Data Workflows

ELIZABETH WICKES, DATA CURATOR
RESEARCH DATA SERVICE
UNIVERSITY OF ILLINOIS URBANA CHAMPAIGN
Workflow Workshop Goals

• Know
  • the tools you use
  • the stuff you use
  • where it all lives
  • where it all goes

• Learn
  • How your project workflow works
  • Points where you need clarification
  • How your collaboration with others could be improved

• Practice
  • Mapping out your workflow
Materials

• Preferred:
  • A few pieces of paper
  • A pencil and/or pens in several colors
  • Post it notes in as many colors as you can find

• Minimally:
  • A piece of paper and a writing instrument

• Alternatively:
  • Your imagination
What data do you have?

**Input**
- Source data
- Data from other people

**Process**
- Temporary files
- Intermediate datasets

**Output**
- Output data
- Data for other people
- Data that goes into reports or other final products
And what do you do to it?

Input
- Ingest

Process
- Clean
- Train
- Test

Output
- Analysis
- Write up
- Backup
So how do you science?

Input data
- Investigate data
- Make test data
- Train a model
- Test the model
- Save stats
- Clean the data again
- Write scripts
- Check the algorithm
- Get other data in
- Clean the data
- Make charts
- Join in other data

Output data
- Output data
- Output data
So how do you science?

Input data → investigate data → make a model

→ make some charts
→ join in other data
→ clean the data again

Output data

SCIENCE.
So how do you science?

- Input data
- Input data
- Input data

SCIENCE.

- Investigate data
- Join in data
- Make some data
- Make some charts
- Clean the data
- Train a model
- Test the model
- Get other data
- Check the algorithm
- Write some scripts
- Save stats
- Analysis
- Clean the data again

Output data
Output data
Output data

Don’t forget about us!
Publications

Input data
Input data
Input data

RESEARCH Data Service
But what do I do?

- We’re going to cover an activity to help you think about your projects
  - Can be used **prospectively**
    - to help plan
  - Or **retrospectively**
    - to pick up the pieces
Choose a project

• Something you’re just wrapping up?
• Something you’re in the middle of?
• Something you’re planning for next year?
Activity: Workflow Map

• The intention is not to capture every detail of your workflow, but to help you get a feel for the big picture and points where you may need clarification or other help.

• Default to thinking very high level and generalized

• Remember to use specific, short, and meaningful names you’ll understand 6 months from now
Approaching an initial workflow

• Think about these 3 questions:
  1. What kind of evidence will help answer your research question?
     ◦ Be as specific as possible, but don’t be afraid to generalize at this stage.
  2. What will you do?
     ◦ Use verbs: read, write, script, compute, process, document, etc.
  3. What will you make?
     ◦ Use nouns or named entities: numbers, words, data, graphics, articles, metadata, databases, etc.
### The Board & the Pieces

<table>
<thead>
<tr>
<th>What you <strong>make</strong></th>
<th>Digital objects</th>
<th>or</th>
<th>Objects for you</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical objects</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>What you <strong>do</strong></th>
<th>Activity/Action</th>
</tr>
</thead>
</table>

| What you **use**  | Source object/data | Tool you use |
Make this your own

• You know what you do best
• Use your own voice and words
• Just be sure you’ll be able to understand them later
• So document your changes, maybe?
Start with your activities:
lay out about 5-7 big yellow stickies in a row in the center, and
write down what you will do – action statements please

Harvest data
Split data pkgs up
Explore data & QA
Extract desired values
Do SCIENCE! & math

Fine to be very general about activities. The point is to note that you’ll do them! Also fine to end your workflow at a meaningful breakpoint.
Then think about order, location, etc. Reorder them as necessary. Write down any data sources or other errata that would be helpful context.
Harvest data

resources that are **made**

+ Split data pkgs up
+ Explore data & QA

Each activity note makes a column

resources that are **used**.

We’ll do the resources **used first**.
Think **first about the data resources** you’ll be using for each activity, and place a **small yellow sticky** in the associated column naming either the data source or the data file used in the process.

**Harvest data**

**Split data pkgs up**

**Explore data & QA**

**Extract desired values**

**Do SCIENCE! & math**

**OAI-PMH datastore**

**Data pkgs from **

**Split data files**

**Split data files**

**My clean data????**

You might be **unsure about** the resource or **there might not be a resource**
Second, use a **small pink sticky** note to note the **tool you use**. Examples might be a database system, a script you have, a module, or a software package.

<table>
<thead>
<tr>
<th>Harvest data</th>
<th>Split data pkgs up</th>
<th>Explore data &amp; QA</th>
<th>Extract desired values</th>
<th>Do SCIENCE! &amp; math</th>
</tr>
</thead>
</table>
| OAI-PMH datastore | Data pkgs from ➜ | Split data files | Split data files | My clean data???
| scrape.py | Split.py | lxml | lxml | pandas | pandas | R?? |
Harvest data
Split data pkgs up
Explore data & QA
Extract desired values
Do SCIENCE! & math
OAI-PMH datastore
Data pkgs from_split data files
Split data files
My clean data???
split data files
pandas
pandas
R??

Use as many as you need. Okay to repeat!
XML chunk files

Access metadata

Note the data products that you’ll be making

Use another color to distinguish another kind of data type or purpose (e.g. if that data will go to another human)

Harvest data

Split data pkgs up

Explore data & QA

Extract desired values

Do SCIENCE! & math

OAI-PMH datastore

Data pkgs from

Split data files

Split data files

My clean data????

scrape.py

Split.py

pandas

pandas

R???
Indiv.
XML	
  files

Jupyter
notebook
Aggreg.
data file

Aggreg.
data file

Document-
tation

XML	
  chunk
files

Access
metadata

Harvest
data

Split data pkgs up

Explore
data & QA

Extract desired values

Do SCIENCE! & math

OAI-PMH
datastore

Data pkgs from

Split data
files

Split data
files

My clean
data????

scrape.py

Split.py

lxml

lxml

pandas

pandas

R??

notes

Make a note if you’re unsure
Start out very general if you need
Use the **red stickes** to note any **pain points or questions**

Then add who can help or answer your question.
Now take another look

• Are there deadlines you can trace back and add?

• Looking at the stuff that you are making:
  • What folders do you need?
  • Where should those folders be?
  • What should your file names be?

• Looking at the tools you use:
  • What documentation do you need about them to understand your project in a few years or for another person to take it up?
  • Do you need to save/backup the software or scripts to include as a reference in a future project?

• Add annotations to your board to indicate this. Use the back of your worksheet to document the folder structure.