Breakout Session 1: Virtual Participants

1. (Host: Marlon Pierce)

- Brief introduction of virtual participants
    - Marlon Pierce
    - Robert Jacob
    - Shahani Weerawarana
    - Steven Browdy
    - *Others - please add yourself...*
- Review of questionnaire and how it can be revisited.
    - Motivation for the background questions
- **Background Questions**

1. Improving the efficiency of a complicated or error-prone set of computational steps or tasks is a common problem in research. Do you have these types of problems in your research? How do you address them? What are the strengths and weaknesses of your current approach? What improvements would you like to see? (eg, over using simple shell scripts to manage computations)
    a. *This question addresses both the complicated and error-prone nature of computational steps. Marlon explained the general background and motivation behind this question.*
    b. *Robert suggested a re-wording of this question since the reference to computational steps may "scare-off" lab scientists and other non-computational scientists.*
    c. Possible rewording (Robert): Improving the efficiency of a complicated or error-prone set of steps or tasks (in the lab or in computational work) is a common problem in research....
2. Another common problem in research is reproducibility: a researcher wants to make sure that the steps used to produce a specific result in his/her work (or students' work) can be reproduced by reviewers, collaborators, new team members, etc. How do you address this problem currently? What are the strengths and weaknesses of your current approach? What improvements would you like to see? (e.g. all steps are in the papers, additional steps are described online, I make available all analysis code.)
    a. *According to Steve, currently reproducibility is handled by an ISO standard (19115) for metadata standards. Very popular in Europe, and many other research centers.*
    b. *Clarification by Marlon that this is particularly addressed towards scientists who have not looked into this aspect in-depth.*
    c. *(Steve) This question can also be answered from a non-metadata perspective.*
    d. *(Steve) Would be nice to have some tool that can capture such information automatically, particularly when information changes. (Marlon) Similar to "executable science publications". Maturity of these technologies maybe questionable at this point.*
3. How do you train new team members (such as students) to use your team's codes and other tools? How much of this do you think could be better supported for reproducibility,

efficiency, etc (e.g. one-on-one training, internal web site, lab manual)?

   a. *Particularly addressed with respect to integrating new team members with respect to training and ramping up.*
   b. *General consensus is that this is a valid and useful question.*

4. Do you use standard data products produced by third party groups in your research? Do you see the need for better documentation of the steps used to produce these data sets? Do you need to modify the processing steps or have alternative processing steps that you would like to apply?
   a. *Different groups do different types of versioning. From the workflow point of view, how are data products created and continuously evolved?*
   b. *(Robert) Clarification on whether last question is for data processing. (Marlon) Yes, about the data pipeline.*
   c. *(Robert) Possible rewording of last question: "Do you need to modify the processing steps for the data or have alternate processing steps that you would like to apply?"*
   d. *(Robert) should also remove "standard" from this question.*
   e. *(Steve) For the long-tail scientists, the change in processing steps is most likely to occur. Everyone would want to do things in their own way, instead of hooking into an existing workflow. Is there such a existing way to do so? (Marlon) It may not exist now. But maybe in the future (3-5 year time period). It may produce a simple solution.*
   f. *(Steve) Individual scientists may use their own scripting language. In the future, it maybe useful to have some methodology where they could plug in these workflows in the future. (Marlon) Preservation of the unique and personal could be done alongside the general.*
   g. *(Steve) What is the general context? Workflows as middleware or workflows on the desktop. (Marlon) Probably both. Metadata should ideally be captured from both aspects.*

5. Does your group produce standard data products used by other groups? How do you document the processing steps that you use to produce the data products? How do you address changes to the processing steps (versioning, for example)?
   a. *Inverse of the previous question (in a sense).*
   b. *(Question) What is meant by standard data product in this case? (Marlon) Actually, the use of a data product by a group that is not your own - may not exactly be a "standard". For instance raw data that supports a particular data community (long-tail scientists).*
   c. *(Robert) What about an individual scientist publishing their own data on the Web? Would that be a "data product"? (Marlon) Yes.*
   d. *(Suggestion) Maybe the word "standard" should be removed from the question to reduce the confusion, since any data product is under consideration within this question.*
   e. *NEON is a good example of an organization that is providing infrastructure for individual scientists to share their results.*

6. Do you work with researchers in other fields on multidisciplinary problems? How do you communicate processes at various stages (planning, development, operation)? What are the challenges in collaborating on those processes? (e.g. using email, slides, wikis, face-to-face meetings)
    a. *(Marlon) This question addresses the whiteboard paradigm where multiple scientists from different disciplines collaborate and interact towards solving a comprehensive problem.*
    b. *How to answer "communicate processes"? (Marlon) I guess, "planning phase" in terms of creating and running infrastructure for collaboration. How do you decide what you are going to do collectively? Across domains? This question probably needs a use case or further explanation. For instance, data mining of environmental data collected by partners, scientists in the field, scientists who store the results with metadata, data mining scientists and scientists who analyze etc.*
    c. *(Robert) It is still not too clear. A sample answer may clarify the intention. (Marlon) Maybe the question could be reformulated by some scientists involved in this paradigm.*
7. Do you have scalability problems in your research, such as accessing, moving, storing, sharing, and processing or mining very large data sets? Please describe them. Data could be both input (such as observational data) and output. What are your approaches to solving these problems?
    a. *This question is very clear.*
    b. *(Steve) Even long-tail scientists sometimes have the need to access big datasets produced by major research groups or agencies. There is a thrust towards moving the process towards the data instead of moving the large datasets towards the processing problem. What are the ways in which the processing can be moved towards the large datasets. It may help solve some of these scalability related problems.*
    c. *Use case from Steve: relatively small data sets from oceanography (physical oceanography, surface meteorological data). Quality control tests require comparing to climatological models have to say about the geographical region and time. These models can be large even if the data provider supports subsetting.*
    d. *This use case comes up often in GEOSS. Satellite data, imagery data, etc are used by multidisciplinary data. GEOSS uses interoperability standards, OGC services, etc. Other examples: Abu Dhabi government's "Eye on Earth" initiative (not the same as the Microsoft partnered initiative)* http://www.eyeonearthsummit.org/summit/about.