**DATA MANAGEMENT IN THE RESEARCH ENVIRONMENT, OF COURSE**

**Angela Clark-Hughes**

RSMAS Librarian | Librarian Associate Professor

University of Miami

 Rosenstiel School of Marine & Atmospheric Science (RSMAS) Library

**Abstract:**

After years of building the infrastructure for data literacy and data management through various library information sessions, faculty consultations, patron surveys, and workshops, the University of Miami Libraries (UML) proposed a formal, credited, graduate level course to be taught in spring 2016 at the Rosenstiel School of Marine & Atmospheric Science (RSMAS). The course was entitled *Data Management in the Research Environment*, to be taught by Tim Norris (CLIR Postdoctoral Fellow at UML) and Angela Clark-Hughes (RSMAS Librarian Associate Professor). The course was unanimously approved by the RSMAS curriculum committee as a 500/600 level two-credit course, open to all graduate students. As a new elective course, enrollment was small with two Ph.D. students, three Masters of Professional Science (MPS) students, and one Research Staff member auditing. This paper will focus on the review and assessment of that course; from the curriculum structure and design (including the implementation of two Data Carpentry Workshops), the learning curves and objectives, the teaching strategies employed, to the final student evaluations of the course. In addition, a summary of lessons learned and next steps for this course, as well as a comprehensive literature review, are also covered.

**Keywords:** Data management, data literacy, data services, data course

The field of Library Information Science has changed considerably in the wake of technological advancements. From a limited access card catalog to an openly accessible online catalog; from a carefully crafted “quick search” guided by a DIALOG blue sheet to a broad search on Web of Science then casually filtered to near perfection; from a physical reference collection to a digital collection of links on a webpage. Any librarian with ten or more years of service could easily pinpoint the day when they realized the profession had changed in a dramatic way.

As the learning and research environments in which we worked changed, so did librarians. Librarians, specifically academic librarians, modified the structure of patron services to extend beyond building hours and developed collection philosophies to include varying material formats. Librarians adapted to the social phenomenon of push/pull information and reference services as we embarked on new challenges to meet the needs of our users. In this evolutionary process to stay in step with the world around us we enlarged our territory. In part our territorial expansion was internal, purposefully crafted, and guided by the How-to-do-it manual for Librarians. However, some parts of our expansion were external and inadvertent as we rose to a new level in response to the changes in the research climate

7

Data management is just one of those new level concepts that logically landed on the library’s doorstep because libraries and librarians have historically and successfully managed many other things. As information scientists, we’ve been the searchers, the collectors, the catalogers, the curators, the

examiners, the evaluators, the advisors, and the advocators of many things for centuries, therefore the manager’s role is befitting. Although the term “data management” is relatively new, managing data is not a new concept or recent practice. It has been and is a necessary component of any research endeavor regardless of its scale. However, managing data, like any applied process, can be done well or poorly. It can be done differently across the various disciplines of study, and can be improved upon as the technology, tools, and skill sets to manage data improve. It is for these reasons that the process of managing data and/or the concept of data management has not been static or inflexible, but has seemingly metastasized.

So, for many academic libraries data management has simply come into our purview within the last ten years. In marine and atmospheric science libraries and centers, we’ve had to address the data deluge on at least some level. Even if our approach was to batten down the hatches and await its passing, there was no denying that the data gauntlet had been thrown down.

The challenge early on for the RSMAS Library was in defining “data” and assessing a feasible service of management. There were and still are a variety of definitions for data, some inclusive of everything, some exclusive of one thing or another. RSMAS chose to adopt a broad working definition of research data – essentially anything that our researchers considered their data. This definition allowed us to include any type of observational, experimental, model or simulation data, as well as compiled data, in our service assessment, whether it was physical or digital in format. We consulted the profusion of literature written on and about data management, data curation and preservation, data literacy, data science, and e-science. We also took note of the variations in who is creating data, who is or is not managing data, who is taking on the responsibility for storing and curating data, and who has created access points to data.

The RSMAS Library journey into research data management began unpretentiously with contributions to NASA’s *Global Change Master Directory (GCMD)*, with initial funding secured for hiring a few graduate students to take the lead on this project. Although the Library’s role was relegated to the entering of datasets compiled by a select group of RSMAS faculty, we were in fact dealing with data management on some level, albeit in the middle.

However, beginning as the middleman not only gave us insight into metadata and discipline specific standards and protocols, it gave us a clearer view of the data lifecycle. We also clearly understood that not all of the research data being produced at RSMAS was being processed through the Library for ingestion into the GCMD, nor could it have been, given our limitations. However, if the Library was managing only a small percentage of the datasets, then where was the larger percentage of the datasets being deposited? Our conclusion was that they could be anywhere. The library, although it was the information center, was not the center of that information and researchers had already taken the lead as prompted by various funding agencies to deposit their data in various repositories.

By 2008 we’d recognized a disconnection in the level of access to our research output. As forecast, there was a flood of scientific data. Technological advances increased computing power, decreased data

storage costs, and broadened sharing capabilities among researchers. Repositories like the GCMD were

8

taking root inside and outside of academic institutions and government agencies. With the myriad of data repositories offering open and immediate access, RSMAS librarians and information professionals looked to improve our chances of identifying and accessing a dataset created in the future as well those created in the past.

The University of Miami Libraries formulated an E-Science Task Force in 2009 to investigate relational models and look at the emerging needs for data training, services and staffing at UM. This group was comprised of the science and social science librarians, the Head of Systems, the Head of Digital Projects, the Metadata Librarian, the Associate Dean of Libraries, and the RSMAS Librarians. With varying perceptions of data, data production, and data usage among the library professional, we first needed to know what perceptions our researchers had; whether they were all creating data; whether they were all depositing their data into repositories; whether they were sharing their data; how the library could assist; and what data services could we provide.

**The Internal Environmental Scan Summary**

To address some of the lingering questions that the libraries had concerning the data being produced at the University of Miami, the E-Science Task Force constructed a thirty-question survey and selected full-time RSMAS faculty as our initial pilot population. We chose to use *SurveyMonkey®* to administer this survey online and to guarantee anonymity to all faculty participants. Thirty-five RSMAS faculty members responded to the online survey; however, not all respondents answered all the questions.

Given the range of disciplines of our survey respondents, it was not surprising that they reported an array of commercial software used as their main sources for data collection or analysis (Microsoft Word, Excel and Access, Matlab, SAS, Fortran, GrADs, Sigmaplot, Sigmastat, and others). Of the respondents, 84.4% reported that the main storage platform for their data was locally (in their labs or offices); 12.5% reported using an external data center, and 6.3% reported using another University platform (non-RSMAS).

Although we had prior knowledge that our researchers were depositing their data in various repositories like Woods Hole, PANGAEA, NOAA, GCMD, and others, 62% of the respondents indicated that they weren’t depositing their data in any repository. In addition, 57.6% reported that their data could not be re-created if lost or destroyed. However, when asked if their current funding agencies required data archiving, preserving or depositing, 46.9% answered “Yes.”

The survey results certainly answered some of our questions. However, it also raised some questions and prompted us to delve deeper into the methods of data production and the quantity of data produced at RSMAS verses the formatted and usable data.

**Data Curation Profile Interviews**

The next steps for the *E-Science Task Force* was to conduct our own version of the Data Curation Profile interviews like the ones being conducted at Purdue University Libraries and the Distributed Data Curation Center (DDCC).

Continuing to use the RSMAS Faculty for our pilot, we conducted a handful of interviews using the tools provided by the DDCC. The questions in the profile were designed to open the discussion on the data

9

being produced by each faculty member interviewed. Through these detailed discussions, the E-Science Task Force hoped to assess potential service needs and ascertain the likelihood of there being a final data product or dataset that could be deposited in an established repository or perhaps in our own institutional repository.

Question 10 of the Curation Profile asked, “How should the data be made accessible?” The responses to this type of open-ended question occasionally prompted the suggestion that unprocessed data be stored within the library, with librarians providing the access when requested. Although this model has worked for many libraries as well as the RSMAS Library as a temporary model of access for data notebooks, slides, floppy disk, and hard drives, it fails as a long-term solution to data management due to limitations on space availability, environmental control, and the necessity to update and maintain software for future retrieval.

The curation profile interviews and the environmental survey convinced us that a large amount of data lived on spreadsheets and local databases, on personal servers and hard drives, in offices and labs throughout the RSMAS campus. In our attempt to define our role in this era of data service we realized that RSMAS Library did not have the infrastructure for data storage, nor the staffing levels and expertise to process data for access. However, we could stay abreast of the current trends and issues to make recommendations for appropriate storage facilities or storage options. We could offer best practices, suggest appropriate repositories and data centers for the varying types of data, and support the planning of future research and data management planning with offerings of the DMPTool, ArcGIS, SAS, and SPSS*.*

**Data Literacy Symposium**

With a rough outline of the data services that we could immediately and feasibly offer at the RSMAS Library, I was given the opportunity to attend the Data Information Literacy Symposium(DIL) held at Purdue University in 2013 to hear firsthand what other libraries were doing to formally implement data services at their institutions. The DIL Symposium was designed to assist librarians in cultivating strategies and approaches for developing their own data programs. It was the culmination of a two year funded collaboration between Cornell, Purdue, University of Minnesota, and University of Oregon Libraries with each institution tasked with assessing the needs of their graduate students to manage, use, and curate their research data; interviewing faculty and graduate students in various disciplines to identify specific data needs; and responding to the needs with educational programming of varying types (mini courses, online courses, embedded librarianship, workshops, readings).

Twelve core competencies were developed as a result of this project to establish a baseline for a data literacy program, which have been outlined and elaborated on in their publication (Carlson et al. 2011; Carlson & Johnston 2015):

• Introduction to database and data formats

• Discovery and Acquisition of Data

• Data Management and Organization

• Data Conversion and Interoperability

• Quality Assurance

• Metadata

10

• Data Curation and Re-use

• Cultures of Practice

• Data Preservation

• Data Analysis

• Data Visualization

• Ethics, including citation of data

Implementing the core competencies and the lessons learned from the DIL Symposium, we began using every opportunity to raise awareness and advance the cause of data literacy, data management, and data services in our library workshops, one-on-one consultations, and classroom instruction sessions. Thanks to several of our colleagues who published their curriculum through Creative Commons and made their work available online we were able to assemble a portfolio of information to pass on our faculty, but specifically to our graduate students, lab managers, teaching assistants, and staff researcher.

**The Future and the Past**

Spreading the word that properly preparing your data promotes compliance for funders and repositories and facilitates better access was just the beginning of our new data service model. Since we had long since moved away from the data input model due to the loss of student funding, this was our opportunity to extend ourselves as liaisons to data resources and professionals in data management. However, our responsibility was two-pronged: to move forward and proclaim the best practices and guidelines for managing data in the research environment, while reaching back to reclaim access to the data dispersed in various repositories and provide access to relevant data published in technical reports.

To accomplish our two-pronged service model, we expanded our team, our skills, and our approach to providing instruction. In the spring of 2016, I teamed up with Dr. Timothy Norris, a CLIR Postdoctoral Fellow hired as UML’s Data Curation Specialist, to co-teach a 2-credit graduate level course entitled, *Data Management in the Research Environment*. The foundational material and curriculum for this course was compiled from various sources and gleaned from colleagues who openly shared their outlines, guides, presentations, and even their time as guest lecturers. However, the most significant influences and contributors have been included in the references.

**Feedback/Lessons Learned and Next Steps**

The first data course at RSMAS was considered a success with five students enrolled in the course, three Masters students and two Ph.D.s. As with any course curriculum, we realized that changes to the lessons and reading assignments would have to be made constantly to incorporate the real-time data policies and procedures, related news and events in data management, relevant readings, and any technology shifts. However, the overall feedback and course reviews were very positive, and the suggestions ranged from additional lab time needed for the software and programming lessons, to creating course levels of data management, an intermediate level and an advanced level. The advanced data management course would likely encompass more of the data carpentry instruction for R, Python, and other software like GitHub, OpenRefine, MATLAB, or LaTex.

A course designed to teach best practices for managing data in a research environment was a clear beginning for UM Libraries to offer our services, but would not and could not appeal to all who need or

11

desire to know more about data management. So in the summer of 2016 the UM Libraries launched our Data Services webpage to provide students, research staff, lab technicians, lab managers, and faculty with resource links and contacts to manage their data and stay informed.

The data course at UM/RSMAS will have its second semester in spring 2017. The content will be retooled and updated to reflect the changes in our US administration, and the lessons learned will be incorporated as much as possible. But is our desire to continue teaching the 2-credit graduate level course to promote and foster a culture of good data management at UM/RSMAS for as long as we can. However, we also realize the need to begin parts of the data and data management discussions much earlier, perhaps on a junior or senior undergraduate level.

**References**

Briney, K. (2015). *Data management for researchers: organize, maintain and share your data for research success*. Exeter: Pelagic Publishing, UK.

Carlson, J., Fosmire, M.; Miller, C.; and Sapp Nelson, M. (2011). "Determining data information literacy needs: a study of students and research faculty". *Libraries Faculty and Staff Scholarship and Research.* Paper 23. <http://docs.lib.purdue.edu/lib_fsdocs/23>

Carlson, J., & Johnston, L. R. (2015). *Data information literacy: librarians, data, and the education of a new generation of researchers*. West Lafayette, Indiana: Purdue University Press, 271p.

Corti, L., Van den Eynden, V., Bishop, L., & Woollard, M. (2014). *Managing and sharing research data: a guide to good practice*. Washington, DC: Sage.

Fecher, B., Friesike, S., & Hebing, M. (2015). What drives academic data sharing? *PLoS ONE*, 10(2): e0118053 <https://dx.doi.org/10.1371/journal.pone.0118053>.

Kitchin, R. (2014). *The data revolution*.Washington DC: Sage.

Neiswender, C. (2010). Introduction to metadata. In: The MMI guides: navigating the world of marine metadata. <http://marinemetadata.org/guides/mdataintro>.

Posner, M. (2013). Embarrassments of riches: managing research assets. <http://miriamposner.com/blog/?p=982>.

Rauber, A., Asmi, A., van Uytvanck, D. & Proll, S. (2015). Data citation of evolving data: recommendations of the working group on data citation. Research Data Alliance.

<https://rd-alliance.org>.

Simperler, A. & Wilson, G. (2015). Software carpentry- get more done in less time. <http://arxiv.org/abs/1506.02575>.

Whitmire, A. L. (2015). Implementing a graduate-level research data management course: approach, outcomes, and lessons learned. *Journal of Librarianship and Scholarly Communication*, 3(2) [doi:10.7710/2162-3309.1246](http://dx.doi.org/doi%3A10.7710/2162-3309.1246).

**Resources**

Data One Education Modules

<https://www.dataone.org/education-modules>

12

Purdue Libraries, DIL Instruction Program Planning

<http://guides.lib.purdue.edu/dil>

<http://www.datainfolit.org/>

OSU Libraries, Research Data Management

<http://guides.library.oregonstate.edu/research-data-services>

University of Minnesota, Data Management Course

[https://sites.google.com/a/umn.edu/data-management-workshop-series/](https://sites.google.com/a/umn.edu/data-management-workshop-series/home-1?pli=1)

MANTRA, Research Data Manage Training

<http://datalib.edina.ac.uk/mantra/>

RDMRose Learning Materials

<http://www.sheffield.ac.uk/is/research/projects/rdmrose>

Lamar Soutter Library, New England Collaborative Data Management Curriculum

<http://library.umassmed.edu/necdmc/index>

13