

# Japan-Vietnam Research Collaboration

## Data Documentation and Codebook

Your Name

2025-11-07

### Table of contents

<b>1 Overview</b>	<b>3</b>
1.1 Dataset Description . . . . .	3
1.2 Data Processing . . . . .	3
1.3 Key Limitations . . . . .	3
<b>2 Variable Codebook</b>	<b>4</b>
2.1 Identifiers and Basic Information . . . . .	4
id . . . . .	4
DOI . . . . .	4
year . . . . .	4
title . . . . .	4
2.2 Impact and Quality Indicators . . . . .	4
cited . . . . .	4
quartile . . . . .	5
sjr_score . . . . .	5
2.3 Collaboration Variables . . . . .	5
coop . . . . .	5
n_countries . . . . .	6
affiliations . . . . .	6
2.4 Funding Variables . . . . .	6
fund . . . . .	6
Regional Funders . . . . .	6
Funder Sectors . . . . .	7
n_regions . . . . .	7
n_sectors . . . . .	7
bilateral_funding . . . . .	8
2.5 Authorship Variables . . . . .	8
First Author Location . . . . .	8
Author Counts . . . . .	8
prop_vn_authors . . . . .	8
Leadership Flags . . . . .	9
collab_type . . . . .	9
2.6 Open Access . . . . .	10
OA . . . . .	10
2.7 Disciplinary Classification . . . . .	10
Macro-Areas . . . . .	10
Subject Areas . . . . .	11
2.8 Sustainable Development Goals (SDGs) . . . . .	11
SDG Variables . . . . .	11
n_sdg . . . . .	12



# 1 Overview

## 1.1 Dataset Description

**Source:** Scopus bibliometric database

**Query:** AFFILCOUNTRY(japan AND viet\*)

**Date extracted:** October 20, 2025 at 17:30 GMT+7

**Document type:** Journal articles only

**Time period:** 1972–2025

**Total records:** 9,982 articles

**Total columns:** 83

**Variable categories:**

- Identifiers & temporal: 7 (id, DOI, year, phase, post\_2009, post\_2014, post\_vju)
- Impact/Quality: 3 (cited, quartile, sjr\_score)
- Collaboration: 3 (coop, n\_countries, affiliations)
- Funding: 15 (fund, 6 regions, 3 sectors, n\_regions, n\_sectors, bilateral\_funding)
- Authorship: 10 (fa\_vn, fa\_jp, fa\_o, vn\_led, jp\_led, n\_vn\_authors, n\_jp\_authors, n\_vn\_jp\_authors, prop\_vn\_authors, collab\_type)
- Institutional: 1 (vju\_affiliated)
- Open Access: 1 (OA)
- Disciplines: 31 (4 macro-areas + 1 multidisciplinary + 26 subject areas)
- SDGs: 18 (17 individual SDGs + n\_sdg)
- Other: 1 (title)

## 1.2 Data Processing

The raw Scopus export was processed through multiple stages:

1. **Filtering:** Restricted to journal articles (primary research outputs)
2. **Merging:** Integrated external data sources:
  - Journal-level subject classifications (Scopus sources file)
  - SJR scores and quartiles (SCImago portal)
  - Funding information (manual Scopus filter exports by region/sector)
  - SDG alignment (text2sdg package with SDGO system)
3. **Variable construction:** Derived collaboration metrics, authorship patterns, and aggregated indicators
4. **Quality control:** Validated country extraction, standardized names, checked for duplicates

**Processing script:** data\_processing.R

**Output file:** data/processed\_data.csv

## 1.3 Key Limitations

- **SJR missingness:** ~6% of articles could not be matched to SJR data due to title variants
- **Affiliation parsing:** Author counts based on affiliation strings may contain errors or inconsistencies

## 2 Variable Codebook

### 2.1 Identifiers and Basic Information

#### id

**Type:** Integer

**Description:** Unique row identifier (1 to 9982)

**Missing:** 0

#### DOI

**Type:** Character

**Description:** Digital Object Identifier

**Missing:** 267 (2.7%)

#### year

**Type:** Integer

**Description:** Publication year

**Range:** 1972 – 2024

**Missing:** 0

Table 1: Summary statistics: year

N	Missing	Missing %	Mean	SD	Median	Min	Max
9982	0	0	2017.44	6.06	2019	1972	2024

#### title

**Type:** Character

**Description:** Journal title (source)

**Unique values:** 3230

**Missing:** 1

### 2.2 Impact and Quality Indicators

#### cited

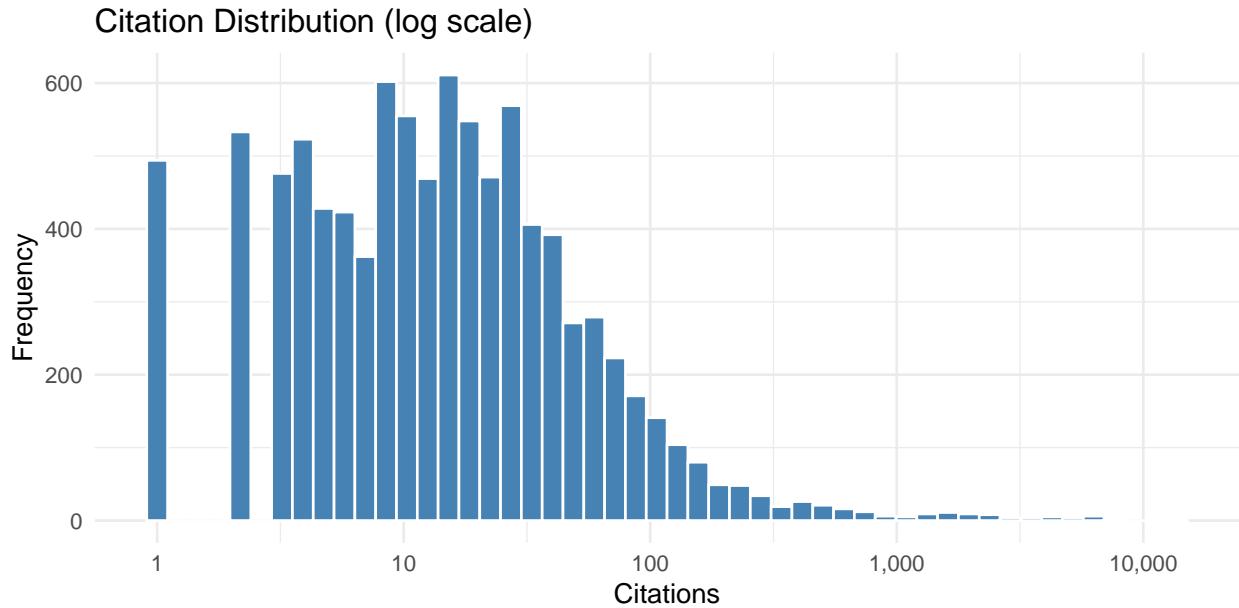
**Type:** Integer

**Description:** Total citation count as of October 2025

**Range:**  $0 - 1.3697 \times 10^4$

Table 2: Summary statistics: cited

N	Missing	Missing %	Mean	SD	Median	Min	Max
9982	0	0	49.8	323.72	12	0	13697



### quartile

**Type:** Factor (4 levels)

**Description:** Best SJR quartile ranking

**Levels:** Q1 (highest) to Q4 (lowest)

Table 3: Frequency distribution: quartile

Category	Frequency	Percentage
Q1	5241	52.5
Q2	2605	26.1
Q3	1138	11.4
Q4	328	3.3
NA	670	6.7

### sjr\_score

**Type:** Numeric

**Description:** SCImago Journal Rank score (continuous measure of journal prestige)

Table 4: Summary statistics: sjr\_score

N	Missing	Missing %	Mean	SD	Median	Min	Max
9312	670	6.7	1.25	1.93	0.78	0.1	22.61

## 2.3 Collaboration Variables

### coop

**Type:** Factor (2 levels)

**Description:** Collaboration type based on unique country count

**Levels:**

- bilateral: Exactly 2 countries (Japan + Vietnam)
- multilateral: More than 2 countries

Table 5: Frequency distribution: coop

Category	Frequency	Percentage
bilateral	5352	53.6
multilateral	4588	46.0
NA	42	0.4

#### n\_countries

**Type:** Integer

**Description:** Number of unique countries involved in collaboration

Table 6: Summary statistics: n\_countries

N	Missing	Missing %	Mean	SD	Median	Min	Max
9982	0	0	5.36	11.52	2	1	196

#### affiliations

**Type:** Character

**Description:** Semicolon-separated list of author affiliations (raw)

**Note:** Used for deriving country and authorship variables

## 2.4 Funding Variables

#### fund

**Type:** Factor (2 levels)

**Description:** Overall funding presence

**Levels:** Not funded, Funded

Table 7: Frequency distribution: fund

Category	Frequency	Percentage
Funded	5696	57.1
Not funded	4286	42.9
NA	0	0.0

#### Regional Funders

Binary indicators (0/1) for funding from each region:

- **asian:** Asian countries (excluding Japan)
- **eu:** European Union countries
- **int:** International organizations (e.g., World Bank, UN agencies)
- **jap:** Japan
- **us:** United States
- **vn:** Vietnam

Table 8: Funding by region

Region	Articles	Percentage
JAP	4,071	40.8
VN	1,470	14.7

Region	Articles	Percentage
ASIAN	1,111	11.1
US	909	9.1
EU	765	7.7
INT	248	2.5

## Funder Sectors

Binary indicators (0/1) for funding from each sector:

- **pub**: Public/government funding
- **uni**: University funding
- **ind**: Industry/private sector funding

Table 9: Funding by sector

Sector	Articles	Percentage
Public	5,252	52.6
University	1,273	12.8
Industry	352	3.5

## n\_regions

Type: Factor

Description: Number of distinct funder regions (0–6)

Table 10: Frequency distribution: n\_regions

Category	Frequency	Percentage
0	4286	42.9
1	4062	40.7
2	934	9.4
3	230	2.3
4	396	4.0
5	74	0.7
NA	0	0.0

## n\_sectors

Type: Factor

Description: Number of distinct funder sectors (0–3)

Table 11: Frequency distribution: n\_sectors

Category	Frequency	Percentage
0	4286	42.9
1	4612	46.2
2	987	9.9
3	97	1.0
NA	0	0.0

### **bilateral\_funding**

**Type:** Binary (0/1)

**Description:** Japan AND Vietnam co-funding (both jap=1 and vn=1)

Table 12: Bilateral co-funding (Japan + Vietnam)

Bilateral Funding	Articles	Percentage (%)
No	9,228	92.4
Yes	754	7.6

## 2.5 Authorship Variables

### First Author Location

Binary indicators (0/1) identifying first author affiliation:

- **fa\_vn:** First author from Vietnam
- **fa\_jp:** First author from Japan
- **fa\_o:** First author from other country

Table 13: First author location

Location	Articles	Percentage
Vietnam	3,138	31.4
Japan	4,544	45.5
Other	2,301	23.0

### Author Counts

- **n\_vn\_authors:** Count of Vietnamese author affiliations
- **n\_jp\_authors:** Count of Japanese author affiliations
- **n\_vn\_jp\_authors:** Total Vietnam + Japan authors (sum of above)

Table 14: Author count statistics

Variable	Mean	Median	Max
n_vn_authors	1.71	1	38
n_jp_authors	2.27	1	41
n_vn_jp_authors	3.98	3	66

### **prop\_vn\_authors**

**Type:** Numeric (0–1)

**Description:** Proportion of Vietnamese authors = n\_vn\_authors / n\_vn\_jp\_authors

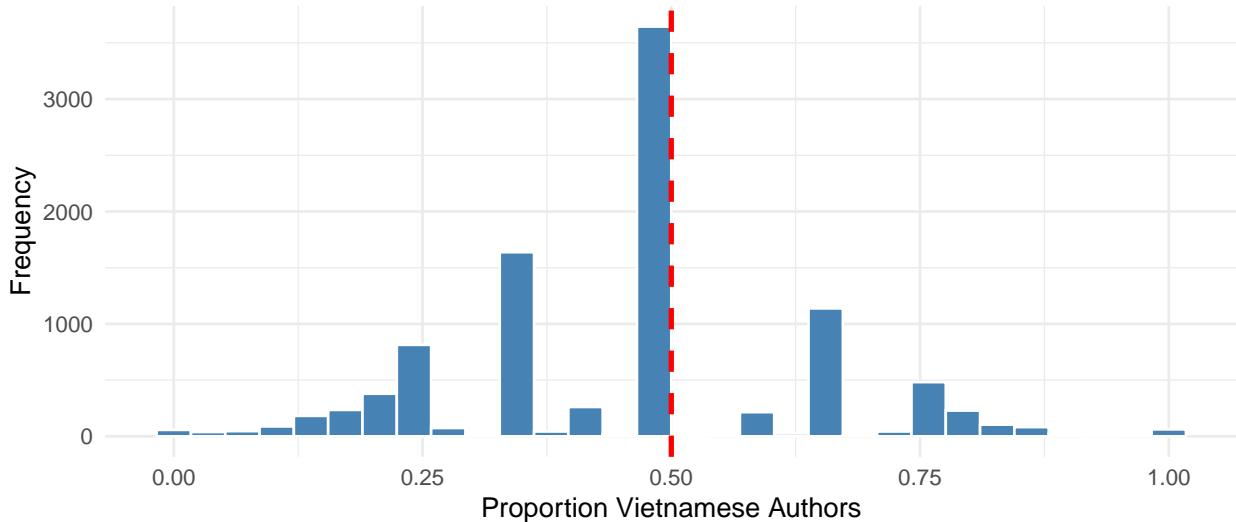
**Interpretation:** 0.5 = perfectly balanced, <0.5 = JP-dominated, >0.5 = VN-dominated

Table 15: Summary statistics: prop\_vn\_authors

N	Missing	Missing %	Mean	SD	Median	Min	Max
9971	11	0.1	0.46	0.19	0.5	0	1

## Distribution of Vietnamese Author Proportion

Red line = perfect balance (0.5)



## Leadership Flags

Binary indicators (0/1) based on first author location:

- **vn\_led**: Vietnamese-led article ( $fa\_vn = 1$ )
- **jp\_led**: Japanese-led article ( $fa\_jp = 1$ )

Table 16: Research leadership distribution

Leadership	Articles	Percentage
VN-led	3,138	31.4
JP-led	4,544	45.5
Other-led	2,301	23.0

## collab\_type

**Type:** Factor (4 levels)

**Description:** Collaboration intensity classification

**Levels:**

- VN-dominated:  $prop\_vn\_authors > 0.6$
- JP-dominated:  $prop\_vn\_authors < 0.4$
- Balanced:  $prop\_vn\_authors$  between 0.4–0.6
- Other: First author from third country

Table 17: Frequency distribution: collab\_type

Category	Frequency	Percentage
Balanced	4164	41.7
JP-dominated	3607	36.1
Other	11	0.1
VN-dominated	2200	22.0
NA	0	0.0

## 2.6 Open Access

### OA

Type: Factor (2 levels)

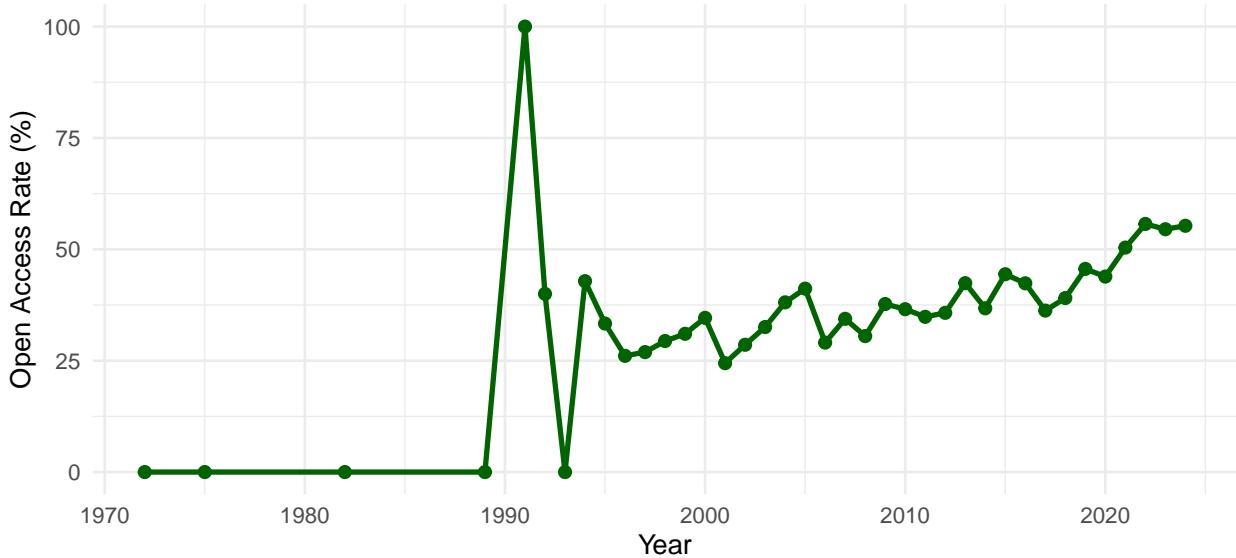
Description: Open access status

Levels: Not OA, OA

Table 18: Frequency distribution: OA

Category	Frequency	Percentage
Not OA	5486	55
OA	4496	45
NA	0	0

### Open Access Adoption Over Time



## 2.7 Disciplinary Classification

### Macro-Areas

Binary indicators (0/1) for broad disciplinary categories:

- **LS**: Life Sciences
- **SS**: Social Sciences
- **PS**: Physical Sciences
- **HS**: Health Sciences
- **mult**: Multidisciplinary

Note: Papers can belong to multiple macro-areas.

Table 19: Distribution across macro-areas

Macro_area	Articles	Percentage
Physical Sciences	5,580	55.9
Life Sciences	3,490	35.0
Health Sciences	2,336	23.4
Social Sciences	902	9.0

Macro_area	Articles	Percentage
Multidisciplinary	387	3.9

## Subject Areas

26 specific subject areas (binary 0/1 indicators):

Table 20: Top 15 subject areas

Subject	Articles	Percentage
Medicine	2,134	21.4
Agriculture & Biological Sciences	1,980	19.8
Engineering	1,723	17.3
Environmental Science	1,511	15.1
Physics & Astronomy	1,369	13.7
Biochemistry	1,321	13.2
Materials Science	1,197	12.0
Chemistry	933	9.3
Computer Science	918	9.2
Immunology & Microbiology	738	7.4
Earth & Planetary Sciences	718	7.2
Social Sciences	553	5.5
Mathematics	504	5.0
Chemical Engineering	486	4.9
Energy	396	4.0

## 2.8 Sustainable Development Goals (SDGs)

### SDG Variables

17 binary indicators (0/1) mapping articles to UN Sustainable Development Goals:

- **SDG\_1** through **SDG\_17**: Individual SDG flags
- **n\_sdg**: Total number of SDGs addressed per article

**Mapping method:** text2sdg package using SDGO system on article abstracts

Table 21: SDG coverage across all articles

SDG	Articles	Percentage
3. Good Health	4,513	45.2
11. Sustainable Cities	2,892	29.0
8. Decent Work	2,018	20.2
10. Reduced Inequalities	1,936	19.4
15. Life on Land	1,910	19.1
16. Peace & Justice	1,760	17.6
14. Life Below Water	1,533	15.4
7. Affordable Energy	1,389	13.9
9. Industry/Innovation	1,200	12.0
6. Clean Water	1,143	11.5
13. Climate Action	1,124	11.3
4. Quality Education	1,109	11.1
2. Zero Hunger	1,009	10.1
17. Partnerships	782	7.8
12. Responsible Consumption	676	6.8

SDG	Articles	Percentage
5. Gender Equality	636	6.4
1. No Poverty	37	0.4

### n\_sdg

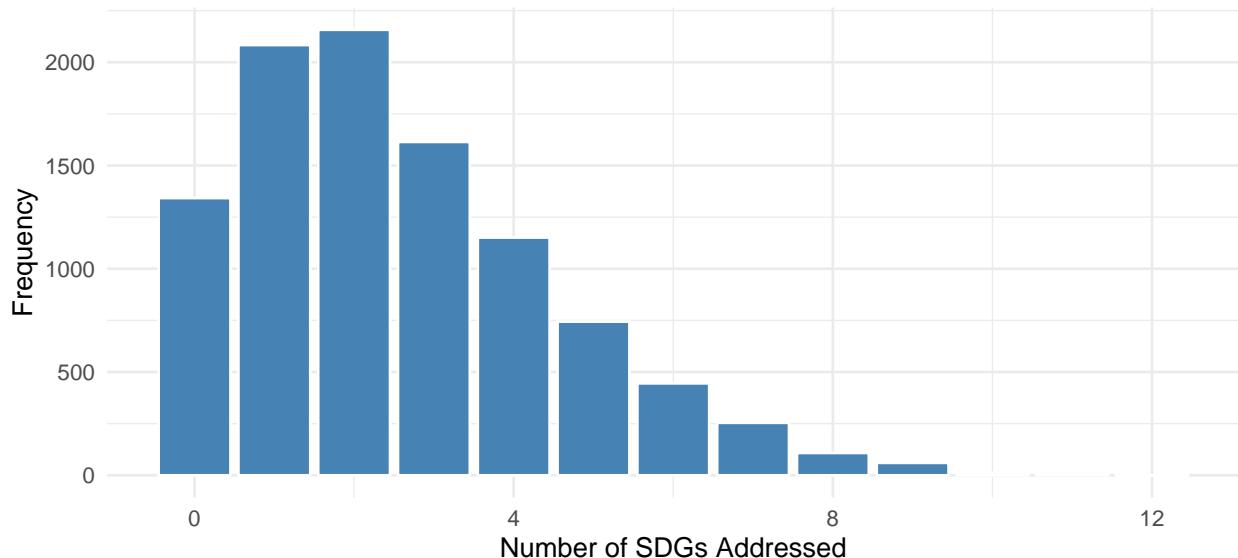
Type: Integer

Description: Total number of SDGs addressed per article (0–17)

Table 22: Summary statistics: n\_sdg

N	Missing	Missing %	Mean	SD	Median	Min	Max
9982	0	0	2.57	1.98	2	0	12

### Distribution of SDG Coverage per Article



### 3 Session Information

```
R version 4.5.2 (2025-10-31)
Platform: aarch64-apple-darwin20
Running under: macOS Tahoe 26.0.1

Matrix products: default
BLAS:    /System/Library/Frameworks/Accelerate.framework/Versions/A/Frameworks/vecLib.framework/Versions/
LAPACK:  /Library/Frameworks/R.framework/Versions/4.5-arm64/Resources/lib/libRlapack.dylib;  LAPACK versi

locale:
[1] C.UTF-8/C.UTF-8/C.UTF-8/C.C.UTF-8/C.UTF-8

time zone: Asia/Ho_Chi_Minh
tzcode source: internal

attached base packages:
[1] stats      graphics   grDevices utils      datasets   methods    base

other attached packages:
[1] scales_1.4.0    knitr_1.50       lubridate_1.9.4 forcats_1.0.0
[5] stringr_1.5.2   dplyr_1.1.4     purrr_1.1.0     readr_2.1.5
[9] tidyverse_2.0.0  tibble_3.3.0    ggplot2_4.0.0   tidyverse_2.0.0

loaded via a namespace (and not attached):
[1] bit_4.6.0        gtable_0.3.6     jsonlite_2.0.0  crayon_1.5.3
[5] compiler_4.5.2   tidyselect_1.2.1  parallel_4.5.2  yaml_2.3.10
[9] fastmap_1.2.0    R6_2.6.1        labeling_0.4.3  generics_0.1.4
[13] pillar_1.11.0   RColorBrewer_1.1-3 tzdb_0.5.0    rlang_1.1.6
[17] stringi_1.8.7   xfun_0.53       S7_0.2.0       bit64_4.6.0-1
[21] timechange_0.3.0 cli_3.6.5       withr_3.0.2    magrittr_2.0.3
[25] digest_0.6.37   grid_4.5.2       vroom_1.6.5    hms_1.1.3
[29] lifecycle_1.0.4 vctrs_0.6.5     evaluate_1.0.5 glue_1.8.0
[33] farver_2.1.2    rmarkdown_2.29   tools_4.5.2    pkgconfig_2.0.3
[37] htmltools_0.5.8.1
```

---

End of Databook

Table 23: Complete variable list

Variable	Type	Description
id	Integer	Unique row identifier
DOI	Character	Digital Object Identifier
year	Integer	Publication year (2000-2025)
title	Character	Journal title
cited	Integer	Citation count
quartile	Factor	SJR quartile (Q1-Q4)
sjr_score	Numeric	SJR score
OA	Factor	Open access status
coop	Factor	Bilateral/multilateral
n_countries	Integer	Number of countries
affiliations	Character	Raw affiliation string
fund	Factor	Funding status
asian	Binary	Asian funder
eu	Binary	EU funder
int	Binary	International funder
jap	Binary	Japan funder
us	Binary	US funder
vn	Binary	Vietnam funder
pub	Binary	Public sector funder
uni	Binary	University funder
ind	Binary	Industry funder
n_regions	Factor	Number of funder regions
n_sectors	Factor	Number of funder sectors
bilateral_funding	Binary	Japan + Vietnam co-funding
fa_vn	Binary	First author Vietnam
fa_jp	Binary	First author Japan
fa_o	Binary	First author other
vn_led	Binary	VN-led paper
jp_led	Binary	JP-led paper
n_vn_authors	Integer	Count VN authors
n_jp_authors	Integer	Count JP authors
n_vn_jp_authors	Integer	Total VN+JP authors
prop_vn_authors	Numeric	Proportion VN authors (0-1)
collab_type	Factor	Collaboration intensity (4 levels)
LS	Binary	Life Sciences
SS	Binary	Social Sciences
PS	Binary	Physical Sciences
HS	Binary	Health Sciences
mult	Binary	Multidisciplinary
[26 subjects]	Binary	Specific subject areas
SDG_1 to SDG_17	Binary	Individual SDG flags
n_sdg	Integer	Total SDGs addressed