Publishing data and code

Oregon Data Science Collaborative

Spring 2022

Why publish data?

Why publish code?

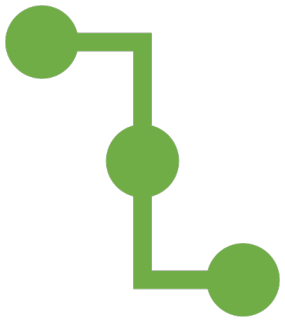
Carrot and stick motivations

- Open science values
- Improve reuse
- Data citations
- Software citations

- Funder requirement
- Journal requirement

Cultural change

Some barriers to data + code publication



- Time and resources needed for publication process
- Expectations are not clear
- Costs are personal, benefits are diffuse and general

- Fear of being scooped
- Fear that data and code are not perfect

Refresher on data repositories

Landscape of data repositories

Collections of repositories

DataOne

Re3

Project specific

UN's IPCC data center

Institutional

UO Scholars Bank

General

Zenodo

Figshare

Dryad

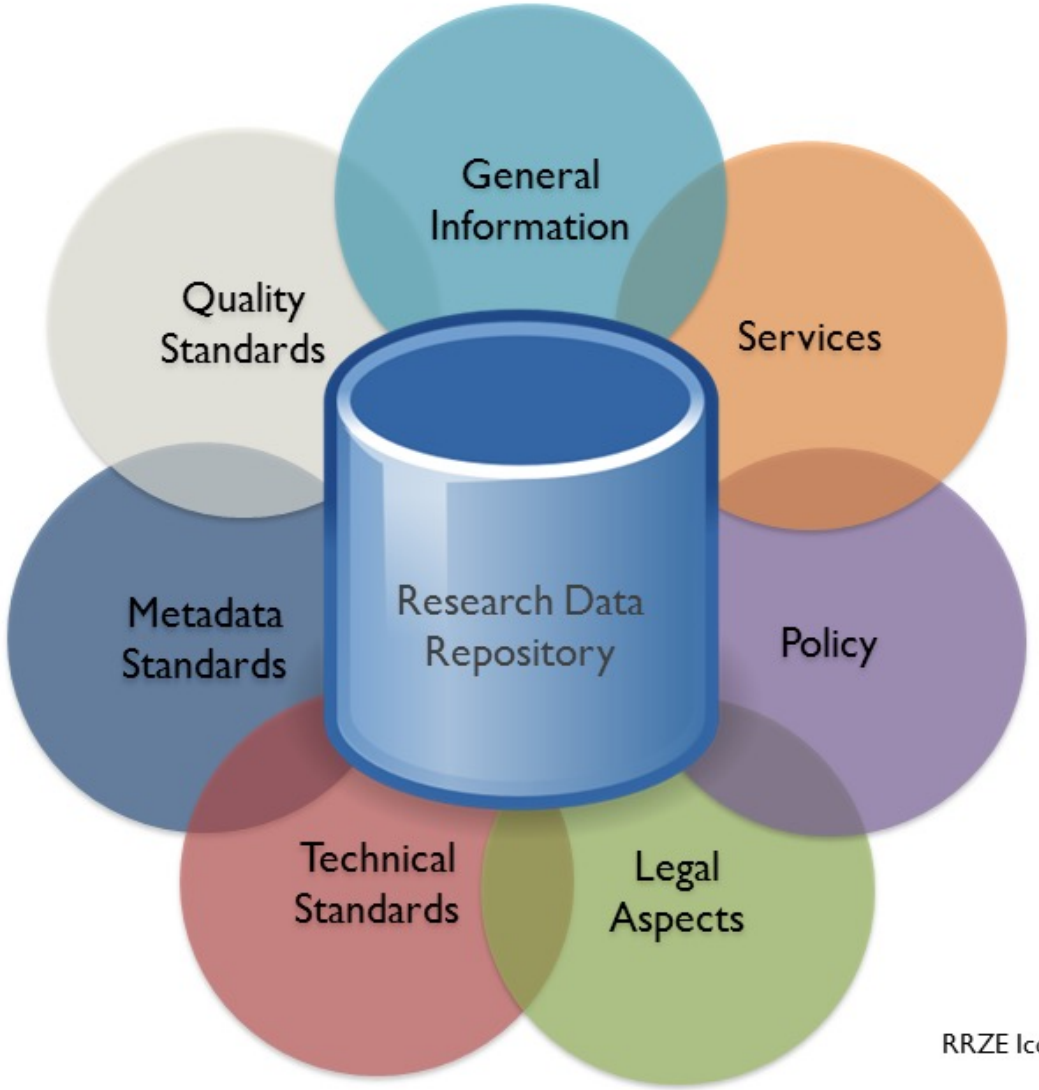
Discipline-specific

Environmental Data Initiative

Funder-specific

NSF's BCO-DMO

Different repositories have different features



RRZE Icon Set (CC: BY-SA)

Some data repos house specific products

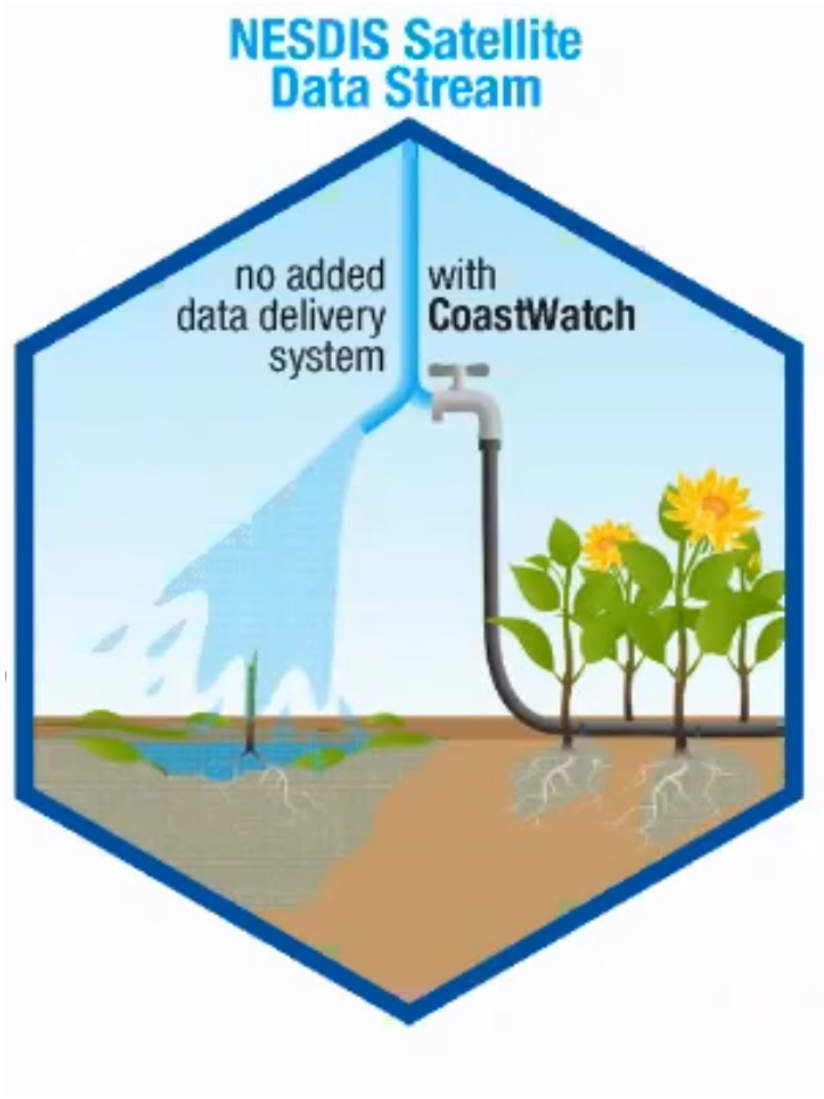
- Datasets attached to a particular paper or project
- Protocols and other related material can also be housed in repositories

Open Science Framework

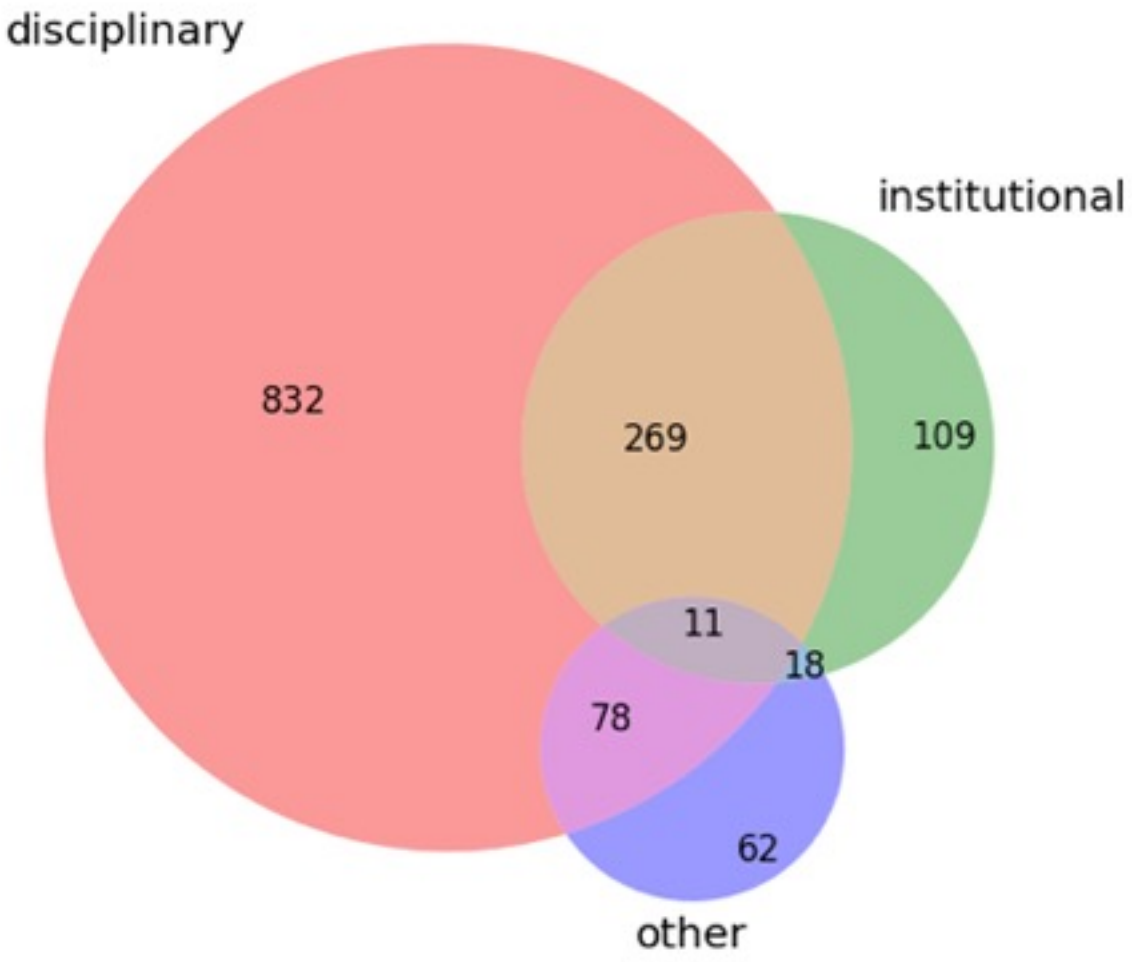
https://osf.io/

Some data repos provide data streams

- Especially pertinent for big data
- Repositories can manage, curate, and process raw data to provide the most useful outputs to users
- Example – [CoastWatch Data Portal]



Repositories serve multiple functions



The Landscape of Research Data Repositories in 2015

Exercise - Finding datasets from a publication

Ancestry-inclusive dog genomics challenges popular breed stereotypes

Publishing in data repos

Landscape of data repositories

Project specific

Collections of repositories

Institutional

General

Discipline-specific

Funder-specific

Landscape of data repositories

Cost?
Who pays?

Accessibility

Longevity

Curation and management

Process – integration with peer review?
Timeline to publication

License

File size,
data types

Where to start?

- Disciplinary and community standards
  - Scholarly associations often publish journals – those journals may have data and code guidelines that represent the discipline
- Funder guidelines
  - NSF Data Policy
  - NIH Data Policy
- Data management plan
- Data availability statements in prior similar publications

Exercise – data sharing guidelines

I. Disciplinary standards in theory
   - Look up a scholarly society in your discipline that publishes journals
   - What are the guidelines for publishing data and code?
   - How easy is it to find that information?

II. Disciplinary standards in practice
   - Look up a recent publication from one of the journals you identified above
   - Does the publication have data and code available?
   - What repository holds the data and/or code?
   - How easy is it to access the data?
   - What kind of metadata is available to help interpret the data?

Open science journals have specific policies

Public Library of Open Science

- Authors must share a minimal data set
- Currently, collaborating with Dryad, FigShare, and OSF to improve

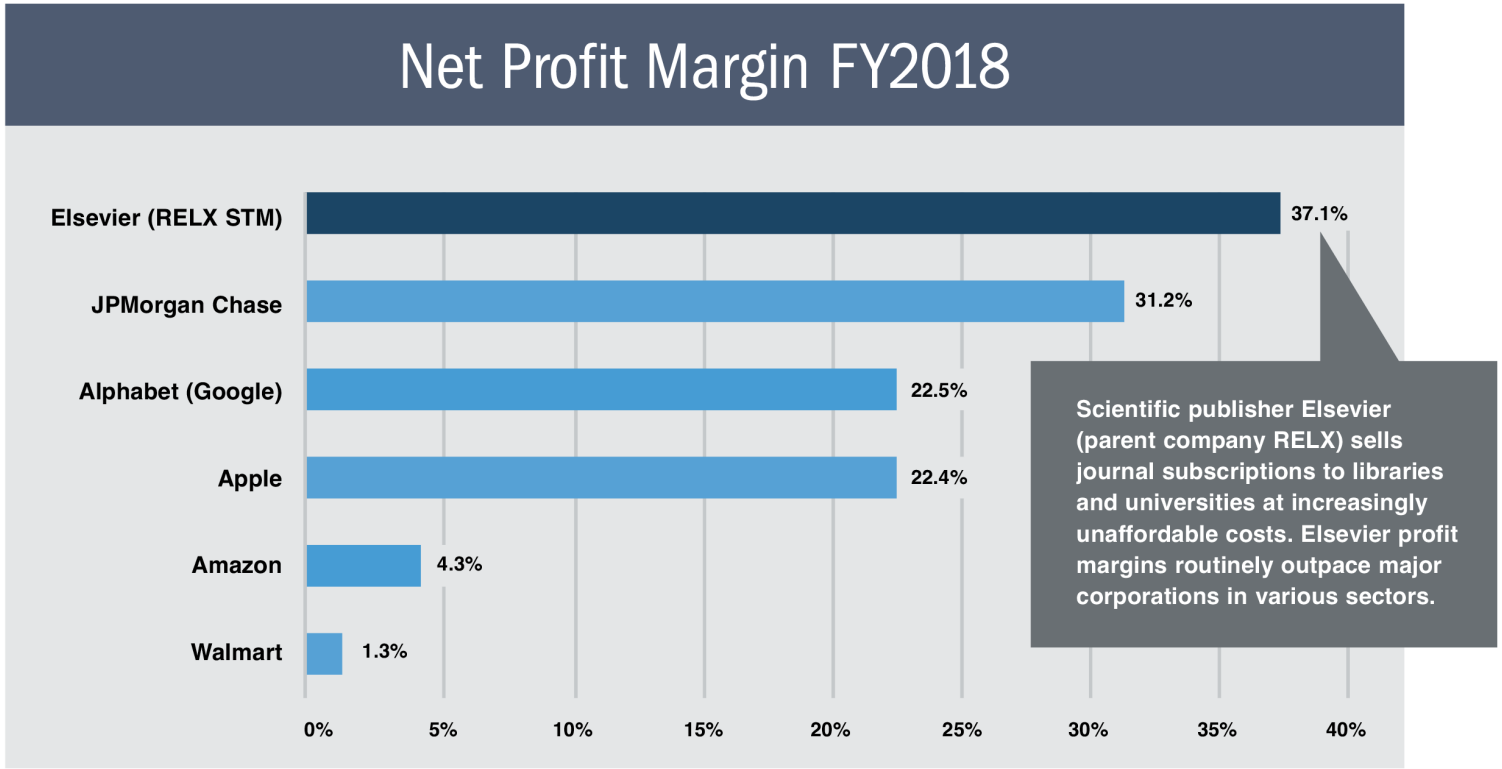
Open science vs. open access

- A new model for scientific publishing
- Free to read AND free to re-use
- Does NOT mean free to publish – OA charges can be prohibitive
- Many traditional publishers now have open access options (for journals or articles)
- OA necessary but not sufficient for open science – may actually make it harder for some scientists to publish

Explainer on Open Access from PhD Comics

A sidenote on academic publishing

- Academic publishing is wildly profitable for publishers and wildly expensive for authors and universities
- A great deep dive from The Guardian for more info



Calculated as (Net Profit/Total Revenue) X 100, using data in Elsevier's 2018 annual report and Standard & Poor's NetAdvantage database.

Source: UNC Libraries

Licenses for data and code

- Important to include a license when you publish so that others know how they can use and cite your work
- Repositories may have a default license, usually you can modify
- Creative Commons license options, e.g. CC-BY, CC-0
  - Guide to license options from Figshare
- Open source licenses for software, e.g. MIT, BSD 3-Clause
  - Some guidelines from The Open Source Guides (curated by GitHub)
  - Add a license to GitHub projects, especially if you will archive the repo for sharing

Citing datasets

- Citations for data and code are part of the open science ecosystem
  - Cite your published data DOI in your paper
- Most repositories will have a recommended citation format, journals and citation styles also have particular formats
- Software packages can also be cited

Project data vs. publication data

- The same larger dataset may serve multiple publications or projects
- Sometimes you want to publish specific analyses vs a general database
- No one size fits all data repo – but you don’t want to publish the same data in multiple places

What about GitHub?

GitHub repo ≠ data repository

- GitHub is a commercial product that can disappear at any time
- HOWEVER, you can archive your GitHub repo with Zenodo and receive a DOI that is a persistent identifier
- Guidelines for linking your GitHub repo to Zenodo

Versioning data and code

- Many (most?) repositories support versioning so that you can update a dataset over time
- You can note in the metadata if the dataset is expected to change or if this is the final version
- Can also embargo data prior to publication of a manuscript, etc.

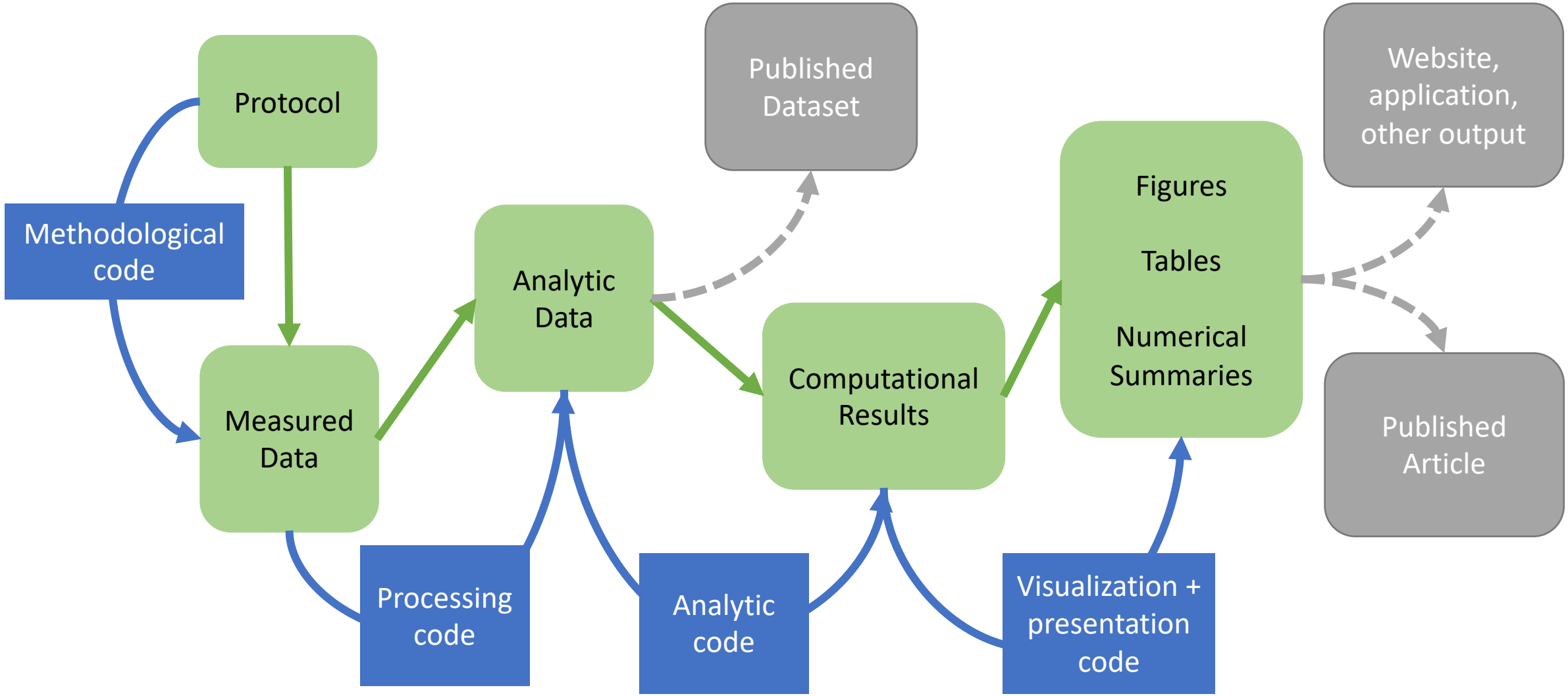
Fitting data publication into your workflow

Data management plans

An ounce of prevention is worth a pound of cure

- Reproducible research skills should facilitate data publication
  - Project organization + workflow
  - Version control
  - Metadata
- Data management plans formalize many of these elements *early in the research process*
- Guide from UO has tips for writing a DMP – doesn't have to be formal

Exercise - where does data publication fit?



Adapted from R. Peng