RUTGERS
University Libraries

The place to go, when you need to know!

Data Management and RUresearch

a discussion of data management principles in connection with NSF Data Management Plans and the data services provided by the RUresearch Data Services Team

Ryan Womack and Aletia Morgan

February 29, 2012

rwomack@rutgers.edu aletia.morgan@rutgers.edu

Why Worry about Data?

- You want to comply with a funding mandate like the NSF Data Management Plan.
- You want to write a better Data Management Plan than this.
- You don't want to end up like this.

Data Management Benefits the Research Community

The NSF is concerned with preserving data for two reasons:

- Data sharing - In NSF terms, this is the peer-to-peer sharing of research data for verification of results and extension of research
- Data dissemination - The broader distribution of data to a general audience

Accountability for the use of public funds is another driver for open access to data from funded research. Other funders (e.g., NIH) have similar policies, with the goal of providing access to research data so that future research can build on earlier work.

This challenge is becoming a hot topic for research. See the recent issue of *Science Magazine* on Dealing with Data, and the special issue of *D-Lib Magazine* on Research Data for further discussions on the challenge of managing research data in various disciplines.

The NSF Data Management Plan

The NSF Data Management Plan (DMP), now required of all grant applicants, may include

- the types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project; *[data model]*
- the standards to be used for data and metadata format and content (where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies); *[metadata]*
- policies for access and sharing including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements; *[policies]*
- policies and provisions for re-use, re-distribution, and the production of derivatives; *[policies]* and
- plans for archiving data, samples, and other research products, and for preservation of access to them. *[preservation]*

Each directorate has developed its own guidelines that complement and may take precedence over any general statements from the NSF.

Data Management Benefits YOU

Close attention to data management of your project will improve the effectiveness and efficiency of your research in many ways:

- Developing a *data model* will help you to organize your data and relate it to other data stores and research questions.
- Complete *metadata* using the vocabulary of your discipline will allow your data to be discovered and reused.
- Creation of *policies* will protect your rights and those of research subjects, and ensure proper use of your data.
- Proper *preservation* of all data and associated documents will ensure that your research is fully available in the future.

We will return to these topics in the context of NSF and RUresearch in just a moment.

Further Benefits

Dimension 1	Direct Benefits	Indirect Benefits (Costs Avoided)
	-New research opportunities -Scholarly communication/access to data -Re-purposing and re-use of data -Increasing research productivity -Stimulating new networks/collaborations -Knowledge transfer to industry - Increasing skills base of researchers/students/staff -Increasing productivity/economic growth -Verification of research/research integrity -Fulfilling mandate(s)	-No re-creation of data -No loss of future research opportunities -Lower future preservation costs -Re-purposing data for new audiences -Re-purposing methodologies -Use by new audiences -Protecting returns on earlier investments
Dimension 2	**Near-Term Benefits**	**Long-Term Benefits**
	-Value to current researcher & students -No data lost from Post Doc turnover -Short-term re-use of well curated data -Secure storage for data intensive research -Availability of data underpinning journal articles	-Secures value to future researchers & students -Adds value over time as collection grows and develops critical mass -Planned management from an early stage in the research lifecycle is ultimately more cost-effective than late intervention (providing proper selection of what to keep is done)
Dimension 3	**Private Benefits**	**Public Benefits**
	-Benefits to sponsor/funder of research/archive -Benefits to researcher -Fulfil grant obligations -Increased visibility/citation -Commercialising research	-Input for future research -Motivating new research -Catalysing new companies and high skills employment

Table from the Keeping Research Data Safe Factsheet, produced by JISC (UK).

Resources to Support Data Management Plan Development

Checklists that can help you identify key elements of your data management plan, but do not let them be a substitute for your own language that emphasizes the unique characteristics of your research.

- Resources at Rutgers:
  - Data Management Guide prepared by Ryan Womack, RUcore Research Data Manager;
  - Research Data Guide site by the Office of the Vice President for Research, including a checklist prepared by Aletia Morgan, at data.rutgers.edu (OVPR).
- Selected resources from peer institutions:
  - Checklist from MIT Libraries;
  - Elements and Framework of DMP from ICPSR, site also includes a webinar on the topic;
  - DMPTool online template developed by the University of California Curation Center of the California Digital Library, based on work by the UK Digital Curation Center (DCC).

It is critical to review and conform to the specific guidance and requirements described by the program requirements and relevant directorate guidelines.

The RUresearch Data Team

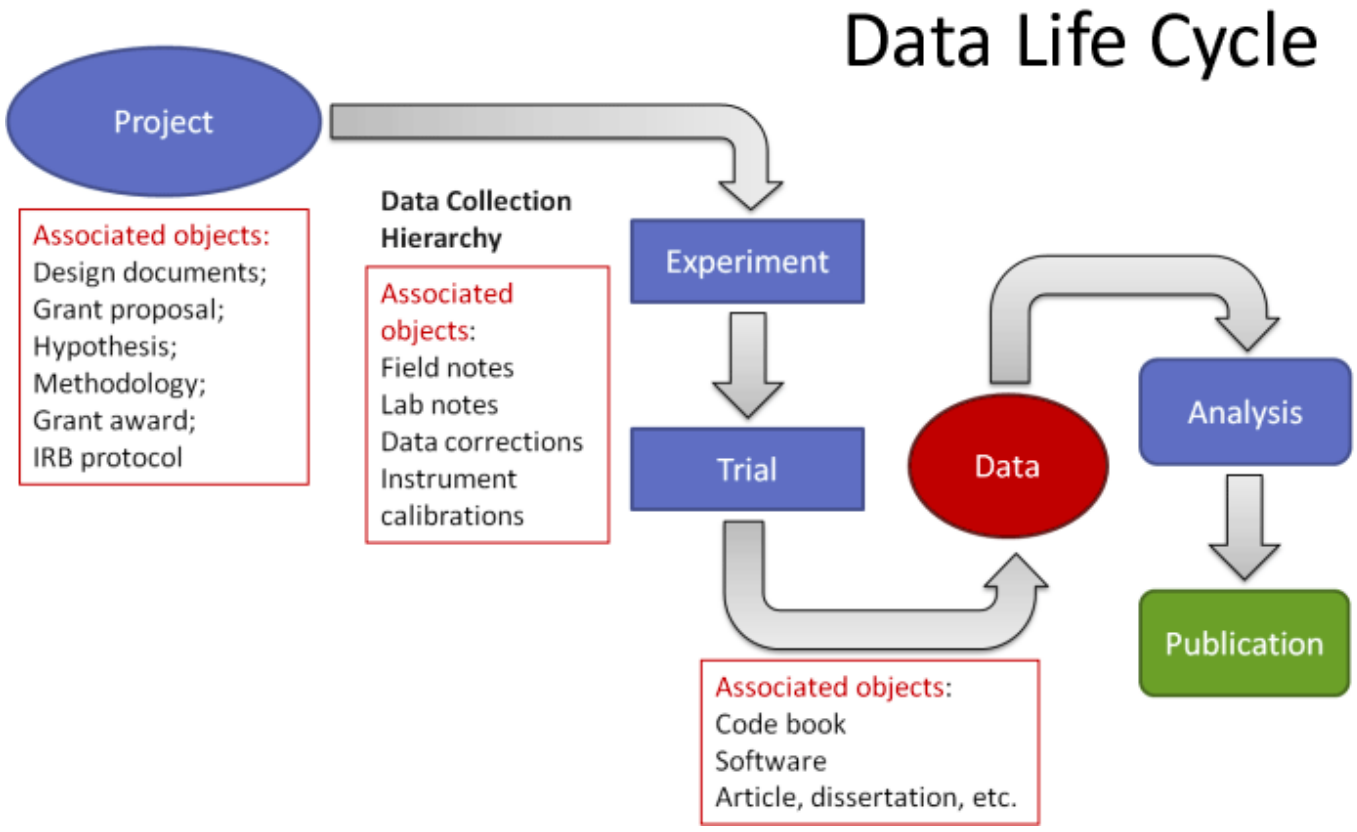
RUresearch has been developed to support Rutgers researchers in the management of their generated research data, utilizing the resources and best practices developed with RUcore, the Rutgers University Community Repository. RUresearch provides the following services:

- Consulting on Data Management Plans and data best practices
- Permanently archiving data in the RUresearch data portal
- Work on larger and more complex data needs in grant-funded projects.

The RUresearch Data Team consists of experienced digital information professionals who work with data, including programmers and developers, metadata experts, and librarians with disciplinary expertise. Some of our members write and manage grants and serve as peer reviewers and consultants for granting agencies, including the National Science Foundation.

The Data Life Cycle

Your Data does not exist in isolation. It is always in process, as illustrated by the *data life cycle.*



The Data Model for your Project

Developing a *data model* for your project supports your work in several ways:

- Identifying entities - the data model identifies all the entities involved in your research that will result in data
- Identifying attributes - once the entities are identified, the attributes represent variables that your data will capture
- Identifying relationships - the relationships between entities are critical in determining both the structure and flow of your data

The data model allows the major data stores and their connections to be visualized. This kind of planning identifies the critical features that the metadata and the structure of the data must support. The data is viewed as an integral part of the desired research process, not simply a final output.

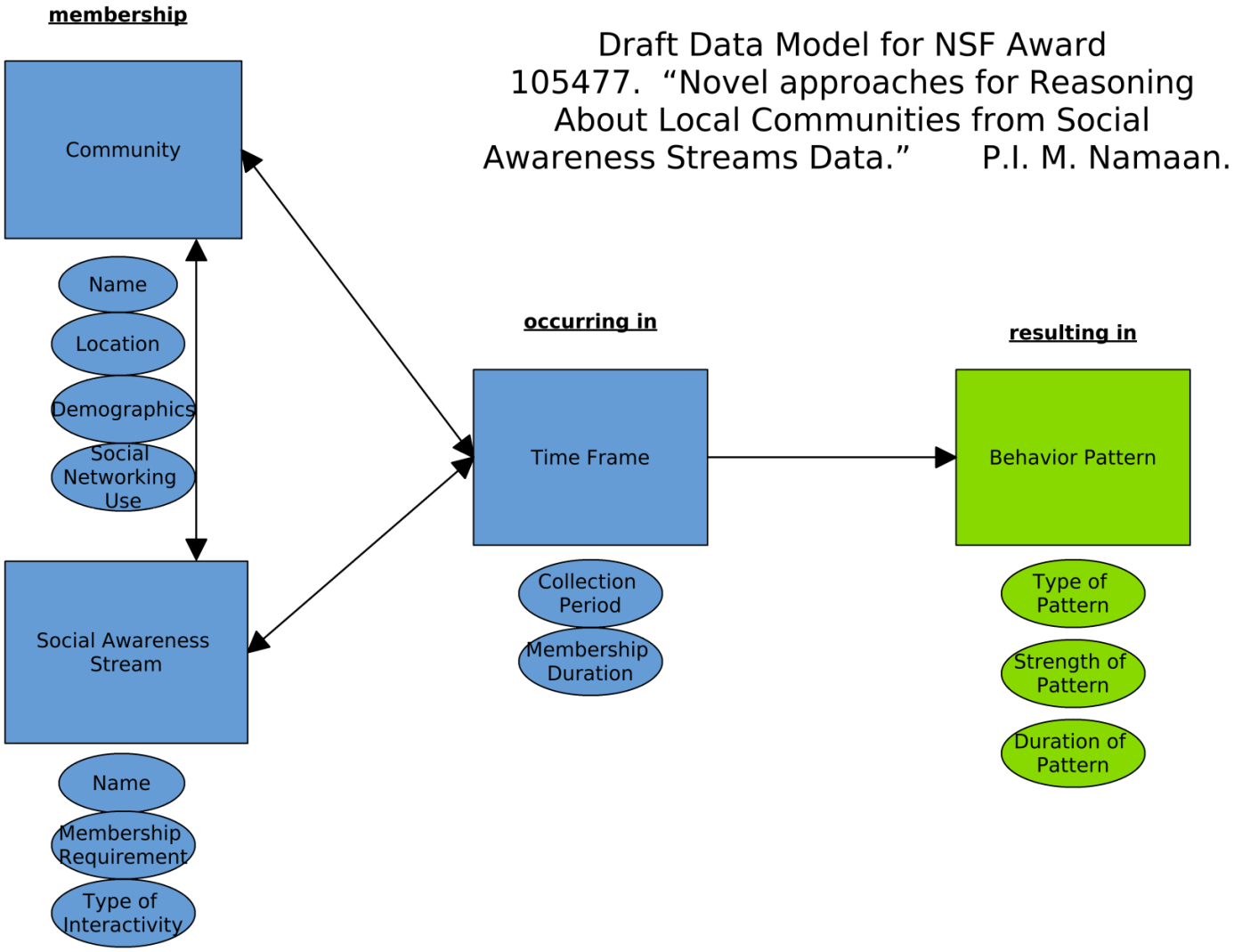
Data Model Support from RUresearch

The RUresearch Data Team will work with you on your data model to develop:

- A core representation that provides a shared understanding of your data.
- A blueprint for designing the metadata for your data.
- an entity-relationship diagram that can help us isolate what is important to describe about your data, to ultimately support searching and display requirements.

Members of the RUresearch Data Team can meet with you to discuss your research project. During the initial consultation, we will work with you to understand and specify your data model, which will provide the basis for the data and supporting documentation that will ultimately be archived in RUcore or another location of your choosing.

A Sample Data Model



What is *Metadata?*

- Metadata is "data about data." Metadata is essential first to organize and manage data, and second to allow for data's discovery and reuse.
- Good metadata is responsive to the information needs of its user community. It captures the information most important for that community, using terminology that is accurate, current and meaningful to that community. It also needs to be consistently applied and shareable with a broader community.
- Metadata standards evolved to enable consistency and broader sharing of information. One of the oldest and most well-known is Dublin Core, a 15-element metadata standard that is widely used. Many research communities have evolved their own standards, such as Darwin Core for biological specimens and DDI (data documentation initiative) for survey data. Here is an example of metadata from DRYAD, and another from ICPSR.

Metadata and RUresearch

- RUresearch employs a very flexible, sophisticated event-based metadata implementation that supports many different metadata standards but is largely independent of any one standard. We can display and export records in many different standards, including the standard your community uses.
- RUresearch metadata contains descriptive elements describing the content of objects (subject, keywords, variables, extent), technical metadata (filetypes, date stamps, revision information), and rights metadata (copyright, licensing, access).
- The RUresearch team can provide advice on appropriate metadata standards and existing controlled vocabularies. As a fee-based service, we can also design customized metadata that can support many standards or serve as a community standard, specific to your project's needs.
- For projects preserved in RUresearch, we will create metadata for all objects in a collection. We can also create metadata for data hosted elsewhere. These are both fee-based services, except for small projects.

What do we mean by *Policies?*

- Licensing, Access Control, and more

It is important to consider the *policies* governing the use of your data.

- Data itself may not be copyrighted, but its use may be licensed. In order to promote use, you may wish to use an Open Data license.
- Your data may also have privacy and confidentiality concerns.
- Your initial IRB protocol may specify certain treatment of the data.

RUresearch can maintain separate licensing and access control information for each component of data, allowing complete control over these aspects of your data. Archiving supplementary documents, such as IRB protocols and contracts helps ensure proper use of the data in the long term. For example, we can maintain a confidential IRB file, response data, and agreements with study participants in a "dark" archive, while simultaneously making public use data and descriptive information available.

Support for Data Preservation

It may be appropriate for you to archive your data in a disciplinary repository or other location. This is an excellent way to provide visibility for your data, but does not guarantee long-term preservation. If you, your research group, or your department are archiving the data, are you maintaining a local and an offline backup?

RUresearch employs "industry best practices" for digital file preservation, including:

- Multiple backups and restoration practices including online, nearline, offline and offsite storage of files.
- Continuous file integrity checks, such as checksum assignment and checking.
- Persistent identifiers that use metadata to continuously locate a file, even if it is moved during routine storage reallocation. When you reference a citation URL, you can be confident that the le will be retrieved.
- Storage of files in multiple formats. One or more "canonical" formats that are vendor independent and conform to non-proprietary standards are employed whenever feasible. The original file format is also always maintained.

Benefits of Data Preservation

There are other advantages to preserving data with RUresearch.

- RUresearch is non-exclusive. In fact, we can help you organize your data and documentation so that your submissions to other repositories are easier.
- RUresearch maintains original files and formats, creates non-proprietary formats, and migrates file types forward over time.
- All objects receive a permanent URL that allows access at the item level, allowing a durable citation to your data. Indexing by Google, Google Scholar, and other search engines, increases access to and use of your research.
- Our institutional commitment to permanent preservation will outlast many archives dependent on contingent funding.
- RUresearch stores not only data, but associated codebooks, documentation, instrumentation information, and can easily support audio and video associated with projects. RUresearch can also store software code used to analyze data.
- RUresearch is local. We are always available to work with you as your data needs change, and can make additions to the archived data as needed. We will work with you throughout the entire data life cycle.

RUresearch and Large Projects

- Our mission is to collect the significant intellectual output of the university. Individual data sets that involve simple cataloging and storage can be accepted at no cost. If you are going to write us into your data management plan, run it by us first!
- RUresearch provides free consultation on your data management plan or grant, but managing data for a large research project involves significant work and planning that will generally require a fee for service.
- Services include creation and customizing of metadata, data portals, and archiving data and associated documents and software.
- The fee can come through cost recovery charges in the grant budget, either as a data management fee or through the involvement of library faculty and staff as co-P.I.s or researchers on the grant, with associated line item cost recovery.
- This will be a one-time, cost recovery only fee based on the amount of work and effort anticipated for the life of the project.
- Data will be preserved and made accessible for the long term at no additional cost beyond any one-time initial fee.

Demonstration

- The standard RUresearch Portal site
- A customized portal – the Video Mosaic Collaborative

The Video Mosaic Collaborative shows the power of data organized using customized metadata and search tools.

References

- Data Management Checklist (MIT)
  http://libraries.mit.edu/guides/subjects/data-management/
- Data Management Guide (Rutgers University Libraries)
  http://libguides.rutgers.edu/datamanagement
- data.rutgers.edu (Rutgers ORSP)
  http://data.rutgers.edu
- D-Lib magazine - special issue on research data
  http://www.dlib.org/dlib/january11/01contents.html
- Disciplinary Repositories (Purdue D2C2)
  http://d2c2.lib.purdue.edu/OtherRepositories.php
- DMPTool
  http://dmp.cdlib.org/
- Elements and Framework of DMP (ICPSR)
  http://www.icpsr.umich.edu/icpsrweb/ICPSR/dmp/index.jsp

Additional References

- How to License Research Data (Digital Curation Center)
  http://www.dcc.ac.uk/resources/how-guides/license-research-data
- Keeping Research Data Safe
  http://www.beagrie.com/krds.php
- My Data Management Plan - a satire
  http://ivory.idyll.org/blog/may-10/data-management.html
- NSF Data Management Plan
  http://www.nsf.gov/bfa/dias/policy/dmp.jsp
- NSF Data Management Plan FAQs
  http://www.nsf.gov/bfa/dias/policy/dmpfaqs.jsp
- Open Data Commons
  http://opendatacommons.org
- Science magazine - special issue on Dealing with Data
  http://www.sciencemag.org/site/special/data/
- Stolen laptop contains cancer cure data
  http://news.cnet.com/8301-17938_105-20028475-1.html

What do you need?

Let's discuss your data needs...

Contact us at rucore_research@email.rutgers.edu

or visit http://rucore.libraries.rutgers.edu/research to initiate a data consultation.