Process data from reproducibility service

Lars Vilhuber

2024-04-29

> Note: The PDF version (https://aeadataeditor.github.io/processing-jira-process-data/README.pdf) of this document is transformed by manually printing from a browser.

Citation

> Vilhuber, Lars. 2024. "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], 2024-04-29. https://doi.org/10.3886/E117876V5 (https://doi.org/10.3886/E117876V5)

```
doi = {10.3886/E117876V5},
url = {https://doi.org/10.3886/E117876V5},
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 = {2024}}
```

Requirements

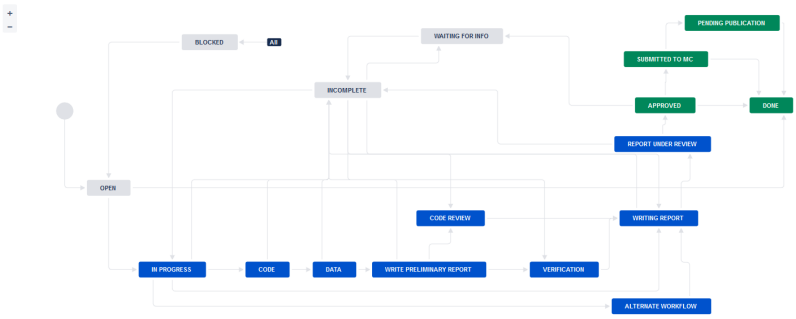
This project requires

- R (last run with R 4.2.3)
  - 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



Workflow stages

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 2023-12-09.

Anonymized data

We subset the raw data to variables of interest, and substitute random numbers for sensitive strings. This is done by running `02_jira_anonymize.R` . The programs saves both the confidential version and the anonymized version.

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

```
##
## > source(here::here("programs", "config.R"), echo = TRUE)
##
## > process_raw <- TRUE
##
## > download_raw <- TRUE
##
## > extractday <- "12-09-2023"
##
## > firstday <- "2022-12-01"
##
## > lastday <- "2023-11-30"
##
## > basepath <- here::here()
##
## > setwd(basepath)
##
## > jiraconf <- file.path(basepath, "data", "confidential")
##
## > if (Sys.getenv("HOSTNAME") == "zotique3") {
## +     jiraconf <- paste0(Sys.getenv("XDG_RUNTIME_DIR"), "/gvfs/dav:host=dav.box.com,ssl=true/dav/Office  ..." .
.. [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]
##
## > jira.conf.plus.rds <- file.path(jiraconf, "jira.conf.plus.RDS")
##
## > assignee.lookup.rds <- file.path(jiraconf, "assignee-lookup.RDS")
##
## > mc.lookup.rds <- file.path(jiraconf, "mc-lookup.RDS")
##
## > if (file.exists(here::here("programs", "confidential-config.R"))) {
## +     source(here::here("programs", "confidential-config.R"))
## + }
##
## > source(here::here("global-libraries.R"), echo = TRUE)
##
## > ppm.date <- "2023-11-01"
##
## > options(repos = paste0("https://packagemanager.posit.co/cran/",
## +     ppm.date, "/"))
##
## > global.libraries <- c("dplyr", "stringr", "tidyr",
## +     "knitr", "readr", "here", "splitstackshape", "boxr", "jose")
##
## > 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]
##
## > results <- sapply(as.list(global.libraries), pkgTest)
##
## > 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]
```

```
## Joining with `by = join_by(ticket)`
```

Publishing data

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

```
source(file.path(programs,"10_jira_anon_publish.R"),echo=TRUE)
```

```
##
## > source(here::here("programs", "config.R"), echo = TRUE)
##
## > process_raw <- TRUE
##
## > download_raw <- TRUE
##
## > extractday <- "12-09-2023"
##
## > firstday <- "2022-12-01"
##
## > lastday <- "2023-11-30"
##
## > basepath <- here::here()
##
## > setwd(basepath)
##
## > jiraconf <- file.path(basepath, "data", "confidential")
##
## > if (Sys.getenv("HOSTNAME") == "zotique3") {
## +     jiraconf <- paste0(Sys.getenv("XDG_RUNTIME_DIR"), "/gvfs/dav:host=dav.box.com,ssl=true/dav/Office  ..." .
.. [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]
##
## > jira.conf.plus.rds <- file.path(jiraconf, "jira.conf.plus.RDS")
##
## > assignee.lookup.rds <- file.path(jiraconf, "assignee-lookup.RDS")
##
## > mc.lookup.rds <- file.path(jiraconf, "mc-lookup.RDS")
##
## > source(here::here("global-libraries.R"), echo = TRUE)
##
## > ppm.date <- "2023-11-01"
##
## > options(repos = paste0("https://packagemanager.posit.co/cran/",
## +     ppm.date, "/"))
##
## > global.libraries <- c("dplyr", "stringr", "tidyr",
## +     "knitr", "readr", "here", "splitstackshape", "boxr", "jose")
##
## > 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]
##
## > 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.be.Fully.Reproducible)  .... [TRUNCATED]
##
## > jira.conf.subtask <- jira.anon.raw %>% filter(subtask !=
## +     "") %>% select(ticket, subtask) %>% separate_longer_delim(subtask,
## +     delim = ", ..." ... [TRUNCATED]
##
## > jira.anon <- jira.anon.raw %>% filter(!is.na(mc_number_anon)) %>%
## +     anti_join(jira.conf.subtask) %>% select(ticket, date_created,
## +     date_u .... [TRUNCATED]
```

```
## Joining with `by = join_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

```
## Rows: 15 Columns: 2
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (2): name, label
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

name	label

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.
Resolution	Resolution associated with a ticket at the end of the replication process.
reason.failure	A list of reasons for failure to fully replicate.
MCStatus	NA
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	Resolution	reason.failure	MCStatus	MCR
AEAREP-4863	2023-12-08	2023-12-08	1	AER:Insights	Open	Stata	Yes	Software used	No			CA	
AEAREP-4863	2023-12-08	2023-12-08	1	AER:Insights	Open		Yes	openICPSR Project Number	No			CA	
AEAREP-4863	2023-12-08	2023-12-08	1	AER:Insights	Open		Yes	Manuscript Central identifier	No			CA	
AEAREP-4862	2023-12-08	2023-12-08	2	AER	Open	Python	Yes	Software used	No			CA	
AEAREP-4862	2023-12-08	2023-12-08	2	AER	Open		Yes	openICPSR Project Number	No			CA	
AEAREP-4862	2023-12-08	2023-12-08	2	AER	Open		Yes	Journal	No			CA	

Lab members during this period

We list the lab members active at some point during this period. This still requires confidential data as an input.

```
source(file.path(programs,"03_lab_members.R"),echo=TRUE)
```

```
##
## > source(here::here("programs", "config.R"), echo = TRUE)
##
## > process_raw <- TRUE
##
## > download_raw <- TRUE
##
## > extractday <- "12-09-2023"
##
## > firstday <- "2022-12-01"
##
## > lastday <- "2023-11-30"
##
## > basepath <- here::here()
##
## > setwd(basepath)
##
## > jiraconf <- file.path(basepath, "data", "confidential")
##
## > if (Sys.getenv("HOSTNAME") == "zotique3") {
## +     jiraconf <- paste0(Sys.getenv("XDG_RUNTIME_DIR"), "/gvfs/dav:host=dav.box.com,ssl=true/dav/Office  ..." .
.. [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]
##
## > jira.conf.plus.rds <- file.path(jiraconf, "jira.conf.plus.RDS")
##
## > assignee.lookup.rds <- file.path(jiraconf, "assignee-lookup.RDS")
##
## > mc.lookup.rds <- file.path(jiraconf, "mc-lookup.RDS")
##
## > if (file.exists(here::here("programs", "confidential-config.R"))) {
## +     source(here::here("programs", "confidential-config.R"))
## + }
##
## > source(here::here("global-libraries.R"), echo = TRUE)
##
## > ppm.date <- "2023-11-01"
##
## > options(repos = paste0("https://packagemanager.posit.co/cran/",
## +     ppm.date, "/"))
##
## > global.libraries <- c("dplyr", "stringr", "tidyr",
## +     "knitr", "readr", "here", "splitstackshape", "boxr", "jose")
##
## > 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]
##
## > results <- sapply(as.list(global.libraries), pkgTest)
##
## > exclusions <- c("Lars Vilhuber", "Michael Darisse",
## +     "Sofia Encarnacion", "Linda Wang", "Leonel Borja Plaza",
## +     "User ", "Takshil Sachdev ..." ... [TRUNCATED]
##
## > lookup <- read_csv(file.path(jirameta, "lookup.csv"))
```

```
## Rows: 2 Columns: 2
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (2): Assignee, Name
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
##
## > jira.conf.plus <- readRDS(jira.conf.plus.rds)
##
## > lab.member <- jira.conf.plus %>% filter(date_created >=
## +     firstday, date_created < lastday) %>% filter(Assignee !=
## +     "") %>% filter(!Assig .... [TRUNCATED]
```

```
## Joining with `by = join_by(Assignee)`
```

```
##
## > write.table(lab.member, file = file.path(basepath,
## +     "data", "replicationlab_members.txt"), sep = "\t", row.names = FALSE)
##
## > external.member <- jira.conf.plus %>% filter(External.party.name !=
## +     "") %>% mutate(date_created = as.Date(substr(Created, 1,
## +     10), "%m/ ..." ... [TRUNCATED]
##
## > write.table(external.member, file = file.path(basepath,
## +     "data", "external_replicators.txt"), sep = "\t", row.names = FALSE)
```

There were a total of 45 lab members over the course of the 12 month period.

R session info

```
sessionInfo()
```

```
## R version 4.2.3 (2023-03-15)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 22.04.2 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
## LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/libopenblasp-r0.3.20.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=en_US.UTF-8
##  [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] jose_1.2.0            openssl_2.0.6         boxr_0.3.6
##  [4] splitstackshape_1.4.8 here_1.0.1            readr_2.1.4
##  [7] knitr_1.42            tidyr_1.3.0           stringr_1.5.0
## [10] dplyr_1.1.0
##
## loaded via a namespace (and not attached):
##  [1] pillar_1.8.1      bslib_0.4.2       compiler_4.2.3    jquerylib_0.1.4
##  [5] tools_4.2.3       bit_4.0.5         digest_0.6.31     jsonlite_1.8.4
##  [9] evaluate_0.20     lifecycle_1.0.3   tibble_3.2.0      pkgconfig_2.0.3
## [13] rlang_1.1.0       cli_3.6.0         rstudioapi_0.14   parallel_4.2.3
## [17] yaml_2.3.7        xfun_0.43         fastmap_1.1.1     withr_2.5.0
## [21] askpass_1.1       generics_0.1.3    vctrs_0.5.2       sass_0.4.5
## [25] hms_1.1.2         bit64_4.0.5       rprojroot_2.0.3   tidyselect_1.2.0
## [29] data.table_1.14.8 glue_1.6.2        R6_2.5.1          fansi_1.0.4
## [33] vroom_1.6.1       rmarkdown_2.20    purrr_1.0.1       tzdb_0.3.0
## [37] magrittr_2.0.3    htmltools_0.5.4   ellipsis_0.3.2    utf8_1.2.3
## [41] stringi_1.7.12    cachem_1.0.7      crayon_1.5.2
```