> Note: The PDF version of this document is transformed by manually printing from a browser.

Citation

> Vilhuber, Lars. 2022. "Process data for the AEA Pre-publication Verification Service." *American Economic Association [publisher]*. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2022-05-16. https://doi.org/10.3886/E117876V3

```
@techreport{10.3886/e117876v3,
  doi = {10.3886/E117876V3},
  url = {https://www.openicpsr.org/openicpsr/project/117876/version/V3/view},
  author = {Vilhuber,  Lars},
  title = {Process data for the AEA Pre-publication Verification Service},
  institution = {American Economic Association [publisher]},
  series = {ICPSR - Interuniversity Consortium for Political and Social Research},
  year = {2022}
}
```

Requirements

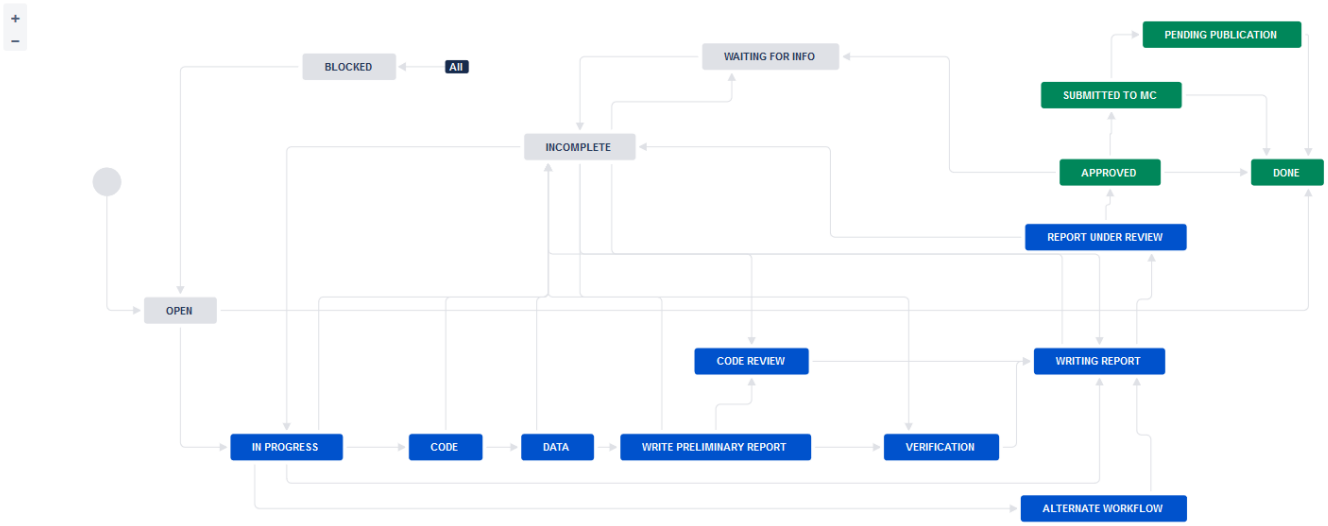
This project requires

- R (last run with R 4.0.4)
  - package `here` (>=0.1)

Other packages might be installed automatically by the programs, as long as the requirements above are met, see Session Info.

Data

The workflow



Raw process data

Raw process data is manually extracted from Jira, and saved as

- `export_MM-DD-YYYY.csv` (for detailed transaction-level data)

The data is not made available outside of the organization, as it contains names of replicators, manuscript numbers, and verbatim email correspondence.

At this time, the latest extract was made 2021-12-09.

Anonymized data

We subset the raw data to variables of interest, and substitute random numbers for sensitive strings. This is done by running `01_jira_anonymize.R`. The programs saves both the confidential version and the anonymized version.

```
source(file.path(programs,"01_jira_anonymize.R"),echo=TRUE)
```

```
##
## > rm(list = ls())
##
## > gc()
##           used (Mb) gc trigger (Mb) max used (Mb)
## Ncells  722127 38.6    1209221 64.6  1209221 64.6
## Vcells 1351474 10.4    8388608 64.0  1938374 14.8
##
## > source(here::here("programs", "config.R"), echo = TRUE)
##
## > process_raw <- TRUE
##
## > download_raw <- TRUE
##
## > extractday <- "12-09-2021"
##
## > firstday <- "2020-12-01"
##
## > lastday <- "2021-11-30"
##
## > basepath <- here::here()
##
## > setwd(basepath)
##
## > jiraconf <- file.path(basepath, "data", "confidential")
##
## > if (Sys.getenv("HOSTNAME") == "zotique3") {
## +     jirconf <- paste0(Sys.getenv("XDG_RUNTIME_DIR"), "/gvfs/dav:host=dav.box.com,ssl=true/dav/Office o ..." ... [TRUNCATED]
##
## > jiraanon <- file.path(basepath, "data", "anon")
##
## > jirameta <- file.path(basepath, "data", "metadata")
##
## > images <- file.path(basepath, "images")
##
## > tables <- file.path(basepath, "tables")
##
## > programs <- file.path(basepath, "programs")
##
## > temp <- file.path(basepath, "data", "temp")
##
## > for (dir in list(images, tables, programs, temp)) {
## +     if (file.exists(dir)) {
## +     }
## +     else {
## +         dir.create(file.path(dir))
## +     }
##  .... [TRUNCATED]
##
## > mran.date <- "2021-01-01"
##
## > options(repos = paste0("https://cran.microsoft.com/snapshot/",
## +     mran.date, "/"))
##
## > pkgTest <- function(x) {
## +     if (!require(x, character.only = TRUE)) {
## +         install.packages(x, dep = TRUE)
## +         if (!require(x, charact .... [TRUNCATED]
##
## > pkgTest.github <- function(x, source) {
## +     if (!require(x, character.only = TRUE)) {
## +         install_github(paste(source, x, sep = "/"))
## +      .... [TRUNCATED]
##
## > if (file.exists(here::here("programs", "confidential-config.R"))) {
## +     source(here::here("programs", "confidential-config.R"))
## + }
##
## > global.libraries <- c("dplyr", "tidyr", "splitstackshape")
##
## > results <- sapply(as.list(global.libraries), pkgTest)
```

```
## Loading required package: splitstackshape
```

```
##
## > exportfile <- paste0("export_", extractday, ".csv")
##
## > if (!file.exists(file.path(jiraconf, exportfile))) {
## +     process_raw = FALSE
## +     print("Input file for anonymization not found - setting global  ..." ... [TRUNCATED]
##
## > if (process_raw == TRUE) {
## +     jira.conf.raw <- read.csv(file.path(jiraconf, exportfile),
## +         stringsAsFactors = FALSE) %>% rename(ticket = .... [TRUNCATED]
```

Publishing data

Some additional cleaning and matching, and then we write out the file

```
source(file.path(programs,"02_jira_anon_publish.R"),echo=TRUE)
```

```
##
## > source(here::here("programs", "config.R"), echo = TRUE)
##
## > process_raw <- TRUE
##
## > download_raw <- TRUE
##
## > extractday <- "12-09-2021"
##
## > firstday <- "2020-12-01"
##
## > lastday <- "2021-11-30"
##
## > basepath <- here::here()
##
## > setwd(basepath)
##
## > jiraconf <- file.path(basepath, "data", "confidential")
##
## > if (Sys.getenv("HOSTNAME") == "zotique3") {
## +     jirconf <- paste0(Sys.getenv("XDG_RUNTIME_DIR"), "/gvfs/dav:host=dav.box.com,ssl=true/dav/Office o ..." ... [TRUNCATED]
##
## > jiraanon <- file.path(basepath, "data", "anon")
##
## > jirameta <- file.path(basepath, "data", "metadata")
##
## > images <- file.path(basepath, "images")
##
## > tables <- file.path(basepath, "tables")
##
## > programs <- file.path(basepath, "programs")
##
## > temp <- file.path(basepath, "data", "temp")
##
## > for (dir in list(images, tables, programs, temp)) {
## +     if (file.exists(dir)) {
## +     }
## +     else {
## +         dir.create(file.path(dir))
## +     }
##  .... [TRUNCATED]
##
## > mran.date <- "2021-01-01"
##
## > options(repos = paste0("https://cran.microsoft.com/snapshot/",
## +     mran.date, "/"))
##
## > pkgTest <- function(x) {
## +     if (!require(x, character.only = TRUE)) {
```

```
## +         install.packages(x, dep = TRUE)
## +         if (!require(x, charact .... [TRUNCATED]
##
## > pkgTest.github <- function(x, source) {
## +     if (!require(x, character.only = TRUE)) {
## +         install_github(paste(source, x, sep = "/"))
## +      .... [TRUNCATED]
##
## > global.libraries <- c("dplyr", "tidyr", "splitstackshape")
##
## > results <- sapply(as.list(global.libraries), pkgTest)
##
## > jira.anon.raw <- readRDS(file.path(jiraanon, "temp.jira.anon.RDS")) %>%
## +     rename(reason.failure = Reason.for.Failure.to.Fully.Replicate) %>%
## + .... [TRUNCATED]
##
## > jira.conf.subtask <- jira.anon.raw %>% select(ticket,
## +     subtask) %>% cSplit("subtask", ",") %>% distinct() %>% pivot_longer(!ticket,
## +     nam .... [TRUNCATED]
##
## > jira.anon <- jira.anon.raw %>% select(ticket, mc_number_anon) %>%
## +     distinct(ticket, .keep_all = TRUE) %>% filter(mc_number_anon !=
## +     is.n .... [TRUNCATED]
```

```
## Joining, by = "ticket"
```

```
##
## > saveRDS(jira.anon, file = file.path(jiraanon, "jira.anon.RDS"))
##
## > write.csv(jira.anon, file = file.path(jiraanon, "jira.anon.csv"))
```

Describing the Data

The anonymized data has 15 columns.

Variables

```
##
## ── Column specification ────────────────────────────────────────────────────────
## cols(
##   name = col_character(),
##   label = col_character()
## )
```

name	label
ticket	The tracking number within the system. Project specific. Sequentially assigned upon receipt.
date_created	Date of a receipt
date_updated	Date of a transaction
mc_number_anon	The (anonymized) number assigned by the editorial workflow system (Manuscript Central/ ScholarOne) to a manuscript. This is purged by a script of any revision suffixes.
Journal	Journal associated with an issue and manuscript. Derived from the manuscript number. Possibly updated by hand
Status	Status associated with a ticket at any point in time. The schema for these has changed over time.
Software.used	A list of software used to replicate the issue.
received	An indicator for whether the issue is just created and has not been assigned to a replicator yet.
Changed.Fields	A transaction will change various fields. These are listed here.
external	An indicator for whether the issue required the external validation.
subtask	An indicator for whether the issue is a subtask of another task.
Resolution	Resolution associated with a ticket at the end of the replication process.

name	label
reason.failure	A list of reasons for failure to fully replicate.
MCRecommendation	Decision status when the issue is Revise and Resubmit.
MCRecommendationV2	Decision status when the issue is conditionally accepted.

Sample records

ticket	date_created	date_updated	mc_number_anon	Journal	Status	Software.used	received	Changed.Fields	external	subtasl
AEAREP-2812	2021-12-08	2021-12-08	979	AEJ:Applied Economics	In Progress	Stata,R	No	Software used,Status	No	NA
AEAREP-2812	2021-12-08	2021-12-08	979	AEJ:Applied Economics	Assigned		No	Assignee,Status	No	NA
AEAREP-2812	2021-12-08	2021-12-08	979	AEJ:Applied Economics	Open		No	openICPSR Project Number	No	NA
AEAREP-2812	2021-12-08	2021-12-08	979	AEJ:Applied Economics	Open		No	DCAF_Access_Restrictions	No	NA
AEAREP-2812	2021-12-08	2021-12-08	979	AEJ:Applied Economics	Open		NA	Journal	No	NA
AEAREP-2812	2021-12-08	2021-12-08	979		Open		NA	Manuscript Central identifier	No	NA

Lab members during this period

We list the lab members active at some point during this period.

```
source(file.path(programs,"lab_members.R"),echo=TRUE)
```

```
##
## > rm(list = ls())
##
## > gc()
##           used (Mb) gc trigger  (Mb) max used  (Mb)
## Ncells  856638 45.8    1491065  79.7  1491065  79.7
## Vcells 1699284 13.0   17175092 131.1 21455074 163.7
##
## > source(here::here("programs", "config.R"), echo = TRUE)
##
## > process_raw <- TRUE
##
## > download_raw <- TRUE
##
## > extractday <- "12-09-2021"
##
## > firstday <- "2020-12-01"
##
## > lastday <- "2021-11-30"
##
## > basepath <- here::here()
##
## > setwd(basepath)
##
## > jiraconf <- file.path(basepath, "data", "confidential")
##
## > if (Sys.getenv("HOSTNAME") == "zotique3") {
## +     jirconf <- paste0(Sys.getenv("XDG_RUNTIME_DIR"), "/gvfs/dav:host=dav.box.com,ssl=true/dav/Office o ..." ... [TRUNCATED]
##
## > jiraanon <- file.path(basepath, "data", "anon")
##
## > jirameta <- file.path(basepath, "data", "metadata")
##
## > images <- file.path(basepath, "images")
##
## > tables <- file.path(basepath, "tables")
```

```
##
## > programs <- file.path(basepath, "programs")
##
## > temp <- file.path(basepath, "data", "temp")
##
## > for (dir in list(images, tables, programs, temp)) {
## +     if (file.exists(dir)) {
## +     }
## +     else {
## +         dir.create(file.path(dir))
## +     }
##  .... [TRUNCATED]
##
## > mran.date <- "2021-01-01"
##
## > options(repos = paste0("https://cran.microsoft.com/snapshot/",
## +     mran.date, "/"))
##
## > pkgTest <- function(x) {
## +     if (!require(x, character.only = TRUE)) {
## +         install.packages(x, dep = TRUE)
## +         if (!require(x, charact .... [TRUNCATED]
##
## > pkgTest.github <- function(x, source) {
## +     if (!require(x, character.only = TRUE)) {
## +         install_github(paste(source, x, sep = "/"))
## +      .... [TRUNCATED]
##
## > global.libraries <- c("dplyr", "tidyr", "splitstackshape")
##
## > results <- sapply(as.list(global.libraries), pkgTest)
##
## > jira.conf.plus <- readRDS(file = file.path(jiraconf,
## +     "jira.conf.plus.RDS"))
##
## > lab.member <- jira.conf.plus %>% filter(Change.Author !=
## +     "" & Change.Author != "Automation for Jira" & Change.Author !=
## +     "LV (Data Edit ..." ... [TRUNCATED]
##
## > write.table(lab.member, file = file.path(basepath,
## +     "data", "replicationlab_members.txt"), sep = "\t", row.names = FALSE)
```

There were a total of 47 lab members over the course of the 12 month period.

R session info

```
sessionInfo()
```

```
## R version 4.0.4 (2021-02-15)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 20.04.3 LTS
##
## Matrix products: default
## BLAS/LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/libopenblasp-r0.3.8.so
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=C
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] splitstackshape_1.4.8 readr_1.4.0          knitr_1.31
## [4] tidyr_1.1.2           stringr_1.4.0        dplyr_1.0.4
##
## loaded via a namespace (and not attached):
```

```
##  [1] rstudioapi_0.13   magrittr_2.0.1    hms_1.0.0         tidyselect_1.1.0
##  [5] here_1.0.1        R6_2.5.0          rlang_0.4.10      highr_0.8
##  [9] tools_4.0.4       data.table_1.13.6 xfun_0.21         cli_2.3.0
## [13] DBI_1.1.1         htmltools_0.5.1.1 ellipsis_0.3.1    yaml_2.2.1
## [17] rprojroot_2.0.2   digest_0.6.27     assertthat_0.2.1  tibble_3.0.6
## [21] lifecycle_1.0.0   crayon_1.4.1      purrr_0.3.4       vctrs_0.3.6
## [25] glue_1.4.2        evaluate_0.14     rmarkdown_2.6     stringi_1.5.3
## [29] compiler_4.0.4    pillar_1.4.7      generics_0.1.0    pkgconfig_2.0.3
```