**Data Management in the Research Environment - Syllabus**

**Instructors**: Dr. Timothy Norris – Postdoctoral Fellow Miami University Libraries

Angela Clark – Librarian Associate Professor RSMAS Library

**2 Credits:** 600 level course

**Texts and Materials:**

There are no required texts for this course. The readings are either available online or will be made available through the course website.

**Software:**

We will be using some very basic software packages all of which are open-source (no license fee). You will need to install the following programs on your computer.

* Open refine (previously google refine) - <http://openrefine.org/>
* Git Bash – see: <https://software-carpentry.org/v5/setup.html>
* Text editor (choose one) – any OS: [Gedit](https://wiki.gnome.org/Apps/Gedit); Windows: [Notepad++](https://notepad-plus-plus.org/); Mac: [TextWrangler](http://www.barebones.com/products/textwrangler/)

**Course Description:**

The purpose of the course is to develop understandings of research data in broader spatial and temporal contexts--known as the research data lifecycle, to introduce several practical tools for digital scholarship, and to encourage early adoption of best practices in research data management. *The course will provide students with strategies to increase productivity (efficiency), enable proper data stewardship (security), and help the student exceed data management expectations/requirements in the research environment (compliance).* This is a practical course: students are required to produce a data management plan for their specific research endeavor, OR to prepare and deposit data into a discipline specific repository (other projects subject to instructor approval will be considered). The class is open to all graduate students in all disciplines.

**Prerequisites:**

**Students should have a good idea of what their MS/PhD research will be.** If no project is identified at time of enrollment admission will be based upon instructor approval.

**Measureable Learning Outcomes:**

1. Describe data lifecycle models and how they inform data management planning
2. Identify file formats, data types, data levels and relevant software and understand how they inform data management and preservation.
3. Design best practices for file system organization and file naming conventions to serve sound data storage, backup, and preservation strategies.
4. Gain practical experience in discovering, acquiring, and cleaning data.
5. Produce documentation and metadata for research data to facilitate discovery and re-use.
6. Evaluate legal and ethical implications for data access and sharing strategies.
7. Identify discipline specific or institutional data repositories and prepare data for deposit.

**Methods of Instruction:**

Lecture-discussion, demonstration, and practical out-of-class and in-class assignments. Students are expected to participate in the lectures, discussions, and in-class projects. We expect that students have little or no prior data management experience/training, but students with previous data management experience will also benefit from the course.

**Evaluation of Student Performance:**

Students will be assigned a letter grade (A-F) based on the following:

|  |  |
| --- | --- |
| Homework Assignments | 20% / 50 pts. |
| Class Participation | 10% / 25 pts. |
| Midterm Project | 30% / 75 pts. |
| Final Project | 40% / 100 pts. |
| Total | 100% / 250 pts. |

**Assignments**

There will be several homework assignments throughout the semester. This will include reading and writing assignments, online tutorial assignments, and data wrangling assignments.

**Class Participation**

Engagement in class activities through speaking, contributing comments to the class web-page, or participating in group projects is an important aspect of this class

**Midterm Project**

Create a data curation profile of your or your laboratories research. This includes thinking through and responding to your choice of three questions from the data curation profile toolkit (Purdue Universities, see links in schedule below).

**Final Project**

Create a data management plan for your research project OR prepare/curate data for deposit into a discipline specific repository.

**Weekly Schedule:**

All topics/dates/assignments are subject to revision/change based on instructor decisions and any other *force majeur* that may take place during the quarter.

Week 1: Course Introduction and Overview: What is data and why are we here?

Introduction to the course, the instructors and to each other as students. Share out the students’ research contexts and ideas about the data they will use in the course.

* Explain what data in the academic research environment are
* Show the value of data management in the research context
* Understand the federal policy context for data management and sharing

Assigned Reading:

* Holdren, J (2013). *Increasing Access to the Results of Federally Funded Scientific Research*. Office of Science and Technology Policy, Executive Office of the President. <https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf>
* National Science Foundation. *Dissemination and Sharing of Research Results*. <http://www.nsf.gov/bfa/dias/policy/dmp.jsp> (see both the sharing and data management sections).

Optional Reading:

* Posner, (2013). “Embarrassments of riches: managing research assets.” <http://miriamposner.com/blog/?p=982>.
* Karasti, H., Baker, K. S., & Halkola, E. (2006). Enriching the Notion of Data Curation in E-Science: Data Managing and Information Infrastructuring in the Long Term Ecological Research (LTER) Network. Computer Supported Cooperative Work (CSCW), 15(4), 321-358. <https://dx.doi.org/10.1007/s10606-006-9023-2>.

Week 2: The Research Data Lifecycle

Describe several data lifecycle models and how they inform data management planning.

* Explain the data lifecycle with attention to understanding “data” and “management”
* Show the variety of approaches to data management planning
* Critically evaluate existing Data Management Plans

Assigned Reading:

* Open Data Policy—Managing Information as an Asset. <https://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf>
* Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age (2009). “Promoting the Stewardship of Research Data” (Chap 4) in *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age*. National Academies Press, Washington D.C. <http://www.nap.edu/catalog/12615/ensuring-the-integrity-accessibility-and-stewardship-of-research-data-in-the-digital-age>
* Choose a DMP to read from <https://dmptool.org/public_dmps>. NOTE: “They are not vetted for quality, completeness, or adherence to funder guidelines.”
* If your research lab has an existing DMP bring it to class

Assignment #1: write a one-page evaluation of the strengths and weaknesses of a DMP (can be from your research lab)

Week 3: Data Types, Stages, and Formats [ TURN IN ASSIGNMENT #1 ]

Describe the difference between data types, data stages, and data formats and then relate these concepts in the research environment.

* Recognize different data types and structures and their related software and/or scripts
* Categorize data by its stage in the research cycle
* Identify file formats and their related software and/or scripts for different kinds of media
* Plan file format practices/choices for both research workflows and for preservation.

Assigned Reading:

* Kitchin, R (2014). “Conceptualizing Data” in Kitchin, R *The Data Revolution.* Washington DC: Sage. <http://uk.sagepub.com/sites/default/files/upm-binaries/63923_Kitchin_CH1.pdf>
* UK Data Archive (2015). Create and Manage Data: formatting Your Data: File Formats and Software. <http://www.data-archive.ac.uk/create-manage/format/formats>. (follow “Recommended File Formats” link as well).
* Library of Congress (2015). Recommended Formats Statement 2015-2016. <https://www.loc.gov/preservation/resources/rfs/RFS%202015-2016.pdf>.

Week 4: Data Organization (software carpentry: BASH shell)

Design best practices for file system organization and file naming conventions.

* Manipulate file system structure and organization through the BASH command-line.
* Write a shell script to rename all files in a folder.

Assigned Reading:

* Frazer, Meghan (2013). An Elevator Pitch for File Naming Conventions. <http://acrl.ala.org/techconnect/post/an-elevator-pitch-for-file-naming-conventions>.
* Alexandra Simperler, Greg Wilson (2015). Software Carpentry get more done in less time. <http://arxiv.org/abs/1506.02575>.

Assignment #2: Software Carpentry (2015). The Unix Shell. [ DO AS MUCH AS YOU CAN BEFORE COMING TO CLASS ON THIS DAY ] <http://swcarpentry.github.io/shell-novice/>.

Week 5: Data Storage, Backup and Security

Design data storage and backup strategies for the stages of your research project.

* Identify different hardware solutions
* Identify online solutions
* Describe advantages and disadvantages of different strategies

Assigned Reading:

* + Backing Up Data (UK Data Archive):<http://www.data-archive.ac.uk/create-manage/storage/back-up>
  + Ruggiero and Heckathorn (2012). Data Backup Options. United States Computer Emergency Readiness Team. <https://www.us-cert.gov/sites/default/files/publications/data_backup_options.pdf>.
  + Version Control and Authenticity (UK Data Archive):<http://www.data-archive.ac.uk/create-manage/format/versions>

Optional Reading:

* + Gallagher. “The great disk drive in the sky: how web giants store big -- and we mean big -- data.” <http://bit.ly/1Ll8gl7>.
  + Rosenthal, “Estimating storage costs.” <http://blog.dshr.org/2013/11/estimating-storage-costs.html#more>.

Week 6: Data Curation Profiles

Introduce the data curation profile.

* Maintain cross-disciplinary conversations about data management
* Critical analysis of current practices in data management

Assigned Reading:

* Download the Data Curation Profiles Toolkit from <http://docs.lib.purdue.edu/dcptoolkit/>. Read the introduction and skim the rest.
* Download and use/modify the interview guide template from <http://bit.ly/1Ll8gl7>.
* Choose one data curation profile to read from <http://docs.lib.purdue.edu/dcp/>.

Midterm Project Assignment: Choose three questions from the data curation profile toolkit relevant to your work. Write responses to the questions (1-3 paragraphs each response).

Week 7: Open Lab

Perform data curation interviews and follow-up. Begin drafting data curation profiles for midterm project.

Week 8: Metadata and Contextual Details [ TURN IN MIDTERM PROJECT ]

Understand the importance of quality descriptions of data and associated software and/or scripts, especially across temporal scales and disciplinary boundaries.

* Document data throughout the research endeavor.
* Produce metadata that facilitates discovery and re-use.

Assigned Reading:

* Riley, “Seeing Standards.” <http://www.dlib.indiana.edu/~jenlrile/metadatamap/>. Download and study this poster in depth.
* Star, Castro, *et al* (2015)*.* Achieving human and machine accessibility of cited data in scholarly publications. *PeerJ Computer Science* 1:e1<https://dx.doi.org/10.7717/peerj-cs.1>.

Optional Reading:

* Curdt, Constanze, Dirk Hoffmeister, Guido Waldhoff, Christian Jekel, and Georg Bareth. "Scientific Research Data Management for Soil-Vegetation-Atmosphere Data—The TR32DB." *International Journal of Digital Curation* 7, no. 2 (2012): 68-80.<http://www.ijdc.net/index.php/ijdc/article/view/220/295>
* [ for ecology: read the introduction through to 2.3.4 ] KNB (no date). “Ecological Metadata Language (EML) Specification.” The Knowledge Network, <https://knb.ecoinformatics.org/#external//emlparser/docs/eml-2.1.1/index.html>.
* [ for marine biology ] Neiswender (2010). "Introduction to Metadata." In *The MMI Guides: Navigating the World of Marine Metadata*. <http://marinemetadata.org/guides/mdataintro>.
* [ for atmospheric science: read the introduction through 2.6.2 ] Eaton, *et al* (2011). NetCDF Climate and Forecast (CF) Metadata Conventions. <http://cfconventions.org/Data/cf-conventions/cf-conventions-1.6/build/cf-conventions.pdf>.
* [ for marine geology: read the introduction through chapter 6 ] SEG (2015). SEG Y rev 2 Data Exchange Format. <http://www.seg.org/documents/51956/6062543/SEG-Y+rev+2+Draft+Jan+2015>
* [ for geographic data: read the introduction through page 31 ] FGDC (2000). Content Standard for Digital Geospatial Metadata Workbook, Version 2.0. <https://www.fgdc.gov/metadata/documents/workbook_0501_bmk.pdf>.

Assignment #3: Choose one of the following: (1) Create metadata for your research data using an existing schema (2) create a metadata schema for your data.

\*\*\*\*\*\*\*\*\*\*\*\*\*\* SPRING BREAK \*\*\*\*\*\*\*\*\*\*\*\*\*\*

Week 9: Versioning and Git (software carpentry) [ TURN IN ASSIGNMENT #3 ]

Become familiar with git and github as versioning tools.

* Understand the difference between git and github
* Create a git repository, make changes, commit the changes and push the changes to github.

No class meeting on Monday of week 9. Instead there will be two scheduled two-day “Software Carpentry” workshops held (see <https://software-carpentry.org/>):

* An R based workshop at the Shalala Student Center Senate Room on the Gables Campus on Monday and Tuesday March 14-15, 8:30am – 3:00pm (see <http://mkcor.github.io/2016-03-14-university-of-miami/>).
* A Python based workshop at the Newman Alumni Center on the Gables Campus on Monday and Tuesday March 28-29, 9:00 – 4:00.

Attendance to the workshops are not required, but is highly encouraged that at least one morning or afternoon session is attended. The more you can attend, the better it will be.

On Wednesday we will do an hands-on lab exercise is based on the Software Carpentry lesson (Tuesday morning part of the R based workshop): <http://swcarpentry.github.io/git-novice/>

Week 10: Introduction to R

Introduction to using R in the research environment

* Load data into the R environment
* Calculate averages
* Make a simple graphic

This is a hands on lab week based on the Software Carpentry Lesson: <http://swcarpentry.github.io/r-novice-inflammation/>

Week 11: Sharing and Reuse: Intellectual Property and Licensing

Evaluate how ownership and copyright of data and associated software and/or scripts impact access and sharing strategies.

* Know University of Miami Policy
* Know your rights and obligations in the research environment

Assigned Reading:

* Miami and other IP policy / data policy. Search for ‘data’ and scan relevant sections. <https://umshare.miami.edu/web/wda/facultysenate/FacultyManual.pdf>.
* Madison (2011). “Knowledge Curation.” Notre Dame Law Review, Vol. 86, p. 1957, 2011; U. of Pittsburgh Legal Studies Research Paper No. 2011-13. Available at SSRN:<http://ssrn.com/abstract=1848086>

Optional Reading:

* Boyle (2003). The Second Enclosure Movement and the Construction of the Public Domain. Law and Contemporary Problems, 66:33(Winter/Spring), 33-74. <http://scholarship.law.duke.edu/lcp/vol66/iss1/2/>.
* David (2008). The Historical Origins of ‘Open Science’: An Essay on Patronage, Reputation and Common Agency Contracting in the Scientific Revolution. *Capitalism and Society* 3(2), Article 5. <https://dx.doi.org/10.2202/1932-0213.1040>.

Assignment #4: Nguyen. “Using Google Refine to clean messy data.” [ DO THE TUTORIAL BEFORE CLASS AND SEND THE FINAL FILE TO BOTH INSTRUCTORS (.csv) ] <http://www.propublica.org/nerds/item/using-google-refine-for-data-cleaning>.

Week 12: Finding and Cleaning Data, Data Entry and Quality Control [ TURN IN ASSIGNMENT #4 ]

Identify difficulties in acquiring and cleaning data.

* Understand challenges of acquiring data online
* Use online tools for data cleaning
* Know best practices for data entry

If time permits we will do an introductory lab on Structured Query Language (SQL) or Python based on the software carpentry lessons: <http://swcarpentry.github.io/sql-novice-survey/>.

Week 13: Repository Identification / Preparation for Preservation / Unique Identifiers

Identify appropriate discipline specific or institutional data repositories

* Prepare data for deposit.
* Cite deposited data correctly

Assigned Reading:

* Rauber, Asmi, van Uytvanck and Proll (2015). Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation. Research Data Alliance. <https://rd-alliance.org/filedepot?cid=262&fid=667>.
* Fecher, B., Friesike, S., & Hebing, M. (2015). What Drives Academic Data Sharing? PLoS ONE, 10(2). <https://dx.doi.org/10.1371/journal.pone.0118053>.

Assignment #5: Explore <http://www.re3data.org/> and identify three potential repositories for your data deposit.

Week 14: The Semantic Web and Open Lab [ TURN IN ASSIGNMENT #5 ]

Understand how the content of this course fits into the broader context of the semantic web.

* Know tables, trees, and triples framework for understanding online data
* Describe the theory behind linked open data

Assigned Reading:

* Berners-Lee, T, Hendler, J and Lassila, O (2002) The Semantic Web. *Scientific American Special Online Issue***:** 24-30. Blackboard > Course Documents > Materials > Week 14.

Optional Reading:

* Berners-Lee, T, Nigel Shadbolt and Wendy Hall (2006). The Semantic Web Revisited. *IEEE Computer Society.* Blackboard > Course Documents > Materials > Week 14.

The second meeting this week will be time with instructors to work on final projects. Specifically students will exchange their final projects in pairs and provide immediate feedback.

Finals Week: [ TURN IN FINAL PROJECT ]