Webinar:
Reproducibility Librarianship in Practice

SPEAKERS

Vicky Steeves

Librarian for Research Data Management and Reproducibility

New York University

vicky.steeves@nyu.edu

Birgit Schmidt

Co-chair of LIBER's Research Data Management Working Group

Göttingen State and University Library

bschmidt@sub.uni-goettingen.de

Reproducibility Librarianship in Practice

Vicky Steeves
Librarian for Research Data Management & Reproducibility
New York University

Slides: goo.gl/7V3Fv8

A little bit about me first?

- Pronouns: she/hers
- 1st degree was in computer science, then librarianship!
- Openness is life (open source, OER, open access, etc)
- My job is dual appointment between NYU's Center for Data Science and Division of Libraries
- Want to empower researchers to carry out best practices in RDM and reproducibility

Basically, I want to work with researchers & librarians to make fully reproducible work preservable, discoverable, usable, and freely accessible

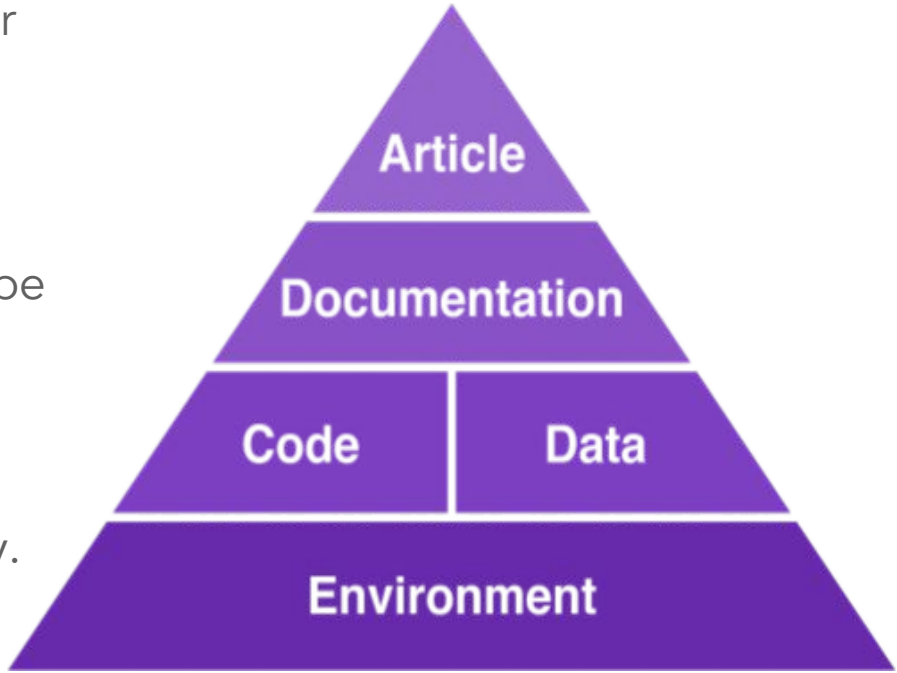
What problem am I trying to solve with my services?

Research isn't being efficiently managed or made reproducible

Much of the time, the workflow & processes aren't reproducible, the findings (data, code, etc.) aren't managed efficiently, and as a result, we all suffer.

Defining reproducibility on a spectrum

Reviewable Research: Sufficient detail for peer review & assessment.
Replicable Research: Tools are available to duplicate the author's results using their data.
Confirmable Research: Main conclusions can be attained independently without author's software.
Auditable Research: Process & tools archived such that it can be defended later if necessary.
Open/Reproducible Research: Auditable research made openly available.



Challenges in Reproducibility

Workload & Time Challenges

It is a time commitment to get data and code ready to share, and to share it

Otherwise known as…

the Incentive Problem

Reproducibility takes time, and is not always valued by the academic reward structure

"Insufficient time is the main reason why scientists do not make their data and experiment available and reproducible."
Carol Tenopir, Beyond the PDF2

"77% claim that they do not have time to document and clean up the code."
Victoria Stodden, Survey of the Machine Learning Community – NIPS 2010

Challenges in Reproducibility

Technical Obsolescence

Technology changes affect the reproducibility

Normative Dissonance[1]

Espoused values don't always match practice

Otherwise known as…

the Pipeline Problem

Reproducibility requires skills that are often not included in most curriculums!

"It would require huge amount of effort to make our code work with the latest versions of these tools." Collberg et al., Repeatability and Benefaction in Computer Systems Research, University of Arizona TR 14-04

[1] https://www.ncbi.nlm.nih.gov/pubmed/19385804

Even if runnable, results may differ...

The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements

We investigated the effects of data processing variables such as FreeSurfer version (v4.3.1, v4.5.0, and v5.0.0), workstation (Macintosh and Hewlett-Packard), and Macintosh operating system version (OSX 10.5 and OSX 10.6). **Significant differences** were revealed between **FreeSurfer version v5.0.0 and the two earlier versions**. [...] About a factor two smaller differences were detected between **Macintosh and Hewlett-Packard workstations** and between **OSX 10.5 and OSX 10.6**

My usual day-to-day work

Education

- Regular library classes
- Embedded classes in for-credit courses
  - E.g. Repro-R for data science class
- Special workshops
  - E.g. RCR
- Openly available & licensed materials
- guides.nyu.edu/data-management

Outreach

- Meet with everyone in the library
  - Liaison librarians
  - Digital archivists
  - IT
- Researchers
  - 1:1
  - group/lab

Tools/Infrastructure

- ReproZip, FLOSS tool for full reproducibility
- Taguette, FLOSS qual analysis tool
- Collaborations with software engineers in CDS
- Teaching using only OSS tools, use only open infra, etc.

Education

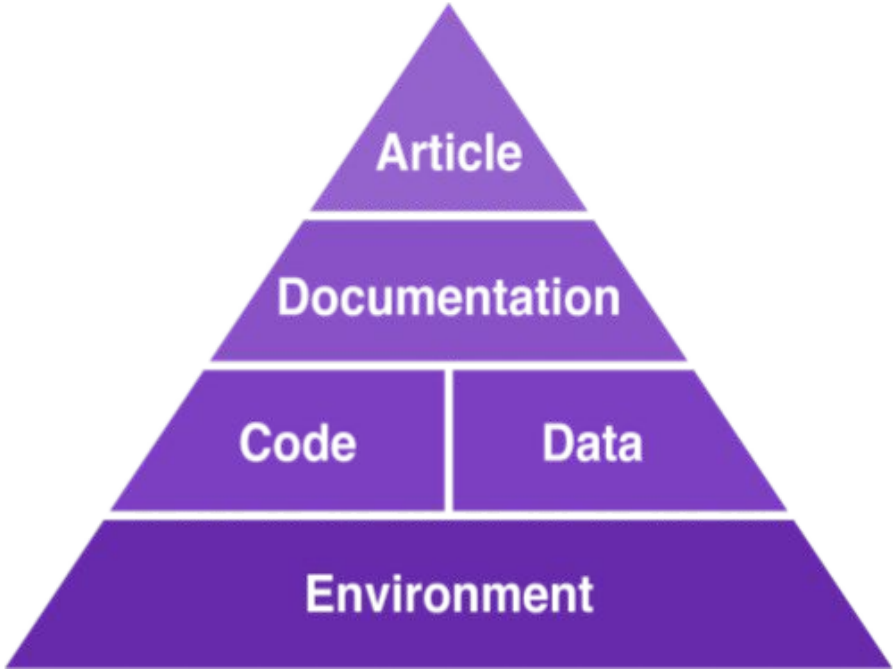
Teaching the folks about reproducibility & RDM!

- 3 offices for consultations:
  - Center for Data Science
  - Bobst Library
  - NYU Tandon CS department
- Openly available/licensed teaching materials: github.com/nyu-dataservices & gitlab.com/VickySteeves
- Keep news at: Reproduciblescience.org & @ReproFeed

Lesson #1:
You can't have any sort of reproducibility without good **data** and **project management**.

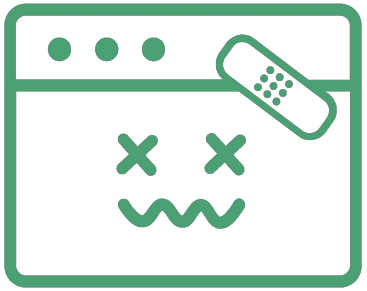
Where is my emphasis in teaching reproducibility?

- The first thing is teaching data management
  - Cannot have any kind of reproducibility without data management
- Reproducibility is on a spectrum (see pyramid next to me)
- Reproducibility requires a lot of skills that aren't generally taught in curricula -- so my classes emphasize starting small and scaling up



EXAMPLE SLIDE -- playfulness works!

Challenges in Reproducibility



- People make mistakes--and it impacts their research
- It's good to have other people check out your data and analyses--it's like having a copy editor for your data!

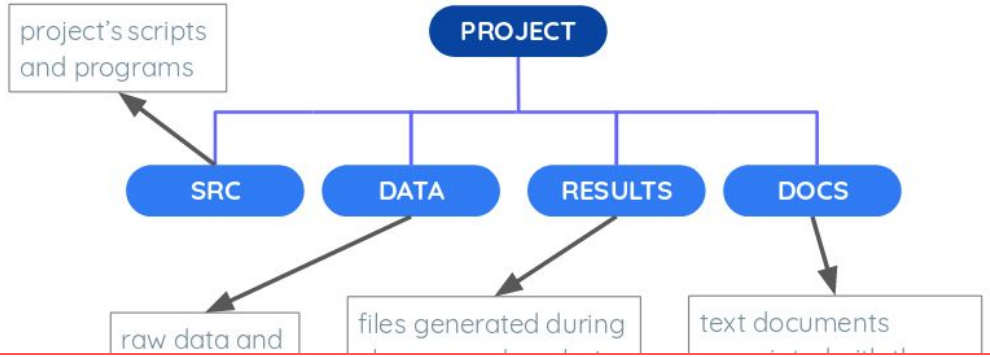
- It's *hard* to keep track of what version of what was used
- Software get updates, and these changes can disrupt reproducibility

EXAMPLE MATERIALS - literate programming

1.2 Project & data management

Some basic tenants of good project etiquette acrosss domains:

- Put each project in its own directory, which is named after the project.
- Put text documents associated with the project in the `doc` folder.
- Put raw data and metadata in the `data` folder. These data are **read-only**!
- Files generated during cleanup and analysis in a `results` folder.
- Put any code or scripts for the project in the `src` folder.
- Name all files to reflect their content or function, with NO special characters (!@#$%^*) or spaces! Use underscores or dashes, A-Z, and numbers!



My favourite outcomes from teaching reproducibility

- My patrons …
  - Take something immediately from the lesson and apply it in their daily work
    - Usually, it's storage related (e.g. 3+1 rule), but sometimes it's things like switching to literate programming! Or using version control!
  - Ask me questions after the fact to keep growing their skills
  - Read a little more interdisciplinary work around RDM + repro

Outreach

Making sure people know these services exist & we can help them!

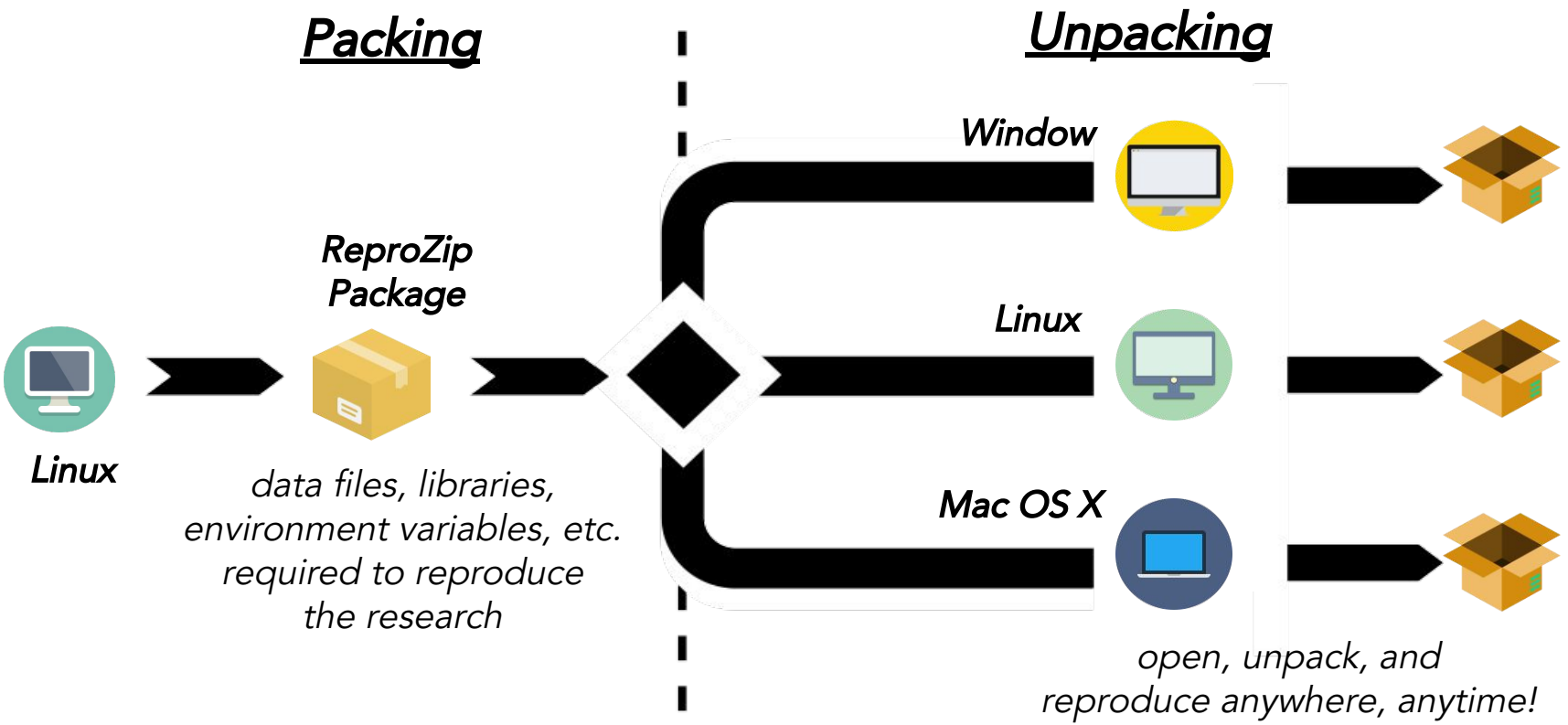
- Met with every liaison librarian
  - Learn about their depts' research & they learn about repro/RDM services
- Met with researchers 1:1 or in labs
  - Learn about their research & they learn about repro/RDM services
- Reproducibility Symposium: researchers showed tools and workflow on how working reproducibility has helped them

Tools & Infrastructure

Building new ways for researchers

- ReproZip: OSS tool for full reproducibility
  - Build open user-facing materials
  - Creates easily shareable, citable object
- OSF for Institutions @ NYU
  - 40 + public projects & 90 + private projects
  - Easier to propagate best RDM practices
- Taguette: FLOSS qualitative data analysis

ReproZip The Reproducibility Packer!



OSF for Institutions @ NYU

Login to the Open Science Framework with NYU credentials, associate projects with NYU and get on the landing page: https://osf.nyu.edu

Seamless for users to get access to NYU resources (e.g. Google Drive)

TAGUETTE 

- https://taguette.fr && https://gitlab.com/remram44/taguette
- Like “baguette” but with a “t” -- a play on the phrase “tag it!”
- Free & open source qualitative analysis tool -- qual software needs to be more accessible!

With TAGUETTE you can:

- Import a lot of different text formats (.md, .pdf, .txt, .docx, .rtf, .epub, .mobi, +)
- Highlight words, sentences, or paragraphs and tag them with the codes *you* create -- do your qualitative coding
- Export tagged documents, highlights for a specific tag, highlights for all tags, and a list of tags with their descriptions.

What generally works

- Horror stories
  - It really does motivate...
- Weirdly, all the stuff I learned in library school… (not weird)
  - Open file formats!
  - Documentation!
  - Backups!
- Hope for the future
  - Give them small steps/wins and ramp up

What generally doesn't

- Jargon (of course)
  - I give them definitions up front in plain-language
- Starting very technical
  - I do a lot with computational repro, so this can be hard...
- Value judgements -- reproducibility takes time + effort and isn't valued by T&P

Data Management + Reproducibility + Libraries

Emerging field of data librarianship has sparked changes in what librarians are offering:

- Data management/data services
  - Data/information literacy
- Grants support (data management plan review)
- Repository services, especially data repositories
  - Publishing and licensing support

Supporting reproducible research practices is a natural next step for our services supporting research -- I hope to see many librarians developing in tandem with engineers, humanists, and scientists (social/life/physical)!

Some light reading --

- Steeves, Vicky. Reproducibility Librarianship
- Sayre, Franklin and Riegelman, Amy. The Reproducibility Crisis and Academic Libraries
- Vitale, Cynthia R.H. Is Research Reproducibility the New Data Management for Libraries?
- Barba-group reproducibility syllabus: https://hackernoon.com/barba-group-reproducibility-syllabus-e3757ee635cf
- Ten Simple Rules for Reproducible Computational Research: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003285
- Five selfish reasons to work reproducibly: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-015-0850-7
- How scientists fool themselves – and how they can stop: https://www.nature.com/news/how-scientists-fool-themselves-and-how-they-can-stop-1.18517

Summary

- Generally, good data management leads to reproducibility (but doesn't guarantee it!), so a lot of reproducibility classes start there
  - *You* can start there!
- Work reproducibly in small ways & ramp up!
- Start with FLOSS tools & work up (if you want)!

Contact Info

Get the slides: goo.gl/7V3Fv8

Email me: vicky.steeves@nyu.edu

Tweet me: @VickySteeves

Toot me (mastodon): @vickysteeves@octodon.social

Teaching materials: gitlab.com/VickySteeves & github.com/nyu-dataservices

THANKS!

Questions?

Please put them in the chat box.

Slides and a recording will be sent to all registered delegates.