

Book Data Linkage Statistics

This notebook presents statistics of the book data integration.

Setup

```
library(tidyverse, warn.conflicts=FALSE)
library(arrow, warn.conflicts=FALSE)
library(jsonlite)
```

— Attaching core tidyverse packages — tidyverse 2.0.0 —

| | | | |
|-------------|-------|-----------|-------|
| ✓ dplyr | 1.1.4 | ✓ readr | 2.1.4 |
| ✓ forcats | 1.0.0 | ✓ stringr | 1.5.1 |
| ✓ ggplot2 | 3.4.4 | ✓ tibble | 3.2.1 |
| ✓ lubridate | 1.9.3 | ✓ tidyr | 1.3.0 |
| ✓ purrr | 1.0.2 | | |

— Conflicts — tidyverse_conflicts() —

✖ dplyr::filter() masks stats::filter()

✖ dplyr::lag() masks stats::lag()

i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

Attaching package: ‘jsonlite’

The following object is masked from ‘package:purrr’:

flatten

I want to use `theme_minimal()` by default:

```
theme_set(theme_minimal())
```

And default image sizes aren’t great:

```
options(repr.plot.width = 7,
        repr.plot.height = 4)
```

Load Link Stats

We compute dataset linking statistics as `gender-stats.csv` as part of the integration. Let’s load those:

```
link_stats = read_csv("book-links/gender-stats.csv")
glimpse(link_stats)
```

Rows: 46 Columns: 4

— Column specification —

Delimiter: ","

chr (2): dataset, gender

dbl (2): n_books, n_actions

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.

Rows: 46

Columns: 4

\$ dataset <chr> "LOC-MDS", "LOC-MDS", "LOC-MDS", "LOC-MDS", "LOC-MDS", "LOC-...

\$ gender <chr> "no-author-rec", "no-book-author", "unknown", "female", "mal...

\$ n_books <dbl> 305932, 598118, 1102373, 736825, 2414561, 73411, 71367, 3112...

\$ n_actions <dbl> NA, NA, NA, NA, NA, NA, NA, 392413, 67273, 47065, 18583, 466570,...

Now let's define variables for our various codes. We are first going to define our gender codes. We'll start with the resolved codes:

```
link_codes = c('female', 'male', 'ambiguous', 'unknown')
```

We want the unlink codes in order, so the last is the first link failure:

```
unlink_codes = c('no-author-rec', 'no-book-author', 'no-book')
```

```
all_codes = c(link_codes, unlink_codes)
```

Processing Statistics

Now we'll pivot each of our count columns into a table for easier reference.

```
book_counts = link_stats %>%
  pivot_wider(id_cols=dataset, names_from=gender, values_from=n_books) %>%
  replace(is.na(.), 0) %>%
  mutate(total=rowSums(across(-dataset)))
glimpse(book_counts)
```

Rows: 7

Columns: 9

\$ dataset <chr> "LOC-MDS", "BX-I", "BX-E", "AZ14", "AZ18", "GR-I", "G...

\$ `no-author-rec` <dbl> 305932, 11557, 5703, 157140, 243065, 62332, 61224

\$ `no-book-author` <dbl> 598118, 10872, 5439, 166954, 151433, 746315, 734560

\$ unknown <dbl> 1102373, 31126, 15152, 244288, 309432, 113312, 111379

```

$ female      <dbl> 736825, 71367, 40089, 247131, 314233, 225786, 223527
$ male        <dbl> 2414561, 102840, 58599, 550289, 669333, 338392, 334124
$ ambiguous    <dbl> 73411, 9573, 5670, 24293, 28300, 19115, 18917
$ `no-book`    <dbl> 0, 34856, 17403, 866957, 1138027, 0, 0
$ total        <dbl> 5231220, 272191, 148055, 2257052, 2853823, 1505252, 1...

```

```

act_counts = link_stats %>%
  filter(dataset != "LOC-MDS") %>%
  pivot_wider(id_cols=dataset, names_from=gender, values_from=n_actions) %>%
  replace(is.na(.), 0) %>%
  mutate(total=rowSums(across(-dataset)))
glimpse(act_counts)

```

Rows: 6

Columns: 9

```

$ dataset      <chr> "BX-I", "BX-E", "AZ14", "AZ18", "GR-I", "GR-E"
$ female       <dbl> 392413, 138698, 4905977, 12113491, 81501002, 35570655
$ unknown      <dbl> 67273, 24009, 2222075, 4980474, 10844936, 3717539
$ `no-book`    <dbl> 47065, 19828, 3852790, 9908335, 0, 0
$ `no-author-rec` <dbl> 18583, 7129, 1115464, 3347274, 3541627, 1031982
$ male         <dbl> 466570, 182935, 7090496, 15541958, 69838115, 33133627
$ ambiguous    <dbl> 116881, 46970, 909797, 2005910, 23242103, 14015868
$ `no-book-author` <dbl> 18935, 7246, 2329752, 2758328, 29517391, 11094188
$ total        <dbl> 1127720, 426815, 22426351, 50655770, 218485174, 98563...

```

We're going to want to compute versions of this table as fractions, e.g. the fraction of books that are written by women. We will use the following helper function:

```

fractionalize = function(data, columns, unlinked=NULL) {
  fracs = select(data, dataset | all_of(columns))
  if (!is.null(unlinked)) {
    fracs = mutate(fracs, unlinked=rowSums(select(data, all_of(unlinked))))
  }
  totals = rowSums(select(fracs, !dataset))
  fracs %>% mutate(across(!dataset, ~ .x / totals))
}
fractionalize(book_counts, link_codes) %>% glimpse()

```

Rows: 7

Columns: 5

```

$ dataset      <chr> "LOC-MDS", "BX-I", "BX-E", "AZ14", "AZ18", "GR-I", "GR-E"
$ female       <dbl> 0.1702787, 0.3320847, 0.3354447, 0.2318300, 0.2378214, 0.324...
$ male         <dbl> 0.5580000, 0.4785348, 0.4903272, 0.5162181, 0.5065723, 0.485...
$ ambiguous    <dbl> 0.01696513, 0.04454506, 0.04744373, 0.02278891, 0.02141833, ...
$ unknown      <dbl> 0.2547561, 0.1448354, 0.1267844, 0.2291630, 0.2341879, 0.162...

```

And a helper function for plotting bar charts:

```

plot_bars = function(data, what="UNSPECIFIED") {
  tall = data %>%
    pivot_longer(!dataset, names_to="status", values_to="fraction")
  codes = c(all_codes, "unlinked")
  codes = intersect(codes, unique(tall$status))
  tall = tall %>% mutate(status=ordered(status, codes))
  ggplot(tall) +
    aes(y=dataset, x=fraction, fill=status) +
    geom_col(position=position_stack(reverse=TRUE), width=0.5) +
    geom_text(aes(label=if_else(fraction >= 0.1,
                                sprintf("%.1f%%", fraction * 100),
                                "")),
              position=position_stack(reverse=TRUE, vjust=0.5),
              colour="white", fontface="bold") +
    scale_fill_brewer(type="qual", palette="Dark2") +
    ylab("Dataset") +
    xlab(paste("Fraction of", what)) +
    labs(fill="Author Gender")
}

```

Resolution of Books

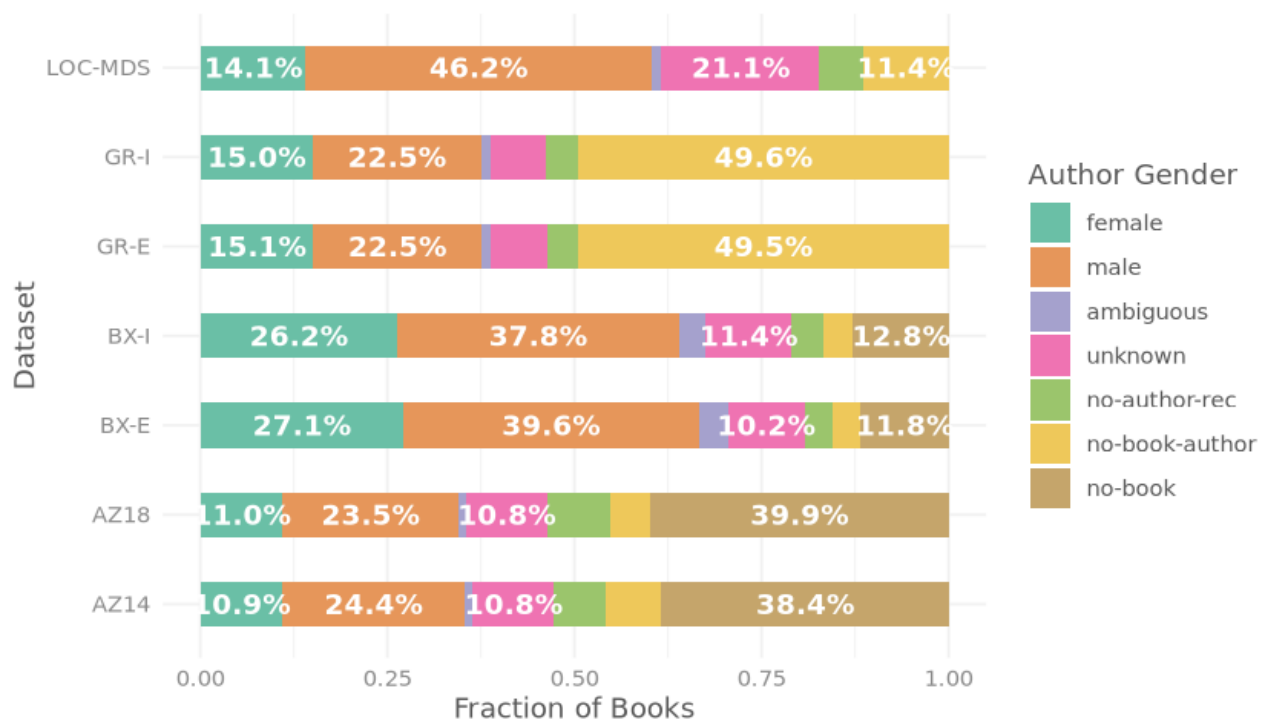
What fraction of *unique books* are resolved from each source?

```
book_counts %>% fractionalize(all_codes)
```

A tibble: 7 × 8

| dataset | female | male | ambiguous | unknown | no-author- rec | no-book- author | no-book |
|---------|-----------|-----------|-------------|------------|-------------------|--------------------|-----------|
| <chr> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> |
| LOC-MDS | 0.1408515 | 0.4615675 | 0.014033247 | 0.21072962 | 0.05848196 | 0.11433624 | 0.0000000 |
| BX-I | 0.2621946 | 0.3778229 | 0.035170156 | 0.11435352 | 0.04245916 | 0.03994254 | 0.1280571 |
| BX-E | 0.2707710 | 0.3957921 | 0.038296579 | 0.10234035 | 0.03851947 | 0.03673635 | 0.1175442 |
| AZ14 | 0.1094928 | 0.2438087 | 0.010763155 | 0.10823322 | 0.06962179 | 0.07396994 | 0.3841103 |
| AZ18 | 0.1101095 | 0.2345391 | 0.009916523 | 0.10842719 | 0.08517172 | 0.05306321 | 0.3987728 |
| GR-I | 0.1499988 | 0.2248075 | 0.012698870 | 0.07527776 | 0.04140968 | 0.49580735 | 0.0000000 |
| GR-E | 0.1506520 | 0.2251918 | 0.012749616 | 0.07506684 | 0.04126354 | 0.49507626 | 0.0000000 |

```
book_counts %>% fractionalize(all_codes) %>% plot_bars("Books")
```

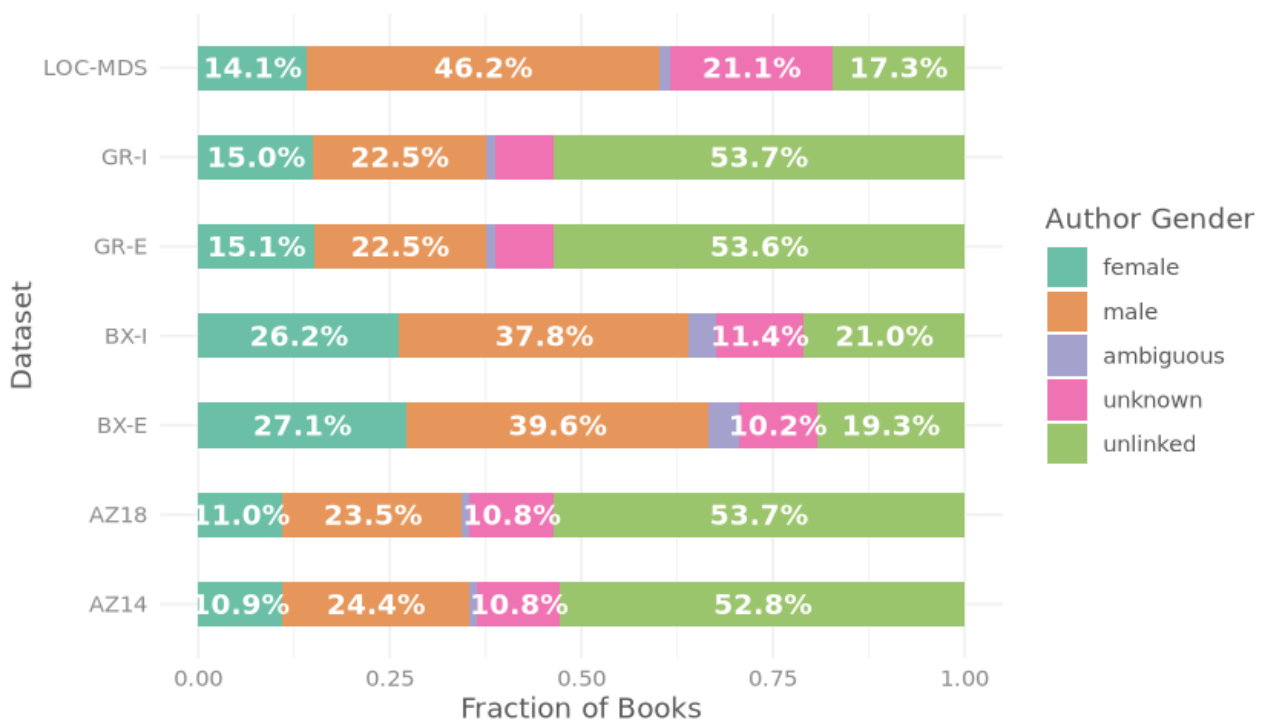


```
book_counts %>% fractionalize(link_codes, unlink_codes)
```

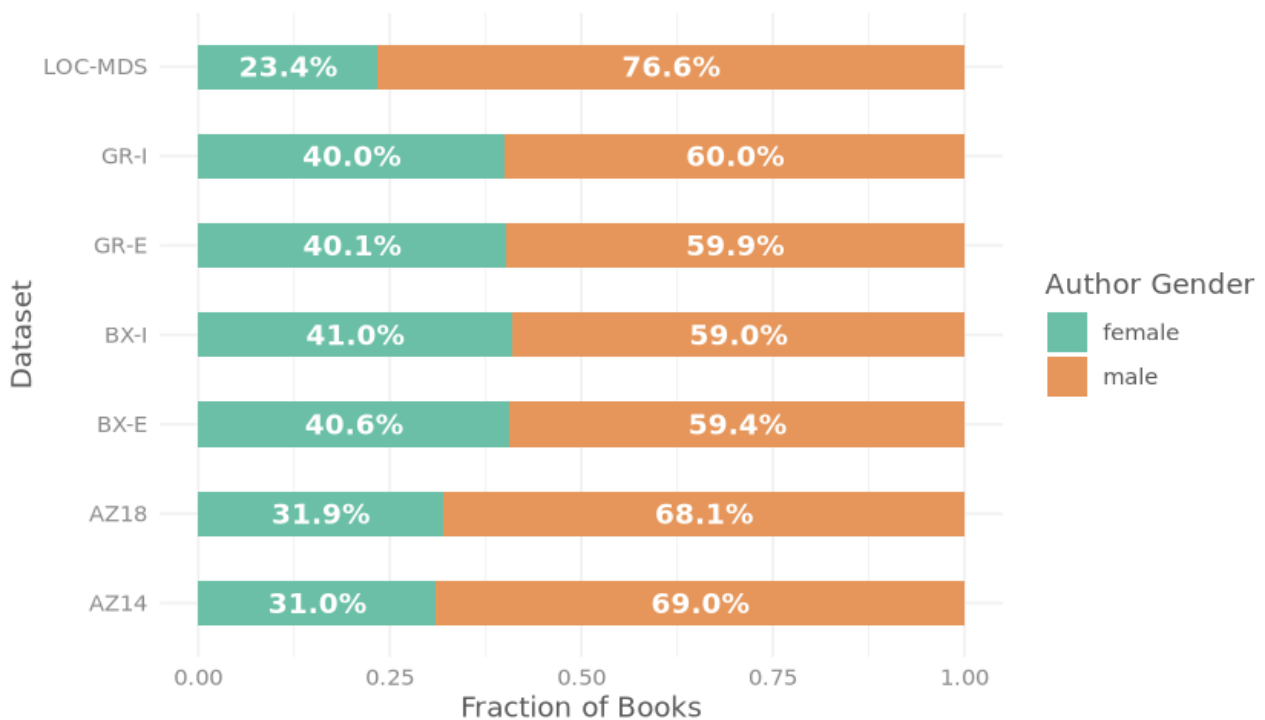
A tibble: 7 × 6

| dataset | female | male | ambiguous | unknown | unlinked |
|---------|-----------|-----------|-------------|------------|-----------|
| <chr> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> |
| LOC-MDS | 0.1408515 | 0.4615675 | 0.014033247 | 0.21072962 | 0.1728182 |
| BX-I | 0.2621946 | 0.3778229 | 0.035170156 | 0.11435352 | 0.2104588 |
| BX-E | 0.2707710 | 0.3957921 | 0.038296579 | 0.10234035 | 0.1928000 |
| AZ14 | 0.1094928 | 0.2438087 | 0.010763155 | 0.10823322 | 0.5277021 |
| AZ18 | 0.1101095 | 0.2345391 | 0.009916523 | 0.10842719 | 0.5370077 |
| GR-I | 0.1499988 | 0.2248075 | 0.012698870 | 0.07527776 | 0.5372170 |
| GR-E | 0.1506520 | 0.2251918 | 0.012749616 | 0.07506684 | 0.5363398 |

```
book_counts %>% fractionalize(link_codes, unlink_codes) %>% plot_bars("Books")
```



```
book_counts %>% fractionalize(c('female', 'male')) %>% plot_bars("Books")
```



Resolution of Ratings

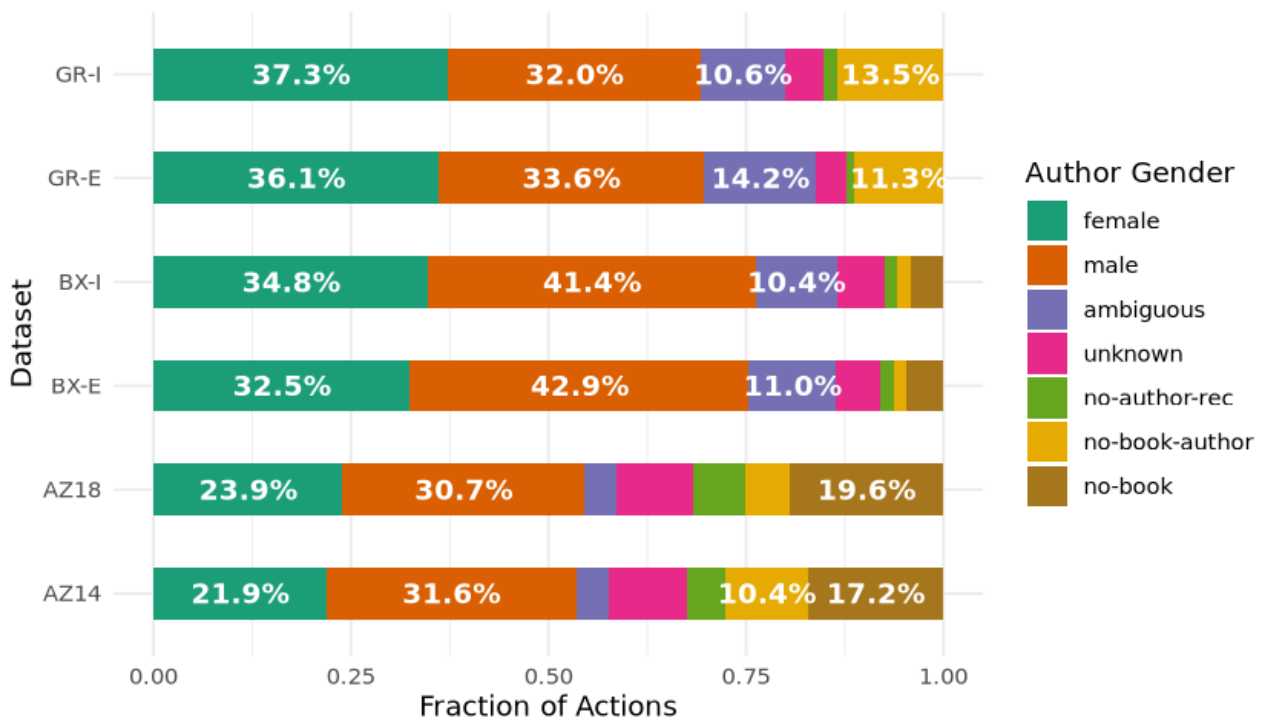
What fraction of *rating actions* have each resolution result?

```
act_counts %>% fractionalize(all_codes)
```

A tibble: 6 × 8

| dataset | female | male | ambiguous | unknown | no-author-rec | no-book-author | no-book |
|---------|-----------|-----------|------------|------------|---------------|----------------|------------|
| <chr> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> |
| BX-I | 0.3479702 | 0.4137286 | 0.10364363 | 0.05965399 | 0.01647838 | 0.01679052 | 0.04173465 |
| BX-E | 0.3249605 | 0.4286049 | 0.11004768 | 0.05625154 | 0.01670279 | 0.01697691 | 0.04645572 |
| AZ14 | 0.2187595 | 0.3161681 | 0.04056821 | 0.09908322 | 0.04973899 | 0.10388458 | 0.17179745 |
| AZ18 | 0.2391335 | 0.3068152 | 0.03959885 | 0.09831997 | 0.06607883 | 0.05445240 | 0.19560131 |
| GR-I | 0.3730276 | 0.3196469 | 0.10637840 | 0.04963694 | 0.01620992 | 0.13510020 | 0.00000000 |
| GR-E | 0.3608894 | 0.3361641 | 0.14220089 | 0.03771706 | 0.01047019 | 0.11255838 | 0.00000000 |

```
act_counts %>% fractionalize(all_codes) %>% plot_bars("Actions")
```



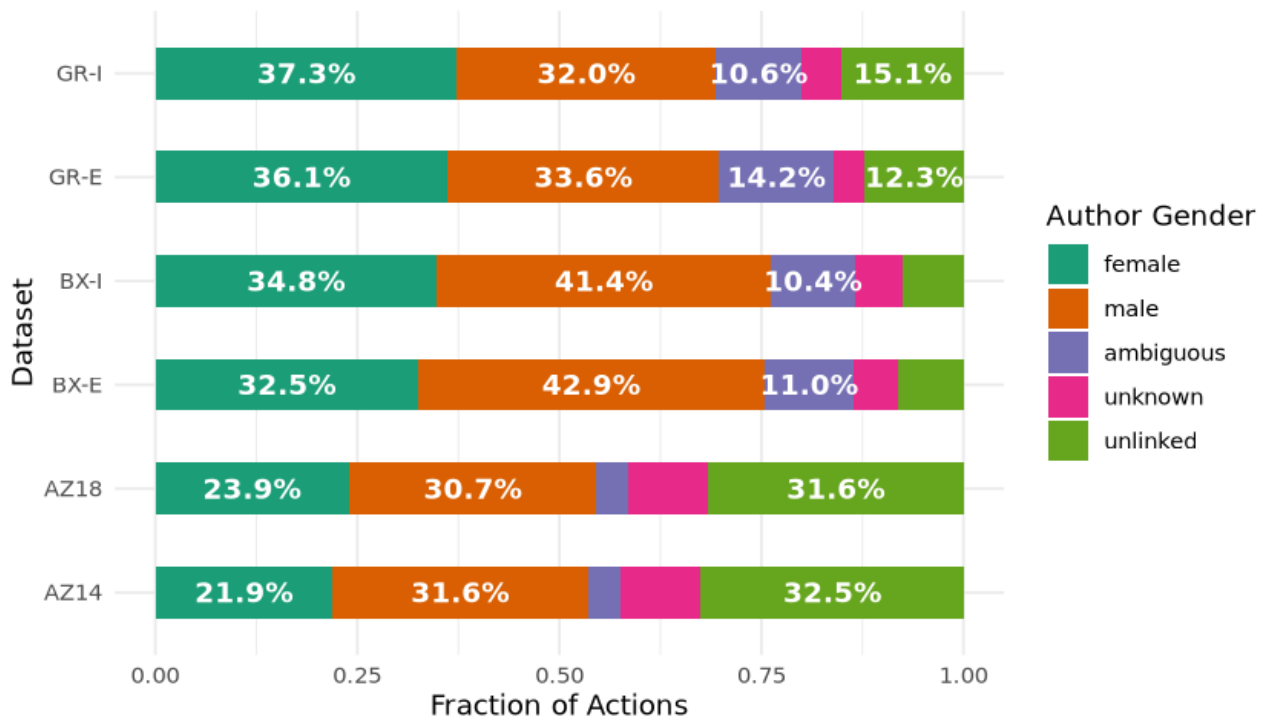
```
act_counts %>% fractionalize(link_codes, unlink_codes)
```

A tibble: 6 × 6

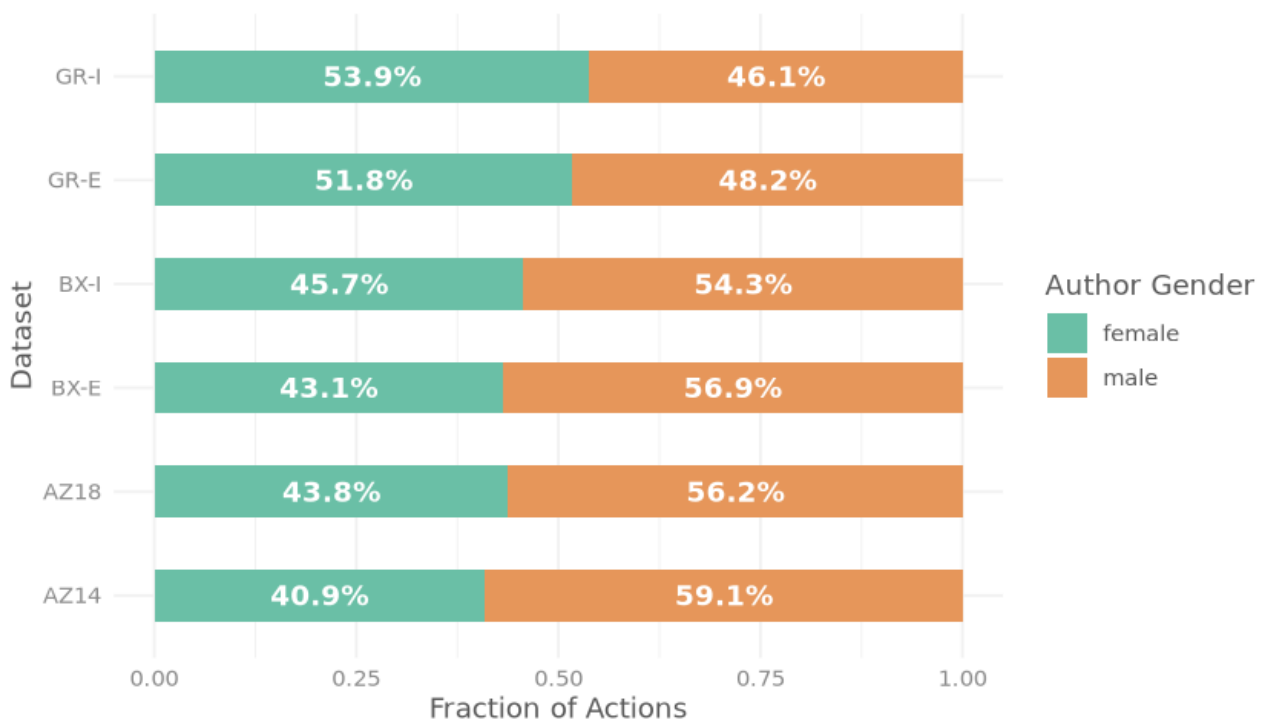
| dataset | female | male | ambiguous | unknown | unlinked |
|---------|-----------|-----------|------------|------------|------------|
| <chr> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> |
| BX-I | 0.3479702 | 0.4137286 | 0.10364363 | 0.05965399 | 0.07500355 |
| BX-E | 0.3249605 | 0.4286049 | 0.11004768 | 0.05625154 | 0.08013542 |
| AZ14 | 0.2187595 | 0.3161681 | 0.04056821 | 0.09908322 | 0.32542102 |

| dataset | female | male | ambiguous | unknown | unlinked |
|---------|-----------|-----------|------------|------------|------------|
| <chr> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> |
| AZ18 | 0.2391335 | 0.3068152 | 0.03959885 | 0.09831997 | 0.31613254 |
| GR-I | 0.3730276 | 0.3196469 | 0.10637840 | 0.04963694 | 0.15131012 |
| GR-E | 0.3608894 | 0.3361641 | 0.14220089 | 0.03771706 | 0.12302856 |

```
act_counts %>% fractionalize(link_codes, unlink_codes) %>% plot_bars("Actions")
```



```
act_counts %>% fractionalize(c('female', 'male')) %>% plot_bars("Actions")
```

Metrics

Finally, we're going to write coverage metrics.

```
book_linked = eval(quote(male + female + ambiguous), envir=book_counts)
book_coverage = book_linked / book_counts$total
book_coverage = setNames(book_coverage, book_counts$dataset)
book_coverage
```

LOC-MDS: 0.616452185149927 **BX-I:** 0.675187643970594 **BX-E:** 0.70485968052413
AZ14: 0.364064718048144 **AZ18:** 0.354565086902727 **GR-I:** 0.387505215073622
GR-E: 0.38859335014231

```
json = toJSON(
  as.list(book_coverage),
  auto_unbox=TRUE,
)
write_file(json, "book-coverage.json")
```