National Center for Charitable Statistics

Evolving Nonprofit Sector Data Infrastructure: New Resources and Tools

ARNOVA 2024

Tiana Marrese • PhD Candidate @ UPenn
Thiya Poongundranar • Data Scientist @ Urban
Jesse Lecy • ASU + Urban

Business Master File (BMF)

Sampling framework for 990 research

New UNIFIED format

- 1.9M active nonprofits + 1.6M historic
- standardized geographies
- better validation of org attributes

NCCS
· URBAN · INSTITUTE ·

990 CORE Series

Best panel for longitudinal financial analysis

- Coverage from 1989-2023
- Separate panels for public charities, private foundations, and other 501c nonprofits
- (how many variables?)

New HARMONIZED format

- Variable names standardized over time
- Geographies standardized + crosswalks available
- Use consistent organizational attributes

NCCS
· URBAN · INSTITUTE ·

Standardized Geographies

Legacy files had inconsistent geographies

- Now uses a single geography based on the most recent address we have for nonprofits to make panels consistent
- Lat-Lon and geographic IDs in BMF and CORE files

GEO Crosswalk framework

- Crosswalks for 13 different geographies – allow for easy data aggregation or to merge outside datasets
- Pre-compiled panel of census variables from 1990-2020

NCCS
· URBAN · INSTITUTE ·

NTEE Fields

National Taxonomy of Exempt Entities

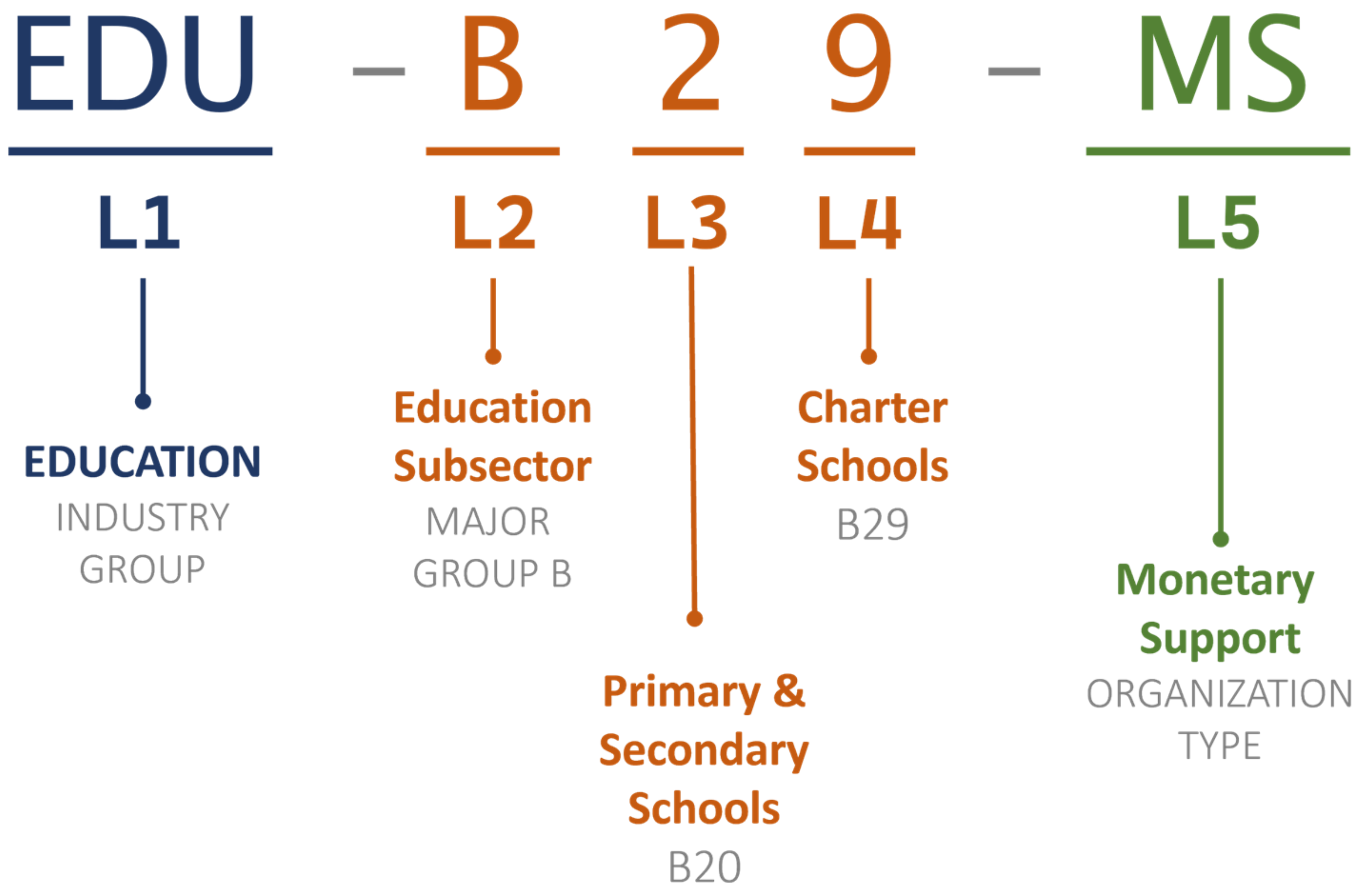
- Industry classification system for the nonprofit sector

Updated NTEEV2 Format

- Designed to make sampling easier
- Separates org types and activities
- Dictionary is built into the NCCS R package

NCCS
· URBAN · INSTITUTE ·

NTEE V2




```
parse_ntee(
  ntee.group   = "all",
  ntee.code    = "A2x",
  ntee.orgtype = "all"  )
```

```
#>     "ART-A26-MT"  "ART-A20-PA"
#>     "ART-A25-PA”  "ART-A20-MS"
#>     "ART-A26-MM”  "ART-A23-RG"
#>     "ART-A20-NS"  "ART-A20-RG"
#>     "ART-A24-RG"  "ART-A25-RG"
#>     "ART-A26-RG"  "ART-A27-RG"
```

990 EFILE Database

Most comprehensive 990 dataset

- 20 times more variables than the CORE series, including text fields and all schedules
- XML has been parsed into 125 CSV tables
- Data from 2010 to 2023, coverage grows over time

Efiling became mandatory in 2022

- 660k 990 filers
- 350k 990EZ filers
- 180k 990PF filers

NCCS
· URBAN · INSTITUTE ·

990 EFILE Database

YEAR	990	990EZ	990PF	990T
2007	17	17	0	0
2008	87	114	20	0
2009	33,311	15,470	2345	0
2010	123,026	63,326	25249	0
2011	159,504	82,048	34597	0
2012	179,688	93,750	39933	0
2013	198,856	104,375	45887	0
2014	218,620	116,417	53442	0
2015	233,520	124,894	58815	0
2016	243,903	130,485	62988	0
2017	261,612	139,146	68950	0
2018	271,442	149,384	80138	0
2019	283,741	152,669	87805	0
2020	323,393	172,020	116,484	23,302
2021	355,254	219,703	129,136	24,575
2022	**663,940**	**349,484**	**176,778**	**38,610**
2023	235,492	266,856	179,826	8866

First year efiling is mandatory

Nonprofit Trends Survey

Annual Surveys from 2020 onward

- Operational insights that better capture sector trends beyond what is available on 990 forms
- Representative national sample of organizations
- Public use data files

Multi-Institution Collaboration

- Urban Institute: NCCS + Teresa Derrick-Mills
- American University: Lewis Faulk
- George Mason University: Mirae Kim, Alan Abramson
- Georgia Tech: Calton Pu
- NSF Funding + Ongoing Philanthropic Support

NCCS
· URBAN · INSTITUTE ·

Political Action Committees (527 orgs)

Form 8871 database of PACs

- Nonprofits that have filed to act as political action committees

Form 8872 activities

- Donations made to nonprofits
- Expenditures on political campaigns

Also see Schedule C: Lobbying Activities

- Available in the EFILE database

NCCS
· URBAN · INSTITUTE ·

Regulations Project: Legal Compendium Dataset

V1: 2016 — Cindy Lott, Faisal Sheikh, Karin Kunstler Goldman, Belinda Johns, Marcus Gaddy and Maura R. Farrell

V2: 2019 — Cindy Lott, Mary Shelly, Nathan Dietz, Put Barber, Rob Greenleaf

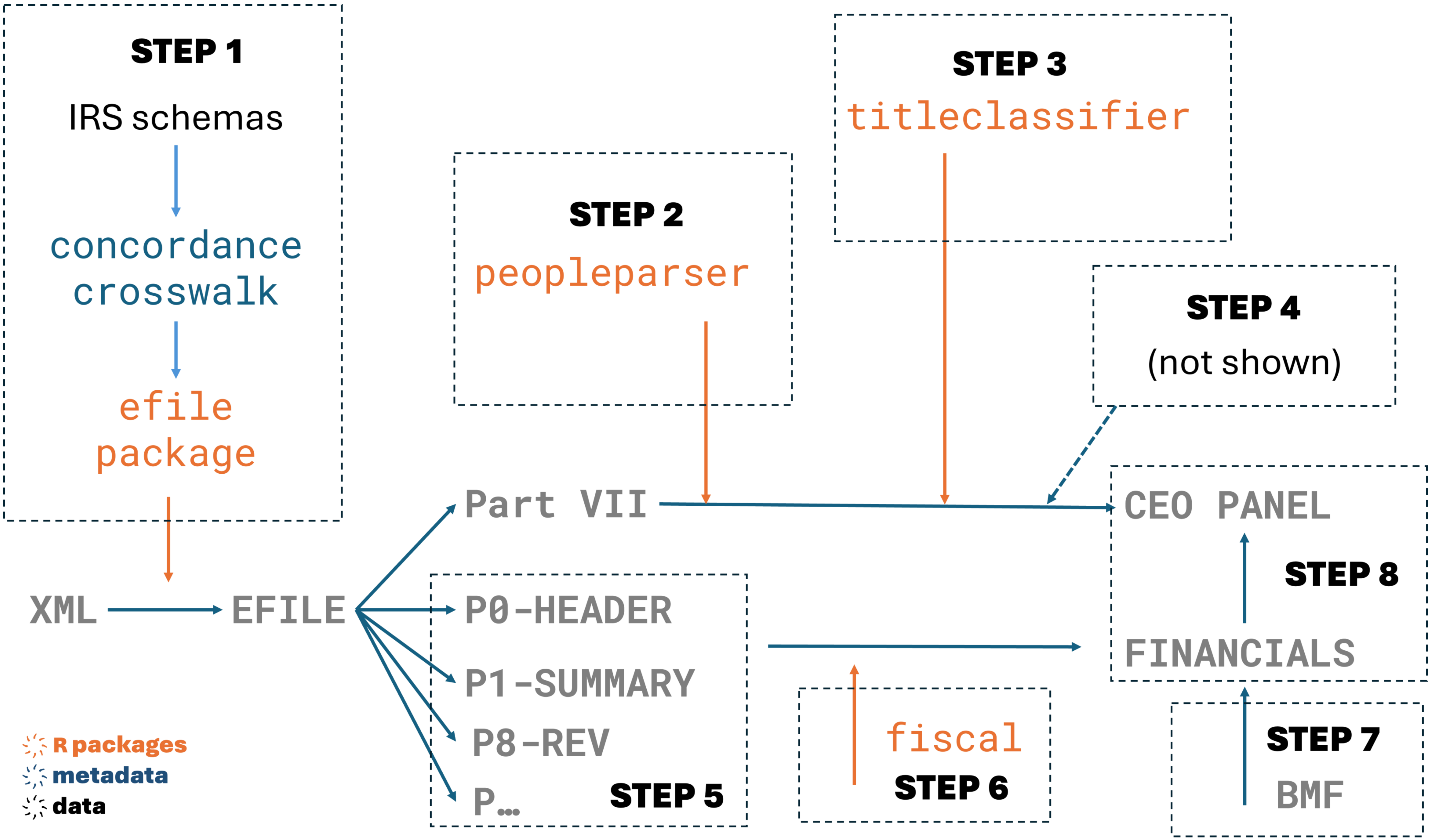
V3: 2024 — Mary Shelly, Cindy Lott, Ethan Roberson

V4: 2024 — Elizabeth Boris, Teresa Harrison, Jesse Lecy

NCCS
· URBAN · INSTITUTE ·

Putting it all together:

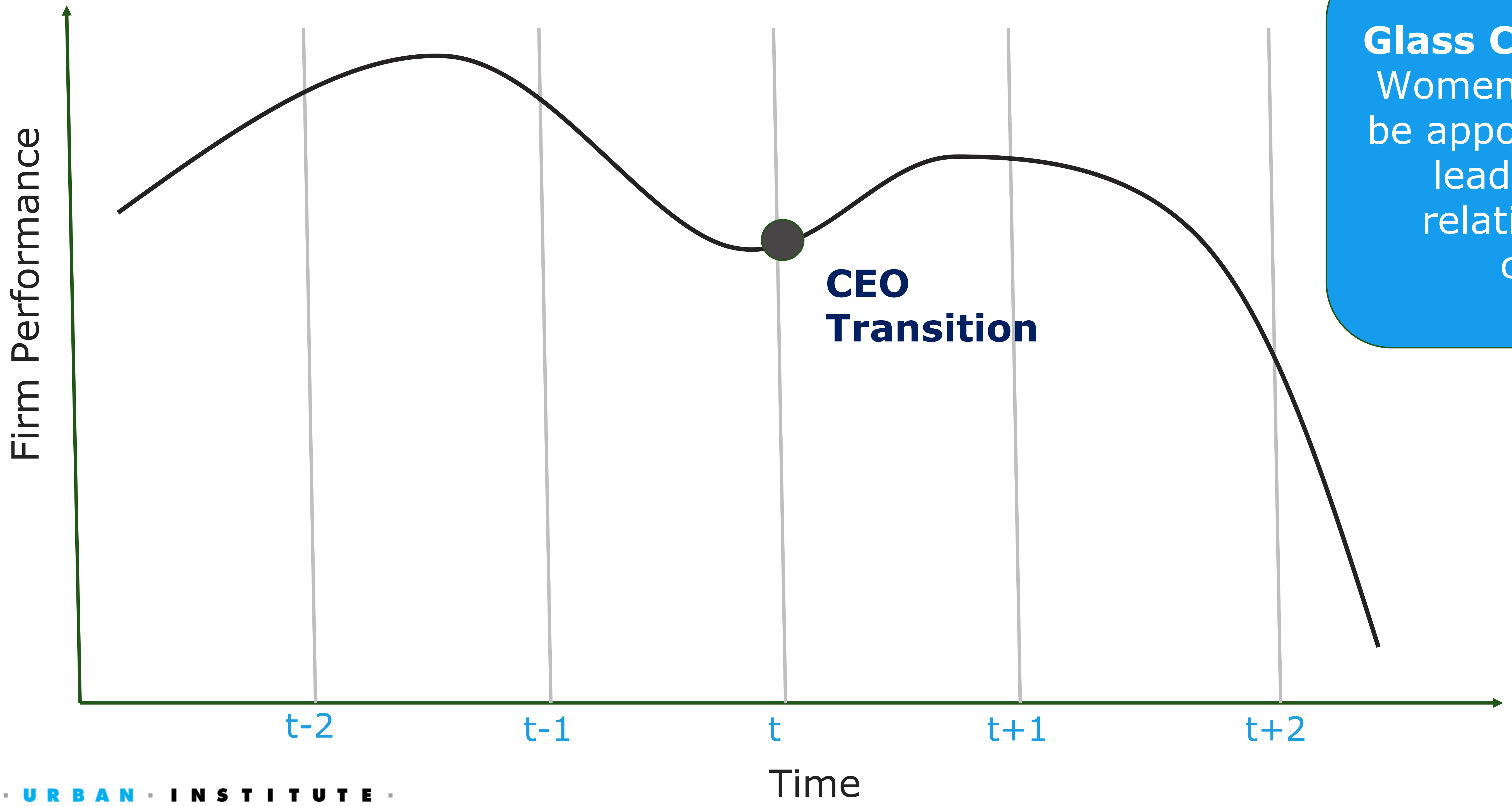
modular and transparent data engineering workflows


STEP 1
IRS schemas
concordance
crosswalk
efile
package
STEP 2
peopleparser
STEP 3
titleclassifier
STEP 4
(not shown)
Part VII
CEO PANEL
STEP 8
XML
EFILE
P0-HEADER
P1-SUMMARY
P8-REV
P...
STEP 5
fiscal
STEP 6
FINANCIALS
STEP 7
BMF
R packages
metadata
data

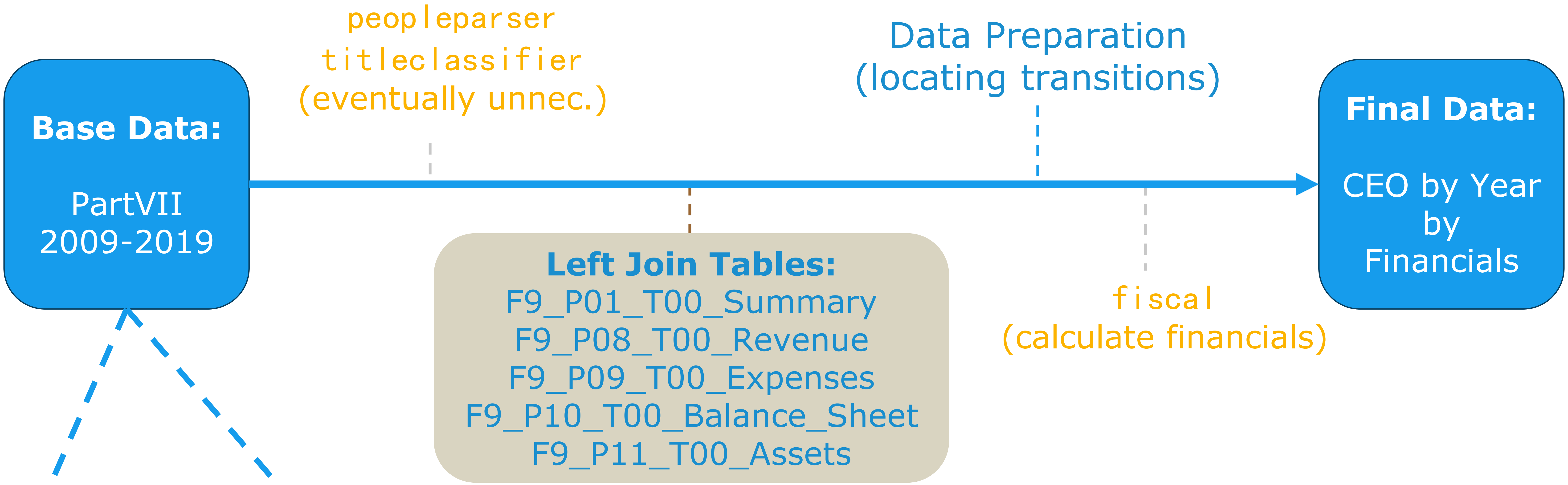
Nonprofits at the Glass Cliff?

A Research Example

The Big Idea



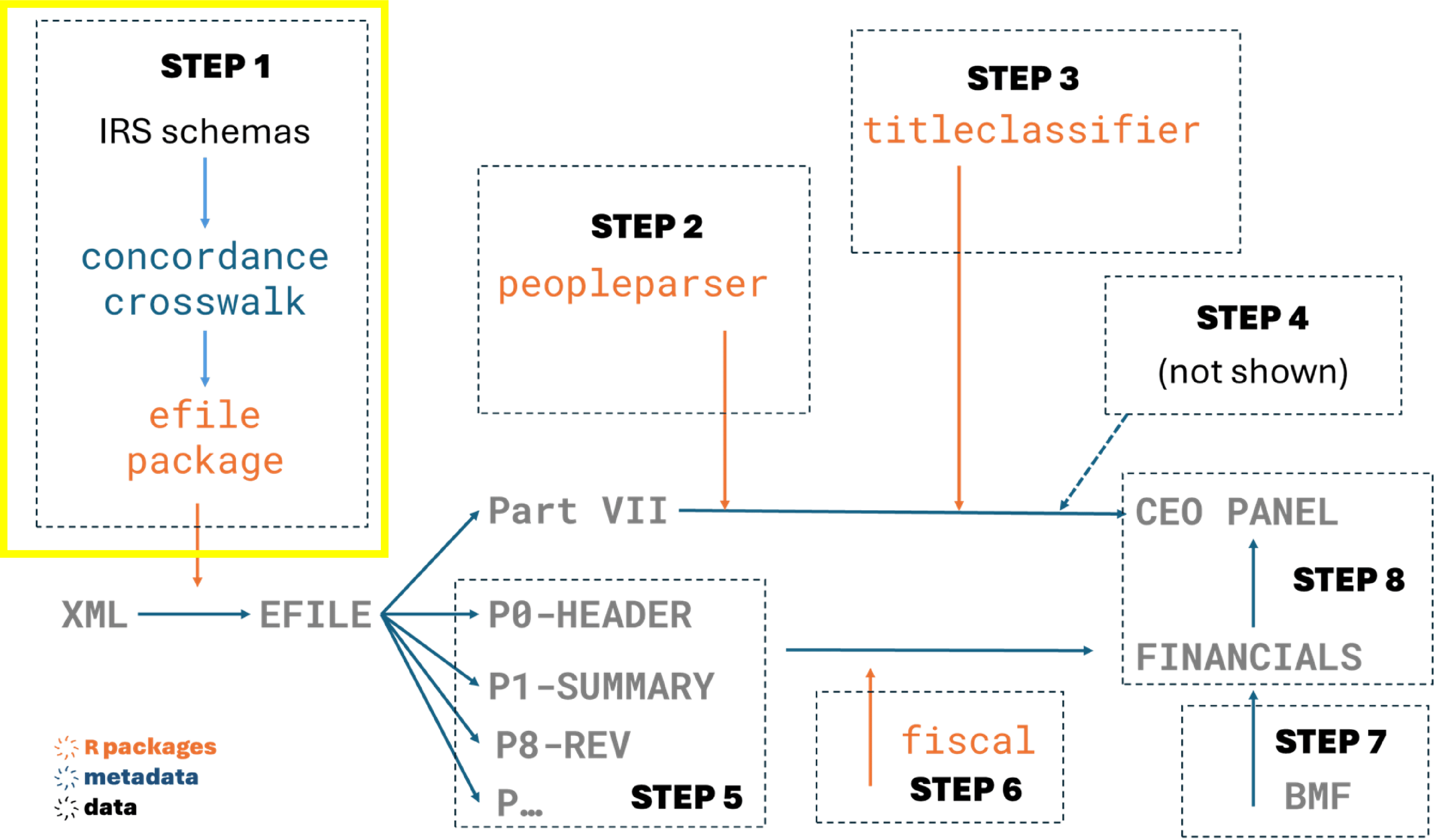
Glass Cliff Phenomenon: Women are more likely to be appointed to precarious leadership positions relative to their male counterparts



EIN	NAME	TAXYR	FORMTYPE	F9_07_COMP_DTK_NAME_PERS	F9_07_COMP_DTK_TITLE
562503325	Helena Historical Society	2018	990EZ	Rhonda Hungerford	Member at Large
562503325	Helena Historical Society	2018	990EZ	Patricia Kear-Ross	Member at Large
581970876	Oregon Park Baseball Association Inc	2018	990	Bob Martel	Treasurer
581970876	Oregon Park Baseball Association Inc	2018	990	Jennifer Bramlett	Secretary
581970876	Oregon Park Baseball Association Inc	2018	990	Wayne Brown	President
250972074	YOUNGSTOWN VOLUNTEER FIRE DEPARTMENT & RELIEF ASSOCIATION	2018	990	BRIAN SCHMUCKER	PRESIDENT
250972074	YOUNGSTOWN VOLUNTEER FIRE DEPARTMENT & RELIEF ASSOCIATION	2018	990	JASON BLOOM	VICE PREIDENT

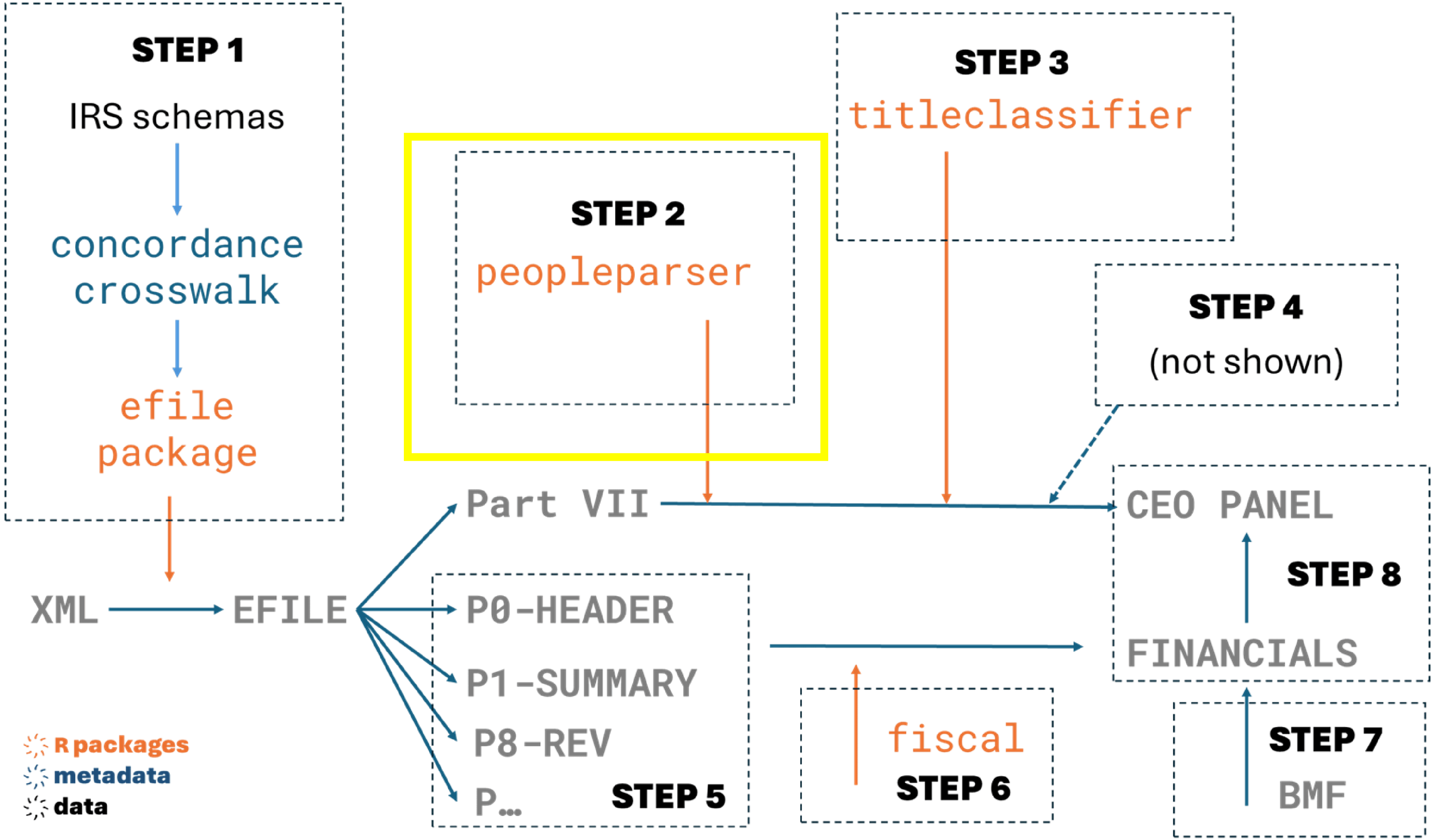
Step 1

```r
#Pre-built functions:
nodc <- "https://raw.githubusercontent.com/Nonprofit-Open-Data-Collective/"
repo <- "arnova-2024/refs/heads/main/"
file <- "functions.R"
source( paste0( nodc, repo, file ) )
```



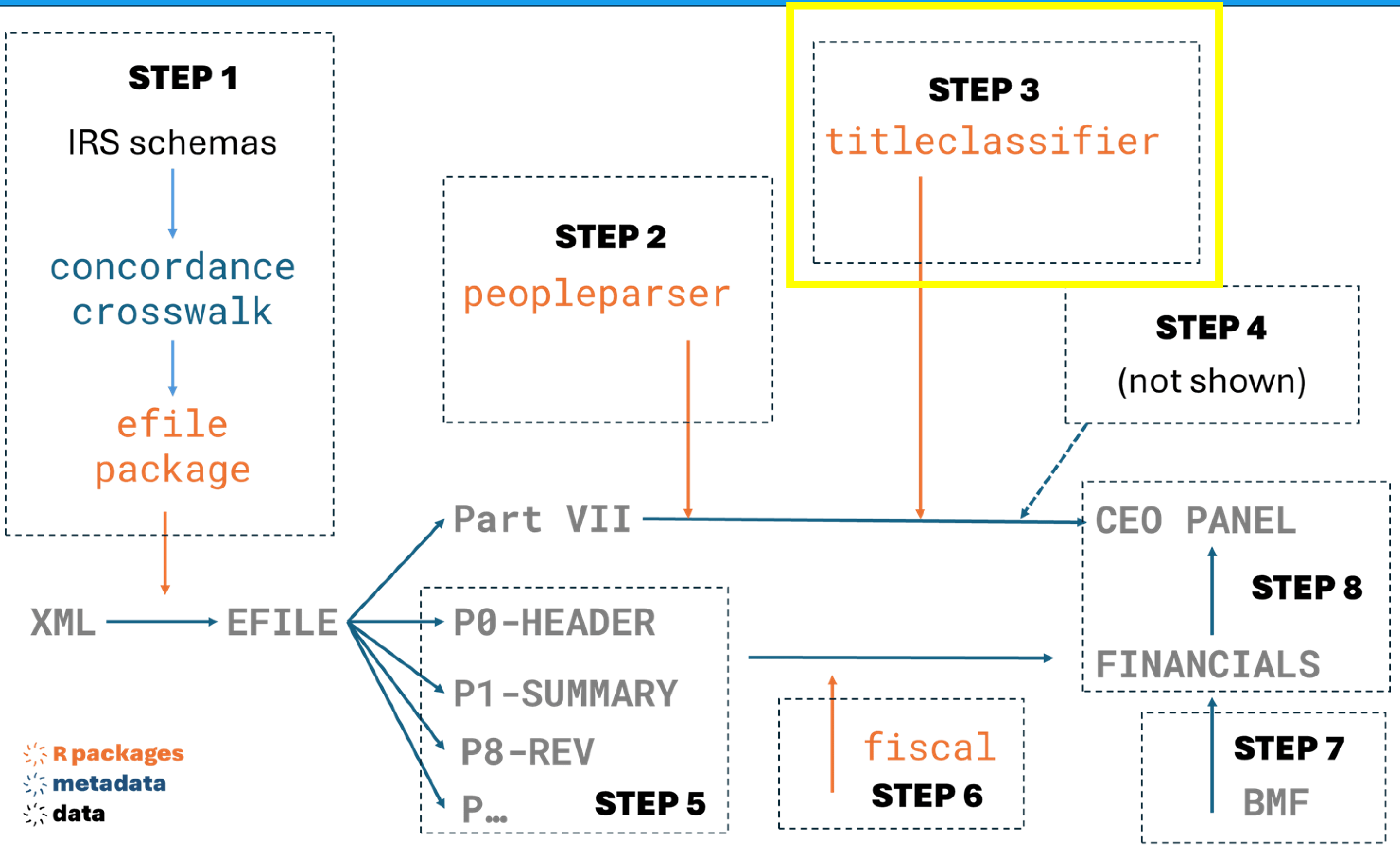
Step 2

```
df_sub <- read.csv(file = "PARTVII_10employees_2009_2010.csv")
#Now let's use peopleparser to clean the names. This function can take a bit to run so let's further limit
our df to only unique names
df_unique_names <- df_sub %>% select(F9_07_COMP_DTK_NAME_PERS) %>%
  distinct()
#Now doing people parser
df_names <- parse.names(df_unique_names$F9_07_COMP_DTK_NAME_PERS)
#We can left_join  these names back onto the df
df_sub <- df_sub %>% left_join(df_names, by = c("F9_07_COMP_DTK_NAME_PERS" = "name"))
```



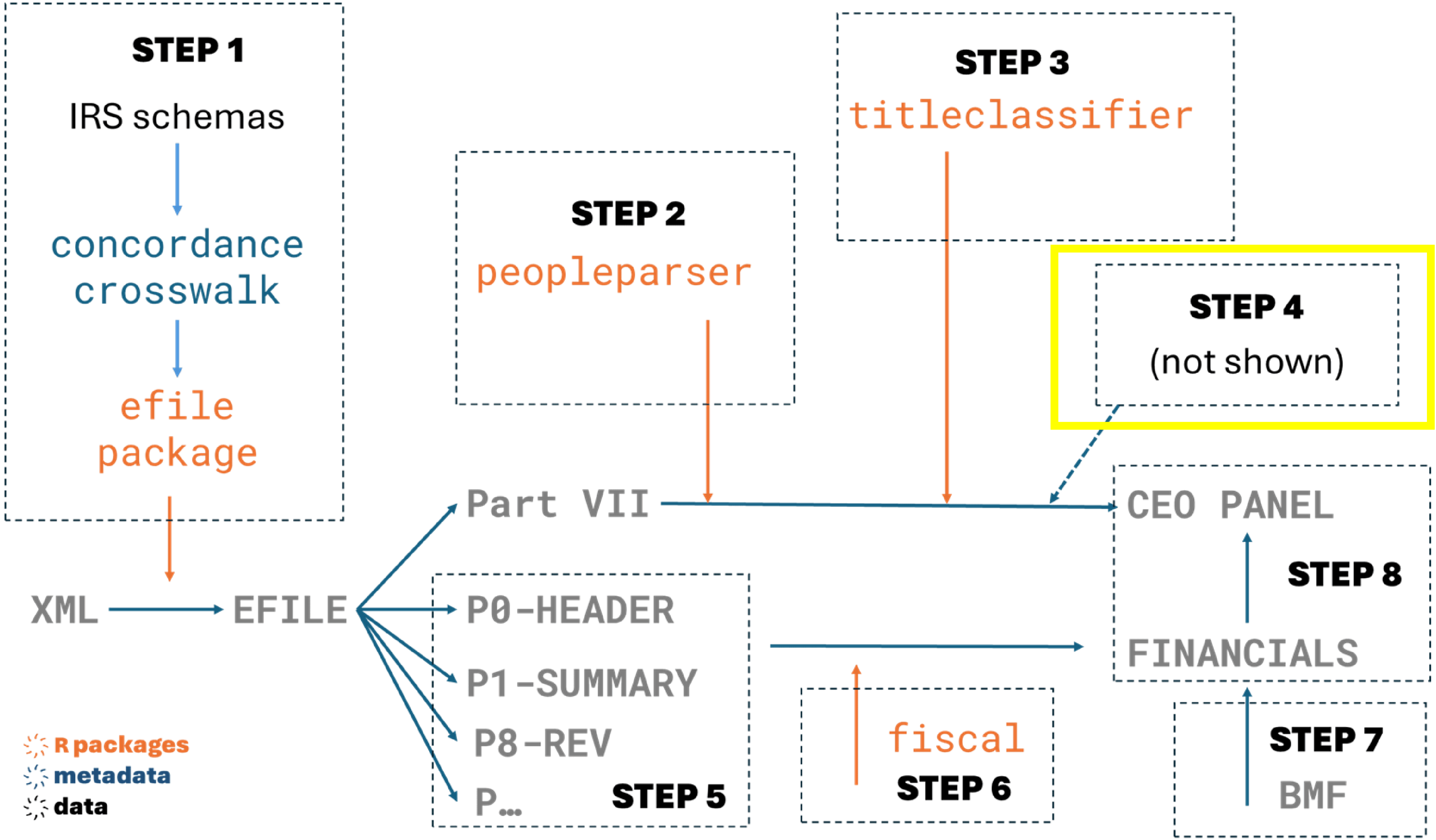
Step 3

```
df_sub <- read.csv(file = "PARTVII_10employees_2009_2010_names.csv")

#Running the title classifier now
df_titles <-  df_unique_titles %>%
  standardize_df() %>%
  remove_dates() %>%
  standardize_conj() %>%
  split_titles() %>%
  standardize_spelling() %>%
  gen_status_codes() %>%
  standardize_titles() %>%
  categorize_titles()

write.csv(df_sub_names_titles, "PARTVII_10employees_2009_2010_names_titles.csv")
```



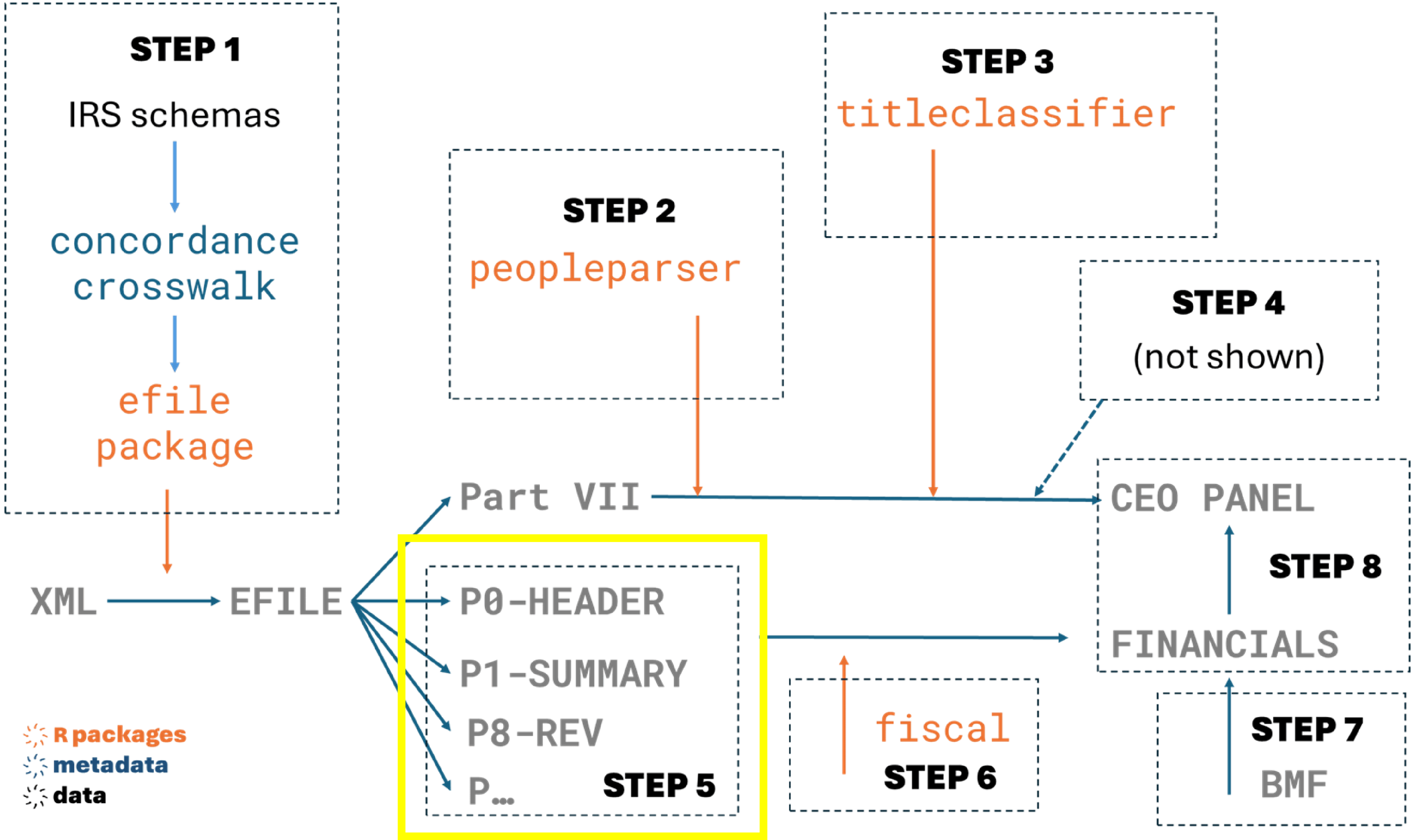
Step 4

```
ceo_trans_10    <- read.csv("toy_CEO_trans_10EIN.csv" )
ceo_trans_1000 <- read.csv("toy_CEO_trans_1000EIN.csv")
```



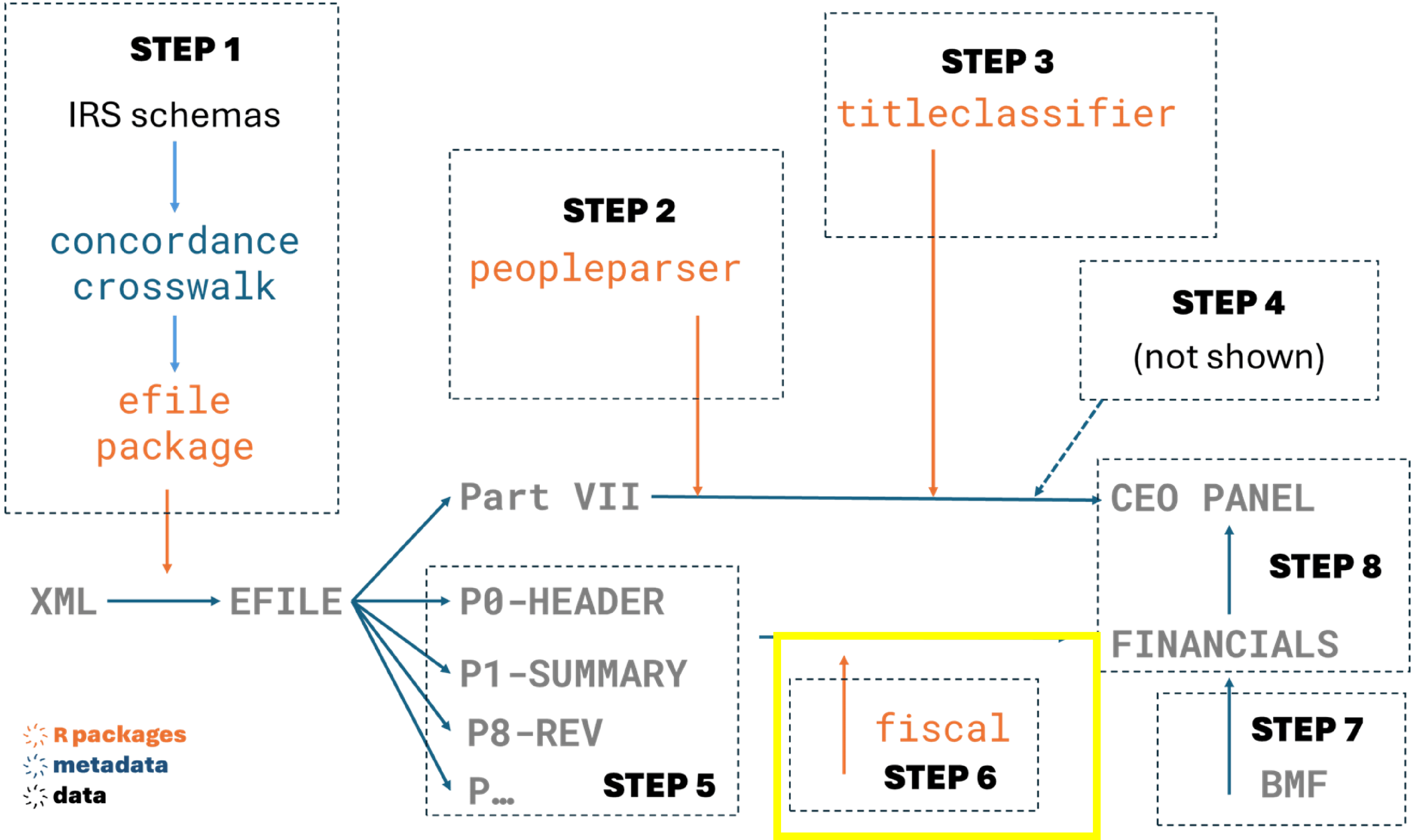
Step 5

```
summary <- read.csv("F9-P01-T00-SUMMARY-SAMPLE-10.csv")
revenue <- read.csv("F9-P08-T00-REVENUE-SAMPLE-10.csv")
expenses <- read.csv("F9-P09-T00-EXPENSES-SAMPLE-10.csv")

#let's join everything together:
ceo_trans_10_fncl <- ceo_trans_10 %>% left_join(summary, by = c("TAXYR" = "TAX_YEAR",
                                                                "EIN" = "ORG_EIN"))
ceo_trans_10_fncl <- ceo_trans_10_fncl %>% left_join(revenue, by = c("TAXYR" = "TAX_YEAR",
                                                                "EIN" = "ORG_EIN"))
ceo_trans_10_fncl <- ceo_trans_10_fncl %>% left_join(expenses, by = c("TAXYR" = "TAX_YEAR",
                                                                "EIN" = "ORG_EIN"))
```

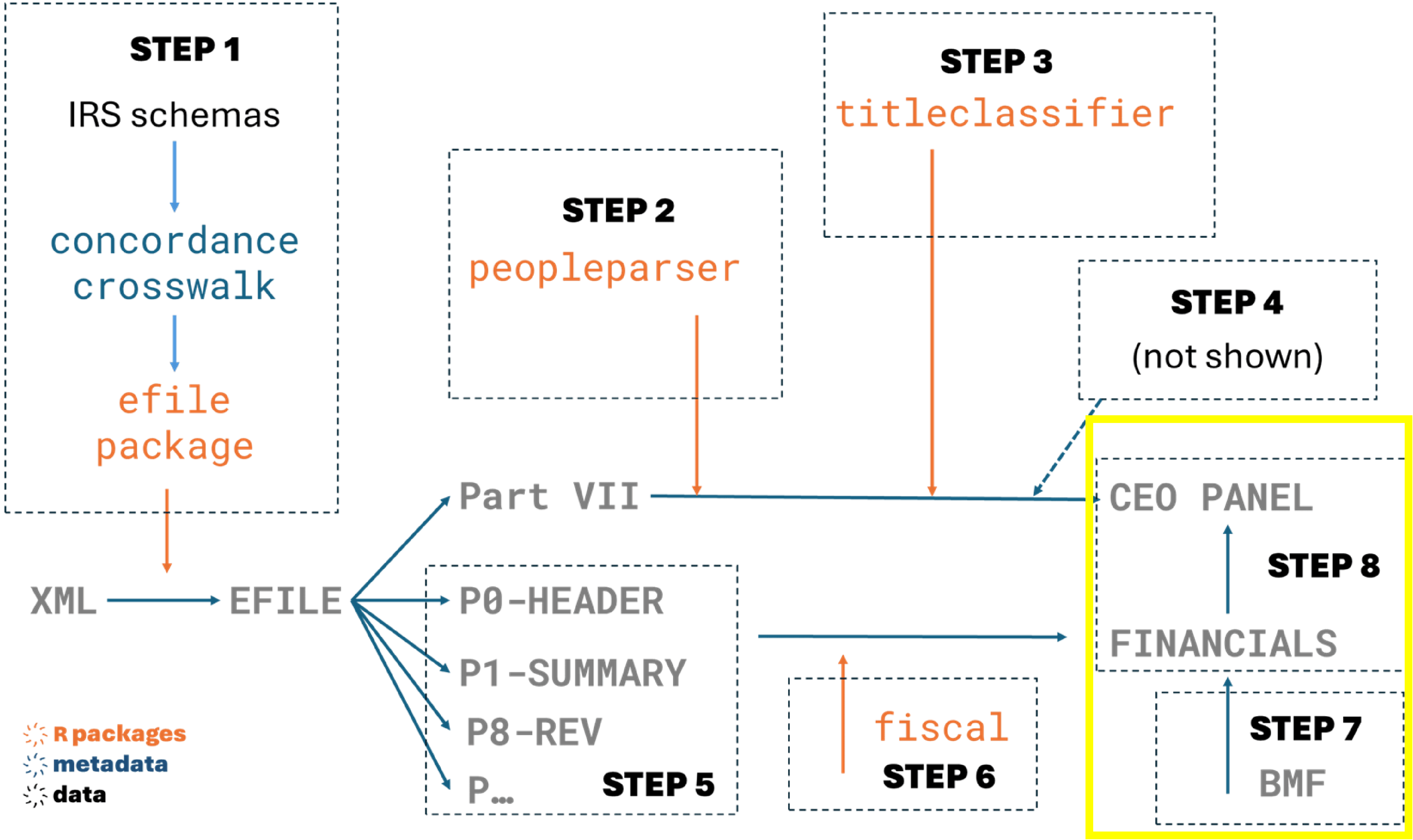


Step 6

```
ceo_trans_1000_fncl <- get_podpm(ceo_trans_1000_fncl)
```

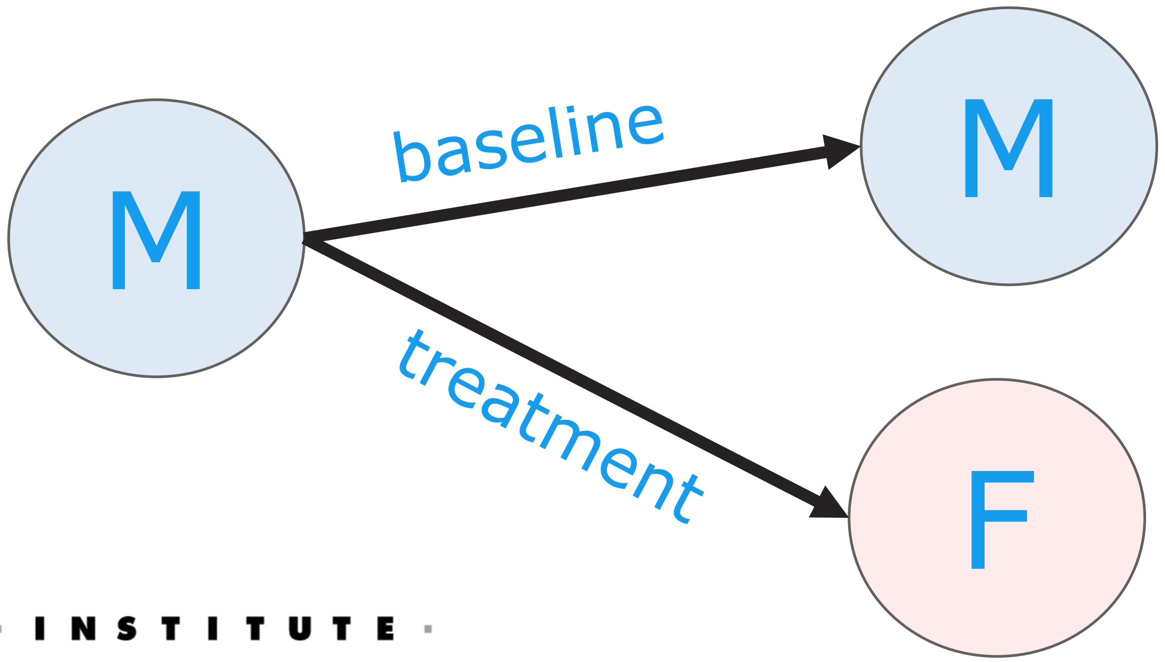


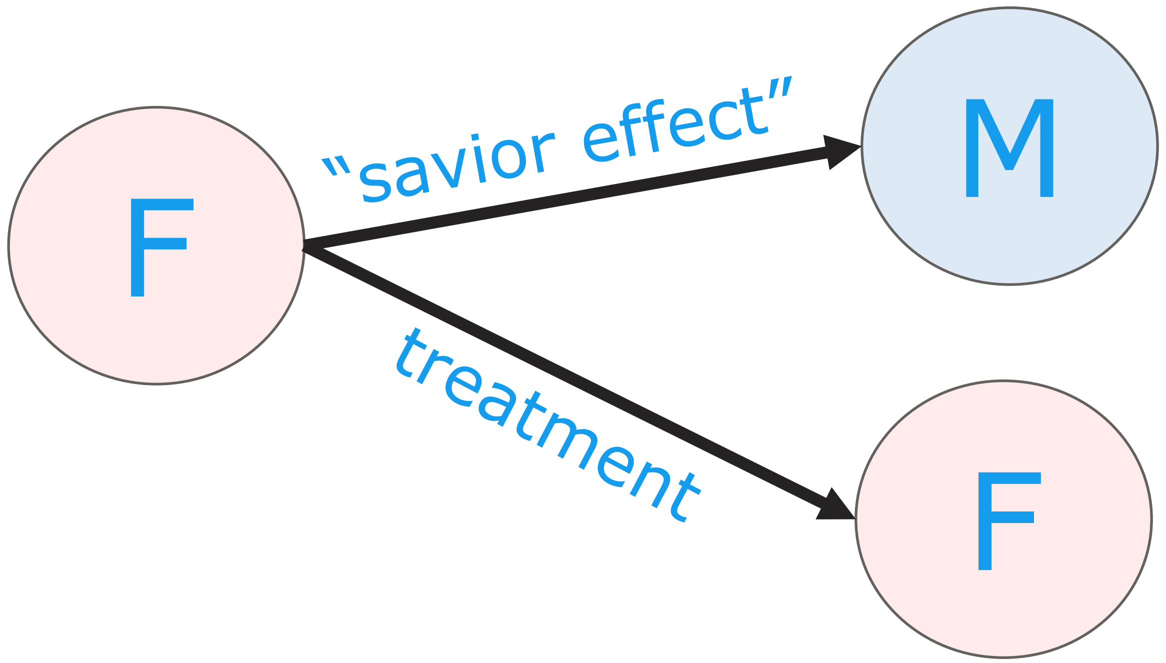
Step 7 & 8


STEP 1
IRS schemas
concordance
crosswalk
efile
package
STEP 2
peopleparser
STEP 3
titleclassifier
STEP 4
(not shown)
XML
EFILE
Part VII
P0-HEADER
P1-SUMMARY
P8-REV
P...
STEP 5
fiscal
STEP 6
CEO PANEL
STEP 8
FINANCIALS
STEP 7
BMF
R packages
metadata
data

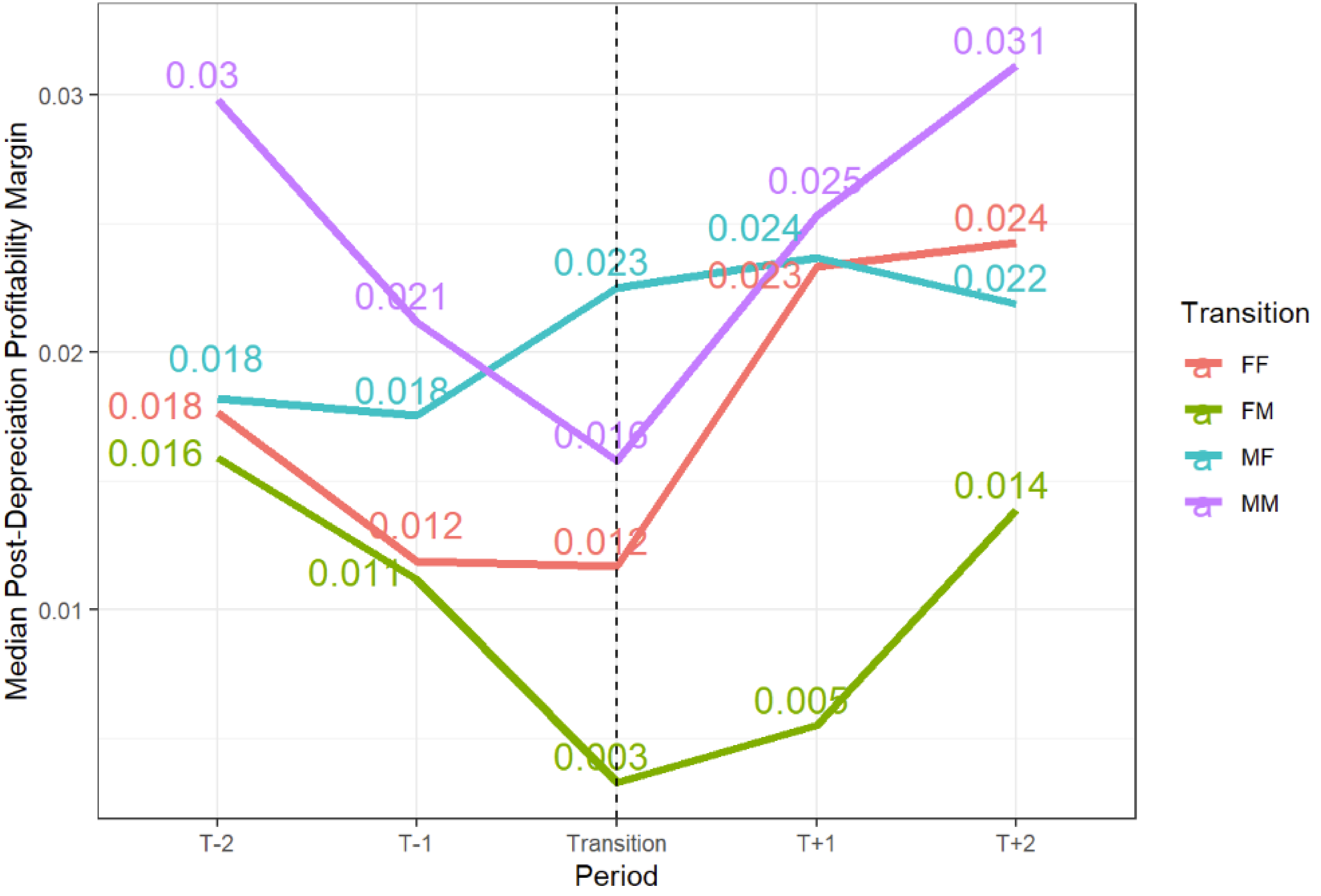
Processed Data Visual

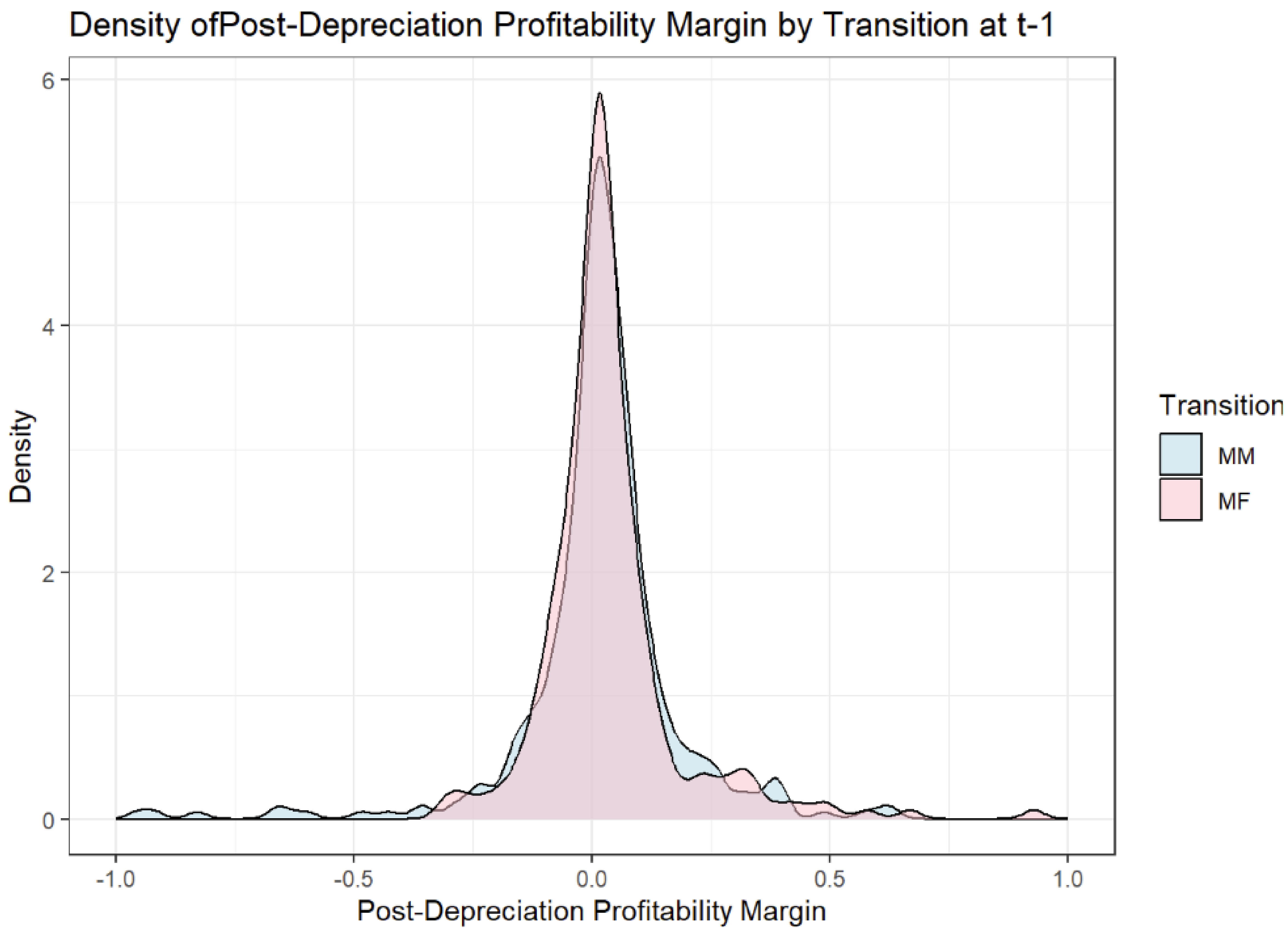
EIN	TAXYR	CEO.1.TM2	CEO.1.TM1	CEO.1	CEO.2	CEO.1.TP1	CEO.1.TP2
10024245	2015	JOHN PORTER	JOHN PORTER	JOHN PORTER	NA	DEB NEUMAN	DEB NEUMAN
10196194	2016	NORMAND DUBREUIL	NORMAND DUBREUIL	NORMAND DUBREUIL	COLE TUCKER	COLE TUCKER	COLE TUCKER
10206603	2012	DONNA STECKINO	DONNA STECKINO	KERRY WOOD	NA	KERRY WOOD	KERRY WOOD
10206603	2014	KERRY WOOD	KERRY WOOD	KERRY WOOD	NA	JENNIFER HOGAN	JENNIFER HOGAN
10211478	2015	JOHN KUROPCHAK	JOHN KUROPCHAK	JOHN KUROPCHAK	NA	SHIRAR PATTERSON	SHIRAR PATTERSON





Median Post-Depreciation Profitability Margin by Transition Type




Density ofPost-Depreciation Profitability Margin by Transition at t-1
Density
Post-Depreciation Profitability Margin
0
2
4
6
-1.0
-0.5
0.0
0.5
1.0
Transition
MM
MF