ICPSR/Retraction Watch, First Look

Ashvin Narayan

26 September 2025

DOIs available?

```
# What proportion of entries in each table have DOIs (ICPSR, Retraction Watch)
c(length(icpsr$DOI[which(!is.na(icpsr$DOI))]) / length(icpsr$DOI),
  length(retractions$OriginalPaperDOI[which(!is.na(retractions$OriginalPaperDOI))]) /
    length(retractions$OriginalPaperDOI))
```

```
## [1] 0.4917015 0.9444636
```

```
# Duplicate DOIs in retractions?
retractions |>
  drop_na(OriginalPaperDOI) |>
  add_count(OriginalPaperDOI) |>
  filter(n > 1) |>
  distinct() |>
  select("Record.ID", "Title", "OriginalPaperDOI", "n") |>
  nrow()
```

```
## [1] 10821
```

```
# Retraction Watch database contains 'duplicate' DOIs - if a paper
# was corrected then retracted, there are separate entries for the
# correction and the retraction

# pivot_wider before merging?

# How many entries for each type of retraction reason?
retractions |>
  drop_na(RetractionNature) |>
  group_by(RetractionNature) |>
  summarize(n())
```

```
## # A tibble: 4 x 2
##   RetractionNature      `n()`
##   <chr>                 <int>
## 1 Correction             1362
## 2 Expression of concern  3440
## 3 Reinstatement           158
## 4 Retraction            61681
```

```
# Duplicate DOIs in ICPSR?

icpsr |>
  drop_na(DOI) |>
```

```
  add_count(DOI) |>
  filter(n > 1) |>
  distinct() |>
  select("Reference.ID", "Title", "DOI", "n") |>
  nrow()
```

```
## [1] 0
```

```
# Doesn't look like it! Thank you data librarians :)
```

Article Types

```
# For entries without DOIs in ICPSR, what type of materials are they?

# 'Type.of.Work' available for what proportion of icpsr entries?
icpsr |>
  drop_na(Type.of.Work) |>
  select("Reference.ID", "Title", "DOI", "Type.of.Work") |>
  nrow() / nrow(icpsr)
```

```
## [1] 0.09660275
```

```
# 'Type.of.Work' available for what proportion of icpsr entries
# that have a DOI?
icpsr |>
  drop_na(DOI) |>
  drop_na(Type.of.Work) |>
  select("Reference.ID", "DOI", "Type.of.Work") |>
  nrow() / nrow(filter(icpsr, is.na(DOI)))
```

```
## [1] 0.005351468
```

```
# ... that don't have a DOI?
icpsr |>
  filter(is.na(DOI)) |>
  drop_na(Type.of.Work) |>
  select("Reference.ID", "Title", "DOI", "Type.of.Work") |>
  nrow() / nrow(filter(icpsr, is.na(DOI)))
```

```
## [1] 0.1846997
```

```
# 'Type.of.Work' appears to be available for < 10% of entries -
# around 5.4% of entries with a DOI and 18.5% of those w/o a DOI

# TODO: fix NA values - diff NA values associated w diff columns

# How many entries associated with each type of work?
icpsr |>
  drop_na(Type.of.Work) |>
  filter(Type.of.Work != "(unknown)") |>
  group_by(Type.of.Work) |>
  summarize(count = n()) |>
  arrange(desc(count))
```

```
## # A tibble: 270 x 2
##    Type.of.Work              count
##    <chr>                     <int>
##  1 "Dissertation"             7988
##  2 "Thesis"                   1885
##  3 "dissertation"              120
##  4 "unpublished manuscript"    106
##  5 "Mimeograph"                 82
##  6 "Working paper"              82
##  7 "Instrument "                70
##  8 "Association Paper"          53
##  9 "[Data Profile; Website]"    50
```

```
## 10 "[Preprint]"                46
## # i 260 more rows
```

```
# Needs to be standardized .. capitalization differences interfering
```

```
# 'ArticleType' unavailable for what proportion of RW entries?
retractions |>
  select(Record.ID, ArticleType) |>
  drop_na(ArticleType) |>
  nrow()
```

```
## [1] 66641
```

```
# ArticleType available for all RW entries, it looks like!

# How many entries associated with each type of work in RW?
retractions |>
  select(Record.ID, ArticleType) |>
  drop_na(ArticleType) |>
  unnest(c(ArticleType)) |>
  group_by(ArticleType) |>
  summarize(count = n()) |>
  arrange(desc(count))
```

```
## # A tibble: 26 x 2
##    ArticleType                count
##    <chr>                      <int>
##  1 Research Article           45084
##  2 Conference Abstract/Paper  13280
##  3 Clinical Study              2970
##  4 Review Article              2581
##  5 Meta-Analysis                875
##  6 Case Report                  868
##  7 Book Chapter/Reference Work  576
##  8 Article in Press             519
##  9 Letter                       517
## 10 Commentary/Editorial         443
## # i 16 more rows
```

```
# These overlap - ArticleType is a list-type column
```

Publishers

```
# Publication name available for what proportion of entries?

# Secondary.Title includes publication name if entry is
# a journal/newspaper
icpsr |>
  select(Reference.ID, Secondary.Title) |>
  drop_na(Secondary.Title) |>
  nrow() / nrow(icpsr)
```

```
## [1] 0.8259025
```

```
# Now looking at Publisher - contains info on gov't
# bureaus etc responsible for the publication
# Approx equivalent to 'Institution' in RW data?
icpsr |>
  select(Reference.ID, Publisher) |>
  drop_na(Publisher) |>
  nrow() / nrow(icpsr)
```

```
## [1] 0.3594607
```

```
# Around 36% of icpsr entries list the publisher

# What about those entries without DOI?

# Secondary.Title
icpsr |>
  filter(is.na(DOI)) |>
  select(Reference.ID, Secondary.Title) |>
  drop_na(Secondary.Title) |>
  nrow() / nrow(filter(icpsr, is.na(DOI)))
```

```
## [1] 0.6794059
```

```
# Publisher
icpsr |>
  filter(is.na(DOI)) |>
  select(Reference.ID, Publisher) |>
  drop_na(Publisher) |>
  nrow() / nrow(filter(icpsr, is.na(DOI)))
```

```
## [1] 0.6396239
```

```
# Around 64% of entries without a DOI have a publisher listed

# What publications appear frequently in ISPCR?

# Secondary.Title
icpsr |>
  select(Reference.ID, Secondary.Title) |>
  drop_na(Secondary.Title) |>
  group_by(Secondary.Title) |>
  summarize(count = n()) |>
  arrange(desc(count))
```

```
## # A tibble: 19,705 x 2
##    Secondary.Title                                                      count
##    <chr>                                                                <int>
##  1 ProQuest Dissertations and Theses                                     1021
##  2 Journal of Marriage and Family                                         971
##  3 American Journal of Public Health                                      842
##  4 Journals of Gerontology, Series B: Psychological Sciences and Social S~   649
##  5 Social Forces                                                          635
##  6 Social Science Quarterly                                               627
##  7 annual meeting of the American Political Science Association           618
##  8 American Sociological Review                                           586
##  9 American Journal of Political Science                                  583
## 10 American Political Science Review                                      556
## # i 19,695 more rows
```

```
# Publisher
icpsr |>
  select(Reference.ID, Publisher) |>
  drop_na(Publisher) |>
  group_by(Publisher) |>
  summarize(count = n()) |>
  arrange(desc(count))
```

```
## # A tibble: 6,988 x 2
##    Publisher                                                            count
##    <chr>                                                                <int>
##  1 American Society of Criminology                                        965
##  2 United States Department of Justice, National Institute of Justice     879
##  3 U.S. Department of Health and Human Services, Administration for Child~   631
##  4 National Bureau of Economic Research                                   603
##  5 United States Department of Justice, Bureau of Justice Statistics      600
##  6 American Sociological Association                                      573
##  7 United States Department of Education, Office of Educational Research ~   448
##  8 University of Michigan                                                 428
##  9 Substance Abuse and Mental Health Services Administration              420
## 10 University of North Carolina at Chapel Hill, Carolina Population Center   419
## # i 6,978 more rows
```

```
# Publication name available for what proportion of entries in RW?

# Secondary.Title includes publication name if entry is
# a journal/newspaper
retractions |>
  select(Record.ID, Journal) |>
  drop_na(Journal) |>
  nrow() / nrow(retractions)
```

```
## [1] 1
```

```
# All entries list a 'Journal' - column just holds the source
# of the article (incl. journals, books, serials, etc)

# What journals are most common in RW?

retractions |>
  select(Record.ID, Journal) |>
```

```
  drop_na(Journal) |>
  unnest(c(Journal)) |>
  group_by(Journal) |>
  summarize(count = n()) |>
  arrange(desc(count))
```

```
## # A tibble: 8,496 x 2
##    Journal                                                                count
##    <chr>                                                                  <int>
##  1 Journal of Intelligent & Fuzzy Systems                                  1566
##  2 2011 International Conference on E-Business and E-Government (ICEE)     1280
##  3 PLoS One                                                                1224
##  4 2011 5th International Conference on Bioinformatics and Biomedical Eng~  1084
##  5 Journal of Healthcare Engineering                                       1074
##  6 Computational and Mathematical Methods in Medicine                      1067
##  7 Computational Intelligence and Neuroscience                             1028
##  8 BioMed Research International                                            953
##  9 Security and Communication Networks                                      949
## 10 Journal of Physics: Conference Series                                    878
## # i 8,486 more rows
```

Merging

```
# Initial brief attempt at a merge

merged <- inner_join(
  x = (icpsr %>%
         filter(!is.na(DOI)) %>%
         select(Reference.ID, Title, DOI) %>%
         unnest(c(DOI))),
  y = (retractions %>%
         filter(!is.na(OriginalPaperDOI)) %>%
         select(Record.ID, Title, OriginalPaperDOI) %>%
         unnest(c(OriginalPaperDOI))),
  by = join_by(DOI == OriginalPaperDOI)
)

print(merged)
```

```
## # A tibble: 9 x 5
##   Reference.ID Title.x                                    DOI   Record.ID Title.y
##          <int> <chr>                                      <chr> <list>    <list>
## 1        11313 Worlds Apart? The Reception of Genetical~ 10.1~ <chr [1]> <chr>
## 2        15930 Nonmetropolitan sex-role ideologies: A l~ 10.1~ <chr [1]> <chr>
## 3        29220 School social bonds, school climate, and~ 10.1~ <chr [1]> <chr>
## 4        87821 Suicide after natural disasters           10.1~ <chr [1]> <chr>
## 5       140614 Electronic cigarette use and myocardial ~ 10.1~ <chr [1]> <chr>
## 6       140614 Electronic cigarette use and myocardial ~ 10.1~ <chr [1]> <chr>
## 7       142038 The verdict is in: How did they decide? ~ 10.1~ <chr [1]> <chr>
## 8       144102 Chronic adolescent marijuana use as a ri~ 10.1~ <chr [1]> <chr>
## 9       155116 Tuning in, not turning out: Evaluating t~ 10.1~ <chr [1]> <chr>
```

```
# Is 5 observations around what we'd expect??
# Slightly concerning might be a data cleaning problem
```