Data Management Tips and Tricks for Geographers

—

Hannah Gunderman, Research Data Management Consultant
Carnegie Mellon University Libraries

Who am I?

B.S., M.A., Ph.D. Geography
Biogeography, cultural geography, GIS
Library and information sciences postdoc
Research data management for all of CMU

Tell me about yourselves: what area of geography is your specialty? Faculty, staff, student, industry?

Research Data in Geography

Any data used in the creation, support or validation of research findings and publications.

Basic types:

- Observational data
- Experimental data
- Simulation data
- Derived or compiled data

Quantitative or qualitative

All subdisciplines of geography! Including cultural

Intermediate results and pre-processed data also important

Research Data Management

RDM is the process of creating organized, documented, accessible, and reusable research data

Helps with research organization and comprehensibility

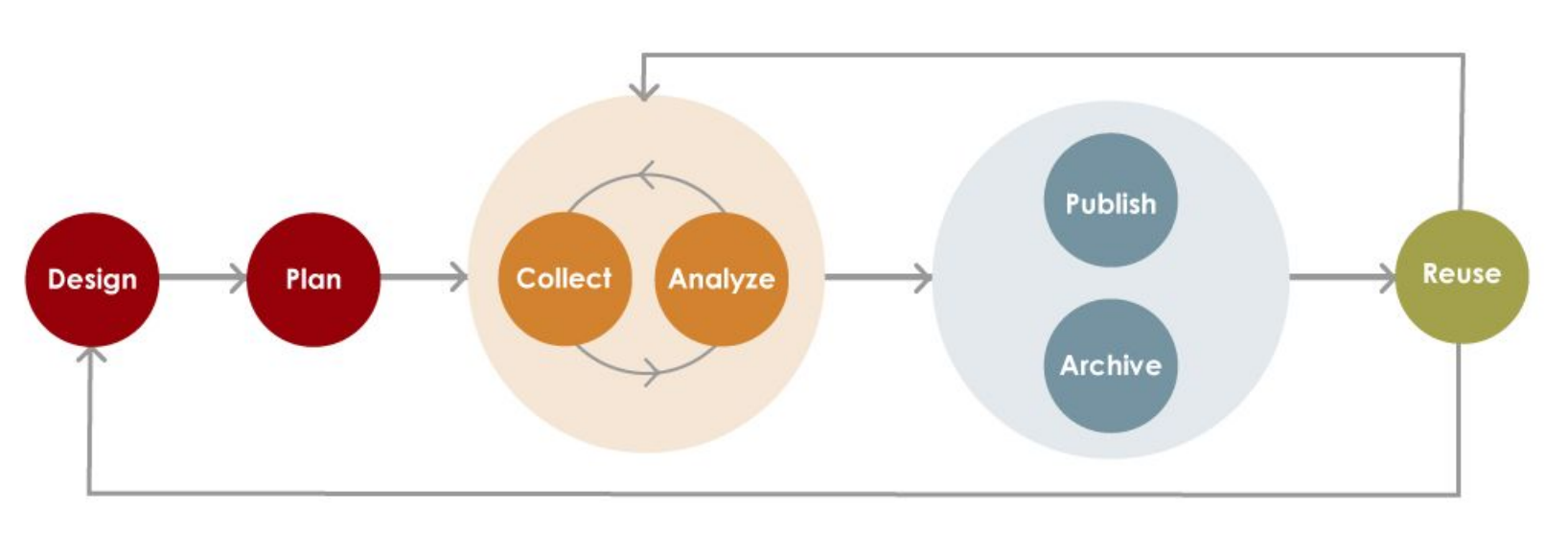
Communicating results to colleagues and others

Improved discovery and access

Meets funder and publisher mandates

Makes it easier to share your research with others - collaborators and public access

RDM Across the Research Lifecycle



Workshop Structure

Sources of data across geographic sub-disciplines: audience survey

Recommended practices for data management in following stages of research:

- Planning
- Collecting data
- Data storage and publishing

Last portion of workshop: open tool exploration

What are your data?

Data in Human Geography

Interview transcripts

Census data

Photographs

Landmarks

Books

Fieldwalk notes

Data in Geographic Information Science

Shapefiles

Code

Coordinates

Addresses

Algorithms

Maps

Data in Physical Geography

Fieldwork data

Laboratory observations

Collected data from instruments

Natural environment: trees, rivers, clouds, etc.

Open data/reused data

Surveys

Planning Stage

Make a data management plan (DMP)!

- What data will be generated in this project?
- What formats will be used (Excel, MySQL, jpg, etc.)?
- What information about the data will need to be captured so that others can understand it?
- How should the data be organized and named?
- How will the data be published or archived at the end of the project?

A DMP is a living document!

Research rarely goes exactly how we think it will

Incorporate data “quality checks” into your workflow

Try to stick with it!

DMPTool

https://dmptool.org/

Data Collection and Analysis Stage

3-2-1 technique:

- 3 copies of your data
- 2 different formats (e.g. laptop, external hard drive)
- 1 off-site back-up or in the cloud (e.g. Google Drive, Box, etc.)

File Naming Conventions (FNC)

Create your FNC by identifying key elements of the project, e.g. date of creation, author's name, project name, or section

Have a code book or data dictionary (Evernote is a good option for human/cultural geographers)

Have a readme file that lists all files and any file hierarchy

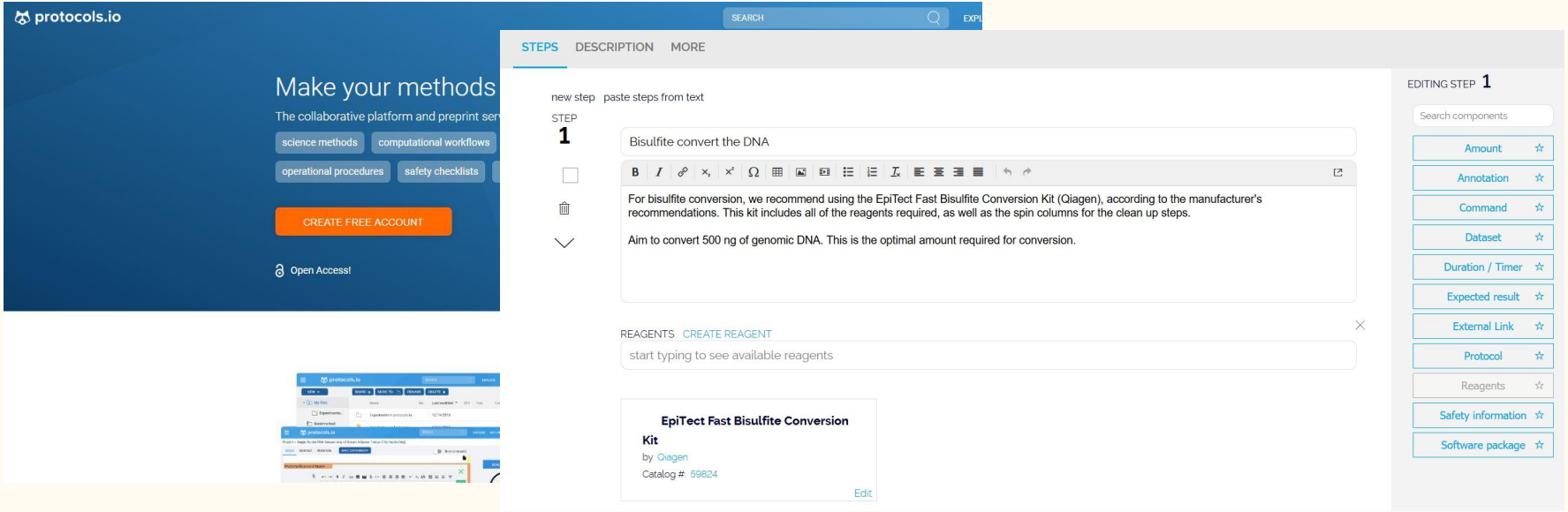
Documentation

Documenting the dataset allows others to understand and reproduce your work.

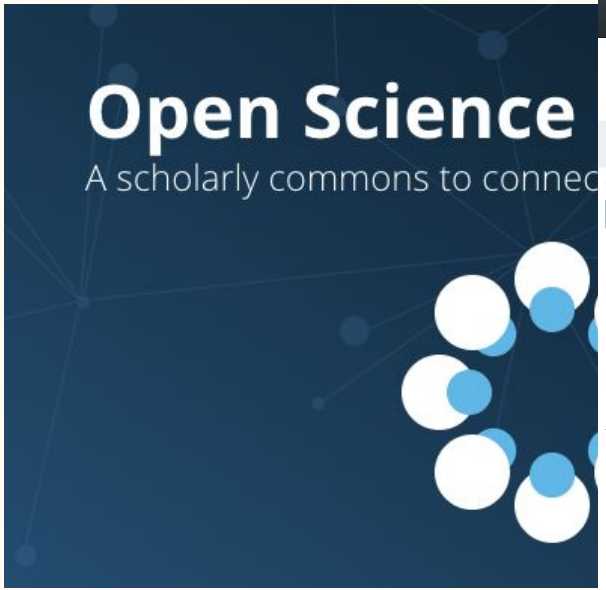
Documentation should include your step-by-step processes (what you did) and place the work in a larger context (why you did it).

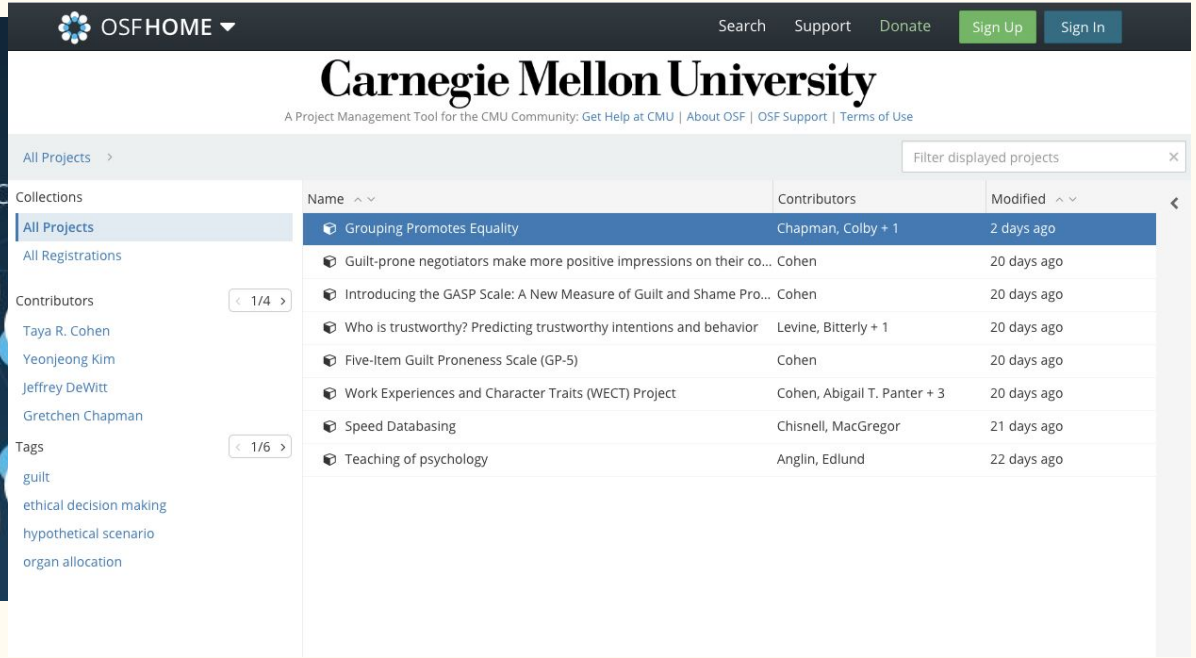
Electronic Lab Notebooks (ELNs) are a great way to do this (for any discipline), such as LabArchives

Protocols.io



Open Science Framework

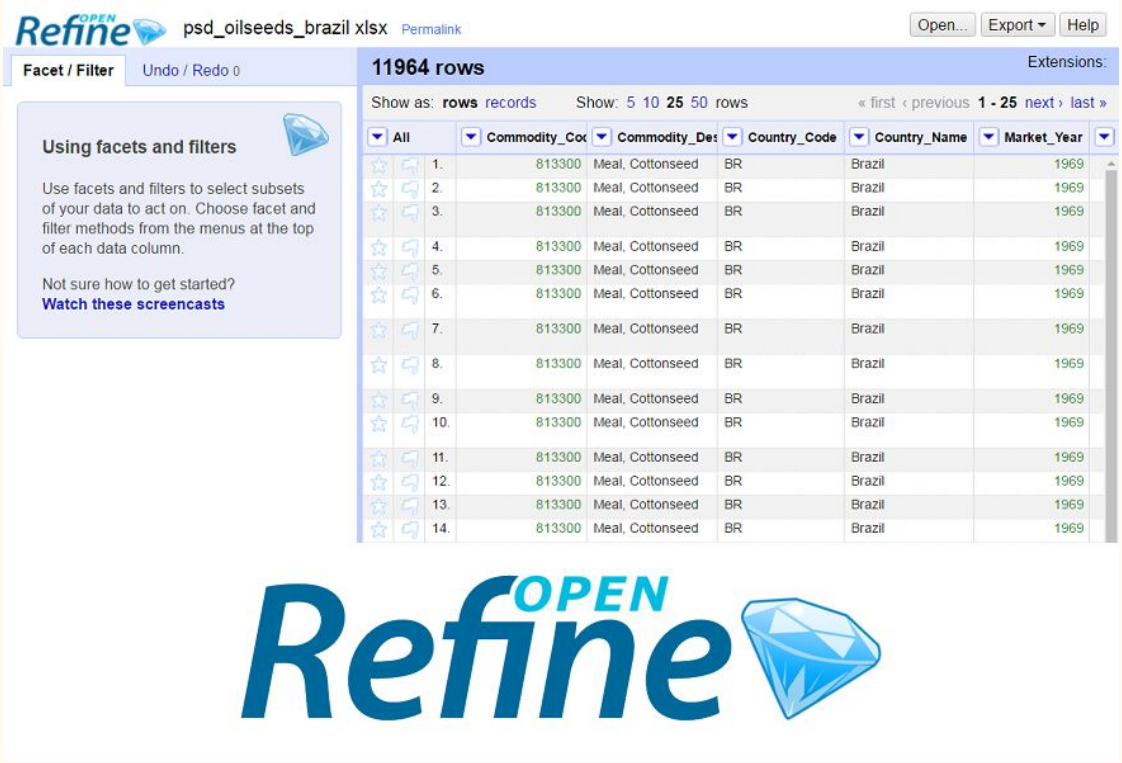




Geospatial data in spreadsheets

ID	Name	SurMat	lat	long
402	Elm Street	asphalt	38.8951	-77.0364
10	Main street	ashpalt	38.8956	-77.0983
343	Elm street	concrete	38.8951	-77.0364
563	Route 47	dirt	38.9355	-74.2944
103	route 47	Concrete	389.355	74.2945

Consistent Data Quality Checks



Metadata - data about your data

Metadata is data that describes a dataset:

What is the data?

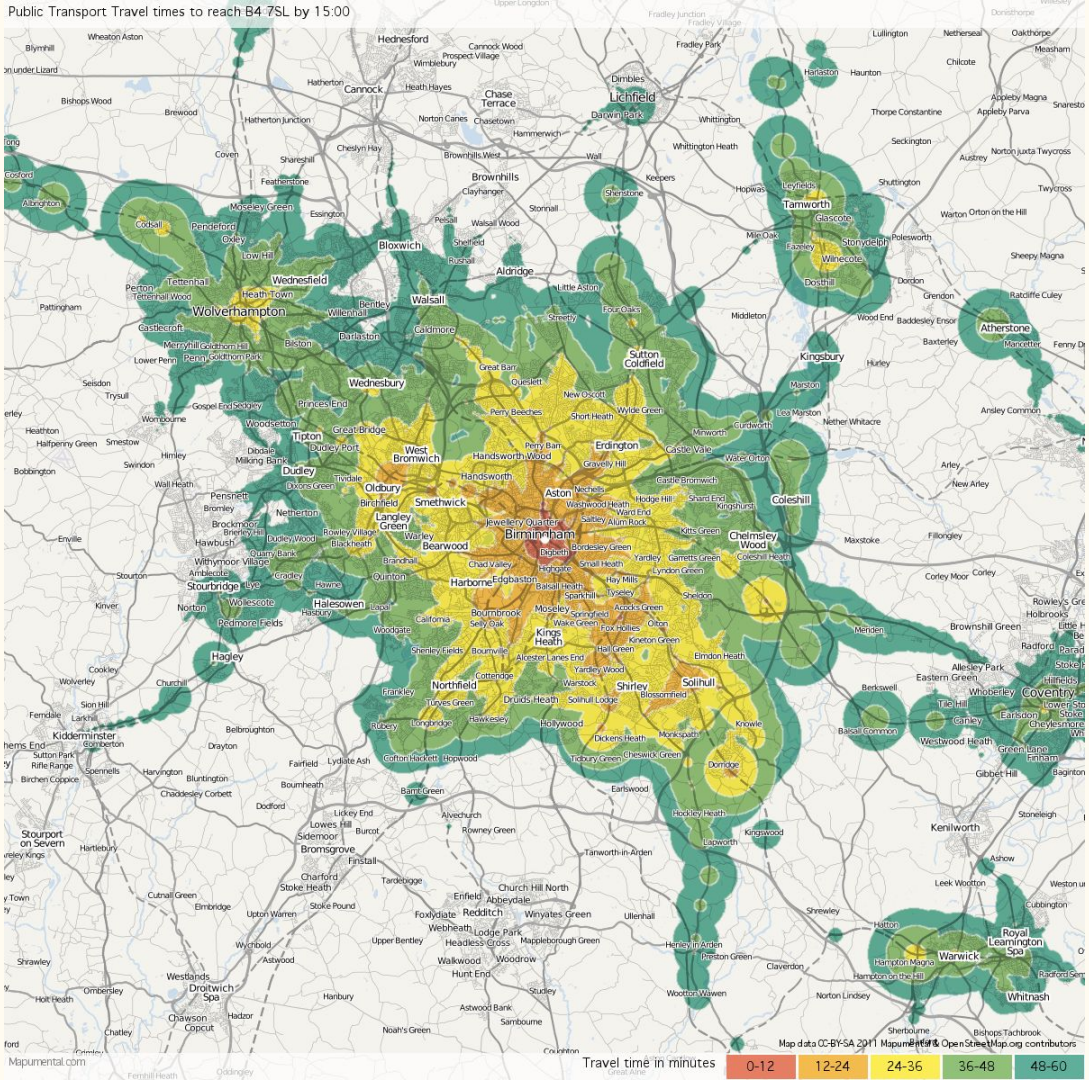
Who created it?

How may it be used?

What generated it?

Most repositories require some basic metadata record.

It is a good practice to build metadata into your collection and analysis workflow!



Open data? Crowdsourced data?

Coordinate system?

GIS software?

Time period?

Data rights?

Layer Metadata

Group layers by...

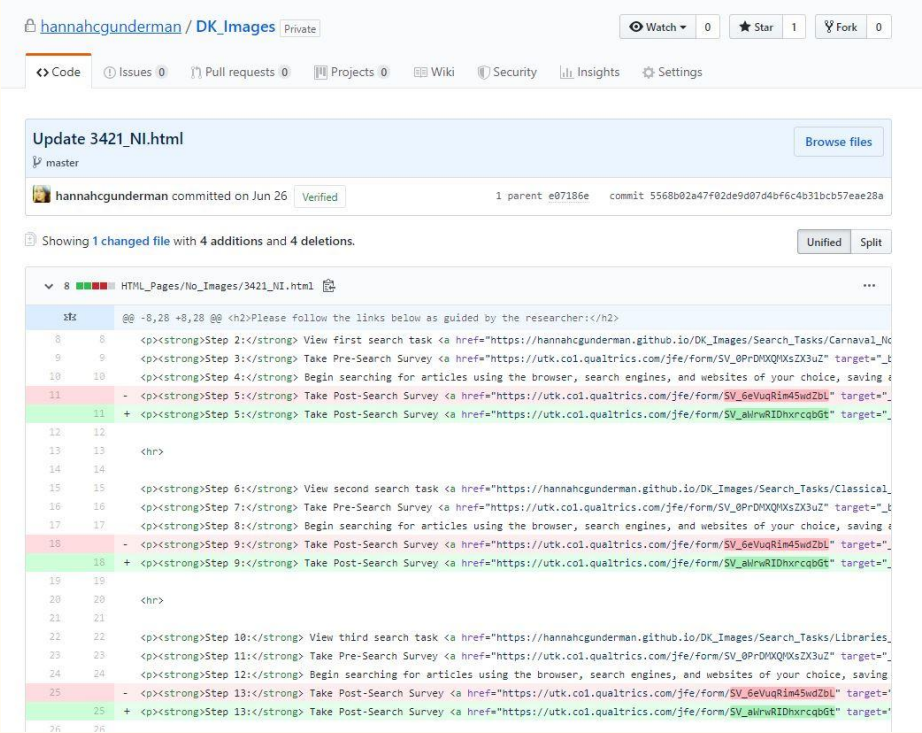
class

Layers | Settings | About

Group by	Value
Vector	
Airports	
provider	ogr
proj4	+proj=longlat +datum=WGS84 +no_defs
layername	Airports
id	Airports20160419171103194
geom	POINT
fields	ICAO String 4:0 NAME String 254:0 USE Integer 6:0 USE_DESC String 254:0 ELEVATION Integer 6:0
extent	-177,-55 : 179,83
crs	EPSG:4326
class	Vector
capabilities	Add Features, Delete Features, Change At...

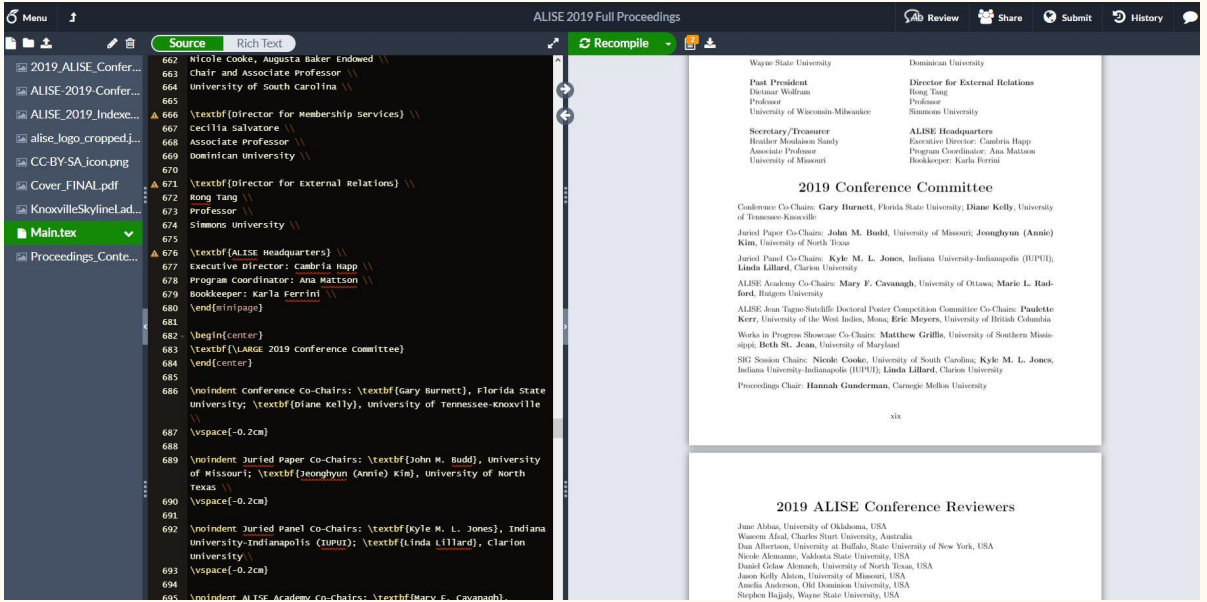
Version control (don’t assume you’ll remember what you did!

GitHub is a great platform for version control



Storage and Publication

Overleaf



Repositories: General vs Specific

Journals/funders may mandate deposits into certain repositories

Many things may influence your choice of repository type

Check with information science experts in your discipline! (librarians)

General Repositories

Discipline-specific repositories

Recommended Data Repositories from Scientific Data:
https://www.nature.com/sdata/policies/repositories

Depending on your discipline, your data may have greater visibility in discipline-specific repositories! Always check with experts in your field.

Remember your librarians!

Open Tool Exploration

DMPTool
Evernote
protocols.io
Open Science Framework
LabArchives
Overleaf
GitHub
OpenRefine