Data Management

Introduction to High Performance Computing

HARVARD
COUNTWAY LIBRARY

Instructors

Julie Goldman

Research Data Services Librarian

Meghan Kerr

Archivist and Records Manager

Radhika Khetani

Training Director, Research Scientist

Meeta Mistry

Research Scientist

Kathleen Keating

Research Computing Consultant

Slides: https://datamanagement.hms.harvard.edu/class-materials

HARVARD MEDICAL SCHOOL | Data Management Working Group

Countway Library of Medicine
An Alliance of the Harvard Medical School and Boston Medical Library

Center *for the* History *of* Medicine

Harvard Chan Bioinformatics Core

hms | hsdm
office for postdoctoral fellows

HARVARD MEDICAL SCHOOL | OFFICE FOR Academic and Research Integrity

Department of Systems Biology

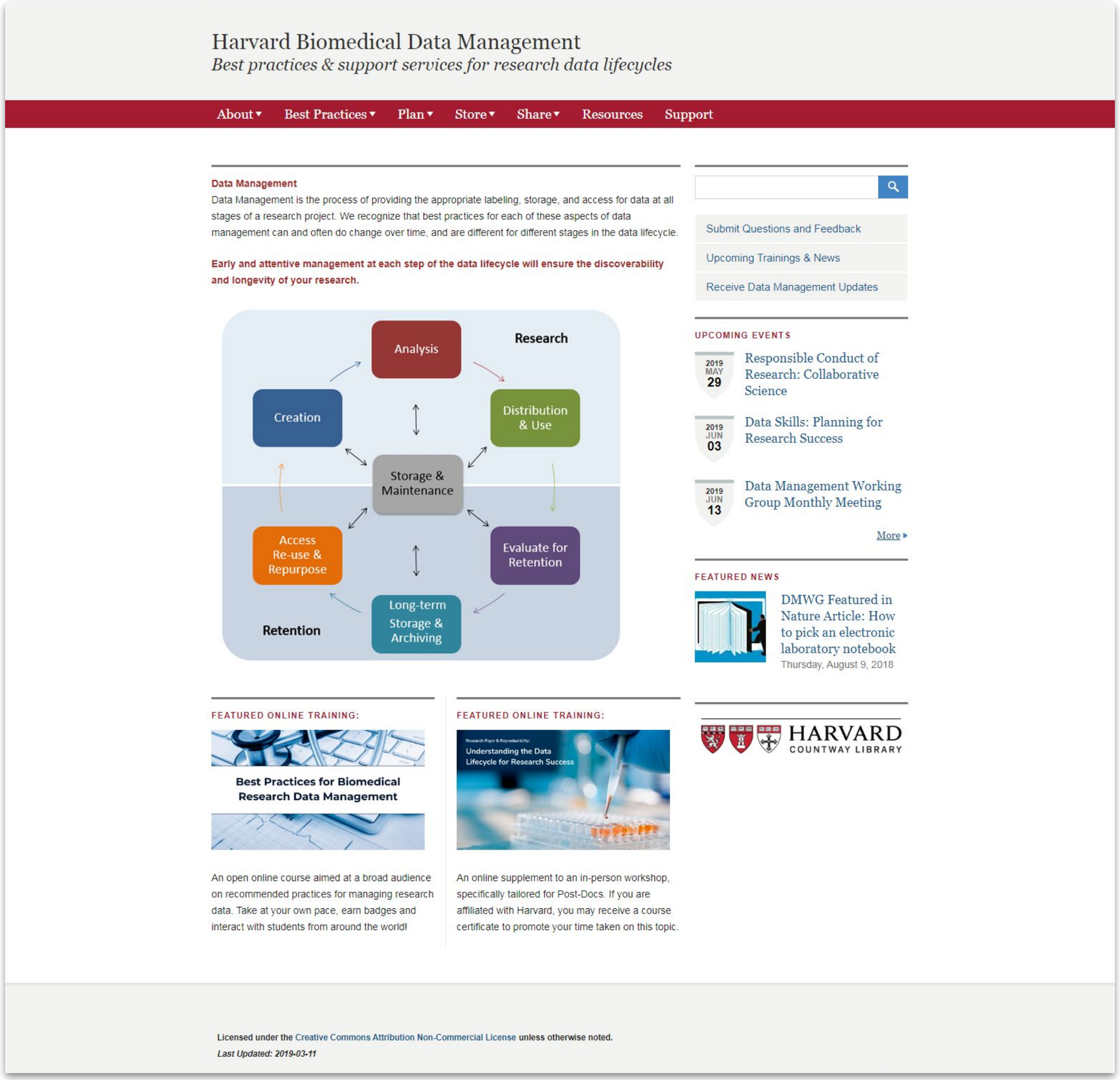
HARVARD MEDICAL SCHOOL
Research Information Technology Solutions - RITS

HMS Information Technology

ICCB-Longwood Screening Facility

DRSC/TRiP Functional Genomics

HiTS
Harvard Program in Therapeutic Science



Harvard Biomedical Data Management

Best practices & support services for research data lifecycles

Data Management

Data Management is the process of providing the appropriate labeling, storage, and access for data at all stages of a research project. We recognize that best practices for each of these aspects of data management can and often do change over time, and are different for different stages in the data lifecycle.

Early and attentive management at each step of the data lifecycle will ensure the discoverability and longevity of your research.

Submit Questions and Feedback
Upcoming Trainings & News
Receive Data Management Updates

UPCOMING EVENTS

2019 MAY 29 Responsible Conduct of Research: Collaborative Science

2019 JUN 03 Data Skills: Planning for Research Success

2019 JUN 13 Data Management Working Group Monthly Meeting

More ▸

FEATURED NEWS

DMWG Featured in Nature Article: How to pick an electronic laboratory notebook

Thursday, August 9, 2018

HARVARD COUNTWAY LIBRARY

FEATURED ONLINE TRAINING:

Best Practices for Biomedical Research Data Management

An open online course aimed at a broad audience on recommended practices for managing research data. Take at your own pace, earn badges and interact with students from around the world!

FEATURED ONLINE TRAINING:

Research Rigor & Reproducibility: Understanding the Data Lifecycle for Research Success

An online supplement to an in-person workshop, specifically tailored for Post-Docs. If you are affiliated with Harvard, you may receive a course certificate to promote your time taken on this topic.

Last Updated: 2019-03-11

Harvard Biomedical Data Management Website

https://datamanagement.hms.harvard.edu

Why Manage Data?

- Running the same workflow can be labour intensive
- Manual manipulation of data files:
  - is often not captured in documentation
  - is hard to reproduce
  - is hard to troubleshoot, review, or improve
- Hard to find poorly organized, documented data
- Hard to analyze poorly recorded workflows

https://datamanagement.hms.harvard.edu/overview

Why HPC?

- High Performance Computing makes workflows more efficient
  - If you work with a lot of data or you have really complex computations, scheduling scripts reduces computation time
  - Automated workflows makes you more productive and also improves the reproducibility of your work by allowing you to save and repeat them
- Using a command line interface to work with files
  - Every step can be captured in the shell script and allow reproducibility and easy troubleshooting
- Offers storage space for active data files and shared drives for sharing data between labs

Training Materials

https://tinyurl.com/hpc-july11

Workshop Outline

Lessons	Estimated Duration
Intro to High-Performance Computing	25 min
Intro to O2	55 min

Tying it Together

Why Data Management:

Not a prerequisite of HPC, but data should be organized in a clear and predictable manner.

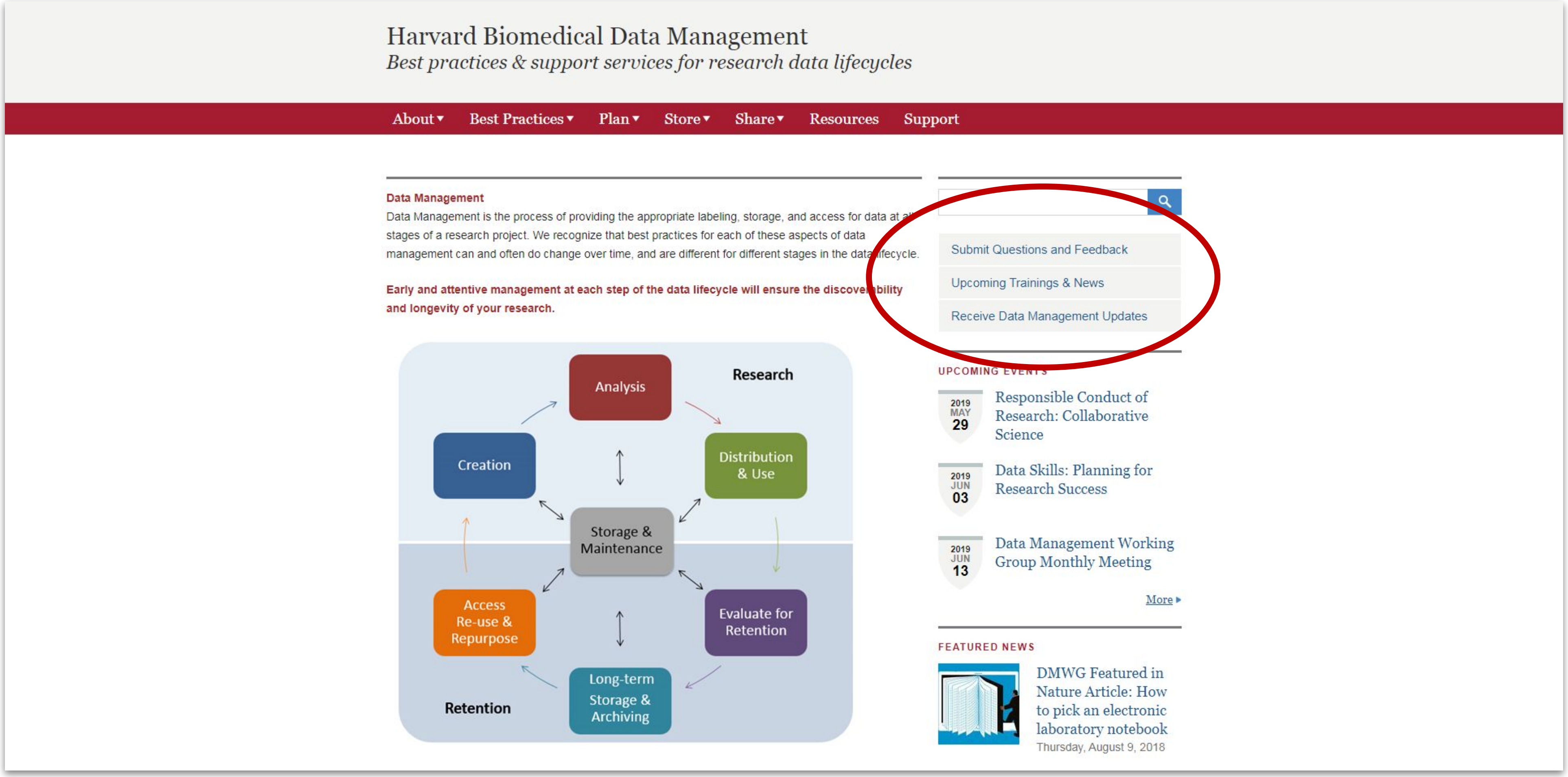
Taking the time to structure your research data and filenaming conventions in a consistent and predictable manner is certainly a significant step towards getting the most out of data analysis.

Why HPC:

Allows you to reduce computation time and help make analyses more efficient. Using a cluster offers advantages such as: speed, volume, efficiency, cost, and convenience.

Automate repetitive tasks and capture small data manipulation steps that are normally not recorded to make research reproducible.

Questions?



Harvard Biomedical Data Management

Best practices & support services for research data lifecycles

Data Management

Data Management is the process of providing the appropriate labeling, storage, and access for data at all stages of a research project. We recognize that best practices for each of these aspects of data management can and often do change over time, and are different for different stages in the data lifecycle.

Early and attentive management at each step of the data lifecycle will ensure the discoverability and longevity of your research.

Submit Questions and Feedback
Upcoming Trainings & News
Receive Data Management Updates

UPCOMING EVENTS

2019 MAY 29	Responsible Conduct of Research: Collaborative Science
2019 JUN 03	Data Skills: Planning for Research Success
2019 JUN 13	Data Management Working Group Monthly Meeting

More ▸

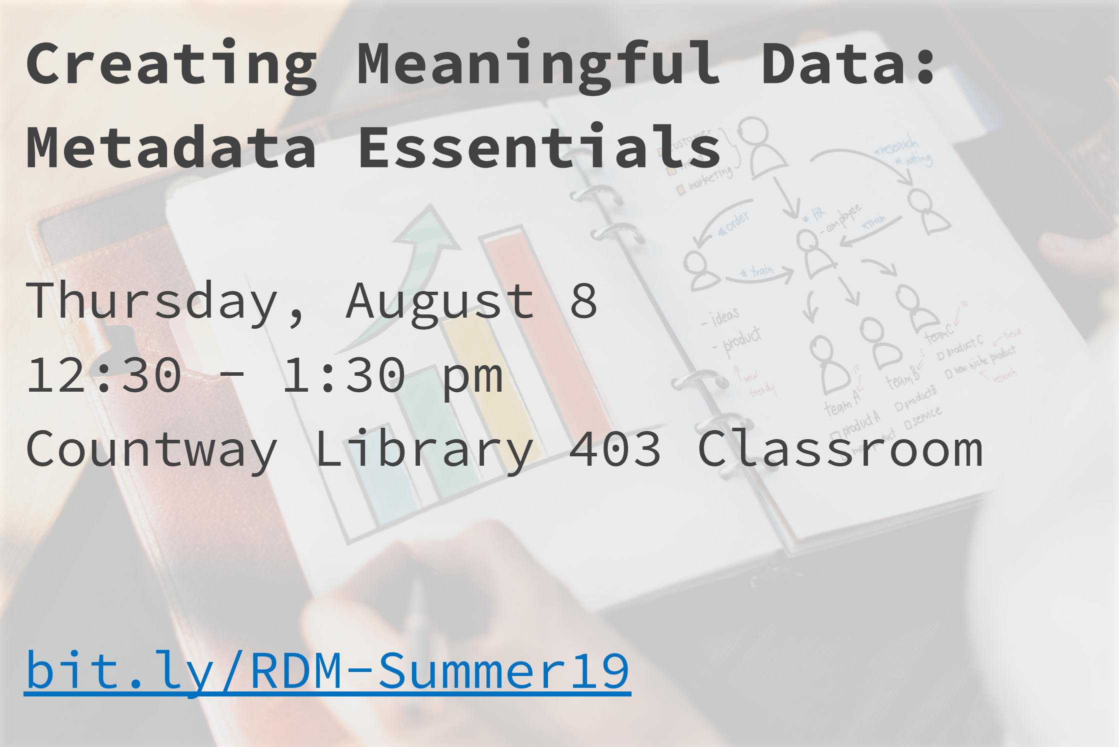
FEATURED NEWS

DMWG Featured in Nature Article: How to pick an electronic laboratory notebook

Thursday, August 9, 2018

https://datamanagement.hms.harvard.edu

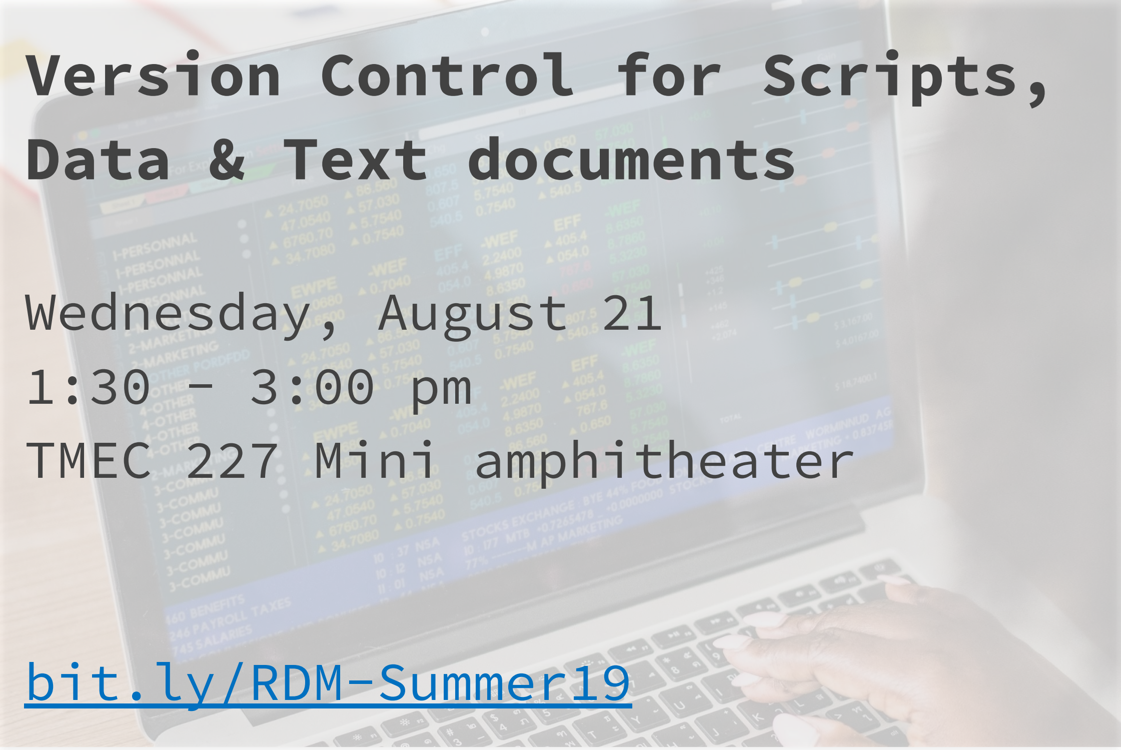
Upcoming Workshops / Seminars



Creating Meaningful Data: Metadata Essentials

Thursday, August 8
12:30 - 1:30 pm
Countway Library 403 Classroom

bit.ly/RDM-Summer19



Version Control for Scripts, Data & Text documents

Wednesday, August 21
1:30 - 3:00 pm
TMEC 227 Mini amphitheater

bit.ly/RDM-Summer19

bit.ly/rdm-survey