Research Computing Advisory Committee

Minutes Mar 4, 2016 (taken by Erik Deumens)

Present: Paul Avery (Skype), S. Balachandar, Peter Barnes, Hai-Ping Cheng, Erik Deumens, Rafa Munoz-Carpena, Richard Hennig, Lauren McIntyre, Alberto Riva, Plato Smith, Jack Stenner, Laurie Taylor

Updates

- HiPerGator 2.0 is now in early user mode since Mar 1st. See https://www.rc.ufl.edu/services/computation/hipergator/ for details. April 15 will be the start of the migration of users from HiPerGator to HiPerGator 2.0 for production use. The migration is expected to be complete by the end of May.
- ResVault is also used by a small number or early users and is expected to go in full production early May. See https://www.rc.ufl.edu/services/restricted-data/researchvault/ for details.
- The Oak Hammock Big Data lecture series is going on this month and next: Each lecture will start at 10:00 am and last 45 minutes, with 15 minutes for questions.
    - March 18 Forrest Masters
    - March 25 George Michailidis
    - April 1 Sara Gonzalez
    - April 8 Paul Avery
    - April 15 Patrick Tighe
    - April 22 Richard Hennig
- Plato Smith is the new UF Data Management Librarian and a new member to the RCAC.

Discussion

Data policy development

One important issues with data management is that data becomes stale and manual investigation to sort data into what must be kept, what must be backed up, what can be deleted is too expensive and time-consuming.
A simple process and best practice where data gets tagged in some way at the time of creation to designate ultimate disposition would help. Because then an automatic process can be run when the data owner decides is a good time to execute and e.g. remove obsolete files.
The tagging can be by extra metadata files, it can be in the file naming convention. There are many ways to implement this, but the value for managing data is great.

Retaining data indefinitely

Granting agencies often require that data be available after the work is done. At least for 2 or 3 years depending on the agency. Prof. McIntyre asks what infrastructure UF has to meet this need?

The Libraries operate the Institutional Repository (IR). There is no cost to the researcher; the infrastructure and operation is funded by the Libraries. It now holds 120 TB of data and is mostly used for thesis and similar sized data. Multi-TB data sets are too large for this infrastructure.
Some researchers buy storage capacity from the grant from Research Computing for the required 2 or 3 years and then RC makes the data available through a suitable web interface developed with the research team. This is part of the pubapps service of RC. This is an open and unsolved problem.

Organization

Next meeting will be on Monday April 4, 2016 at the usual location in NPB 2205 from 1:30 – 2:30 pm.