Data Curation Initiative
University of Miami

2015-16 Report and Recommendations

Dr. Timothy B. Norris
CLIR Postdoctoral Fellow in Data Curation

Sarah L. Shreeves
Associate Dean of Digital Strategies

University of Miami Libraries

October 2016

About the Authors:

Dr. Timothy Norris is a CLIR (Council for Library and Information Resources) postdoctoral fellow in Data Curation. He joined the University of Miami Libraries in the summer of 2014 after completing his doctoral degree in Environmental Studies at the University of California at Santa Cruz. His research is data intensive with a focus on the use of geographic information systems and applications in sustainable governance of natural resources. At the University of Miami this work has expanded to include thinking on the sustainable governance of information resources—data.

The CLIR postdoctoral appointment in the University of Miami Libraries is the first of its kind in the history of the University. In 2008 several institutional actors at the University of Miami recognized the need to develop services around the curation of research data. Hosting a CLIR postdoctoral fellow was suggested as a way forward and the position was created as soon as was possible. This postdoctoral research fellowship is part of a much larger network of CLIR fellows across the United States. Please see http://www.clir.org/ for more information.

Sarah L. Shreeves is the Associate Dean for Digital Strategies at the University of Miami Libraries. She provides leadership and direction for the UML digital infrastructure and technology planning. This includes overseeing the Libraries' digital production program and infrastructure development for all of the Libraries content management systems and repositories, providing leadership within the Libraries on the creation and curation of digital objects for learning and research, and ensure a robust technical infrastructure to support a wide range of digital scholarship and scholarly publishing.

Prior to coming to the University of Miami in 2015, she was the Coordinator for the Illinois Digital Environment for Access to Learning and Scholarship (IDEALS), a set of services and collections supporting scholarly communication (including the institutional repository) at the University of Illinois at Urbana-Champaign. She also served as the co-Coordinator for the Scholarly Commons, a space for expert, interdisciplinary research support services and open workshops for faculty and graduate students to develop skills in areas such as digital content creation, management of research data, understanding copyright issues and author rights, and working with geospatial and numeric data. She was responsible for working with faculty, students, and researchers on a range of scholarly communication issues including author rights, open access, theses and dissertations, data management, and data curation. She serves on the Steering Committee for the Open Repositories Conference, and on the Advisory Committee for the Digital Library Federation.

Data Curation Initiative
University of Miami

2015-16 Report and Recommendations

Contents

Executive Summary

Data curation—the management of data throughout its research lifecycle—is now considered a necessary collaborative effort to be organized at and among research institutions worldwide. At the University of Miami several institutional actors including the Libraries, Information Technology, and the Center for Computational Science recognized the need to start building services around the stewardship of research data. In early 2015 a formal data curation initiative was launched under library leadership with four main components: socialization of the initiative within the research community, assessment of data curation practices and needs, piloting data management/curation services, and programmatic development of data services.

The socialization and assessment phases used a mixed-method anthropological research design to develop an understanding of data curation practices across all three University of Miami campuses. The socialization work started prior to the 2015 initiative launch and is still ongoing. The bulk of the assessment phase was undertaken from March to June of 2015 and consisted of many informal conversations and over twenty formal semi-structured faculty interviews. Two principal results have emerged from these processes: a deeper institutional knowledge of data curation practices and a growing sense of community around data curation at the University of Miami.

Data curation develops in an uneven manner, with some scholars and research groups exercising good curation practices within their projects; however, these often tend to be ad-hoc solutions that may not serve for long term access to and preservation of research data. We also see research projects with little or no curation practices in place. Long-term data curation requires human and machine readable descriptions of data—metadata—as well as a good understanding of the principles of digital preservation; there is a lack of human resources and knowledge to implement these practices. In addition to these lacunae, the institutions that govern data curation, understood as rules of the game and community based norms, are underdeveloped at the University of Miami.

We make six recommendations in Section VII:

- Establish a Research Data Service, a collaboration of the Libraries, Information Technology, the Office of Research, and the Center for Computational Science to provide research data curation infrastructure and services.
- Establish clear policies and guidelines for research data governance and stewardship.
- Develop or license research data repository infrastructure.
- Develop data management services including consultation on data management plans, policies, external research data repositories, and sensitive data.
- Develop curricula for data management including workshops and credit-bearing courses.
- Continued assessment and socialization of data curation practices.

The report that follows elaborates on the findings of the first phases of the initiative and provides context and detail for the subsequent recommendations. Continued effort to strengthen the data curation initiative will leverage opportunity for collaboration and community building across all scales of the academy and ensure that the University of Miami remains a leader among research institutions.

I. Introduction

Data curation—the management of data throughout its research lifecycle—is now considered a necessary endeavor to be organized at, and among, research institutions worldwide. This need is driven two principal factors. First, cultural norms within the academy are emerging around improved data management practices that enable more productive research and embrace data sharing (the carrot). Second, many funding agencies now require data management plans for data sharing and data preservation in order to leverage investments in research (the stick). Overarching both of these drivers is rapid technological advance which is in turn changing the practice of science. At the University of Miami institutional actors including the Libraries, Information Technology, the Center for Computational Science, and several individual members of the research community recognized the need to start building services around data curation as early as 2008. This recognition started conversation which spawned collaboration, and in early 2015 a formal data curation initiative was launched with a presentation to the Academic Computing Advisory Committee (ACAC) on January 26th. The initiative has four main components:

- Socialization of the initiative within the research community
- Assessment of data curation practices and needs at the University of Miami
- Piloting data management/curation services
- Programmatic development of services based on needs assessment and pilot projects

This preliminary report covers work undertaken for the first three components: socialization, assessment, and the pilot projects. Initial results from these processes are shared and the report ends with recommendations for the fourth phase, the programmatic development of data services. For the purposes of brevity, there is no formal reflection on the socialization, and instead the report focuses on the assessment of curation needs. After a description of the methodology used, the report is broken into five parts:

- A reflection on what *data curation* means across the academy and within the University of Miami research community.
- An exploration of the tensions and synergies between sharing, backup, ownership and the value of research data.
- A description of ongoing data curation and pilot projects at the University of Miami.
- A review of data governance and research data policy both at the University of Miami and at a broader inter-institutional level.
- Six recommendations for the programmatic development of data services at the University of Miami.

The results align with existing information and data science literature on data curation within the academy writ large; data curation practices at the University of Miami show many similarities to those at other research intensive universities where assessments have also been undertaken. On a practical level, the assessment process proves invaluable in two ways. First, as a case study of data curation practices here at the University of Miami the results are critical to administrators to move the Data Curation Initiative forward. Second, the assessment process itself is perhaps more valuable than the data it produced. Through continued engagement and assessment, the initiative is further socialized within the research community and best practices in data curation will be disseminated and adopted.

II. Methodology

The research design for the data curation initiative is based on participatory research methods drawn from the disciplines of Anthropology and Geography. The principal assessment method used was the in-depth semi-structured interview with a sample drawn from the research community at the University of Miami. This principal methodology was complimented with many informal conversations with graduate students and faculty, two formal presentations at the Center for Computational Science (CCS) and the Calder Library at the Miller School of Medicine, and several graduate seminars on "Data Wrangling in the Research Environment" (ongoing). The research also involved low-level engagement with several ongoing research projects at CCS, the Rosenstiel School of Marine and Atmospheric Sciences (RSMAS), the School of Nursing and Health Studies (SOHNS), and within the library system itself (see the Pilot Projects below).

This mixed-method approach to anthropological research allows for a better triangulation on the human behavior being investigated. While no statistical inferences can be made with this research design, the desired outcome is a "thick description" of the practices observed; perhaps better said, an in depth assessment of the behavior.[1] This research design is considered participatory and emphasizes process over product. It is participatory in the sense that the researcher cannot be considered an objective observer as in laboratory experiments, but instead 'participates' in the very system that is being observed. While this approach embodies an analytical technique, it is also a community building tool and thus gives at least as much importance to process as it does to the final research product.[2]

To inform this approach a literature review of current trends in data curation within the field of Library and Information Science was undertaken. Topics reviewed included data sharing practices, data repositories, data ownership and property rights, open data, research data services, data management plans, data curation best practices, and the broader political economy of research data. The review provided insights into broad data curation trends within the academy. In addition to this engagement with background material, a scan of research data policies across 17 research intensive universities—including the ten peer institutions identified by the libraries—was undertaken (please see the Data Governance section below).

The sample for the semi-structured interviews was drawn using the snowball sampling technique in which key informants suggest appropriate and (hopefully) willing individuals for the formal interviews. The initial key informants included Dr. John Bixby at the Miller School of Medicine, Dr. Ben Kirtman at the RSMAS, Dr. Brian Blake on the Gables Campus, and several members of ACAC. Outside of this group several department heads were also consulted, such as Dr. Helena Solo-Gabriel of Engineering, Dr. Kenneth Broad of the Abess Center, and Dr. Sigman Splichal in the School of Communication. During the period of initial contact and coordination with the key informants an interview guide was developed and an exemption proposal for Internal Review Board (IRB) was prepared. For the interview guide an initial set of 78 questions were drawn from six existing institutional data curation assessments, including the 2009 survey completed at RSMAS.[3] This initial collection of questions was refined to a set of eight general questions with

[1] For an outline of this approach see Geertz (1973).

[2] This research instrument was chosen over a standardized survey for two reasons. First, several existing studies from different research institutions already show general data curation trends and it is likely that these results would simply be reproduced through yet another survey. Second, it is well known that survey participants—the research community—tire of answering questions and resent the process; this is *not* a desirable outcome of the assessment.

[3] The main resources were drawn from existing library scholarship (Carlson, 2010; Fish, 2010; Jones, Ross, & Ruusalepp, 2009; SFU, 2013; Tenopir et al., 2011; Van Tuyl, 2014). Other sources include personal communications from colleagues (Chris Erdmann (Harvard),

follow-up prompts. This refined draft was then field tested with several researchers prior to the submission of the IRB exemption proposal. The proposal with the final interview guide was submitted on February 25th 2015 and the exemption was granted by the IRB on March 8th 2015.

During the months of March, April and May of 2015 twenty-three formal interviews were conducted with researchers from 15 different departments and 10 different schools. The departments visited were: Computer Science, English, Geography and Regional Studies, History, Mathematics, Physics, Political Science, Journalism and Media Management, Civil Architectural and Environmental Engineering, Electrical and Computer Engineering, Music Education and Music Therapy, Atmospheric Sciences, Marine Biology & Ecology, Marine Ecosystems and Society, Architecture, Developmental Neuroscience, Law, and the Library. The interviewees selected from these departments constitute a broad and shallow sample of the research community at the University of Miami.

What follows is a preliminary analysis of current data curation practices and future data curation needs within the research community at the University of Miami constructed using the research methods described above.

III. Initial Observations

The term 'data curation' means many different things within the research community. While the term 'curation' is relatively straightforward to define, it is the term 'data' that poses difficulties in its definition. In the context of current research, data can be anything: algorithms, actual code, numerical data, processed data, analyzed data, and so on. Data can be constituted of model input or model output; this understanding can be extended to include data as research input and research output. Data can include video recordings, audio recordings, artifact scans, photographs, material things, and other media resources. Data can be literally be anything we can imagine, create, collect, write, and communicate.[4] Understood thus, there is very broad spectrum of what investigators talk about when they speak of data or data curation (see Box 1).

Very generally we can understand data curation as the management of data throughout its lifecycle, including the collection, management, processing, analysis, and dissemination of research data products. Each of these steps is essential to the well-established knowledge creation process, yet novel modes of dissemination and sharing of research data gave rise to the term 'data curation' within academic circles in the 1990s.[5] In the past data sharing and publication within the academic circles remained either within the published article or in private communication channels between collaborating scientists.[6] While this tradition remains firmly in place, some observers suggest that with the advent of computer based methods in the practices of science and those of scholarly communication, the practices of creating and disseminating knowledge are also shifting rapidly.[7] Much as these changes may define what data curation means broadly, they are also a foci of the assessment of data curation at the University of Miami. Indeed, drawn from the interviews and other informal conversations in the assessment and socialization processes described above, we can observe shifting practices in scholarly communication and data sharing and thus gain a better understanding of what data curation means to investigators at UM.

Sarah Pickle (USC), and Jack Reed (Stanford)) and less formalized shared experiences from other universities and research organizations (University of Minnesota, Penn State, and Bepress).

[4] One researcher at the RSMAS has tubes of sediment cores that need curation services.

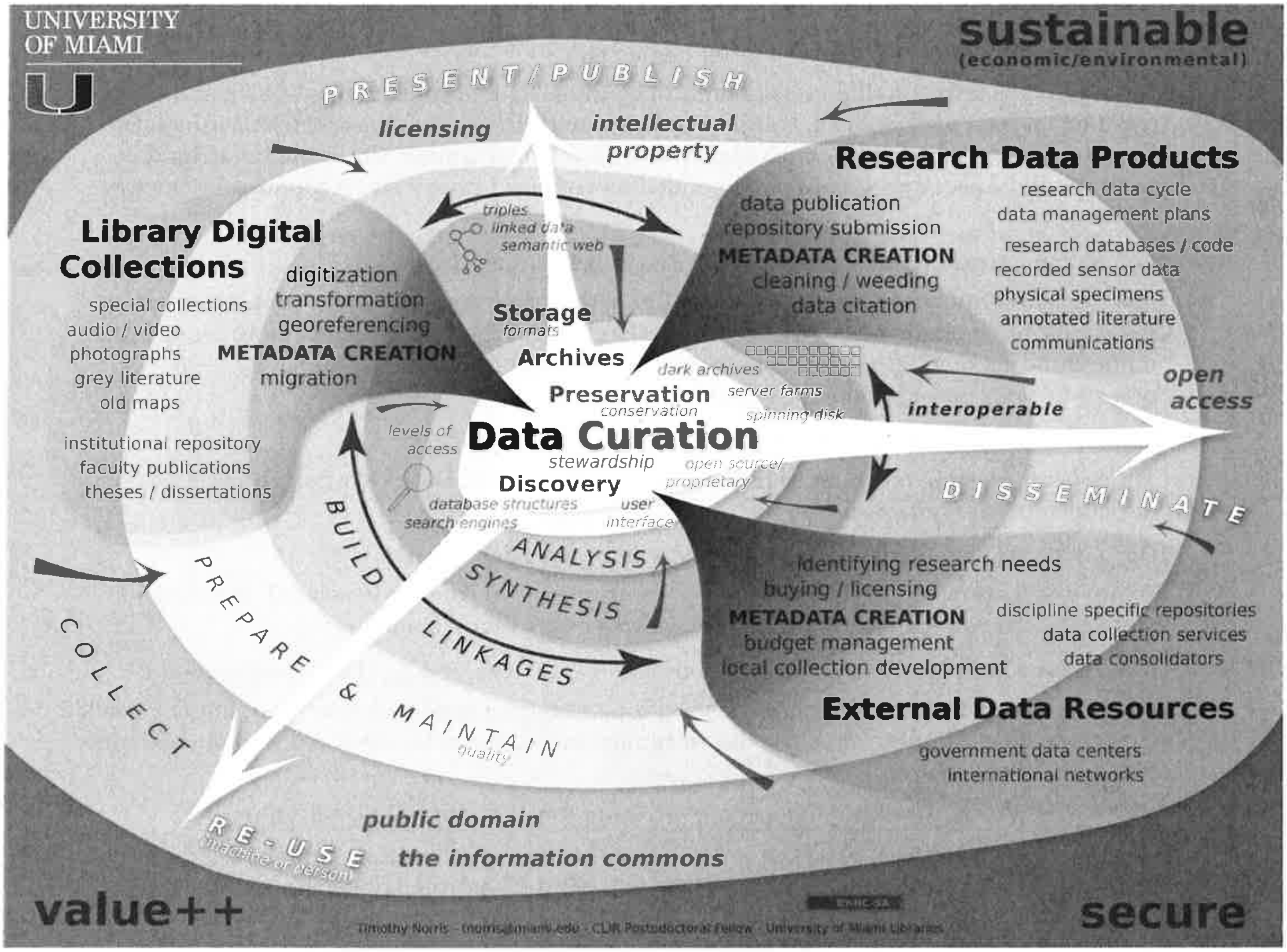
[5] See Zorich (1995) for a brief history of the term.

[6] David (2008) provides an excellent history of open science and data sharing within the Scientific Revolution.

[7] The relatively lengthy experience from the Long-Term Ecological Research network (LTER) provides interesting insights into this process (Karasti, Baker, & Halkola, 2006).

Box 1: The Data Curation Mind Map

To complement and inform the data curation assessment this data curation mind map was developed. This map was inspired by the widespread use of the term 'data curation' with often unclear conceptions of what it meant. It is the product of many conversations in the library, within the CLIR community, and with research faculty across the three campuses. What started as a doodle has become a tool for starting conversations and clarifying how to understand the practices and institutions that comprise data curation. It was featured in the California Digital Library's "Data Pub" blog (http://datapub.cdlib.org/2015/09/14/data-curation-viz/) and a more elaborate poster based on the same image won best poster at the 2016 International Data Curation Centre's annual conference in Amsterdam (see http://bit.ly/269Yr3F for the poster and http://bit.ly/2ebfAE8 for the award).



see https://youtu.be/yKvxWk5wwUU for an animated version

What follows are the most important of these observations:

- Data are both a research input and a research output.

  This may be obvious: like a good musician combines melodies from the past to create a new masterpiece, scholars use data from previous work to create new knowledge.

  Data about the relationships between research input and research output are often as important as data itself. The bibliography is the investigator's legacy. In this sense nothing has changed from historical academic practice, but the tools to construct "bibliographies" or "linked data" are changing and hence the historical practice is also changing.

- Big data are the exception.

  Research labs with big data tend to be an exception. While these groups sometimes have well developed data management practices, difficult problems do arise and solutions may be ad-hoc.

- The 'long tail' of research data is often neglected.

  Research labs with small data sets tend to be invisible to data curation efforts and are often not served well. Indeed, this was the case with the chosen assessment method: the few exceptional laboratories with big data drew much attention while the investigators with small datasets were not recommended as potential interview candidates by the key informants.

- Long-term curation and preservation for data is lacking.

  Most researchers within the community know that they don't have the time, the human resources, or the knowledge to adequately take care of their data. There is an inadequate understanding of what actions and infrastructure are needed to support long term preservation of research data.

- Human resources are the limiting factor in good data curation.

  There is no possible machine or technology to curate data. Data curation is based on difficult human-made decisions.

- Good data management crosses personal and professional boundaries.

  While it is a standard sage practice to separate personal from professional, this division is more difficult as technology further penetrates every living moment.

- Most researchers want to share data, but some are frustrated by DMP requirements.

  Sharing within research communities is a long-standing academic tradition, but the 'stick' that the federal funding agencies use to require researchers to develop data management plans is not always received well.

- Competition in the research environment can both limit *and* encourage sharing.

  Sharing can make research more competitive through collaboration, but the fear of not finishing first in the publication race limits sharing. These tensions are often discipline specific.

- There is a tension between rescuing the past and capturing the future.

  Data curation for late-career researchers who have much data which may get lost requires large investment. We must also invest in early-career researchers to ensure good data curation in the future.

IV. Tensions and Synergies

Several broad tensions and synergies were made visible through the assessment process. What follows is an exploration of two paired relationships that show these characteristics: first the synergies and tensions around the sharing and the backing up of data are explored, and second, the synergies and tensions around the value and the ownership of data are considered. This dual pairing of relationships is an abstraction of a much more complex web of relations observed. For example, practices of sharing are constantly evolving within discipline specific contexts, yet this axis of change cannot be described with the chosen abstractions. Another shortcoming of this approach is that all of the themes mentioned, sharing, backup, value, ownership, and so on, have complex relations to the political economy of academic publishing, yet these relationships are not explicitly described.[8] The paired relationships can inform our thinking as the Data Curation Initiative moves forward here at the University of Miami.

a. Tensions and Synergies I - sharing and backup

From a practical perspective sharing and backup strategies are two of the principal points of intervention for improving data curation practices. There is a large variance in these practices in the research community at the University of Miami, which range from exemplarity to problematic. What follows is a snapshot of the major trends (see Box 2).

We should also be clear that when we talk about 'back-up' here that ensuring that data is safely stored and backed-up is just the start to long term preservation of the data. *Backing up data does not mean that it is preserved.*

- Approximately one third of the researchers interviewed maintain blogs of some nature (share). Often times the content is copied from already existing files (backup).

  The blogs include teaching materials, e-journals, grey literature or post-prints of professional literature. Often these materials are hosted on WordPress or Piazza and paid for by the researcher. This is a growing trend; personal servers are going to become standard: as we move to the future we will all rent low-cost server space to create our own web presence, to share work, and to keep extra copies.

- Some research labs/groups use external drives and/or the 'personal cloud' to manage their research data (share, backup)

  In most cases hard drives are purchased by the investigator although there are a few exceptions where the backup/sharing device is purchased with grant money. While this can also be considered a form of data preservation, the longevity of these solutions is unknown (as examples: drives will break, data organization tends to be ad-hoc with no indexing).

- Many researchers are using cloud services such as google drive, box, and spider-oak (mostly share, some backup).

  Investigators are experimenting with these cloud-based technologies and solutions tend to be ad-hoc. There are initial efforts in the field of data curation to set standards for these technologies. The general feeling is that they should NOT be used for backup. There may also be privacy concerns with some of the cloud based services (e.g. Dropbox, Google Drive, iCloud, OneDrive and so on) while other cloud based services have better security policies (e.g. Box or SpiderOak).

[8] As an example of this analytical gap we can consider sharing and publication of research results as a continuum: at one end is sharing with a future self, and at the other end is publishing within peer reviewed journals run by for-profit publishing houses. This continuum and the respective relationships between sharing and publishing are poorly rendered.

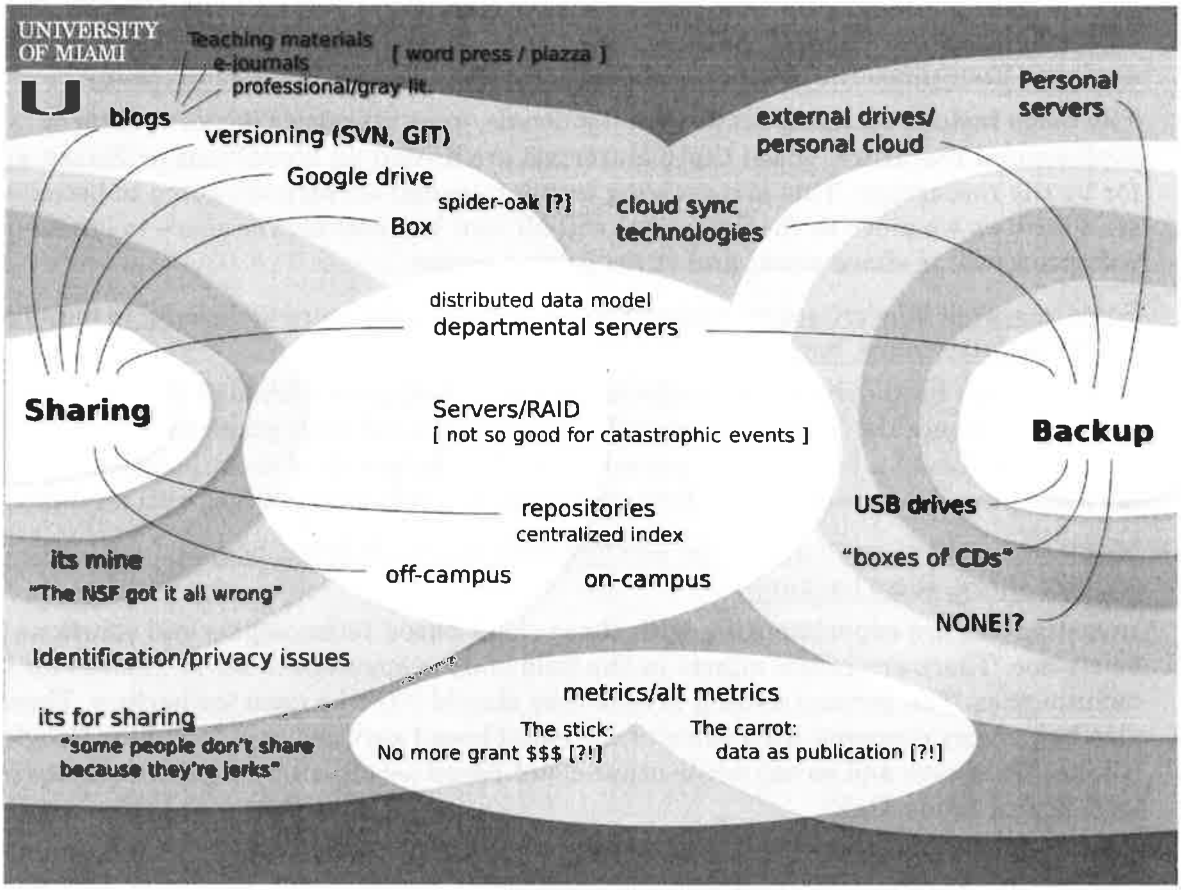
- Some researcher labs/research groups maintain departmental and/or broader disciplinary servers/repositories (share, backup and [maybe] preserve).

  This is not a widespread practice at the University of Miami. The few observed solutions do not have good information about broad preservation issues in information science such as the regular use of checksums to monitor 'bitrot', paying attention to file format viability, or the need to de-accession data as needed. Sustainability of these (grant driven) efforts is a major issue. Some good examples are:

  - At RSMAS there is an NSF funded project to set up (and maintain?) a server to share atmospheric data (research output and teaching resources).
  - At the Miller School of Medicine there are several NIH funded projects with 'repositories' for research data products both from within the Miller School and from other research institutions.
  - On the Gables campus in the engineering department servers are maintained to share data within the department.

Box 2: Sharing and Backup

This mind map was developed to better understand the tensions and synergies between data sharing and data backup. The mapping between sharing and backup is far from exact, but the development of the visualization was an excellent heuristic exercise.



- Very few researchers are using versioning software like GitHub or SVN (share, collaborate, and version).

  Some lab groups doing advanced computer work are using these versioning/ sharing services. It is likely that this versioning, sharing, and backup model will be adopted by more research groups in the future; we should note, however, that these do not provide preservation services.

- Backup strategies vary greatly in the research community and need improvement (backup).

  Less than half of the researchers interviewed have backup strategies sufficient to provide access to the data at some future date. Some researchers report using thumb-drives, emailing to self, and google drive as their backup strategy. Few researchers have standardized and methodical backup practices.

- Opinions about data sharing come from a broad spectrum of normative stances on data ownership (share).

  On the private end of the spectrum investigators claim that "its mine" [the data] or "the NSF got it all wrong." At the public end of the spectrum investigators complain that "some people don't share because they're jerks" (actual quotes from the interviews). More moderate positions acknowledge how ownership can be much more nuanced and how identification and privacy issues affect sharing.

- There is limited conversation around the 'carrots' for data sharing, data publication and data preservation (share and preserve).

  Few researchers speak about data publication as a way to increase impact metrics (the carrot). This includes a noticeable lack of knowledge about DOIs, ORCIDs and other persistent identifiers that are used in calculating impact metrics.[9] Any change to increase data sharing/publication incentives will necessarily be cultural (data publications on the CV as respected) and technical (how the carrot functions).

b. Tensions and Synergies II – value and ownership

Conceptions of value and ownership tend to shape the society we live in. Both of these words have vast literatures, ranging from "objective" academic works to the more "subjective" politically active works. No matter where we place ourselves on this spectrum of thought on value and ownership, consideration of the tensions and synergies between the two is necessary and useful. Indeed, these are the primary considerations that inform sound data governance. A snapshot of the major issues observed throughout the assessment process follows.

- Data inputs and outputs must be considered (value and ownership)
  - The monetary value of input data ranges from shared, to self-created, to purchased/licensed (low to high). This monetary value does not necessarily correlate with use or re-use value.
  - The re-use value of output data depends on several factors: number of process and analysis steps (high vs low level data), extent of documentation (metadata), standardized versus ad hoc version control, open source versus proprietary software and formats used, and the financial investment for the creation or capture of the data.
  - The monetary or exchange value of data outputs varies with market demand.

[9] For an explanation of ORCIDs and DOIs see http://orcid.org/ and https://www.doi.org/ respectively.

- Metadata creation adds value to both data inputs and data outputs, but very few researchers are aware of this relationship (value and ownership).

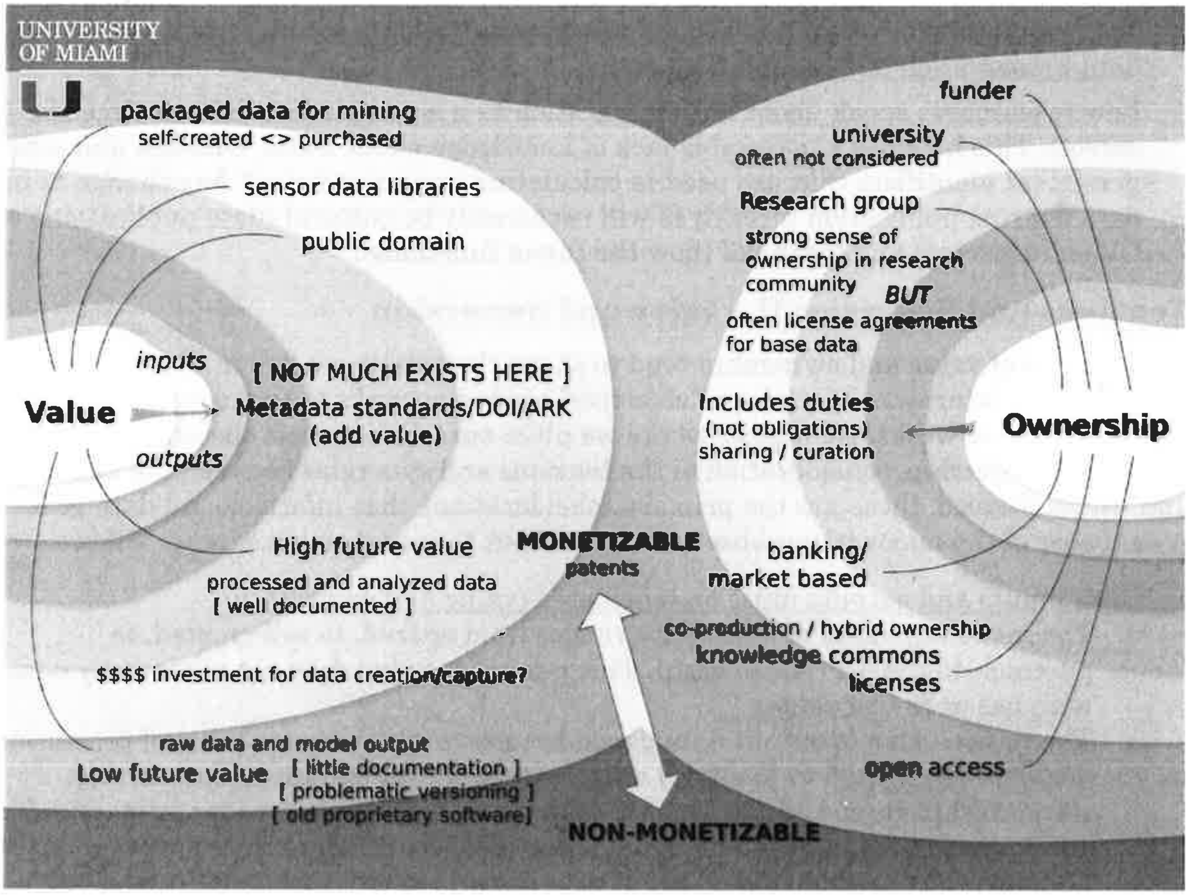
  Transcribed metadata creates little value for the individual researcher who retains the metadata in their memory, but instead creates greater value for the research community writ large (perhaps even some imagined future set of users). This tension between individual and community is hard to overcome, much like the tension between private and communal property in economic systems.

- Data produced through research on a university campus is owned by the institution, but most researchers consider data created as their property (ownership).

  At the University of Miami there is no written definition of data in the faculty manual, nor is there a clear written policy on the ownership of research data products. This may cause confusion between the research community and the university at some future date.

Box 3: Tensions between Value and Ownership

This visualization was developed to better understand the relationships between value and ownership. The mapping between value and ownership is not exact, particularly with future values of research data. Value can be assigned in ways other than monetary.



- Research groups have a strong sense of ownership of their data produced, yet often ownership is mixed (value and ownership).

  There is often a mix of ownership for input data held by a research lab. Data inputs may range from purchased to gifted and have several other ownership regimes between. In some cases it is unclear as to who really owns the data inputs and thus who really can claim ownership to the derivative products.

- Ownership of data *must* range from "private" to "open access"

  There is a banking market based model for ownership of research data in which data is bought, sold and traded. There is a knowledge commons model where licenses exist to govern access to data for defined user groups. There is an open access or open data model in which data is shared with no cost. All of these models will be useful in the future.[10]

- Data is both monetizable and non-monetizable.

  This division is largely driven by understandings of value, particularly exchange value vs other kinds of value. Data about the earth and the natural environment are not necessarily monetizable in our current economic system, but there is great use value in this data. Other data such as genome sequences are highly monetizable through the pharmaceutical industry.

V. Pilot Projects

Throughout the socialization and assessment processes several pilot projects were identified as good data curation experiments to build upon. Unlike the assessment process, there is no research design for the selection of the pilots or the method used in the pilots, but instead the approach is one of collaborative learning through the evaluation of opportunities and returns. The pilots are a learning process for the investigators involved as there is a sharing of information and practices from the library to the research group. The experience also serves as a learning process within the library on how to engage with researchers and develop and/or provide curation services to these research endeavors. These collaborations continue to grow and inform the creation of infrastructure to provide data services at the University of Miami.

The level of library commitment to the projects was initially and intentionally very low. In development work it is dangerous to create expectations that cannot be met due to lack of resources. The development of data services at an institution of higher learning is no exception to this rule. The collaborations often started as a consulting process where the Library learned about the research and then provided specific data management feedback to the investigator depending on the identified needs. In some exceptional cases the relationship have grown and are ongoing. In some cases, perhaps the ideal, the needed services were provided and the relationship has waned. In some less than ideal situations the collaborations never formed; often this shortcoming was due to a lack of time. Indeed, as these collaborations formed difficult decisions about where to dedicate resources were made: it is not possible nor desirable to curate all the data that exists at the University of Miami.

A brief summary of some of the pilot projects follows. The projects are categorized into general types (with the understanding that the boundaries are sometimes fuzzy):

[10] The work of Elinor Ostrom and Charlotte Hess is groundbreaking in thinking about ownership regimes for knowledge resources (Hess & Ostrom, 2006). Also see Stiglitz (1999).

- Data purchase or license negotiation
  - The Library has long provided access to large sets of research data through the Inter-University Consortium for Political and Social Research (ICPSR), but has begun to identify data sets for specific needs on campus.
  - In the fall of 2014 the Library failed to negotiate the purchase of high resolution digital elevation model (DEM) for Havana, Cuba. The cost was prohibitive. Both the investigators and the Library learned much about this purchasing process, nevertheless.
  - In the spring and summer of 2015 the Library negotiated a subscription with CINEP for the *Archivo Digital De Prensa* news database from Columbia (http://onbase.cinep.org.co/AppNet/). The database is open for use by any investigator at the University of Miami for one year.
  - Since the arrival of the Libraries' GIS Services Librarian, the Library has begun to negotiate more access to datasets. In addition, with the addition of a Data Services Librarian by 2017, more attention will be focused on this need.

- Organization of research data products for deposit into data repositories
  - The first data object deposited in the University of Miami institutional repository is a collection of tweets that followed the announcement in December of 2014 that the United States would normalize diplomatic relations with Cuba. The dataset was curated by Natalie Baur of the Cuban Heritage Collection and deposited in early 2015. http://dx.doi.org/10.17604/M6RP4D
  - The second data object deposited into the repository at the University of Miami is a dataset from the Physics department authored by Stanislav Lazopulo (a graduate student) and co-authored by Dr. Sheyum Syed. The data was curated by the research group and followed recommended best practices from the Library. http://dx.doi.org/10.17604/M6WC7P
  - The Library has have identified more datasets for curation and deposit, but lack the human resources to curate them properly. For example the LIBQUAL dataset under Sharyn Ladner of the Libraries and the NetZero Water Project under Dr. Jim Englehardt of the Department of Engineering. Many more datasets have been identified.

- Use of the Scholarly Repository for deposit of research data
  - The Library is depositing data objects into the University of Miami institutional repository based on the proprietary BePress repository service to which the university subscribes (http://scholarlyrepository.miami.edu/).

- Provision of DOIs (digital object identifiers) to investigators for curated datasets
  - In 2015 the University of Miami Libraries negotiated a contract with the University of California Digital Library's EZID service (http://ezid.cdlib.org/). The Libraries can now mint DOIs for digital objects hosted on the University of Miami campuses.
  - Dr. John Bixby and Dr. Vance Lemon are investigating how they can use the service for their nerve cell regeneration repositories.
  - The first DOI minted at the University of Miami was created in August of 2015 for the Miami Affordability Project (MAP), a collaboration between the Office of Civic and

  - Community Engagement, The Libraries and the Center for Computational Science. http://dx.doi.org/10.17604/M6159M
  - See also the curated datasets above.

- Collaboration with investigators who are building data repositories and/or curating valuable data sets
  - The library has an ongoing and successful collaboration with the SOHNS measures library under Dr. Victoria Mitrani. The project seeks to make the collection of survey instruments collected by Dr. Mitrani a publically accessible resource (see the product here: http://www.miami.edu/sonhs/index.php/elcentro/research/measures_library/). Information Technology and Blackboard Services are additional collaborators. The library's engagement with this project is advisory.
  - The library is engaged with an ongoing and growing collaboration with Chris Mader and the Software Engineering Group at the Center for Computational Science. This is a powerful collaboration that spans several projects including the Pulley Ridge project under Dr. Cowen at RSMAS, the Drone Survey project under Dr. Adib Cure in the School of Architecture, amongst other projects.
  - We have identified more potential collaborations, but currently lack the human resources to pursue these relationships.

- Data management curriculum development and delivery
  - Six graduate seminars for "Data Wrangling in the Research Environment" have been organized and delivered at RSMAS, Richter Library, and in the Engineering Department (Fall 2014), and in the departments of History, International Studies, and the Abess Center for Ecosystem Science and Policy (spring of 2015). These have been very well received and, besides being a great teaching opportunity, they are a powerful community building device. More seminars are planned.
  - As a collaboration between the RSMAS library and the Richter Library a new two-credit class for data management is under development. The 600 level class was taught at RSMAS in the Spring of 2016 and will be taught again in the spring of 2017 (see Appendix D for the syllabus an online materials).
  - The Library is running "train the trainer" sessions for the fall of 2016. These will be collaborative learning opportunities with the subject librarian staff.
  - Tim Norris is now a trained instructor for software carpentry (http://software-carpentry.org/) and taught one workshop at UM in the spring of 2016. This effort is led by the Center for Computational Science. Norris taught another workshop at FIU in the fall of 2016. More workshops for UM are currently being programmed by CCS.

VI. Data Governance

The governance structure—the policies, guidelines, and procedures related to the stewardship of research data—is underdeveloped and essentially made up of a patchwork of policies. This is somewhat typical of research-intensive universities, though a growing number have established policies specific to research data in recognition of the unique forces (faculty and institutional interests, funder and publisher requirements for management and sharing, commercialization issues) at play. Currently, policies/guidelines that relate to research data are covered within the Faculty Manual, by the Office of Research, and by UM Information Technology.

University of Miami Faculty Manual:

Research data is discussed in two places in the 2016-17 University of Miami Faculty Manual: in the Policy on Inventions, Intellectual Property, and Technology Transfer section and in the Policies and Procedures of the University of Miami Relating to Allegations of Misconduct (see https://umshare.miami.edu/web/wda/facultysenate/FacultyManual.pdf).

- Data is included as an "Innovation" in the Policy on Inventions, Intellectual Property, and Technology Transfer:

  > Innovations: patentable or un-patentable inventions, discoveries, processes, compositions, research tools, data, ideas, databases, know-how, copyrightable works that are not scholarly or artistic Creations and tangible property, including biological organisms, engineering prototypes, drawings, and software created, conceived or made by Applicable Personnel within their normal duties (including clinical duties), course of studies, field of research or scholarly expertise or making more than Incidental Use of University's resources. (2016-17 Faculty Manual, p. 136)

  Innovations are specifically called out in the policy as owned by the University (p. 137). In addition, the policy states:

  > Applicable Personnel are required to record all research data and information accurately and clearly and to keep all such data in a permanent and retrievable form. In addition, with regard to a patentable Innovation, original laboratory data must be kept for the life of the patent. Tangible property, including biological materials, chemical compounds, etc., must be securely stored. All of the foregoing are the University's property. Exceptions to these requirements may be adopted in writing by the TTPC. (p. 141)

- Data is also discussed within the Policies and Procedures of the University of Miami Relating to Allegations of Misconduct:

  > Research record means the record of data or results that embody the facts resulting from scientific inquiry, including but not limited to, research proposals, laboratory records, both physical and electronic, progress reports, abstracts, theses, oral presentations, chapters, books, audio or video tapes, CDs, internal reports, journal articles, and any documents and materials provided to the University or to a University official by a respondent in the course of the research misconduct proceeding. (p. 125)

  In this section there is also a requirement that data be retained that is somewhat inconsistent with the statement directly above from page 141.

  > In order to respond to allegations regarding the integrity of any published report, adequate records of the original protocols and research records, including all raw data, must be preserved for at least seven years (or longer if required by the funding agency), so they can be made available for inspection." (p 124)

University of Miami Office of Research:

The UM Office of Research also maintains a number of policies and guidelines. Three policy documents directly address research data: the Research Policies / Guidelines Handbook, the Data Sharing Guideline, and the Data Management Template.

- The Research Policies / Guidelines Handbook mentions research data multiple times. http://uresearch.miami.edu/documents/UM_Research_Policy_Guidelines_Handbook_(07_25_16).pdf
    - Policies and Procedures of the University of Miami Relating to Allegations of Misconduct in Research (p.14-23).

      This is generally the same as the Policies described in the Faculty Manual except that it does ***not*** state the requirement to maintain data and protocols for seven years.
    - Export Control and Technology Management (p.67-73)

      This policy includes data within its purview and references several other policies related to data that are administered – or jointly administered – by UM Information Technology.
- The Data Sharing Guideline focuses on a Data Use Agreement for the sharing of data between institutions yet the guideline is limited in scope. It does not speak directly to sharing data within research data repositories, for example. http://uresearch.miami.edu/research-resources/research-guidelines/data-sharing
- The Data Management Template is meant as a guide for researchers to write a data management plan, but is generic and not tailored for specific funding agencies. http://uresearch.miami.edu/documents/Data_Management_Plan_Template.pdf

UM Information Technology:

UMIT maintains or jointly maintains the following policies that address data: the Data Classification Policy, the Electronic Data Protection and Encryption Policy, and the Electronic Data Quality Policy for Clinical Research. These are all broadly applicable and address more than just research data; they also apply to educational data, institutional data, and so on. They are important pieces of the patchwork of policies for research data nevertheless.

- The Data Classification Policy Covers how data is classified in terms of sensitivity and risk. http://www.miami.edu/index.php/a110_data_classification_policy/
- The Electronic Data Protection and Encryption Policy covers when data should be encrypted. https://umshare.miami.edu/web/wda/itciso/POL-UMIT-A175-014-01-Electronic_Data_Protection_and_Encryption_Policy.pdf
- The Electronic Data Quality Policy for Clinical Research outlines the principles for ensuring data quality for clinical research. It is maintained by UMIT, the Office of Research, and UHealth Information Technology. http://uresearch.miami.edu/documents/Electronic_Data_Quality_Policy_for_Clinical_Research.pdf

Research Data Policy at the University of Miami

The challenge with this patchwork of policies (beyond the small inconsistencies) is that they do little to recognize the current governance of research data in a broader context. With funders and a growing number of publishers requiring that research data be made openly available, researchers may find the above policies confusing in terms of what rights they might have to comply and whether they can place an open license on the data as often required. There may also be disciplinary expectations, and copyright limitations particularly where data is derived from other sources (for example, a corpus of digitized novels as an input for a text mining project in the digital humanities).

Research Data Policy at Peer Institutions

To better understand the research data policy landscape, a small environmental scan was undertaken across seventeen research institutions. The scan shows that the policy gaps identified at the University of Miami are not unique. Indeed, amongst our peer institutions, the University of Miami is neither ahead of the game nor amongst the laggards (see Table 1). The scan identified policy language and policy components including, but not limited to: definitions of research data, definitions of responsible parties, ownership, custodial responsibilities, retention policy, access policy, governance of cases when the principal investigator leaves the university, governance of disputes, and take-down policy.[11]

Table 1: Institutional data policy scan across peer and non-peer institutions

Peer Institutions	last revised	responsible party	ownership stated	"research data" defined	data retention (min.)	data access	PI moves	disputes	take-down	**notes**
Case Western Reserve University *	2000	X		X	3 years	X	X	X	X	
New York University *	2010	X	X	X	3 years	X	X		X	
University of Rochester *	2014	X	X	X	3 years	X	X			
University of Miami	*2014*				*7 years*					*in faculty manual*
Emory University *	2007	X			7 years		X			in faculty manual
Brandeis University *	2003		X							in IP policy
University of Southern California *	2001		X							in IP policy
Carnegie Mellon University *										
Syracuse University *										
Tulane University *										
Vanderbilt University *										
Non-peer Institutions										
University of Kentucky	2011	X	X	X	5 years	X	X	X	X	
Johns Hopkins University	2008	X	X	X	5 years	X			X	
Duke University	2007	X	X	X	5 Years	X				
The University of Edinburgh	2011	X				X			X	not really a policy
University of Pittsburg	2009	X	X	X	7 years	X	X			
Pennsylvania State	2003	X	X	X	5 years	X	X			set of guidelines

We recommend that the Office of Research, UM Innovation, Libraries, and UMIT should develop a policy document specific to the governance of research data. The policy should reference the other policies as outlined above. A draft Research Data Policy can be found in Appendix B: University of Miami Research Data Policy [DRAFT].

Recommendations

Six concrete steps are identified as ways to move forward with the data curation initiative:

[11] These are drawn from Briney, Goben, & Zilinski, 2015, several less formal working group documents within the library community, and careful study of existing policies.

- **Establish a cross-unit collaboration and supporting budget for research data curation infrastructure and services.**

  Because research data curation has so many facets, support needs to be a collaborative effort across the Office of Research, the UM Libraries, and UMIT, with further support from groups like the Center for Computational Science. We recommend a working group be formed to develop a single researcher-facing Research Data Service that provides researchers with a consistent set of services, policies, and practices.

  The Research Data Service would be a phased implementation that would encompass the development of the recommendations made below.

  In Appendix A, we have outlined the initial areas that would need budgetary support; the working group would need to provide more specifics. The numbers given here are only estimates.

- **Establish clear policies and guidelines for research data.**

  Develop a policy document specific to the governance of Research Data at the University of Miami as a collaborative effort between the Office of Research, the Office of Research Administration, UM Innovation, the Libraries, and UM Information Technology. A draft Research Data Policy can be found in Appendix B: University of Miami Research Data Policy [DRAFT]. The language in this draft was modeled after existing policies at several research institutions; see Appendix C.

- **Develop data repository infrastructure as a collaboration**

  Develop or license infrastructure for an institutional data repository (including discovery layer) for research data products that have no disciplinary home. Provide support for research groups who create and maintain discipline specific repositories across campus, when possible. Both efforts should include the provision of DOIs (digital object identifiers).

  Three exemplary efforts to use as models can be found at Johns Hopkins University, the University of Illinois Urbana-Champaign, and Purdue University:
  - Johns Hopkins Data Archive Dataverse Network: https://archive.data.jhu.edu/dvn/
  - University of Illinois Data Bank: https://databank.illinois.edu/
  - Purdue University Research Repository: https://purr.purdue.edu/

- **Develop data management services**

  Build research data services at the University of Miami as an cross-unit institution-wide effort led by the library and centered on the following programmatic components:
  - Use and support of the data repository infrastructure described above. Include also assessment and consultation on the use of external data repositories which may be better suited for some research data.
  - A set of data management planning services that includes resources for DMP writing, review and submission. Note that the library is currently providing this service on an ad-hoc basis with plans to expand and formalize the service.
  - Assessment and improved implementation of access to data management, analysis and visualization tools available to the research community on all three campuses.
  - Consultation on policy, rights, and issues related to sensitive data.

- A web presence for data services at the University of Miami. This will be necessarily simple at first due to limited resources, but will be designed for scalability as the data curation initiative grows. See http://library.miami.edu/datacuration/.
- Several exemplary models can be found in the institutional landscape of higher education:
  - Research Data Management at Johns Hopkins University http://dmp.data.jhu.edu/
  - Illinois Research Data Services: http://researchdataservice.illinois.edu/
  - Data Management Services Stanford University Libraries https://library.stanford.edu/research/data-management-services
  - Data Management MIT Libraries: http://libraries.mit.edu/data-management/

- **Develop curricula for data management**

  Develop data management curricula for a diverse set of audiences as follows:
  - A pilot 600 level two-credit course to be taught at RSMAS in the 2016 and 2017 spring semesters (see appendix D for a brief description).
  - Continue to deliver and further develop department level data management seminars designed primarily for grad students.
  - Continued use of the Software Carpentry / Data Carpentry workshops – or augmentation of these – in order to grow comfort level with critical tools for data management such as R.
  - Develop library training modules for the subject liaison librarians for a series of workshops/seminars in the fall and spring 2016-2017.
  - Several exemplary efforts to create similar curricula will guide this process:
    - Research Data MANTRA – University of Edinburgh http://datalib.edina.ac.uk/mantra/.
    - DataONE Education Modules: https://www.dataone.org/education-modules.
    - Data Management Training Clearinghouse: http://dmtclearinghouse.esipfed.org/

- **Continued assessment and socialization**

  Continue the interview process into the indefinite future with two principal purposes. First to continue the socialization of the initiative and second to continue the identification of ongoing data curation practices at the University of Miami. Liaison librarians will be integrated into this process as much as possible.

Appendix A: Budget Considerations – Research Data Services

Establishing a Research Data Service is both a technology and staff intensive effort. The following outline describes the areas which would need attention in terms of budget.

1. Staffing: We recommend that the Research Data Service be managed by a director with data management experience and one data curation/technical specialist support staff. These positions would be in the Libraries and would need $125-$175,000/annum. Staffing for Research Data Services vary by institution; a large research university like the University of Illinois at Urbana-Champaign has five full time staff (a director, two data curators, one developer, and one post-doctoral fellow). Emory University has the equivalent of two FTE.

2. Establish a Research Data Repository for the sharing and publication of data sets as increasingly required by funders and publishers. Options for this are:
   - License software such as Figshare (https://figshare.com/). Based on initial conversations with Figshare in early 2016, costs range between $72-90,000 annually depending on storage needs.
   - Install open-source software such as Dataverse (http://dataverse.org/institutions). This would require server support from UMIT as well as developer support to implement and manage the software. This option would need one half-time developer with support between $25-30,000/annum.
   - Develop repository in-house. This would require server support from UMIT and at least two developers to develop, implement, and manage the software. The developers would need support from $100-120,000/annum.
   - Some combination of these.

3. Storage and technical infrastructure costs would depend on the options above.

4. Membership in supporting organizations: EZID for DOI minting services (digital object identifier – permanent identifiers for digital objects located on the Internet), ORCID (open researcher and contributor ID), and APTrust (long term preservation). The Library already pays membership with EZID ($2,500/year), uses discretionary funds to finance APTrust membership and storage costs (estimated at $22,000 for 2018), and is negotiating with ORCID for an affiliated membership ($5,000/year)

Table 2: Estimated Initial and Annual Expenditures: Research Data Services

Item Description	One time Costs	Annual Costs
1. Staffing the Research Data Service		$125,000-$175,000
2. Repository Development		$25,000-$120,000
3. Storage and Technical Infrastructure	unknown	unknown
4. Membership in Supporting Organizations		$29,500
ESTIMATED TOTALS	unknown	$200,000-400,000

Appendix B: University of Miami Research Data Policy [DRAFT]

Introduction: Creating, sharing, and retaining data and records of research (hereafter referred to as Research Data) is a central aspect of the scholarly research process. In most cases research at the University of Miami (hereafter referred to as the University) is funded through agreements between the University, the principal investigator and external sponsors. Thus both the University and the Principal Investigator have rights and responsibilities with respect to the Research Data. These rights and responsibilities govern the ownership of, the access to, the use of, and the retention of the Research Data. The purpose of this document is to set forth these rights and responsibilities so that the University can substantiate research findings when necessary, protect intellectual property rights, and ensure compliance with federal and sponsor regulations and requirements.

Definitions: The term "Research Data" in this document refers to information recorded and/or collected for research performed at or under the auspices of the University regardless of the form or the media upon which it is recorded. This term includes, but is not limited to, computer programs (code and documentation), computer databases, instrumental outputs, raw numerical results, original biological or environmental samples, photographs, digital images, films, protocols, graphs, and other deliverables produced under sponsored agreements. Research Data also includes any records related to the design, conduct or reporting of the research that would be necessary to reconstruct the reported research results. Research data can be intangible (statistics, findings, conclusions, etc.) and tangible (notebooks, printouts, etc.).

The term "Principal Investigator" in this document refers to the individual who bears primary responsibility for technical, programmatic, fiscal, and administrative requirements of the research project. The Principal Investigator can be faculty, staff, a student, a post-doctoral fellow, or a visiting scientist. The Principal Investigator is responsible for the collection, management, and retention of the Research Data and should adopt an orderly system of data organization for the duration of individual research projects. It is also the Principal Investigators responsibility to communicate clearly how the data management system works to all members of the research team including administrative personnel.

The term "University" in this document refers to the University of Miami. The University has responsibilities and obligations related to the custody of Research Data that include, but are not limited to, the following: to ensure compliance with federal grant requirements with respect to the retention of Research Data, to ensure compliance with the terms of sponsored project agreements including clinical trial agreements, to protect the rights of access for students, postdoctoral fellows, and other research collaborators, and to facilitate investigations such as those for misconduct or conflict of interest.

Ownership of Research Data: The University is the principal owner of the Research Data and retains rights of access with the understanding that information that would violate confidentiality of sources or subjects should not be disclosed. The Principal Investigator is the principal custodian of the Research Data for the University and holds the Research Data in trust for the University. The Principal Investigator will retain all rights of access to, control over, and use of the Research Data.

Research Data Retention: The Principal Investigator is the responsible person who must preserve all of the *relevant* Research Data for at least three years after the final closeout of the research project or publication of the research results, whichever occurs last (or longer if required by the funding agency). The *relevant* Research Data is deemed that which is necessary for the duplication of the research process to validate the final results. A longer retention period may be deemed necessary if 1) the retention is necessary for the protection of intellectual property resulting from the work, 2) if there are any charges or allegations of misconduct the data must be retained until all charges are fully resolved, and 3) if the data constitute part of a students work towards a degree, the data must be retained until the degree is awarded or it is clear that the student has abandoned the work. After the period of Research Data retention is over, the destruction or preservation of the research data remains at the discretion of the principal investigator.

Access to Research Data: To enable for the university to meet its responsibilities with respect to Research Data, the Principal Investigator must provide access to the Research Data upon reasonable request, to the University, its official bodies, or the external funding agencies or journals, or other external regulatory agencies. This obligation continues even after the Principal Investigator leaves the University. The Principal Investigator is also obligated to provide access to the Research Data to co-investigators and collaborators including, but not limited to, students, postdoctoral fellows and investigators from other institutions. Terms of data sharing and/or custody arrangements should be determined by the investigators at the time of joining the research project preferably as a data use agreement.

Transfer When Investigators Leave the University: In the event that investigators other that the Principal Investigator leave the University they may take copies of the Research Data unless restricted by terms of any applicable agreement with either the sponsor of the research or the Principal Investigator. In the event that the Principal Investigator leaves the University the original Research Data may be transferred with the approval of the Vice Provost for Research, and with written agreement form the Principal Investigator's new institutions that guarantees: 1) its acceptance of custodial responsibilities for the data outlined above, and 2) the University's access to the data, should that become necessary. The University may refuse to permit the transfer of original Research Data for any reason, may impose conditions beyond those stipulated in this document, and may request that copies of the Research Data remain with the University. When the University permits the Principal Investigator to leave the University with the original Research Data, it is with the understanding that the Principal Investigator must retain the Research Data for the period as stipulated above and must provide access to the Research Data by the University as stipulated above.

All questions relating to this policy should be directed to the Vice Provost for Research.

Appendix C: Sources for University of Miami Research Data Policy (Appendix B)

This DRAFT data policy document found in Appendix B draws *directly* from the following institutional data policies (i.e. some phrases are used word for word):

University of Rochester (2014). "Access to and Retention of Research Data." Last accessed on 11/15/15 at https://www.rochester.edu/orpa/_assets/pdf/policy_retent.pdf.

New York University (2010). "Policy on Retention of and Access to Research Data." Last accessed on 11/15/15 at http://www.nyu.edu/content/dam/nyu/research/documents/OSP/PolicyonResearchData030110.pdf.

Case Western University (2000). "University Policy on Custody of Research Data." Last accessed on 11/15/15 at https://research.case.edu/files/University_Policy_On_Custody_Of_Research_Data.pdf.

Emory University (2007). "Policy 7.9.02.C Access to and Retention of Scientific Research Protocols and Data." Last accessed on 11/15/15 at http://policies.emory.edu/policy/index_pdf.php?policy_number=7.9.

This DRAFT data policy document found in Appendix B draws *indirectly* from the following institutional data policies (i.e. scope, format, and intent informs the content):

Johns Hopkins University (2008). "Johns Hopkins University Policy on Access and Retention of Research Data and Materials." Last Accessed on 11/15/15 at http://provost.duke.edu/wp-content/uploads/FHB_App_P.pdf.

Stanford University (2015). "Research Policy Handbook, section 1.9 Retention of and Access to Research Data." Last accessed on 11/15/15 at https://doresearch.stanford.edu/policies/30294/print.

Appendix D: Course Description: Data Management in the Research Environment

Instructors: Dr. Timothy Norris – Postdoctoral Fellow Miami University Libraries
Angela Clark – Librarian Associate Professor RSMAS Library

2 Credits: 600 level course (30 student maximum)

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; Windows: Notepad++; Mac: 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. Professional MS students can use instructor prepared datasets if none are identified at the time of registration.

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.

NOTE: the complete course outline and reading list is located at:
http://tibbben.github.io/teaching.data.literacy/UM_DataManagementClass/DMClassSyllabus.html

References:

Berners-Lee, T., Shadbolt, N., & Hall, W. (2006). The Semantic Web Revisited. *IEEE Intelligent Systems, 21*(3). doi:http://dx.doli.org/10.1109/MIS.2006.62

Briney, K., Goben, A., & Zilinski, L. (2015). Do You Have an Institutional Data Policy? A Review of the Current Landscape of Library Data Services and Institutional Data Policies. *Journal of Librarianship and Scholarly Communication, 3*(2). doi:http://dx.doli.org/10.7710/2162-3309.1232

Carlson, J. (2010). The Data Curation Profiles Toolkit: Interviewers Manual. *Data Curation Profiles Toolkit,* Paper 2. doi:http://dx.doli.org/10.5703/1288284315651

David, P. A. (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. doi:http://dx.doli.org/10.2202/1932-0213.1040

Fish, E. (2010). *RSMAS Facutly Pilot Data Survey.* University of Miami E-Science Working Group. Miami.

Geertz, C. (1973). Thick description: towards an interpretive theory of culture. In C. Geertz (Ed.), *The interpretation of cultures: selected essays* (pp. 3-30). New York: Basic Books.

Hess, C., & Ostrom, E. (Eds.). (2006). *Understanding Knowledge as a Commons: from theory to practice.* Cambridge, Massachusetts: The MIT Press.

Jones, S., Ross, S., & Ruusalepp, R. (2009). Data Audit Framework Methodology. *Data Asset Framework,* draft for discussion, version 1.8. Retrieved from http://www.data-audit.eu/DAF_Methodology.pdf

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. doi:http://dx.doli.org/10.1007/s10606-006-9023-2

SFU. (2013). *Research Data Web Survey.* Simon Fraser University Libraries. Vancouver.

Stiglitz, J. (1999). Knowledge as a Global Public Good. In I. Kaul, I. Grunberg, & M. A. Stern (Eds.), *Global Public Goods* (pp. 308-325). New York: UNDP.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., . . . Frame, M. (2011). Data Sharing by Scientists: Practices and Perceptions. *PLoS ONE, 6*(6). doi:http://dx.doli.org/10.1371/journal.pone.0021101

Van Tuyl, S. (2014). CMU Faculty Research Data Management Survey (Carnegie Mellon University). (personal communication).

Zorich, D. M. (1995). Data management: Managing electronic information: Data curation in museums. *Museum Management and Curatorship, 14*(4), 430-432. doi:http://dx.doli.org/10.1016/0260-4779(96)84690-5