# STEP 1 – Cleaning the Data

Once 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, open the **Git project** in the *TargetAssessmentReport* directory and 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 it and manually check for acronyms/term

* If any exist, check the source material
* Make a copy of the most up-to-date **terms file** in the */Support\_Input\_Data* directory, updating the date in its name
* If any new acronyms/terms detected seem *country-specific* or are in a language *other than English*, 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**

Change:

* Language (lang), in **DATA Import** cell 1
  + … if needed, add another language option
* Country name (dta): “**data\_{country\_name}\_{date}.xlsx**”, in **DATA Import** cell 2

What this does:

* *Non-English* terms/acronyms in the targets are replaced in **DATA tweaks** cell 2;
* *Non-English* targets are translated (with a glossary and GPT) in **DATA tweaks** cells 4-5;
* *Country-specific* English terms/acronyms are added to a wider list of terms, and are all replaced in the targets in **DATA tweaks** cell 9-10;
* The country name is removed from the targets in **DATA tweaks** cell 11;
* The analysis itself is run in **MODEL *Running***cell 1.

The end result will be saved as “**UNDP\_themes\_AI\_Assessment\_GOT4o-mini\_{Country}\_{Date}.csv**” into */2025\_Themes\_Results*

# STEP 3 – Running the Target-to-Target Analysis

# STEP 4 – Running the Summary

# STEP 5 – Cleaning the Quantitative Analysis

# STEP 6 – Building the Report