

Diagnostic review on the use of methodological approaches and evaluation methods in UN evaluations

– Progress Presentation

Prof. Dr. Steffen Eckhard

Daniel Baumann

Sebastian Korb




Dec 30th, 2024

Objectives of the Review

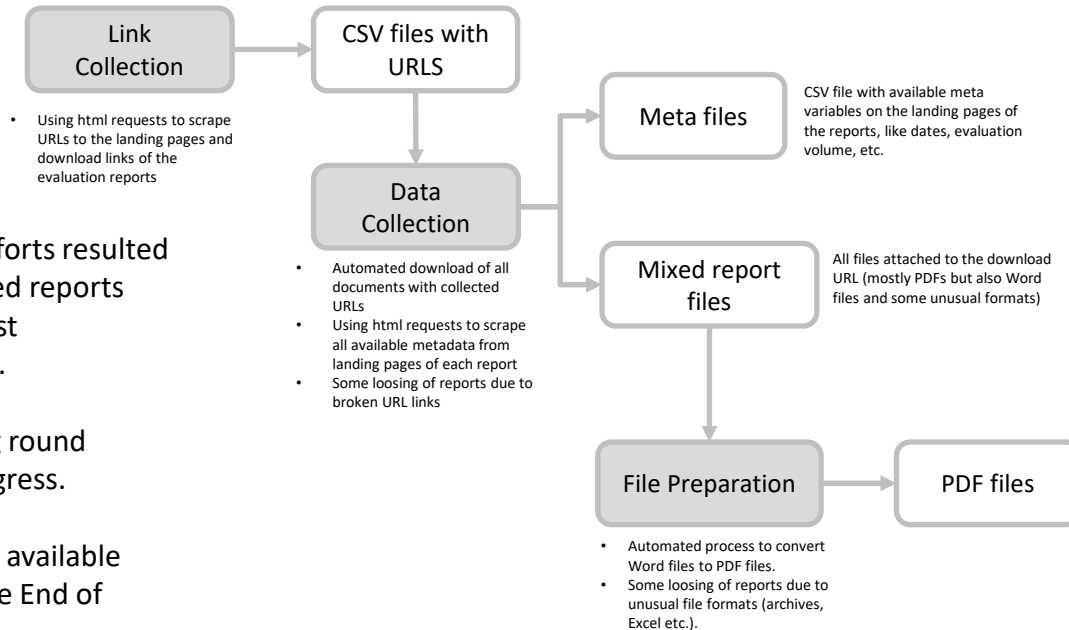
- **Mapping Methodological Diversity:** identifying and classifying the use of advanced evaluation methods
- **Trend Analysis:** examining temporal trends of method usage
- **Agency Comparisons:** identifying differences in methodological practices across UNEG member agencies
- **Insight Development:** illustrating impacts of methodological choices
- **Support Decision-Making:** provide data-driven insights on effective and innovative evaluation practices

PROGRESS PRESENTATION




Pipeline Phase 1: Data Collection and File Preparation

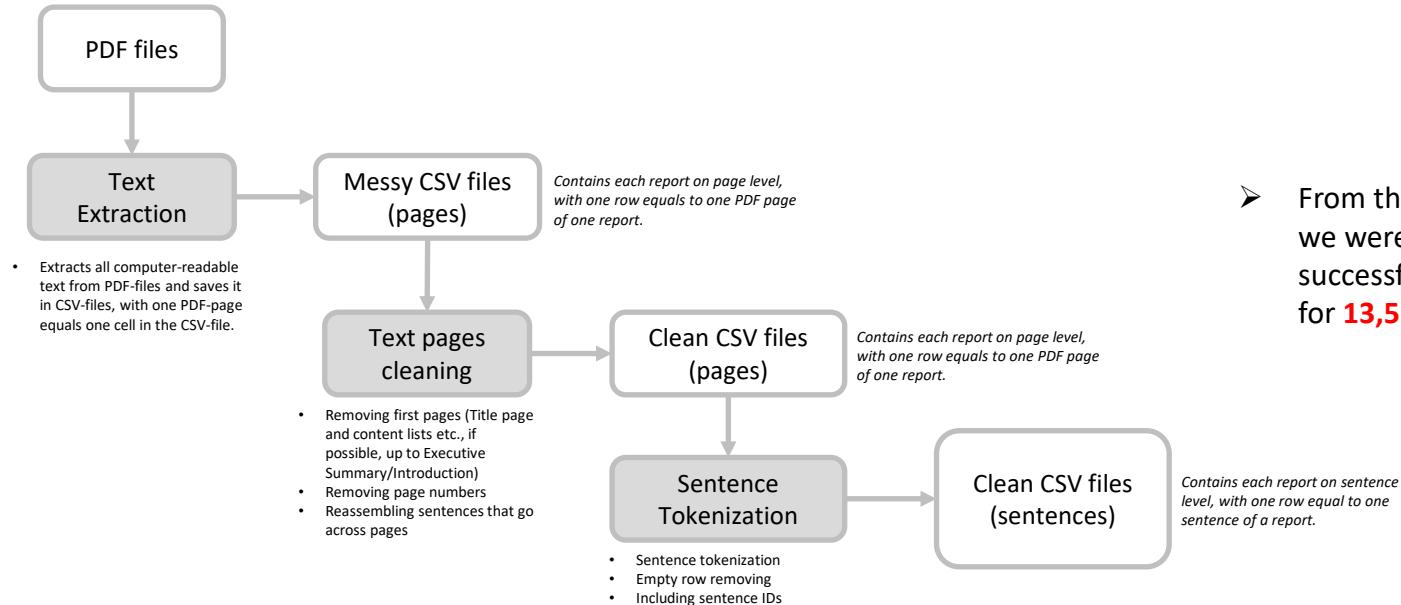
Meaning of Box Colors	
	In-/Output files
	Pipeline processes
	Not finished

- First scraping efforts resulted in **13,635** scraped reports from the 21 most represented IOs.
- Second scraping round currently in progress.
- Goal: include all available reports up to the End of 2024.



Pipeline Phase 2: Text Extraction and Cleaning

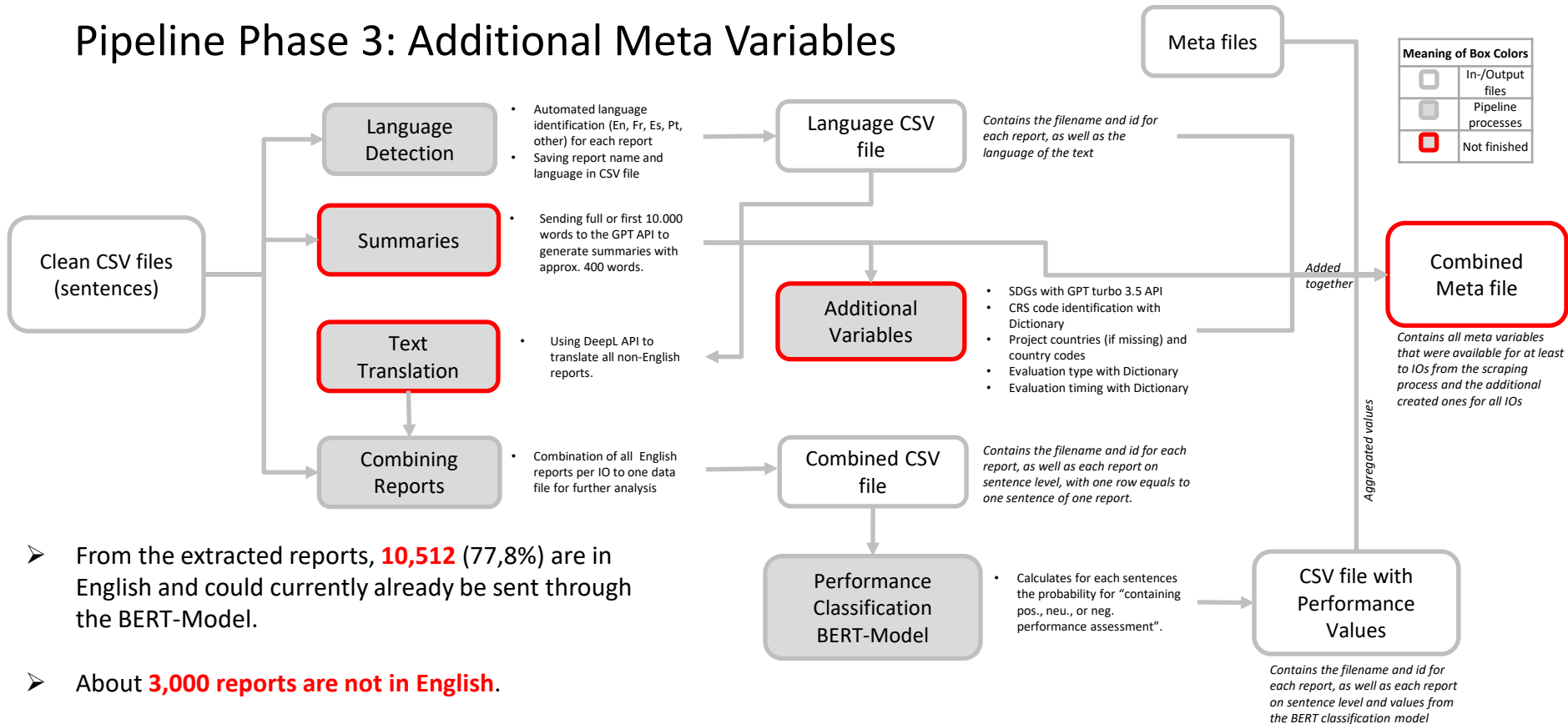
Meaning of Box Colors	
	In-/Output files
	Pipeline processes
	Not finished



➤ From the scraped reports, we were currently able to successfully extract the text for **13,518** (99.1%).

PROGRESS PRESENTATION

Pipeline Phase 3: Additional Meta Variables



➤ From the extracted reports, **10,512** (77,8%) are in English and could currently already be sent through the BERT-Model.

➤ About **3,000 reports are not in English.**

Evaluation Dataset – 1st Overview

- After initial scraping and preprocessing phases, we have a dataset with the following variables for the English reports from **21 IOs**

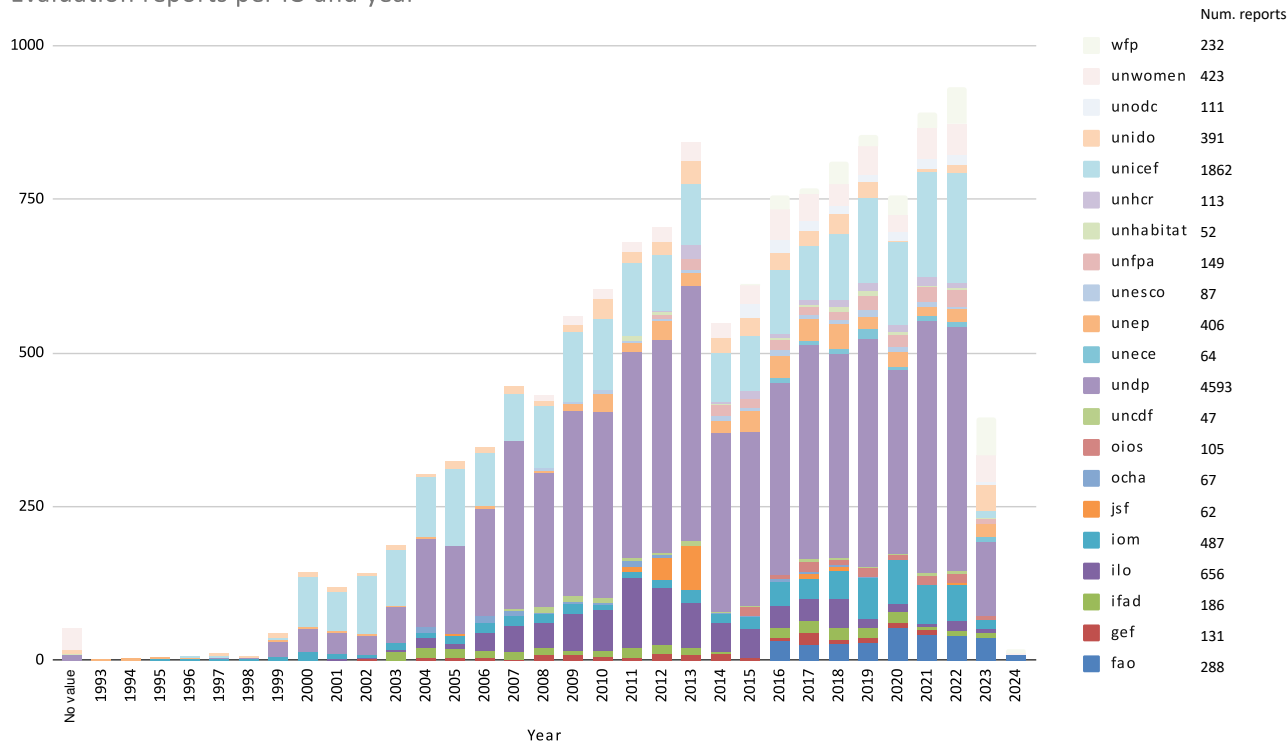
Variable	N
Report title	10,434
Summary	<i>Not done by now</i>
Year	10,391
IO	10,434
Language	10,434
Report Length	10,434
Country	9,445
Country Code (ISO3)	8,800
Evaluation Type	<i>Not done by now</i>
Evaluation Timing	<i>Not done by now</i>
Sector (CRS)	<i>Not done by now</i>
SDG	<i>Not done by now</i>
Positivity Share	10,422

PROGRESS PRESENTATION

Evaluation Dataset – 1st Overview

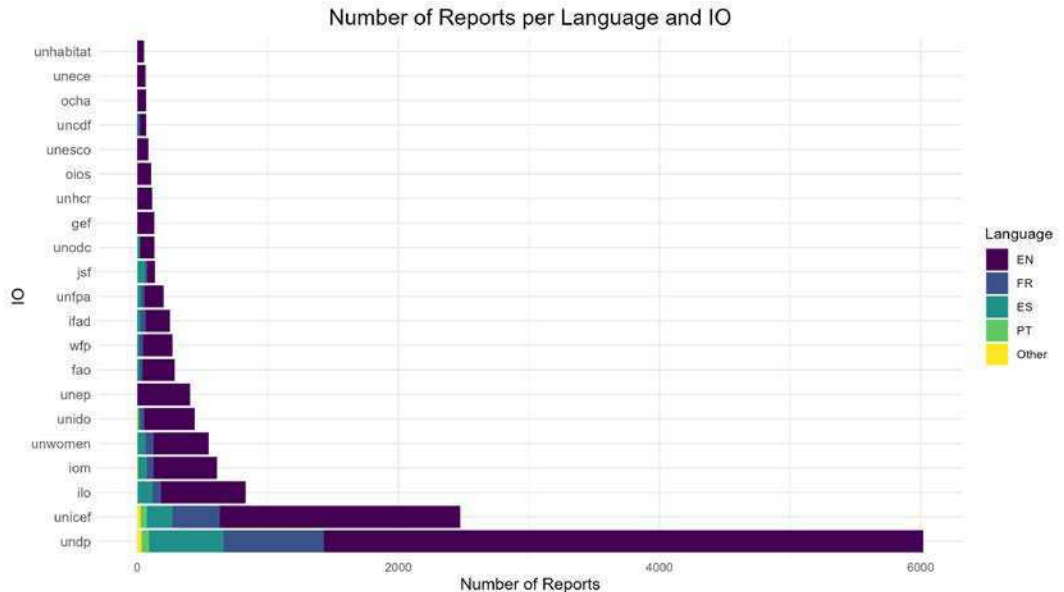
- Big differences in the number of reports between IOs.
- All IOs are included with at least 50 downloadable evaluation reports.
- Second scraping round will fill the dataset up to the end of 2024.

Evaluation reports per IO and year

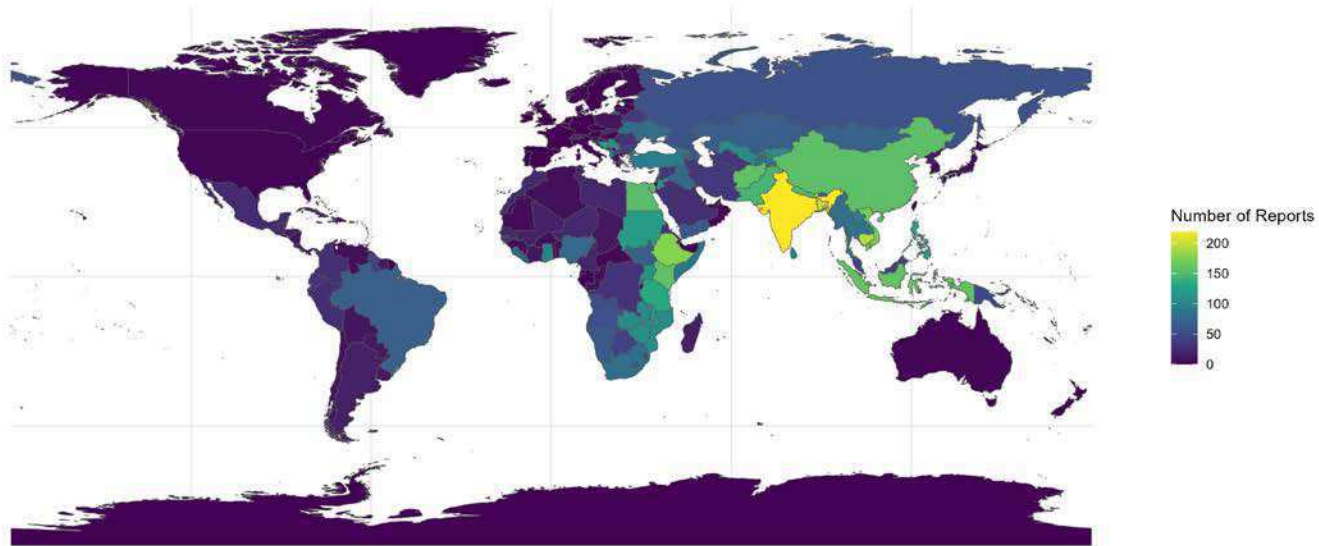


Evaluation Dataset – 1st Overview

- Even though most reports are already in English, about **3,000** reports are in other languages (mostly Spanish, French or Portuguese)
- DeepL costs for translation:
 - On average 225,000 characters per report
 - 20 € per 1 mio. characters
 - 675 mio. characters = € 13,500 or **US\$ 14,000**



PROGRESS PRESENTATION

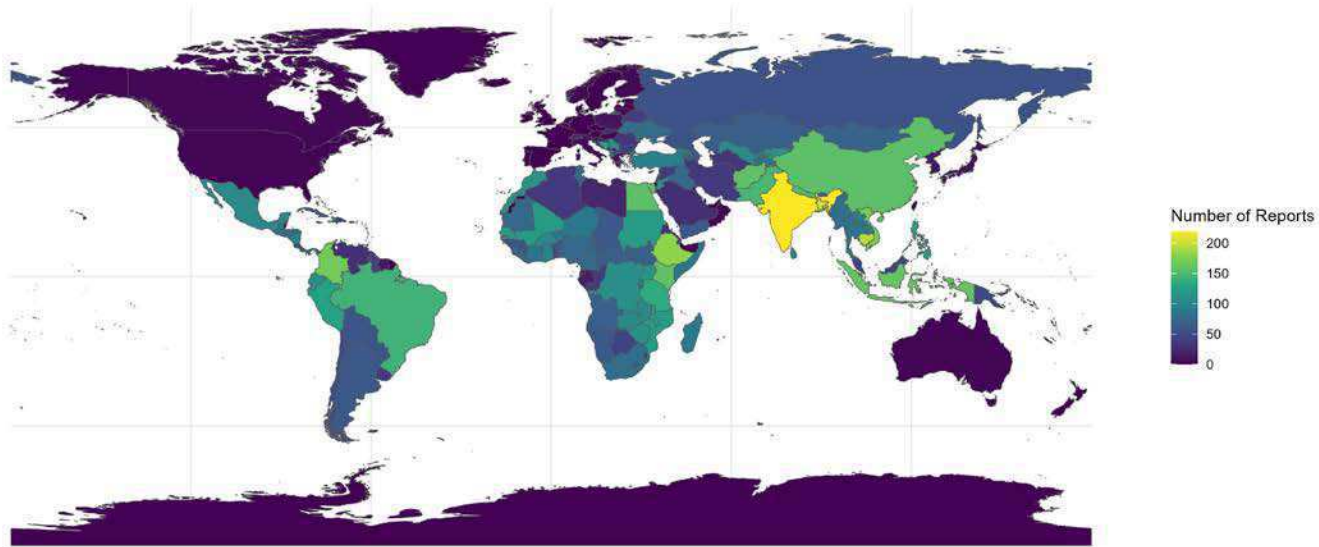
Report Distribution by Country **before Translation**

Excluded projects with multiple or no countries of implementation: 1365

- Western Africa and Latin America are **severely underrepresented**

PROGRESS PRESENTATION

Report Distribution by Country **after Translation**



Excluded projects with multiple or no countries of implementation: 1423

- Including all available reports would **reduce the bias** over countries significantly.

Evaluation Dataset – Findings: Performance of evaluated activities

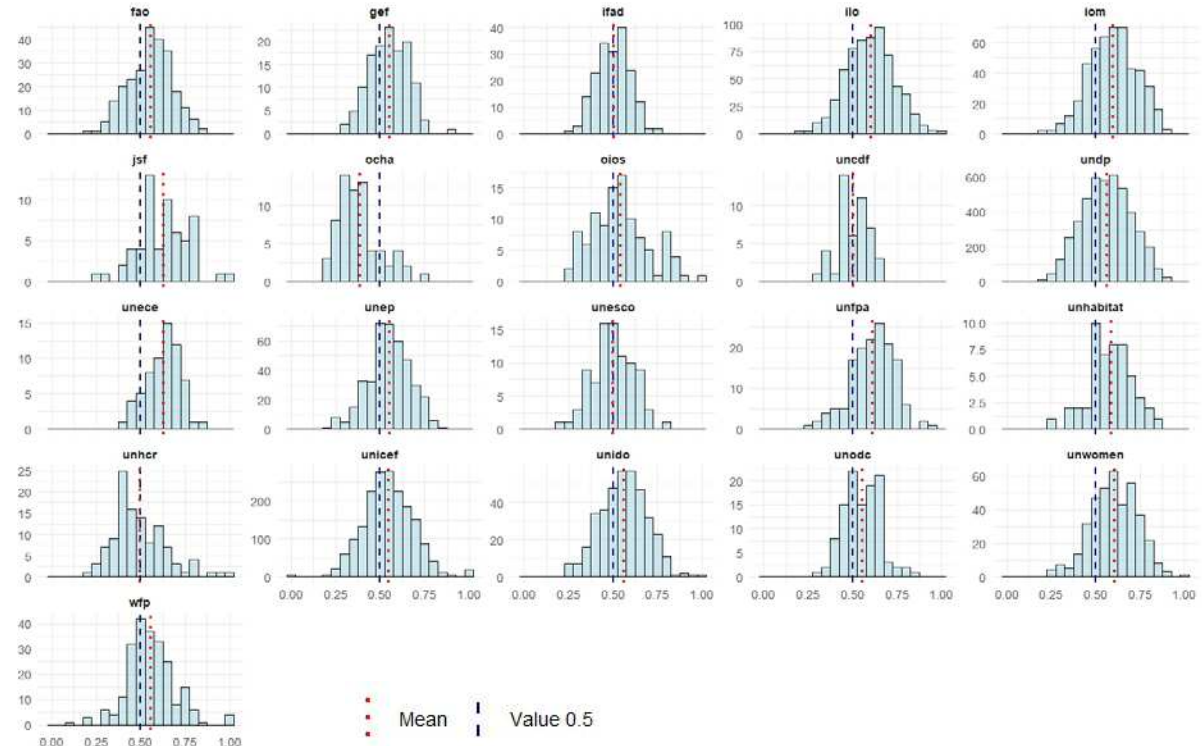
- Measuring the performance of evaluated activities per report
- Measure based on publication by Eckhard et al (2023)
 - Share of positive to negative “assessment sentences” per report (**positivity share**)
 - Validation shows very high correlation with human performance annotation



Eckhard, S., Jankauskas, V., Leuschner, E., Burton, I., Kerl, T., & Sevastjanova, R. (2023). The performance of international organizations: a new measure and dataset based on computational text analysis of evaluation reports. *The Review of International Organizations*, 18(4), 753-776. <https://link.springer.com/article/10.1007/s11558-023-09489-1>

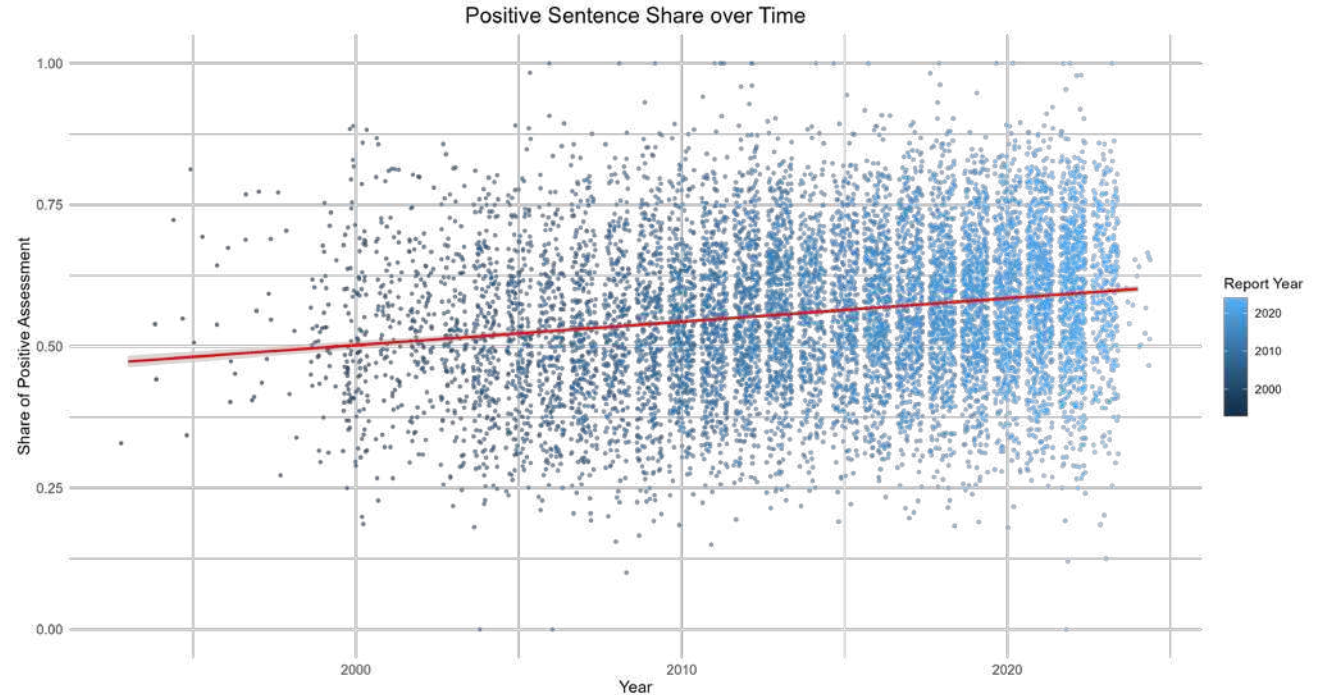
Evaluation Dataset – Findings: Performance Distribution Overall

- Positivity share is overall **normally distributed** with a slight positive bias.



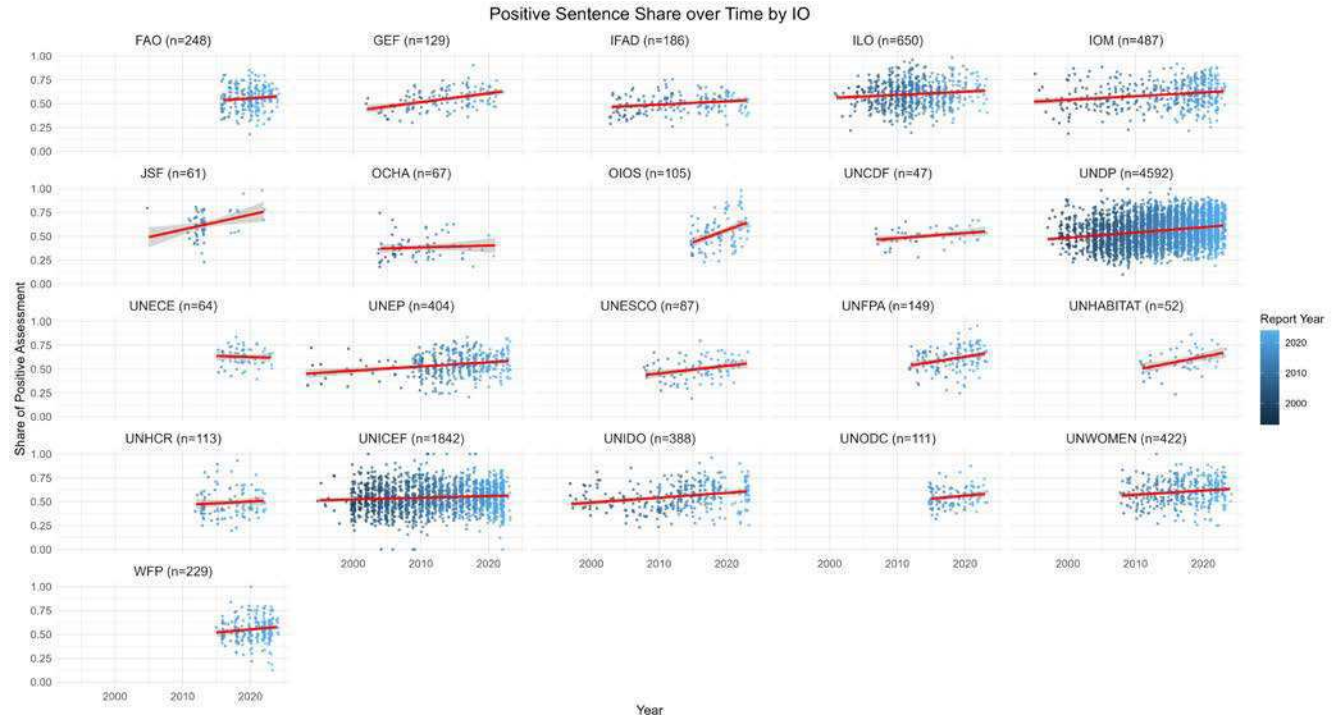
Evaluation Dataset – Findings: Performance Distribution Overall

- We see a general **increase in project performance** assessment in the evaluation reports over time.



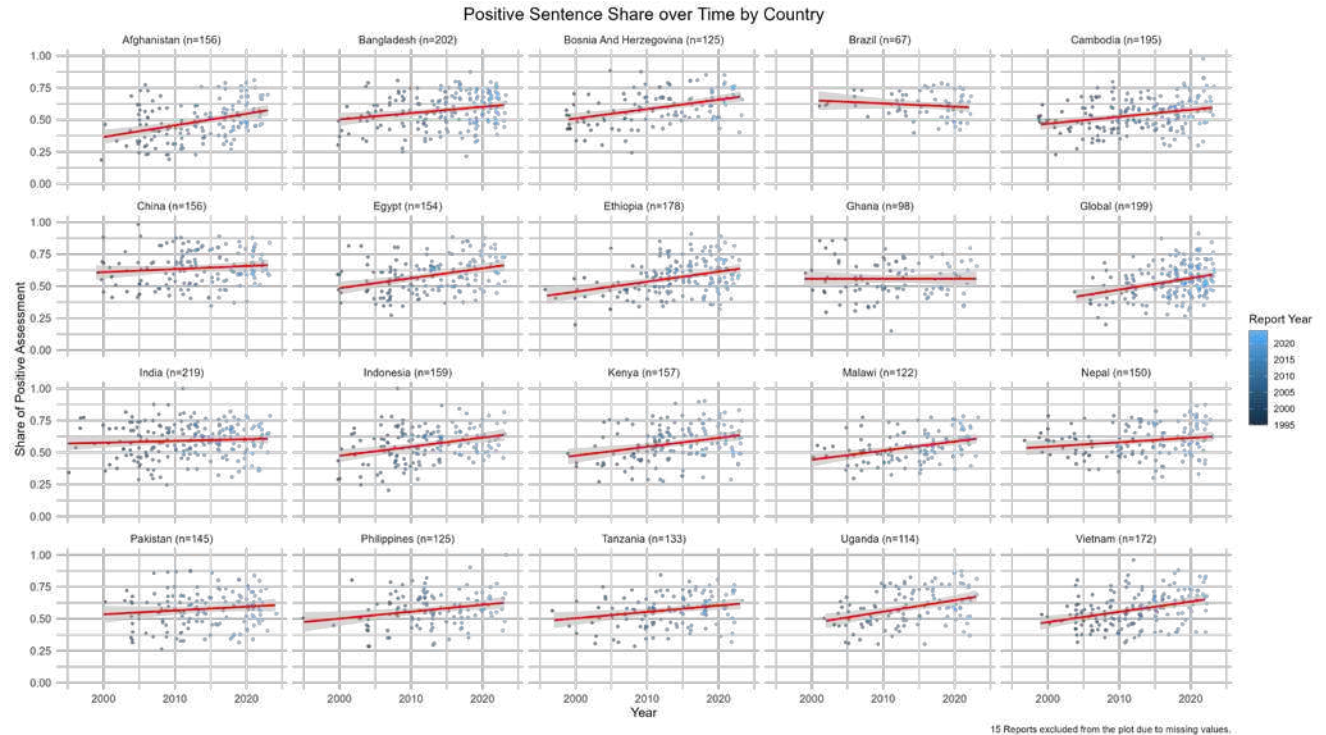
Evaluation Dataset – Findings: Performance Distribution Across IOs

- This general trend is also visible when inspecting the different IOs.
- IOs with smaller case numbers show more extreme trends, but this is to be interpreted with caution.



Evaluation Dataset – Findings: Performance Distribution by Top 20 Countries

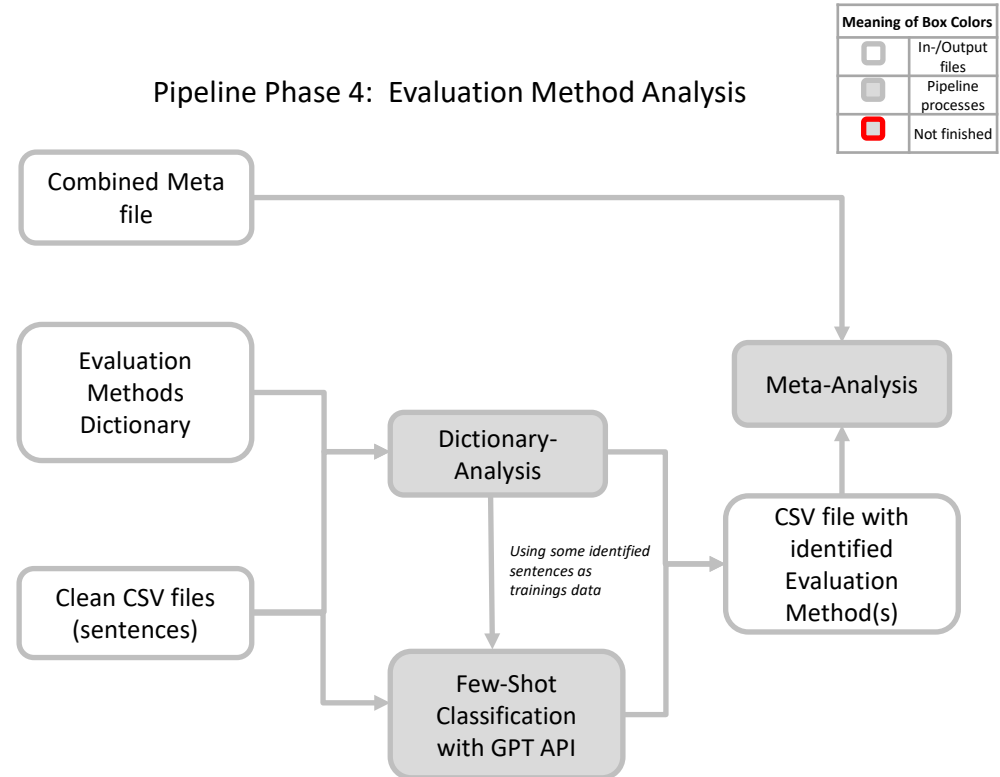
- This general trend is also visible when inspecting the different countries of implementation.



Next Steps

- Complete report scraping
- Creation of report summaries and additional variables
- Different **options to continue** with identification of evaluation methods in the reports:
 - Dictionary based classification:
 - Using predefined set of keywords to detect specific methods in the reports
 - Few-shot model with GPT API:
 - Using some identified sentences of the dictionary analysis to train the model
 - Model then able to also identify less explicitly mentioned methods
 - A combination of the two

Pipeline Phase 4: Evaluation Method Analysis



PROGRESS PRESENTATION

Timetable

Steps	Task	Explanation	Days required	Dec	Jan	Feb	Mar	Apr	May	June	July
1	Overview	Complete overview of the list of UNEG evaluation reports: how many reports per IO are included, which IOs are included, do the URLs to the landing pages and to the reports work, etc. We estimate that only about 16,000 reports are available (of 26,000 listed in UNEG database)	2								
2	Scraping reports	Write a script to retrieve the reports for each IO. As we suspect that some of the links will not work, we need to double-check the overview pages of each IO's evaluation unit page. For IOs that have less than 100 reports, we download manually.	3								
3	Scraping meta data	In addition to the reports and the metadata already included in the UNEG list, scrape as much metadata as possible for each report from the respective landing pages of the IOs.	2								
4	Converting File Formats	We will not receive PDF files for all reports. We will write a script that converts all non-PDF files to PDF files to facilitate text extraction. Reports in very unusual or non-text formats or with corrupted files will be excluded during this process.	1								
5	Text extraction	Extracting the text from the PDF files. Since we will get a large variety of different formatting of the texts in the files, it is very unlikely that we will be able to preserve much of the structure (such as chapters, paragraphs, etc.) of the texts.	5								
6	Text cleaning	Clean up the extracted texts. If possible, we exclude the title pages, tables of contents, abbreviation tables, etc., as these contain little relevant information for the text analysis. We also split the texts at sentence level for later text classification applications.	9								
7	Performance assessment classification	Classify all sentences of the remaining reports with our BERT Large Language Model for the classification of performance assessment in evaluation reports (positive, neutral, negative).	5								
8	Creating final text corpus	After the performance classification, we create the final text corpus from all remaining reports at sentence level with the classification values for each sentence.	2								
9	Merging scraped meta data	From the metadata lists collected in step 3 from the various IO web pages, we create a large metadata list for all reports that remain in our corpus. We will keep as many variables as possible, even if they are not available for all IOs.	2								
10	Cleaning meta data	Check and if necessary, adjust important variables such as the reporting year or the country of the project/program for uniform spellings (add ISO-3 country codes), to avoid problems in later analyses.	4								
11	Creating additional meta variables from the texts	Extracting additional variables from the content of the reports. This could be a summary of each report, which can then be used for further tasks, the type of assessment (for some reports included in the UNEG metadata), the assessment timing (mid-term or ex-post) and a sectoral classification.	9								
12	Developing measures for methods extraction	Dictionary-based measures: Development and validation of dictionaries for methods mentioned in the Compendium of evaluation methods. Analysis and creation of data report and visualizations. Alternative classification approach with a few-shot machine learning based measure (GPT API)	20								
13	Presentation and Report	PowerPoint presentation with graphs and writeup of the report	6								
14	Sharing Code	Compiling code and ensuring transparent code documentation	2								
Total			72								