

# Preserving Survey Data

Lars Vilhuber    Laurel Krovetz

2026-06-14



## Preserving Survey Data

Lars Vilhuber   Laurel Krovetz

2026-06-14



[labordynamicsinstitute.github.io/tutorial-preserving-survey/presentation/](https://labordynamicsinstitute.github.io/tutorial-preserving-survey/presentation/)



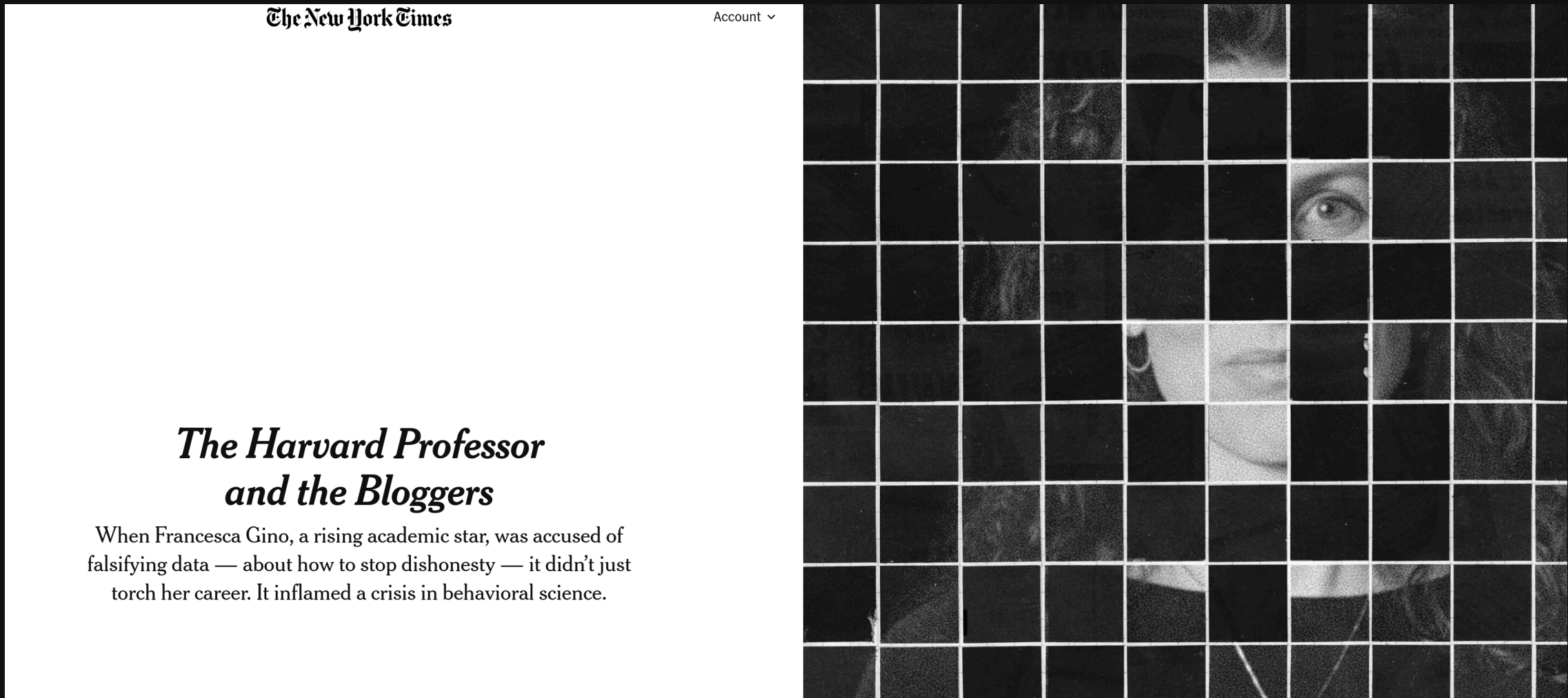
# The problem of credibility



**How can we know that a data source is reliably obtained?**



# Consider the case of Gino

A screenshot of a New York Times article page. The page is white with black text. At the top left is the New York Times logo, and at the top right is an 'Account' dropdown menu. The main content area is mostly blank, with a large grid of dark squares overlaid on the right side. The grid is composed of many small squares, some of which are slightly offset or missing, creating a fragmented effect. The text of the article is visible on the left side of the grid.

The New York Times Account ▾

## *The Harvard Professor and the Bloggers*

When Francesca Gino, a rising academic star, was accused of falsifying data — about how to stop dishonesty — it didn't just torch her career. It inflamed a crisis in behavioral science.

Francesca Gino



# The case of Gino

- Francesca Gino was a tenured professor at Harvard Business School, writing on honesty (!)



# The case of Gino

- Several articles were investigated by third parties (Data Colada, in particular <sup>1</sup>), and found to be problematic



The screenshot shows the Data Colada website. At the top center is the logo, which features a blue cocktail glass with a pink umbrella and a straw, containing orange liquid and the numbers 18, 20, 2, and 6. To the right of the glass, the word "DATA" is written in large, bold, black letters, and "COLADA" is written below it in smaller, orange letters. Below the logo is the tagline "Thinking about evidence, and vice versa" in a dark blue font. A horizontal line separates the header from the navigation menu, which includes the links "HOME", "TABLE OF CONTENTS", "FEEDBACK POLICY", and "ABOUT". Below the navigation menu is the article title "[118] Harvard's Gino Report Reveals How A Dataset Was Altered" in a large, orange font. To the right of the article title, there is a yellow sidebar with the text "GET CO" and "you@".

**DATA COLADA**  
Thinking about evidence, and vice versa

[HOME](#) [TABLE OF CONTENTS](#) [FEEDBACK POLICY](#) [ABOUT](#)

**[118] Harvard's Gino Report Reveals How A Dataset Was Altered**

GET CO  
you@

# The case of Gino

- At least one of them had manipulated data **AFTER** it had been collected, **BEFORE** it had been analyzed.

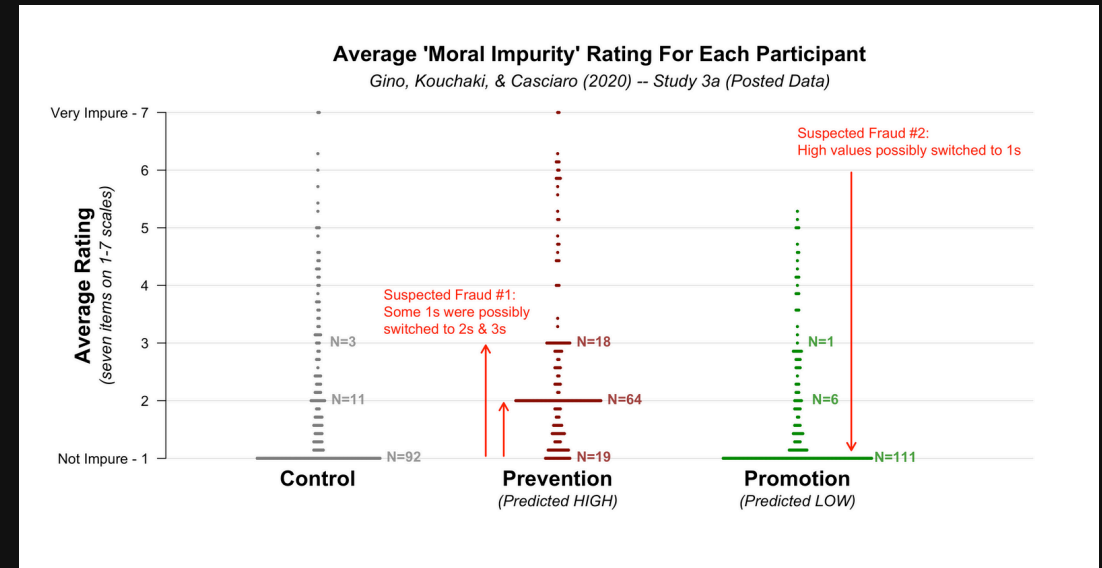
How The Moral Impurity Data Were Altered In the Promotion Condition

ID	Impure1	Impure2	Impure3	Impure4	Impure5	Impure6	Impure7
233	0	0	0	0	0	0	0
200	-2	-2	-2	-2	-2	-2	-3
447	-3	-3	-5	-3	-5	-6	-5
471	-3	-4	-5	-4	-5	-6	-5
335	-4	-3	-6	-4	-5	-6	-6
319	0	0	0	0	0	0	0
199	-2	-4	-3	-4	-5	-2	-4
30	0	0	0	0	0	0	0
498	-4	-3	-5	-2	-2	-4	-4
237	0	0	0	0	0	0	0
118	-5	-5	-6	-4	-4	-5	-5
120	-4	-4	-4	-4	-4	-4	-4
204	0	0	0	0	0	0	0
309	0	0	0	0	0	0	0
589	-4	-4	-5	-4	-5	-6	-5
220	0	0	0	0	0	0	0
251	0	0	0	0	0	0	0
248	-5	-5	-6	-4	-5	-6	-4
364	-6	-6	-6	-6	-6	-6	-6
376	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0
268	0	0	0	0	0	0	0
290	-3	-2	-3	-2	-2	-4	0
47	0	0	0	0	0	0	0
454	-3	-2	-4	-3	-4	-4	-4
441	-4	-3	-2	-3	-4	-5	-3
538	-6	-6	-6	-6	-6	-6	-6
8	0	0	0	0	0	0	0

Participant IDs in the Posted Data

Participant 200's 7<sup>th</sup> moral impurity rating was decreased by 3

All of Participant 538's moral impurity ratings were decreased by 6

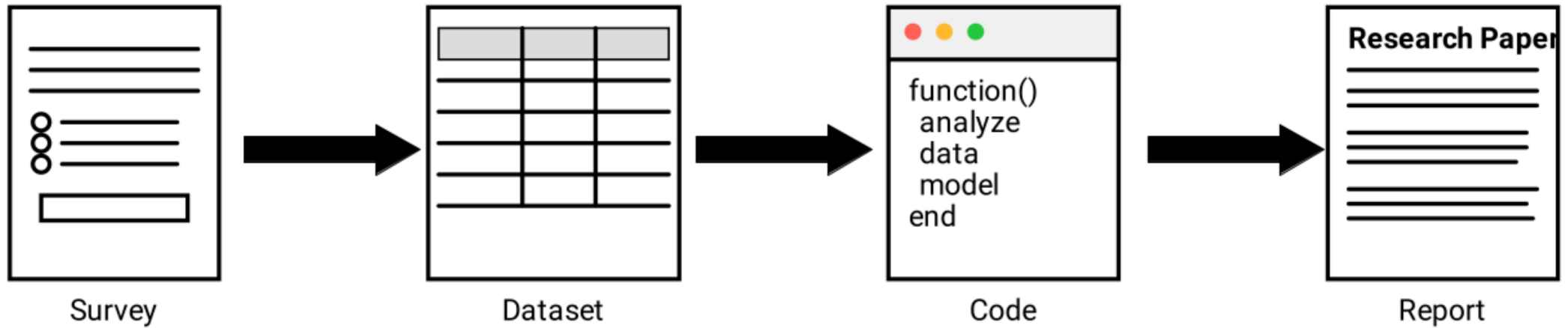


Results of manipulation

# A generic survey workflow

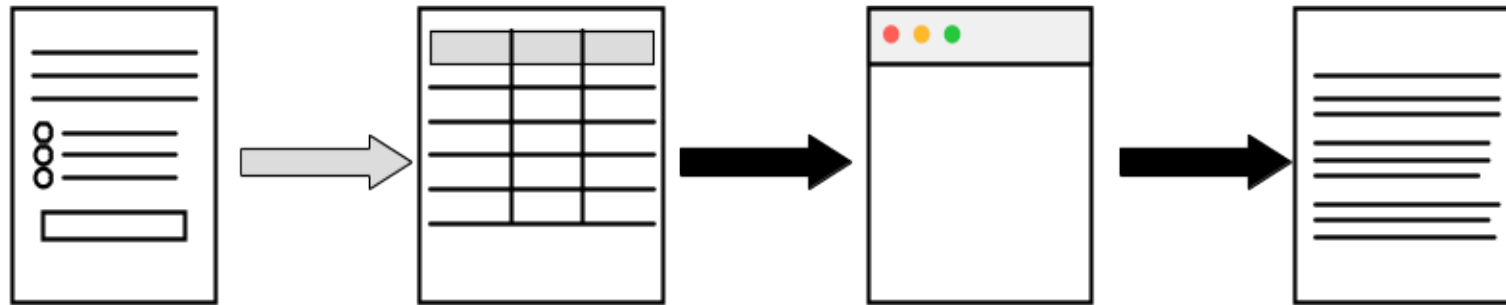


# Generic survey processing



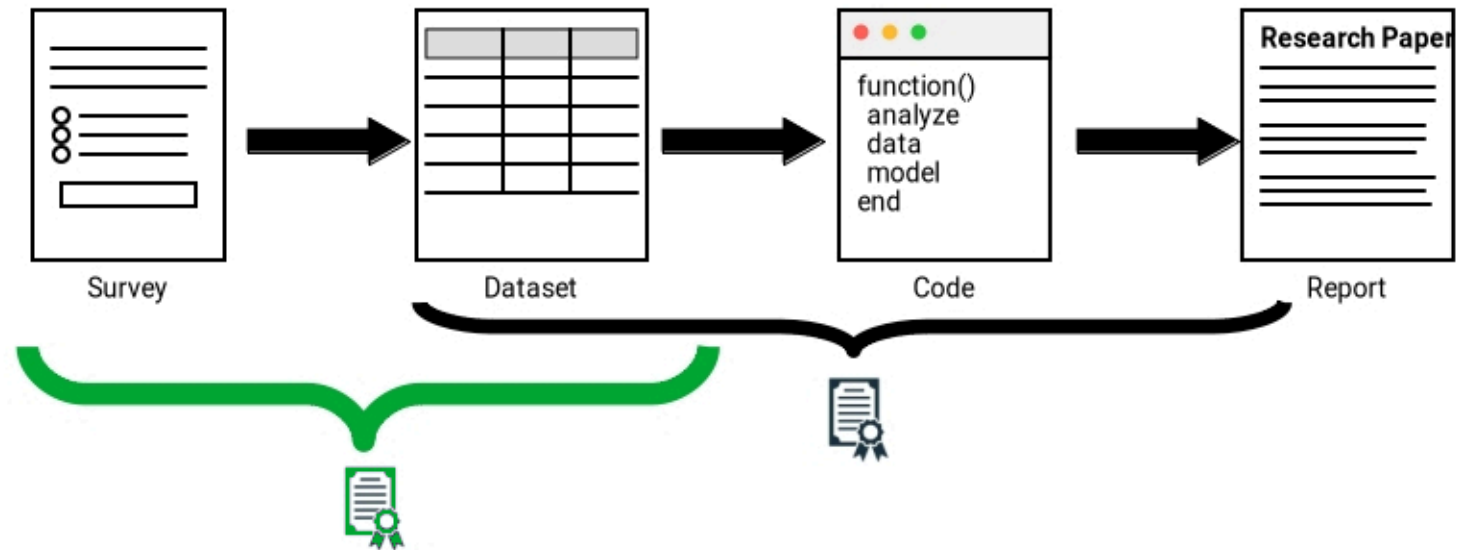
Generic survey processing

# Requiring transparency in academia



Generic survey processing

# Where we are headed

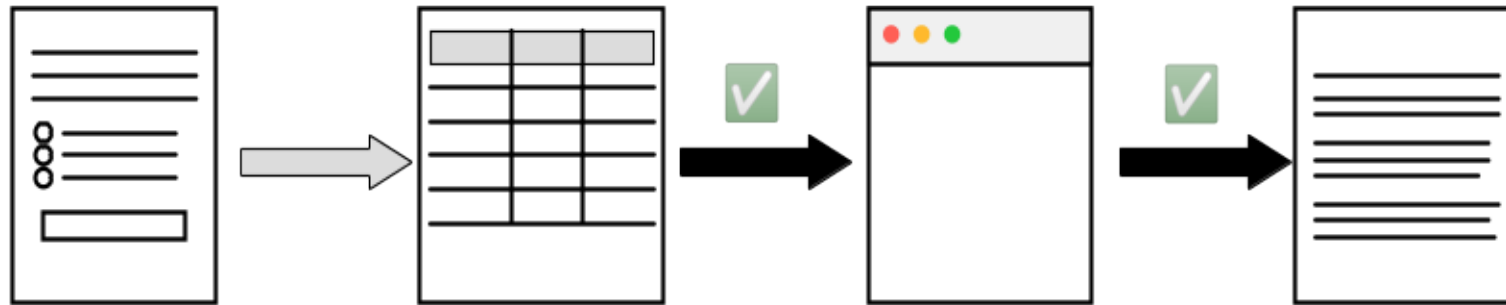


Certified survey processing

# Modern verification processes



# Verifying transparency in academia



Generic survey processing

# Verification by journals

- **Provision** (publication of materials) provides transparency
- **Verification** (running the analysis again - computational reproducibility) compensates for **mistrust/absence of trust**

# Which journals

- American Economic Association (8)
- Econometric Society (3)
- Canadian Journal of Economics (1)
- Royal Economic Society (2)
- Western Economic Association International (1)
- European Economic Association (1)
- Review of Economic Studies (1)
- Journal of the European Economic Association (1)
- Journal of Political Economy (3)
- American Journal of Political Science (1)
- American Political Science Review (1)







Data and Code Availability Standard

---

## Journals

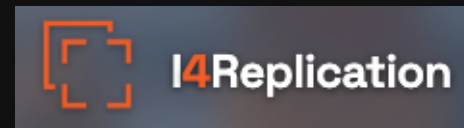
The following journals endorse the Data and Code Availability Standard.

1. American Economic Journal: Applied Economics  
2. American Economic Journal: Economic Policy  
3. American Economic Journal: Macroeconomics  
4. American Economic Journal: Microeconomics  
5. American Economic Review  
6. American Economic Review: Insights  
7. Canadian Journal of Economics 
8. Econometrica 
9. Econometrics Journal
10. Economic Inquiry 
11. Economic Journal 
12. Journal of Economic Literature  
13. Journal of Economic Perspectives  
14. Journal of Political Economy 
15. Journal of the European Economic Association 
16. Quantitative Economics 
17. Review of Economic Studies 



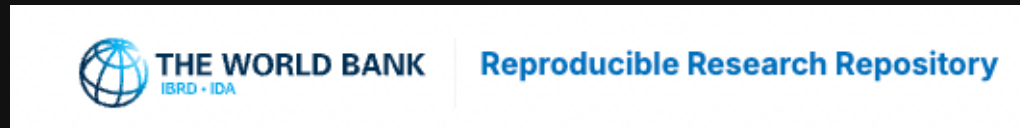
# Verification by others

- Pre-publication: **cascad**
- Post-publication: **Data Colada, Institute for Replication**



# Verification by institutions

- World Bank



World Bank RRR<sup>2</sup>

 A screenshot of the J-PAL website. The header includes the J-PAL logo (Abdul Latif Jameel Poverty Action Lab) and navigation links: Blog, Careers, Courses, For Affiliates, Support J-PAL, and a search icon. Below the header are links for EVALUATIONS, RESEARCH RESOURCES, POLICY INSIGHTS, and EVIDENCE TO POLICY. A secondary navigation bar contains "About", "Offices", and "Sectors". The main content area features the title "Research Transparency and Reproducibility" in orange. The text below states: "For over 15 years, J-PAL has been a leader in making research more transparent. In 2008, J-PAL was one of the first organizations to have researchers publish their data on the [Harvard Dataverse](#), a repository for published scientific data. In 2009, we developed a [hypothesis registry](#), which was the precursor to the [American Economic Association's \(AEA\) registry for randomized controlled trials](#) where researchers can publish their study design." A second paragraph mentions: "J-PAL works closely with other organizations that promote research transparency, including the [Berkeley Initiative for Transparency in the Social Sciences \(BITSS\)](#), [Center for Open Science](#), [Innovations for Poverty Action \(IPA\)](#), and [International Initiative for Impact Evaluation \(3ie\)](#), among others." On the right side, there is a "PAGE CONTENT" section with a dropdown arrow, containing "Core activities", "Contacts", and "Read more about our work".

J-PAL again?



# Outline of the tutorial



# Basic

- how to process,
- de-identify, and
- analyze survey data
- publish data

# Expansion

- how to download automatically
- how to preserve automatically
- how to do so in a credible and transparent fashion.

# In a nutshell

We'll use an API to retrieve the data, show you how to clean and strip the data of confidential information and non-consenting responses, and use another API to preserve the data.



# Goals

- **Create a survey** (in Qualtrics) for data collection.
- Load the latest responses from the server (**using an API**)
- Clean and process the data to **remove non-public data automatically.**
- **Preserve** shareable data in a trusted repository
- Later, **publish those data** with a credible record of when it was first preserved!

# Some notes on Qualtrics

- **Create a survey** (in Qualtrics) for data collection.
- Load the latest responses from the server **using an API**

There is not much in this tutorial that requires Qualtrics.

- You could do this with SurveyCTO, LimeSurvey, or any other system that has an API.
- You could do this with Google Forms, if you have linked it to a Google Sheet.
- You could do this with a **lab experiment** system that stores data in an SQL database

# Some notes on removal of PII

- Clean and process the data to **remove non-public data automatically.**

It is important to remove any PII or confidential information as soon as possible.

# Some notes on removal of PII

- That **may not always be feasible**. For instance, if you need geolocation to merge in contextual data, or compute distances, then some data processing may unavoidably require access to sensitive data.
- But any data that is ***not needed should be removed*** early on.
- This is **not irreversible**: if you later find that you need more data elements, you can always re-process the raw data, stored on Qualtrics, until your IRB requires that you delete those data.

# Some notes on Preservation vs. Sharing

- **Preserve** shareable data in a trusted repository
- Later, **publish those data** with a credible record of when it was first preserved!

It is important to distinguish

- **preserving data** from
- **publishing** data, and possibly
- **sharing data with collaborators**



# Preservation

- Preservation != publication, != sharing
- In fact, preservation may mean: not very accessible at all!
- Preservation is intended to maintain data for tens, even hundreds of years
  - Preservation may involve curation: active transformation of the data for improved accessibility

# Sharing data

- Shared on a personal website
- Sharing a Dropbox link
- Posting it on OSF as a project

All useful for sharing, but do not preserve the data

# Test

- Who has a Github account?

# Test

- Who has a Github account?
- How long does it take you to delete your entire Github repository, forever?

# Demonstrating the core steps in R



# The core steps

We will first walk through the core steps you can do **by hand**, in R:

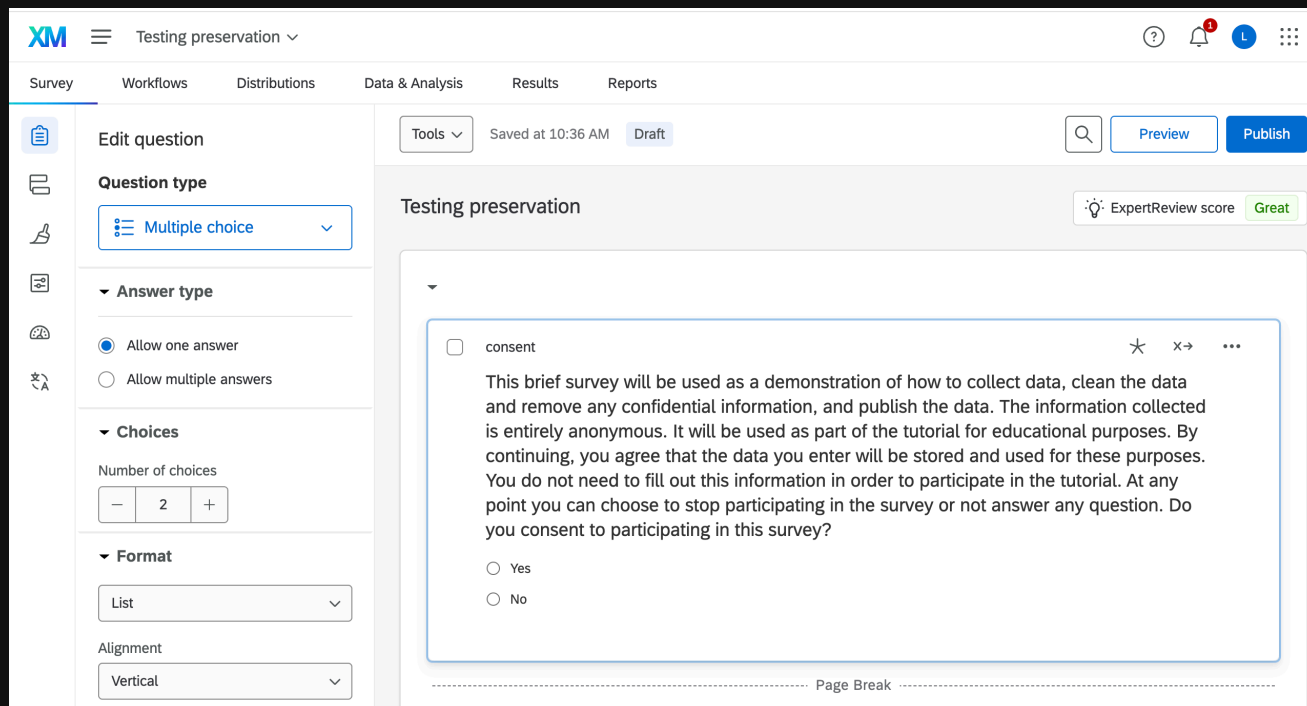
1. **Collect** — create the survey and gather responses (in Qualtrics)
2. **Download** — export the responses from the web interface
3. **Analyze** — load, clean, and process the data in R
4. **Publish** — save the cleaned data, ready to share

**This is not a full Qualtrics tutorial!**



# Creating a survey in Qualtrics

You'll typically have access to a Qualtrics account through your university or organization. Then it is easy to construct a survey using the web tool.



The screenshot displays the Qualtrics web tool interface for editing a survey question. The top navigation bar includes the XM logo, a menu icon, and the current survey name "Testing preservation". The main navigation tabs are Survey, Workflows, Distributions, Data & Analysis, Results, and Reports. The left sidebar contains icons for question types and a list of question types, with "Multiple choice" selected. The main editing area shows the question title "Testing preservation", a search icon, and buttons for "Preview" and "Publish". The question content is a consent form with the following text: "This brief survey will be used as a demonstration of how to collect data, clean the data and remove any confidential information, and publish the data. The information collected is entirely anonymous. It will be used as part of the tutorial for educational purposes. By continuing, you agree that the data you enter will be stored and used for these purposes. You do not need to fill out this information in order to participate in the tutorial. At any point you can choose to stop participating in the survey or not answer any question. Do you consent to participating in this survey?". Below the text are two radio button options: "Yes" and "No". The interface also shows a "Page Break" indicator at the bottom of the question content area.



# Side-note: Survey definition from Qualtrics

You should not forget to preserve your survey definition!

- Download a `qsf` file to save and transfer survey structure to have a backup survey template. Export as `.qsf` in `Tools` in Qualtrics.
- Can also export survey as Word doc in `Tools` in Qualtrics. Choose this option to get a well-formatted document.



# Side-note to side-note: Confidential data in survey definitions!

It is possible that your survey definition itself contains information that you are not allowed to publish:


- You might be running the survey with a firm, and the firm does not want to be identified
- You are asking questions about specific products, and the product names are confidential

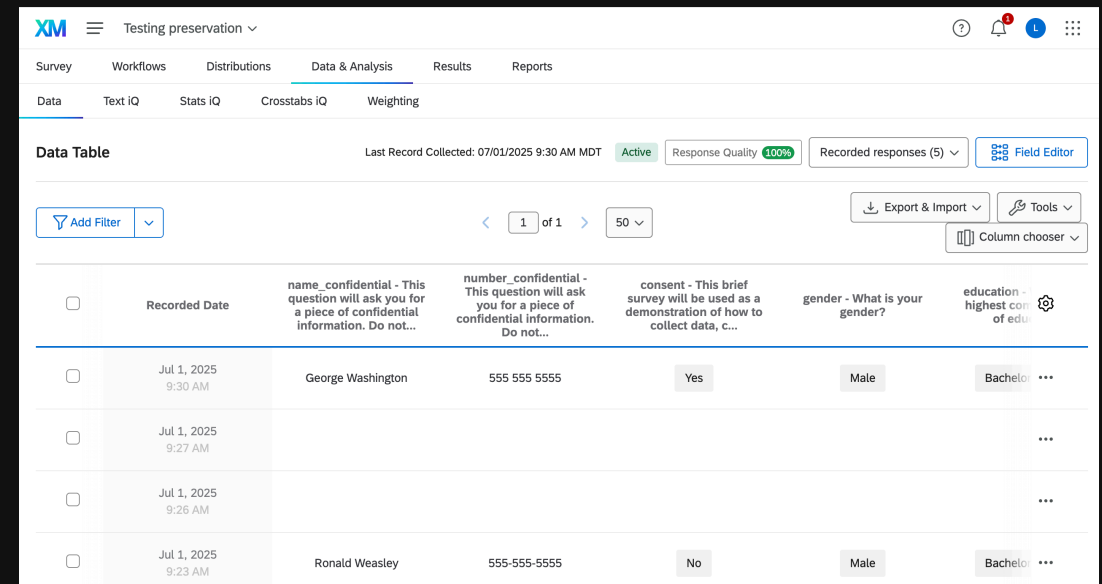
It is actually **hard** to de-identify a `qs_f` file. We will not try to do this here, but you should be aware of this issue.

# Take our survey



# Survey responses in Qualtrics

Responses can be easily checked at a glance in the **Data** and **Analytics** tab. 



The screenshot shows the Qualtrics Data Table interface. The top navigation bar includes 'Survey', 'Workflows', 'Distributions', 'Data & Analysis' (selected), 'Results', and 'Reports'. Below this, there are sub-tabs for 'Data', 'Text IQ', 'Stats IQ', 'Crosstabs IQ', and 'Weighting'. The main area displays a 'Data Table' with the following details:

- Last Record Collected: 07/01/2025 9:30 AM MDT
- Active
- Response Quality: 100%
- Recorded responses (5)
- Field Editor
- Export & Import
- Tools
- Column chooser

The table has 6 columns and 5 rows of data. The columns are:

- Recorded Date
- name\_confidential - This question will ask you for a piece of confidential information. Do not...
- number\_confidential - This question will ask you for a piece of confidential information. Do not...
- consent - This brief survey will be used as a demonstration of how to collect data, c...
- gender - What is your gender?
- education - highest com of edu

The data rows are:

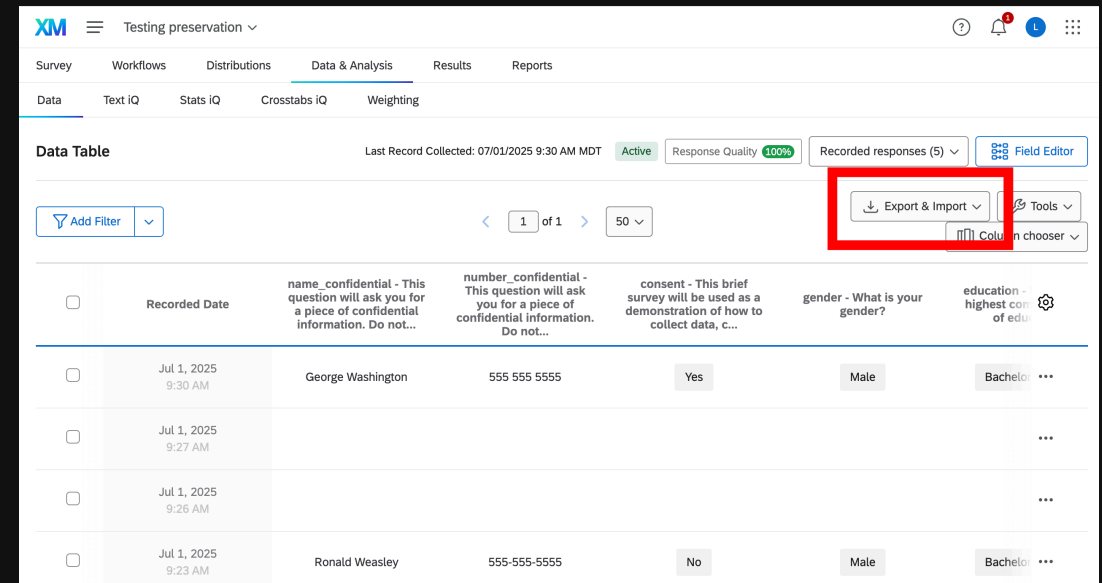
Recorded Date	name_confidential	number_confidential	consent	gender	education
Jul 1, 2025 9:30 AM	George Washington	555 555 5555	Yes	Male	Bachelo...
Jul 1, 2025 9:27 AM					...
Jul 1, 2025 9:26 AM					...
Jul 1, 2025 9:23 AM	Ronald Weasley	555-555-5555	No	Male	Bachelo...

Qualtrics data interface

# Downloading data

You can download data directly from this page

- If you do this only once, downloading manually is fine.
- Do it 2-3 times, you **may** want to program it!



The screenshot shows the Qualtrics Data Table interface. The 'Export & Import' button is highlighted with a red box. The table displays the following data:

Recorded Date	name_confidential - This question will ask you for a piece of confidential information. Do not...	number_confidential - This question will ask you for a piece of confidential information. Do not...	consent - This brief survey will be used as a demonstration of how to collect data, c...	gender - What is your gender?	education - highest com of edu
Jul 1, 2025 9:30 AM	George Washington	555 555 5555	Yes	Male	Bachelo...
Jul 1, 2025 9:27 AM					...
Jul 1, 2025 9:26 AM					...
Jul 1, 2025 9:23 AM	Ronald Weasley	555-555-5555	No	Male	Bachelo...

Qualtrics data download

# Download options

You can download data directly from this page

- Do it 2-3 times, you **may** want to program it!

## Download a data table

CSV TSV Excel XML SPSS Google Drive User-submitted files

---

### Comma separated values

This is a .csv file that can be imported into other programs. Each value in the response is separated by a comma and each response is separated by a newline character. If your responses contain special characters and you will open this export in Microsoft Excel we recommend using the TSV export. Qualtrics CSV exports use UTF-8 encoding, which Excel will not open correctly by default.

[Learn more](#)

Download all fields

Values or labels

Export values

Export labels

[More Options](#)



# Downloaded data

The data downloaded depends on parameters chosen. For instance, downloading as CSV with default settings yields

```

1 StartDate,EndDate,Status,Progress,Duration (in seconds),Finished,RecordedDate,Response
2 Start Date,End Date,Response Type,Progress,Duration (in seconds),Finished,Recorded Date
3 "{\"ImportId\":\"startDate\",\"timeZone\":\"America/New_York\"}\", \"{ \"ImportId\": \"endDa
4 2025-07-01 11:13:44,2025-07-01 11:14:18,IP Address,100,34,True,2025-07-01 11:14:19,R_5
5 2025-07-01 11:23:01,2025-07-01 11:23:28,IP Address,100,26,True,2025-07-01 11:23:28,R_5

```

# Loading the downloaded data into R

You downloaded the responses from the Qualtrics web interface (previous slide). Now read that exported file into R.

```
1 datesuffix <- "June+16,+2026_15.35"  
2 fileprefix <- "Testing+preservation_"  
3 filename <- paste0(fileprefix, datesuffix, ".csv")
```



# Some data org hygiene

We want to be careful about managing our data structure:<sup>3</sup>

```
1 # Path to the file you downloaded from Qualtrics
2 datapath <- here::here("data")
3 rawdatapath <- file.path(datapath, "raw-confidential")
4 confdatapath <- file.path(datapath, "confidential")
5 cleandatapath <- file.path(datapath, "clean")
6 metadatapath <- file.path(datapath, "metadata")
```



# Minor thing

Let's ensure that these paths all exist!

```
1 for (path in list(rawdatapath, confdatapath, cleandatapath, metadatapath)) {  
2   if (!dir.exists(path)) {  
3     dir.create(path, recursive = TRUE)  
4     message("Created directory: ", path)  
5   } else {  
6     message("Directory already exists: ", path)  
7   }  
8 }
```

Directory already exists: /home/runner/work/tutorial-preserving-survey/tutorial-preserving-survey/data/raw-confidential

Created directory: /home/runner/work/tutorial-preserving-survey/tutorial-preserving-survey/data/confidential

Created directory: /home/runner/work/tutorial-preserving-survey/tutorial-preserving-survey/data/clean

Created directory: /home/runner/work/tutorial-preserving-survey/tutorial-preserving-survey/data/metadata



# Loading the downloaded data into R

- Any CSV reader works too, with some adjustments.

```
1 library(readr)
2 # discard the two Qualtrics metadata rows
3 data.raw <- read_csv(file.path(rawdatapath, filename), skip = 3,
4                       col_names = FALSE)
5 # read the header separately to get column names
6 header <- read_csv(file.path(rawdatapath, filename),
7                    n_rows=0)
8 colnames(data.raw) = colnames(header)
```



# Loading the downloaded data into R

```
1 head(data.raw)
```

```
# A tibble: 6 × 17
```

```
  StartDate      EndDate      Status Progress Duration..in.seconds.
  <dtm>         <dtm>         <chr>    <dbl>      <dbl>
1 2025-07-01 11:13:44 2025-07-01 11:14:18 IP Add...    100      34
2 2025-07-01 11:23:01 2025-07-01 11:23:28 IP Add...    100      26
3 2025-07-01 11:26:40 2025-07-01 11:26:40 Survey...    100       0
4 2025-07-01 11:27:12 2025-07-01 11:27:12 Survey...    100       0
5 2025-07-01 11:30:26 2025-07-01 11:30:47 IP Add...    100      21
6 2025-12-19 12:20:44 2025-12-19 12:21:19 IP Add...    100      35
# i 12 more variables: Finished <lgl>, RecordedDate <dtm>, ResponseId <chr>,
#   DistributionChannel <chr>, UserLanguage <chr>, consent <chr>, age_1 <dbl>,
#   gender <chr>, education <chr>, num_tabs_1 <dbl>, name_confidential <chr>,
#   number_confidential <chr>
```



# Loading with `qualtRics` package

- The `qualtRics`<sup>4</sup> package can read a Qualtrics CSV export directly with `read_survey()`:

```
1 library(qualtRics)
2
3 data.raw <- read_survey(file.path(rawdatapath, filename))
```

# Loading with `qualTRics` package

```
1 head(data.raw)
```

```
# A tibble: 6 × 17
  StartDate      EndDate      Status Progress Duration (in seconds...1
  <dtm>         <dtm>         <chr>      <dbl>      <dbl>
1 2025-07-01 11:13:44 2025-07-01 11:14:18 IP Ad...      100      34
2 2025-07-01 11:23:01 2025-07-01 11:23:28 IP Ad...      100      26
3 2025-07-01 11:26:40 2025-07-01 11:26:40 Surve...      100       0
4 2025-07-01 11:27:12 2025-07-01 11:27:12 Surve...      100       0
5 2025-07-01 11:30:26 2025-07-01 11:30:47 IP Ad...      100      21
6 2025-12-19 12:20:44 2025-12-19 12:21:19 IP Ad...      100      35
# i abbreviated name: 1`Duration (in seconds)`
# i 12 more variables: Finished <lgl>, RecordedDate <dtm>, ResponseId <chr>,
#   DistributionChannel <chr>, UserLanguage <chr>, consent <chr>, age_1 <dbl>,
#   gender <chr>, education <chr>, num_tabs_1 <dbl>, name_confidential <chr>,
#   number_confidential <chr>
```



# Cleaning data

- We filter the data to only include those who consented
- We remove survey preview responses
- (Optionally) remove responses that took place outside the relevant window.
- Remove confidential data (variables `name_confidential` and `number_confidential` in our survey, for example).

```
1 data.confidential <- data.raw |>
2   filter(consent == "Yes") |>
3   filter(Status != "Survey Preview") |>
4   filter(StartDate > QUALTRICS_STIME & EndDate < QUALTRICS_ETIME) |>
5   select(StartDate, EndDate, Status, Finished, RecordedDate,
6         ResponseId, consent, age_1, gender, education,
7         num_tabs_1, name_confidential, number_confidential)
8 data.clean <- data.confidential %>%
9   select(-name_confidential, -number_confidential)
```



# Cleaning data by selection

We could also simply **not select** the confidential data if we don't actually need it.

```
1 data.clean <- data.raw |>
2   filter(consent == "Yes") |>
3   filter(Status != "Survey Preview") |>
4   filter(StartDate > QUALTRICS_STIME & EndDate < QUALTRICS_ETIME) |>
5   select(StartDate, EndDate, Status, Finished, RecordedDate,
6         ResponseId, consent, age_1, gender, education, num_tabs_1)
```



# Using confidential data

We could also (hypothetically) immediately compute variables that rely on confidential data.

```
1 # not run
2 data.clean <- data.raw |>
3   filter(consent == "Yes") |>
4   filter(Status != "Survey Preview") |>
5   filter(StartDate > QUALTRICS_STIME & EndDate < QUALTRICS_ETIME) |>
6   select(StartDate, EndDate, Status, Finished, RecordedDate,
7         ResponseId, consent, age_1, gender, education, num_tabs_1,
8         gps_lat, gps_lon) |>
9   mutate(distance = compute_distance_from_cornell(
10    gps_lat, gps_lon, precision="100m")) |>
11   select(-gps_lat, -gps_lon)
```



# Saving confidential and clean data

- save the confidential data to a **clearly marked** folder (J-PAL policy: an encrypted volume)
- save the cleaned publishable data to a **well-defined** folder.

```
1 # save confidential data NOT for publishing, if needed
2 write.csv(data, file.path(confdatapath, "confidential_data.csv"),
3           row.names = FALSE)
4 saveRDS(data, file.path(confdatapath, "confidential_data.rds"))
5 # saving clean data for publishing
6 write.csv(data, file.path(cleandatapath, "clean_data.csv"),
7           row.names = FALSE)
```



# Descriptive statistics

Now you are ready use your cleaned data for reproducible analyses!

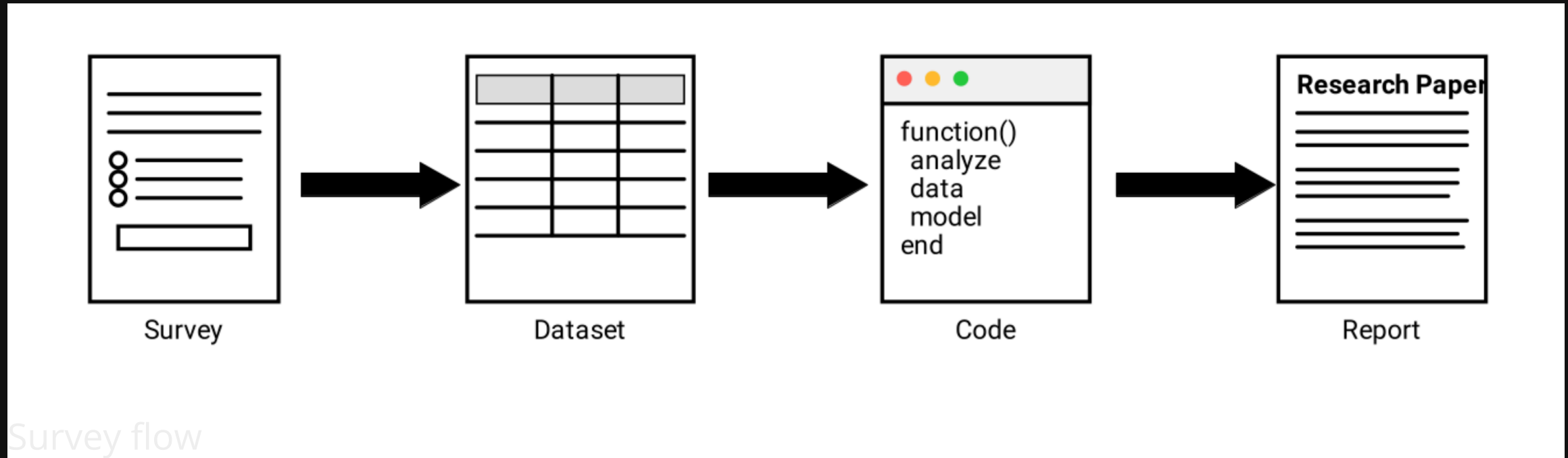
# Stepping it up



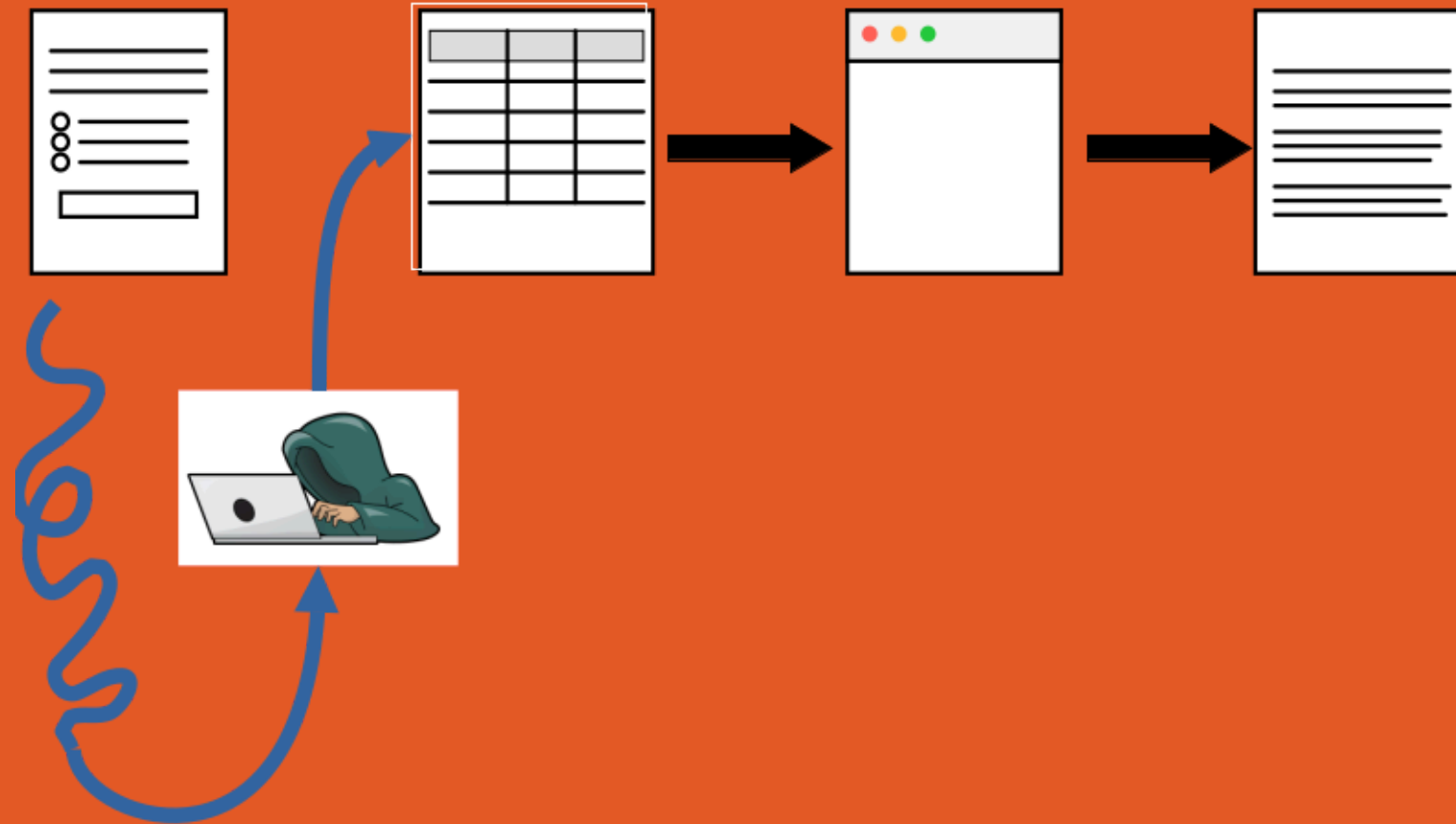
# Stepping up the process

- API,
- checksums,
- trusted systems
- preservation

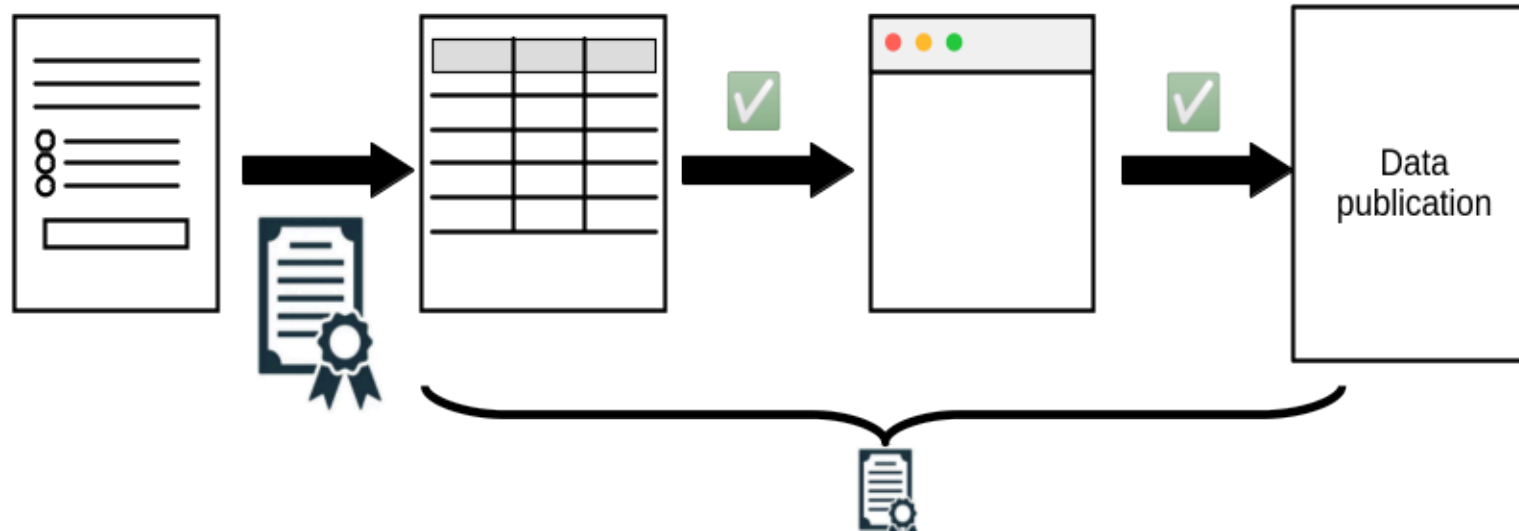
# Credibility of the data flow



# What could go wrong?



# How to CERTIFY the full process?



Survey flow

# Taking it a step further

- Survey tool provider (Qualtrics, etc.) exports data, posts checksum
- Survey tool provider exports data only to institution directly into trusted repository, researchers obtain data from there (with privacy protections)
- Has been discussed by authors behind Data Colada
- Don't hold your breath...

# Using automation



# APIs

- An API (**Application Programming Interface**) is a mechanism that enables two software components to communicate with each other
- APIs can be used to request data or services and get responses without needing to know how the other program works internally
- We will use APIs to streamline and **automate** the processing

# Loading data from Qualtrics using an API

We need to know a few things:

- the URL we want to use, defined by a generic part, and a survey specific part
- these are **public** - no need for secrecy

```
1 # qualtrics URL components
2 QUALTRICS_FULL_URL <- "first part of survey URL"
3
4 QUALTRICS_SURVEY <- "second part of survey URL, usually starts with SV"
```



# Loading data from Qualtrics using an API

We may want to limit the responses we download programmatically. This is not part of API, but of good programming practices.

```
1 # Keep only responses in the desired window of time
2 QUALTRICS_STIME <- ymd_hms("2025-07-01 00:00:01")
3 QUALTRICS_ETIME <- ymd_hms("2025-08-26 23:59:00")
```



# Fetching the data with the API

The API call *replaces the manual download* from before:

```
1 data.raw <- fetch_survey(surveyID = QUALTRICS_SURVEY,  
2                           verbose = TRUE)
```

# BUT: Privacy! Confidentiality!

Can anybody just download these data?

NO!

We need to authenticate, but not by entering a password manually.

That's where the **API token** comes in.

# Fetching the data with the API

We need to set an **API token**, then we can download this.

```
1 if (Sys.getenv("QUALTRICS_API_KEY") != "") {  
2   data.raw <- fetch_survey(surveyID = QUALTRICS_SURVEY, verbose = TRUE)  
3 } else {  
4   stop("Please set the QUALTRICS_API_KEY environment  
5   variable to your API key.")  
6 }
```



# The rest of the pipeline is unchanged

`data.raw` now comes from the API instead of a downloaded file — but the cleaning and saving steps from before are **exactly the same:**

```
1 clean_data <- data.raw |>
2   filter(consent == "Yes") |>
3   filter(Status != "Survey Preview") |>
4   filter(StartDate > QUALTRICS_STIME & EndDate < QUALTRICS_ETIME) |>
5   select(StartDate, EndDate, Status, Finished, RecordedDate,
6     ResponseId, consent, age_1, gender, education, num_tabs_1)
7
8 write.csv(clean_data, file.path(publicdata, "clean_data.csv"),
9   row.names = FALSE)
```



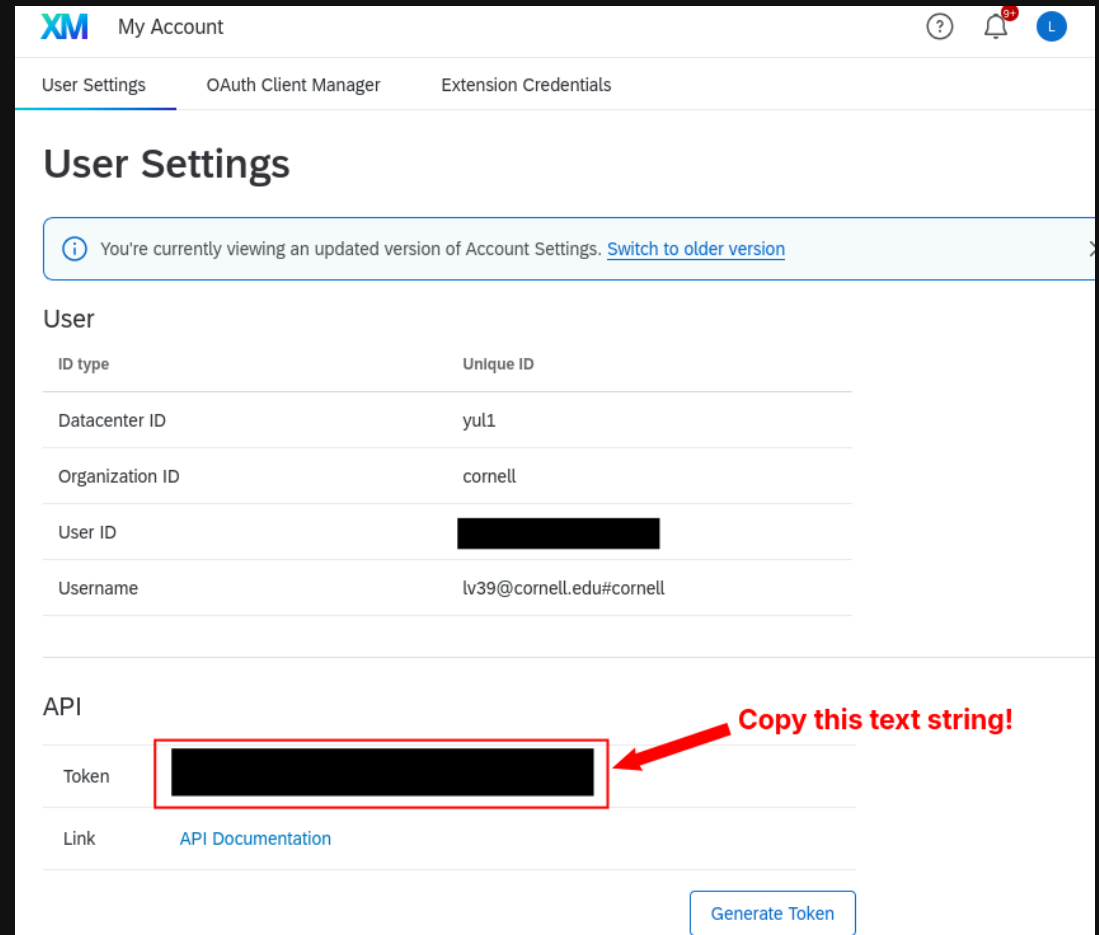
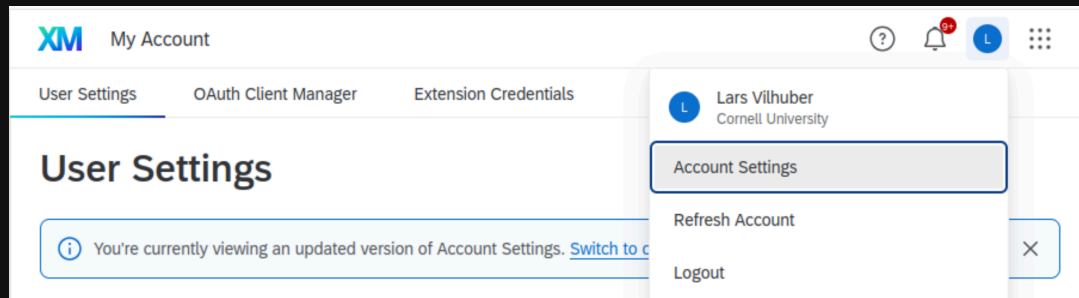
# Not unique to Qualtrics

- LimeSurvey -> `limer`
- SurveyCTO -> `rsurveycto`
- Google Forms -> `googlesheets4`

And of course works just fine in Python (and via Python, could use Stata!)

# Qualtrics and API tokens.

An API token is assigned to your **Qualtrics account**. Where do you find it?



# Setting API tokens

Not specific to the Qualtrics API!

- Set it manually:

```
1 Sys.setenv(QUALTRICS_API_KEY = "ab7ece8b")
```

- Set it using environment variables stored outside your code (e.g., in `.Renvirom` file) - **good for testing**

```
1 # This is .Renvirom
2 QUALTRICS_API_KEY="ab7ece8b"
```



# Setting API tokens

We want to automate on cloud servers!

- Push these “**secrets**” to `GitHub Secrets` and load it in `GitHub Actions` [[link](#)]

# Using API tokens

Now we need to make it available to our code (regardless of where it comes from)

```
1 # Here environment variables are read from .Renviron
2 QUALTRICS_API_KEY <- Sys.getenv("QUALTRICS_API_KEY")
```



# Full code

Now this works both **locally** and on **cloud servers** without any manual interaction!

```
1 QUALTRICS_FULL_URL <- "first part of survey URL"
2 QUALTRICS_SURVEY <- "second part of survey URL, usually starts with SV"
3
4 if (Sys.getenv("QUALTRICS_API_KEY") != "") {
5   data.raw <- fetch_survey(surveyID = QUALTRICS_SURVEY, verbose = TRUE)
6 } else {
7   stop("Please set the QUALTRICS_API_KEY environment
8   variable to your API key.")
9 }
```



# Side-note: Qualtrics API credentials

Qualtrics API credentials cannot be restricted to a single survey.

skip



# Traditional Static API Tokens (**X-API-TOKEN**)

- **The Limit:** You can only have **one** active static API token per user account at a time.
- **The Catch:** If you go to your account settings and generate a new token for a second application, it will **immediately overwrite and invalidate** the old token, breaking your first application.

# OAuth 2.0 Client Credentials

Separate, independent credentials for different applications:

- **Account Settings** > **Qualtrics IDs** > **OAuth Client Manager**.
- Click **Create Client** to generate unique sets of `Client ID` and `Client Secret` credentials.
- You can create **multiple clients** for different applications
- To revoke access for one app, you can delete its specific client without affecting the others.

# BUT: API Key access ALL your surveys

- Regardless of method, the API key can access ALL of your surveys!
- There is **NO** way to restrict that natively.

That's a problem.

# Workaround: “Service Account”

- You can create a specific user, say `jpal-survey-user`
- Requires support from system administrator
- Once the user is created, proceed as before, but when logged in as `jpal-survey-user`!
- As **YOU**, share only the specific survey with `jpal-survey-user` and give it only the permissions it needs

**It's a bit more complicated...**



# Secrets



# Secrets (Github version)

- You will want to keep APIs key safe using **GitHub Secrets**.
- Secrets allow you to store sensitive information in the repository environment.
- Use the secret as an environment variable in the GitHub workflow file.

# Storing secrets in `.Renvi` locally

You already have a `.Renvi` for local development:

```
1 QUALTRICS_API_KEY='something here'
```

- Do not publish this file!
- Do not commit it to Github!<sup>5</sup>

# Storing secrets in Github

- Enter them manually in the GitHub web interface
- Use the `.Renviron` file to set the GitHub Actions secrets with the **Github CLI**:

```
1 gh secret set -f .Renviron
```

```
1  
2 ✓ Set Actions secret DATAVERSE_TOKEN for  
3 ✓ Set Actions secret QUALTRICS_BASE_URL  
4 ✓ Set Actions secret DATAVERSE_SERVER fo  
5 ✓ Set Actions secret QUALTRICS_API_KEY f  
6 ✓ Set Actions secret DATAVERSE_DATASET_D
```



# Using secrets in GitHub Actions

In GitHub workflows, set your environment variables:

```
1 echo "QUALTRICS_API_KEY=${{ secrets.QUALTRICS_API_KEY }}" >> $GITHUB_ENV
```



# Using secrets in Github Actions

- R code does **not** need to be adapted!

```
1 # Same R code as before!  
2 if (Sys.getenv("QUALTRICS_API_KEY") != "") {  
3   data.raw <- fetch_survey(surveyID = QUALTRICS_SURVEY, verbose = TRUE)  
4 } else {  
5   stop("Please set the QUALTRICS_API_KEY environment  
6   variable to your API key.")  
7 }
```



# Checksums



# Checksums can be used to demonstrate consistency

- A **checksum** is a single value calculated from a data file that can be used to verify the integrity of the file.
- The `sha256` algorithm is commonly used for this purpose.

# Checksums can be used to demonstrate consistency

- In R, the package `digest`<sup>6</sup> is available.

```
1 library(digest)
2 # Calculate the checksum of a file
3 digest(trees)
```

```
[1] "370a7132861fb520bd721d9bcbe008a4"
```

```
1 digest(trees, algo="sha256")
```

```
[1] "70823d6c0cebdc582b388f3ec56930bd2ec5dd176272032d949b93319d74d17b"
```



# Adding checksums to the data download

```
1 if (Sys.getenv("QUALTRICS_API_KEY") != "") {  
2   data.raw <- suppressMessages(fetch_survey(surveyID = QUALTRICS_SURVEY, verbose = FALSE))  
3   data.raw.sha256 <- digest::digest(data.raw, algo = "sha256")  
4   message("Checksum of the downloaded data: ", data.raw.sha256)  
5   # Write checksum to a file  
6   writeLines(data.raw.sha256, file.path(metadata.path, "data.raw.sha256"))  
7 } else {  
8   stop("Please set the QUALTRICS_API_KEY environment  
9   variable to your API key.")  
10 }
```

Checksum of the downloaded data:

a0e2146acc752debff66a670fe654c7618c45bacf6b8c634e58a21d5999fd222



# How does that help?

- Subsequent downloads can verify that the download is the same as originally downloaded!

```
1 # Read the original checksum from file
2 original.sha256 <- readLines(file.path(metadata.path, "data.raw.sha256"))
3 message("Original checksum from file: ", original.sha256)
```

Original checksum from file:

a0e2146acc752debff66a670fe654c7618c45bacf6b8c634e58a21d5999fd222

```
1 # Redownload data
2 if (Sys.getenv("QUALTRICS_API_KEY") != "") {
3   data.raw <- suppressMessages(fetch_survey(surveyID = QUALTRICS_SURVEY, verbose = FALSE))
4   data.raw.sha256 <- digest::digest(data.raw, algo = "sha256")
5   message("Checksum of the downloaded data: ", data.raw.sha256)
6 }
```

Checksum of the downloaded data:

a0e2146acc752debff66a670fe654c7618c45bacf6b8c634e58a21d5999fd222



# How does that help?

- Subsequent downloads can verify that the download is the same as originally downloaded!

```
1 # Compare the checksums
2 if (original.sha256 == data.raw.sha256) {
3   message("Checksums match! Data integrity verified.")
4 } else {
5   warning("Checksums do NOT match! Data may have changed/ been altered.")
6 }
```

Checksums match! Data integrity verified.

# Preservation



# Credibility of Survey Data

- You run a study using the **PSID**. Do you **trust** the downloaded data?

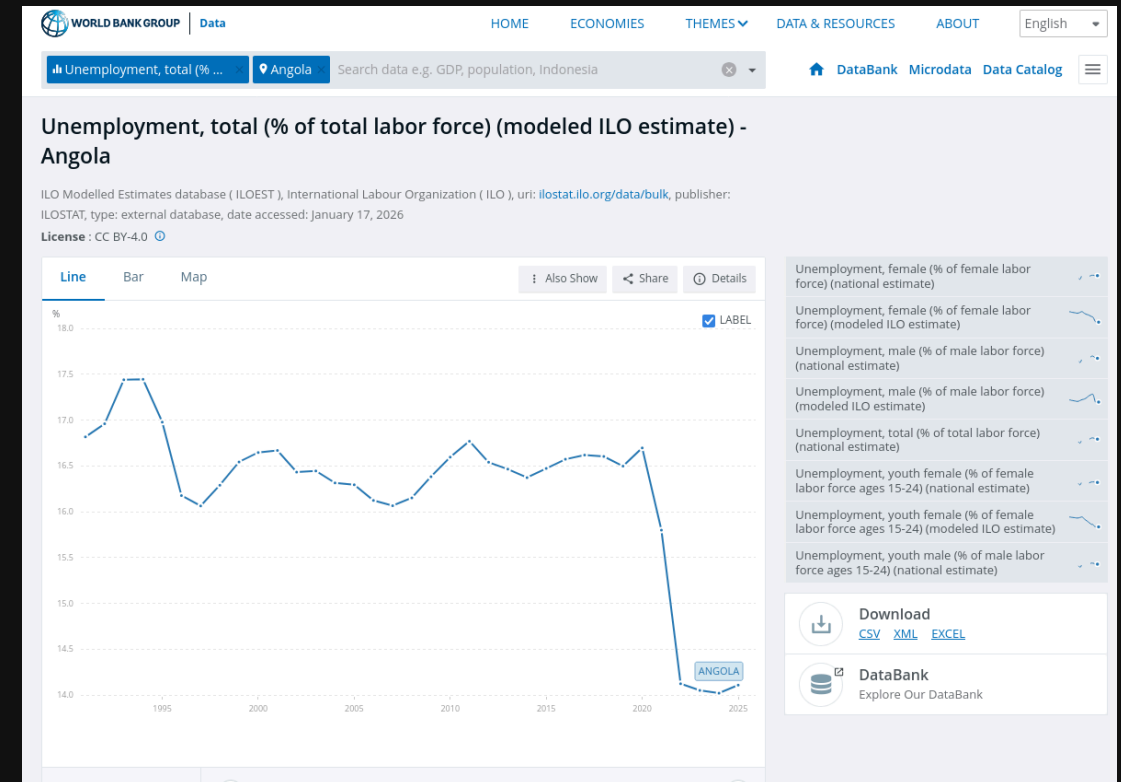


The screenshot shows the PSID Variable Search interface. At the top, there is a logo for the Institute for Social Research and Survey Research Center, along with the text 'INSTITUTE FOR SOCIAL RESEARCH • SURVEY RESEARCH CENTER' and 'PANEL STUDY OF INCOME DYNAMICS'. Below the logo is a navigation menu with links for 'GETTING STARTED', 'STUDIES', 'DOCUMENTATION', 'DATA', 'PUBS & MEETINGS', 'PEOPLE', and 'NEWS'. A search bar is located below the navigation menu, with the text 'Search name, label, question, and explanation text of all variables' and a 'Search' button. Below the search bar are radio buttons for 'Any words (OR)', 'All words (AND)', and 'Exact phrase'. The main content area is a table with three columns: 'DATA TYPE', 'YEARS', and 'SECTION OF CODEBOOK'. The 'DATA TYPE' column lists various survey components, the 'YEARS' column lists years from 1968 to 2024, and the 'SECTION OF CODEBOOK' column lists search criteria. A 'Quick Links' sidebar is visible on the right side of the page.

DATA TYPE	YEARS	SECTION OF CODEBOOK
<input type="checkbox"/> PSID Family-level	<input type="checkbox"/> 1968 <input type="checkbox"/> 1981 <input type="checkbox"/> 1994 <input type="checkbox"/> 2014	<input type="radio"/> Question or explanation text
<input type="checkbox"/> PSID Individual-level	<input type="checkbox"/> 1969 <input type="checkbox"/> 1982 <input type="checkbox"/> 1995 <input type="checkbox"/> 2015	<input type="radio"/> Variable label
<input type="checkbox"/> Child Development Supplement	<input type="checkbox"/> 1970 <input type="checkbox"/> 1983 <input type="checkbox"/> 1996 <input type="checkbox"/> 2016	<input type="radio"/> Variable name
<input type="checkbox"/> Child Development Supplement Time Diaries	<input type="checkbox"/> 1971 <input type="checkbox"/> 1984 <input type="checkbox"/> 1997 <input type="checkbox"/> 2017	<input checked="" type="radio"/> All
<input type="checkbox"/> Transition into Adulthood Supplement	<input type="checkbox"/> 1972 <input type="checkbox"/> 1985 <input type="checkbox"/> 1999 <input type="checkbox"/> 2019	
<input type="checkbox"/> Family History	<input type="checkbox"/> 1973 <input type="checkbox"/> 1986 <input type="checkbox"/> 2001 <input type="checkbox"/> 2020	
<input type="checkbox"/> Disability and Use of Time	<input type="checkbox"/> 1974 <input type="checkbox"/> 1987 <input type="checkbox"/> 2002 <input type="checkbox"/> 2021	
<input type="checkbox"/> Childhood Retrospective Circumstances Study	<input type="checkbox"/> 1975 <input type="checkbox"/> 1988 <input type="checkbox"/> 2003 <input type="checkbox"/> 2023	
<input type="checkbox"/> Family Rosters and Transfers	<input type="checkbox"/> 1976 <input type="checkbox"/> 1989 <input type="checkbox"/> 2005 <input type="checkbox"/> 2024	
<input type="checkbox"/> Wellbeing and Daily Life	<input type="checkbox"/> 1977 <input type="checkbox"/> 1990 <input type="checkbox"/> 2007	
<input type="checkbox"/> Family Relationship Matrix	<input type="checkbox"/> 1978 <input type="checkbox"/> 1991 <input type="checkbox"/> 2009	
	<input type="checkbox"/> 1979 <input type="checkbox"/> 1992 <input type="checkbox"/> 2011	

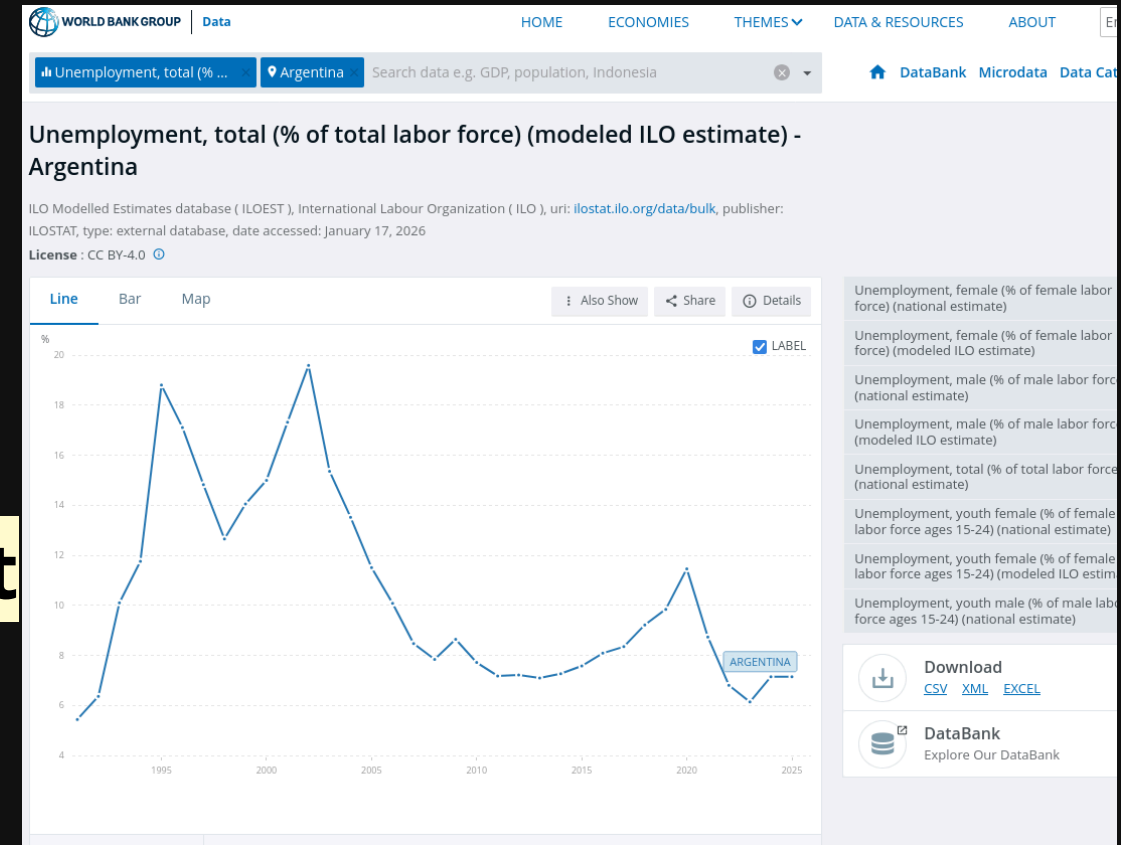
# Credibility of Government Data

- You run a study using the PSID. Do you trust the downloaded data?
- You use unemployment data for **Angola** through **World Bank Data Bank**. Do you **trust** the downloaded data?



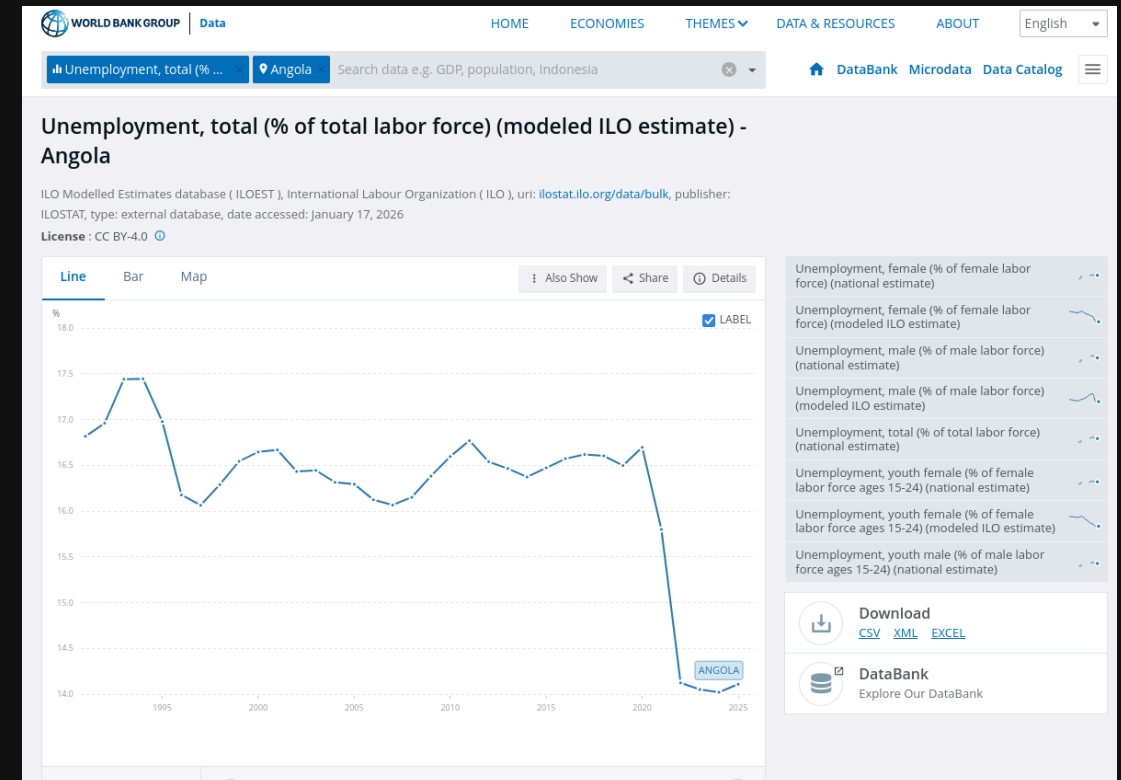
# Credibility of Government Data

- You run a study using the PSID. Do you trust the downloaded data?
- You use unemployment data for **Argentina** through World Bank Data Bank. Do you **trust** the downloaded data?



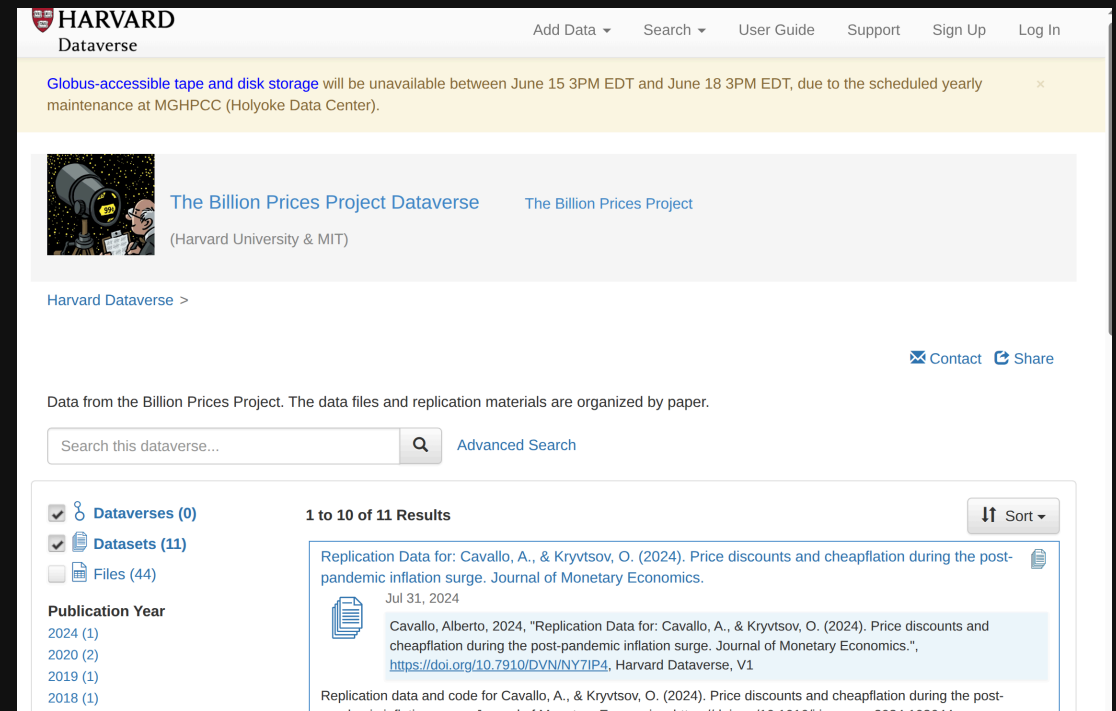
# Credibility of Government Data

- You run a study using the PSID. Do you trust the downloaded data?
- You use unemployment data for **United States** through World Bank Data Bank. Do you **trust** the downloaded data?



# Credibility of Researcher provided data

- You run a study using the PSID. Do you trust the downloaded data?
- You use inflation data for **Argentina** through a **research deposit on Dataverse**. Do you trust the downloaded data?



The screenshot shows the Harvard Dataverse interface. At the top, there's a navigation bar with "Add Data", "Search", "User Guide", "Support", "Sign Up", and "Log In". A yellow banner below the navigation bar states: "Globus-accessible tape and disk storage will be unavailable between June 15 3PM EDT and June 18 3PM EDT, due to the scheduled yearly maintenance at MGHPCC (Holyoke Data Center)." The main content area features the "The Billion Prices Project Dataverse" by Harvard University & MIT. Below this, there's a search bar with the text "Search this dataverse..." and a search icon. To the right of the search bar is a link for "Advanced Search". On the left side, there's a filter menu with "Dataverses (0)", "Datasets (11)", and "Files (44)". Below the filter menu is a "Publication Year" section with a list: "2024 (1)", "2020 (2)", "2019 (1)", and "2018 (1)". The main results area shows "1 to 10 of 11 Results" and a "Sort" dropdown. The first result is "Replication Data for: Cavallo, A., & Kryvtsov, O. (2024). Price discounts and cheapflation during the post-pandemic inflation surge. Journal of Monetary Economics." with a date of "Jul 31, 2024". Below the title is a document icon and the text: "Cavallo, Alberto, 2024, 'Replication Data for: Cavallo, A., & Kryvtsov, O. (2024). Price discounts and cheapflation during the post-pandemic inflation surge. Journal of Monetary Economics.', https://doi.org/10.7910/DVN/NY7IP4, Harvard Dataverse, V1". At the bottom of the result is another line of text: "Replication data and code for Cavallo, A., & Kryvtsov, O. (2024). Price discounts and cheapflation during the post-pandemic inflation surge. Journal of Monetary Economics. https://doi.org/10.1016/j.jmonecon.2024.103644".

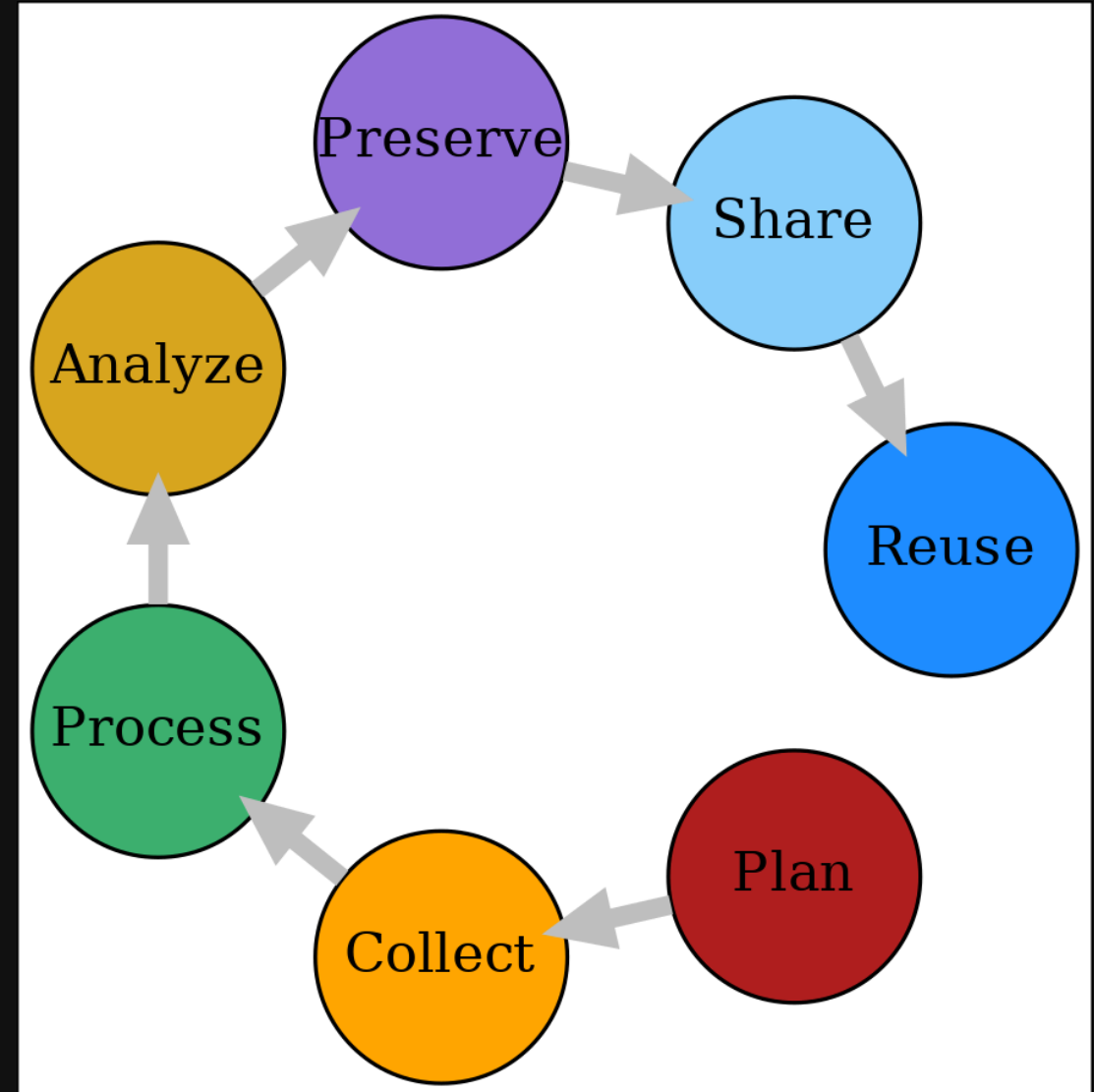
**Let's make you become a mini-  
PSID**



# Timing

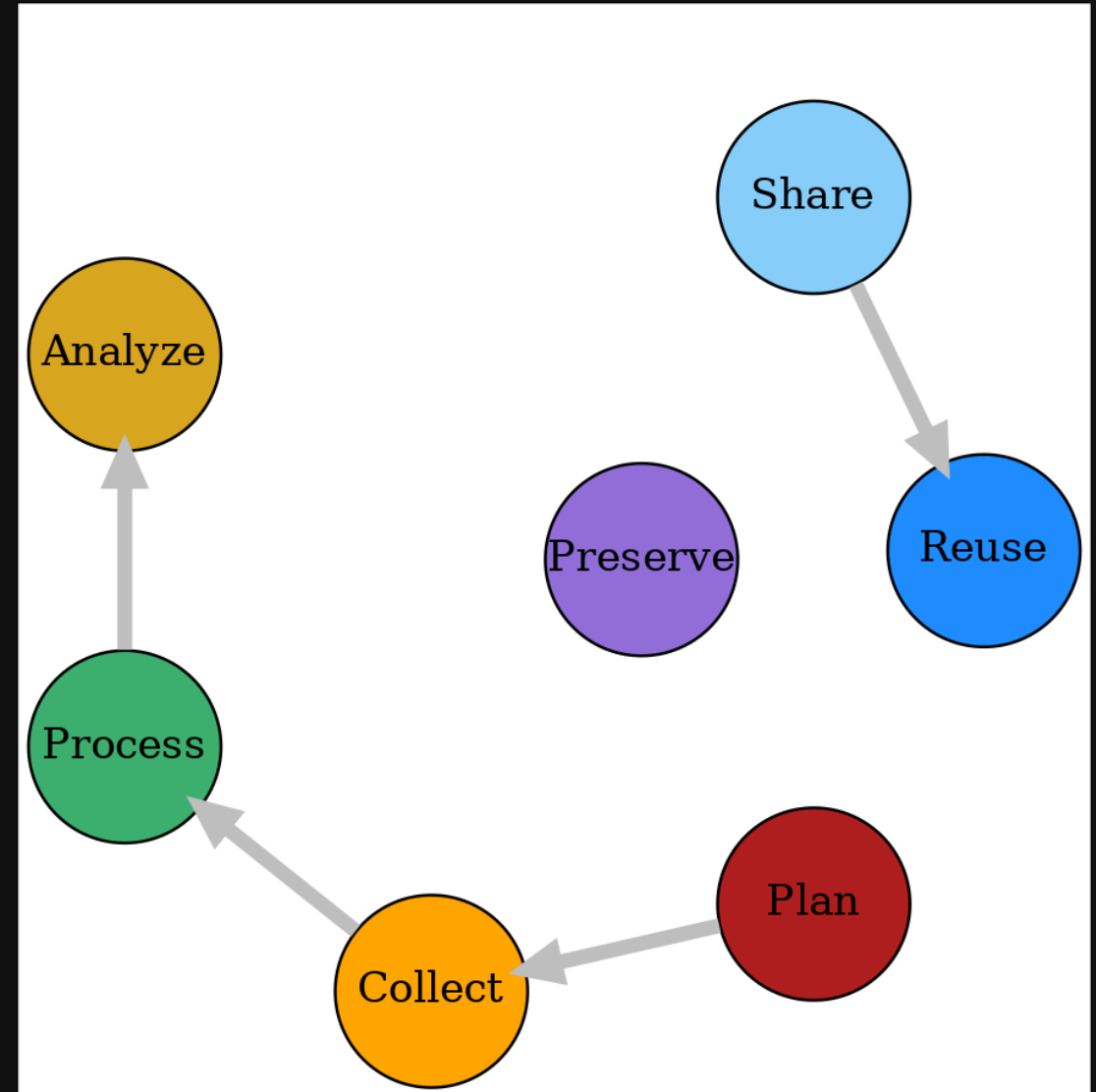
Once you have registered your analysis plan - should the **processing and analysis** really change?

Once you have **collected** the data - is it really going to change?



# Modified Data and Workflow

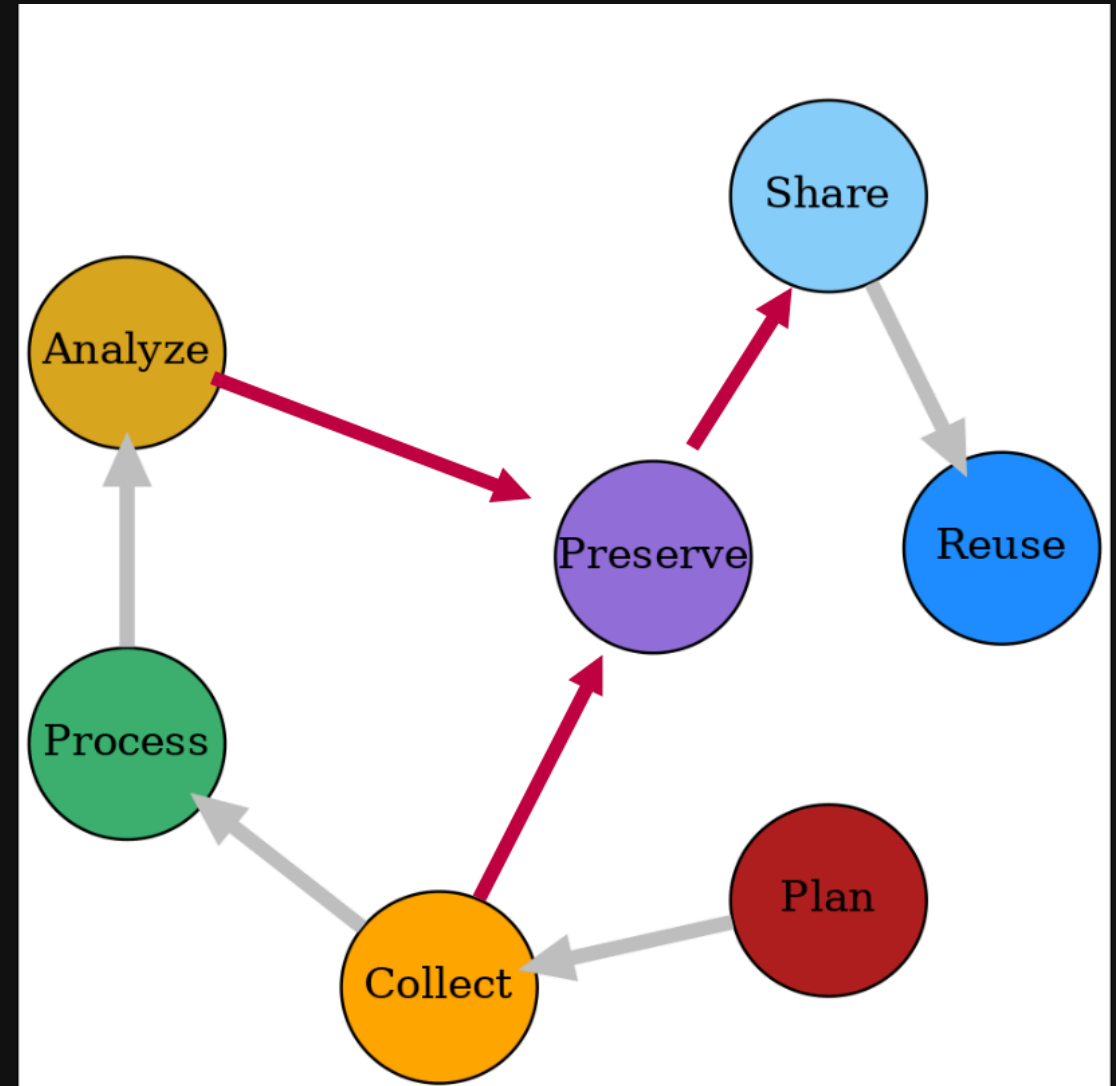
Let's consider the preservation part separately:



# Modified Data and Workflow

Proposal:

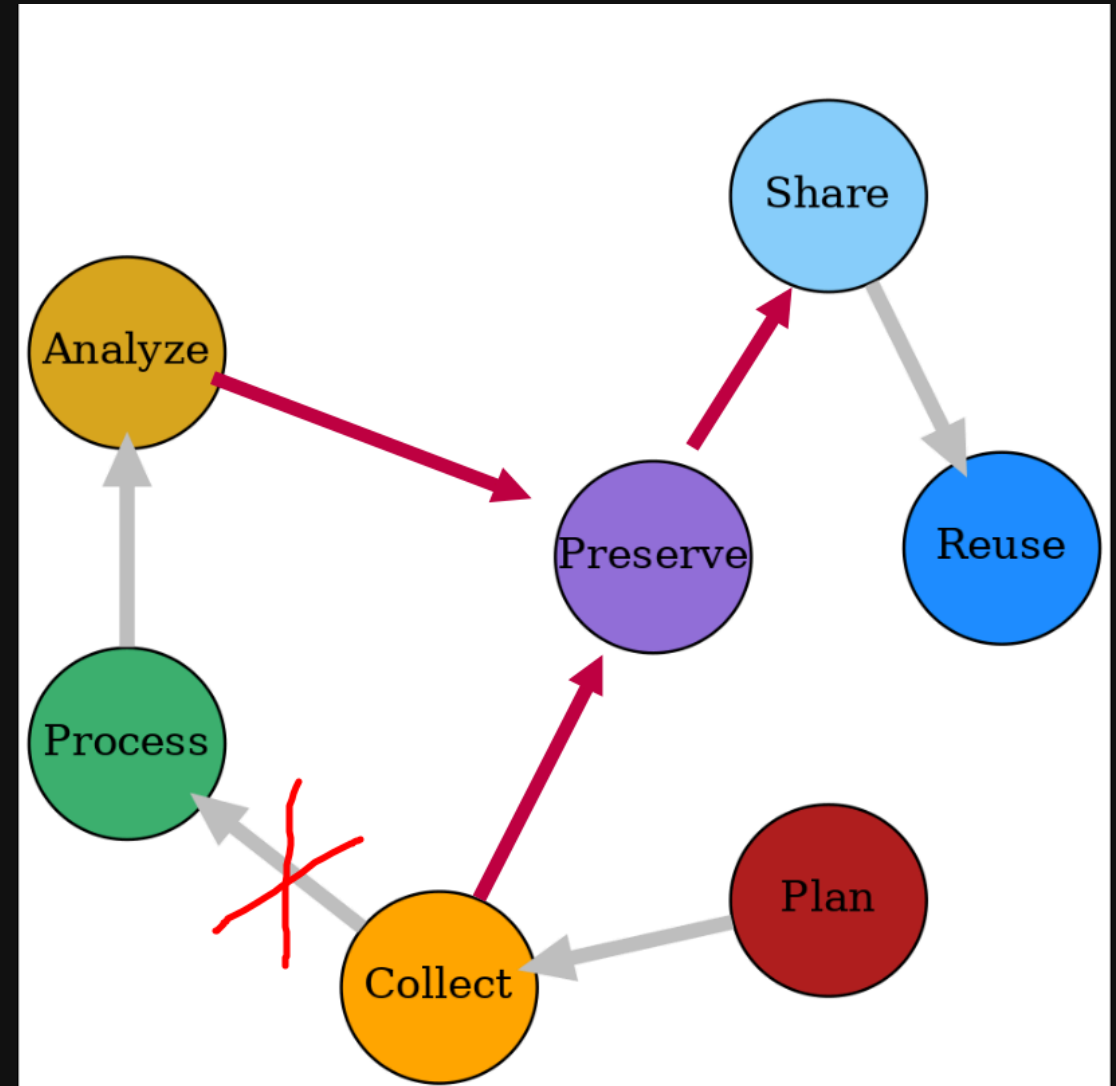
- Preserve as you go



# Modified Data and Workflow

Proposal:

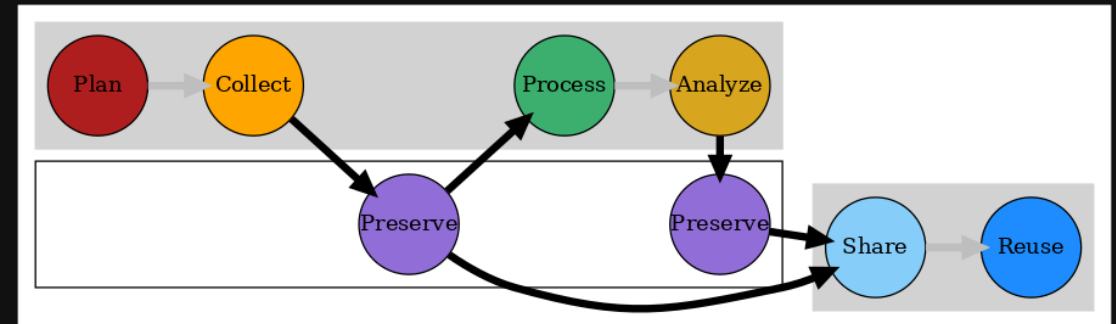
- Preserve as you go
- Use what you preserve



# Modified Data and Workflow

Proposal:

- Preserve as you go
- Use what you preserve



**Note: Doubtful ethics of others...**

**I don't want to be scooped!**

Thus, I'm not going to  
publish my raw data just  
yet!

# What is preservation



# Preservation

- Preservation != **publication**, != **sharing**
- In fact, preservation may mean: **not very accessible** at all!
- Preservation is intended to maintain data for tens, even hundreds of years
  - Preservation may involve curation: active transformation of the data for improved accessibility



# What is publication

**Publication** typically involves making information about the data, as well as the data themselves, available to others.

- Publication can initially mean that only **metadata** (information about the data) is published
- In some cases, it may be that **only** metadata is ever published
- But the metadata will point to how to access the data, how long the data will be preserved, and other salient facts

# This all seems so complicated

- I need to preserve my data for decades!
- I need to manage the application process for decades!
- Where do I get that DOI thing?
- How to I get Google to index my data?

# Options for Preservation (1)

## Trusted Repositories

Journals and institutions have assessed a number of trusted repositories:

- [CoreTrustSeal](#) has a certification process
- [re3data.org](#) lists research data repositories
- [Nature](#), [F1000Research](#), and [PLOS](#) have lists of trusted repositories.
- Always check with your journal for specific restrictions or suggestions.

# Options for Preservation (2)

## Trusted Repositories

- These generally include at least the following:
  - Dryad Digital Repository
  - figshare
  - Harvard Dataverse
  - ICPSR and OPENICPSR
  - Open Science Framework
  - Zenodo
  - Country or region-specific repositories (that nevertheless generally accept depositors from anywhere): **GESIS** (Germany), **Swedish National Data Service (SND)**, **EASY** (Netherlands), **CSIRO** (Australia), etc.
- Many universities have formal document repositories that may be able to assume such a role; talk to your (data) librarian



# What are NOT options for preservation

- Github, Gitlab, Bitbucket, etc.
- Dropbox, Box.com, Google Drive, etc.
- Your personal website
- Your university's departmental website



404. That's an error.

The requested URL /a\_cool\_website was not found on this server. That's all we know.



# 404

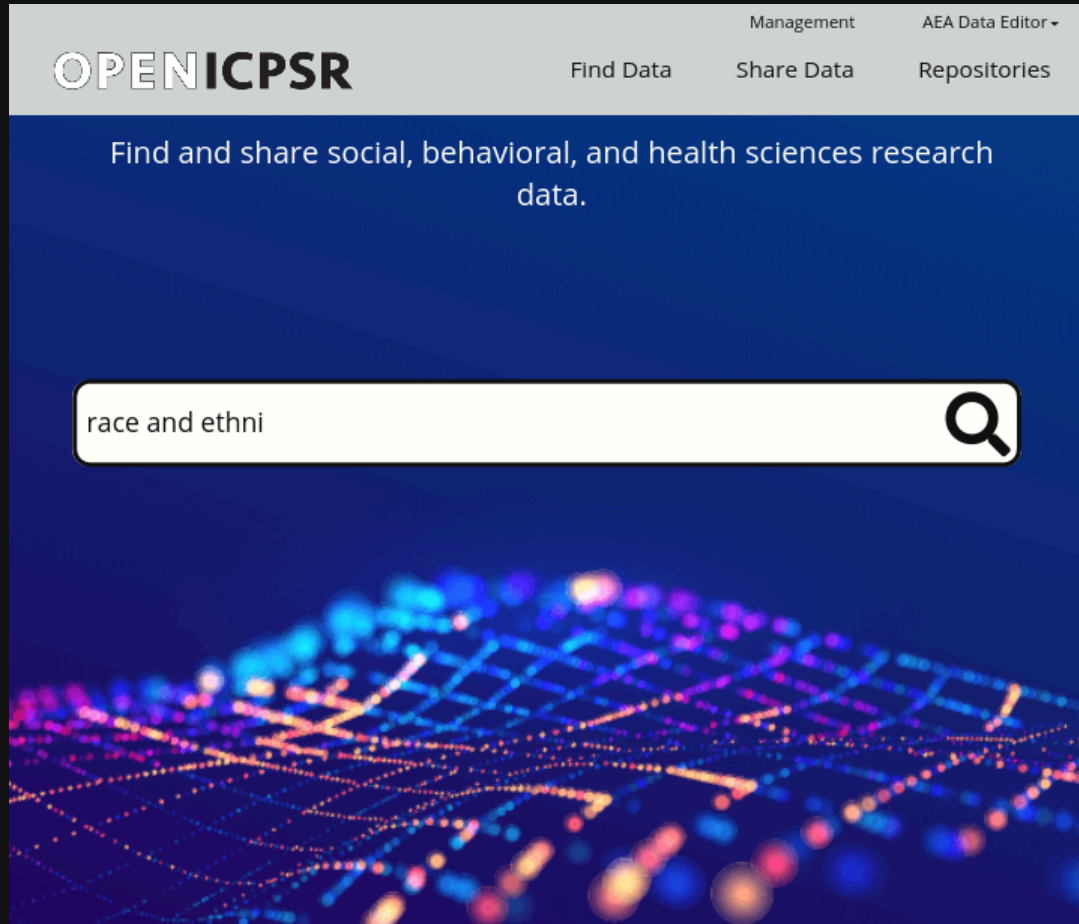
**File not found**

The site configured at this address does not contain the requested file.

If this is your site, make sure that the filename case matches the URL.  
For root URLs (like <http://example.com/>) you must provide an `index.html` file.

# Options for Preservation

In one of my day jobs:



The screenshot shows the OPENICPSR website interface. At the top left is the logo "OPENICPSR". To its right are navigation links: "Management", "AEA Data Editor", "Find Data", "Share Data", and "Repositories". Below the navigation is a blue banner with the text "Find and share social, behavioral, and health sciences research data." A search bar is centered on the banner, containing the text "race and ethni" and a magnifying glass icon. The background of the banner features a glowing blue and purple network of dots and lines. At the bottom of the page, there are logos for "AMERICAN ECONOMIC", "AERA", "IEH", and "NIDA".



# Options for Preservation with API

Dataverse

<https://demo.dataverse.org/dataverse/larstest>

The screenshot shows the Dataverse interface. At the top, there's a navigation bar with 'Add Data', 'Search', 'About', 'User Guide', 'Support', and 'Lars Vilhuber'. Below that, a banner for the 2025 Dataverse Community Meeting is visible. The main content area shows 'Lars Vilhuber Dataverse' (Cornell University) and a search bar. A search result is displayed for 'Summer School Data', dated May 26, 2025, with a draft status. The author is listed as Vilhuber, Lars.

Also Zenodo <https://zenodo.org>

The screenshot shows the Zenodo website. The top navigation bar includes 'Upload', 'Communities', 'Log in', and 'Sign up'. The main content area features 'Featured communities' with a highlighted entry for the 'National COVID Cohort Collaborative (N3C)'. Below this, the 'Recent uploads' section shows a dataset titled 'Gene Ontology Data Archive' uploaded on September 2, 2021. A 'Need help?' section is also visible on the right side.



# Getting started on Dataverse

We will NOT use the regular Dataverse; rather, we will test on the demo site.

- This also works with Zenodo: <https://sandbox.zenodo.org/>
- Check your URL bar! There's often no other indication that this is not the real Zenodo or Dataverse!

# A tutorial of sorts

- Demo Dataverse for Lars

<https://demo.dataverse.org/dataverse/larstest>

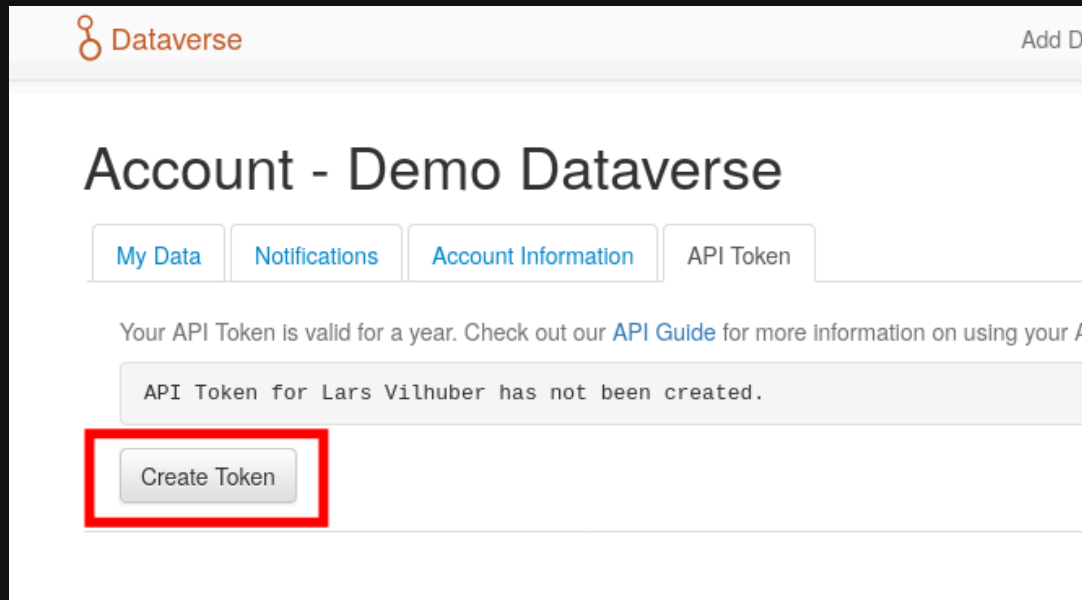
# Remember the API tokens?

```
1 QUALTRICS_API_KEY='something here'  
2 QUALTRICS_BASE_URL='url goes here'  
3 DATAVERSE_TOKEN='token goes here'  
4 DATAVERSE_SERVER='https://demo.dataverse.org'  
5 DATAVERSE_DATASET_DOI='doi goes here'
```

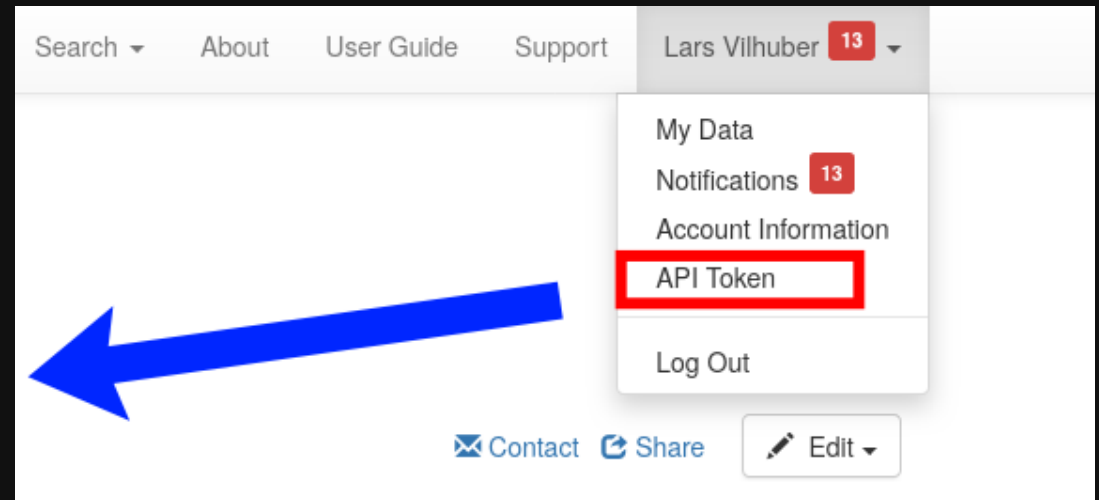
We're going to need the last three here!



# Getting your API keys from Dataverse

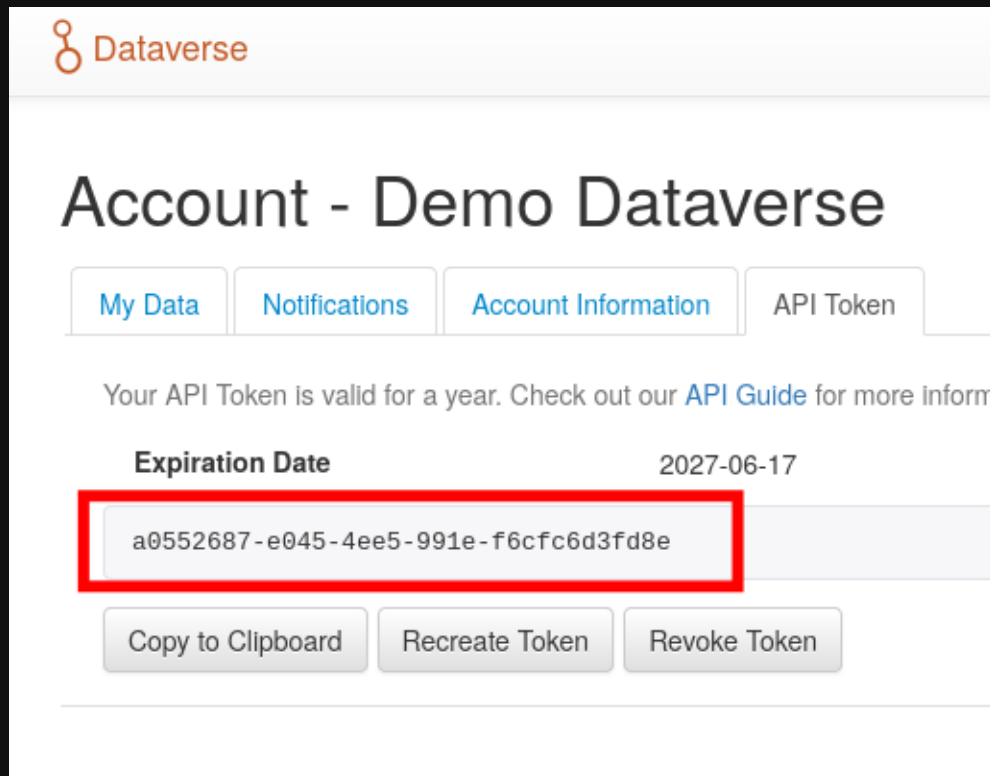


The screenshot shows the 'Account - Demo Dataverse' page. At the top left is the Dataverse logo. Below it, the title 'Account - Demo Dataverse' is displayed. A navigation bar contains four tabs: 'My Data', 'Notifications', 'Account Information', and 'API Token'. Below the tabs, a message states: 'Your API Token is valid for a year. Check out our [API Guide](#) for more information on using your API Token.' Below this message is a text box containing 'API Token for Lars Vilhuber has not been created.' At the bottom left, a 'Create Token' button is highlighted with a red rectangle.



The screenshot shows the user profile dropdown menu for 'Lars Vilhuber' with 13 notifications. The menu items are: 'My Data', 'Notifications 13', 'Account Information', 'API Token' (highlighted with a red rectangle), and 'Log Out'. A large blue arrow points from the 'API Token' menu item towards the left, indicating the next step in the process. At the bottom right, there are links for 'Contact', 'Share', and 'Edit'.

# Adding your API key to your `.Renviron`



The screenshot shows the 'Account - Demo Dataverse' page. The 'API Token' tab is selected. The page displays the expiration date as '2027-06-17' and the API token as 'a0552687-e045-4ee5-991e-f6cfc6d3fd8e', which is highlighted with a red box. Below the token are three buttons: 'Copy to Clipboard', 'Recreate Token', and 'Revoke Token'.

**Expiration Date** 2027-06-17

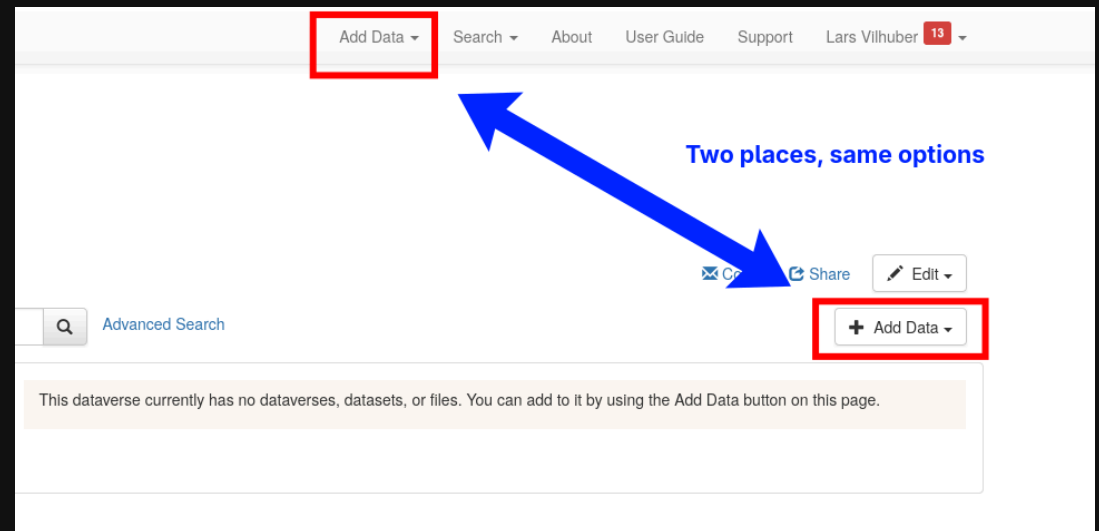
a0552687-e045-4ee5-991e-f6cfc6d3fd8e

Copy to Clipboard Recreate Token Revoke Token

```
1 QUALTRICS_API_KEY='something here'  
2 QUALTRICS_BASE_URL='url goes here'  
3 DATAVERSE_TOKEN='token goes here'  
4 DATAVERSE_SERVER='https://demo.dataverse  
5 DATAVERSE_DATASET_DOI='doi goes here'
```

# We need a “container”

- Dataverse calls this a “dataset”.
- A “dataset” can hold multiple files.
- While this **can be created** via the API, I suggest doing it **manually** (once per project)



# Fill in metadata

Demo Dataverse > Lars Vilhuber Dataverse >

**Host Dataverse** ⓘ

Changing the host dataverse will clear any fields you may have entered data into.

Lars Vilhuber Dataverse

**Dataset Template** ⓘ

Changing the template will clear any fields you may have entered data into.

None

\*Asterisks indicate required fields

**Citation Metadata** ⌵

**Title** ⓘ

Survey of Browser Tabs

Add "Replication Data for" to Title

**Author** ⓘ

**Name** ⓘ

Vilhuber, Lars

**Affiliation** ⓘ

Cornell University

**Point of Contact** ⓘ

**Name** ⓘ

Vilhuber, Lars

**Affiliation** ⓘ

Cornell University

**E-mail** ⓘ

lars.vilhuber@cornell.edu

**Description** ⓘ

This field supports only certain [HTML tags](#).

**Text** ⓘ

These data were collected as part of an exercise in transparent and automated survey processing.

**Annotations:**

- Blue box around Host Dataverse and Dataset Template with arrow: **Probably fine to leave untouched**
- Red box around Author, Point of Contact, and Description with arrow: **Some pre-filled, some need to be filled**

**Deposit Date** ⓘ

2026-06-17

**Files**

Multiple file upload/download methods are available for this dataset. Once you upload a file using one of these methods,

**Upload with HTTP via your browser** ⌵

Select files or drag and drop into the upload widget. Total size of Collection: 2.2 KB. Maximum of 1,000 files per upload limited to 100.0 MB. Ingest is limited to the following file sizes based on their format: rdata: disabled.

**+ Select Files to Add**

**Leave empty!** Drag and drop files here.

An option to upload a folder of files will be enabled after this dataset is created.

**Metadata Tip:** After adding the dataset, click the Edit Dataset button to add more metadata.

**Save Dataset** Cancel

**Annotations:**

- Red box around the file upload area: **Leave empty!**
- Red box around the Save Dataset button: **Save Dataset**



# Uploading data to Dataverse

- You **could** upload data manually, but this is about automation!
- Now that the “container” is ready, we can upload data to it via the API.

# Getting the Identifiers

```

1 QUALTRICS_API_KEY='something here'
2 QUALTRICS_BASE_URL='url goes here'
3 DATAVERSE_TOKEN='token goes here'
4 DATAVERSE_SERVER='https://demo.dataverse
5 DATAVERSE_DATASET_DOI='doi goes here'

```

Lars Vilhuber Dataverse  
(Cornell University)

Demo Dataverse > Lars Vilhuber Dataverse >

Success! -This dataset has been created.

Info -This draft version needs to be published. When ready for sharing, please publish it so that others can see these changes.

## Survey of Browser Tabs

Draft Unpublished

Vilhuber, Lars, 2026, "Survey of Browser Tabs", <https://doi.org/10.70122/FK2/EMAWKA>, Demo Dataverse, DRAFT VERSION

Cite Dataset - Learn about [Data Citation Standards](#).

Publish Dataset -  
Edit Dataset -  
Curation Status -  
Link Dataset  
Contact Owner Share

Description - These data were collected as part of an exercise in transparent and automated survey processing.

Subject - Social Sciences

License/Data Use Agreement CC0 1.0

Dataset Metrics -  
0 Downloads

**We need the DOI**

# Additional controls

Lars Vilhuber Dataverse  
(Cornell University)

Demo Dataverse > Lars Vilhuber Dataverse >

Success! –This dataset has been created.

Info –This draft version needs to be published. When ready for sharing, please **publish** it so that others can see these changes.

## Survey of Browser Tabs

Draft Unpublished

Vilhuber, Lars, 2026, "Survey of Browser Tabs", <https://doi.org/10.70122/FK2/EMAWKA>, Demo Dataverse, DRAFT VERSION

Cite Dataset - Learn about [Data Citation Standards](#).

**Publish Dataset** -  
Edit Dataset -  
Curation Status -  
Link Dataset  
Contact Owner Share

Description - These data were collected as part of an exercise in transparent and automated survey processing.

Subject - Social Sciences

License/Data Use Agreement PUBLIC DOMAIN CC0 1.0

Dataset Metrics - 0 Downloads

**This is where we control WHEN it becomes public!**

Lars Vilhuber Dataverse  
(Cornell University)

Demo Dataverse > Lars Vilhuber Dataverse >

Success! –This dataset has been created.

Info –This draft version needs to be published. When ready for sharing, please **publish** it so that others can see these changes.

## Survey of Browser Tabs

Draft Unpublished

Vilhuber, Lars, 2026, "Survey of Browser Tabs", <https://doi.org/10.70122/FK2/EMAWKA>, Demo Dataverse, DRAFT VERSION

Cite Dataset - Learn about [Data Citation Standards](#).

**Publish Dataset** -  
Edit Dataset -  
Curation Status -  
Link Dataset  
Contact Owner Share

Description - These data were collected as part of an exercise in transparent and automated survey processing.

Subject - Social Sciences

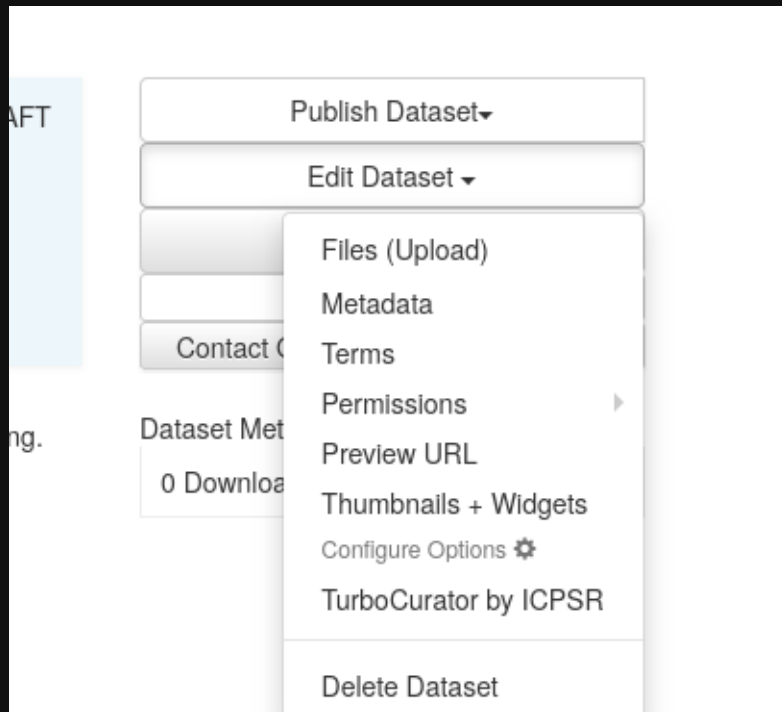
License/Data Use Agreement PUBLIC DOMAIN CC0 1.0

Dataset Metrics - 0 Downloads

**This is where we control who can access the data when published!**



# Licensing and permissions



- **Terms**: For people who wish to download published files
- **Permissions**: For fine-grained access over who can do what before publication

# Permissions

Users/Groups <sup>▲</sup> All the users and groups that have access to your dataverse.

[Assign Roles to Users/Groups](#)

2 Users/Groups

User/Group Name (Affiliation) <sup>▾</sup>	ID <sup>▾</sup>	Role <sup>▾</sup>	Action
Lars Vilhuber (Cornell University)	@lv39	Admin	<i>Role assigned at Lars Vilhuber Dataverse</i>
Lars Vilhuber (Cornell University)	@lv39	Contributor	<a href="#">✕ Remove Assigned Role</a>

Roles <sup>▲</sup> All the roles set up in your dataverse, that you can assign to users and groups.

**Admin** - A person who has all permissions for dataverses, datasets, and files, including approving requests for restricted data.

[AddDataverse](#) [AddDataset](#) [ViewUnpublishedDataverse](#) [ViewUnpublishedDataset](#) [DownloadFile](#) [EditDataverse](#) [EditDataset](#) [ManageDataversePermissions](#) [ManageDatasetPermissions](#) [ManageFilePermissions](#) [PublishDataverse](#) [PublishDataset](#) [LinkDataverse](#) [LinkDataset](#) [DeleteDataverse](#) [DeleteDatasetDraft](#)

**Contributor** - For datasets, a person who can edit License + Terms, and then submit them for review.

[ViewUnpublishedDataset](#) [DownloadFile](#) [EditDataset](#) [DeleteDatasetDraft](#)

**Curator** - For datasets, a person who can edit License + Terms, edit Permissions, and publish and link datasets.

[AddDataverse](#) [AddDataset](#) [ViewUnpublishedDataverse](#) [ViewUnpublishedDataset](#) [DownloadFile](#) [EditDataset](#) [ManageDatasetPermissions](#) [ManageFilePermissions](#) [PublishDataset](#) [LinkDataset](#) [DeleteDatasetDraft](#)

**File Downloader** - A person who can download a published file.

[DownloadFile](#)

**Member** - A person who can view both unpublished dataverses and datasets.

[ViewUnpublishedDataverse](#) [ViewUnpublishedDataset](#) [DownloadFile](#)

You can designate

- who can upload
- who can edit metadata
- who can publish



# Terms and Licenses

- **Licenses** are broad permissions on how to re-use
- Often CC-BY, see <https://creativecommons.org/licenses/>
- **Terms** are more restrictive. Do
  - need to contact somebody
  - need to sign a data use agreement

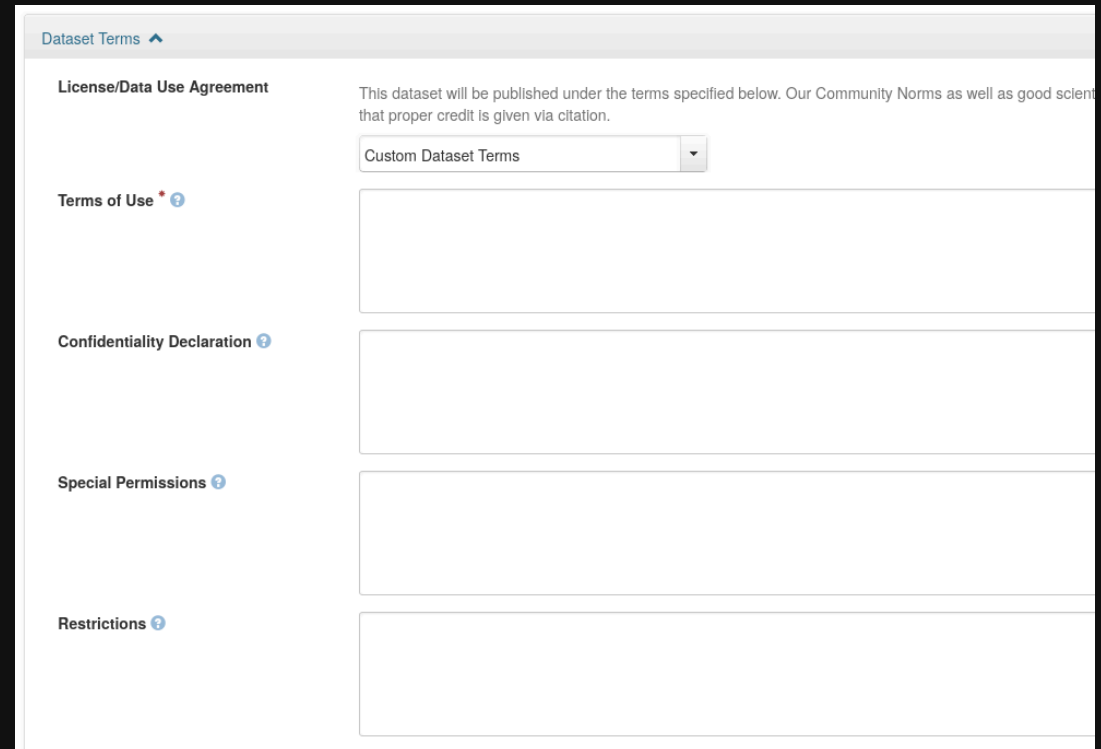
The screenshot shows a web interface for configuring dataset terms. It is divided into two main sections: "License/Data Use Agreement" and "Restricted Files + Terms of Access".

**License/Data Use Agreement:** This section includes a text box stating, "This dataset will be published under the terms specified below. Our Community Norms as well as good scientific practices expect that proper credit is given via citation." Below this is a dropdown menu set to "CC0 1.0" and a "PUBLIC DOMAIN" icon.

**Restricted Files + Terms of Access:** This section contains a "Request Access" checkbox labeled "Enable access request" which is checked. Below it is a "Terms of Access for Restricted Files" field, which is currently empty. At the bottom is a "Data Access Place" field, also empty.

# Terms and Licenses

- You can define custom terms (instead of a standard license)
- Strongly suggest talking with University Counsel!



The screenshot shows a web form titled "Dataset Terms" with a dropdown arrow. The form is divided into several sections:

- License/Data Use Agreement:** Contains the text "This dataset will be published under the terms specified below. Our Community Norms as well as good science that proper credit is given via citation." and a dropdown menu currently set to "Custom Dataset Terms".
- Terms of Use:** A large empty text input field.
- Confidentiality Declaration:** A large empty text input field.
- Special Permissions:** A large empty text input field.
- Restrictions:** A large empty text input field.

# Back to practical matters



# Uploading data to Dataverse via API

- From terminal, with **Python**

```
1 python3 -m venv venv-dv
2 source venv-dv/bin/activate
3 source .Renviron
4 git clone https://github.com/larsvilhuber/dataverse-uploader
5 pip install -r dataverse-uploader/requirements.txt
6 python3 dataverse-uploader/dataverse.py \
7     $DATAVERSE_TOKEN $DATAVERSE_SERVER \
8     $DATAVERSE_DATASET_DOI . -d data/metadata
```

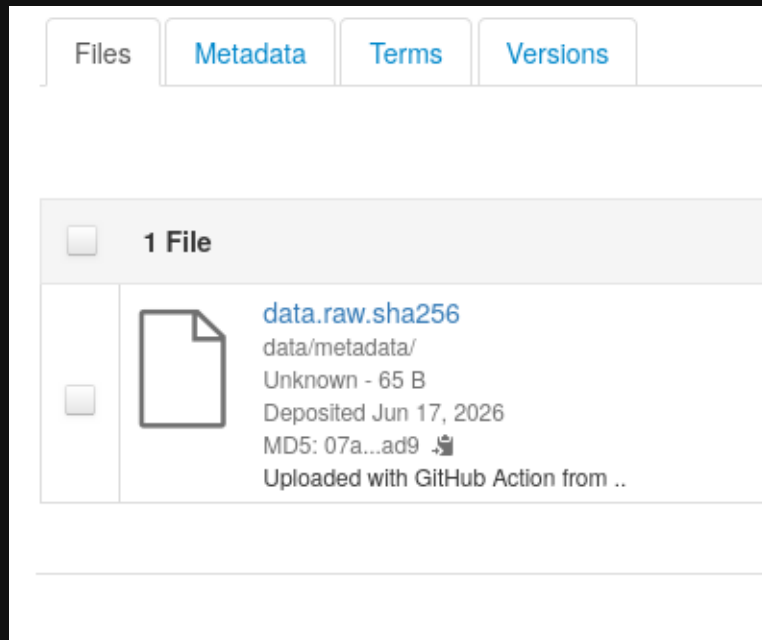


# Results locally

```
1 Connecting to Dataverse server: https://
2 Dataset DOI: doi:10.70122/FK2/EMAWKA
3 Dataset ID: 2695588
4 Found 0 existing files in dataset
5 Deleting 0 existing files...
6
7 Paths to upload: ['./data/metadata']
```

```
1 Scanning path: ./data/metadata
2   Directory: ./data/metadata contains 1
3   Uploading [1]: ./data/metadata/data.
4     Filename: data.raw.sha256
5     Directory label: './data/metadata'
6     Response status: 200
7
8 Total files uploaded: 1
9
10 Done!
```

# Results remotely



- Filename is preserved
- Pathname is preserved!  
*data/metadata*
- MD5 checksum is also present! (less useful for the checksum file!)

# Automatically from Github Actions

- Code for uploading automatically

# Voilà!

We have a workflow that can automatically download from Qualtrics, and in the same move, upload to Dataverse!

Possible improvements:

- immediately publish (no human intervention)
- upload the download logs and checksums together with the data
- run on a schedule

# Putting it all together



# Putting it all together

- We already downloaded from the API, and have created checksum for raw data.
- Let's clean the data, and save local copies
- Then upload the publishable data to Dataverse, and add metadata.

# Cleaning the data

```
1 data.confidential <- data.raw |>
2   filter(consent == "Yes") |>
3   filter(Status != "Survey Preview") |>
4   filter(StartDate > QUALTRICS_STIME & EndDate < QUALTRICS_ETIME) |>
5   select(StartDate, EndDate, Status, Finished, RecordedDate,
6         ResponseId, consent, age_1, gender, education,
7         num_tabs_1, name_confidential, number_confidential)
8 data.clean <- data.confidential %>%
9   select(-name_confidential, -number_confidential)
```



# Save files

```
1 # save files in their locations
2 data.confidential.file <-
3   file.path(confdatapath, "confidential_data.rds")
4 data.clean.file <-
5   file.path(cleandatapath, "clean_data.rds")
6 saveRDS(data.confidential,
7         data.confidential.file)
8 saveRDS(data.clean,
9         data.clean.file)
```



# ... and create checksums

```
1 # Calculate checksums for the saved files
2 confidential_checksum <-
3     digest::digest(data.confidential.file,
4                     algo = "sha256",
5                     file = TRUE)
6 clean_checksum <-
7     digest::digest(data.clean.file,
8                     algo = "sha256",
9                     file = TRUE)
10 # Write checksums to files
11 writeLines(confidential_checksum,
12            file.path(metadatapath, "data.confidential.sha256"))
13 writeLines(clean_checksum,
14            file.path(metadatapath, "data.clean.sha256"))
```



# Analysis

So here are the results so far (2026-06-18):

<b>gender</b>	<b>Frequency</b>	<b>Percent</b>
Male	5	45.45
Female	5	45.45
NA	1	9.09

# By Education

education	Frequency	Percent
Secondary or less	1	9.09
Master's degree	5	45.45
Professional or doctoral degree	4	36.36
NA	1	9.09

# Age

Statistic	Value
Count	10.00
Mean	33.40
Median	28.50
Min	25.00
Max	62.00
Std. Dev.	11.48

# Number of tabs open

Statistic	Value
Count	10.00
Mean	15.80
Median	16.00
Min	2.00
Max	27.00
Std. Dev.	9.19

# State of the data directory

```
1 fs::dir_tree(datapath)
```

```
/home/runner/work/tutorial-preserving-survey/tutorial-preserving-survey/data
├─ clean
│   └─ clean_data.rds
├─ confidential
│   └─ confidential_data.rds
├─ metadata
│   ├── data.clean.sha256
│   ├── data.confidential.sha256
│   └─ data.raw.sha256
├─ raw-confidential
│   ├── README.md
│   └─ Testing+preservation_June+16,+2026_15.35.csv
├─ tutorial-survey.csv
└─ tutorial-survey.rds
```



# Uploading to Dataverse

```
1 # System setup
2 # Need: apt install python3.10-venv
3 python3 -m venv venv-dv
4 source venv-dv/bin/activate
5 git clone https://github.com/larsvilhube
6 pip install -r dataverse-uploader/require
```

```
1 # Do the uploads
2 python3 dataverse-uploader/dataverse.py
3     $DATAVERSE_TOKEN $DATAVERSE_SERVER \
4     $DATAVERSE_DATASET_DOI . -d data/meta
5 python3 dataverse-uploader/dataverse.py
6     $DATAVERSE_TOKEN $DATAVERSE_SERVER \
7     $DATAVERSE_DATASET_DOI . \
8     -d data/clean \
9     --remove false
```

## Change View

Table

Tree

Search this dataset...











Filter by

File Type:All ▾

Access:All ▾

 1 to 4 of 4 Files

<input type="checkbox"/>	 <a href="#">clean_data.rds</a> data/clean/ Gzip Archive - 1.7 KB Deposited Jun 18, 2026 MD5: 159...8cb  Uploaded with GitHub Action from ..
<input type="checkbox"/>	 <a href="#">data.clean.sha256</a> data/metadata/ Unknown - 65 B Deposited Jun 18, 2026 MD5: 7ae...8fb  Uploaded with GitHub Action from ..
<input type="checkbox"/>	 <a href="#">data.confidential.sha256</a> data/metadata/ Unknown - 65 B Deposited Jun 18, 2026 MD5: 88e...0b7  Uploaded with GitHub Action from ..
<input type="checkbox"/>	 <a href="#">data.raw.sha256</a> data/metadata/ Unknown - 65 B Deposited Jun 18, 2026 MD5: 07a...ad9  Uploaded with GitHub Action from ..

# What is next in this space?



# Using 3rd-party trusted systems



# A sketch: Transparency Certified

<https://transparency-certified.github.io/>

**TRACE: Building trust in computational research**

A new approach to computational transparency and reproducibility

Transparency Certified



# Work in progress

- Working with **cascad**, several **INEXDA** members, World Bank, various RDCs
- Relying on external certification of data inputs (data catalogs with metadata, checksums)

# Work in progress

- **SIVACOR**: Scalable Infrastructure for Validation of Computational Social Science Research<sup>7</sup>



**SIVACOR**

Scalable Infrastructure for Validation of Computational Social Science Research



Submit



Documentation

**Does it prevent all  
fraud?**



# Does not prevent all fraud

Toronto researcher loses Ph.D. MIT student makes up firm data

**Exclusive: Psychology researcher loses PhD after allegedly using husband in study and making up data**

Toronto case

[Home](#) | [News](#) | [Assuring An Accurate Research Record](#)

## Assuring an accurate research record

May 16th, 2025

MIT case

# Back to Gino

- A transparent, automated pipeline makes it ***much harder*** to manipulate data **after collection, before analysis** — exactly the Gino failure mode.
- But it does **not** prevent fabricating data at the source, or other forms of misconduct.
- Transparency and preservation raise the cost of fraud and the odds of detection — they are not a silver bullet.

**The end! Thanks for  
your attention.**



# Footnotes

1.

<https://datacolada.org/109>, <https://datacolada.org/110>, <https://datacolada.org/111>,  
<https://datacolada.org/112>, <https://datacolada.org/114>, <https://datacolada.org/118>

2.

Jones, M. (2024). Introducing Reproducible Research Standards at the World Bank. *Harvard Data Science Review*, 6(4). <https://doi.org/10.1162/99608f92.21328ce3>

3. See [my tutorial on handling of confidential data and reproducibility](#)

4.

Ginn J, O'Brien J, Silge J (2024). ***qualtRics: Download 'Qualtrics' Survey Data***. R package version 3.2.1, <https://github.com/ropensci/qualtRics>, <https://docs.ropensci.org/qualtRics/>.

5.

Add `.Renvirom` to your `.gitignore` file to prevent it from being tracked by Git and accidentally pushed to GitHub.

