R Markdown (Policy Doc/IDM Lit Analysis)

Emma Chapman-Banks

2025-01-18

Part 1: Cross-Referencing Policy Documents and Abstracts (Same Country)

Load libraries

```
library("tidyverse")
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library("sf")
```

```
## Linking to GEOS 3.11.0, GDAL 3.5.3, PROJ 9.1.0; sf_use_s2() is TRUE
```

```
library("countrycode")
library("ggrepel")
library("gt")
library("dplyr")
library("tidyr")
```

Read the data

```
overton_data <- read.csv("data/overton_results_expanded.csv")
idm_data <- read.csv("data/Combined_Location_With_Metadata.csv")
```

Before creating this R markdown, I ran the analysis by joining the datasets together then cross-referencing the dois and country location. This returned zero results. At first, I was puzzled, but then I realised it was because the countries might be spelt differently or be abbreviated. Therefore, before joining the datasets, I used the Iso3 country codes and standardised the country names across the data sets

Standardise IDM Data. This code is the same code used in the World Mapping Markdown.

```r
# Define manual mappings only for ambiguous single locations
manual_mapping <- list(
  "England and Wales" = "GBR",
  "England & Wales" = "GBR",
  "Fayoum, Egypt" = "EGY",
  "New York metropolitan area" = "USA",
  "Hebei Province, China" = "CHN",
  "Brasilia, Brazil" = "BRA",
  "Michigan, U.S.A.; New York City, U.S.A." = "USA",
  "Edmonton, Alberta, Canada" = "CAN",
  "South India" = "IND",
  "sub-Saharan Africa" = "unmapped",  # Exclude broad regions
  "southeastern United States" = "USA",
  "Kalutara District; Sri Lanka" = "LKA",
  "Lausanne" = "CHE",
  "Britain" = "GBR",
  "England, UK" = "GBR",
  "U.S." = "USA",
  "Mexico; Southern region of Mexico" = "MEX",
  "Washington State" = "USA",
  "Europe" = "unmapped",
  "Kalutara District" = "LKA"
)

# Process `llm_location` column
idm_data_clean <- idm_data %>%
  filter(!is.na(llm_location) & llm_location != "") %>%  # Remove blanks/NA
  separate_rows(llm_location, sep = ";\\s*") %>%         # Split multiple countries
  mutate(
    llm_location_resolved = recode(llm_location, !!!manual_mapping, .default = llm_location),
    Iso3 = ifelse(
      llm_location_resolved %in% c("BRA", "CAN", "CHE", "CHN", "EGY", "GBR", "IND", "LKA"), # Skip rema
      llm_location_resolved,
      countrycode::countrycode(
        sourcevar = llm_location_resolved,
        origin = "country.name",
        destination = "iso3c",
        warn = FALSE  # Suppress warnings for unmatched values
      )
    )
  ) %>%
  mutate(Iso3 = ifelse(is.na(Iso3), "unmapped", Iso3))

# Add Status column to indicate whether a location is mapped or unmapped
idm_data_clean <- idm_data_clean %>%
  mutate(Status = ifelse(Iso3 == "unmapped", "Unmapped", "Mapped"))

# Identify unmatched values for review
unmatched <- idm_data_clean %>%
  filter(Status == "Unmapped") %>%
  distinct(llm_location_resolved) %>%
  arrange(llm_location_resolved)
```

```r
# Print unmatched values for verification. This is to check and confirm that there hasn't been any coun
print("Unmatched Values:")
```

```
## [1] "Unmatched Values:"
```

```r
print(unmatched)
```

```
## # A tibble: 1 x 1
##   llm_location_resolved
##   <chr>
## 1 unmapped
```

When trying to standarise the country codes for Overton data, I also realised that the new column it creates returns a value of NA if the country code is already written. Thus, I identified which country codes were already listed in the dataset and had to include the manual country mapping and ensure it doesn't pass as NA when doing country code mapping.

```r
# Define manual mappings for unmatched values
manual_country_mapping <- c(
  "IGO" = "IGO",
  "EU" = "EU"
)

# Standardise Overton Data and include manual mapping
overton_data_clean <- overton_data %>%
  mutate(
    country_iso3 = if_else(
      country %in% names(manual_country_mapping),
      manual_country_mapping[country],
      countrycode::countrycode(
        sourcevar = country,
        origin = "country.name",
        destination = "iso3c",
        warn = FALSE
      )
    )
  )
```

Now, cross-reference the two datasets.

```r
# Combine the datasets based on DOI and use the ISO3 country codes for comparison
combined_data <- overton_data_clean %>%
  inner_join(idm_data_clean, by = c("source_doi" = "doi"))

# Filter where the policy document's country matches the IDM abstract's location
matching_countries <- combined_data %>%
  filter(country_iso3 == Iso3)

# Count the number of matches
number_of_matches <- nrow(matching_countries)

# Print the result
print(number_of_matches)
```

```
## [1] 5
```

```
#View the rows where the policy document's country matches the IDM literature's location (cleaned versi
matching_countries_clean <- matching_countries %>%
  select(total_citations, source_title, document_title, country, topics, title)

matching_countries_clean %>%
  head(10) %>%  # Show the first 10 rows (adjust as needed)
  ggplot(aes(x = "", y = source_title, label = document_title)) +
  geom_text() +
  theme_void()
```

Post Office Box 111149 Juneau, AK 99811 Main : 907.465.4855

)RTATION WORKERS AND PASSENGERS FROM COVID-19; CAPS Safety Lessons Learned and R

d, and next steps : remote hearing before the Committee on Transportation and Infrastructure, Ho

Occupational Exposure to COVID−19; Emergency Temporary Standard

```
knitr::kable(matching_countries_clean, caption = "Matching Countries")
```

Table 1: Matching Countries

total_citations	source_type	document_title	country	topics	title
6	House Committees	"Protecting Transportation Workers and Passengers from COVID: Gaps in Safety, Lessons Learned and Next Steps."	USA	HEPA; Cruise ship; COVID-19 pandemic	Estimation of differential occupational risk of COVID-19 by comparing risk factors with case data by occupational group
6	House Committees	PROTECTING TRANSPORTATION WORKERS AND PASSENGERS FROM COVID: GAPS IN SAFETY, LESSONS LEARNED, AND NEXT STEPS	USA	COVID-19 pandemic; COVID-19; Face masks during the COVID-19 pandemic	Estimation of differential occupational risk of COVID-19 by comparing risk factors with case data by occupational group
6	Federal Register	Occupational Exposure to COVID-19; Emergency Temporary Standard	USA	Severe acute respiratory syndrome coronavirus 2; Infection; Mental disorder	Estimation of differential occupational risk of COVID-19 by comparing risk factors with case data by occupational group
6	Government Publishing Office (GPO)	Protecting transportation workers and passengers from COVID : gaps in safety, lessons learned, and next steps : remote hearing before the Committee on Transportation and Infrastructure, House of Representatives, One Hundred Seventeenth Congress, first session, February 4, 2021.	USA	COVID-19 pandemic; Occupational Safety and Health Administration; Face masks during the COVID-19 pandemic	Estimation of differential occupational risk of COVID-19 by comparing risk factors with case data by occupational group
6	State of Alaska	Post Office Box 111149 Juneau, AK 99811 Main : 907.465.4855	USA	Airborne transmission; Medicine; Diseases and disorders	Estimation of differential occupational risk of COVID-19 by comparing risk factors with case data by occupational group

Part 2: Cross-Referencing Policy Documents and Abstracts (Different Country)

Now, let's answer this question: how many policy documents cite IDM literature from a country different from their own?

```r
# Filter where the policy document cites IDM literature from a different country
global_research_citations <- combined_data %>%
  filter(country_iso3 != Iso3)

# Count the number of matches
num_global_citations <- nrow(global_research_citations)

# Print the result
print(num_global_citations)
```

```
## [1] 9
```

```r
#View the rows where the policy document's country does not match the IDM literature's location (cleane
global_research_citations_clean <- global_research_citations %>%
  select(total_citations, source_title, document_title, country, title, llm_location, Iso3)
#print(global_research_citations_clean)

knitr::kable(global_research_citations_clean, caption = "Global Research Citations")
```

Table 2: Global Research Citations

total_citations	source_title	document_title	country	title	llm_location	Iso3
8	European Centre for Disease Prevention and Control	Measles and rubella elimination: communicating the importance of vaccination	EU	Evolutionary game theory and social learning can determine how vaccine scares unfold	England & Wales	GBR
8	Publications Office of the European Union	Measles and rubella elimination : communicating the importance of vaccination.	EU	Evolutionary game theory and social learning can determine how vaccine scares unfold	England & Wales	GBR
8	Government of Italy	11.Poster il morbillo è una malattia seria	Italy	Evolutionary game theory and social learning can determine how vaccine scares unfold	England & Wales	GBR
8	IZA Institute of Labor Economics	Personal Belief Exemptions for School-Entry Vaccinations, Vaccination Rates, and Academic Achievement	Germany	Evolutionary game theory and social learning can determine how vaccine scares unfold	England & Wales	GBR
8	State of Wisconsin	Therapies for Children With Autism Spectrum Disorder: Behavioral Interventions Update	USA	Evolutionary game theory and social learning can determine how vaccine scares unfold	England & Wales	GBR
8	UNESCO	Des maths pour agir : accompagner la prise de décision par la science	IGO	Evolutionary game theory and social learning can determine how vaccine scares unfold	England & Wales	GBR
8	UNESCO	Mathematics for action: supporting science-based decision-making	IGO	Evolutionary game theory and social learning can determine how vaccine scares unfold	England & Wales	GBR
8	World Health Organization	Promotion of behavioural change for health in a heterogeneous population	IGO	Evolutionary game theory and social learning can determine how vaccine scares unfold	England & Wales	GBR

total_citations	source_title	document_title	country	title	llm_location	Iso3
6	Terveyden ja hyvinvoinnin laitos	Koronavirusinfektiot (COVID-19) terveyden- ja sosiaalihuollon työntekijöillä Suomessa 1.2.2020-30.6.2021 : Rekisteripohjainen kohorttitutkimus	Finland	Estimation of differential occupational risk of COVID-19 by comparing risk factors with case data by occupational group	Washington State	USA

Part 3: Abstracts with 1+ countries

Now, let's look at how many papers covered 1 country, 2 countries or 3 countries.

```
# Step 1: Split the countries into separate rows
idm_data_split <- idm_data_clean %>%
  mutate(countries = strsplit(as.character(Iso3), ";")) %>%
  unnest(Iso3) %>%
  mutate(Iso3 = trimws(Iso3))  # Clean up extra spaces

# Step 2: Count the number of countries per paper (group by paper ID)
country_count <- idm_data_split %>%
  group_by(doi) %>%  # Replace `paper_id` with the unique identifier for each paper
  summarize(num_countries = n_distinct(Iso3))

# Step 3: Count how many papers cover 1, 2, 3, etc., countries
coverage_summary <- country_count %>%
  count(num_countries)

# Step 4: Rename columns for better readability
coverage_summary <- coverage_summary %>%
  rename(
    `Number of Countries Covered` = num_countries,
    `Paper Count` = n
  )
# Step 5: View results
print(coverage_summary)
```

```
## # A tibble: 4 x 2
##   ‘Number of Countries Covered‘ ‘Paper Count‘
##                           <int>         <int>
## 1                             1            44
## 2                             2             5
## 3                             3             1
## 4                             6             1
```