Process data from reproducibility service

Lars Vilhuber

2023-05-24

> 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. 2023. “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], 2023-05-24. https://doi.org/10.3886/E117876V3 (https://doi.org/10.3886/E117876V4)

```
@techreport{10.3886/e117876v4,
  doi = {10.3886/E117876V4},
  url = {https://www.openicpsr.org/openicpsr/project/117876/version/V4/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 = {2023}
}
```

Requirements

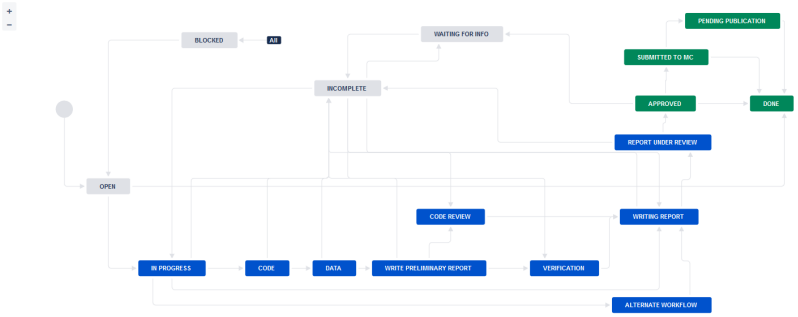
This project requires

- R (last run with R 4.2.2)
  - 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 2022-12-08.

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  866198 46.3    1362272 72.8  1362272 72.8
## Vcells 1557484 11.9    8388608 64.0  2312509 17.7
##
## > source(here::here("programs", "config.R"), echo = TRUE)
##
## > process_raw <- TRUE
##
## > download_raw <- TRUE
##
## > extractday <- "12-12-2022"
##
## > firstday <- "2021-12-01"
##
## > lastday <- "2022-11-30"
##
## > 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]
##
## > mran.date <- "2022-04-22"
##
## > 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]
## [1] "Input file for anonymization not found - setting global parameter to FALSE"
##
## > if (process_raw == TRUE) {
## +     jira.conf.raw <- read.csv(file.path(jiraconf, exportfile),
## +         stringsAsFactors = FALSE) %>% rename(ticket = .... [TRUNCATED]
## [1] "Not processing anonymization due to global parameter."
```

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 <- "12-12-2022"
##
## > firstday <- "2021-12-01"
##
## > lastday <- "2022-11-30"
##
## > 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]
##
## > mran.date <- "2022-04-22"
##
## > 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

```
## 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
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.

name	label
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.
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	subtask	Resolution	reason.failure	MCR
AEAREP-3787	2022-12-08	2022-12-08	1316	AEJ:Economic Policy	Open	Stata,R	No	Software used	No	NA			
AEAREP-3787	2022-12-08	2022-12-08	1316	AEJ:Economic Policy	Open		No	openICPSR Project Number	No	NA			
AEAREP-3787	2022-12-08	2022-12-08	1316	AEJ:Economic Policy	Open		NA	Manuscript Central identifier	No	NA			
AEAREP-3787	2022-12-08	2022-12-08	1316	AEJ:Economic Policy	Open		NA	Journal	No	NA			
AEAREP-3787	2022-12-08	2022-12-08	1316		Open		NA		No	NA			
AEAREP-3786	2022-12-07	2022-12-07	316	AER	Done		No	Status	No	NA	Evaluation only		

Lab members during this period

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

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

```
##
## > rm(list = ls())
##
## > gc()
##           used (Mb) gc trigger  (Mb) max used  (Mb)
## Ncells 1028453 55.0    2007871 107.3  2007871 107.3
## Vcells 1946453 14.9   19765818 150.9 30884089 235.7
##
## > source(here::here("programs", "config.R"), echo = TRUE)
##
## > process_raw <- TRUE
##
## > download_raw <- TRUE
##
## > extractday <- "12-12-2022"
##
## > firstday <- "2021-12-01"
##
## > lastday <- "2022-11-30"
##
## > 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]
##
## > mran.date <- "2022-04-22"
##
## > 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)
##
## > 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 42 lab members over the course of the 12 month period.

R session info

```
sessionInfo()
```

```
## R version 4.2.2 (2022-10-31)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 22.04.1 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] splitstackshape_1.4.8 readr_2.1.3           knitr_1.41
## [4] tidyr_1.2.1           stringr_1.5.0         dplyr_1.0.10
##
## loaded via a namespace (and not attached):
##  [1] highr_0.9         bslib_0.4.1       compiler_4.2.2    pillar_1.8.1
##  [5] jquerylib_0.1.4   tools_4.2.2       bit_4.0.5         digest_0.6.30
##  [9] jsonlite_1.8.4    evaluate_0.18     lifecycle_1.0.3   tibble_3.1.8
## [13] pkgconfig_2.0.3   rlang_1.0.6       cli_3.4.1         DBI_1.1.3
## [17] rstudioapi_0.14   parallel_4.2.2    yaml_2.3.6        xfun_0.39
## [21] fastmap_1.1.0     withr_2.5.0       generics_0.1.3    vctrs_0.5.1
## [25] sass_0.4.4        hms_1.1.2         bit64_4.0.5       rprojroot_2.0.3
## [29] tidyselect_1.2.0  data.table_1.14.6 glue_1.6.2        here_1.0.1
## [33] R6_2.5.1          fansi_1.0.3       vroom_1.6.0       rmarkdown_2.18
## [37] tzdb_0.3.0        purrr_0.3.5       magrittr_2.0.3    ellipsis_0.3.2
## [41] htmltools_0.5.4   assertthat_0.2.1  utf8_1.2.2        stringi_1.7.8
## [45] cachem_1.0.6      crayon_1.5.2
```