Naming Things

HARVARD LONGWOOD MEDICAL AREA
RESEARCH DATA MANAGEMENT WORKING GROUP

We can help!

- Digital Repositories & Manuscript Submission
- NIH Public Access Policy Compliance
- Copyright & Fair Use
- Author Profiles & CV
- Research Data Management & Data Sharing
- Online Collaboration Tools
- Bioinformatics
- Research Assessment & Impact Metrics
- Scholarly Publishing & Manuscript Preparation

HARVARD
COUNTWAY LIBRARY

Publishing & Data Services Team

Scott Lapinski, MS
Associate Director, Scholarly Communication and Open Science Initiatives
Email
617-432-6784

Reddy Gali, PhD
Bioinformatics Support Specialist
Email
617-432-7471

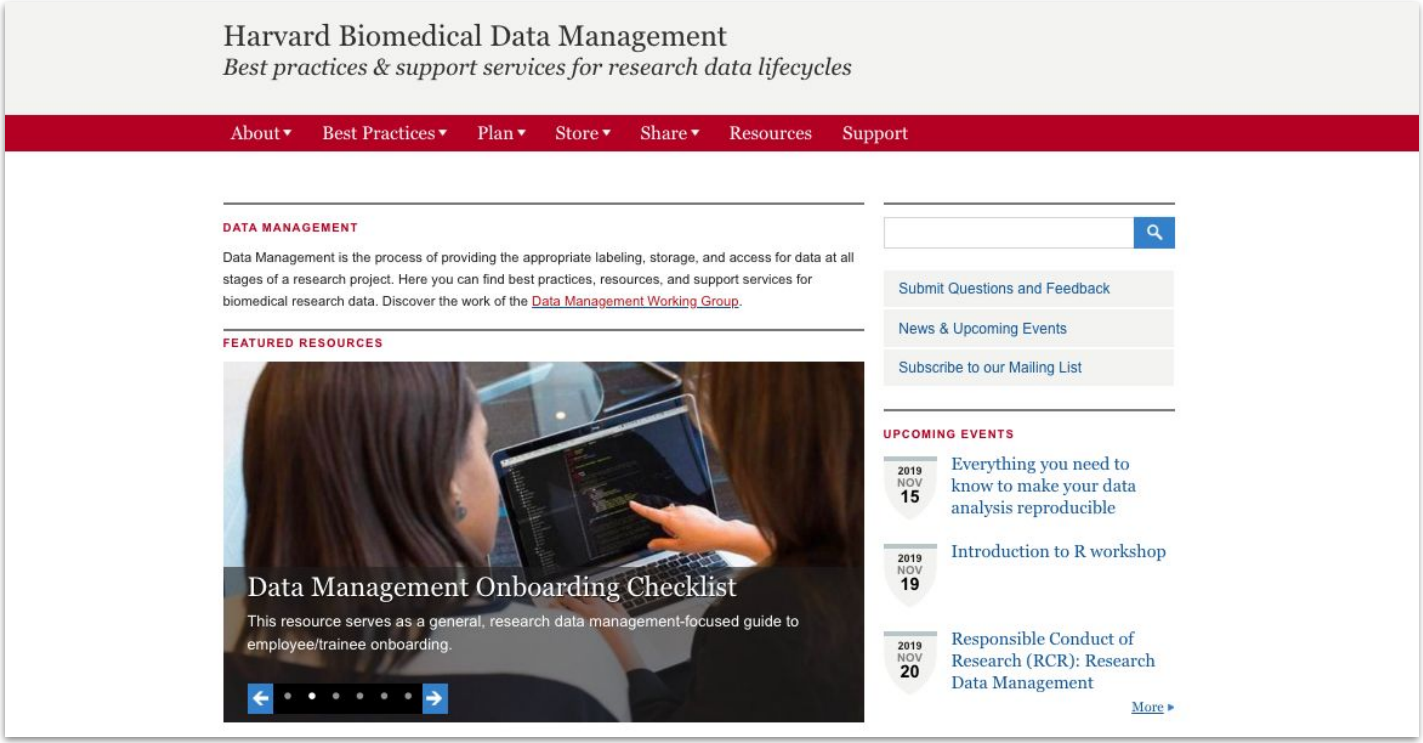
Julie Goldman, MLIS
Research Data Services Librarian
Email

Iris Jahng, MLIS, MA
Digital Scholarship Librarian
Email

Ashley Thomas, MLIS, MA
RDMLA Librarian
Email

Bookmark this website!

https://datamanagement.hms.harvard.edu



Learning Objectives

- Understand why naming conventions is essential for data management
- Learn best practices for file naming
- Design naming conventions for your project or group
- Find out who to contact for assistance

Based on the Data Carpentry Lesson File Organization: Naming
https://datacarpentry.org/rr-organization1/01-file-naming/index.html

"Research data management concerns the organization of data, from its entry to the research cycle through the dissemination and archiving of valuable results. It aims to ensure reliable verification of results, and permits new and innovative research built on existing information."

Whyte, A., Tedds, J. (2011). ‘Making the Case for Research Data Management’. DCC Briefing Papers. Edinburgh: Digital Curation Centre

Why should you care?

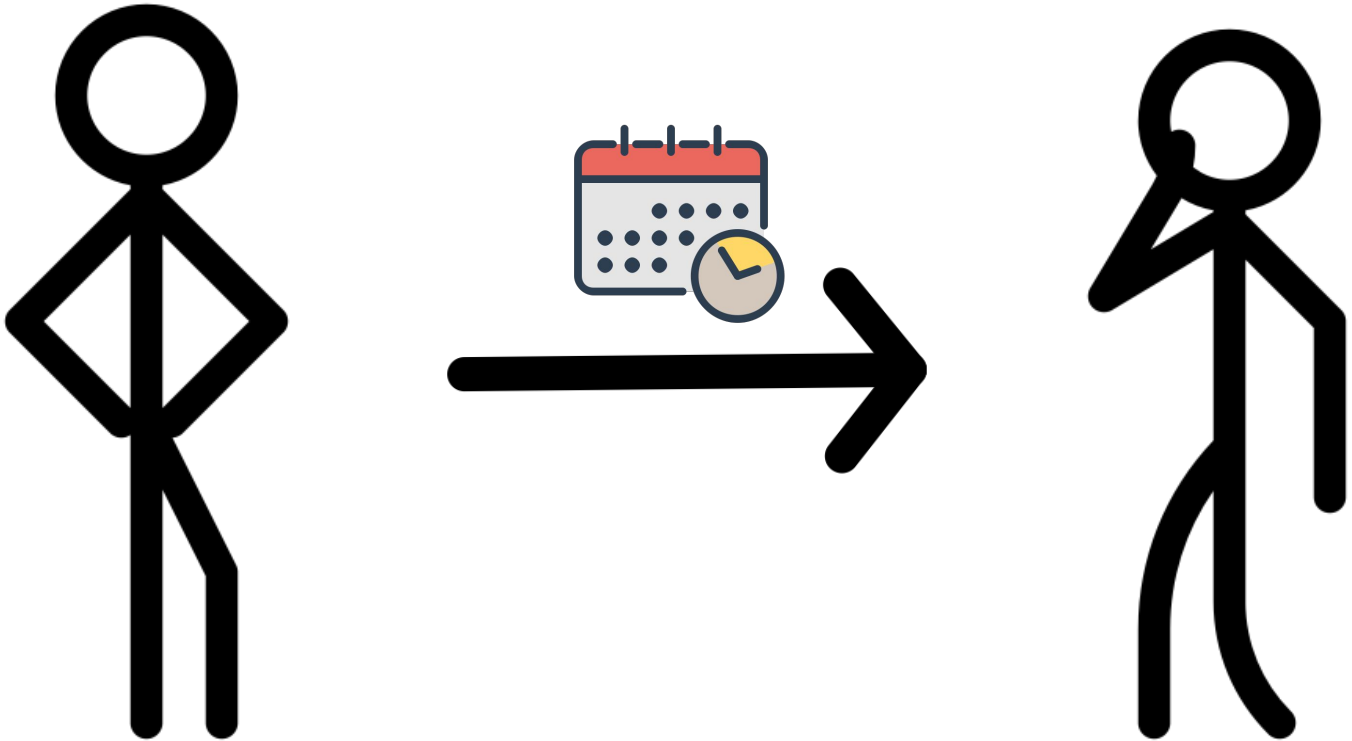
Benefits for yourself, researchers, and science:

- Your future self will thank you!
- Facilitate and ensure seamless team transitions
- Conduct analysis effectively when collaborating with others
- Check and verify research results
- Support FAIR principles: Findable, Accessible, Interoperable, & Reusable

Compliance:

- Be **compliant with University and School policies**
- Be **compliant with funding organizations** that require a data management plan and data sharing
- Be **compliant with journals** that require to submit your data accompanying the article

Naming Things



Three Principles for (File) Names

- Machine readable
- Human readable
- Plays well with default ordering

Examples of really awesome file names

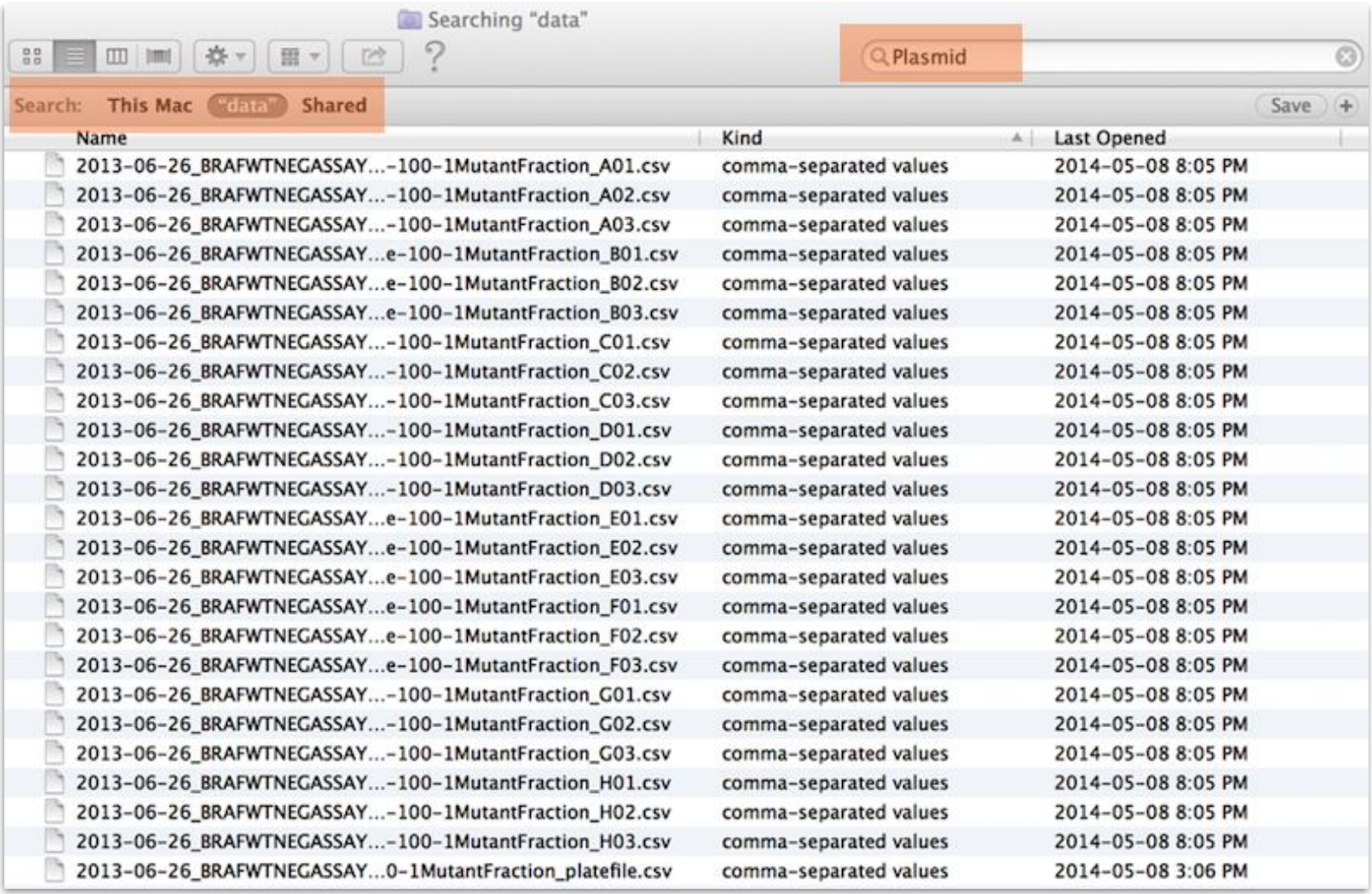
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H01.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H02.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H03.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_platefile.csv
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A01.csv
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A02.csv
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A03.csv
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A04.csv

Machine Readable

- Easy to search for files later
- Easy to narrow file lists based on names
- Easy to extract info from file names

Machine Readable

- Regular expression and globbing friendly! Be sure to avoid:
  - spaces in file names
  - punctuation
  - accented characters
  - case sensitivity (“test” vs “Test”)
- Easy to compute on
  - deliberate use of delimiters


Searching "data"
Plasmid
Search: This Mac "data" Shared
Save
Name
Kind
Last Opened
2013-06-26_BRAFWTNEGASSAY...-100-1MutantFraction_A01.csv comma-separated values 2014-05-08 8:05 PM
2013-06-26_BRAFWTNEGASSAY...-100-1MutantFraction_A02.csv comma-separated values 2014-05-08 8:05 PM
2013-06-26_BRAFWTNEGASSAY...-100-1MutantFraction_A03.csv comma-separated values 2014-05-08 8:05 PM
2013-06-26_BRAFWTNEGASSAY...e-100-1MutantFraction_B01.csv comma-separated values 2014-05-08 8:05 PM
2013-06-26_BRAFWTNEGASSAY...e-100-1MutantFraction_B02.csv comma-separated values 2014-05-08 8:05 PM
2013-06-26_BRAFWTNEGASSAY...e-100-1MutantFraction_B03.csv comma-separated values 2014-05-08 8:05 PM
2013-06-26_BRAFWTNEGASSAY...-100-1MutantFraction_C01.csv comma-separated values 2014-05-08 8:05 PM
2013-06-26_BRAFWTNEGASSAY...-100-1MutantFraction_C02.csv comma-separated values 2014-05-08 8:05 PM
2013-06-26_BRAFWTNEGASSAY...-100-1MutantFraction_C03.csv comma-separated values 2014-05-08 8:05 PM
2013-06-26_BRAFWTNEGASSAY...-100-1MutantFraction_D01.csv comma-separated values 2014-05-08 8:05 PM
2013-06-26_BRAFWTNEGASSAY...-100-1MutantFraction_D02.csv comma-separated values 2014-05-08 8:05 PM
2013-06-26_BRAFWTNEGASSAY...-100-1MutantFraction_D03.csv comma-separated values 2014-05-08 8:05 PM
2013-06-26_BRAFWTNEGASSAY...e-100-1MutantFraction_E01.csv comma-separated values 2014-05-08 8:05 PM
2013-06-26_BRAFWTNEGASSAY...e-100-1MutantFraction_E02.csv comma-separated values 2014-05-08 8:05 PM
2013-06-26_BRAFWTNEGASSAY...e-100-1MutantFraction_E03.csv comma-separated values 2014-05-08 8:05 PM
2013-06-26_BRAFWTNEGASSAY...e-100-1MutantFraction_F01.csv comma-separated values 2014-05-08 8:05 PM
2013-06-26_BRAFWTNEGASSAY...e-100-1MutantFraction_F02.csv comma-separated values 2014-05-08 8:05 PM
2013-06-26_BRAFWTNEGASSAY...e-100-1MutantFraction_F03.csv comma-separated values 2014-05-08 8:05 PM
2013-06-26_BRAFWTNEGASSAY...-100-1MutantFraction_G01.csv comma-separated values 2014-05-08 8:05 PM
2013-06-26_BRAFWTNEGASSAY...-100-1MutantFraction_G02.csv comma-separated values 2014-05-08 8:05 PM
2013-06-26_BRAFWTNEGASSAY...-100-1MutantFraction_G03.csv comma-separated values 2014-05-08 8:05 PM
2013-06-26_BRAFWTNEGASSAY...-100-1MutantFraction_H01.csv comma-separated values 2014-05-08 8:05 PM
2013-06-26_BRAFWTNEGASSAY...-100-1MutantFraction_H02.csv comma-separated values 2014-05-08 8:05 PM
2013-06-26_BRAFWTNEGASSAY...-100-1MutantFraction_H03.csv comma-separated values 2014-05-08 8:05 PM
2013-06-26_BRAFWTNEGASSAY...0-1MutantFraction_platefile.csv comma-separated values 2014-05-08 3:06 PM

Globbing → in Shell

```
Jennifers-MacBook-Pro-3:2014-03-21 jenny$ ls *Plasmid*

2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_A01.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_A02.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_A03.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_B01.csv

...

2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H03.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_plate.csv
```

Regular Expressions → in R

```
> list.files(pattern = "Plasmid") %>% head

[1] "2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_A01.csv"
[2] "2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_A02.csv"
[3] "2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_A03.csv"
[4] "2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_B01.csv"
[5] "2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_B02.csv"
[6] "2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_B03.csv"
```

Breaking it Down

```
> flist <- list.files(pattern = "Plasmid") %>% head

> stringr::str_split_fixed(flist, "[_\\.]", 5)

     [,1]         [,2]             [,3]                                [,4]  [,5]
[1,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "A01" "csv"
[2,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "A02" "csv"
[3,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "A03" "csv"
[4,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "B01" "csv"
[5,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "B02" "csv"
[6,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "B03" "csv"
```

date assay sample set well file type

Human Readable

- Name contains information on content
- Connects to concept of a slug from semantic URLs
- Easy to figure out what something is based on its name

Human Readable

```
Jennifers-MacBook-Pro-3:analysis jenny$ ls -1

01_marshal-data.md
01_marshal-data.r
02_pre-dea-filtering.md
02_pre-dea-filtering.r
03_dea-with-limma-voom.md
03_dea-with-limma-voom.r
04_explore-dea-results.md
04_explore-dea-results.r
90_limma-model-term-name-fiasco.md
90_limma-model-term-name-fiasco.r
helper01_load-counts.r
Helper02_load-exp-des.r
helper03_load-focus-statinf.r
Helper04_extract-and-tidy.r
Makefile
figure
01.md
01.r
02.md
02.r
03.md
03.r
04.md
04.r
90.md
90.r
helper01.r
helper02.r
helper03.r
helper04.r
tmp.txt
```

Which set of file(name)s do you want at 3am before a deadline?

Embrace the Slug

```
01_marshal-data.r
02_pre-dea-filtering.r
03_dea-with-limma-voom.r
04_explore-dea-results.r
90_limma-model-term-name-fiasco.r
helper01_load-counts.r
helper02_load-exp-des.r
helper03_load-focus-statinf.r
helper04_extract-and-tidy.r
```

Easy to figure out what the heck something is based on its name

Plays Well with Default Ordering

- Put something numeric first
- Left pad other numbers with zeros
- Use the ISO 8601 standard for dates

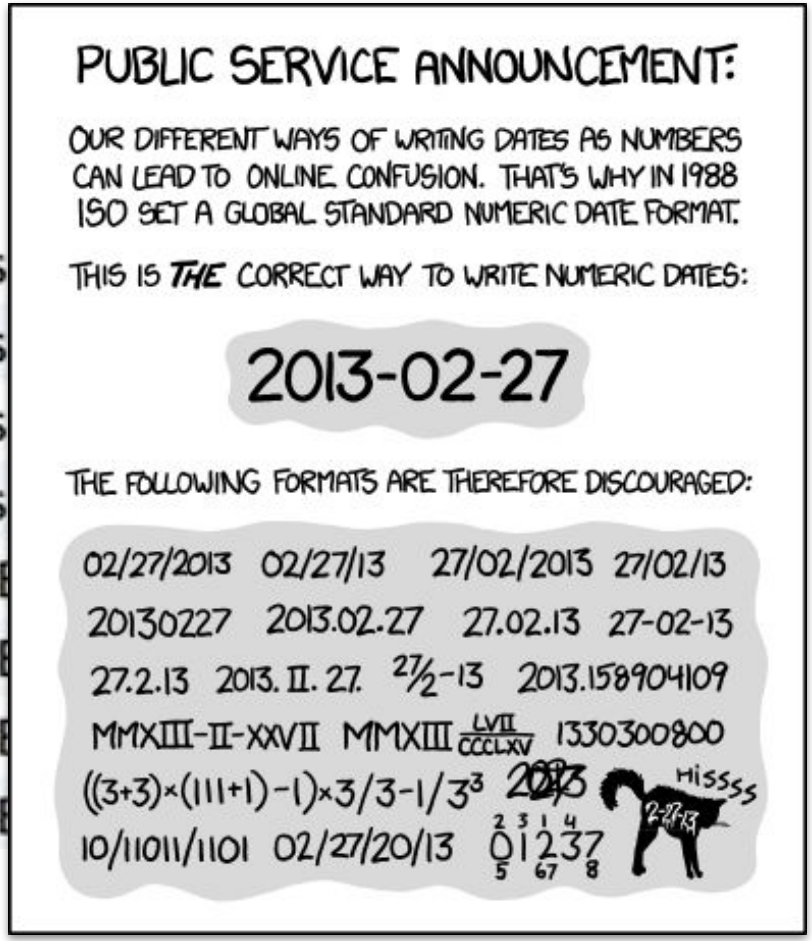
Numbering

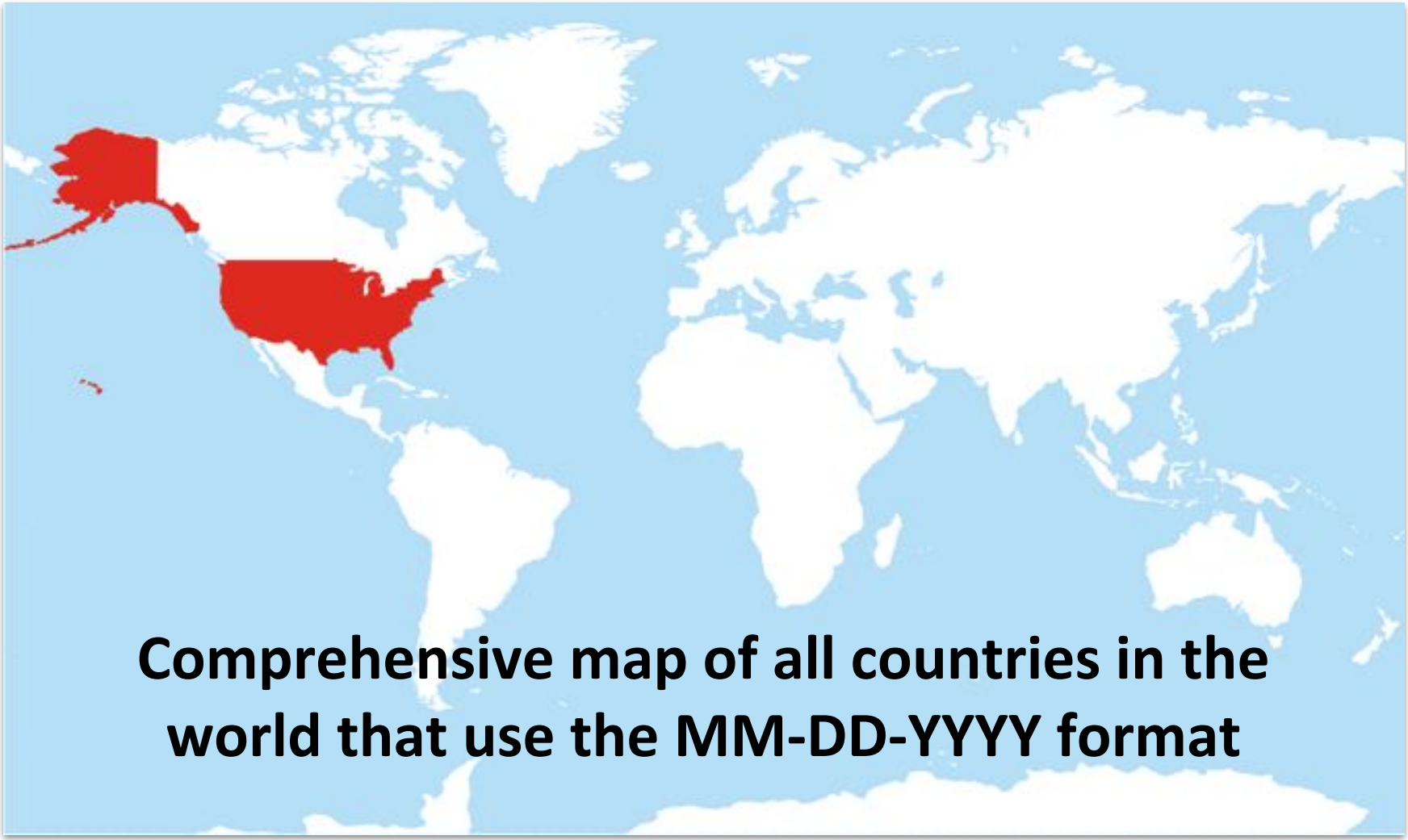
```
01_marshal-data.r
02_pre-dea-filtering.r
03_dea-with-limma-voom.r
04_explore-dea-results.r
90_limma-model-term-name-fiasco.r
helper01_load-counts.r
helper02_load-exp-des.r
helper03_load-focus-statinf.r
helper04_extract-and-tidy.r
```

If you don't left pad, you get this:

```
10_final-figs-for-publication.R
1_data-cleaning.R
2_fit-model.R
```

ISO 8601 Standard




Comprehensive map of all countries in the world that use the MM-DD-YYYY format

Renaming Files

Mac:

1. In the Finder, select and highlight the files you want to change using Shift or Command.
2. Right click/Control + click on the selected files and choose “Rename X Items”.
3. Select one of the rename options: swap out text, add a set text, or apply a custom sequential format.

File Renaming Tool: NameChanger

Windows:

1. In the File Manager, select and highlight the files you want to change using Shift or Control.
2. Right click on the selected files and choose “Rename”. One of the file names will become active for editing.
3. Enter a systematic file name and press enter.
4. All files will be renamed using the chosen file name and numbered sequentially (1) (2), etc.

File Renaming Tool: Rename It!

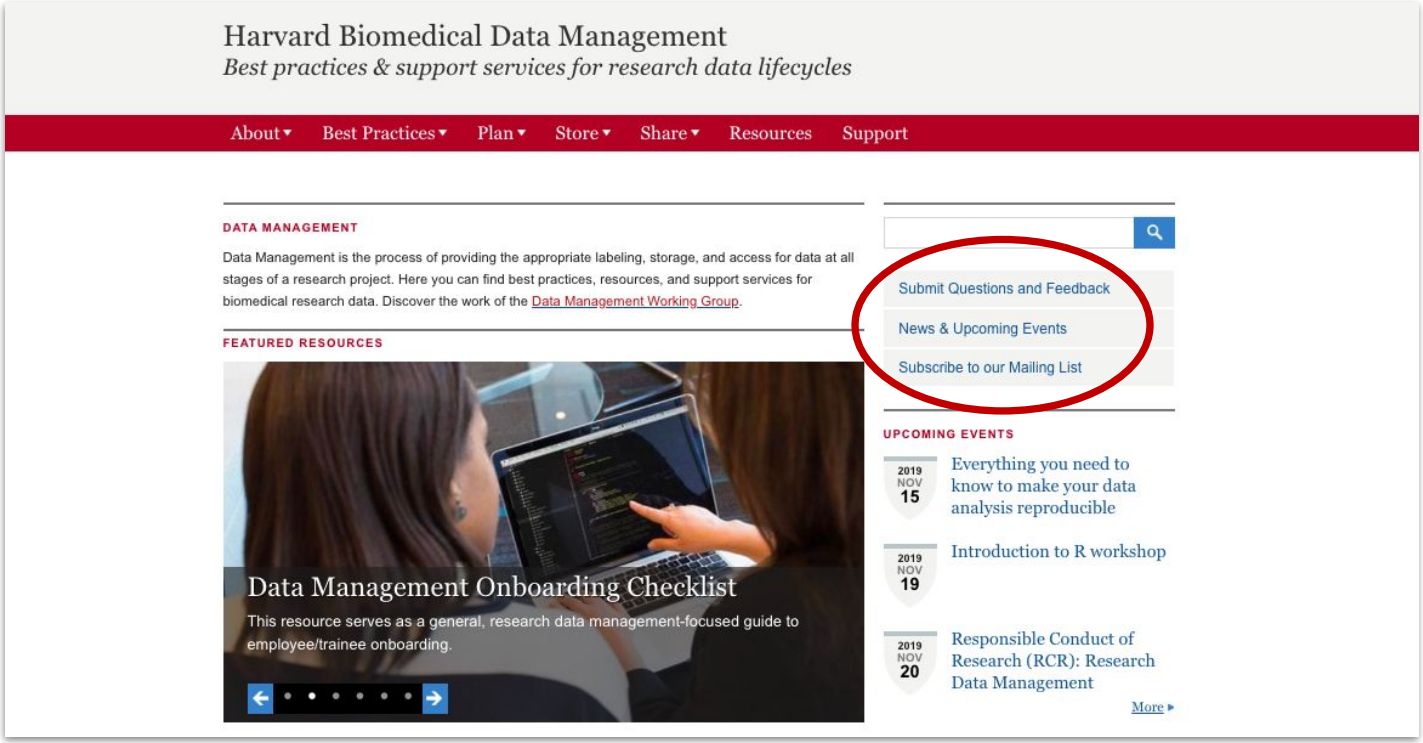
More suggestions: https://datamanagement.hms.harvard.edu/file-naming-conventions

Takeaways

- Remember the three principles for names: machine readable, human readable, and play well with default ordering.
- Check for established file naming conventions in your discipline or in your group. Naming conventions should be documented so that others in your lab or department can follow this standard.
- Payoffs accumulate as your skills evolve and projects get more complex. File names should be descriptive and provide just enough contextual information.
- Check your instruments, software, or other equipment to see if the data outputs can be set with a file naming system.

Submit Questions

https://datamanagement.hms.harvard.edu



Upcoming Data Management Online Seminars

Version Control Basics

Wednesday, May 20, 2020
1:00pm - 1:30pm
Online - Zoom

bit.ly/RDM-Seminars

Tidy Data is Good Data

Tuesday, May 26, 2020
3:00pm - 3:30pm
Online - Zoom

bit.ly/RDM-Seminars

Thank you!

Please fill out this survey

bit.ly/rdm-online