Data Life Cycle subcommittee (RCAC-DLC)
Minutes November 21, 2011 (taken by Erik Deumens)

Present: Sophia Accord, Paul Avery, Peter Barnes, Reed Beaman, Erik Deumens, Rolando Millian, Hannah Norton (Guest), Dave Pokorney, Laurie Taylor, Lois Widmer (Guest)

Report: Summary of Data Life Cycle Management
See presentation slides reproduced at the end of these minutes.

Discussion: Data Life Cycle Management
Reed Beaman alerted the committee of the following request for comment from the Office for Science and Technology Policy:

OSTP Seeks Comments on Public Access to Data and Results of Federally-Funded Research. On November 4 OSTP issued two separate Requests for Information (RFI) seeking recommendations for (1) ensuring "broad public access to the peer-reviewed scholarly publications," and (2) ensuring "long-term stewardship and encouraging broad public access to unclassified digital data" that results from federally funded research. Comments regarding public access to scholarly publications are due January 2, 2012; comments on access to digital data are due January 12, 2012.

The committee agrees that drafting a response will be helpful for our own effort to define the needs for data management. Here is a helpful article from the Society of Scholarly Publishers that contextualizes the OSTP request for information and formulates approaches for responding to it productively:
http://scholarlykitchen.sspnet.org/2011/11/22/realistic-approaches-to-the-us-federal-public-access-rfis/

Sophia Acord contributed the following, relevant observation from a paper she is writing about academic needs and values:

Data sharing is also greatly impeded by scholars' lack of personal time to prepare the data and necessary metadata for deposit and reuse (which includes the sometimes Herculean efforts of converting analog data to digital formats, or migrating old digital formats to new ones). For scholars focused on personal credit, narrowly defined, there is no advantage to spending time (and grant funding) curating data, when that same time can be applied to the next research project and/or publishing books and articles. While data sharing may be facilitated by development of new tools and instruments that ensure standardization (such as in gene sequencing), the idiosyncratic ways in which scholars work, and the extreme heterogeneity of data types in most non-computational fields, do not lend themselves to one-size-fits-all models of data sharing. The escalation of funder requirements (e.g., NSF, NIH) for sharing data management plans points to an important space to track. We predict that faculty will not be doing the work, but rather a new professional class and academic track (perhaps akin to museum curators, specialist librarians, or tool-builders) may emerge to take on these new scholarly roles (cf: Borgman, 2007; Nature, 2008; Science, 2011; Waters, 2004). In sum, until issues of time and peer review are worked out, we predict an uneven adoption of sharing and publishing data openly.

The full paper can be found at http://nms-theme.ehumanities.nl/manuscript/credit-time-and-personality-acord-and-harley.

The reason for trying to define data life cycle management as an activity and develop a process and best practices for it, is that the researcher, as pointed out by Reed, does not always make the right decision, and indeed, may not be able to without further insight, help, or work. The problem is made more complex by the need to share data with collaborators and protect it from access by competitors or the general public during some stages of the research. At other stages some may be deleted and other parts pf the data need to be made public.

In addition, there are no universal standards. Different research communities have developed and are developing standards and best practices for the data relevant to their subject matter. IT providers in support of research, such as Research Computing at UF, and libraries are asked to work with all types of data in an efficient and cost effective way.

Sophia remarks that UC Berkeley is playing with the "Scholar's Box" project. They have a more educational focus, but many of the principles about creating a space for data storage/manipulation/curation are the same: http://raymondyee.net/wiki/ScholarsBox.

The meetings of Dec 5 and Dec 19 have been canceled to be able to work off the backlog of material. The agenda for the spring semester will be developed in the first week of January 2012.