Assignment: Building an Author-Level Dataset from Bibliographic Metadata

Objective

The goal of this assignment is to construct a **replicable, author-level dataset** starting from a bibliographic export (e.g. Scopus) in order to **study the representativeness of scholars who conduct research on financial literacy**.

In particular, the purpose of the project is to build a dataset that allows us to analyze **who produces research on financial literacy**, in terms of:

- gender composition,
- career stage (seniority),
- institutional background and prestige (university rankings).

For **each paper and each author**, you are required to identify and code:

- the **gender** of the author,
- the **seniority / career stage** of the author,
- the **ranking of the author's university affiliation**.

The final dataset will serve as the basis for a descriptive analysis of **representation and concentration in the academic literature on financial literacy**.

Starting Dataset

You will start from a dataset similar to a Scopus export, where:

- each row corresponds to a **paper**,
- multiple authors and affiliations are stored in the same cell,
- the paper identifier is the variable `EID`.

Example: Original Dataset (Paper Level)

EID	Title	Authors	Affiliations	Year
2-s2.0-AAA	FinTech and Growth	Smith, J.; Rossi, M.	University of Oxford; University of Milan	2022
2-s2.0-BBB	Financial Literacy	Garcia, L.; Chen, W.; Müller, T.	University of Madrid; Tsinghua University; University of Bonn	2023

Important: This format is *not* suitable for author-level analysis, because multiple authors are stored in a single row.

Target Dataset

You must transform the data into an **author × paper dataset**.

- Each row must correspond to **one author in one paper**.
- The identifier of the paper must be `EID`.

Example: Final Dataset (Author Level)

Note: In the final dataset you must include *both* ranking variables: `Ranking QS` and `Ranking Repec`. If a ranking is not available for a given affiliation in one of the sources, leave it missing (e.g. `NA`) and document it.

EID	Author name	Author pos.	Gender	Seniority	Affiliation	Country	Ranking QS	Ranking Repec
2-s2.0-AAA	Smith, J.	1	Male	Associate Prof.	Univ. of Oxford	UK	3	12
2-s2.0-AAA	Rossi, M.	2	Male	Assistant Prof.	Univ. of Milan	IT	316	145
2-s2.0-BBB	Garcia, L.	1	Female	Full Prof.	Univ. of Madrid	ES	171	98
2-s2.0-BBB	Chen, W.	2	Male	PhD Student	Tsinghua Univ.	CN	25	60
2-s2.0-BBB	Müller, T.	3	Male	Postdoc	Univ. of Bonn	DE	227	110

Variables to Construct

Gender

You are allowed to use **automatic tools or software** (e.g. name–gender classifiers, APIs, or packages) to obtain a first guess of the author's gender.

However:

- all results must be **manually checked and validated**,
- ambiguous or conflicting cases must be **verified individually** using webpages or CVs,
- if gender cannot be confidently identified, it must be coded as `Unknown` and documented.

Seniority / Career Stage

You must classify each author into a career stage (e.g. PhD student, Postdoc, Assistant Professor, Associate Professor, Full Professor).

- You must define **explicit classification rules**.
- The rules must be applied **consistently**.
- Ambiguous cases must be documented.

University Ranking

For each author affiliation, you must collect rankings from **both** sources (when available):

- **QS World University Rankings** → store in `Ranking QS`
- **RePEc** (for economics/business institutions) → store in `Ranking Repec`

You must store the (numeric) ranking position in each variable. If a ranking is not available in one source, leave it missing (e.g. `NA`) and document it.

On Software and Hand-Coding

You are allowed to use any software you want (R, Python, Stata, Excel, etc.) to:

- clean and reshape the dataset,
- parse authors and affiliations,
- merge external sources (e.g. rankings),
- check consistency and produce the final files.

However, you should be aware that a **substantial part of the work will necessarily require manual hand-coding**, in particular:

- verifying gender assignments,
- identifying seniority / career stage,
- verifying affiliations and current positions.

This is **expected and normal**. The goal of the project is not only automation, but also careful and transparent data construction.

Replicability Requirement (Crucial)

You may work:

- manually,
- using software (R, Python, Stata, Excel, etc.).

In all cases, you must:

- provide all intermediate files,
- provide all scripts or detailed procedures,
- ensure that the final dataset can be reproduced from scratch.

Weekly Progress Reports

Each week you must prepare a short progress report answering:

- What did you do during the past week?
- What problems did you encounter?
- How did you solve them?
- What will you do next week?

Submission procedure:

- Each weekly report must be saved in the appropriate shared **OneDrive folder**.
- The same report must also be **sent to me by email**.

These reports are part of the evaluation.

Final Step (If Time Allows)

Once the dataset is complete, you may be asked to produce basic descriptive statistics aimed at documenting the **representativeness of the field**, such as:

- gender composition of authors,
- seniority distribution,
- concentration in top-ranked institutions (QS and/or RePEc),
- differences by author position.

Final Deliverables

You must submit:

- the final author-level dataset,
- all scripts and intermediate files,
- a short documentation file describing sources and rules,
- the full set of weekly progress reports.

Rules

- Do not invent data.
- Flag and document all uncertain cases.
- Transparency and replicability matter more than speed.