INTRODUCTION TO DATA MANAGEMENT FOR RESEARCHERS

Grace Agnew, Special Advisor for Strategic Initiatives, RUL
Forough Ghahramani, Associate Director, RDI2
Ron Jantz, Digital Library Architect
Laura Palumbo, Science Data Specialist
Ryan Womack, Data Librarian
May 15, 2018

WHY DATA MANAGEMENT?

- Organize your work
- Preserve for future reuse
- Impact
- Credibility
- Future collaborators and your future self will need guidance on how to use the data you create today.

YOUR DATA

Some basic practices:

Keep raw data pristine and separate from any working data

Document your variables and data collection as you work

Write down anything you yourself would forget when revisiting the project 3 years later in response to a query

That will be the same thing other users need too!

Don’t work in Excel [if you can] or other manual editing environment

You should write down all your steps if you are doing this

Better to use code or an environment that will at least record your steps

FILE ORGANIZATION AND NAMING

- Adopt a file naming convention that is structured and meaningful to you
  - Date, location, method, subject
  - All can be encoded into file name so that file is self-explanatory
- File directory structure should also be logical and well-thought out
  - Separate raw data, modified data, code, documentation, other material
- The larger the project, the greater the benefits of planning for organization and communicating the plan to all project participants

DOCUMENTATION

Readme

Codebook

What do others need to use your data?

Document early and often!

STORAGE AND BACKUP

Rule of 3 => 2 onsite 1 offsite

Dropbox is not backup, it is sync!

Consider Future File Usability – Open Formats

DISTRIBUTION AND SHARING

- OSF
- Dataverse
- OpenICPSR
- Zenodo
- Opensciencedatacloud
- Re3data
- Disciplinary repositories like OOI and PDB
- And Open Data Licensing

Providing Access to Interdisciplinary Research

Major Virtual Data Collaboratory Goals

- Provide researchers, educators, and entrepreneurs from across a broad range of disciplines and scientific domains seamless access to data and tools.
- Enable researchers to develop and apply advanced data management and analysis tools for high impact scientific applications.
- Train the next generation of researchers with deep disciplinary expertise and a high degree of competence in leveraging data, infrastructure, and tools.

VDC development is supported by a major NSF grant and is a collaboration between Penn State, Rutgers, and Temple.

Finding and Using Data – Most Important

- Who created the data, methodology, and time frame.
- Salient aspects of the data and how can it be used.
- Understanding how the data was collected and analyzed.
- Determining the credibility of the data
- Collaborators may want to limit access.
- Concerns about reuse and whether users will understand the context.

Finding Collaborators

- Informal groups
- Researchers who have received grants in similar areas.
- Conference networking – smaller conferences are best!
- Analysis of citations to find the “gate-keepers” of the field in question.

REPRODUCIBILITY AND COLLABORATION

Standards such as DOI and ORCID enable identification and reuse

Literate Programming

Github

Jupyter

Rpubs

WHAT FORCES AFFECT INTERDISCIPLINARY RESEARCH? A DIALOG WITH RESEARCHERS

What is your primary area of research and what are the most important disciplines outside of your primary area?

Describe your approach to doing interdisciplinary research.

- How do you locate a person to get more information?
- How do you find a resource in another discipline?
- How do you determine if the information will be useful and credible?
- In locating an number of results, how do you select the most useful?

WHAT FORCES AFFECT INTERDISCIPLINARY RESEARCH? A DIALOG WITH RESEARCHERS

How do you initiate a search and what are the most important keywords?

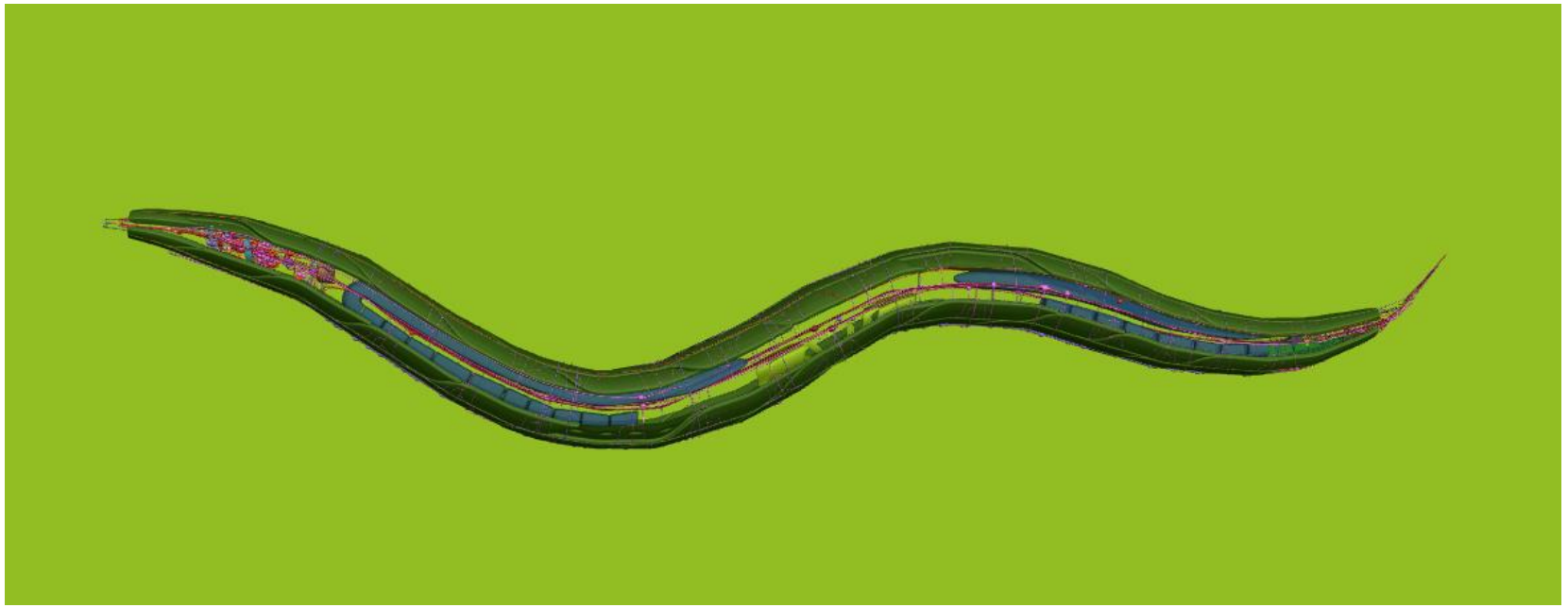
What communities of collaboration are the most important for you?

What are the most important tools that you use in your research?

What are the obstacles that prevent you from carrying out your research?

What are your suggestions for tools and resources that would improve your research?

THE ROUNDWORM
(959 CELLS AND 302 NEURONS)



Exploring and Simulating Virtual Life

(See openworm.org)

Research disciplines: biologists, computer scientists, histologists, mathematicians, molecular geneticists, neuroscientists