ESIP winter meeting
Data Stewardship committee meeting
1/10/2014

Attendees: Ruth Duerr, Curt Tilmes, Rama, Sarah Ramdeen, Helen Convers, Anne Wilson, Nic Webber, Bruce Barkstrom, Mark Parsons, Bob Downs, Steve Kempler (and three others).
ONline - Nancy Hoebelheinrich, Mo Khayat,

Notes:
Ruth showed everyone how to find the Preservation collaboration area on the wiki. The meeting notes section have gotten log - Ruth suggested trimming this. The page also has information about telecons and getting involved.

Ruth is the new cluster chair, Curt is phasing out but will act as a co-chair. Denise Hills will also act as a co chair. Sarah is our student fellow taking notes.

We have a number of resources including activities.

We are an active group, with monthly telecons and side groups which have had their own calls. We are going to do planning for the next six months worth of activities, and we have a planning list which has not been updated in a while. With the goal of only addressing a few thing we think we can get done by the summer meeting. THis group has wanted to do a lot of things but volunteer time means we can't always complete everything though we have done a lot of them.

Activities list - **it is a google doc which will be linked** h**ere**

This list has not been updated in a while.

Use cases
The use case activities have a lot in common with what the PROV-ES group is doing.

Ruth asked if there were any funding possibilities to continue the Use case activity including interviewing possible researchers. Nic offered a suggestion - Data Curation Profile Projects as a tool that might be useful in talking to users. Mark had suggested targeting friendly users instead of leaving this open, in order to get user engagement. Sarah will continue to work on this project on the side as she has time. Bruce said he might consider.

Someone asked about the goal of some of these activities - Ruth said some of these things could be presented as an EOS paper. Ruth asked who the audience was at the AGU events and Mark asked if they were writing a PCCS paper - Ruth said yes, one for DLIB, and that it was to create awarness in the library community, which is a different focus.

Bruce said if there was not a funding source it is really hard to find volunteers to think about the

problem. He thought that the conversation should be amongst users - real scientists as opposed to librarians who are the audience of DLIB. Ruth said that as Nic pointed out, we can target individuals. Mark was not sure what the point of a publication would be. Bruce made a comment about use loading in archives and numerical demographics and a reasonable scenario of how people spend their time and forecasting in the future. Rama said in this case the purpose of the use cases are different, to influence the improvement of the PCCS and justification for it as well. A summary of that would be a broad discussion of that and if you are using earth science data, and want to reuse them, here are some scenarios that defend why you need the information. Mark thought it might be more in an informatics journal ESSI instead. To the community. Rama said also the science community, to show how they engage the community. Mark asked if that was the PCCS paper. Curt said this is a step away from that and gap analysis. ANd Rama said if we go in to a standard, this is a justification of why this is a standard. Ruth summarized - where should this be published and summarized the utility of this project, to flesh out the PCCS and start gathering information to say why the PCCS is useful. Steve asked - we are putting emphasis on use cases, but is there a good cross range of users using these use cases? It would be nice to know we are hitting all types of users including early career. Curt says we are struggling with that - who to reach out to and how to get them to contribute. Steve said that we have to fill in the categories and then find the people. Bob said - is there a away to a way to find the gaps in our use cases - to find the knowns and the unknowns. Nic said, there is a larger point of using the use cases and standards development and the usefulness of use cases. And what is the gap - make some commentary on the use case process. Steve said - finding gaps and throwing out things we don't need. Someone said that was like the big data conversation - and we can find a more conceptualized model and extend to a broader focus. Steve said - in a science publication, a great study would be a user case study which highlights the different categories that have been defined. And maybe that is a thing to proposed to the federation to fund this effort. And key in to the user needs analysis is doing, and have some senergy - you defined these categories, that might lend themselves to us.

Ruth said that Rama can speak to some of that, in the document that NASA created it does speak to some of these requirements. Rama spoke about how this started in ESIP and moved to NASA, and it is internally being used in NASA at headquarters, they sent a note to all the flight project scientists asking them to send a planning document to headquarters which included - assume your mission will end, what is your close out plan, phase f. As part of that, besides taking care of re-entry of the craft, how are you going to do the final processing of your data and share it with the dacts, and it referenced the PCCS and how the groups are going to address these specs. Bruce asked how long it took to develop this, Mark clarified if you have a data set and a mission, how long would it take to fill out the PCCS? Bruce said both - to develop and to employ. Rama said two years, a few people months, and a workshop report - 3 day workshop with man hours. So Bruce said 3-6 people months to create and as a project manager to fill it out? Rama said maybe a person week maybe? It is the actual work involved with the instrument teams. And with GLASS. These teams use this and followed the guidelines and gathered materials and of the order of 2000 documents they had to sort, and narrow down to what they had to give to dac, which is now holding on to these. So maybe 3 man months to

do that. Curt said the PCCS said you have to archive the data, where do you draw the line for that? Steve addressed Marks comment about hurdles. And how people did not prepare for PCCS delivery when they started, the information needed for the PCCS is scattered. And now they are doing it. He mentioned a specific group where they are meeting to discuss it before they start their project, and it will go faster with the mission having not ended when they review the process. This will make it useful and routine. This is in level 1 requirements for new missions like SMAP.

Mark - NASA is now adopting this, and should be publicized. And that would be a good EOS paper. That would be a good point to make. Those scientists are planning the missions as we speak. Mark - not only is it a good idea, it is the law! Ruth thought that the important point was why this is important … and equally importance that it is being used. Steve - people believe they are part of the requirements. So make it more concrete and believable. Put it out there, Curt - can NASA assign a DOI to the specifications? Someone said - as a data user side, at the end of the day, not sure if it would be used (?) and Ruth mentioned that the PCCS has a great justification in the document as to why it is included. IIke if there was a glitch in teh system. There was a scale from 1-5 and each item on the list has a justification for why it would be kept and how important it was. That is part of the story. Rama - the final document removed the priority thing because low priority might lead people to not include information. Bob said that would be good to include in an article, this is important because… Rama said the matrix has more information in it then was used in the final standard by NASA. SOmeone said something about data uncertainty. And Ruth said this is why EarthCube might succeed or fail - you dont have a lot of this stuff. BUt your results might not be so great. Someone else said that if it is so cumbersome that you can't access it you might as well not have it. Ruth said she admits, the requirements on a nasa mission are different than someone who bought a data logger. For a NASA mission you have a lot of calibration information for your data logger. It is a different scale than the individual and filling in a category might be trivial. mark said that it also depends on the user, they might not use much of this but it builds trust. Someone said - you have to have guidance so that there is no commitment. Curt suggested a targeted telecon on this topic instead of continuing the discussion now.

Mark summarized the paper topic and get to a telecon later
THe pccs is important, it is so important it is a nasa requirement and comes from a science need as evidence by the use cases which justify this, that this is the information that is needed, so useful that NASA is including them. Bob suggested being able to point to the use cases that highlight these points. Mark said if this is an EOS article it is a quick, lunchtime read and it should be something to be considere.d Someone suggested we develop the abstract and outline first. And maybe that can be at a telcon.

Mo said - GES-DISC is publishing a paper about the way they implement the PCCS, in EOS. Mo offered for him and his colleagues to help Sarah.

Mo Khayat Comment on Article:

For the HIRDLS preservation at the GES-DISC we actually have an AGU Eos article that will be published in the Feb 2014 time frame to publicise the preservation issues that we encountered.. That article might be of help for the Eos article under discussion here.
My DISC colleagues and I can help Sarah with her article. I volunteer myself and my colleagues for that!!

Article Title: "Just Take Those Old Records Off the Shelf: GES DISC Manages HIRDLS Data Preservation:"
James Acker, NASA's Goddard Space Flight Center, james.g.acker@nasa.gov

Data Collections Structure
Bruce - basically from this stand point, we had a good discussion with a small group, and out of that we have a conclusion that we have enough information to create a fairly short paper in earth science informatics to do inventory accounting consistent with OAIS reference model. Well before the middle the month or year. And we had a discussion about uniqueness and we might be able to get a longer preliminary article when you have replication and equivalence of data objects. So two potential articles by the middle fo the year.

PCCS
Denise is testing to see if it can be extended to cores but it has not reached finalization of vetting yet. Rama mentioned finishing the paper - the opinion piece is almost done and the DLIB piece is nearly done as well. Just Ruth's section is missing at this point. Bob thought it was nearly done, and Rama mentioned that Denise was going to write something as well. The Science piece is still waiting for reviewers? Well it is an editorial thing but they still review it.

Citations
Ruth asked if Mark was still lead on this, and can report on this. Mark said there was not a lot of ESIP activities on citations, but he and Ruth have been working at another initiative at RDA. There are 25 actual groups, not just individuals who have been working on this. they have come up with data citation principles, sort of a step backwards from our guidelines, but we are getting past that level of competition between groups, and this is beyond the earth sciences and sciences. Those principles were put out for comments and that closed at teh end of the year, and they are compiling them and there will be a workshop in FEb at the digital curation meeting where they will be refined into finalized principles. This is important because we are working as one voice. In the next six months we need to engage with that activity and maybe have ESIP endorse this effort. Also we have to fight a rearguard action,AGU just posted a backward policy on which does not include data citation. THere have been some conversation on the mailing list about that, and Ruth on ESSI has been tasked to address that. What Mark and Ruth are going to do is write up something about why citation is important and get a change in AGU statement. They want to get Bernard (?) to help with his clout. We have guidelines that were endorsed by ESIP two years ago on how to develop citations. Someone asked about old papers and data -

Mark said that we have discussed this… Curt pointed to the three missions on the activities spreadsheet - why what and how. And Mark is not volunteering to do anything except fight this rearguard action. Someone else in the audience asked about the citation specifics and how to connect to the data and the author. Curt said it has been published. And Rama said it was on the commons.

Nic said this would be nice from ESIP for them to make a statement on AGU's policy. and AMS is discussing the AGU policy and ESIP should respond soon. Mark asked for any publications that justify data citations to send them to him. Curt asked if we can publish an opinion piece through ESSI? Which is something that Ruth had been tasked with as part of the ESSI community. She is not sure what the AGU group who wrote that was thinking and is trying to get a hold of them first.

Someone said we should make more use of the weight of the ESIP federation as opposed to individual members, if there is a way to leverage that. Mark said, to even get started we need two paragraphs to the publication committee that says data citation is important and in this way, once there is push back we would rally the full force of these groups. That will take time, so we should do this first and then gather the groups. This could be something the general assembly, and before we blow this up we get the publication committee the time to respond.

Ruth added that she will be at some workshop representing ESIP and will send out a call so she can have a list of names or icons of everybody who is implementing the data citation guidelines. They reflect high level principles and are at these centers, internationally as well.

Data management training
We have a set of brochures. In addition to the brochures, she has some changes to the template for the educational modules on the commons, to change to schema.org and would like feedback on that if it is useful. and if it is, she will do that to the rest of the modules. She can give us one to look at . Curt said to send out an email as there are not a lot of us here. Ruth also said we can verify if google is seeing these correctly and if they are we are good to go. Nancy will send a request out in email. Ruth said she has gotten interest in her own community on how to use these. and training her graduate students. This person had a suggestion for when we create more modules - not just the rules but actual examples. If doing type x do it this way. So directory structures, file names, data formats and contents of spreadsheets. Real concrete examples. Curt suggested the carpentry effort from the plenary. Mark said Kevin Ashley is interested in collaborating in that effort (he is from the DCC). And Mark recommended that someone follow up with him, it would not be him but someone else from that group. Ruth said she would do that. Nancy is also happy to work on this kind of connection.

Physical objects Stewardship
STill in progress - by Denise. CODATA is doing its own thing, spinning off a group and is there someone from here that might be interested in working on it, with physical objects. This will be chaired by Kerstin.

Prove ES
also addressing use cases

Information quality
Has graduated to its own activity - it is its own working group.

Data decadal survey
Data Study has been graduated to become its own group

Monthly telecons
Curt and Nancy have not been able to attend, so Ruth would like to change the times. Sarah is gathering times from key players and will send out a poll.

Curt said there should be an article where success stories from data citation use would be important. Even if it was a small sample, not just reproducibility, but other practices and see what other people are doing. Bruce said that it related to his work and showed a slide with various project stages. He also made some comments on how long it would take to reconstruct someones data. And showed how many years it might take with a large scale project! Probably as much time as people took to gather the data in the first place. Ruth agrees that a small real study with statistical significance would be important as to how well it was to reproduce the data. Rama said even if looking at 100 papers last year, and look at them and see if in fact you can identify the data that was used. Sarah suggested collaborating with EarthCube (and Erin on this).

Action items -
Ruth will email Kevin Ashley regarding the management training modules and joining with DCC. Nancy offered to help (see note above in the minutes).

Sarah will organize a telecon call for developing an EOS paper on use cases

Ruth and Mark are working on an ESSI response to the data citation issues - people will need to send any articles or publications on justifying this to Mark.

Talking to Erin about collaborating with EarthCube on a data tracking/citation investigation as mentioned by Curt.