Introduction to Research Data Management

Jordan Bratt, PhD
Yale Library DHLab

Ted Ellsworth, PhD
Yale Library StatLab

November 18, 2025

`Research Data Management`: organization, documentation, storage, and preservation of the data resulting from the research process

Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process… The outcomes from good data management and stewardship, therefore, are high quality digital publications that facilitate and simplify this ongoing process of discovery, evaluation, and reuse in downstream studies. — Wilkonson, M., et al. “The Fair Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016)

Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process… The outcomes from good data management and stewardship, therefore, are high quality digital publications that facilitate and simplify this ongoing process of discovery, evaluation, and reuse in downstream studies. — Wilkonson, M., et al. “The Fair Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016)

What is FAIR Data?

Findable Accessible Interoperable Reusable

Findability: Can someone *locate* your data? *(Is it discoverable?)*

Accessibility: Can someone *acquire* your data? *(Is is available to others?)*

Interoperability: Can your data *talk to* other data and systems? *(Is it a universal language?)*

Reusability: Can someone use your data for other purposes? (Is it adaptable and reproducible?)

Data Management Plan (DMP): document that outlines how data will be managed before, during, and after a research project.

- `DMPs` are **living-documents**
- `DMPs` are strongly encouraged and sometimes **required** (e.g. NSF, NIH)

DMP Tools

Want to learn more about how to write DMPs? Check out the DMP tool and our Library’s research guide on Research Data Management

Findability

The findability of your data and code can be divided into three important areas:

1. **Rich Metadata**
2. **Persistent Identifiers**
3. **Indexed in a Catalog or Database**

1.Rich Metadata

Metadata are structured statements about your data. They should be `'generous and extensive'`, including descriptive information about the context, quality and condition, and characteristics of your data.

It is important to ensure that metadata can be understood by humans and easily processed by machines.

Types of Metadata:

1. **Administrative Metadata**:This focuses on the management and administration of the dataset. This would encompass information related to:

- rights, use, data creation period, version, collaborators, funders, etc.

Types of Metadata:

2. **Descriptive or Citation Metadata**: This data allows users to discover and identify the dataset. This includes:

- author(s), title, abstract, keywords, persistent ID, related publications, etc.

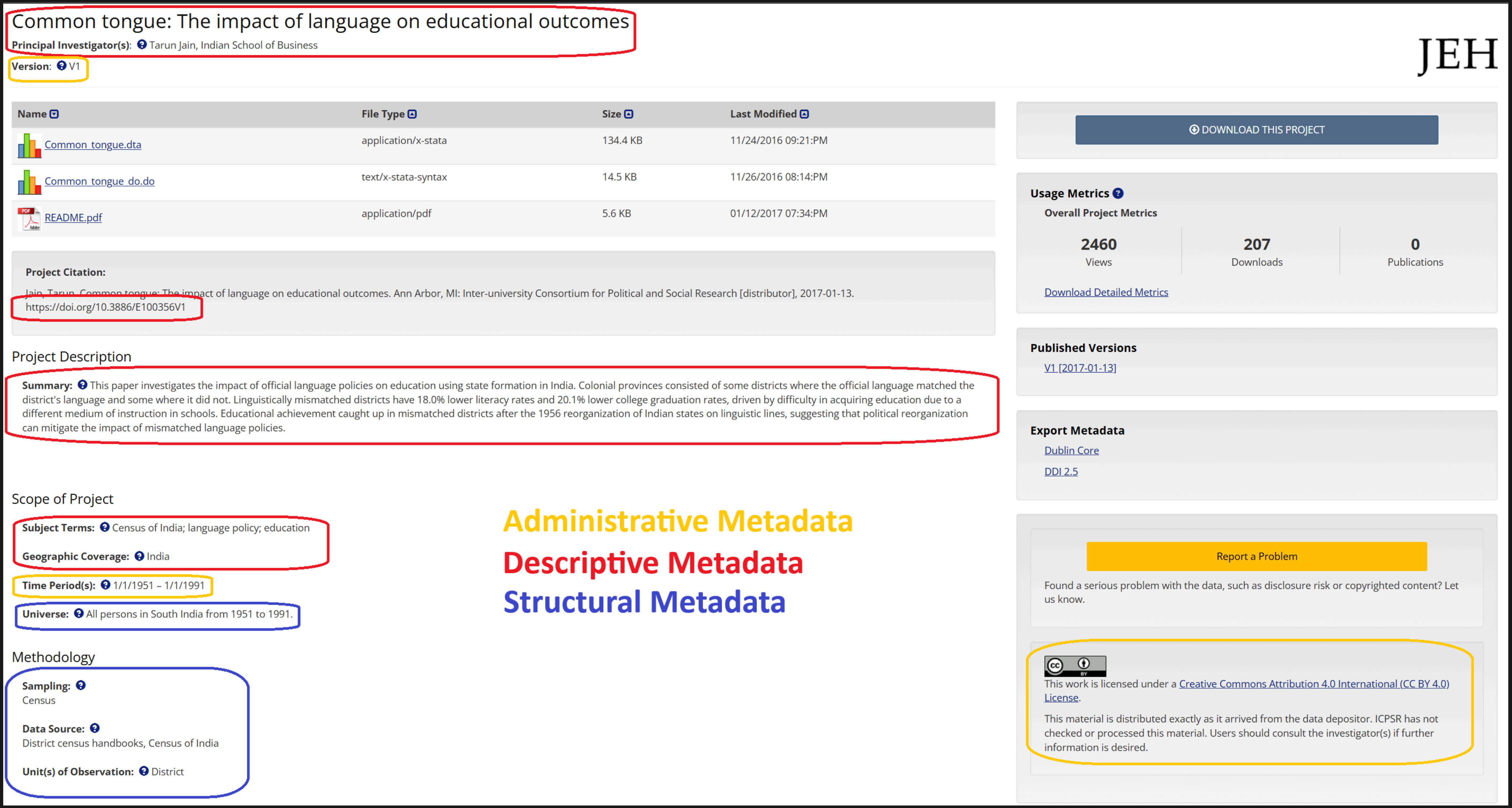
Descriptive metadata should be added to/updated continuously throughout the project.

Types of Metadata:

3. **Structural Metadata**: This focuses on how the data came about and is internally organized. This data describes:

- unit of analysis, collection method, sampling procedure, sample size, categories, variables, etc.

Structural metadata should be added to/updated continuously throughout the project.



Common tongue: The impact of language on educational outcomes

Principal Investigator(s): Tarun Jain, Indian School of Business

Version: V1

JEH

Name	File Type	Size	Last Modified
Common_tongue.dta	application/x-stata	134.4 KB	11/24/2016 09:21:PM
Common_tongue_do.do	text/x-stata-syntax	14.5 KB	11/26/2016 08:14:PM
README.pdf	application/pdf	5.6 KB	01/12/2017 07:34:PM

Project Citation:

Jain, Tarun. Common tongue: The impact of language on educational outcomes. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2017-01-13. https://doi.org/10.3886/E100356V1

Project Description

Summary: This paper investigates the impact of official language policies on education using state formation in India. Colonial provinces consisted of some districts where the official language matched the district's language and some where it did not. Linguistically mismatched districts have 18.0% lower literacy rates and 20.1% lower college graduation rates, driven by difficulty in acquiring education due to a different medium of instruction in schools. Educational achievement caught up in mismatched districts after the 1956 reorganization of Indian states on linguistic lines, suggesting that political reorganization can mitigate the impact of mismatched language policies.

Scope of Project

Subject Terms: Census of India; language policy; education

Geographic Coverage: India

Time Period(s): 1/1/1951 – 1/1/1991

Universe: All persons in South India from 1951 to 1991.

Methodology

Sampling:
Census

Data Source:
District census handbooks, Census of India

Unit(s) of Observation: District

Administrative Metadata

Descriptive Metadata

Structural Metadata

Usage Metrics

Overall Project Metrics

2460	207	0
Views	Downloads	Publications

Download Detailed Metrics

Published Versions

V1 [2017-01-13]

Export Metadata

Dublin Core

DDI 2.5

Found a serious problem with the data, such as disclosure risk or copyrighted content? Let us know.

This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) License.

This material is distributed exactly as it arrived from the data depositor. ICPSR has not checked or processed this material. Users should consult the investigator(s) if further information is desired.

Example of different types of metadata

The quality of the metadata has a huge impact on the *discoverability* and *reusability* of your data. It is best practice to use a `'metadata standard'` and/or an ontology.

A `'metadata standard'` is a discipline specific set of guidelines for consistent data description and management. They outline what content should be included, what syntax should be used, and what controlled vocabulary is included. They provide the properties we use to make our structured statements.

A common metadata standard in the social sciences is the `'Data Documentation Initiative'`. DDI is an international standard for describing surveys, questionnaires, statistical data files, and study-level information.

Other widely used metadata standards include `'Machine-Readable Cataloging (MARC)'` for libraries and `'Dublin Core'` for general-use and bibliographic information. Exploring the repository you wish to use can help you identify the metadata standard(s) they utilize and accept.

How/Where do I apply a metadata standard?

You should maintain a document that includes the metadata for your research data. This can be done using a CSV or XML file, where each row represents a data file or object, and the columns represent the metadata standard elements.

Tip

Some repositories allow you to upload a metadata file to automatically fill in the fields, while others require you to enter the information manually. For instance, OpenICPSR can import a metadata file in the DDI format, whereas Dataverse necessitates manual metadata entry.

There are many more metadata standards depending on the discipline and the type of data. Two helpful resources for exploring and identifying pertinent metadata standards are:

- Research Data Alliance (RDA) Metadata Standards Catalog
- FAIRsharing Standards

2. A Persistent Identifier

A `persistent identifier (PID)` is a long-lasting reference to a digital resource, providing the information required to reliably identify, verify, and locate the research data, thereby eliminating many misunderstandings.

Below is a list of different persistent identifiers with the first two being the most common:

- ORCID: Open Researcher and Contributor ID
- DOI: Digital Object Identifier
- ARK: Archival Resource Key
- PURL: Persistent URL
- RAiD: Research Activity Identifier

Which PID should I use?

The most common PID for datasets is the DOI. However, varioust guides are available to help you determine which PID is best suited for your dataset. Additionally, reviewing published datasets in your field can help identify common PIDs.

The important thing to remember is that all PIDs are meant to solve the same problem.

How do I obtain a PID for my dataset?

A PID is assigned to your dataset when it is deposited in an authorized repository that registers PIDs. You can select a repository based on your discipline or institution.

- General-purpose repositories: Zenodo or Figshare
- Institutional repository: Yale Dataverse
- Discipline repository: OpenICPSR

Resource: Search re3data.org for a list of suitable repositories

At what stage should I obtain a PID? Which data needs a PID?

A PID should be assigned to your data once it becomes shareable, citable, or is publicly disseminated. This is important to ensure that the research can be reliably located and identified over the long term.

Typically, this assignment occurs at the time of publication, but it may also happen when a new version is released.

3. Indexed in a Catalog or Database

Identifiers and rich metadata descriptions alone will not ensure “findability” for your data. If the availability of the dataset is not known, then no one (and no machine) will discover it.

The dataset should be indexed in a searchable resource, such as a data catalog or repository.

Consider publishing the dataset in various journals, some examples include:

- Journal of Computational Social Science
- Research Data Journal for the Humanities and Social Sciences
- Journal of Open Humanities Data

Accessibility

Once a user discovers your data and the software needed to use it, they need to know how to access the data.

In this context, “accessible” does not mean universally open. A good rule of thumb is data should be as open as possible, and as closed as necessary.

An important aspect to consider at this stage is the `CARE principles` for Indigenous Data Governance. This framework introduces an ethical and governance layer to data management, particularly concerning data related to Indigenous peoples. It focuses on centering their rights, well-being, and control over their own information.

- C: Collective Benefit
- A: Authority to control
- R: Responsibility
- E: Ethics

Data should be retrievable by their PID using a standardized Communication Protocol

Data retrieval should not require specialized or proprietary tools or communication methods. Most users retrieve data by “clicking on a link.” There are other protocols that can be used to access your data including File Transfer Protocol (FTP).

Highly Sensitive Data

If you are working with sensitive data, you can establish a `'contact protocol'`. This provides an email and/or phone number of a contact person who can provide access to the data as appropriate. Contact protocols should be clearly described in the metadata.

Additionally, you can work with the Office of Research Administration to address concerns around sensitivity, privacy, etc.

The protocol provides for an authentication and authorization procedure when necessary

The specific conditions for data availability should be clearly stated. For restricted data, a repository may require users to create an account to authenticate their identity and ensure compliance with the data use agreements established by the data owner.

Metadata should be available even when the data is not

Access to data can be restricted for two primary reasons:

1. Privacy/Sensitivity
2. Dataset degradation or loss

Whatever the reason for the limited or restricted access to the data, metadata should continue to be available. Storing and maintaining metadata is much easier and it allows individuals associated with the original research to be contacted by interested users.

Interoperability

Legal/Ethical	Organizational
Technological	Semantic

1. Legal/Ethical Interoperability

This layer is about the rules, laws, and agreements governing your data. It defines who can use your data and for what purpose.

For a researcher, this means asking:

- What are my `'IRB restrictions'`?
- Am I working with data covered by `'HIPAA'` or `'GDPR'`?
- What does my `'Data Use Agreement (DUA)'` say?

A researcher at University A receives de-identified patient data from Hospital B.

The `'Data Use Agreement'` is the key legal document. It explicitly states:

- Data can only be used for the `'specified research project'`.
- Data cannot be shared with `'third parties'` without written consent.
- Data must be destroyed `'five years'` after project completion.

2. Organizational Interoperability

This layer is about aligning processes and expectations between collaborating groups. It ensures everyone works together in a coordinated way.

For a researcher, this organizational interoperability means:

- Defining roles and `'responsibilities'` within a team/lab
    - Data `storage`: All raw sequence files stored on A’s server
    - Data `processes`: Lab B will upload data within`48 hours` of a successful sequence run
    - Data `access`: named individuals have read/write privileges
- Documenting workflows and data-exchanges with other institutions.
- Creating a `'Memorandum of Understanding (MoU)'` to define interoperability

3. Semantic Interoperability

This is the most common failure point in research. It’s about ensuring the *meaning* of your data is clear and unambiguous.

A new grad student is asked to analyze data from a previous project. They receive this file:

`turnout_final_v2_AM.csv`

precinct_id	value	treatment
101	68	1
102	71	2
103	66	1

What does any of this mean?

The Filename

```
turnout_final_v2_AM.csv
```

- The filename is not descriptive. Who is AM? Is this really the final version? Which election’s turnout?

The Column Headers

precinct_id	value	treatment
101	68	1
102	71	2
103	66	1

- The column headers are vague.
- What is `value`? Is it percentage turnout or raw vote count?
- What does `treatment` represent?

Jenny Bryan’s Rules for Naming Things

Names should be:

1. machine readable
2. human readable
3. sorted in a useful way

Use consistent, deliberate structure.

Bad 👎

Filename
`Turnout Data GA Nov 2024.csv`
`figure1.png`
`county economic data rev2.RData`

Good 👍

Filename
`2024-11-05_turnout-general_georgia-county.csv`
`fig01_turnout-by-income_2024-general.png`
`2025-01-30_analysis_economic-indicators_county.RData`

A Better Structure

A good filename acts as a preview of the contents.

`2024-11-05_turnout-general_georgia-county.csv`

We can parse this with code or our eyes:

`[date]_[data-type]_[election-type]_[geography]_[unit].csv`

This structure makes files easy to find, sort, and process programmatically.

Applying the Rules

A better filename: `2024-11-05_turnout-by-ad-campaign_georgia-precinct.csv`

With better column names:

precinct_id	turnout_percentage	ad_campaign_group
101	68.5	1
102	71.2	2
103	66.9	1

But what does `1` in `ad_campaign_group` mean?

The Final Step: The Data Dictionary

To achieve full semantic interoperability, we use a **‘data dictionary’** (or codebook), a separate file that defines the data. This is often a `README.md` or `codebook.md` file.

codebook.md

```
# Codebook for Georgia Turnout Experiment

**File:** `2024-11-05_turnout-by-ad-campaign_georgia-precinct.csv`
**Date Created:** 2025-01-15
**Contact:** J. Doe (jdoe@university.edu)

*   **precinct_id**: Unique identifier for the electoral precinct, assigned by the

*   **turnout_percentage**: The percentage of registered voters in the precinct wh

*   **ad_campaign_group**: The experimental group assignment for the precinct's me
    *   `'1'` = **Treatment Group**: Exposed to the new digital ad campaign.
    *   `'2'` = **Control Group**: Not exposed to the new digital ad campaign; rec
```

Use persistent identifiers: Incorporate persistent identifiers (like DOIs) and link to related resources to make data discoverable and understandable.

Use standard protocols: Use pre-existing schemas and vocabularies and meta-data to communicate data origins and compatabilities

4. Technical Interoperability

This layer is about the nuts and bolts—the hardware and software. It ensures that machines can talk to each other and read the files.

Key considerations for researchers: * Using `open, non-proprietary file formats` (e.g., .csv, .txt).

* Documenting software versions and dependencies.

* Using `'APIs'` (Application Programming Interfaces) in a standardized way.

A researcher shares their data in two formats.

1. `data.csv`: A Comma-Separated Value file.
    - **Pro:** Can be opened by any software (R, Python, Excel, Google Sheets, a simple text editor). It is `maximally interoperable`.
2. `data.spss`: A proprietary IBM SPSS Statistics file.
    - **Con:** Can only be opened by someone who has a `specific, often expensive, software license`. This creates a significant barrier.

Always prefer open formats for data sharing.

Consider the Needs of Researchers After You

- Does anything in the file require my computer to run?

R

```r
sessionInfo()
```

Python

```python
import session_info
session_info.show()
```

To guarantee your code runs forever, you must lock package versions.

In R…

```r
# 1. Initialize renv in your project
renv::init()

# 2. As you install packages, renv tracks them
install.packages("dplyr")

# 3. Save the exact state of your library to renv.lock
renv::snapshot()

# A collaborator just needs to run:
renv::restore() # This installs the exact package versions
```

In Python…

```
# Create a file with all packages and their exact versions
pip freeze > requirements.txt

# A collaborator just needs to run:
pip install -r requirements.txt
```

Reusability

A primary condition for reusability is rich documentation and metadata!

good documentation should provide `context`

- Why was the data collected?
- How was it collected?
- What limitations are present?

This information should live in the `README.txt` at the top level of your project!

If people don’t know how they are allowed to reuse the data, they simply won’t. A license removes all ambiguity.

Creative Commons licensing types:

- `CC0`: Public domain dedication. No rights reserved. Maximum reusability.
- `CC BY`: Attribution. Others can use your data for any purpose, as long as they credit you.
- `CC BY-NC`: Only non-commercial use.
- `CC BY-ND`: Attribution. No derivatives.

Data is only reusable if has clear `provenance!`

Provenance is the documented history of your data. It answers:

- Where did the `raw data` come from?
- What exact steps transformed it into `processed data`?
- What software versions and packages were used to generate the results?

Pro-tip: Write Scripts with Explicit Inputs/Outputs!

```python
import pandas as pd
from pathlib import Path

data_dir = Path(".") / "data"
raw_file = "2025-10-01-covid19-infections-raw.csv"
processed_file = raw_file.replace("-raw.csv", "-processed.csv")

data_frame = pd.read_csv(data_dir / raw_file)

def clean_covid_data(df):
    df = df.drop_duplicates()
    df['date'] = pd.to_datetime(df['date'])
    df[df.select_dtypes(include=['float64', 'int64']).columns] = df.select_dtypes(
    df['state'] = df['state'].str.upper()
    return df.sort_values(['date', 'state'])

covid_state = clean_covid_data(data_frame)

```

Putting everything together…

```
georgia-voter-turnout-replication/
│
├── README.md               # Top-level guide: explains the project, data, and how to
run it.
├── LICENSE.md              # Standard license for code and non-restricted data
(e.g., MIT, CC-BY).
├── figures_and_tables.csv  # Maps scripts to specific outputs (e.g., "figure_1.png"
-> "code/03_figure-generation/make_turnout_map.R").
│
├── 📁 data/
│   ├── 📁 L1_raw_restricted/
│   │   └── 📄 ACCESS_NOTE.txt     # IMPORTANT: Explains that raw voter data is NOT
included due to privacy laws.
│   │
```

README.txt

```
# README: Replication for "The Impact of Polling Place Consolidation on Voter Turnout
in Georgia"

**Author Contact Information:**
Jane Doe, j.doe@university.edu

---

### **IMPORTANT DATA & ACCESS NOTE:**

This replication package contains all code and public-facing data required to
reproduce the main analyses, tables, and figures in the published article.

However, due to Georgia state law and a Data Use Agreement (DUA) with the Secretary
```

Contact us!

Jordan Bratt, Digital Humanities Lab,
jordan.bratt@yale.edu

Ted Ellsworth, Statistical Support Services (StatLab),
ted.ellsworth@yale.edu

Brandon Miliate, Research Data Management Librarian,
brandon.miliate@yale.edu

Feedback Form