# STEP 1 – Cleaning the Data

One a country has sent the **data** **file** (.xlsx or .csv), update the [check-list](https://undp.sharepoint.com/:x:/s/NatureClimate/EYTAsAwnhJ5JidEcRoB4j7wBRfTeJJvDLd0vEYEC3dopxA?e=H0tUDh&CID=6682b80a-af86-3013-0f7b-5a95c13d9b77) and download it, locally

Locally, open the **Git project** in the *TargetAssessmentReport* directory

Place **data** **file** in the */data/countries/{country\_name}/original* directory

… openit with Excel and quickly check (for):

* Merged and NA cells, swapped columns, column titles that are not, …
* Language
* The relevant tab/sheet

Return to *TargetAssessmentReport* and open “**data\_cleaning.R**”

* Specify:
  + Country name (cty)
  + If there are merged cells (mrgd)
  + Language (lang)
  + Relevant tab/sheet (sheet)
* Make specific changes to the data in the “Country changes” sections
* Checks:
  + single country name
  + 3 types of targets
  + uniquely identified target names
  + odd characters

The end result will be saved as “**data\_{country\_name}\_{date}.xlsx**” into */data/countries/{country\_name}*

Open the file and manually check for acronyms

* if any exist, check the source material
* make a copy of the **terms file** in the */Support\_Input\_Data* directory, updating the date in its name
* if the acronyms seem country-specific, create a new tab with the country name and add them there; if not, just add them to the “Terms” tab/sheet

Open Azure: on the left-side pane, select “Data”, then click the “Linked” section in the Data pane, drop *undp\_ngd-syn-bio-diversity-dev (Primary – undpngddlsdev01)* down and select *sandbox-bpps-bio-diversity (Primary)*

Upload “**data\_{country\_name}\_{date}.xlsx**” into the */Similarity\_Assessment\_Data* directory and – if any acronyms were detected – add the **terms file** to the */Support\_Input\_Data* directory

# STEP 2 – Running the Themes Analysis

Make sure that

* the **themes and terms files** in the */Support\_Input\_Data* directory are the most up-to-date, like in *TargetAssessmentReport/data*
* the **glossaries files** in the */Support\_Input\_Data* directory are the most up-to-date, like in *TargetAssessmentReport/data/glossaries*

Back on the left-side pane of Azure, select “Develop”, then in the Develop pane head to “Notebooks” and drop *Nature-Climate (2025)* down and open **JP\_Nature\_Climate\_themes\_UNDP\_GPT4o-min\_v5\_14Mar25**