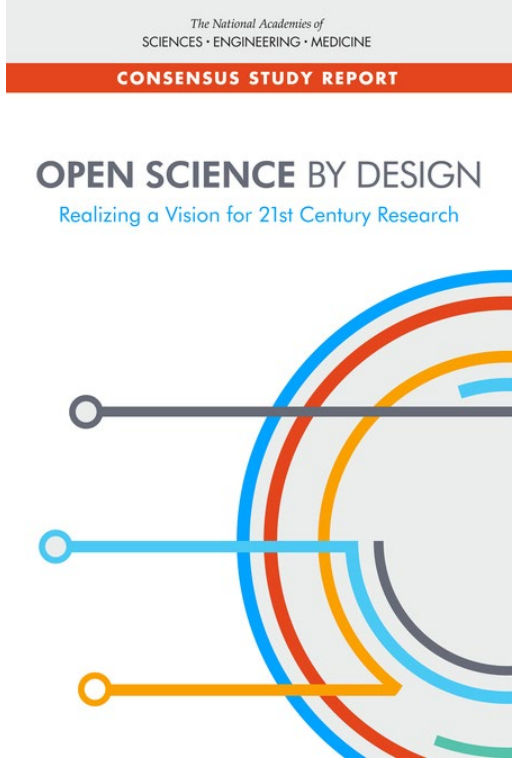
Mainstreaming **Metadata** into **Research Workflows** to advance **Reproducibility** and **Open Geographic Information Science**

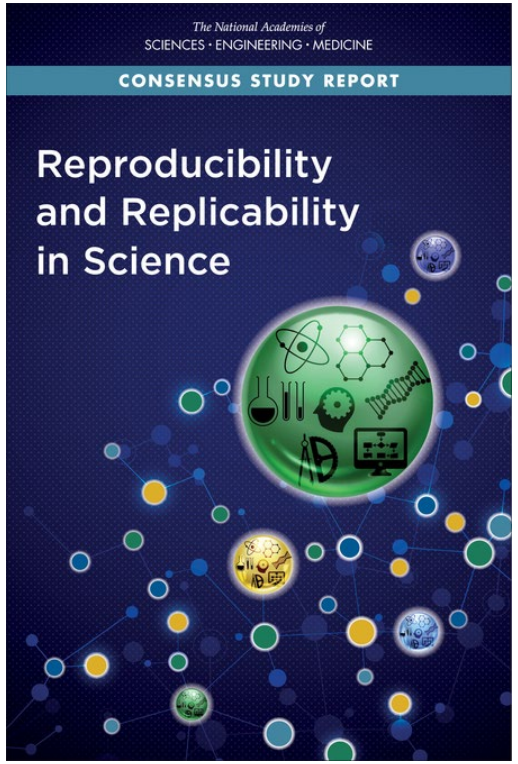
Joseph Holler (Middlebury College) and Peter Kedron (Arizona State University)

Conference: FOSS4G 2022, Firenze
Funding: National Science Foundation BCS-2049837
Publication: https://doi.org/10.5194/isprs-archives-XLVIII-4-W1-2022-201-2022
GitHub: github.com/HEGSRR/foss4g-2022-metadata
OSF: osf.io/52j8s/

Motivation for Metadata

- Enhance the reproducibility of geographic research
- Increase the pace and credibility of knowledge production in the geographic sciences
- Facilitate more efficient & open research life cycles

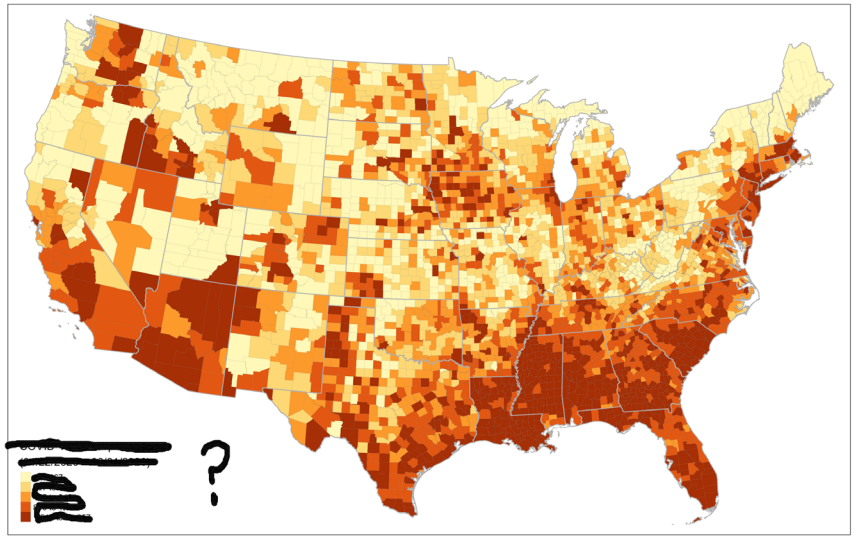




Context

I am familiar with "reproducibility", and my research is reproducible!

Metadata?
No, I have never used that…



7 Reproduction or Replication Studies

github.com/HEGSRR/

1. RPr-Chakrabory-2021
2. RPr-Malcomb-2014
3. RPr-Mollalo-2020
4. RPr-Vijayan-2020
5. RPr-Saffary-2020
6. RPr-Kang-2020
7. RPl-DiMaggio-2021

Three Points

1. Open Science and Reproducibility require standardized metadata.

2. Researchers use, create, and modify information about their research projects and research data throughout the research life cycle.

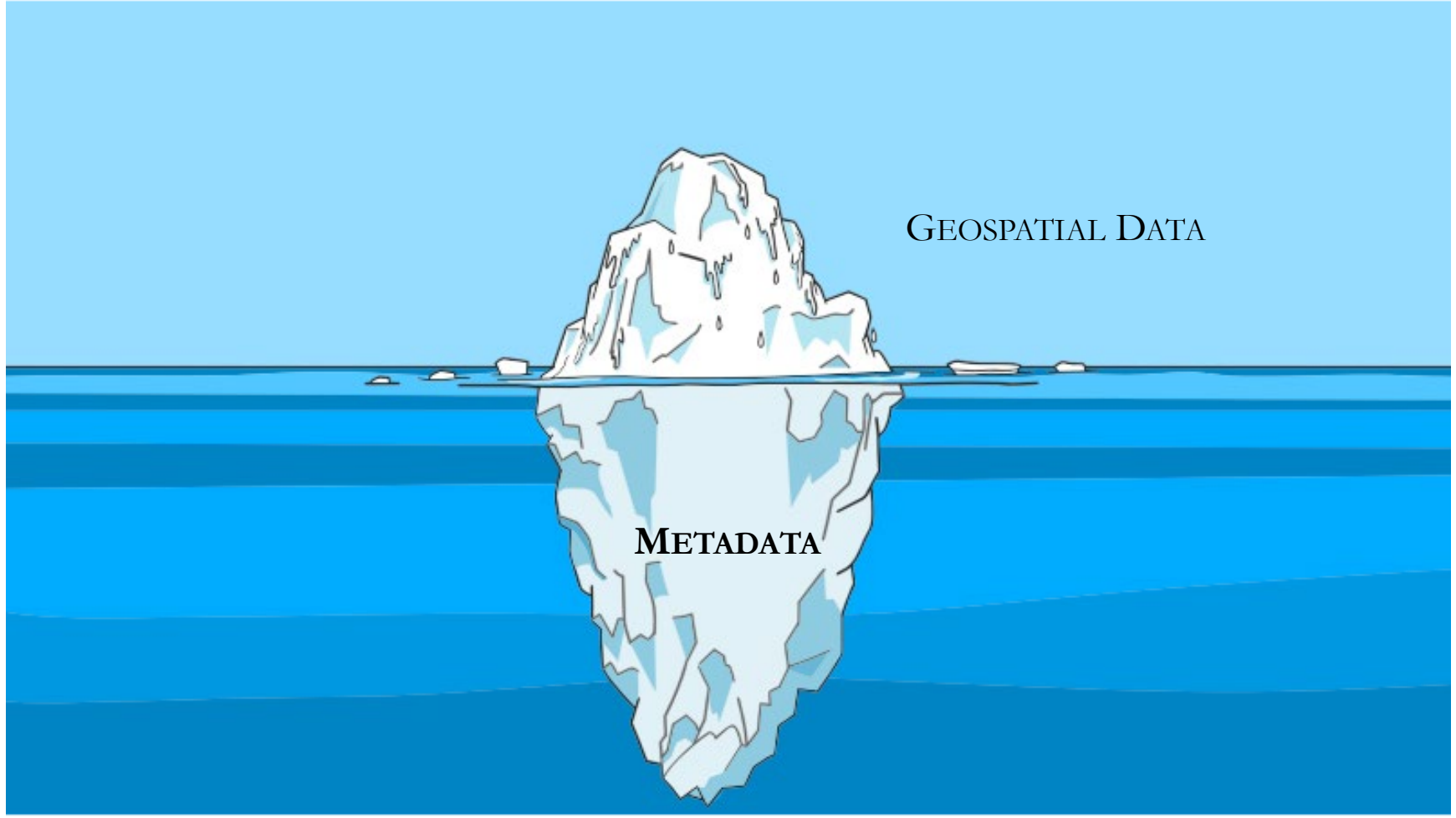
3. We need better open source geospatial software to support metadata-rich research

Reproducibility > repeating computations

	Same Methods	Varied Methods
Same Data	Reproduction (Verification)	Reanalysis
Different Data	(Direct) Replication	Extension

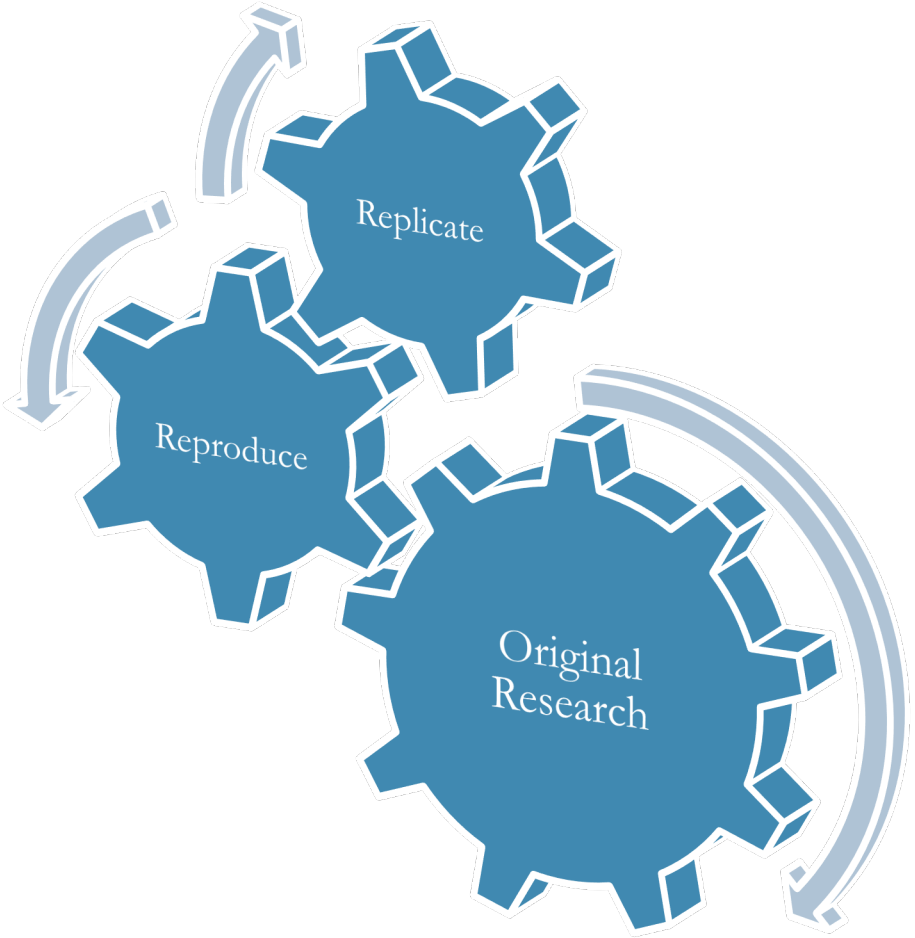
Christensen, Freese and Miguiel (2019)

Geospatial Metadata: information about spatial data



Metadata for Reproducibility & Open Science

- Social & ontological context (Schuurman and Leszcynski 2006, Comber et al 2008)
- Metadata is an ethical issue (Tullis and Kar 2021)
- FAIR open data (Wilkinson et al 2016)
  - F indable
  - A ccessible
  - I nteroperable
  - R eusable
- 5-Star Reproducibility (Wilson et al 2021)
  - ☆ data, code, and license
  - ☆ ☆ some metadata & provenance
  - ☆ ☆ ☆ complete & structured metadata and provenance
  - ☆ ☆ ☆ ☆ international standards for data and metadata
  - ☆ ☆ ☆ ☆ ☆ processing environment



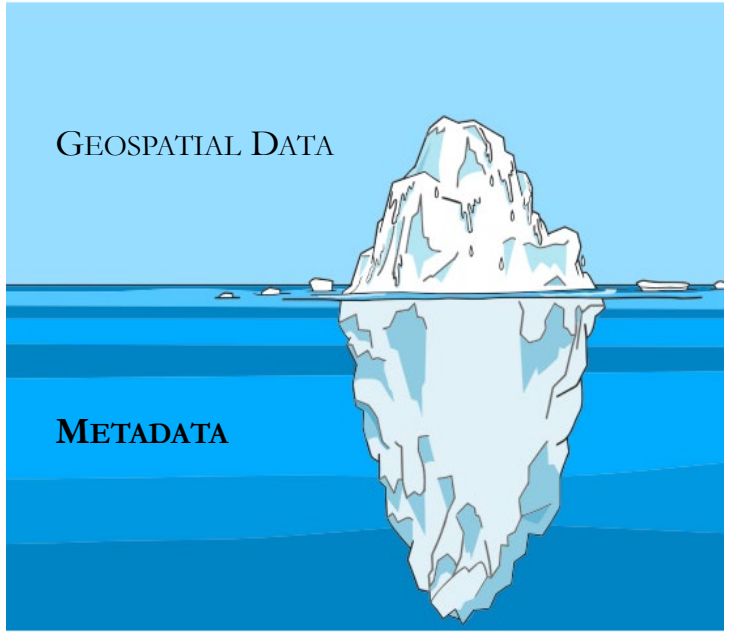
Standards

- Spatial Data Infrastructures
  - FGDC: Federal Geographic Data Committee
  - INSPIRE: Infrastructure for Spatial Information in Europe

- Standards Organizations
  - ISO: International Organization for Standardization
  - DCMI: Dublin Core Metadata Initiative
  - OGC: Open Geospatial Consortium

Recommended Standards

- ISO 19115 for geographic data
- Dublin Core for research projects



ISO 19115	Dublin Core
Dataset name	Title
Abstract, Purpose	Description
Topic Category	Subject Keywords
Unique Identifier	Identifier
Date	Date
Contact / Responsible Parties	---
Credit, Citation	Contributors, Creator, Publisher
Constraints	Rights
Distribution and Format	Type
Spatial Representation	Type
Extent (spatial & temporal)	Coverage
Spatial Resolution	
Temporal Resolution	
Content information (attributes, measurements)	
Data Quality, Usage	
Lineage	Source, Provenance

Three Points

1. **Open Science and Reproducibility require standardized metadata.**

2. Researchers use, create, and modify information about their research projects and research data throughout the research life cycle.

3. We need better open source geospatial software to support metadata-rich research

Open Science Research Life Cycle

Open Science Research Life Cycle

Provocation → Ideation → Knowledge Generation → Validation → Dissemination → Preservation

Literature Review → New Idea

Open Science Research Life Cycle

Provocation → Ideation → Knowledge Generation → Validation → Dissemination → Preservation

Research Planning & Prototyping

Open Science Research Life Cycle

Provocation → Ideation → Knowledge Generation → Validation → Dissemination → Preservation

Collect / Generate Data & **Metadata**

Open Science Research Life Cycle

Provocation → Ideation → Knowledge Generation → Validation → Dissemination → Preservation

Interpretation, Working Papers & Conferences

Open Science Research Life Cycle

Provocation → Ideation → Knowledge Generation → Validation → Dissemination → Preservation

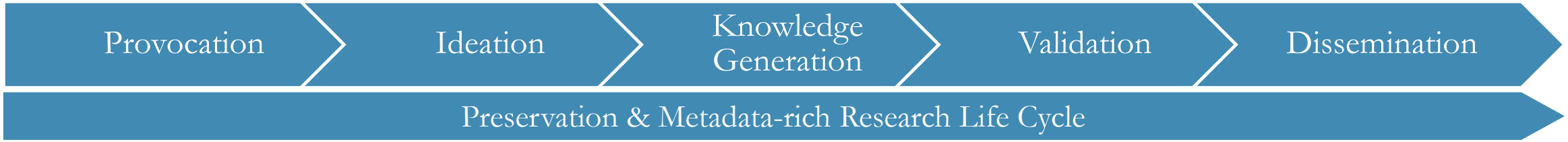
Peer Review, Publication

Open Science Research Life Cycle

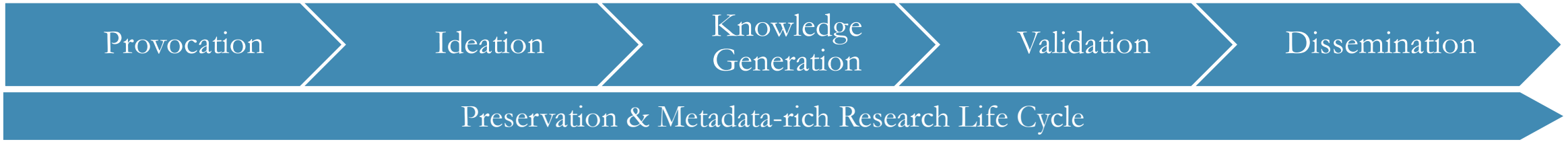
Provocation → Ideation → Knowledge Generation → Validation → Dissemination → Preservation

Public Archive with finalized **Metadata**

Open Science Research Life Cycle



Open Science Research Life Cycle



Research Compendium Template

Project with ***Metadata***

- \ Data
  - \ Raw
    - \ Public
    - \ Private
  - \ Derived
    - \ Public
    - \ Private
  - \ ***Metadata***
- \ Docs (Reports, Manuscript, Presentation)
- \ Procedures
  - \ Code (computational notebook)
  - \ Environment
  - \ Protocols
- \ Results (figures, maps, tables, model outputs)

Open Science Research Life Cycle

Provocation → Ideation → Knowledge Generation → Validation → Dissemination

Preservation & Metadata-rich Research Life Cycle

Research Compendium Template

Project with ***Metadata***

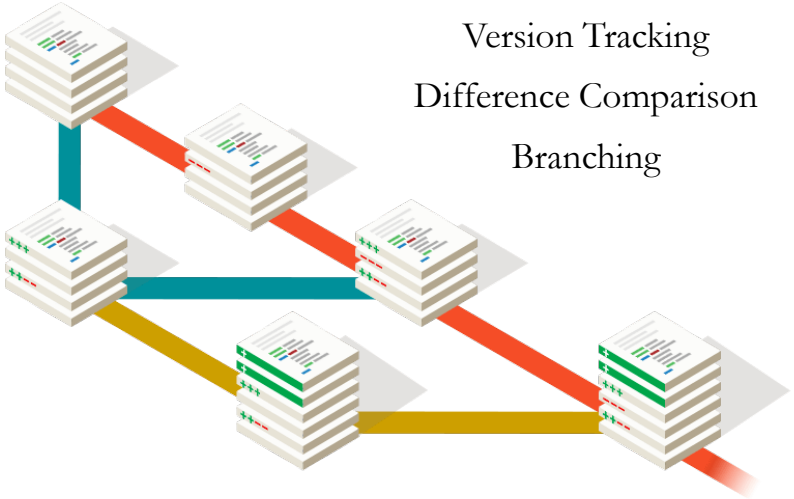
- \ Data
  - \ Raw
    - \ Public
    - \ Private
  - \ Derived
    - \ Public
    - \ Private
  - \ ***Metadata***
- \ Docs (Reports, Manuscript, Presentation)
- \ Procedures
  - \ Code (computational notebook)
  - \ Environment
  - \ Protocols
- \ Results (figures, maps, tables, model outputs)

Git Repository

Version Tracking

Difference Comparison

Branching



```
528  559  race_gee <- geeglm(
529       -  covid_rate ~ white_pct + black_pct + native_pct + asian_pct + other_pct,
     560  +  covid_rate ~ z_white_pct + z_black_pct + z_native_pct + z_asian_pct + z_other_pct,
530  561     data = gee_data, # data frame
531  562     id = id, # cluster IDs
532  563     family = Gamma(link = "log"),
     @@ -537,7 +568,7 @@ race_gee <- geeglm(
537  568  # coef() extracts coefficients table from the summary, same as $coefficients
```

Open Science Research Life Cycle

Provocation → Ideation → Knowledge Generation → Validation → Dissemination

Preservation & Metadata-rich Research Life Cycle

Research Compendium Template

Project with ***Metadata***

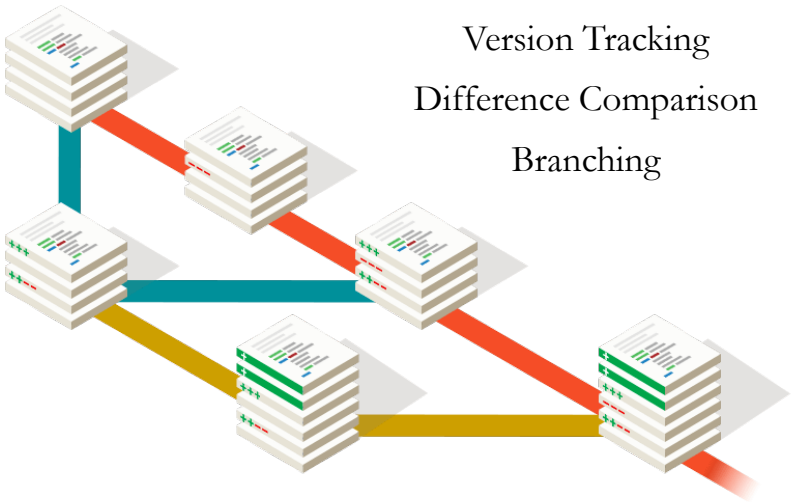
- \ Data
  - \ Raw
    - \ Public
    - \ Private
  - \ Derived
    - \ Public
    - \ Private
  - \ ***Metadata***
- \ Docs (Reports, Manuscript, Presentation)
- \ Procedures
  - \ Code (computational notebook)
  - \ Environment
  - \ Protocols
- \ Results (figures, maps, tables, model outputs)

Git Repository

Version Tracking

Difference Comparison

Branching



```
528  559  race_gee <- geeglm(
529       -  covid_rate ~ white_pct + black_pct + native_pct + asian_pct + other_pct,
     560  +  covid_rate ~ z_white_pct + z_black_pct + z_native_pct + z_asian_pct + z_other_pct,
530  561     data = gee_data, # data frame
531  562     id = id, # cluster IDs
532  563     family = Gamma(link = "log"),
     @@ -537,7 +568,7 @@ race_gee <- geeglm(
537  568  # coef() extracts coefficients table from the summary, same as $coefficients
```

OSF

DOI

Link to Git Repository

Register pre-analysis plan and final report

Research Compendium Template in Action

HEGSRR / RPr-Chakraborty-2021 Public

Not

generated from HEGSRR/HEGSRR-Template

Code | Issues | Pull requests | Actions | Projects | Security | Insights

main | 1 branch | 0 tags | Go to file | Code

josephholler finalized report 33013f9 27 days ago 192 commits

data	update state data management and covid map	last month
docs	finalized report	27 days ago
procedure	final report revisions	27 days ago
results	added tables to the results to compare the computational environments	last month
.gitignore	Update .gitignore	2 months ago
CITATION.cff	Update CITATION.cff	3 months ago
LICENSE	change to BSD-3 license	12 months ago
r_project.rproj	Initial commit	14 months ago
readme.md	mon morning updates	2 months ago
template_readme.md	Initial commit	14 months ago

Reproduction of Chakraborty 2021 analysis of unequal distribution of COVID-19 for people with disabilities

This study is a replication of:

> Chakraborty, J. 2021. Social inequities in the distribution of COVID-19: An intra-categorical analysis of people with disabilities in the U.S. *Disability and Health Journal* 14:1-5. DOI:10.1016/j.dhjo.2020.101007

Abstract

> The original paper is a national scale study of the relationship between COVID-19 incidence and disability characteristics (by demographic) in the United States. The paper aims to determine whether COVID-19 incidence is more significant in counties with larger proportions of socio-demographically disadvantaged people with disabilities, based on race, ethnicity, poverty status, and biological sex.

Authors

- Joseph Holler
- Drew An-Pham
- Peter Kedron
- Derrick Burt
- Junyi Zhou

Repository Documents

> Link your reports, manuscripts, presentations, publication DOIs, preregistrations, etc. here. Delete this instruction and unused list items from your final repository. Adjust the file names and paths and add additional items as necessary.

- OSF Project: https://doi.org/10.17605/OSF.IO/S5MTQ
- Preregistration: https://doi.org/10.17605/OSF.IO/MJXHD
- Publication: t.b.d.
- Pre-analysis plan: docs/report/preanalysis.pdf
- Study report: docs/report/report.pdf
- Computed R Markdown notebook: docs/report/01-RPr-Chakraborty.pdf or docs/report/01-RPr-Chakraborty.html

Repository Contents

The contents of this repository are outlined in three tables:

- Data: data/data_metadata.csv
- Procedures: procedure/procedure_metadata.csv
- Results: results/results_metadata.csv
- Processing Environment: procedure/environment/r_environment.txt

The template_readme.md file contains more information on structure and rationale of this research template repository, as well as important references and licenses.

Research Compendium Template in Action: FGDC XML

main RPr-Chakraborty-2021 / data / metadata / metadata_acs.xml Go to file ...

emilyzhou112 adding metadata files Latest commit 58855ac on Jun 17 History

1 contributor

228 lines (228 sloc) 9.96 KB Raw Blame

```xml
<?xml version='1.0' encoding='UTF-8'?>
<metadata>
  <idinfo>
    <citation>
      <citeinfo>
        <origin>US Census Bureau</origin>
        <pubdate>20191219</pubdate>
        <title>County Level Poverty and Disability Data, Derived from 2014-2018 American Community Survey 5-year Estimates</title>
        <geoform>vector digital data</geoform>
        <onlink>https://www.census.gov/</onlink>
        <lworkcit>
          <citeinfo>
            <origin>US Census Bureau</origin>
            <pubdate>20191219</pubdate>
            <title>2014-2018 American Community Survey 5-year Estimates</title>
            <geoform>vector digital data</geoform>
            <onlink>https://www.census.gov/</onlink>
          </citeinfo>
        </lworkcit>
      </citeinfo>
    </citation>
```

Research Compendium Template in Action: CSV Index

main RPr-Chakraborty-2021 / data / data_metadata.csv Go to file ...

emilyzhou112 revise metadata and lit review Latest commit 35ae126 on Jun 27 History

3 contributors

18 lines (18 sloc) 2.11 KB Raw Blame

Search this file...

	file_path	file_name	format	sources	metadata
2	raw\public	covidcase080120.gpkg	Geopackage	raw\public\chakraborty\Aug1data\Aug1data.shp	
3	raw\public	covidcase080120.csv	Comma-separated values	raw\public\chakraborty\Aug1data\Aug1data.shp	
4	raw\public	acs.gpkg	Geopackage	TidyCensus Package call to ACS 2018 5-year	metadata_acs.xml
5	raw\public	disability_raw.csv	Comma-separated values	ACS 2018 5-year	ACSST5Y2020.S1810_metadata.csv
6	raw\public	poverty_raw.csv	Comma-separated values	ACS 2018 5-year	ACSDT5Y2020.C18130_metadata.csv

Three Points

1. Open Science and Reproducibility require standardized metadata.
2. **Researchers use, create, and modify information about their research projects and research data throughout the research life cycle.**
3. We need better open source geospatial software to support metadata-rich research

Open Geographic Information Metadata Systems

Must have metadata editing functionality…

- 3 Desktop GIS: QGIS, GRASS, SAGA
- 2 Spatial Data Science Packages: R geometa, Python pygeometa
- 1 Catalogue: GeoNetwork
- 1 Content Management: GeoNode
- 2 Metadata Authoring: Metadata Wizard, mdEditor (mdeditor.org)
- 1 Executable Research Compendium Tools: o2r-meta (o2r.info)

Metadata Software Needs

- Easy to use
  - Start-up
  - Graphical user interface
- Open Standards
  - International metadata standards
  - Standardized encoding
- Automation
  - Cataloguing / searching
  - Geographic metadata
  - Attribute metadata
  - Validation
  - Provenance

Results of Metadata Software Evaluation

Software	Easy Start	GUI	Standards	Encoding	Catalogue	Auto – Geographic	Auto – Attribute	Validate	Provenance
QGIS	✓	✓		XML	✓	✓	✓	✓	
SAGA	✓				✓	✓	✓		✓
GRASS	✓	✓	✓	XML		✓			
GeoNetwork		✓	✓✓	XML				✓	
GeoNode		✓	✓✓	XML		✓	✓		
mdEditor	✓✓	✓	✓	JSON				✓	
Metadata Wizard	✓✓	✓	✓	XML		✓	✓	✓	
Geometa			✓	XML				✓	
PyGeometa			✓	XML, YAML					
o2r-meta				XML, JSON	✓	✓		✓	

Three Points

1. Open Science and Reproducibility require standardized metadata

2. Researchers use, create, and modify information about their research projects and research data throughout the research life cycle

3. **We need better open source geospatial software to support metadata-rich research**

- Easy to use
  - Start-up
  - Graphical user interface
- Open Standards
  - International metadata standards
  - Standardized encoding
- Automation
  - Cataloguing / searching
  - Geographic metadata
  - Attribute metadata
  - Validation
  - Provenance
- Extensible

Questions, corrections, comments, and collaborations welcome!

www.github.com/HEGSRR
osf.io/c5a2r/
josephh@middlebury.edu

Thank You