**Data Management/Curation Task Force[[1]](#footnote-1)**

**Wed., Nov. 13, 2013, 1-2pm; Library West 419**

**Members :** Hannah Norton, Laurie Taylor, Rolando Garcia-Milian, Denise Bennett, Val Minson, Joe Aufmuth, David Schwieder, Blake Landor, Mark Sullivan, Sara Russell Gonzalez, Erik Deumens, Robert Ferl, and Cecilia Botero; *Invited:* Matt Gitzendanner and Aaron Gardner

**Draft Agenda**

**Updates and Discussion:**

* Discussion/updates from work and activities of interest at UF and external
  + Research Computing Day
  + Internal Library presentation
* Work towards Year One Report, next steps with strategic recommendations, noting specific problems/projects
* 11/27 meeting: Mark Sullivan presenting on the IR@UF data support now, and possible for futures
  + Question/discussion topic: Tool/portal to provision dinky databases and data websites (similar to what REDCAP does for clinical trial surveys)

**Notes Toward Year One Report**

* Survey: recommendation for the libraries to do an annual survey
  + Consistent focus on data, Research Computing, HiPerGator, and HPC
  + Added/changed sections as needed for data activities
* Possible recommendations for librarians right now
  + For ALL LibGuides, add DATA page
    - Include links to subject repositories
    - Include link to IR@UF
    - Include link to Data website
    - Include link to DMPTool
* Review of Big Data, Little Data event evaluations and feedback (overall for future events, trainings, and for reporting key points)
* Dinky Databases[[2]](#footnote-2)
* ETDs and data
* IR and ORCID
* Visualization
* Creating data project page[[3]](#footnote-3) to promote understanding of what “data” means for different research fields, options for managing it, and way to connect/provide context for case studies
* Creating case study listing for research data for projects/groups that are excellently managed (and thus link to resources on campus that are supporting)
* Support for collaboration among/across groups
  + DH Academic Production Specialist in SPOHP
  + Relevant existing as well as new positions with the Informatics Institute
  + Different areas, different staff hiring classifications, etc.; including support for collaboration across most closely related IT groups (Research Computing, Libraries’ Digital Development, ICBR IT Team…?)[[4]](#footnote-4)
* Identification of specific ways to reach data goals
* Organize recommendations by functions, places, etc. For instance, perhaps establishing “data labs” as places/staffing/times at different library branches, like InfoCommons in LW.
* Activities related to Research Computing outreach (in collaboration with RCAC)
* Define specific needs/goals and create supports or a plan so that librarians can support, use, and promote the IR@UF for data, when applicable/appropriate
  + Developing guide on when and what type of data to submit to the IR
  + Planning for how the group can approach future work, possibly for developing specific cases (as with DVN) for campus needs, and then translating into functional and non-functional requirements specifications for data support within the IR, or elsewhere (with defining when, where, and how appropriate, etc.)
* Plan for integration with existing trainings and new trainings (SobekCM and data, reference and data,
* Plan for integration with existing related library events (GIS Day, DH Day, InfoCommons events, etc.)
* Plan for integration with courses, and for developing new courses
  + Integration with existing classes (e.g., team-taught classes with Subject & Library Faculty; examples: Preserving Archives, Graduate Research Methods, etc.)
  + Integration with new courses in development (e.g., team-taught classes with Subject & Library Faculty; examples: Digital History Lab; Introductory Concepts in Research Computing)
  + Data courses
    - Collaboration with Sara and the Instruction Committee for possible new data literacy course (could, in some ways, parallel support for the information literacy course); possible new group on this for next year?
* Group structure: recommend remaining as-is, changes, additions? Changing regular meeting days/times?
  + What goals can be met through connecting to related groups and how should this be done?
    - Possible example: Connecting with Sara as a member of DMCTF and the Library Instruction Committee for support on a Data Literacy course like that of the Information Literacy course.
  + As a task force or as a committee, with task forces?
  + Already best, or how to best connect for questions on: copyright and rights; records management; born digital records group in special collections; etc.?
  + New structures? Possible examples: Content Stewardship at PSU (<http://www.diglib.org/archives/5288/>), Center for Digital Research and Scholarship at Columbia, etc.

**Upcoming events scheduled and to be discussed/planned**

* Nov. 12, 10-11:30am, Smathers Library 1A: 20-30 minutes during the SRRS Meeting  
  Librarians presenting within the libraries on activities to date, expected future activities; discussion points/recommendations may include:
  + All LibGuides to include links to relevant repositories; possible updates to position descriptions; affiliate faculty status in subject faculty units…
* Nov. 20: GIS Day[[5]](#footnote-5)
  + In future years, combining data event like 10/3 with GIS day?
* Zotero workshops (citation management software for data in bibliographic databases and connects to many tools for text/data mining)
* DMPTool, scheduling hands-on training
* Workshop for outreach for HiPerGator

**Resources**

* Meetings: alt. Wed.; HSC Library C2-41, Library West 429, Marston Science Library L107

**Ongoing**

* Planning and supporting different informational, training, and outreach activities and events on data and related resources like HiPerGator
* Workshops (types for different groups: researchers, and data service providers); known needs:
  + DMP Tool for Librarians (and other Data Liaisons/Supporters to be identified)
  + DMP Tool and creating a plan
  + Possible workshop: Primer on Data Management, 2 hour version, expanded primer within 2 day workshop, co-taught with teaching faculty in-field; expanded primer within lab-style courses as with research and methods courses, etc.

**Deadlines/Events**

* November:
  + Presenting to libraries; work on survey result analysis; RC Day; GIS Day
  + Work towards larger Year One report and strategic directions/recommendations
  + Quarterly report due for July-September
* 2014 January:
  + Quarterly report due for October-December
  + Year One Report, draft due to group[[6]](#footnote-6)
* 2014 February:
  + Year One Report due to Deans of the Libraries
  + Future surveys/data gathering for feedback on data needs with possible questions[[7]](#footnote-7)

**Initial Draft for Discussion**

The initial draft notes below are towards a possible course to aid in translation competency with data (for working with Data Scientists, no prereqs, not necessarily heavily technical, etc.). The course could draw on theories of the database age, procedural rhetoric, data provenance for reproducible research, and help frame questions and learning for changes in working, thinking, and doing scholarship and research overall in the Data Age. Readings could include Manovich, Bogost (*Persuasive Games*: “practice of authoring arguments through processes […], through the authorship of rules of behavior, the construction of dynamic models” (29).

**Introductory Concepts in Research Computing**

**3 Credits**

**Fall/Spring, or Summer A/B compressed course**

**Undergraduate/Graduate sections possible (at what level?)**

**Purpose**

Working with “Big Data”, or large numbers of digitized texts, images, sounds, and other information sets enables students and researchers to ask new and exciting questions in their fields. This research is termed ‘computational’ because it involves harnessing computer power to examine more sources than is possible by any individual or team. This course is an introduction to the basic concepts that will enable students to collaborate with computer scientists to develop or support computational research projects in different fields. The primary goals of the course are to help researchers to determine what types of data modeling tools to use for their research, and to provide an introduction to associated computing concepts. This course will not teach or involve computer programming.

**Prerequisites:** None. Anyone interested in using computers for research is encouraged to attend.

**Format**

Classes will be part lecture, discussion, and guided inquiry with hands-on examples to work through different concepts and learn different programs. Students will produce a computational research proposal at the end of the course.

**Course Content**

Overview of Research Computing, Common Uses and Tools

What are Data, and Where Do They Come From?

Computer Simulations

The Monte Carlo Method

GIS

Data Mining and OCR

Visualizations and Everything Else

Unit Operations

Procedural Rhetoric

Grounded Theory Approaches to Analysis (Functional and non-Functional Requirements)

Introduction to \*nix and Shell Scripting

Other Systems Operations

Overview of Applications, Programming Languages, and Libraries used in Research

Commercial Software Examples

Open Source Software

Package Managers

Scripting Languages

High Performance Compiled Languages

Brief Overview of [Parallel Computing](https://computing.llnl.gov/tutorials/parallel_comp/#Whatis) Techniques and Resources

GPUs and Moving Data

On-Campus resources: HiPerGator

Data Management

Data Storage and Curation

Ethics of Big Data

1. Data Management: <http://www.uflib.ufl.edu/datamgmt> & DMCTF resources: <http://ufdc.ufl.edu/AA00014835/> [↑](#footnote-ref-1)
2. <https://docs.google.com/spreadsheet/ccc?key=0AoYPOTobTSykdDNUeEprN01DQzlBdE10T19BRnRLRUE#gid=0> [↑](#footnote-ref-2)
3. Possibly like that for DH: <http://cms.uflib.ufl.edu/DigitalHumanities/UFDigitalHumanitiesProjects> [↑](#footnote-ref-3)
4. To support connecting “application engineers” as those in roles that are neither research engineers nor enterprise systems, possessing skills from programming, system administration, database administration/engineering, system analysis, application development, etc. They work both at web scale (many dependencies and interactions, large scale and live environment, etc.) and with work defined in relation to research needs (exacting standards, data loss/corruption unacceptable, needs for provenance auditing and etc.), often acting as translators and glue people, and ideally positioned for collaboration with the libraries for campus-wide data support. Leta Hunt, Marilyn Lundberg, and Bruce Zuckerman (“Getting beyond the common denominator,” <http://llc.oxfordjournals.org/content/26/2/217.full>) provide an extended explanation of application engineers. [↑](#footnote-ref-4)
5. <http://guides.uflib.ufl.edu/geog> [↑](#footnote-ref-5)
6. See charge and notes: Draft proposed recommendations as whitepapers for review/approval/implementation to include: Recommendations for the Libraries’ campus-level role in support of data management and curation; proposing a corresponding framework and resources for library support of the data life cycle; recommending the role of the institutional repository and research computing in storing, finding, and accessing working and final data, and linking publications to supporting data; and, recommending a framework for liaisons and subject specialists to incorporate data instruction and consultation into their workflows. Outline with detailed plan for training and other supports based on information gathered during Focus Groups, survey, and other activities; plan for ideal (more resources) and for conservative (current resources); Outline with detailed information on how the IR fits in the overall supports for data; and same for other applicable resources that can be used/leveraged as is now, and detailed information on how to enhance or make best fit [↑](#footnote-ref-6)
7. Possible questions:   
   -- How would you like authenticated users to be able to interact with the data on-line, if you were to make it available? [Download only; Search on site, no download; Run statistical analysis across my data; etc.]

   -- What type of data visualizations would you like authenticated users to have access to regarding my data on-line? [A, B, C, D, etc., write-in]   
   -- If you (or other authenticated users) could add individual records through a form on the online system, would you transfer the data to the system and rely on it for working access and long-term preservation? [↑](#footnote-ref-7)