doi? Ruth said if it is on the commons, it will get one.

PCCS - presentation earlier this week. Mark also said they have to get the paper out. Nature might be willing to accept it. Was hoping Josh was here, he wants to give up to go with EOS. But Ruth said we need to have both. And Mark said we need to talk outside of our own world. Nature is broader and they are slightly different articles. The EOS would be the PCCS and NASA's efforts and specifications and where they are applying it etc. Actual implementation. The Nature article is this is an important topic, pay attention. And here is an example of how one community handled it and how can you handle it in your own community. Mark said Nature thought it was relevant but worth submitting as a different type of piece. Need to boil it down to 500 words or something like that. And the editors we need to hit up. Mark will follow up on that. But if the paper needs to be rewritten, Bob can take a stab at it, once we have more information. Ruth will help as well. Contact relevant editors, then shorten. Denise will help as well.

Data Stewardship Training - Nancy's summary of where she is at (she could not attend). She is in the evaluation stage. and is looking at metrics from before and after schema.org. And is doing some marketing efforts. Is working with the educational committee as well.

Issues - (see slides for list). Some of the modules are getting old and need to be updated. Do we update them, retire them?

And we tried to get a grant to evaluate the effectiveness of the modules but we did not get it. We are shy on resources to do that evaluation. Bob suggested a new grant to do more modules. There are lots of topics to be covered. Mark said, isn't there a room for a landscape scan? Seeing what else is being done in different disciplines and by libraries for data curation. Do we have a harmonizing effort? Anne said software carpentry is doing data management training - Ruth said that is Erin. But she is working with Mozilla to do it and is reusing some of our documents.

Nancy needs help with this, if anyone is interested let us know. Data carpentry is a great partnership - maybe we as a committee can help, and make a real partnership with Mozilla. We need to have a telecon and have Erin talk. Sky said they are working on something like this for USGS, and he will beat the bushes there to get us involved in that process.

Ge Peng Stewardship Maturity matrix

Summary of talk given yesterday for those who were not there. Work done between NC state and NCDC. We all know the value of stewardship for data after they are used and we become custodians of that data. As a data center, many questions come out of this, like congressional inquiries with compliance with laws, or businesses that want to invest money. And common data format etc. Or modelers who want to use a data set for computations.

There is not currently a practice for how to do this that is holistic. Reviewed in flowchart, what they have done so far. There are 4 key areas. And focused on the relevant areas they identified. ... went over who could use the matrix and the various scenarios related to that. Current stage is a working draft and next steps is to get ESIP involved with this process. The presentation is uploaded in the node for yesterdays session. Best practices on data quality (http://commons.esipfed.org/node/2369 [7]).

Sky had been looking in to being a trustworthy ISO certified digital repository, and some will never make that level. So now they are taking that list to achieve this standard, and making something similar to this project. To figure out how to grade and evaluate these facilities. And something like this would be really helpful for us and we would like to be able to work together on this. NCSID (sp) also worked on something like this as well.

Discussion continued with questions about data centers etc.

Hopes to have use cases in the future. Right now it is generic.

Ruth - was thinking that if this group took this on to ESIP-ize it, we can look at the terminologies and phrases to make it fit a broader audience.

Matt mentioned the iSchool at Illinois (or Syracuse?) which has developed something similar and recently put up a detailed website. (he can share this with those interested) But this is a more operational level then their model. And this can be something to look at as well. Sky suggested we co own this task with others.

Mark this is a big question right now, how do we measure quality of what we produce. He thinks maybe an RDA working group, but that might make it so generic that it is like this thing made at Syracuse.

Bob asked if it is generic in this way, so that as community standards change, this can stay static and conforming to the community standards would be different. So score of 4 today, but measured in 10 years the score might only be a 3.

Ge addressed these points, as people all get to level three, they might have to change the scale to remain consistent.

Mastpha (sp?) - said this is a discussion they have had in his group from the data center perspective.

Ruth, they have concepts of levels of service at NCSID because they don't have the resources to do high level for all of the materials. We can do a good job with this one, but others came in in a non standard format and we don't have the funds, but we will archive it but not put in a lot of service on it. Wouldn’t want to be ranked differently because of these differences in funding. There is a minimum level of service.

Mastpha (sp?) - yes a minimum level and then it is up to the user and what they want to use it for to determine the quality. So is it really up to the data center to provide this evaluation? Ruth - I am not saying this is a perfect answer. reviewed some of the headers and what they mean.

Mark talked about the differences between data centers respo bilities and what will improve the quality of the data. The data is better taken care of and more useful, doesn’t go to the quality of the data. And then there are scientific views, that this data is valid. And the reuse question can only be answered by the re-user themselves. And these are all distinct questions, and the community is bundling them all together. And this will be useful to help with that.

Vicky has suggested this go on the list of activities. Ruth agreed.

Only other new topic - software citations. Vicky said we do need to talk about it and connects to the broader topic and talking to other groups.

We dont have time to cover working with other groups, and will move that to another telecon too.

Rama - PCCS, we want to take it to the next level of going to a standard. Started working with the ISO group. And a new work item is needed. Denise suggested she needs to brief to this group the outcomes of her efforts on physical items. should the standard that move forward be remote sensing only, or can we generalize it? Rama says move forward with remote sensing and then do another piece, but Ruth wants to have a telecon first to get more information. Someone said it would be good to have it more generalized.

Actions:
Outcome/Action items:

- Matt M. will work on the data citations for users along with the work he is doing for AMS
- Ruth will crib from the AGU guidelines to fill in the section for publishers.
- Ruth will get Hook to attend a telecon and report on PROV-ES
- Mark (and Denise, and Bob) will follow up on the Nature article on PCCS
- Ruth will get Erin to join a telecon to discuss the partnership with Mozilla regarding the educational modules.
- Sky will look into what is currently being done by the USGS in the area of best practices and educational materials (see above).
- Telecon on how to stay more connected to other clusters/committees etc.
- Telecon by Denise to review the differences in views on PCCS

Session Leads:

Name: Ruth Duerr [8]
Organization(s): The Ronin Institute [9]

Notes takers:

Name: Sarah Ramdeen [10]
Organization(s): School of Information and Library Science UNC-CH [11]
Email: ramdeen@email.unc.edu [12]

Teaser: During this session we will review the status of the activities of the committee since the winter meeting and make plans for the remainder.
Accepted:

Source URL: https://commons.esipfed.org/node/2351

Links
[1] https://commons.esipfed.org/node/2351
[2] https://commons.esipfed.org/2014SummerMeeting
[3] https://commons.esipfed.org/session-type/breakout
[4] https://commons.esipfed.org/collaboration-area/data-preservation
[5] https://commons.esipfed.org/collaboration-area/preservation-and-stewardship
[6] https://docs.google.com/document/d/1lPFHUXmAXA4eOtRAD4-aUqVgPA0EOvkhbvFjwMMdVY4/edit?pli=1
[7] http://commons.esipfed.org/node/2369
[8] https://commons.esipfed.org/node/295
[9] https://commons.esipfed.org/taxonomy/term/2011
[10] https://commons.esipfed.org/node/557
[11] https://commons.esipfed.org/taxonomy/term/373
[12] mailto:ramdeen@email.unc.edu