Becoming an economist

Data on economic PhD

Alexandre Truc, Aurelien Goutsmedt, Thomas Delcey

Introduction

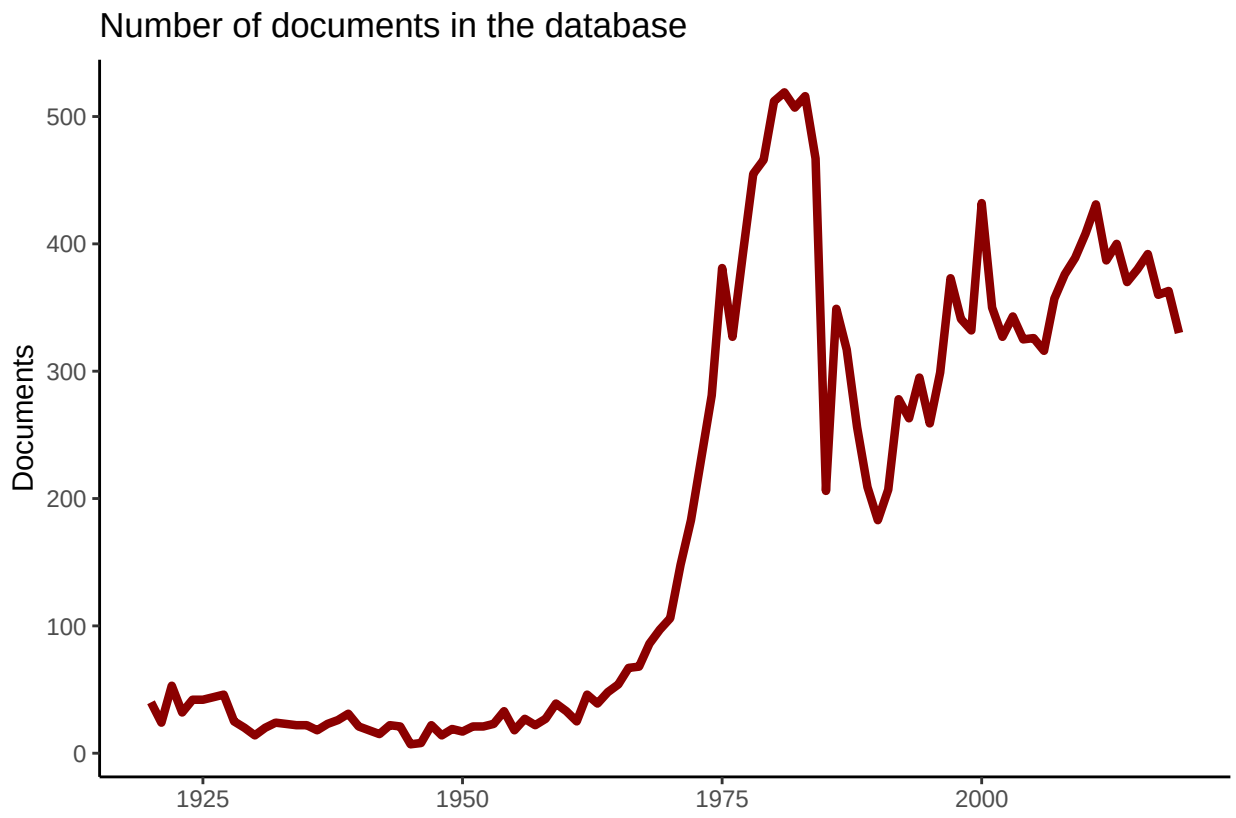
Data

A look at french Ph.D.

Documents

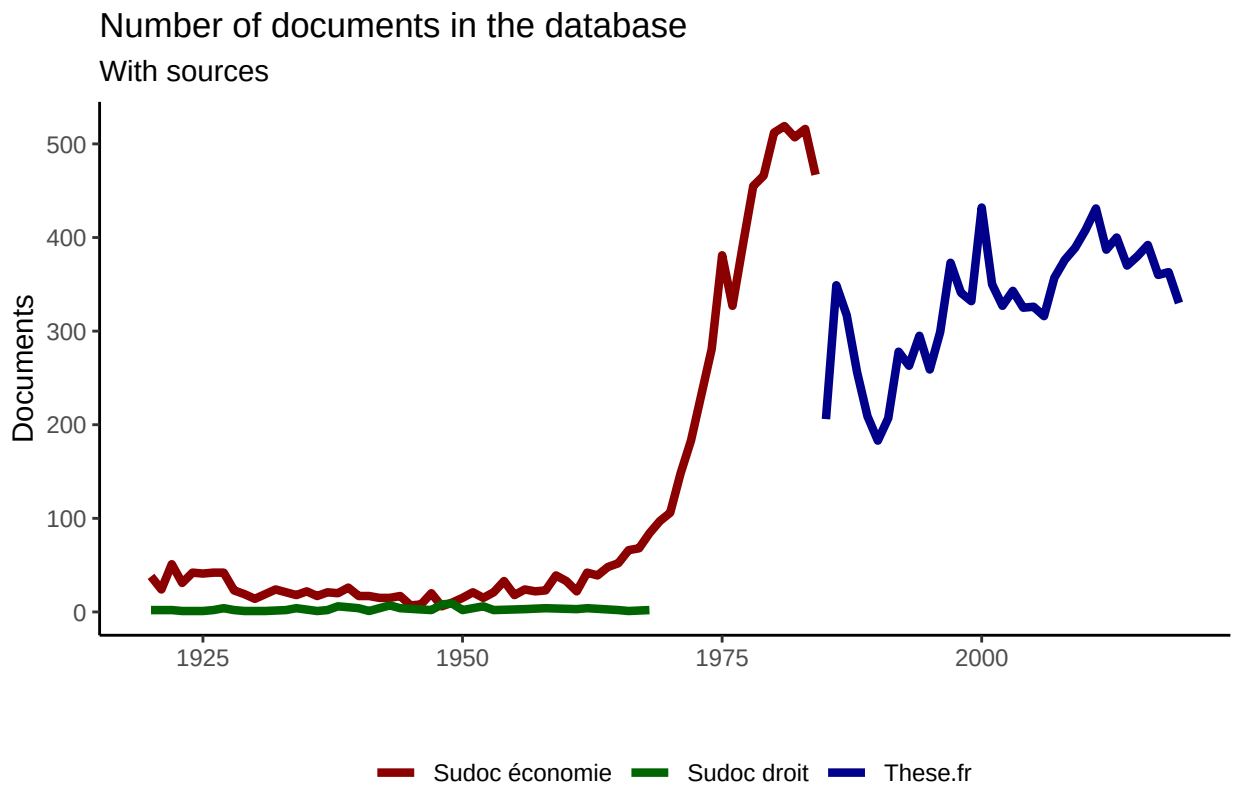
Distribution over time

```
thesis_table %>%
    filter(date < 2020) %>%
    count(date) %>%
    ggplot() + geom_line(aes(x = as.numeric(date), y = n), colour = "darkred",
    linewidth = 1.5) + labs(y = "Documents", x = "", title = "Number of documents in the database") +
    theme_classic()
```



Sources of data

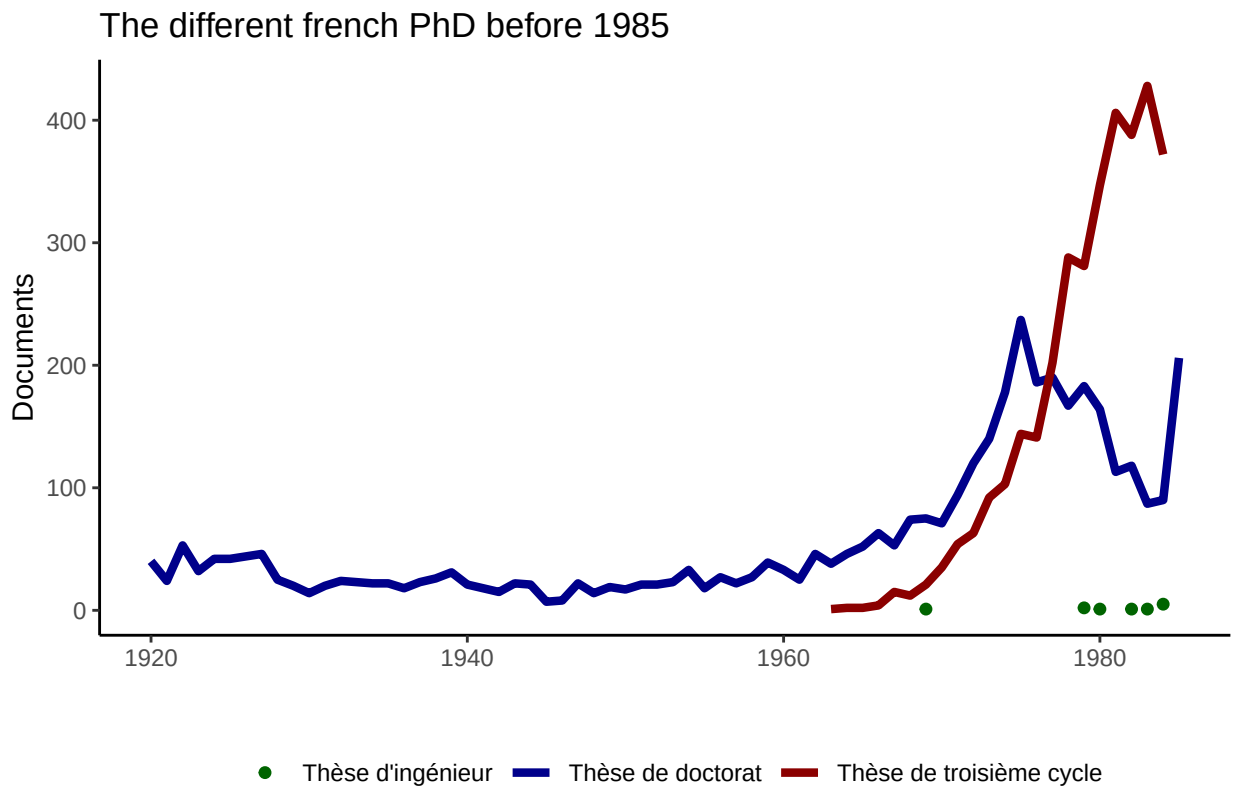
```
thesis_table %>%
    filter(date < 2020) %>%
    group_by(date, database) %>%
    summarise(n = n()) %>%
    ggplot() + geom_line(aes(x = as.numeric(date), y = n, colour = database),
    linewidth = 1.5) + scale_color_manual(name = "", values = c(sudoc = "darkred",
    thesefr = "darkblue", sudoc_law = "darkgreen"), labels = c("Sudoc économie",
    "Sudoc droit", "These.fr")) + labs(y = "Documents", x = "",
    title = "Number of documents in the database", subtitle = "With sources",
    caption = "Note: PhD in economics were defended in law faculty before 1968") +
    theme_classic() + theme(legend.position = "bottom")  #'she/he is an economists and did an econ PHD
```



Note: PhD in economics were defended in law faculty before 1968

The heterogeneity of French PhD before 1984

```
eval_thesis_kind <- thesis_table %>%
    ungroup() %>%
    select(doc_id, thesis_type, date) %>%
    mutate(thesis_type2 = case_when(str_detect(thesis_type, "3e cycle") ~
        "Thèse de troisième cycle", str_detect(thesis_type,
        "ingénieur") ~ "Thèse d'ingénieur", TRUE ~ "Thèse de doctorat")) %>%
    count(thesis_type2, date) %>%
    filter(date < 1986)

ggplot(data = eval_thesis_kind, aes(x = as.integer(date), y = n,
    color = thesis_type2, linetype = thesis_type2, shape = thesis_type2),
    ) + geom_line(linewidth = 1.5) + geom_point(size = 1.6, fill = "darkgreen") +
    scale_linetype_manual(name = "", values = c(`Thèse de troisième cycle` = "solid",
        `Thèse de doctorat` = "solid", `Thèse d'ingénieur` = NA)) +
    scale_shape_manual(name = "", values = c(`Thèse de troisième cycle` = NA,
        `Thèse de doctorat` = NA, `Thèse d'ingénieur` = 21)) +
    scale_color_manual(name = "", values = c(`Thèse de troisième cycle` = "darkred",
        `Thèse de doctorat` = "darkblue", `Thèse d'ingénieur` = "darkgreen")) +
    labs(x = "", y = "Documents", title = "The different french PhD before 1985",
        caption = "Note: The variable 'Thèse de doctorat' includes the classic PhD and the 'Thèse d'Eta
    theme_classic() + theme(legend.position = "bottom")
```



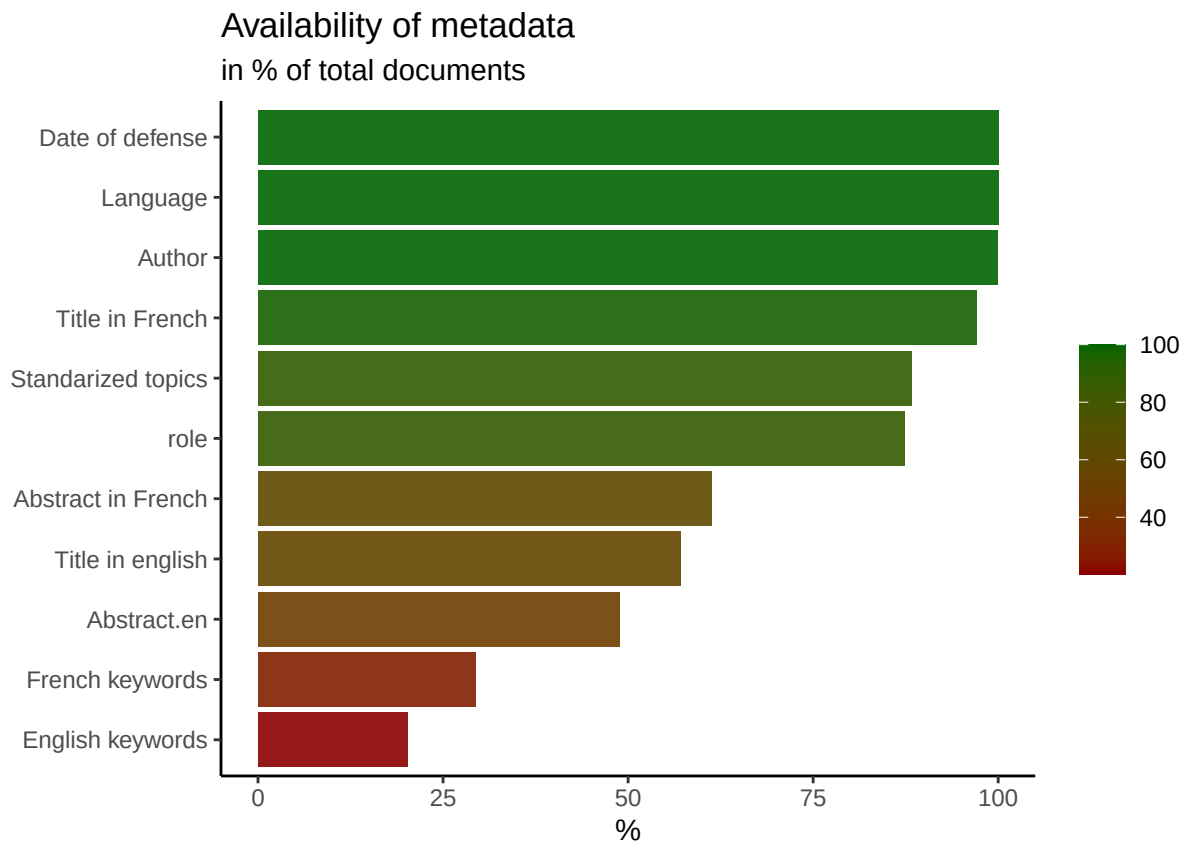
Note: The variable 'Thèse de doctorat' includes the classic PhD and the 'Thèse d'Etat'

Metadata

Global data

```
metadata <- thesis_table %>%
    left_join(people_table %>%
        filter(role == "supervisor") %>%
        select(doc_id, role)) %>%
    rename(`Date of defense` = date, Language = language, Author = first_author,
        `Title in French` = title.fr, `Standarized topics` = topics_standarized,
        `Abstract in French` = abstract.fr, `Title in english` = title.en,
        Abstract.en = abstract.en, `French keywords` = topics_author.fr,
        `English keywords` = topics_author.en)

eval_global <- as.data.frame(colSums(!is.na(metadata %>%
    select(-doc_id, -database, -thesis_type)))/nrow(metadata) *
    100) %>%
    rownames_to_column("variables") %>%
    rename(na = 2)

eval_global %>%
    mutate(variables = fct_reorder(variables, na)) %>%
    ggplot() + geom_col(aes(x = variables, y = na, fill = na,
    position = "identity"), alpha = 0.9) + coord_flip() + scale_fill_gradient(name = "",
```

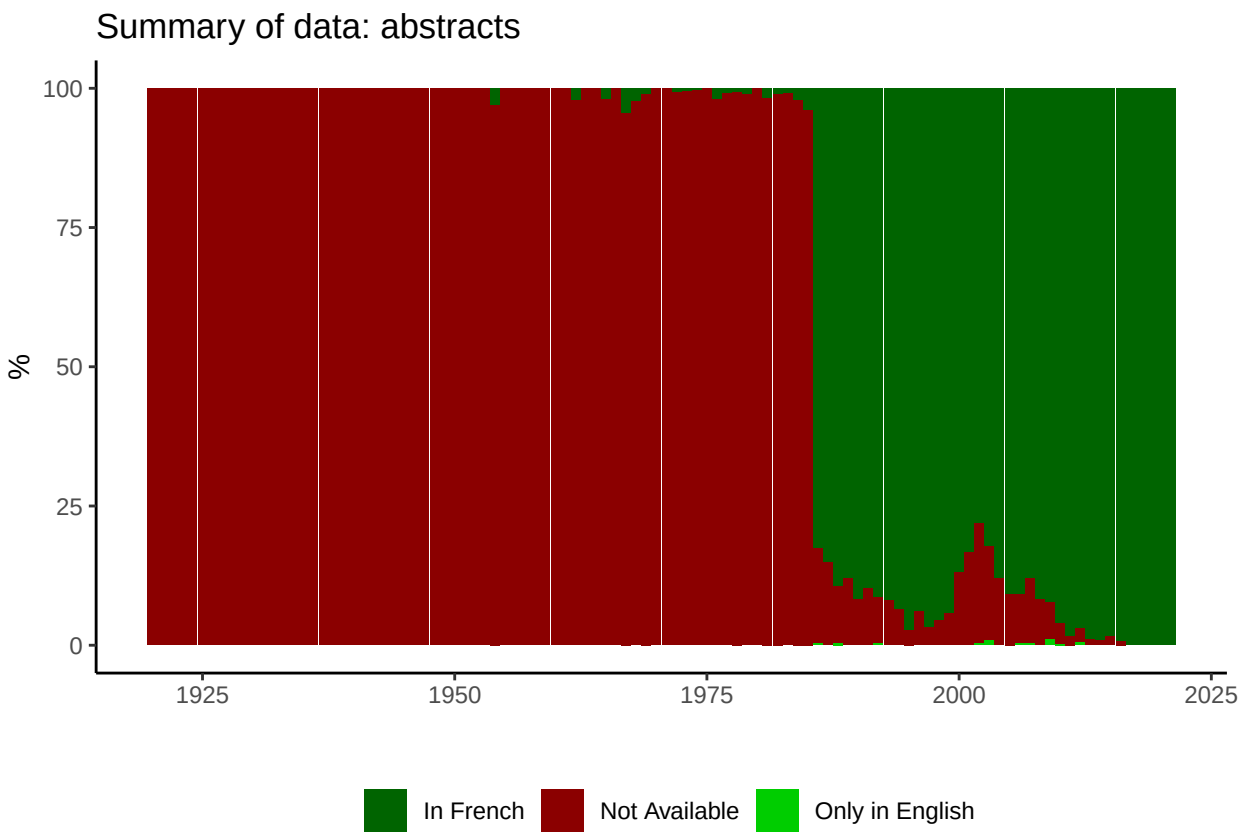
```
    low = "darkred", high = "darkgreen") + labs(x = "", y = "%",
    title = "Availability of metadata", subtitle = "in % of total documents") +
    theme_classic()
```



Textual data

```
abstract_eval <- thesis_table %>%
    ungroup() %>%
    select(date, doc_id, abstract.fr, abstract.en) %>%
    add_count(date, name = "n_thesis") %>%
    mutate(na_ab = case_when(is.na(abstract.fr) & is.na(abstract.en) ~
        "Not Available", is.na(abstract.fr) & !is.na(abstract.en) ~
        "Only in English", !is.na(abstract.fr) ~ "In French")) %>%
    add_count(date) %>%
    add_count(na_ab, date) %>%
    select(date, na_ab, n, nn) %>%
    mutate(nnn = nn/n * 100) %>%
    unique()

abstract_eval %>%
    ggplot() + geom_col(aes(x = as.integer(date), y = nnn, fill = na_ab),
    position = "stack") + scale_fill_manual(name = "", values = c(`Not Available` = "darkred",
```

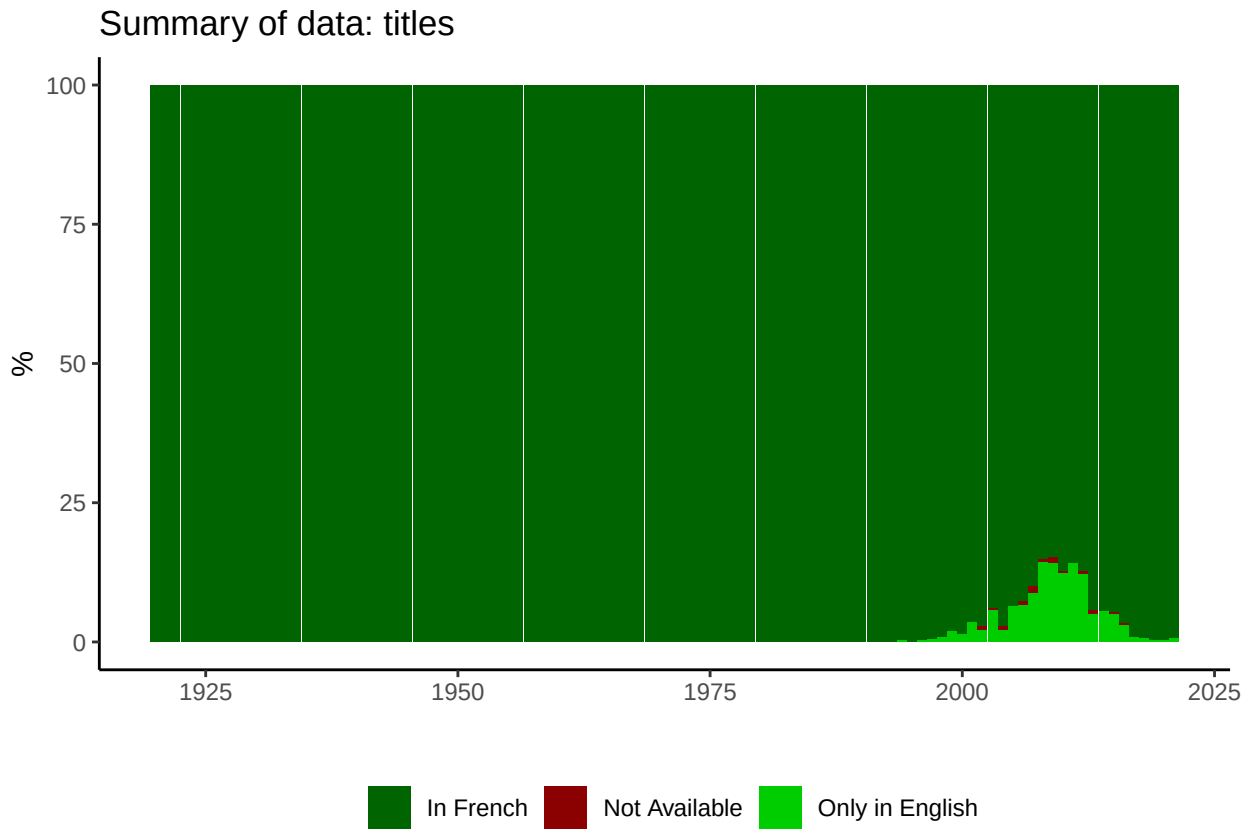
```
    `In French` = "darkgreen", `Only in English` = "green3")) +
    labs(x = "", y = "%", title = "Summary of data: abstracts") +
    theme_classic() + theme(legend.position = "bottom")
```



Abstracts

```
title_eval <- thesis_table %>%
    ungroup() %>%
    select(date, doc_id, title.fr, title.en) %>%
    add_count(date, name = "n_thesis") %>%
    mutate(na_title = case_when(is.na(title.fr) & is.na(title.en) ~
        "Not Available", is.na(title.fr) & !is.na(title.en) ~
        "Only in English", !is.na(title.fr) ~ "In French")) %>%
    add_count(date) %>%
    add_count(na_title, date) %>%
    select(date, na_title, n, nn) %>%
    mutate(nnn = nn/n * 100) %>%
    unique()


title_eval %>%
    ggplot() + geom_col(aes(x = as.integer(date), y = nnn, fill = na_title),
    position = "stack") + scale_fill_manual(name = "", values = c(`Not Available` = "darkred",
    `In French` = "darkgreen", `Only in English` = "green3")) +
```

```
    labs(x = "", y = "%", title = "Summary of data: titles") +
    theme_classic() + theme(legend.position = "bottom")
```



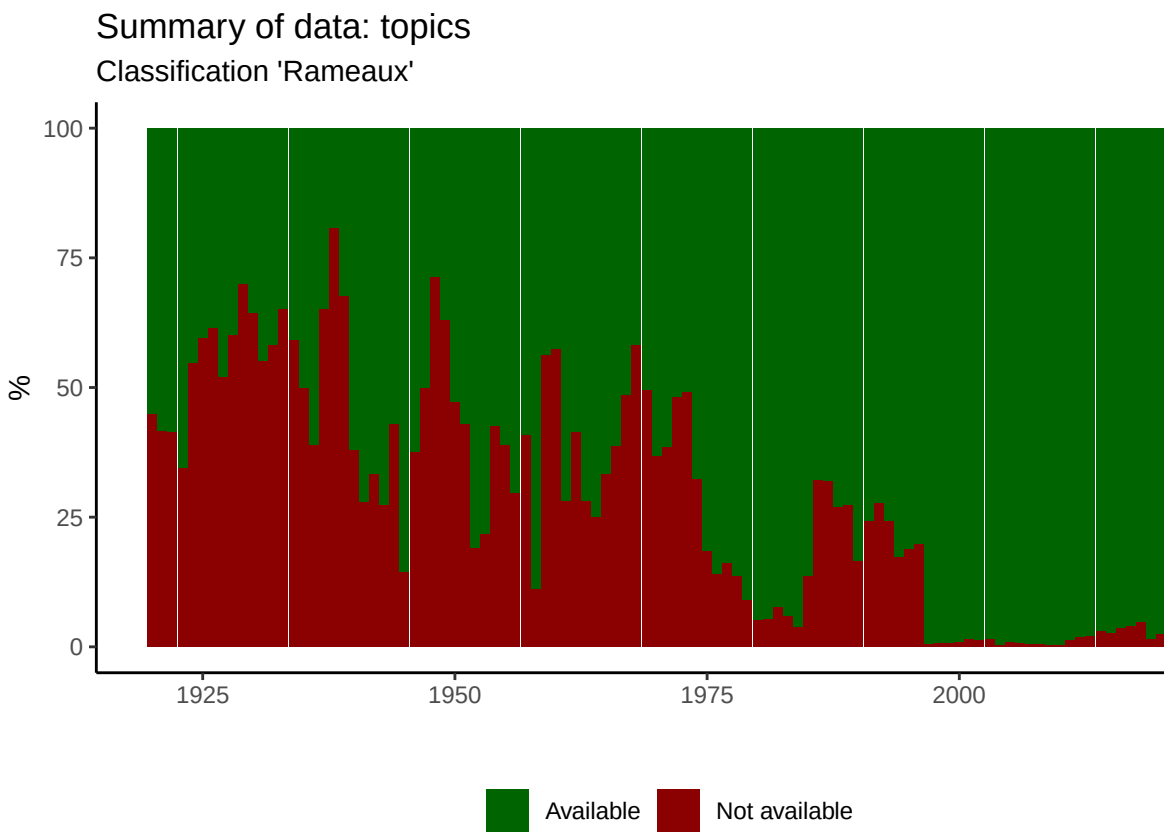
Titles

```
library(knitr)
library(kableExtra)

thesis_table %>%
    filter(date > 1985) %>%
    sample_n(3) %>%
    select(title.fr, topics_standarized) %>%
    kbl() %>%
    kable_minimal() %>%
    kable_styling(full_width = F, latex_options = c("scale_down",
        "HOLD_position"))
```

title.fr	topics_standarized
Energie et économie agro-alimentaire du manioc au Congo : dualisme et possibilités de réduction	[[Manioc – Congo (République)]]Agriculture et énergie – Congo (République)
Modélisation et prospective de la demande de mobilité	Mobilité spatiale – Aspect économique||Transport – Choix des modes||Taxe sur le dioxyde de carbone – Aspect économique – Modèles économétriques
Classe ouvriere et relations corporatives en uruguay- 1930 - 1985 : reproduction de la force de travail, etat et systeme politique.-	NA

```
topic_eval <- thesis_table %>%
    mutate(na_topic = ifelse(is.na(topics_standarized), "Not available",
```

```
        "Available")) %>%
    add_count(date) %>%
    add_count(na_topic, date) %>%
    select(date, na_topic, n, nn) %>%
    mutate(nnn = nn/n * 100) %>%
    unique()


topic_eval %>%
    ggplot() + geom_col(aes(x = as.integer(date), y = nnn, fill = na_topic),
    position = "stack") + scale_fill_manual(name = "", values = c(`Not available` = "darkred",
    Available = "darkgreen")) + labs(x = "", y = "%", title = "Summary of data: topics",
    subtitle = "Classification 'Rameaux'") + theme_classic() +
    theme(legend.position = "bottom")
```

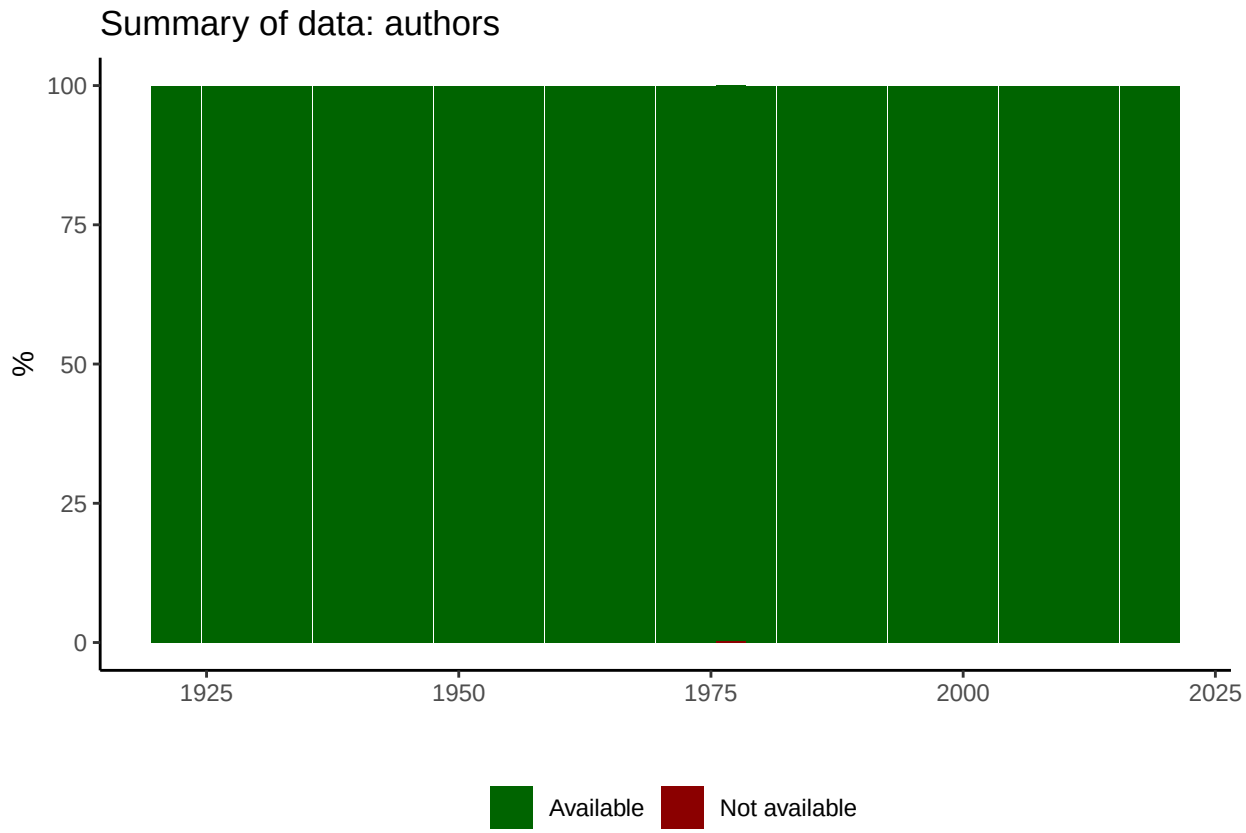


Standarized topics

People

```
eval_author <- thesis_table %>%
    left_join(people_table %>%
        filter(role == "author")) %>%
    mutate(na_author = ifelse(is.na(role), "Not available", "Available")) %>%
    add_count(date) %>%
    add_count(na_author, date) %>%
```

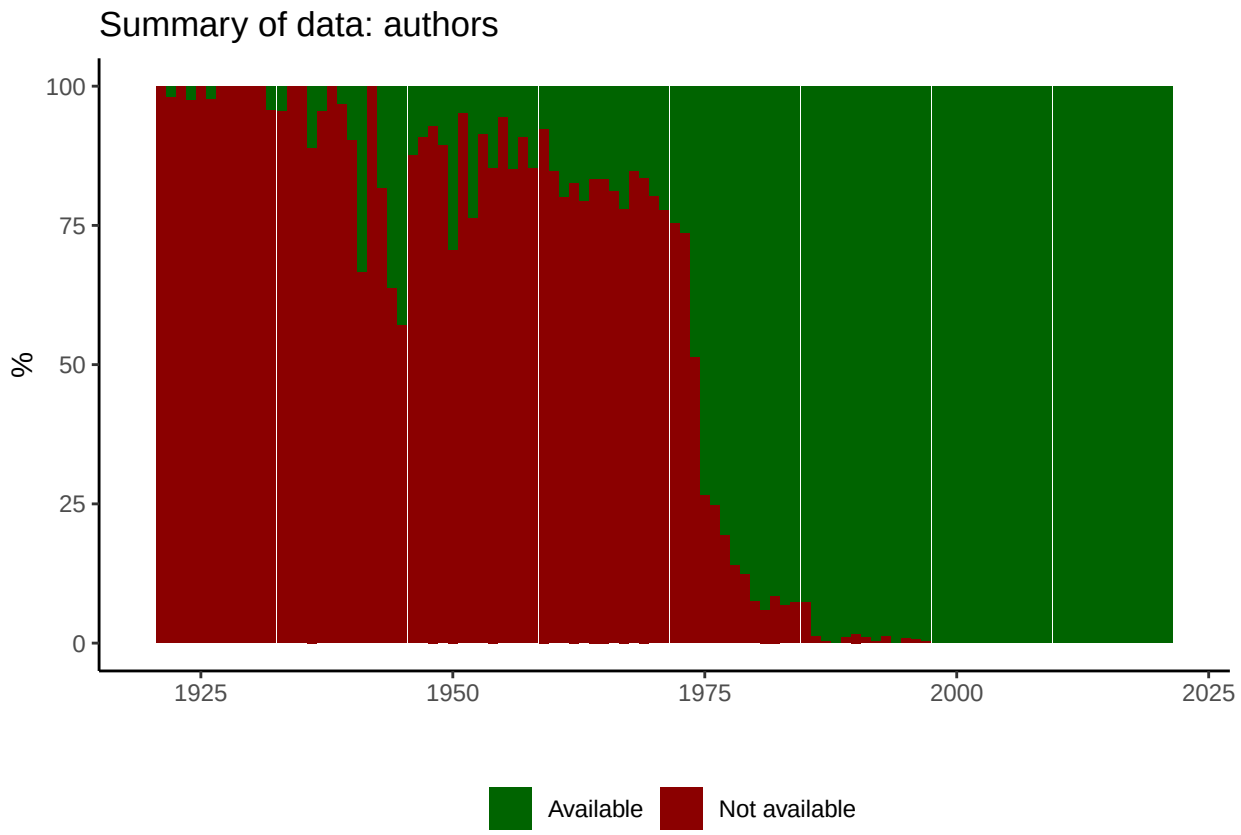
```
    select(date, na_author, n, nn) %>%
    mutate(nnn = nn/n * 100) %>%
    unique()

eval_author %>%
    ggplot() + geom_col(aes(x = as.integer(date), y = nnn, fill = na_author),
    position = "stack") + scale_fill_manual(name = "", values = c(`Not available` = "darkred",
    Available = "darkgreen")) + labs(x = "", y = "%", title = "Summary of data: authors") +
    theme_classic() + theme(legend.position = "bottom")
```



Authors

```
eval_supervisor <- thesis_table %>%
    left_join(people_table %>%
        filter(role == "supervisor")) %>%
    mutate(na_sup = ifelse(is.na(role), "Not available", "Available")) %>%
    add_count(date) %>%
    add_count(na_sup, date) %>%
    select(date, na_sup, n, nn) %>%
    mutate(nnn = nn/n * 100) %>%
    unique()

eval_supervisor %>%
    ggplot() + geom_col(aes(x = as.integer(date), y = nnn, fill = na_sup),
```

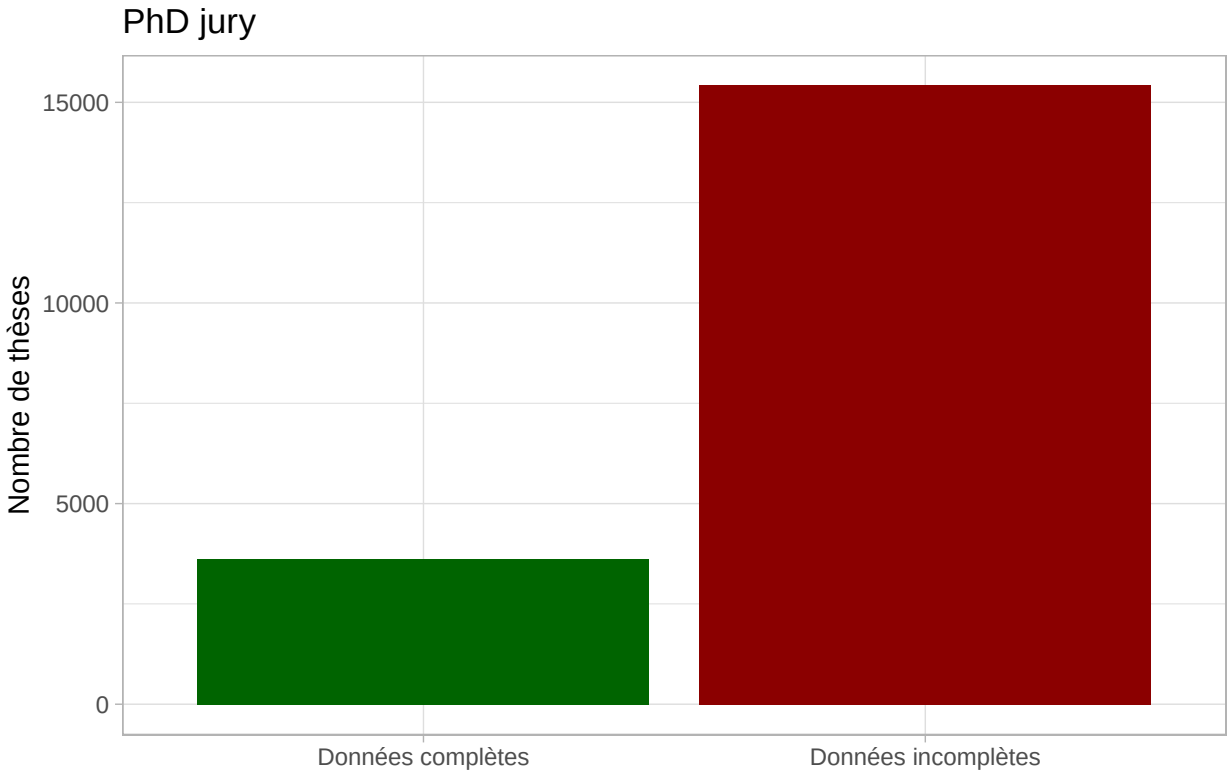
```
    position = "stack") + scale_fill_manual(name = "", values = c(`Not available` = "darkred",
    Available = "darkgreen")) + xlim(c(1920, 2022)) + labs(x = "",
    y = "%", title = "Summary of data: authors") + theme_classic() +
    theme(legend.position = "bottom")
```



Supervisor

```
complete_data <- people_table %>%
    left_join(select(thesis_table, doc_id, date)) %>%
    mutate(role = ifelse(role == "referee", "member", role)) %>%
    mutate(name = paste(prenom, nom)) %>%
    complete(doc_id, role)

check_if_complete <- complete_data %>%
    group_by(doc_id, date) %>%
    summarise(author = sum(role == "author" & !is.na(nom)), member = sum(role ==
        "member" & !is.na(nom)), supervisor = sum(role == "supervisor" &
        !is.na(nom)))

complete <- check_if_complete %>%
    group_by(doc_id, date) %>%
    filter(author > 0 & member > 0 & supervisor > 0) %>%
    count() %>%
    filter(!is.na(date))
```

```
incomplete <- check_if_complete %>%
    group_by(doc_id, date) %>%
    filter(!(author > 0 & member > 0 & supervisor > 0)) %>%
    count() %>%
    filter(!is.na(date))

ggplot() + geom_bar(data = complete, aes(x = "Données complètes",
    y = n), stat = "identity", fill = "darkgreen") + geom_bar(data = incomplete,
    aes(x = "Données incomplètes", y = n), stat = "identity",
    fill = "darkred") + labs(x = "", y = "Nombre de thèses",
    title = "PhD jury", caption = "Note: A complete PhD jury is defined as PhD having at   least one au
    theme_light()
```

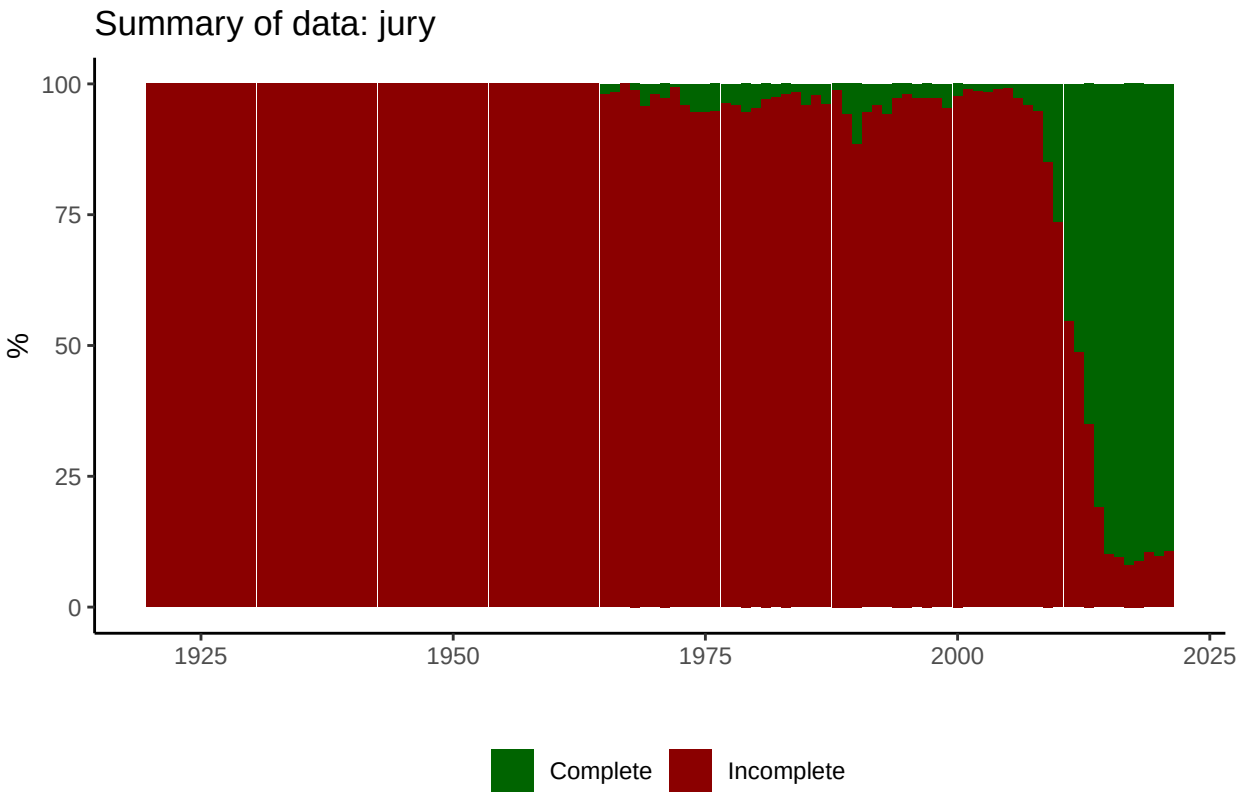


Note: A complete PhD jury is defined as PhD having at least one author, one supervisor and one jury member

Jury

```
complete_data <- people_table %>%
    left_join(select(thesis_table, doc_id, date)) %>%
    mutate(role = ifelse(role == "referee", "member", role)) %>%
    mutate(name = paste(prenom, nom)) %>%
    complete(doc_id, role)

check_if_complete <- complete_data %>%
    group_by(doc_id, date) %>%
    summarise(author = sum(role == "author" & !is.na(nom)), member = sum(role ==
        "member" & !is.na(nom)), supervisor = sum(role == "supervisor" &
        !is.na(nom)))
```

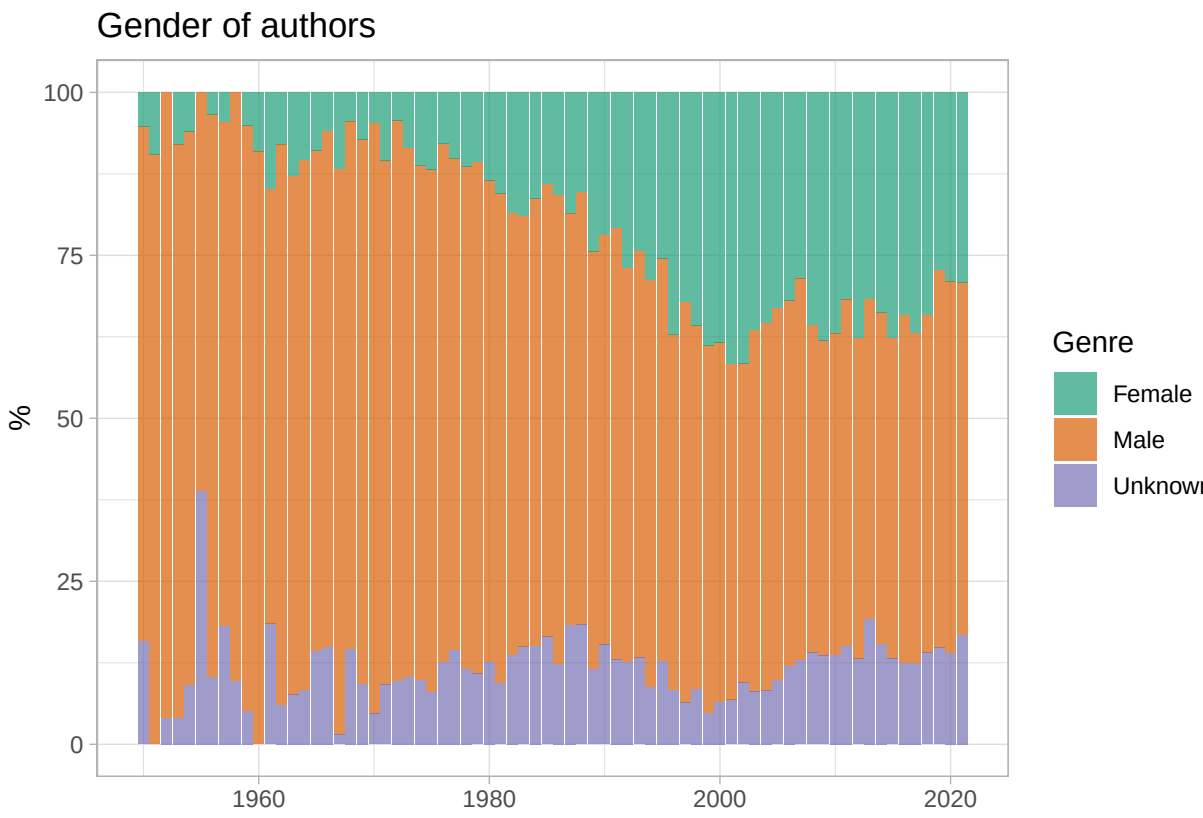
```
complete <- check_if_complete %>%
    group_by(doc_id, date) %>%
    filter(author > 0 & member > 0 & supervisor > 0) %>%
    ungroup() %>%
    count(date) %>%
    filter(!is.na(date)) %>%
    mutate(data = "Complete")


incomplete <- check_if_complete %>%
    group_by(doc_id, date) %>%
    filter(!(author > 0 & member > 0 & supervisor > 0)) %>%
    ungroup() %>%
    count(date) %>%
    filter(!is.na(date)) %>%
    mutate(data = "Incomplete")

eval_jury <- rbind(incomplete, complete) %>%
    group_by(date) %>%
    mutate(total = sum(n)) %>%
    group_by(data) %>%
    mutate(n_perc = n/total * 100) %>%
    group_by(date) %>%
    mutate(control = sum(n_perc))

eval_jury %>%
    ggplot() + geom_col(aes(x = as.integer(date), y = n_perc,
    fill = data), position = "stack") + scale_fill_manual(name = "",
    values = c(Incomplete = "darkred", Complete = "darkgreen")) +
    labs(x = "", y = "%", title = "Summary of data: jury", caption = "Note: A complete PhD jury is defi
    theme_classic() + theme(legend.position = "bottom")
```



Note: A complete PhD jury is defined as PhD having at least one author, one supervisor and one jury member

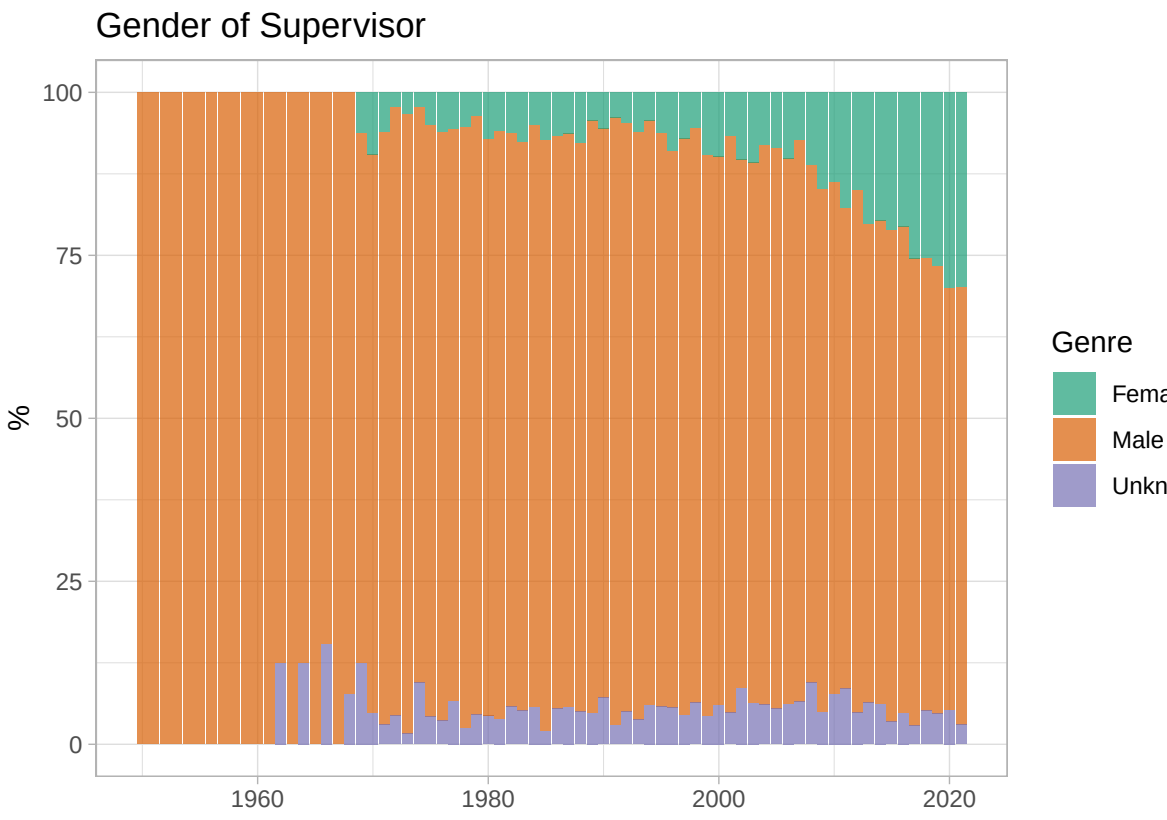
Gender

```
# plot
people_table <- readRDS(here(FR_cleaned_data_path, "people_table.RDS")) %>%
    as.data.table()

author_gender <- people_table[role == "author"]

hist_author_gender <- author_gender[, .N, .(gender_cleaned, date)]
hist_author_gender[, tot := sum(N), date]
hist_author_gender[, share := N/tot * 100]

ggplot(data = hist_author_gender[date >= 1950], aes(x = as.numeric(date),
    y = share, fill = gender_cleaned)) + geom_bar(stat = "identity",
    alpha = 0.7) + scale_fill_brewer(name = "Genre", palette = "Dark2") +
    theme_light() + labs(x = "", y = "%", title = "Gender of authors")
```



Authors gender

```
supervisor_gender <- people_table[role == "supervisor"]
hist_supervisor_gender <- supervisor_gender[, .N, .(gender_cleaned,
    date)]
hist_supervisor_gender[, tot := sum(N), date]
hist_supervisor_gender[, share := N/tot * 100]

ggplot(data = hist_supervisor_gender[date >= 1950], aes(x = as.numeric(date),
    y = share, fill = gender_cleaned)) + geom_bar(stat = "identity",
    alpha = 0.7) + scale_fill_brewer(name = "Genre", palette = "Dark2") +
    theme_light() + labs(x = "", y = "%", title = "Gender of Supervisor")
```

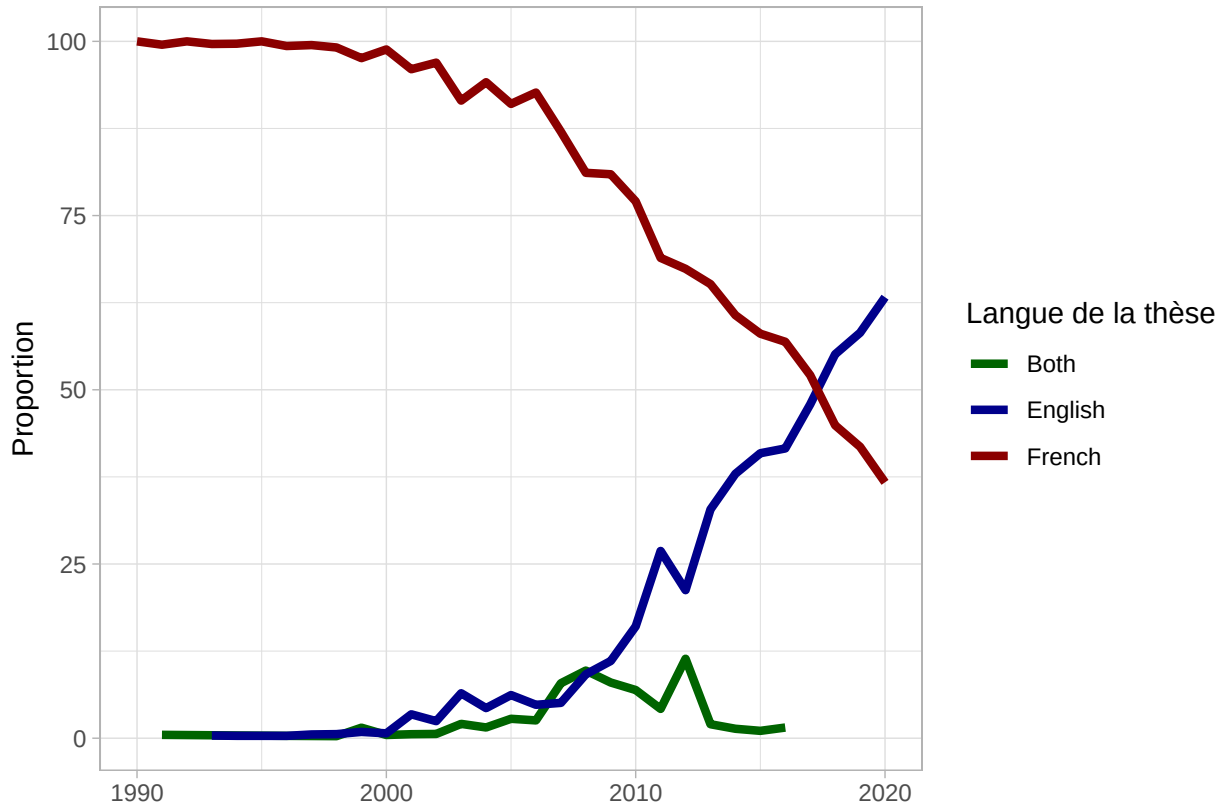


Supervisor gender

An application: the internationalization of French economics through PhD data

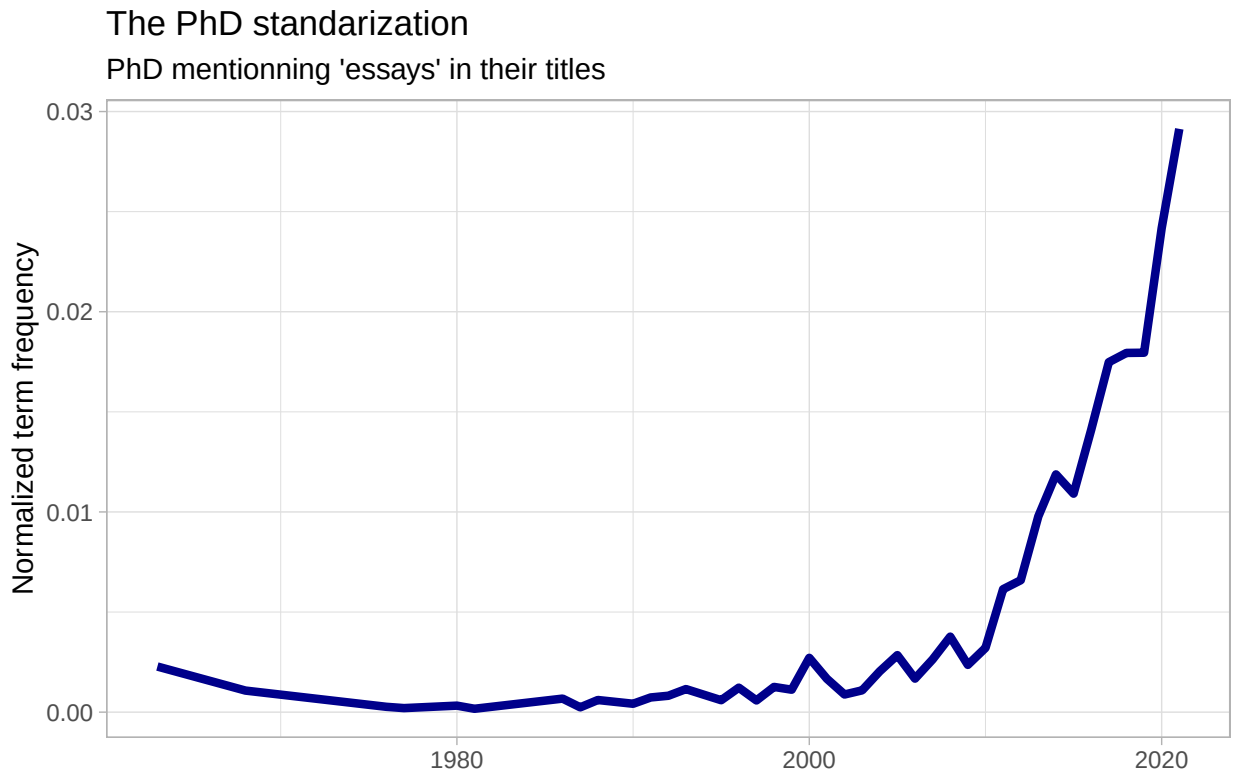
Language of PhD

```
distribution_languages <- thesis_table %>%
    filter(str_detect(language, "^fr$|^en")) %>%
    filter(date %in% c(1990:2020)) %>%
    count(language, date) %>%
    group_by(date) %>%
    mutate(n_perc = n/sum(n) * 100) %>%
    mutate(language = case_when(language == "fr" ~ "French",
        language == "en" ~ "English", language == "enfr" ~ "Both"))

distribution_languages %>%
    ggplot() + geom_line(aes(x = as.numeric(date), y = n_perc,
    colour = language), linewidth = 1.5) + scale_color_manual(name = "Langue de la thèse",
    values = c(French = "darkred", English = "darkblue", Both = "darkgreen")) +
    ylab("Proportion") + xlab("") + theme_light()
```

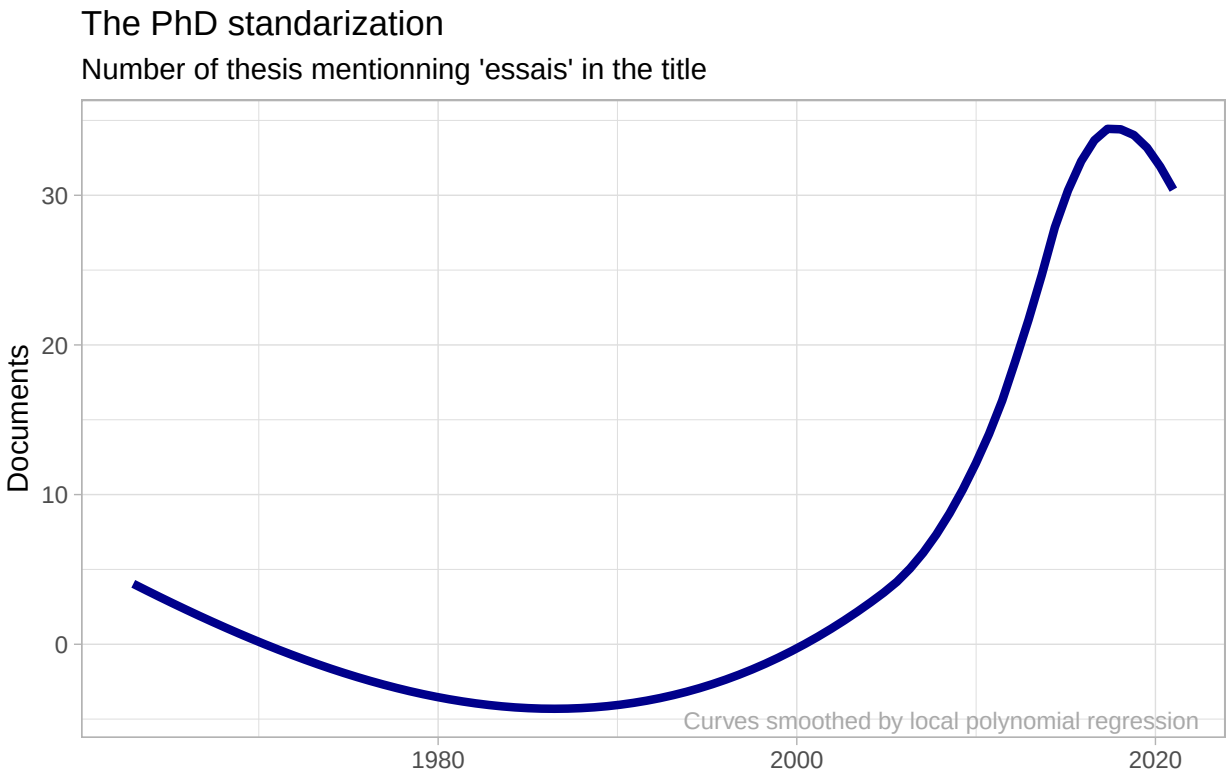


Standarization of PhD format

```
format_thesis <- thesis_table %>%
    select(title.fr, date) %>%
    group_by(date) %>%
    unnest_tokens(input = title.fr, output = token) %>%
    add_count(date) %>%
    filter(str_detect(token, "essais")) %>%
    add_count(date)

format_thesis %>%
    ggplot() + geom_line(aes(x = as.numeric(date), y = nn/n),
    linewidth = 1.5, colour = "darkblue") + labs(x = "", y = "Normalized term frequency",
    title = "The PhD standarization", subtitle = "PhD mentionning 'essays' in their titles",
    caption = "Note: Frequency normalized by the size of the corpus") +
    theme_light()
```




```
thesis_table %>%
    select(title.fr, date) %>%
    filter(str_detect(title.fr, "essais")) %>%
    add_count(date) %>%
    ggplot() + geom_smooth(aes(x = as.numeric(date), y = n),
    method = "loess", se = F, linewidth = 1.5, colour = "darkblue") +
    labs(x = "", y = "Documents", title = "The PhD standarization",
        subtitle = "Number of thesis mentionning 'essais' in the title",
        caption = "Note: Frequency normalized by the size of the corpus") +
    theme_light() + ggplot2::annotate("text", x = max(as.numeric(thesis_table$date)),
    y = -Inf, hjust = 0.95, vjust = -0.5, label = "Curves smoothed by local polynomial regression",
    size = 3, color = "darkgrey")
```



Note: Frequency normalized by the size of the corpus