Methodology for measuring and estimating funding to data and statistics

Yu Tian, PARIS21 Archita Misra, PARIS21

Contents

Acknowledgements

Draft methodological note prepared by Yu Tian and Archita Misra (PARIS21) under the supervision of Rajiv Ranjan (PARIS21). The authors are grateful to Eric Swanson and Lorenz Noe (Open Data Watch) and Simon Lange (OECD) for their valuable review and feedback. Further comments to the authors (Yu.Tian@oecd.org; Archita.misra@oecd.org) are welcome.

1 Background

PARIS21 produces the Partner Report on Support to Statistics (PRESS) annually to report on trends in support to statistics. The methodology is applied retrospectively for all previous years to ensure comparability over time. This document presents the methodology.

2 Monitoring funding to statistics with accuracy

This section provides information on how to monitor support to statistics and how the data of the PRESS, which is also used for reporting SDG indicator 17.19.1 ("Dollar Value of all resources made available to strengthen statistical capacity in developing countries"), is generated.

Prior to PRESS 2018, the PRESS only focused on borrowing countries of the International Development Association.[1] Since 2018, the PRESS covers the commitments received by all countries throughout the report to align the findings with the SDG indicator 17.19.1: "Dollar value of all resources made available to strengthen statistical capacity in developing countries". Commitments were used as the main measurement instead of disbursements from the beginning of PRESS in 2006, when commitment data entries were made available more consistently. See section 4.2 for a discussion on using commitments vs disbursements in detail.

The PRESS aims to provide a full picture of international support to statistics. To achieve this goal, it mainly takes advantage of two data sources:

2.1 OECD's Creditor Reporting System (CRS)

The Organisation for Economic Co-operation and Development (OECD)'s Creditor Reporting System (CRS) records data from OECD Development Assistance Committee (DAC) members (donors) and some non-DAC donors. This provides a comprehensive account of Official Development Assistance (ODA). Donors report to the CRS using specific codes for the sectors targeted by their aid activity. Statistical Capacity Building (SCB) is designated by the sector code 16062.[2] Each activity reported in CRS can only be assigned with one of the over 100 purpose codes.[3] While CRS is one of the most reliable and comprehensive databases

[1] Eligibility for IDA support depends first and foremost on a country's relative poverty, which is defined as GNI per capita below an established threshold and updated annually ($1,185 in the fiscal year 2021). IDA also supports some countries, including several small island economies, that are above the operational cut-off but lack the required creditworthiness to borrow from the International Bank for Reconstruction and Development (IBRD). For more information, see: https://ida.worldbank.org/en/about/borrowing-countries

[2] Until 2019, this purpose was vaguely defined as "Both in national statistical offices and any other government ministries". However, after a successful campaign to improve the description, this purpose is now defined as "All statistical activities, such as data collection, processing, dissemination and analysis; support to development and management of official statistics including demographic, social, economic, environmental, and multi-sectoral statistics; statistical quality frameworks; development of human and technological resources for statistics, investments in data innovation."

[3] In recent years, CRS reporters can also assign multiple voluntary purpose codes to the same project. Code 16062 is not a voluntary code. See the CRS code list for more information: https://www.oecd.org/dac/financing-sustainable-development/development-finance-standards/dacandcrscodelists.htm

that accounts for aid flows, there are some concerns that need to be addressed while compiling support for data and statistics, such as: the cross-cutting nature of projects with statistical components, limited reporter knowledge about the code, the assignment of some ODA for data and statistics under other codes, lack of granularity in reporting, etc. These issues are covered in detail in PARIS (2019). PARIS21 is seeking to reduce this usually downward bias using a text analysis methodology (see section 3).

The CRS identifies a project donor by looking at the source of the funding. Countries are identified as donors if the flow is directly between them and the recipient country (type 1 in Fig. 1), or if the flow is earmarked for a certain project and channeled through multilateral organisations (type 3 in Fig. 1). If a project is funded by un-earmarked core contributions to multilateral organisations, the donors are marked as the multilateral organisations (types 2 and 4 in Fig. 1).

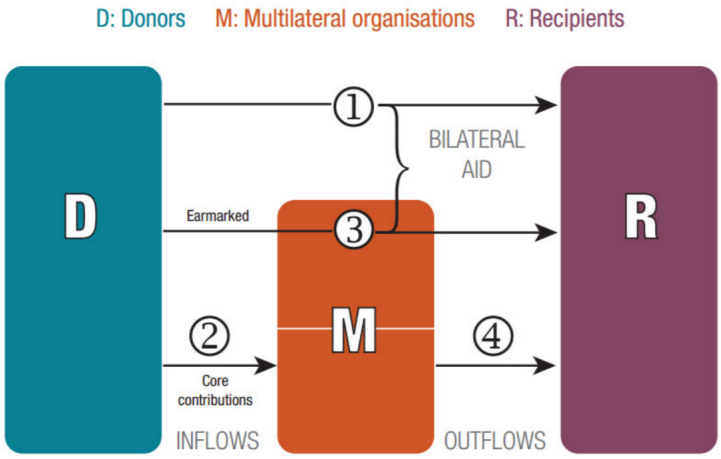

Figure 1: Flow of official aid in CRS.

2.2 PARIS21's annual online survey

The PARIS21 Secretariat supplements the data from the CRS with an annual online survey that is completed by a global network of respondents, mostly non-DAC donors. The survey covers a subset of the variables collected in the CRS, as well as some additional variables specific to data and statistics. Responding to the online survey is voluntary and offers an opportunity for respondents to share information about their statistical activities. Respondents include non-DAC members, including non- DAC donor countries, multilateral organisations, regional statistical training institutes, and other philanthropic organisations. The percentage of these projects in the final PRESS database has decreased in recent years, as many multilateral organisations have improved the granularity of their reporting to the CRS, making these data equally useful as data collected from the PRESS survey. To reduce the burden on donors, these multilateral organisations are no longer required to fill in the PRESS survey.

3 Identifying data and statistical projects using machine learning

The developed method to identify data and statistical projects is based on a two-step procedure that analyzes project titles in the first step by detecting pertinent keyword (**A**) and evaluates project's detailed descriptions using a machine learning approach (**B**).

3.1 Reading the CRS data

After downloading all .txt files for the years 2006 - 2020 from the official OECD data base, the fully merged data set is stored.

3.2 Preparing the data

Here, the process of preparing the data is outlined (see Fig. 2 for a comprehensive overview).

1. **Reducing the full CRS data set**

   A known characteristic of Canadian reporting in the CRS data base is that both project titles and long descriptions[4] are reported in both official languages in the format "Englisch/French". To avoid misclassification and misidentification due to the presence of both languages, the French part was dropped. Additionally, the full data set was reduced to 16 necessary variables to avoid heavy computational load of the full 96-variable data set.

2. **Adding text identifiers**

   i. *Text cleaning*: First of all, the titles and descriptions were lowercased and cleaned by removing all numbers and punctuation signs in an effort to prepare the text for the creation of unique text identifiers. This is done to avoid unnecessary inclusion of projects that differ only slightly (e.g. by a number or comma).

```
library(tm)

# Define function to clean titles
clean_titles <- function(title){
  title <- title %>%
    removeNumbers %>%
    removePunctuation(preserve_intra_word_dashes = TRUE) %>%
    tolower
  return(title)
}

df_crs <- df_crs_raw %>%
  mutate(projecttitle = clean_titles(projecttitle),
         shortdescription = clean_titles(shortdescription),
         longdescription = clean_titles(longdescription))
```

   ii. *Id creation*: Each project title and description is given a specific id in order to be able to analyze only distinct titles and descriptions later on. These were created using a well-known hashing algorithm called "xxHash" that is reasonably fast and exhibits very good collision properties (see https://github.com/Cyan4973/xxHash).

```
library(digest)

df_crs <- df_crs %>%
  rowwise() %>% # use rowwise operations since digest concatenates vector of strings
  mutate(text_id = digest::digest(longdescription, algo = "xxhash32")) %>% # add text_id as ha
```

   iii. *descr2mine*: Due to lazy reporting, frequently the descriptions differ only marginally from the project titles. This would pose a problem to the previously outlined twofold procedure since descriptions that are identical to the project titles would be analyzed twice. Therefore, only distinct

[4] Originally both short and long description present in CRS data; from now one referred to as description

descriptions are used which are identified using the Damerau-Levenshtein-Distance that counts how many alterations it would take to align both texts. The threshold for the maximal distance was set to 10 since this includes spelling mistakes, as well as one-word deviations (e.g. Output: ...).

```
library(stringdist)

# Max string distance underneath which strings can be considered the same/differing just by a
max_string_dist <- 10
df_crs <- df_crs %>%
  mutate(descr2mine = ifelse(stringdist(projecttitle, longdescription) < max_string_dist | str
                             NA,
                             longdescription))
```

iv. *Crating identifiers*: The CRS data set contains information about purpose of the funding flow in form of the purpose code, as well as other valuable information in other markers such as the gender marker (add link to resource) or the certain channel codes (41146 for UN Women). Table 1 lists all added identifiers.

```
df_crs <- df_crs %>%
  mutate(scb = ifelse(purposecode == 16062, 1, 0), # Statistical capcity building identifier
         pop = ifelse(purposecode == 13010, 1, 0), # population policy identifier
         gen_ppcode = ifelse(purposecode %in% c(15170:15180), 1, 0), # add gender purpose code
         gen_marker = ifelse(gender == 1 & !is.na(gender), 1, 0), # add gender marker (0 - no
         gen_donor = ifelse(channelcode == 41146, 1, 0), # all projects from UN Women
         gen_sdg = str_detect(sdgfocus, "^5|,5")) # SDG 5: Gender equality
```

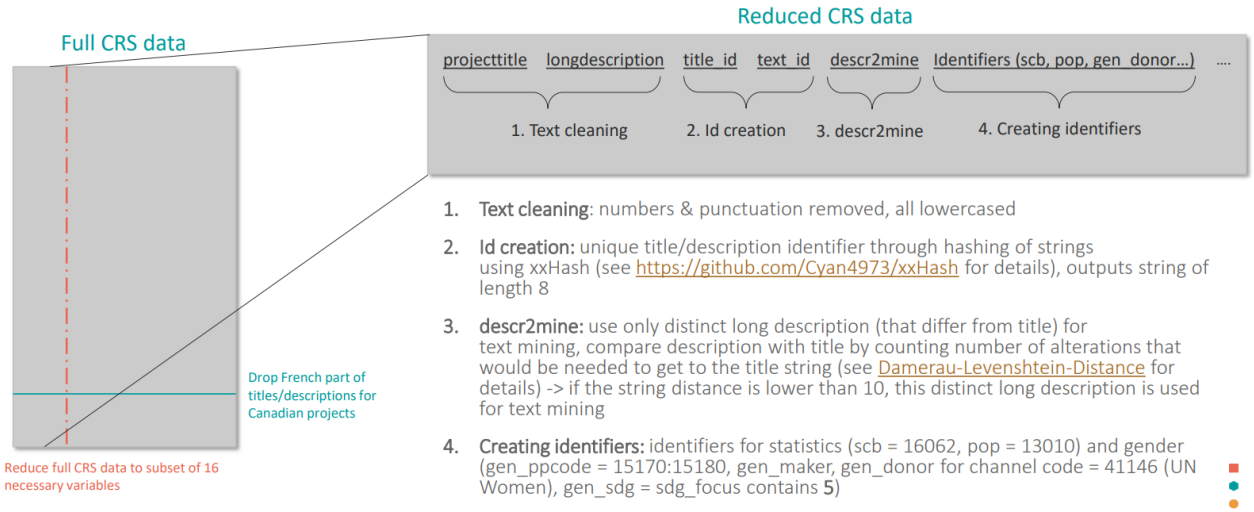

Figure 2: Diagram of the data preparation process.

3.3 A: Title pattern matching

In the following, the process of matching pertinent keywords in project titles is outlined (see Fig. 3 for a comprehensive overview).

1. **Preparing the data**

In the first step, the language of both title and description is detected using Google's Compact Language detector 2 (CLD2). It can detect 83 different languages and exceeds similar language detection engines by as much as 10x in speed. Analyzing the language distribution is crucial to a refined classification since text in every language has to be treated differently, using different keywords for the subsequent title pattern matching and fitting a different machine learning model later on. Therefore, the procedure was applied to projects in English, French, Spanish and German since these make up the majority of detected languages. This was implemented by selecting only projects with combinations of (title_language, long_language) in (en, fr, es, de, NA) x (en, fr, es, de, NA) while excluding the (NA, NA) combination. This combination was excluded since CLD2 in a vast majority of cases detects NA if the text is very short or nonsensical.

To give an overview how many projects are analyzed using this method, this approach encompasses 3.145.387 (90.6%) projects while 23.4020 (6.7%) were excluded for being (NA, NA) projects. That leaves 90241 (2.6%) projects that were excluded because they were either wrongly detected or belong to some minor reporting languages (e.g. Norwegian, Portuguese or Polish with a significant fraction within the 2.6% of excluded projects).

In the second step, duplicated project titles were dropped to analyze these titles only once during the title pattern matching procedure which again reduced computation time.

```
# All languages to include in classification - options: en, fr, es, de
languages <- c("en", "fr", "es", "de")

# Add unique title id and detect language of title and long description
df_crs <- df_crs %>%
  mutate(projecttitle_lower = tolower(projecttitle)) %>%
  rowwise() %>% # use rowwise operations since digest concatenates vector of strings
  mutate(title_id = digest::digest(projecttitle_lower, algo = "xxhash32")) %>% # create title id to
  ungroup() %>%
  mutate(title_language = cld2::detect_language(projecttitle)) %>%
  mutate(long_language = cld2::detect_language(longdescription))

# Use only projects in en, fr, es and de
df_crs <- df_crs %>%
  filter(title_language %in% c(languages, NA) & long_language %in% c(languages,NA)) %>%
  filter(!is.na(title_language) | !is.na(long_language)) # omit projects with both languages NA

# Select necessary columns and drop projects with duplicated title ids
df_crs <- df_crs %>%
  select(title_id, projecttitle, projecttitle_lower, longdescription, title_language, long_language
  filter(!duplicated(title_id))
```

2. **Title pattern matching**

   i. *Clean and lemmatize keyword lists*: For the treatment of the minority languages (French, Spanish and German), the English keyword list for statistics was translated by experts working in the field of official statistics. It contains many aspects of official development assistance in statistics and can be found in Appendix A. The keywords therein are chosen in a way that it is almost certain that a project is at least partly related to statistics if its title contains one of the keywords. The same was done for the English list of acronyms which can differ in other foreign languages. Together with the list for mining projects, the keyword lists were cleaned and lemmatized to guarantee that they will be matched to cleaned and lemmatized words occurring in project titles.

```
# list_keywords_stat, list_acronyms and demining_small_arms previously loaded

# Define lemmatization function
```

```
clean_and_lemmatize <- function (string){
  string <- string %>%
    tolower %>%
    removeWords("'s") %>% # remove possesive s so that plural nouns get lemmatized correctly,
    removeNumbers() %>%
    removePunctuation(preserve_intra_word_dashes = TRUE) %>%
    stripWhitespace %>%
    removeWords(c(stopwords('english'))) %>%
    removeWords(c(stopwords(source = "smart")[!stopwords(source = "smart") %in% "use"])) %>% #
    lemmatize_strings()
}

# Lemmatization for "en"
list_keywords_stat <- clean_and_lemmatize(list_keywords_stat)
demining_small_arms <- clean_and_lemmatize(demining_small_arms)

# Stemming for minority languages "fr", "es" and "de"
list_keywords_stat <- stem_and_concatenate(list_keywords_stat, language = lang)
demining_small_arms <- stem_and_concatenate(demining_small_arms, language = lang)
```

ii. *Clean and lemmatize titles*: Cleaning of project titles was achieved by removing numbers, punctuation and so called "stopwords" (e.g. "and", "the", "for") since they don't contain information towards the classification. Subsequently, words were lemmatized meaning to reduce different forms of a word to its lemma (e.g. "women", "woman's", "woman" -> "woman"). This is very important to guarantee that all various versions are found during the title pattern search. For minority languages however, stemming is used instead of lemmatization since no good lemmatization implementation was available.

```
df_crs <- df_crs %>%
  mutate(projecttitle_clean = ifelse(title_language == lang & !is.na(title_language),
                                     clean_and_lemmatize(projecttitle_lower),
                                     projecttitle_clean)) %>%
```

iii. *Keyword detection*: For every language, the project title was analyzed whether it contains one of the statistical keywords or acronyms. Note that statistical keywords were detected within cleaned and lemmatized titles whereas for acronyms, the original title was used since the lemmatization and stemming algorithms were found to change acronyms.

```
# Create regex for searching titles
list_keywords_stat <- paste0(" ", paste(list_keywords_stat, collapse = " | "),"|^", # words w
                             paste(list_keywords_stat, collapse = " |^"),"| ", # beginning of
                             paste(list_keywords_stat, collapse = "$| "), "$") # end of string

list_acronyms <- paste0(" ", paste(list_acronyms, collapse = " | "),"|^",
                        paste(list_acronyms, collapse = " |^"),"| ", # beginning of string
                        paste(list_acronyms, collapse = "$| "), "$") # end of string

demining_small_arms <- paste0(" ", paste(demining_small_arms, collapse = " | "),"|^",
                              paste(demining_small_arms, collapse = " |^"),"| ", # beginning of s
                              paste(demining_small_arms, collapse = "$| "), "$") # end of string

# Detect stat, acronyms and mining
df_crs <- df_crs %>%
  mutate(match_stat = ifelse(title_language == lang | is.na(title_language),
                             str_detect(projecttitle_clean, list_keywords_stat),
```

```
                              match_stat),
        mining = ifelse(title_language == lang | is.na(title_language),
                        str_detect(projecttitle_clean, demining_small_arms),
                        mining)) %>%
  mutate(match_stat = ifelse(title_language == lang | is.na(title_language),
                             str_detect(projecttitle_lower, list_acronyms) | match_stat,
                             match_stat))
```

iv. *Merging classes for final filter*: The reason to detect also mining projects was to exclude those projects from the statistics filter since expressions like "small arms survey", "survey of landmine situation" make frequent appearances in project titles but are not related to statistics. Hence, only projects for which a statistical keyword was detected but no mining keyword are marked as a statistical project in the pattern matching step.

```
# Exclude mining projects, since they contain survey -> not statistical project
df_crs <- df_crs %>%
  mutate(text_detection_wo_mining = match_stat & !mining) %>%
  mutate(text_detection_wo_mining_w_scb = match_stat | scb)
```

Lastly, the statistics filter is added back to the reduced data set according to the title id. This ensures that all projects with the same title in the reduced data set are marked as statistical by the title pattern matching.

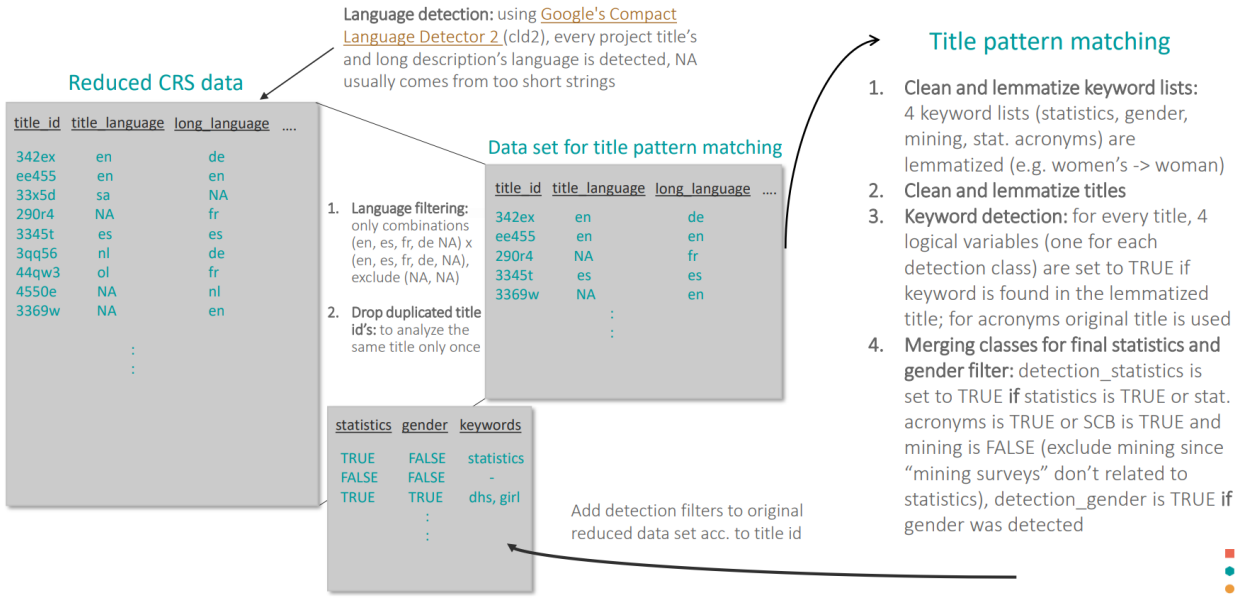

Figure 3: Schematic diagram of the title pattern matching.

3.4 B: Text mining of long descriptions

Lastly, the process of applying a machine learning approach to classify the projects' long descriptions will be explained in detail (see Fig. 4 for a comprehensive overview).

1. **Preparing the data**

i. *Language filtering*: For the preparation of the data, the reduced data set with the additional statistics filter from the pattern matching is filtered according to the description language to ensure that the text mining is applied only to text in one language. Note that there are projects with differing title and description language (frequently English title, minority language description) which is however no problem, since a project's description can be assumed to be statistical even when its title is in another language.

```
lang <- "en"

# Filter only projects with description language lang
df_crs <- df_crs_reduced %>%
  filter(long_language == lang)
```

ii. *Manual filter correction*: For 200 English projects, the description of projects, which were detected as statistical projects by the title pattern matching, were verified manually by experts. It can be the case that a projects title refers to statistics (e.g. "census aid") while its description contains no relevant information towards a classification ("Material and equipment for on the ground operations"). This additional step makes sure that the learning set contains less errors and hence increases the accuracy.

```
# Read manually verified projects
man_verified <- readRDS("./Data/Manually verified/stat_projects_verified.rds")

df_crs <- df_crs %>%
  filter(!is.na(descr2mine)) %>%
  select(text_id, description = descr2mine, longdescription, class_filter = text_detection_wo_m
  left_join(man_verified %>% select(longdescription, match_stat), by = "longdescription") %>% s
  mutate(class_filter = ifelse(!is.na(match_stat), match_stat, class_filter)) %>% # replace cl
  select(-longdescription, -match_stat)
```

iii. *Drop duplicated text ids*: As for the title ids, duplicated text ids are dropped to reduce the computational load during the text mining. In addition, some projects shared a discription but differed in their title. If one of the projects was detected as `TRUE` and one as `FLASE` in step **A**, both of them were discarded to reduce errors in the training set later on.

```
df_crs <- df_crs_reduced %>%
  filter(!is.na(descr2mine)) %>%
  distinct() %>%
  group_by(text_id) %>% # remove all ambiguous projects (same description, one FALSE one TRUE)
  filter(n() == 1) %>%
  ungroup() %>%
  as.data.frame
```

2. **Text mining of long description**

i. *Construct learning and prediction set*: For this machine learning approach, it is necessary to construct a balanced learning set which contains 50% negatively marked (NM) and 50% positively marked (PM) projects. The projects detected in Step **A** are used as the PM projects since it is reasonable to assume that if the title contains statistical keywords, also its description refers to statistics. The NM projects are chosen randomly because it can be assumed that only a small fraction of projects refer to statistics and therefore the probability to introduce error into the learning set is very small. The prediction data set contains simply the rest of the NM projects in the text mining data set.

```r
# Define parameters
frac_pred_set <- 1             # use only x% of full prediction set to speed up for testing
full_learning_percent <- 1     # take only x% of full learning set size if too large for RAM
neg_sample_fraction <- 1       # fraction of NM to PM in learning set

# Get size of PM projects in learning set
size_positive_train <- neg_sample_fraction * full_learning_percent * df_crs %>% filter(class_fi

# Construct prediction set
pred <- df_crs %>%
  filter(class_filter == FALSE | is.na(class_filter)) %>%
  sample_n(size = frac_pred_set * n())

# Error: if size of pred smaller than size of PM projects, not possible to construct training
if(pred %>% filter(!is.na(class_filter)) %>% nrow < size_positive_train) stop("Pred not large e

# Construct training set
learning <- df_crs %>%
  filter(class_filter == TRUE) %>%
  sample_n(size = n()*full_learning_percent) %>%
  rbind(pred %>% filter(!is.na(class_filter)) %>% sample_n(size = size_positive_train)) # add

# Exclude NM projects in training set from pred
pred <- pred %>%
  filter(!text_id %in% train$text_id)
```

ii. *Clean and lemmatize descr2mine*: As previously discussed, only distinct long descriptions (distinct from title) are used to avoid analyzing the same text twice. These are then cleaned and lemmatized to reduce the text to the relevant information.

```r
# Set languages for stemming and lemmatization
stem_languages <- c("de", "fr", "es")
lemma_languages <- c("en")

# Change original description with cleaned description
if (lang %in% lemma_languages) {
  learning$text_cleaned <- clean_and_lemmatize(learning$description)
  print("Start lemmatize pred")
  pred$text_cleaned <- clean_and_lemmatize(pred$description)
  print("Finished lemmatization pred")
} else if (lang %in% stem_languages) {
  learning$text_cleaned <- stem_and_concatenate(learning$description, language = lang)
  pred$text_cleaned <- stem_and_concatenate(pred$description, language = lang)
}
```

iii. *Create DTM matrices*: After splitting the learning set into the training set and testing set in a ration of 80/20, the document term matrix (DTM) is created for the training set. It has all the words that are present in all descriptions of the training data set (terms) as columns and collects their weighted frequency for each project in the respective row. For creating the DTMs of the test data and prediction data, terms occurring in the training data DTM are used which means that the all DTMs share the same columns. This is important for the prediction step later on since the model is only trained on these terms and assigns a relative weight to each of them. Therefore, it can only predict on terms that has already "seen".

```
# Take 80% training data, 20% testing data
dt <- sort(sample(nrow(learning), nrow(learning)*0.8))
train_data <- learning[dt,]
test_data <- learning[-dt,]

# Construct DTMs
train_data_dtm <- train_data$text_cleaned %>% VectorSource() %>% VCorpus() %>% DocumentTermMatr
dictionary_dtm <- Terms(train_data_dtm) # use only terms appearing in training data to constru
test_data_dtm <- test_data$text_cleaned %>% VectorSource() %>% VCorpus() %>% DocumentTermMatri
prediction_data_dtm <- pred$text_cleaned %>% VectorSource() %>% VCorpus() %>% DocumentTermMatri
```

iv. *Training the XGBoost model*: The model is obtained from the regularizing gradient boosting framework XGBoost by fitting the training data. Due to the broad literature on this machine learning approach, a detailed discussion shall be refrained from here. It can be said however that by passing along the training data DTM alongside the correct classification labels, the XGBoost model identifies the most important words appearing in the PM projects and assigns a high importance to them (see Fig. 5 below).

```
# Set the labels for class_filter
label.train <- as.numeric(train_data$class_filter)

# Training parameters
eta_par <- 0.1
nrounds_par <- 5 / eta_par

# Train the model
fit.xgb <- xgboost(data = as.matrix(train_data_dtm), label = label.train, max.depth = 17, eta =
                   nrounds = nrounds_par, objective = "binary:logistic", verbose = 1)
```

v. *Testing and prediction*: The model is then assessed using the test data. Since the model returns a score p_stat in the range from 0 to 1 whether a project's description refers to statistics, different thresholds are tested to see how the model performs (more in Appendix B). Finally, all projects in the prediction set are predicted using the fitted model. If a project receives a score of $p_{stat} \geq 0.9$, it is marked as statistical by the text mining (justification of threshold).

```
# Predict test and pred data
test.xgb <- predict(fit.xgb, as.matrix(test_data_dtm))
pred.xgb <- predict(fit.xgb, as.matrix(prediction_data_dtm))

# Set all projects to 1 for a score higher than 0.9
threshold <- 0.90
test_data <- mutate(test_data, predictions = ifelse(predictions_raw > threshold, 1, 0))
pred <- mutate(pred, predictions = ifelse(predictions_raw > threshold, 1, 0))

# Show accuracy
accurracy <- mean(test_data$predictions == test_data$class_filter)
print(accurracy)
```

vi. *Iteration of step i.-v. for learning set robustness*: In step 1, the 50% NM projects were chosen at random since the probability that statistical project is in this set is very small. However, it could still be the case that the statistical projects are included by chance. This can be almost avoided by repeating steps 1. – 5. with a training set that is constructed using only projects that are predicted not to be statistical with $p_{stat} \leq 0.3$. This threshold is chosen because it makes sure that the training set is only constructed from true NM projects while not being too restrictive and potentially introducing a bias into the training set (e.g. if all projects with $p_{stat} \leq 0.05$ stem

from the agriculture sector). On average, this iterative procedure increases the accuracy by 5% - 10% depending on the size of the prediction set.

```
# Filter projects with low score
pred_negative <- pred %>%
  filter(predictions_raw <= 0.3) %>%
  sample_n(size = size_positive_train) %>%
  select(text_id, description, class_filter)

# Construct new learning set with low-score projects as NM
learning <- df_crs %>%
  filter(class_filter == TRUE) %>%
  sample_n(size = n()*full_learning_percent) %>%
  rbind(pred_negative) %>%
  filter(!is.na(class_filter))

# Construct pred from all NM projects that are not in the training set
pred <- df_crs %>%
  filter((class_filter == FALSE | is.na(class_filter)) & !(text_id %in% pred_negative$text_id)) %>%
  sample_n(size = frac_pred_set * n()) #use only frac_pred_set% to speed up for testing

# Repeat step i. - v.
```

Finally, the text mining filter is added back to the reduced data set according to the text id. This ensures that all projects with the same description in the reduced data set are marked as statistical by the text mining methodology.

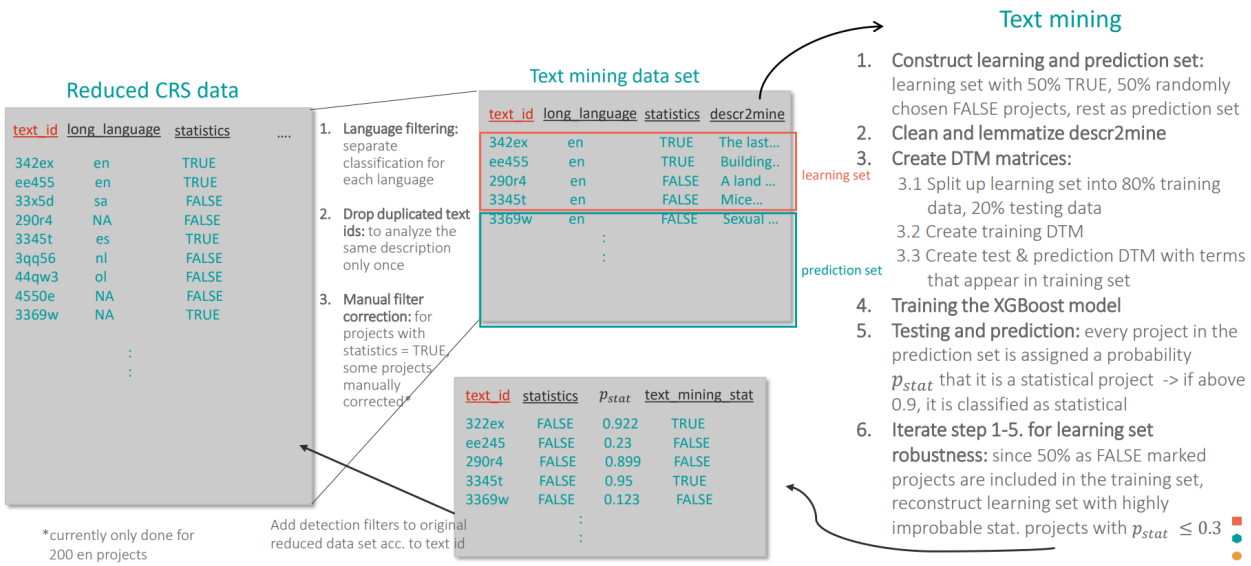

Figure 4: Schematic diagram of the text mining.

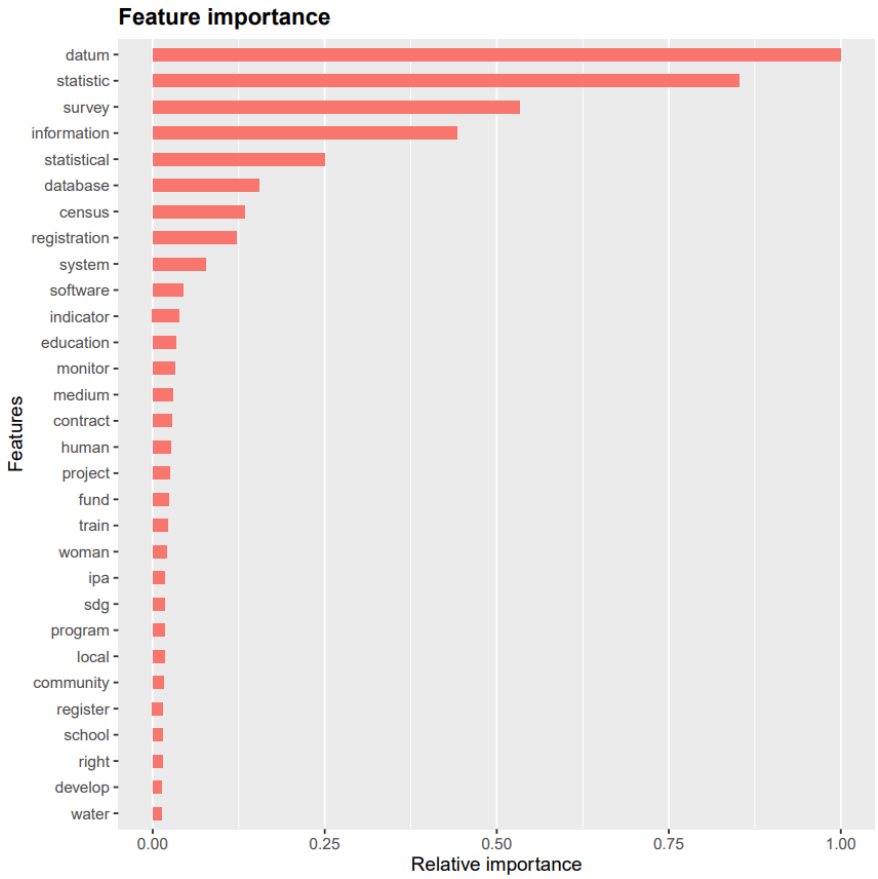

Figure 5: Relative importance assigned to terms appearing in long descriptions.

4 Monitoring funding to statistics with reduced reporting lag

4.1 What is the reporting lag?

The workflows for combining the two main data sources of PRESS are described in Fig. 6.

One key step when merging the PRESS data, reported by both donors and implementors, with the CRS data is avoiding duplication in a donor-implementor-recipient funding flow. To achieve this, the projects are examined against their unique identifier in both sources. The projects reported by implementers (mostly from the PRESS survey) are not counted as contribution of the reporting agencies. These projects are counted as projects by the donor agencies, after duplication checks were applied when merging the projects reported by implementers and the projects reported by donors. As the data and final report of PRESS depend in large part on the CRS database, which has a 12-months lag in coverage, the previous editions of PRESS did not capture timely donor financial flows to statistics, leading to a structural lag in reporting.

This lag meant that in its previous format, PRESS could not provide timely information for partners in data and statistics, including:

- Nowcasting the funding to statistics
- Forecasting funding to statistics

Hence, despite the many improvements in PRESS over the years, the lack of timely aid reporting is a persistent concern among its primary users, especially development aid providers. With a growing interest

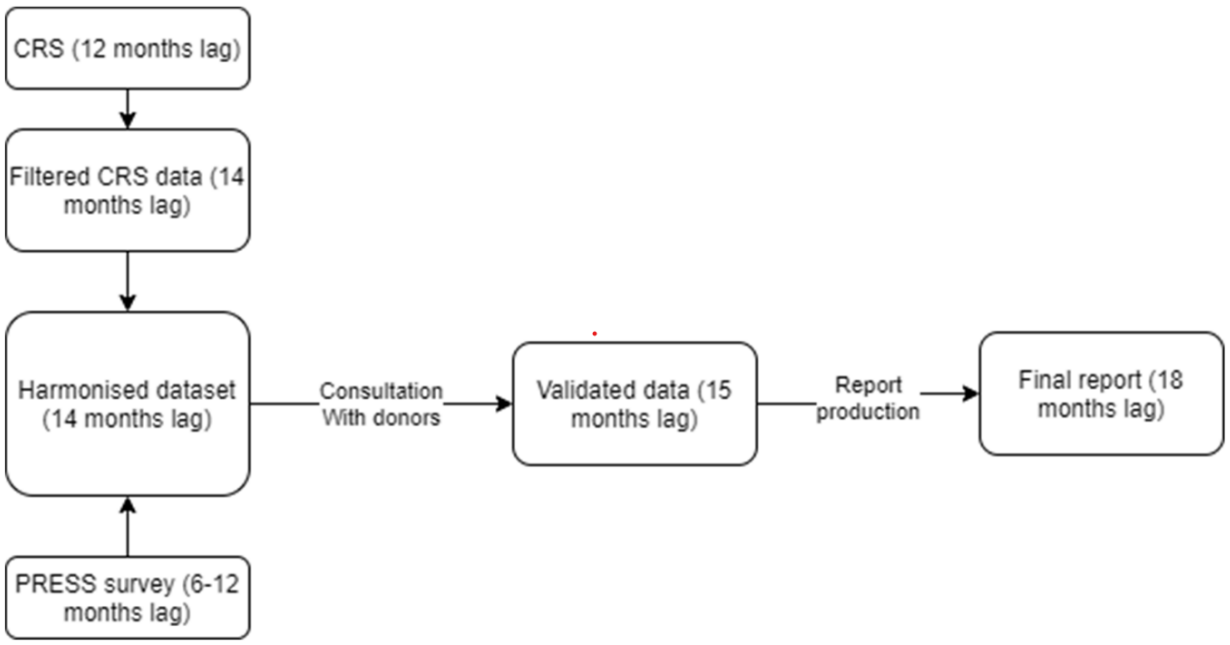

Figure 6: How the lag in the CRS data led to a lag in previous rounds of PRESS.

in supporting data and statistics, there is an increasing demand for timely data to plan activities and projects and coordinate development co-operation efforts. This issue has become particularly urgent in light of the coordinating efforts to fund the Cape Town Global Action Plan for Sustainable Development Data (CT-GAP),[5] as well as in the context of a diverse data ecosystem comprising new actors.

PARIS21 addressed this request in its 2019 annual meeting by introducing the concept of a methodology extension. While the PRESS methodology will still be used to report information until 18 months before the publication, the methodology extension will provide stakeholders with PRESS-like information on more recent periods, therefore reducing the reporting lag significantly. This concept became more relevant in 2020, when the development co-operation community had to face the challenges that arose due to the COVID-19 pandemic in national statistical systems (PARIS21 2020a) and funding to data by domestic and external stakeholders (PARIS21 2020b).

4.2 Estimating up-to-date support to statistics using CRS

While the previous PRESS (from 2008 to 2019) captured the support to data and statistics by looking at global **commitments**[6] to statistics, the annual **disbursements**[7] received by a certain country are also informative for donors and countries when planning their activities, especially those short-term activities financed by a donor's annual or biannual budget. Leveraging this additional variable allows for the estimation of funding to data and statistics received by countries in the current and coming years while still using the

[5] See https://unstats.un.org/sdgs/hlg/cape-town-global-action-plan/

[6] A firm obligation, expressed in writing and backed by the necessary funds, which is undertaken by an official donor. It provides specified assistance to a recipient country or a multilateral organisation. Bilateral commitments are recorded in the full amount of the expected transfer, irrespective of the time required for the completion of disbursements. Commitments to multilateral organisations are reported as the sum of (i) any disbursements in the year reported on, which have not previously been notified as commitments, and (ii) expected disbursements in the following year.

[7] The release of funds to or the purchase of goods or services for a recipient; by extension, the amount spent. Disbursements record the actual international transfer of financial resources, or of goods or services valued at the cost to the donor. In the case of activities conducted in donor countries, such as training, administration, or public awareness programmes, disbursement is assumed to have occurred when the funds have been transferred to the service provider or recipient. These may be recorded as gross (the total amount disbursed over a given accounting period) or net (the gross amount, less any repayments of loan principal or recoveries on grants received during the same period). It can take several years to disburse a commitment.

same base data, i.e., the CRS and PRESS surveys (and many other data sources on development aid, see section 3), which include both variables for each project.

Looking at disbursements instead of commitments to estimate the support to data and statistics has two distinct advantages:

1. Disbursements capture the actual release of funds, so are more useful for donor planning purposes.
2. It can take several years to disburse a commitment and some commitments are never disbursed. Hence, by design, there are more data points available on disbursements than commitments over the same time period. The additional data on disbursements allows for better understanding of financing patterns and donor behaviour, leading to more robust data analysis.

This availability of more data points enables us to estimate support to statistics in the current year (nowcasting) through robust regression analysis. It also provides more substantial evidence of funding trends in the coming years (forecasting). The following sub-sections will focus on how to arrive at these estimates.

4.2.1 Nowcasting: using commitments to predict current disbursements

Given that CRS has a lag of 12 months for reporting both disbursements and commitments, one way we can estimate support to statistics disbursed in the current year is by looking at the relationship between the two variables. The literature on aid predictability indicates that these two variables may be closely related over time. A 2013 study examining aid predictability based on CRS data also shows that commitments have a significant impact on disbursements five years after they were made (Hudson 2015).

For most development projects reported to the CRS and the PRESS survey, both commitment and disbursement data are reported. Even when these variables are not directly reported, however, the missing value can usually be imputed.[8]

Using these two variables, PARIS21 has developed a simple linear regression model to estimate the funding from donors based on historical data at activity level. Regression analysis was conducted to predict current disbursements based on reported commitments, captured by $Average\_Annual\_Spending$.

$$Disbursement = Average\_annual\_spending \cdot k + d$$

$$\text{where} \quad Average\_annual\_spending = \text{Total Project Commitments}/number\_of\_years$$

k is the regression coefficient and d is the error term. The number of years is the difference between the start date and end date of a reported activity. Reported dates are used for activities with missing value in those two variables. The analysis used the most recently accessible data from the CRS[9].

This model shows a correlation between disbursements and average spending. Average annual spending is calculated based on the assumption that commitment without a detailed plan for disbursement will be distributed evenly by year, from the expected start year to the end year of the project.

Table 1: Regression table from the analysis

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	38.455	8.04901	-4.778	0.0139*

[8]For example, the amount for technical support projects that do not have direct monetary transfers can be replaced by a cost estimate by the provider. In cases where the disbursement information is missing, the estimated disbursement amount can be calculated by dividing the unspent commitment amount using the number of years left before the expected end date

[9]See https://stats.oecd.org/DownloadFiles.aspx?DatasetCode=CRS

	Estimate	Std. Error	t value	Pr(>|t|)
Average_Annual_Spending	0.87461	0.05191	12.997	0.0456*

The analysis of CRS data shows a significant correlation (90%) between disbursements and commitments each year. The value of k and the predictability of the model vary depending on the reporting pattern of each donor. For example, while the commitment numbers reported by most donors each year are usually higher than disbursements (Fig. 7 on the left), this is reversed in the case of a few donors (Fig. 7 on the right)[10].

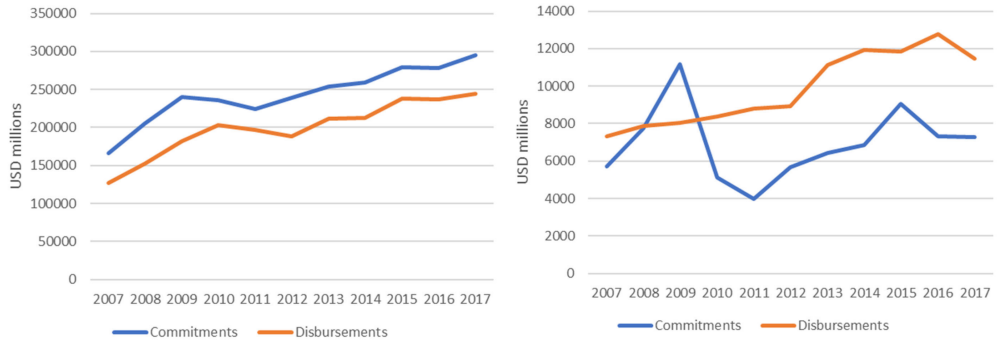

Figure 7: Disbursements vs commitments reported in CRS (all donors) on the left. On the right, disbursements vs commitments in the CRS reported by the UK.

Using the above method, PARIS21 was able to nowcast the funding to statistics in years that the most recent CRS data yet to cover. For instance, although the CRS data available in early 2020 only includes full coverage of official aid until 2018, the nowcast is able to provide information on aid to statistics including 2019 and 2020. This is because the stable relationship between average annual spending and commitments is leveraged, which allows us to estimate the 2019 and 2020 disbursement values, from the 2018 reported commitment values.

As a result, for the first time, the 2020 edition of PRESS presented information on funding to data and statistics up to 2019, as opposed to two years prior as in editions including and before PRESS 2019.[11]

4.2.2 Forecasting: anticipating future funding

The predictability of disbursements and commitments used for nowcasting funding to statistics decreases greatly after the current year, since, for example, many projects which commenced in 2017 will end in 2020. However, this lack of predictability can be partially mitigated by the creation of a forecasting model based on a few well-informed assumptions, leveraging past PRESS data and PARIS21's institutional knowledge on support to statistics for over two decades. These assumptions, which can lead to better forecasting quality, are described below:

1. **Continuation of certain long-standing projects**: We can assume that large projects such as the support to the Demographic Health Survey driven by USAID, IMF's national and regional training on economic statistics, and the World Bank's programme on statistical development will remain stable

[10]This particular reverse correlation can be explained by different factors. Firstly, the financial crisis impacted the continuity of some donors' ODA flow more than others. The significant variation of the exchange rate or inflation rate of a donor could also lead to a sudden change in the converted constant value of aid. In addition, some donors tend to make more long-term commitments than others, resulting in a distribution of disbursements over a long period of time, even after the donors had significantly reduced their overall international aid package

[11]Due to the disruptive effect of the COVID-19 pandemic on the predictability of the model, the nowcasting results for 2020, although produced, were not presented in the PRESS 2020 report.

in the near future. Significant changes on these programmes are also easier to target and detectable. These projects are generally stable and attract similar spends each year. Likewise, the upcoming censuses or major surveys in low-income countries are expected to be funded partially by donors.[12] This information accounts for nearly half of the total amount for data and statistics. Confirmation from donors of the continuation of these projects can further improve the accuracy of this analysis.

2. **Termination or reduction of funding for certain projects**: We can also anticipate the termination or reduction for funding tied to a project based on its specific nature. For instance, the support for censuses is a one-off disbursement and will not reoccur until the next census round. Similarly, if a country become no longer eligible for ODA, graduate from IDA's borrower list or becomes an upper-middle-income country, it is then expected to receive a lower ODA grant and become ineligible for some loans. In those cases, support for statistics might be affected disproportionally, given its low priority.

It is also crucial to state that these predictions can only be accurate if the following additional assumptions are met:

- Development aid providers maintain their current levels of effort
- Existing programmes continue to run
- Commitments are fully disbursed
- There is a response to prioritised needs such as censuses

The estimation is also limited if donors agencies publishing the new initiatives with a lag. The predictability of both nowcast and forecast on funding to data and statistics also relies on aid providers committing to maintaining the transparency and timeliness of their aid data. The forecasting looks at contribution until $n + 2$, given most international organisations' work plan don't go beyond that horizon.

Due to the uncertainty caused by the COVID-19 pandemic, the forecasting results from this methodology were not presented in PRESS 2020. These will appear in the future PARIS21 publications once more evidence becomes available.

In sum, the forecasting estimates should be interpreted with significant caution even if the above model indicates a relative increase in the coming years. According to many historical estimates, the funding gap for data and statistics (i.e. to find the entire CT-GAP) is far from being closed. This gap is likely to be exacerbated by the effects of the ongoing COVID-19 crisis in large parts of the globe

4.3 Expanding the PRESS database

4.3.1 Exploring alternative data sources for aid flows on statistics

Apart from nowcasting and forecasting disbursements to statistics from PRESS data, another way to address the structural lag in aid-flow reporting can be by attempting to remedy the root cause of the problem – the dependency on the CRS database – and searching for more timely information in alternative data sources. PARIS21 has identified three main (types) of alternative data sources, outlined below:

The International Aid Transparency Initiative (IATI) The IATI datastore is the largest alternative database outside of OECD-DAC data for official development assistance. With more than 100 donors reporting to this database, IATI has a much shorter lag than CRS. It also covers more projects by philanthropic foundations. The COVID-19 pandemic and the rising need for coordination has also incentivised aid providers to report to IATI with less delay. However, IATI data suffer from a lack of quality assurance and

[12]The COVID-19 pandemic has brought great uncertainty to this assumption, especially for censuses planned to take place between 2020 and 2021. Although funding for most censuses has been secured, many countries have diverted their national budget to other, more urgent matters (CCSA 2020; UNFPA 2020), resulting in the postponement of external funding

inconsistency within the dataset. Although it uses a similar data structure as the CRS, the reported projects in IATI may not include important granular information, such as the project description. Furthermore, as many donors only committed to reporting to IATI after 2014, the lack of historical data for drawing time series also affects its ability to forecast.

Donors' transparency portals In recent years, global donors have strengthened their efforts in aid transparency. Many donors have developed online data portals or uploaded online datasets to share information on their aid projects, especially the major donors in statistics such as the World Bank, UNDP, USAID, FCDO, IDRC, etc. These datasets usually have a similar density of information as the CRS data and are usually updated more frequently than CRS. However, the majority of donors still lack appropriate portals and public datasets. Furthermore, merging these different datasets is possible, but time intensive.

PARIS21 has been exploring these data sources since 2019 and has accumulated knowledge over this period. For example, the USAID dataset helped PRESS 2019 to identify the USA's support to statistics for the first time; in particular, its effort with the Demographic Health Survey (DHS). PARIS21 has also established a methodology for merging and harmonising the aforementioned datasets. The methodology maps variables in different datasets against each other and uses internal project identifiers to avoid duplications.

Multilateral donors' prospective reporting to the PRESS surve The online PRESS survey (introduced in Sec. 2.2) includes a feature that allows donors to report on future projects. Since many donors have a biannual programme of work, in each year's survey, they are encouraged to provide information on the project they have planned or committed to in the near future. In the survey, donors can verify, edit, and cancel future projects in the next round of reporting. However, previous editions of PRESS did not fully reflect future projects due to the report's focus on accuracy. However, these future projects could still contain valuable information to assist in the projection of aid flows. The methodology developed by PARIS21 leverages these projects as an underutilised existing data source that may not reflect completely on the activities from donors. Nevertheless, it is useful for nowcasting funding.

The comparison of the above data sources can be found in 8. As another important data source in the area of data and statistics, the Eurostat's donor survey is analysed in Box 1.

Box 1: The Eurostat's Donor Survey – and why it is not included in the PRESS database

The annual Donor Survey by Eurostat aims to provide an overview of statistical projects in the respective countries to allow for better planning of assistance in the field of statistics, to benefit from acquired experience, and to avoid overlap. The survey is distributed to donor countries, international organisations, and the recipients (mostly Eurostat member states or partner states) of support to data and statistics.

According to the report of the 2020 round, the Donor Survey's main objectives included:

- facilitating better planning and prioritisation of assistance, especially amongst the donors;
- increasing transparency/visibility for donors and implementing agencies and potentially serving as an "audit" document for beneficiaries;
- providing the beneficiaries with a broad overview of the areas in which they are receiving support, also at regional level, and prioritising their future needs;
- benefitting and learning from the experiences, good practices, and shortcomings of other projects;
- facilitating a dialogue between beneficiaries, donors, and implementing agencies.

At donor country level and on the donors' side, respondents to the Donor Survey overlap with CRS. However, the CRS reporting is coordinated by one national focal point, while the Donor Survey questionnaire is distributed to both the aid agencies and national statistics offices who provide technical support. However,

Data source	No. of observations	Advantages	Disadvantages
Conventional data source			
CRS	More than 2.5 million	• Full coverage of DAC donors' portfolios • Data quality assured by the WP-STAT[12] standards	• 12-month lag • Relatively moderate coverage of multilateral organisations
Alternative data sources			
PRESS survey	4000 to 5000	• Only statistics- or data-related projects are included • Allows reporters to enter future planned projects	• Only covers key multilateral donors in statistics • Detailed information may be omitted by reporters during the reporting process
IATI[13]	More than 500,000	• A shorter lag in reporting than CRS • Reporting from NGOs and philanthropic foundations • Relatively wider coverage on multilateral organisations • More activities reported over the years	• Lack of quality assurance • Incomplete portfolio
Donor transparency portals	The full portfolio for each donor	• Full coverage of donors' portfolios • Well-maintained and updated	• Fewer than ten major donors provide accessible databases

Figure 8: Comparison of data sources.

the Donor Survey data set does not have a good system to harmonise the data from the two sides, nor does it contain the program identifiers, which are essential to avoid duplication when combining information with other sources. Unlike the development of financing databases, the Donor Survey also lacks a mechanism to update previous years' reported projects. Coverage of the survey is also not comprehensive enough to provide additional values to the alternative data sources identified in this document. :::::

4.3.2 Addressing gaps in the alternative data sources

To take advantage of the alternative data sources, PARIS21 combined the information from CRS and online surveys with these alternative data sources to create a more up-to-date, harmonised database on funding for data and statistics. However, this means that two important problems must be addressed, described below:

1. **The completeness problem**: A common weakness of the alternative data sources, compared with CRS, is their incomplete coverage. This is especially apparent in IATI where, unlike CRS, donors may not report their full portfolio, leading to a lack of comprehensive information (a "vertical" information gap i.e. multiple donors but incomplete project information for individual donor). Furthermore, the CRS uses a more "centralized" reporting system for each donor country, whereby information from different agencies providing ODA is gathered under one entity before reporting to CRS as a whole. The IATI, on the other hand, allows different agencies in a single donor country to report their data separately. This implies different reporting patterns for different agencies based on their capacity to do so regularly, and a lack of overall coordination.
2. **The coverage problem**: CRS contains information from over 100 donors. On the other hand, donor transparency portals suffer from a "horizontal" problem, i.e., they usually have complete coverage and contain the full portfolio, but fewer than ten major donors provide access to such open and easy-to-use databases.

PARIS21 has mitigated the horizontal and vertical problems in alternative data sources by harmonising and linking these databases. The alternative sources provide a wider coverage, while the combination of CRS, IATI and donors' databases enhance the completeness of the data. The final data set used in the analyses combines the alternative data sources and PRESS data using project IDs and other identifiers. As a caveat, however, the total number of projects included in the alternative sources still only represents 40% of all number of projects reported in the CRS.

The advantage of using such data sources, such as the timeliness and inclusion of philanthropic foundations, makes them a useful extension of PRESS data, especially when trying to solve the lag issue. However, their weaknesses imply that they are not a substitute for CRS or, by extension, conventional PRESS analysis.

4.3.3 Linking the alternative sources: the new harmonised database

The next step in leveraging the independent alternative data sources described above is to link them with PRESS data and create a new harmonised database of disbursements for data and statistics **at project level.**

Using disbursements as the primary variable to determine the support to statistics is even more beneficial at this stage, and data on annual disbursements is adequately available in most of the sources considered above. The activity-based CRS data, for example, contain nowcasting/forecasting regressions on this new data set by Simon etc. to see value add and/or consistency of our findings! Should we present this? I think you did it and found little to no change right? disbursement information for more than 98% of the projects. Similarly, the PRESS survey for multilateral donors provides specific information on disbursement plans (though it is project-based). The donors' transparency portals are also expenditure/disbursement based.

However, the IATI database is weaker in this regard: it contains an unusually high percentage of negative commitments or disbursements. For example, data downloaded from the IATI database in 2018 contained

negative value commitments or disbursements in 18% of projects reported[13]. In comparison, less that 1% of activities reported in CRS has negative value, mostly associated with loan repayments and correction of previous entries, i.e. not a database error.

For the harmonised database, the missing disbursement values were imputed based on the assumption that the commitments were distributed evenly from the start date to the end date of the project. The negative projects in IATI were corrected or removed by cross-validating against CRS. The duplicated projects were removed based on project identifiers. Consultations with several multilateral donors were conducted to ensure the validity of the final data.

Fig. 9 presents an example of how the new harmonised dataset looks after linking the different sources. For the 15,312 projects in the new datasets, 54% of disbursement activities come from the CRS data, compared with over 73% in the earlier dataset used for PRESS. In the new dataset, the PRESS survey accounts for 27% of projects, while IATI data and donor databases account for 19% of projects. By filtering through the data using recipient country and year, donors can already observe the upcoming funding received by a country for statistical development. It is then easier for them to identify funding gaps in prioritised areas. Hence, the new harmonised database can achieve better diversification of data sources and reduce the dependence on CRS.

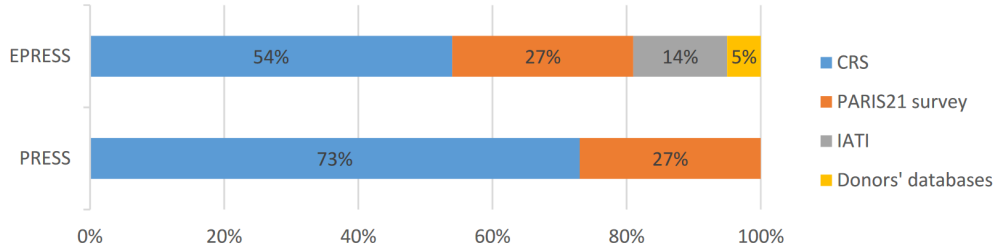

Figure 9: Comparison of PRESS and the new harmonised database for the share of projects in the final data sets, by sources of data, 2016-2018.

4.4 Bringing them together – nowcasting and forecasting with the new harmonised database

PARIS21 applied the same model on nowcasting and forecasting support to statistics (disbursements) but based on the new harmonised database, linking PRESS data with the aforementioned alternative sources. The main findings of this approach were similar to those based on CRS estimations: there is no indication of a systematic increase in funding to statistics in the current or coming years.

5 Conclusion

The nowcast and forecast analysis aims to address the specific reporting lag issue faced by the PRESS, rather than to substitute it. Even with longer lag, PRESS still reports the most reliable and comprehensive data on funding to statistics. The PRESS database continues to serve as the source data for SDG Indicator 17.19.1.

The outputs from nowcast and forecast also vary in their accuracy. While the merged database is as robust as PRESS, and the nowcast results are relatively reliable, the forecast analysis is based on several assumptions. Results produced through the nowcast and forecast analyses can therefore be used for different products to serve different purposes:

[13]Based on data collected in 2019, the future improvements of the IATI database have resolved this issue.

- The nowcast on disbursements and some information provided by the harmonised dataset can be directly presented in PRESS going forward, as a natural extension of its current content, based on existing data sources such as CRS.
- The forecast for funding gaps could be presented in separate policy briefs due to its speculative nature.
- The complete, harmonised database can also be deployed on a platform such as Bern Network's Clearinghouse for Financing Development Data to be used as a dynamic planning tool by development partners.

The immediate next step in improving the new methodology is to strengthen the communication and consultation between PARIS21, donors, and recipients. Benefits from such consultations would lead to further robust results by correcting erroneous information contained in IATI data, validating assumptions on the termination or continuation of projects in assumptions for forecasting, and enhancing data sharing in general. Consultations for this product will also help PARIS21 to shape its work in order to better meet the demand of its stakeholders.

As the COVID-19 pandemic will affect donors' GNI and consequently their ODA, fluctuations in development financing for data and statistics can be expected. Despite the difficulty these anomalies will bring to the analyses, the transparency and timeliness of the information on funding to data and statistics have also become more relevant, as identified in Part I of PRESS 2020. The results from these analyses can play a crucial role in informing and supporting

References

CCSA. 2020. "How COVID-19 is changing the world: a statistical perspective." taken from https://unstats.un.org/unsd/ccsa/documents/covid19-report-ccsa_vol2.pdf.

Hudson, John. 2015. "Promises kept, promises broken? The relationship between aid commitments and disbursements." *Review of Development Finance* 3: 109–20. https://doi.org/10.1016/j.rdf.2013.08.001.

PARIS21. 2020a. "Combating COVID-19 with data: what role for national statistical systems?" Paris: Taken from https://paris21.org/sites/default/files/inlineâĂŘfiles/COVID_Policybrief_Full.pdf?v=2.0.

PARIS21. 2020b. "Partner Report on Support to Statistics 2020." Paris: https://paris21.org/sites/default/files/inline-files/PRESS2020%20Final.pdf.

UNFPA. 2020. "Technical Brief on the Implications of COVID-19 on Census." Taken from https://www.unfpa.org/sites/default/files/resourceâĂŘpdf/Census_COVID19_digital.pdf.

A Appendix - Keyword lists

English

statistical_keywords	demining_keywords	acronyms
statistics	land mine	dhs
survey	small arm	crvs
census	demining	cpi
database	demine	d4d
big data	landmine	paris21
data for decisions		gpsdd
civil registration		odw
civil register		data4sdg
land registration		devinfo
land register		nsds
cadaster		afristat
sdg indicator		ckan
information system		lfs questionnaire
birth registration		
birth register		
business register		
national account		
price index		
production index		
data science		
data for development		
data journalism		
data for education		
education data		
data for health		
peacebuilding data		
global data		
global pulse		
health data		
refugee data		
migration data		
data collection		
action through data		
data project		
open government data		
death regist		
vital registration		
data portal		
data archive		
archive data		
data dissemination		
disseminate data		
microdata		
metadata		
data management		

statistical_keywords	demining_keywords	acronyms
data management		
data documentation		
quality data		
data quality		
access data		
open data		
use data		
produce data		
production data		
data user		
data producer		
data outreach		
data awareness		
data production		
data processing		
data access		
data harmonization		
harmonization data		
mdg indicators		
data standards		
data curation		
curating data		
demographic data		
mdg monitoring		
sdg monitoring		
monitoring mdg		
monitoring sdg		
release data		
data release		
prsp monitoring		
data revolution		

French

statistical_keywords	demining_keywords	acronyms
statistiques	mine terrestre	dhs
enquête	arme de poing	enquêtes démographiques et de santé
recensement	déminage	EDS
base de données	déminer	crvs
données massives		cpi
des données pour les décisions		IPC
état civil		d4d
registre foncier		paris21
cadastre		gpsdd
Indicateur sdg		odw
système d'information		data4sdg
registre des naissances		devinfo
registre des entreprises		nsds
compte national		afristat
indice des prix		ckan
indice de production		enquête emploi
science des données		
données pour le développement		
journalisme de données		
données pour l'éducation		
données sur l'éducation		
données pour la santé		
données sur la consolidation de la paix		
données mondiales		
global pulse		
données de santé		
données sur les réfugiés		
données de migration		
collecte de données		
action par les données		
projet de données		
données gouvernementales ouvertes		
registre des décès		
enregistrement vital		
portail de données		
archives de données		
données d'archives		
diffusion des données		
disséminer les données		
microdonnées		
métadonnées		
gestion de données		
documentation des données		
données de qualité		
qualité des données		

statistical_keywords	demining_keywords	acronyms
qualité des données		
données d'accès		
données ouvertes		
utilisateur de données		
producteur de données		
données de production		
utilisateur de données		
producteur de données		
sensibilisation aux données		
connaissance des données		
production de données		
traitement des données		
accès aux données		
harmonisation des données		
données d'harmonisation		
indicateurs omd		
normes de données		
préservation des données		
conservation des données		
données démographiques		
surveillance des omd		
suivi des omd		
surveillance des odd		
suivi des odd		
données publiées		
publication des données		
surveillance des dsrp		
révolution des données		

Spanish

statistical_keywords	demining_keywords	acronyms
estadísticas	mina terrestre	dhs
encuesta	Arma de fuego corta	encuesta de demografía y salud
censo	desminado	EDS
base de datos		crvs
datos masivos		cpi
datos para la toma de decisiones		IPC
registro civil		d4d
registro de la propiedad		paris21
catastro		gpsdd
indicador sdg		odw
sistema de información		data4sdg
registro de nacimiento		devinfo
registro de empresas		nsds
cuenta nacional		afristat
índice de precios		ckan
índice de producción		lfs
ciencia de los datos		epa
datos para el desarrollo		
periodismo de datos		
datos para la educación		
datos sobre educación		
datos para la salud		
datos sobre la construcción de la paz		
datos globales		
pulso global		
datos sanitarios		
datos de los refugiados		
datos de migración		
recolección de datos		
acción a través de los datos		
proyecto de datos		
datos públicos abiertos		
registro de defunción		
registro vital		
portal de datos		
archivo de datos		
datos de archivo		
difusión de datos		
difundir los datos		
microdatos		
metadatos		
gestión de datos		
documentación de datos		
datos de calidad		
calidad de los datos		

statistical_keywords	demining_keywords	acronyms
calidad de los datos		
acceso a datos		
datos abiertos		
utilizar los datos		
producir datos		
datos de producción		
usuario de datos		
productor de datos		
alcance de los datos		
conocimiento de los datos		
producción de datos		
procesamiento de datos		
acceso a los datos		
armonización de datos		
datos de armonización		
indicadores mdg		
normas de datos		
conservación de datos		
curación de datos		
datos demográficos		
monitoreo de los objetivos del milenio		
monitoreo de los ods		
monitories de los ddm		
seguimiento de los ods		
datos de publicación		
publicación de datos		
seguimiento de los prsp		
revolución de los datos		
diseminación de datos		
diseminar datos		
usuario de estadísticas		
productor de estadísticas		

German

statistical_keywords	demining_keywords	acronyms
Statistik	Landmine	dhs
Umfrage	Handfeuerwaffe	crvs
Volkszählung	Minenentschärfung	cpi
Datenbank	Minenräumung	VPI
big Data	entmienen	d4d
Daten für Entscheidungen		paris21
Melderegister		gpsdd
Grundbucheintrag		odw
Kataster		data4sdg
sdg-Indikator		devinfo
Informationssystem		nsds
Geburtenregister		afristat
Personenstandsregister		ckan
Personenstandsregister		lfs
nationales Konto		Arbeitskräfteerhebung
Preisindex		AKE
Produktionsindex		
Data Science		
Daten für die Entwicklung		
Datenjournalismus		
Daten für Bildung		
bildungsbezogene Daten		
Daten für das Gesundheitssystem		
Daten zur Friedensförderung		
globale Daten		
globaler Puls		
Gesundheitsdaten		
Flüchtlingsdaten		
Migrationsdaten		
Datenerfassung		
Handeln durch Daten		
Datenprojekt		
offene Regierungsdaten		
Sterberegister		
Bevölkerungsstatistik		
Datenportal		
Datenarchiv		
Archivdaten		
Datenverbreitung		
Daten verbreiten		
Mikrodaten		
Metadaten		
Datenverwaltung		
Datendokumentation		
Qualitätsdaten		

statistical_keywords	demining_keywords	acronyms
Qualitätsdaten		
Datenqualität		
Zugangsdaten		
offene Daten		
Daten verwenden		
Daten produzieren		
Produktionsdaten		
Datenbenutzer		
Datenproduzent		
Datenübermittlung		
Datenbewusstsein		
Datenerstellung		
Datenverarbeitung		
Datenzugriff		
Datenharmonisierung		
Harmonisierungsdaten		
mdg-Indikatoren		
Datenstandards		
Datenkuratierung		
Daten kuratieren		
demografische Daten		
mdg-überwachung		
sdg-überwachung		
überwachung von mdg		
überwachung von sdg		
Freigabedaten		
Datenfreigabe		
prsp-überwachung		
Datenrevolution		

B Appendix - Performance tests

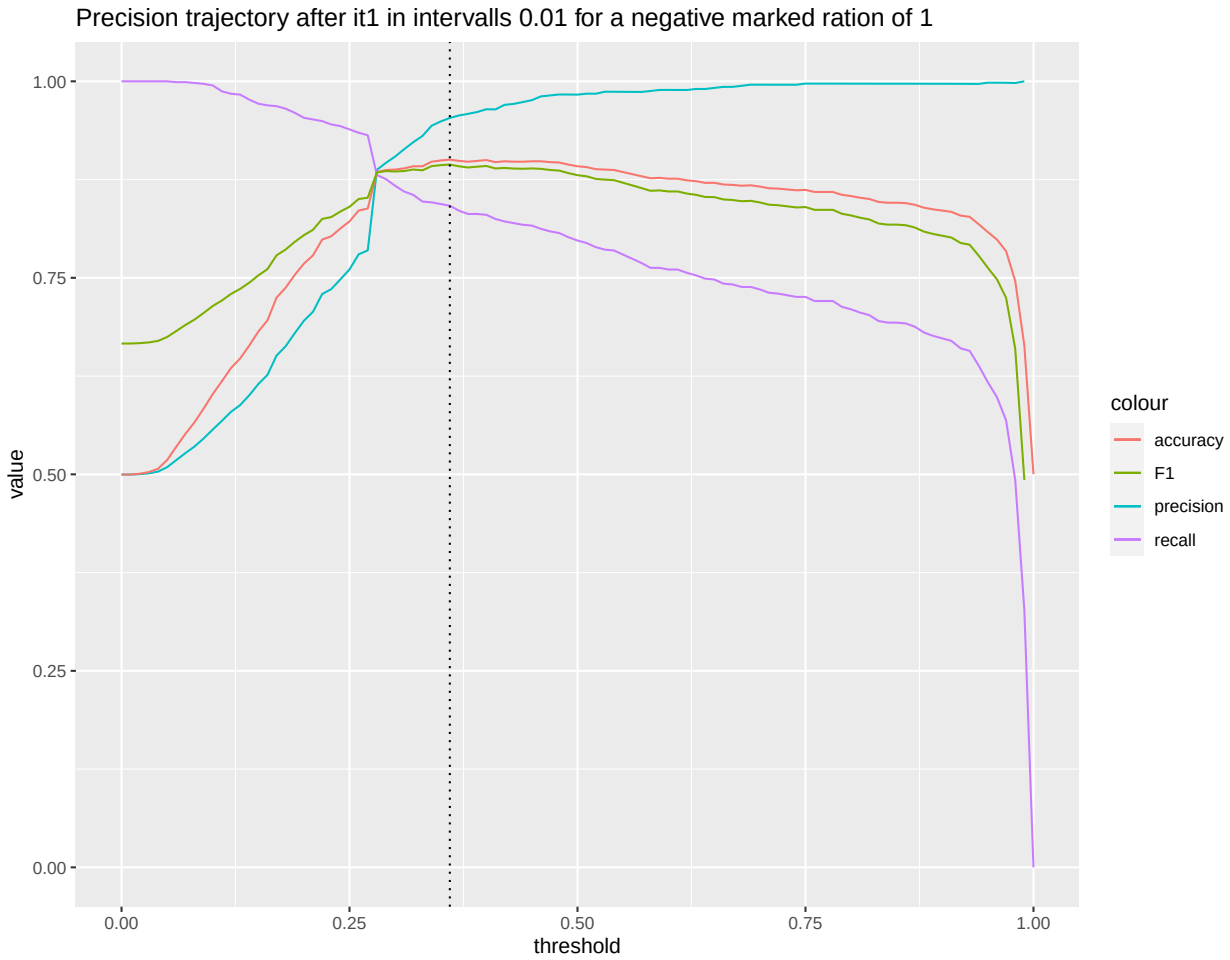

Figure 10: Precision, accuracy, recall and F1 score for the English classification

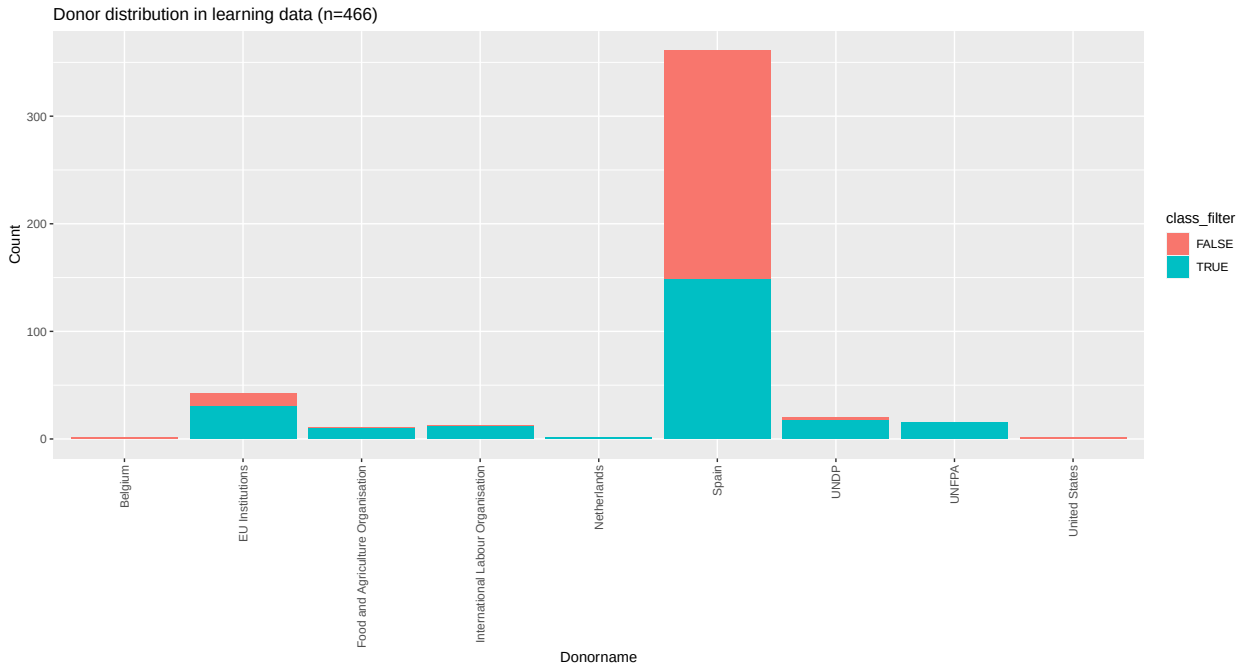

Figure 11: Donor distribution of the learning data for the Spanish classification.

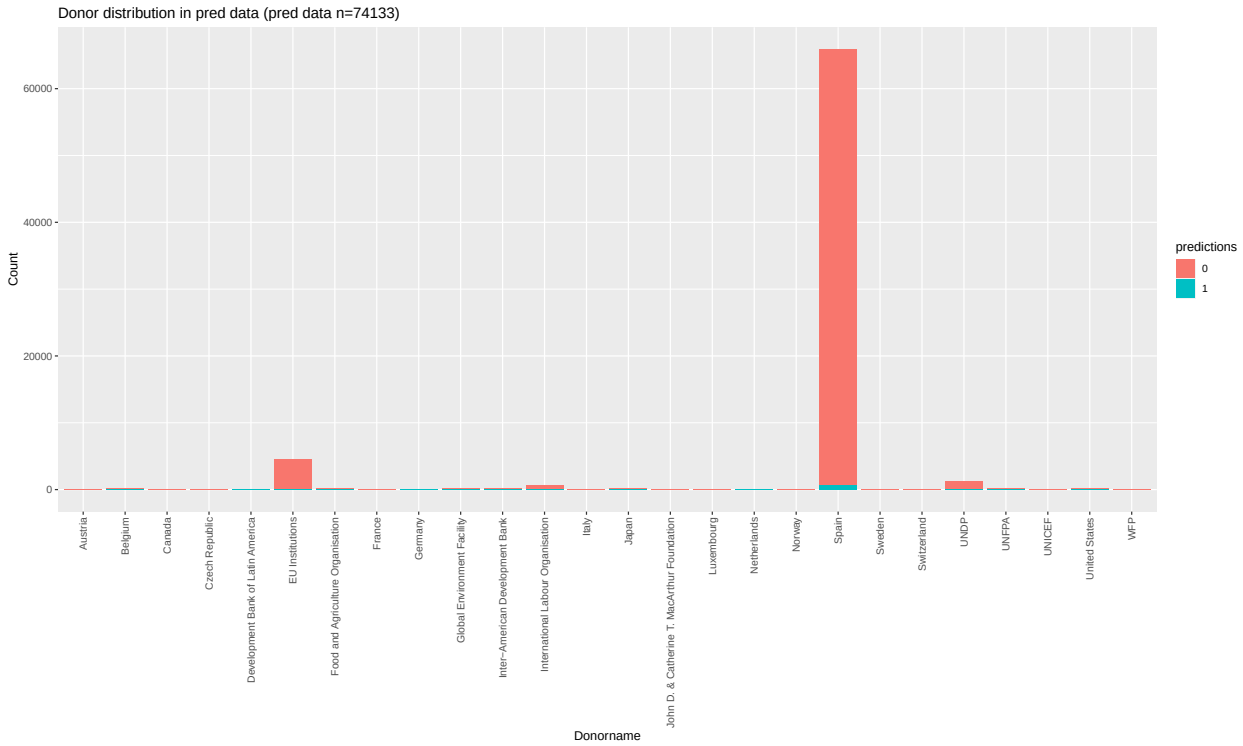

Figure 12: Donor distribution of the prediction data for the Spanish classification.