Title: From Chaos to Consistency: Moving Towards Data Stewardship and Sharing for the Watershed Research Cooperative

Investigators:
Jon Souder, Jeff Hatten, Lisa Ganio, Kevin Bladon

Project Duration:
July 1, 2016 – June 30, 2017

Objectives:
The primary objective for this project is design a WRC data stewardship and management framework that will allow integration and synthesis across disciplines within a particular study, across studies within a discipline, and an overall synthesis across disciplines and sites. Consistent with the WRC MOAs with funders, these data sets need to be available to the Cooperators and others in a timely and transparent manner.

The second objective of the project is to respond to the increasing emphasis on the part of many funders (NSF, NIH, federal resource agencies) that require data management plans and data sharing for projects they fund. As part of this objective, we will begin structuring the WRC datasets to meet ESA publication standards.

Summary of Accomplishments:
In our proposal, we identified four approaches that we would use to fulfill the two project objectives identified above. These are:

a. ***Steering Committee***. As outlined in our proposal, the initial task for the Steering Committee was to prepare a position description (PD) and select a data manager.
   - An opportunity arose for us to work with the Research Data Services group at the OSU Valley Library. We were able to work with Steven Van Tuyl (Data & Digital Repository Librarian) and Dr. Clara Llebot Lorente (Data Management Specialist) to use FWHMF funds to purchase Dr. Llebot Lorente's time to coordinate and prepare the data management plan. This has relieved the Steering Committee (and PIs) of a considerable burden and risk associated with recruiting, housing, and supervising a data manager.
   - Members of the Steering Committee have been identified, and an initial meeting is scheduled for late November, 2016.

b. ***Framework Design***. The goal for the Data Framework Design approach is to identify a database management system that a consensus of the WRC PIs and Cooperators can support. Tasks accomplished to date to achieve this goal are:
   - Dr. Llebot Lorente has met with the WRC Trask PWS team to receive input on their data management process, which is based on the H.J. Andrews LTER and NSF protocols.
   - A consolidated database structure has been created on the CoF's server, consistent with the structure used for the WRC-Trask PWS.
   - Dr. Llebot Lorente has evaluated other data management systems (such as Open Science Framework) for their potential for joint use and long-term archiving of WRC data. This evaluation will be brought to the Steering Committee for their review.
   - Dr. Llebot Lorente has designed a procedure to perform a data inventory, based on a questionnaire that will be distributed to the researchers involved in WRC. The survey will

gather information about the existing data sets, how they are organized, the period of time that they cover, how many versions of the same data set there are, the researchers that are responsible for the data set, active users or active managers, documentation, and level of quality control performed on the data. Iterations between researchers and Dr. Llebot Lorente will be repeated as needed until the data inventory is complete. The data inventory plan also involves personal interviews to the few researchers more heavily involved in data management to get a sense of the workflows that they follow.

- The data obtained from the surveys and interviews will also be used to publish research on data management. The goal of the research is to evaluate the data management practices of this particular group of researchers, and understand their main challenges. We will assess how the behavior of this group can be extrapolated to others. We will also address the potential of data management projects as educational tools, evaluating whether being involved in a data management project like the present one affects the data practices of the researchers that have been involved outside of the project. An extra questionnaire has been developed for this purpose.

c. ***Pilot Using Alsea Revisited Data***. The goal for this activity is to demonstrate the data management approach using the Alsea Revisited dataset. To date, we have achieved:
   - We have completed the initial inventory of the Alsea Revisited datasets, and are currently confirming that it's up to date.
   - The Alsea Revisited PWS datasets on CoF servers have been consolidated into a single folder in the consolidated WRC data folder. This activity required significant effort since datasets were spread over eight different folders in three different groups on the CoF server. We are also working to insure that any datasets residing on individual PIs computers are duplicated in the consolidated dataset.
   - Dr. Catalina Segura has completed a draft QA/QC workup of the Alsea Revisited streamflow data for Flynn, Deer, and Needle Branch Creeks. Beyond the QA/QC, this includes assessing uncertainty on the flow estimates, tagging empty cells, and saving these in an archival format.
   - We are currently working with Terry Bousquet at NCASI to insure that her QA/QC results are incorporated into the Alsea Revisited dataset.

d. ***Develop Future Strategy***. The goal of this activity is to use the results from the Alsea Revisited pilot to develop a strategy, including effort and costs, to transition the WRC datasets into the data stewardship framework. We have not made sufficient headway on the other activities to report any progress on this activity.

Problems, barriers, proposed changes to objectives:
Our agreement with the OSU Valley Library has resolved one of the biggest barriers to project success: finding a well-qualified and supervised data management specialist. This arrangement increases our confidence that we will be able to create a state-of-the-art data management program for the WRC.

As we began consolidating the various paired watershed study datasets, concerns were raised about data ownership and intellectual property rights. Review of the WRC agreements revealed that the MOA with the funders specifically vested in the WRC ownership rights and publication clearance. However, there were no equivalent agreements with the researchers conducting the studies. The WRC Trask PWS PIs has a publications policy that they created in 2011, but this was never approved by the WRC Advisory Committee; there are apparently no publications policies for either the Hinkle or Alsea Revisited PWSs,

although some of the PIs have worked on a draft. At the WRC's 2016 Business Meeting, a Publications Policy Committee was formed to create a unified process for all three PWS and for future data syntheses among the studies.

Planned Work:
In our proposal, we presented a timeline for the project. We anticipate largely following this schedule, with the primary focus on getting the Alsea Revisited data into the archival database structure and developing the data management strategy. During mid-December through mid-March, we expect that Dr. Llebot Lorente will be on maternity leave, but have frontloaded her work to accommodate that absence. However, we may adjust the Steering Committee meeting dates depending upon her availability and project progress.

The planned strategy to complete the data and workflows inventory during the next month is**:**

- Get IRB approval for the surveys and interviews
- Send the survey and start interviews.
- Summarize information gotten from the interviews and redistribute to researchers to detect any missing data. Iterate with the researchers as many times as necessary until the data is complete, via questionnaires or personal visits.
- Summarize information from surveys (data inventory) and interviews (workflows) in a report.

List of names and brief overview of graduate and/or undergraduate engagement in project:
None

List of presentations, posters, etc.:
Data management was one of the four focus areas at the WRC's 2016 Annual Business Meeting on October 4, 2016. Dr. Llebot Lorente presented the structure and outline of the anticipated data management plan to the attendees.

List of publications, thesis citations:
N/A