Research Data Management:
Organize, standardize and control your files

Hilda Teresa Ayala González

March 19, 2019
Graduate Research and Innovation Center (GRIC)
University of Puerto Rico, Mayaguez Campus

This activity is sponsored by the Transformational Initiative for Graduate Education and Research (TIGER), US Department of Education, Title V, Part B, Promoting Postbaccalaureate Opportunities for Hispanic Americans (PPOHA) Program (#P031M140035).
Goals

• Understand the benefits of organizing, standardizing, and controlling your research data & documentation

• Recognize good data management practices for organizing and naming your files

• Learn about strategies and tools to enhance your organization skills
Pre-test

Link: https://bit.ly/2DI5amo

Scan QR Code
Organizing, standardizing, and controlling your research data & documentation

Benefits
Benefits

• Files are easier to find & share
  • By you or your colleagues
• Content is recognizable
  • No need to open to confirm
  • Naming should be enough
• Back ups are easier
  • Locations are set
  • Quantity is more reliable

• Less time looking for things
• Decreased risk of loss
• Increase discoverability
• Easier interpretation
• Less time explaining
• Little to none re-fixing
• Unnecessary duplication
• Faster disposition or sharing
Exercise #1

Getting to know you

Think about your:

- **Responsibility** as a grad student
- **What activities** you perform to accomplish it
- **What tasks** are needed to complete an activity

Draw a diagram of your life

10 minutes
My diagram

**THIS IS ME**

Hilda T.

**THESE ARE MY ACTIVITIES**

- GRIC
- ArchiRED
- EGCTI
- Personal

**THESE ARE MY FUNCTIONS**

- Workshops
- Consultation Committee
- Course
- Professional Development

- Projects
- Publication Committee
- Research
- Special Projects

- Conferences
- External Projects
- Students
- Curriculum

**THESE ARE MY TASKS**

- Online
- Associations
Good recordkeeping

**Before**

- Think ahead

**During**

- Look for similarities

**After**

- Completeness

- Activity = thesis, finances
- Task = literature review, bills
- Type = photos, audio, data
- Subject = trees, frogs, flowers

*Rename files & Start to group content*  
*Re-group as needed through the project*

- New needs according to product
  - Publication, Report
- Will another person manage it?
- Clean your files
  - Select what needs to stay
  - Re-group
  - Make sure the “master” is complete in one place
- Document the decisions
- Create copies

- What will you create
  - Images, text, code
- How many will have access
- For how long do you have to keep it
- How are you going to retrieved it
  - Text search, type, folder, tags
Files

Messy
Humongous quantities
Multiple locations
Different formats
Organization

Strategies
Control location of files
Folder & file structures
Metadata
Organization Location of files

- Have a:
  - Temporary Folder
  - Programs Folder
  - Sharing Folder
  - Archive / Delete

- Avoid duplicate
  - Create shortcuts

- Download & re-locate
  - Is not for storage

Keep them out of the desktop
Organization: Location of files

• Things on the cloud
  • Select ONE cloud services for your project
  • Ensure it has version control/log
  • Create under project account, not personal

• Email is not storage
  • Download & archive important records
  • Create subfolders/tags
  • At the end of a project, download content

If it is analog or a legacy file – digitize it, or migrate!
Organization: Folder & File Structures

Folders
Group related records
Work hierarchically

...BUT they are also
Messy
Have multiple locations
Could be analog / digital
Organization: Folder & File Structures

Strategies

• Analyze your files or possible files that will be generated
• Identify patterns
• Group related items together
  • Project / Activity / Task / Type
• Find something that works for you and your co-workers
• Be consistent across systems & platforms
Organization: Folder & File Structures

• Content in folders should not be too much
  • More than 20? Regroup

• Directory structure should not be too deep
  • Five-Seventy levels work ok.

• Don’t repeat folder/files in same project
  • Sometimes it is ok, like with images, but it’s better to avoid overlapping
  • Be more specific

Drawback: HARD to set up & maintain

Images

NO

Yunque_20180212

YES
Organization: Folder & File Structures

Example: Thesis Project

TIPS: For chronological order, date at the beginning or end, not in the middle
Organize by task
Organization: Folder & File Structures

Example: Administrative Project

TIPS: From generic to specific.
Organize by document type
Another example

TIP: Use numbers to force an order
Organize by project

Image: Nikola Vukovic
Organization: One last recommendation

Complete/add metadata to files

- Properties → Summary
- Tags → Keywords

- Properties → Details
- Tags & Categories

Drawbacks: Descriptions can be lost when transferred
Systems don’t share same description fields
Exercise #2

How are your files distributed through your life

1. Use the diagram from the beginning and list some of folders and files that will be generated.

10 minutes
Organization: Folder & File Structures

THESE ARE MY FOLDERS

- Personal
- Professional Development
- Online
- Coursera

THESE ARE MY FILES

- Certificate
  - Completion_HTA
  - Percentage_HTA
- Recordings
  - Day1_Copyright
  - Day2_Copyright
  - Day3_Copyright
- Copyright
- Readings
  - Jonse_2016
  - Jonse_2017
  - Peter_1985
  - Abstrid_2011
Standardization

Image adapted from: Jorge Cham
# Standardization

Your files look like this?

<table>
<thead>
<tr>
<th>Name</th>
<th>Date Modified</th>
<th>Size</th>
<th>Kind</th>
</tr>
</thead>
<tbody>
<tr>
<td>8q79238e7238dy387dewhiu.png</td>
<td>Today, 6:28 AM</td>
<td>386 KB</td>
<td>PNG image</td>
</tr>
<tr>
<td>87987q98e78293e796e2309e782u.jpg</td>
<td>Today, 6:28 AM</td>
<td>410 KB</td>
<td>JPEG image</td>
</tr>
<tr>
<td>090987.pdf</td>
<td>Today, 6:13 AM</td>
<td>143 KB</td>
<td>Adobe...cument</td>
</tr>
<tr>
<td>896433299.pdf</td>
<td>Today, 6:15 AM</td>
<td>1.6 MB</td>
<td>Adobe...cument</td>
</tr>
<tr>
<td>afcb5e29842a4dde75d47627f506941a.png</td>
<td>Today, 6:25 AM</td>
<td>21 KB</td>
<td>PNG image</td>
</tr>
<tr>
<td>Clemmensen.pdf</td>
<td>10/21/18, 7:30 AM</td>
<td>820 KB</td>
<td>Adobe...cument</td>
</tr>
<tr>
<td>Fig3_Martin_LIBER2017_27-1.jpg</td>
<td>Today, 6:27 AM</td>
<td>95 KB</td>
<td>JPEG image</td>
</tr>
<tr>
<td>FILE revised and uploaded.pdf</td>
<td>Today, 6:14 AM</td>
<td>5.8 MB</td>
<td>Adobe...cument</td>
</tr>
<tr>
<td>gify.gif</td>
<td>Yesterday, 7:38 AM</td>
<td>779 KB</td>
<td>GIF image</td>
</tr>
<tr>
<td>MIT.docx</td>
<td>Today, 6:20 AM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SampleFINAL final october advisor revised.docx</td>
<td>Today, 6:20 AM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Screen Shot 2018-10-31 at 4.52.53 AM</td>
<td>Today, 4:53 AM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Screen Shot 2018-10-31 at 5.09.39 AM</td>
<td>Today, 5:09 AM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Screen Shot 2018-10-31 at 5.10.13 AM</td>
<td>Today, 5:10 AM</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Standardization

- How do you find it?
- How long it take?
- Do you have to open it?

- How many times you have found something days/weeks after
- You thought it was lost
- ... send the wrong version
Standardization

Files
Multiple authors / Provenance
Diverse naming strategies
Duplicate locations
Many versions BIG or SMALL changes

Strategies
File naming
Versioning
Standardization: File Naming

Strategies

• Avoid special characters
  • !@#$%^&*?/\, including accents

• Avoid spaces
  • Use an underscore (_) or dash (-)
  • Use a capital letter, but is not always reliable
Standardization: File Naming

Strategies

• Keep names short and descriptive
  • 25-32 should be enough

• Use standard abbreviations
  • Use a style manual / vary per discipline

• Avoid
  • and / or / the / to / not....

Examples:

YES

Thisiswhat25lookslikehere.doc
Thesearethirtytwogreatcharacters.doc
Report_IRB_UPRM.doc

NO

TheThesisandComments.doc
Standardization: File Naming

Strategies
Include elements that facilitates the recognition of the file without opening it

• Project name or number
• Description of content
• Date of creation or modification
• Name of creator or team
• Version number
• Extension file

Look for conventions in your discipline!

Others

• Experiment name
• Location/spatial
• Time/temporal
• Instrument
• Type of data
• Condition/status
Standardization: File Naming

• As possible, incorporate the date in the name of files
  • Don’t trust the automatic date
  • It can change when opened or transferred

• Benefits
  • Helps organize files in chronological order
  • Could facilitated versioning

• Use a standard
  • ISO 8601  YYYY-MM-DD / YYYYMMDD
  • 2018-11-06 / 20181106

TesisFisica_12012018.doc

Is it December 1st or January 12?
# Standardization: File Naming

If you look at this list, can you tell what they are?

<table>
<thead>
<tr>
<th>Name</th>
<th>Date Modified</th>
<th>Size</th>
<th>Kind</th>
</tr>
</thead>
<tbody>
<tr>
<td>8q79238e7238dy387dewhiu.png</td>
<td>Today, 6:28 AM</td>
<td>386 KB</td>
<td>PNG image</td>
</tr>
<tr>
<td>87987q98e78293e796e2309e782u.jpg</td>
<td>Today, 6:28 AM</td>
<td>410 KB</td>
<td>JPEG image</td>
</tr>
<tr>
<td>090987.pdf</td>
<td>Today, 6:13 AM</td>
<td>143 KB</td>
<td>Adobe...cument</td>
</tr>
<tr>
<td>896433299.pdf</td>
<td>Today, 6:15 AM</td>
<td>1.6 MB</td>
<td>Adobe...cument</td>
</tr>
<tr>
<td>afcb5e29842a4dde75d47627f506941a.png</td>
<td>Today, 6:25 AM</td>
<td>21 KB</td>
<td>PNG image</td>
</tr>
<tr>
<td>Clemmensen.pdf</td>
<td>10/21/18, 7:30 AM</td>
<td>820 KB</td>
<td>Adobe...cument</td>
</tr>
<tr>
<td>Fig3_Martin_LIBER2017_27-1.jpg</td>
<td>Today, 6:27 AM</td>
<td>95 KB</td>
<td>JPEG image</td>
</tr>
<tr>
<td>FILE revised and uploaded.pdf</td>
<td>Today, 6:14 AM</td>
<td>5.8 MB</td>
<td>Adobe...cument</td>
</tr>
<tr>
<td>gifhy.gif</td>
<td>Yesterday, 7:38 AM</td>
<td>779 KB</td>
<td>GIF image</td>
</tr>
<tr>
<td>MIT.docx</td>
<td>Today, 6:20 AM</td>
<td>48 KB</td>
<td>Micros...(.docx)</td>
</tr>
<tr>
<td>SampleFINAL final october advisor revised.docx</td>
<td>Today, 6:20 AM</td>
<td>171 KB</td>
<td>Micros...(.docx)</td>
</tr>
<tr>
<td>Screen Shot 2018-10-31 at 4.52.53 AM</td>
<td>Today, 4:53 AM</td>
<td>329 KB</td>
<td>PNG image</td>
</tr>
<tr>
<td>Screen Shot 2018-10-31 at 5.09.39 AM</td>
<td>Today, 5:09 AM</td>
<td>138 KB</td>
<td>PNG image</td>
</tr>
<tr>
<td>Screen Shot 2018-10-31 at 5.10.13 AM</td>
<td>Today, 5:10 AM</td>
<td>16 KB</td>
<td>PNG image</td>
</tr>
</tbody>
</table>
## Standardization: File Naming

**File name example**

<table>
<thead>
<tr>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project name or number</td>
<td>GeneralLibrary_NEH_2018</td>
</tr>
<tr>
<td>Institution / Lab / Project</td>
<td>HGP_PR_Cruzado_2018</td>
</tr>
<tr>
<td>Experiment name</td>
<td>Invoice_Modernica_20181024</td>
</tr>
<tr>
<td>Description of content</td>
<td>20161109_LabNotes_HTA</td>
</tr>
<tr>
<td>Date of creation or modification</td>
<td>BIOL_Thesis2018_HTA</td>
</tr>
<tr>
<td>Name of creator or team</td>
<td>Image_MIC1_HYP_001</td>
</tr>
<tr>
<td>Instrument</td>
<td>RiosPR_Guayanilla_1</td>
</tr>
<tr>
<td>Location/spatial</td>
<td>RiosPR_Guayanilla_1_AM</td>
</tr>
<tr>
<td>Time/temporal</td>
<td>RiosPR_Guayanilla_1_FieldNotes</td>
</tr>
<tr>
<td>Type of data</td>
<td>ProposalNEH_Submitted</td>
</tr>
<tr>
<td>Condition/status</td>
<td>ProposalNEH_HTA_01</td>
</tr>
<tr>
<td>Version number</td>
<td>ReadMe.txt</td>
</tr>
</tbody>
</table>

What about now?  
*Maybe not all, but better...*
## Standardization: File Naming

<table>
<thead>
<tr>
<th>Content to include in file name</th>
<th>Example</th>
<th>File name example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project name or number</td>
<td>Institution / Lab / Project</td>
<td>GeneralLibrary_NEH_2018</td>
</tr>
<tr>
<td>Experiment name</td>
<td>Full or abbreviated</td>
<td>HGP_PR_Cruzado_2018</td>
</tr>
<tr>
<td>Description of content</td>
<td>Type / Format / Content</td>
<td>Invoice_Modernica_20181024</td>
</tr>
<tr>
<td>Date of creation or modification</td>
<td>YYYYMMDD</td>
<td>20161109_LabNotes_HTA</td>
</tr>
<tr>
<td>Name of creator or team</td>
<td>Initials / abbreviated</td>
<td>BIOL_Thesis2018_HTA</td>
</tr>
<tr>
<td>Instrument</td>
<td>Microscope 1 / Microscope 2</td>
<td>Image_MIC1_HYP_001</td>
</tr>
<tr>
<td>Location/spatial</td>
<td>Place</td>
<td>RiosPR_Guayanilla_1</td>
</tr>
<tr>
<td>Time/temporal</td>
<td>AM / PM / Hour</td>
<td>RiosPR_Guayanilla_1_AM</td>
</tr>
<tr>
<td>Type of data</td>
<td>Data Set / Field Notes / Recording</td>
<td>RiosPR_Guayanilla_1_FieldNotes</td>
</tr>
<tr>
<td>Condition/status</td>
<td>Raw / Anonymized / Processed</td>
<td>ProposalNEH_Submitted</td>
</tr>
<tr>
<td>Version number</td>
<td>Full / Incremental</td>
<td>ProposalNEH_HTA_01</td>
</tr>
<tr>
<td>Extension file</td>
<td>.pdf / .txt / .csv</td>
<td>ReadMe.txt</td>
</tr>
</tbody>
</table>
Standardization: File Naming

• What if you are working on a group, with people from different parts of the World, who use different names to refer to the same thing, or the name means something different, depending on their location....

Talk about it! Make a compromise in favor of mutual understanding.
Standardization: Version control

Avoid
- While active: Draft, Revision Thesis_Final_Draft.doc

- Significant changes
  - Date and version
  - V01, V02, V03
  - POP_20120312_V03

- Minor revisions
  - V01-01
  - POP_20120312_V03-01

Other strategies
- Depending on quantity, use 2 or 3 digits V00 - V000, never 1 digit
- Keep older versions in separate Folder
- Erase when the older versions are no longer needed
- Use track changes!
  - Save as new version
Standardization: One last recommendation

→ Document your conventions
  • If you develop a naming system
  • If it has many acronyms or numbers
  • You need them to make sense, now and later

Remember that others or "10 years you” cannot remember all.
Controlling

- Set time for housekeeping
  - Check consistency, not working? Revise it!

- At the end of a project / semester
  - Revise folder structures
    - Needs are different, folders can be merged

- Cleaned your downloads? Desktop?
  - Try to do it weekly
  - Desktop = more than a week?? Not using it? Move it!
Controlling

Multiple copies → duplicity

**OK**
- When records are needed for context or completeness

**NOT OK**
- When you keep downloading the same thing over and over because you forget where it is (a.k.a. you are lazy)

**Strategy**
- Use shortcuts instead
  - Careful, don’t erase THE only copy

**Strategy**
- Keep a better file naming system & folder structure
Controlling

Work or shared computers?

• Always separate work files – personal files

• AVOID keeping personal stuff
  • If inevitable...keep in a separate folder “Personal-HTA”
  • Maintain “your” stuff there
    • Easier to erase when you are done
    • Protect your privacy
    • Control access
Controlling

Work or shared computers?

• Need to share content? → Create a Sharing folder
Tools
Tools

Batch Renaming

- Replace spaces with (_,) or (-)
- Change ambiguous naming

**PSRenamer**

- Open source
- Written in Java
- MAC / Windows / Unix / Linux

There are many other programs... look for *Batch Renaming* and see what fits best for your needs
**Tools**

Find duplicates

- Identify duplicate content
- Decide what to keep and where

**Easy Duplicate Finder**

- Application
- MAC / Windows
- Has a pro version

**NOTE:** Not everything needs to be erased. Always keep a backup, you can erase things you shouldn’t.
Tools

Versioning
- Keep track of each version
- Know who made the change
- Add comment/changes

**Git / Sourcetree**
- Open source
- Distributed
- MAC / Windows / Linux

NOTE: Look for "version control" there's many alternatives
Group Exercise
**Exercise #3**

**File naming for teams**

*Imagine you work in a team project. Each of you have to review some images and submit your findings. They don’t tell you the name or anything, they just sent you these images.*

- Discuss in the table ALL the possible names that will describe the picture
- Select one file naming convention and write an explanation for it

15 minutes
Post-test

Link: https://bit.ly/2Pa0p5t

Scan QR Code
Where to look for more


- GRIC [Gestión de Datos de Investigación](#) Libguide
- MIT [Data Management File Organization](#) Presentation
- University of Michigan [Research Data Management (Health Sciences)](#) Libguide
- Community Owned Digital Preservation [Tool Registry](#) (COPTR)
Evaluation: http://uprm.libsurveys.com/tiger
Password: 2018

Title: RDM: Organize, standardize and control your files
Hilda Teresa Ayala González
March 19, 2019