Le RGPD facile avec R !

Comment devenir le meilleur ami de votre DPO

Christophe REGOUBY
R user Group Toulouse

AIRBUS

Disclaimer:

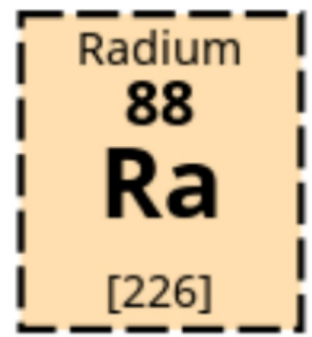
I am not representing my employer AIRBUS in this talk

I cannot confirm nor deny if AIRBUS is using any of the methods, tools, results etc. mentioned in this talk

RGPD : quelques rappels

- April 18 : General Data Protection Regulation (GDPR) Law compliance with
  - Conditions applicables au consentement (Art 7.)*
  - Droit à l'effacement («droit à l'oubli») (Art 17.)*
  - Protection des données dès la conception et protection des données par défaut (Art 25.)
  - Sécurité du traitement (Art 32.)

  * amendes administratives jusqu'à 20M€ , ou 4% du chiffre d'affaires annuel mondial total (Article 83.4)
- Jan 19 : Communication à la personne concernée d'une violation de données à caractère personnel *(Art 34.)*



Source : wikimedia.org

RGPD : La vue du Data-analyste

	Last update	PM	Domain	Functional area	Project short name	Project full name	Comments	Perimeter	Freeze Impact Category 1	ABD scope Category 2	2017 Project Category 3	Bundle for projects during transition Category 4	Handover skills and workload on CAP side Category 5	Type	Phase	Next milestone date	Actual: Global KF	Schedule	Scope KF	Quality KF	Custome	Budget KF
49	21/10/2016		Search	HR	Search Platform Hub Improvement	Search Platform Hub Improvement	1st phase: audit of the HUB search => delivered to Business and Portal team, quick wins are scheduled for implementation. 2nd phase: with more elaborate improvement actions --> depends on budget and decision of Portal team; will be 2017 if positive	IMCB	N	N	N	N/A		Study	Closed	Closed						
50	21/10/2016		Search	HR	Search Platform Log Analytics	Search Platform Log Analytics	Ongoing, showing good results, and a good collaboration with the datalab team!!	IMCB	N	N	Y	N/A		POC	Closed	Closed						
51	16/12/2016		Search	HR	BSVA Open Wiki	Add the Landing Gear open Wiki as source to BS v3	change was needed to capture the whole content, costing an additional 2KEUR. Budget approved by F. Bouix (HKACK); else it's on track, with M10/M10a passed on Dec 14	IMCB	N	N	Y	N/A		Project	M5-M10	19/01/2017						
52	21/10/2016		Search	HR	Search Platform upgrade	Search Platform upgrade	Upgrade Search engine to be able to provide new functionalities Discussion on developer license with IBM => topic closed (13/04/2016) No visibility on roadmap, will be established with POC result. Issue on budget, project not funded by business yet. => postponed to 2017 C231, AT60, 1D10, AW10, DV70, and dependencies with I770, 1J91, I640	IMCB	N	N	Y	N/A		Project	Future	Future						
53	21/10/2016		Search	HR	Search Platform upgrade POC	Search Platform upgrade POC	POC is budgeted and has a roadmap, shall be finished week 28 (shift). Result will determine the complete Search Roadmap for 2016. Delay linked to licence issue => global status is amber, situation has improved, no more blocking point. POC was delivered, we now know the effort and cost to upgrade	IMCB	N	N	N	N/A		POC	Closed	Closed						
54	18/11/2016		Search	Platform governance	Search Platform Big Data POC	Search Platform Big Data POC	Scope study No budget form, was LucidWorks marketing POC Complete => results are ok	IMCB	N	N	N	N/A										
55	21/10/2016		Search	Quality	Search Platform Zamiz	Search Platform Zamiz	Add Airbus Archive as a data source to Search Platform. Budget recovery in progress, not funded yet Not started yet, as no budget transferred yet => Cancelled by Martin as no answer from	IMCB	N	N	N	N/A										
56	25/11/2016		BI	Others	MSN5 - AP1	MSN5 - AP1 (MM)	AP1 = Spares in Germany Pre-M5 meeting. Costs share. Waiting for cost confirmation from Business to pass M5. Budget around 22 ~24 kj, it may be ok. M9 passed w/o amendment 06/10/2016 (Meeting carried out by Gerd) Risk on planning (too tight). BW impacts analysis done (PBA & PBI) by Sopra, not Funded. DFE: M10/M10a passed. M11 postponed due to my unavailability. COP/MIP prepared (06/11/16) DFE: M11 passed/ MIP: 06.11.2016 DFE: MIP successfully done	IMCB	Y	N	N	Sopra										
57	18/11/2016		BI	Others	MSN5 - ARP	MSN5 - ARP (MM, PP, QM)	Planning: M5: 04.07.16 M7: 22.07.16 M9: 15/10/16 M10: 01/12/16 MIP: 21./22.01.2017 After MIP: 22.01. – 26.01. (loading activities on PBI & PBA, + Webfocus) M13: 02.02.2017 (due regulations for hand over to AS) M5 still not done, some activity already started to keep the team. Budget is not officially confirmed by Séverine BRUNET. => M7 may be postponed due to global ARP project. Activity started without any funding for IMCB & Sopra & Steria. DFE: M9 done, Funding OK so far but risk of two CR to be funded.	IMCB	N	Y	Y	Sopra										
58	18/11/2016		BI	Others	MSN5 - P11	MSN5 - P11 (MM,	MIP done. Documentation is ok. Closed	IMCB	N	N	N	Closed										
							Changes might be expected. To be monitored (old systems, documentation missing.... => complex) Documentation is ok. RACI not clear between source system owner & IMCB, list of impacted data provided by source system owner may be not exhaustive														

Dataset preparation/wrangling: Privacy by design

```
# Plan it asynchroneously with Future
metadata_f <- future_map(file_lst,metadata_reader,column="file_metadata")
# Turn in into data.frame and reduce names by 15 characters
metad_df  <- metadata_f %>%
  data.table::rbindlist(fill=TRUE)  %>%
  as_tibble() %>%
  # Privacy by design
  select(-matches("[Aa]uthor|[Cc]reator")) %>%
  select(-matches("X-Parsed-By|X-TIKA:content_handler"),-file_metadata, -starts_with("file_metadata.X-TIKA:orig")) %>
  rename_all(str_sub,start=15L)
```

Dataset preparation/wrangling: Privacy by design

```
findings_raw <- read_csv2(here::here("data/raw/0_ADAPDS_clean.csv"),
                          col_types ="???c?????c?????ccc?c?????????c?????????c?????ccc?c?????????
                          c???????ddd?????????c?????????c?????????c?????????c??", n_max = 100000)

findings <- findings_raw  %>%
  # Privacy by design : remove PII
  select(-ends_with("BY"),-DESIGNSTATUS,-LOCKINGRESPONSIBLE) %>%
  mutate_at(c("ADAPDSNUMBER","ATA","SUBATA"), as.factor) %>%
  # filter business out-of-interest entries
  filter(ADAPDSISSUEINDEX=="A00")  %>%
  # remove all is.na columns : 25 columns removed
  select_if(~mean(is.na(.))<1)
```

Datasets considered - challenges

One Non-Conformity example

** 11.02.2009 10:23:06 **Ernst MULLER** (MULLER) * 1 x Halter Item 114 montieren (B/12) * * HTZ ABS0785B14C * BU L534-66895-000-00 * View 302 * * Folio : LH 141 / FA / 1041 / 5.2 / 01 * 05.03.2009 14:15:35 **John DOE** (DOE) Tel. **040/645 48888** * * EADS- AUG discrepancies are reworked by production and inspected by Mr. **Johnson** W1017. No further work necessary. * EADS-AUG Diskrepanzen wurden durch Produktion nachgearbeitet und geprüft durch Hr. **Johnson B**.-1017. Keine weiteren Arbeiten erforderlich. **

*Note: people names and phone number have been changed

- Mixed of “structured” information and free text
- Mixed of several language:
  - ✓ German names within English text
  - ✓ NC starts in german and then continue in english
- Engineering English : lots of Acronyms

Datasets considered - challenges

One QSR example

Technical memo to be raised referencing positive test specimen results, class of parts etc to justify low risk rating of the parts in question **A Doe** 2 XXX to sign tech memo ref action 1**P Dupont** 3 Investigation to take place as to how to cover the non-conformity in parts already delivered to Airbus or sub-tier ? blanket concession? **A Doe,C Dubois**,**B Smith**

- Diversity in language level
  - ✓From polite business-to-business exchange
  - ✓Via minutes-of-meeting like language
  - ✓To Business-insiders specific English

Une solution : L’anonymisation

Existe-t-il un modèle de **machine-learning model** qui permet d’identifier les PII dans des datasets **variés de language-naturel** sans avoir a créer des rêgles métier specifiques à chaque dataset ?

Result on NC example :

** 11.02.2009 10:23:06 #PERSON (#UserID) * 1 x Halter Item 114 montieren (B/12) * * HTZ ABS0785B14C * BU L534-66895-000-00 * View 302 * * Folio : LH 141 / FA / 1041 / 5.2 / 01 * 05.03.2009 14:15:35 #PERSON (#UserID) Tel. #PHONE * * EADS- AUG discrepancies are reworked by production and inspected by Mr. #PERSON W1017. No further work necessary. * EADS-AUG Diskrepanzen wurden durch Produktion nachgearbeitet und geprüft durch Hr.#PERSON .-1017. Keine weiteren Arbeiten erforderlich. **

Result on QSR example :

Technical memo to be raised referencing positive test specimen results, class of parts etc to justify low risk rating of the parts in question #PERSON 2 XXX to sign tech memo ref action 1 #PERSON 3 Investigation to take place as to how to cover the non-conformity in parts already delivered to Airbus or sub-tier ? blanket concession? #PERSON,#PERSON,#PERSON

Les modèles de NER publics :

- library(cleannlp)
- cnlp_init_corenlp(lang="fr")
  : Stanford CoreNLP
- cnlp_init_spacy(model_name = "fr")
  : spacy models

Entrainés sur la tâche « CoNLL2003 » avec les entités

- LOC
- MISC
- ORG
- PER

English data	LOC	MISC	ORG	PER
Training set	7140	3438	6321	6600
Development set	1837	922	1341	1842
Test set	1668	702	1661	1617

German data	LOC	MISC	ORG	PER
Training set	4363	2288	2427	2773
Development set	1181	1010	1241	1401
Test set	1035	670	773	1195

Sur un dataset de News de Reuters

Train : Août-1996, Test : Déc-1996

Source: Introduction to CoNLL2003 Shared Task
https://arxiv.org/pdf/cs/0306050.pdf

Sur le dataset OntoNotes Release 5.0 (2007-2011, 2,9M mots)

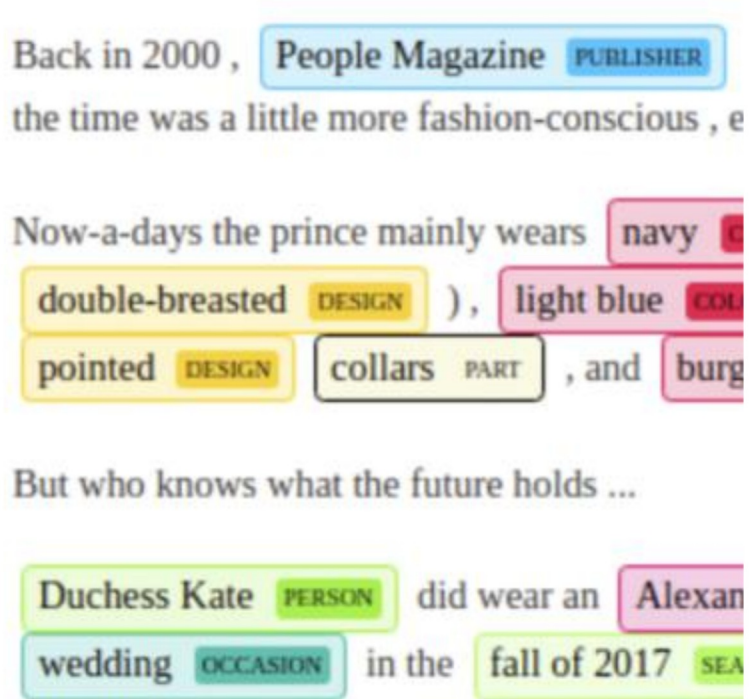
	Arabic	English	Chinese
News	300k	625k	250k
BN	n/a	200k	250k
BC	n/a	200k	150k
Web	n/a	300k	150k
Tele	n/a	120k	100k
Pivot	n/a	n/a	300

Source: LDC OntoNotes Release 5.0
https://catalog.ldc.upenn.edu/LDC2013T19

Qu’est-ce qu’on cherche : la taxonomie des PII « Personal Identifiable Information »

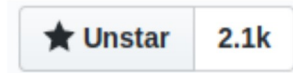
- Name (first name, last name, fullname)
- User ID (windows Id, SAP id..)
- Telephone number
- Organization
- Email address

On va choisir un dataset <u>représentatif</u> à annoter dans
- un training dataset
- un testing dataset



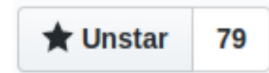
Il faut ré-entrainer ! : Step 1: Choisir l'outil d'annotation

Doccano : https://github.com/chakki-works/doccano




```
$ docker pull chakkiworks/doccano
$ docker run -d --rm --name doccano \
   -e "ADMIN_USERNAME=admin" \
   -e "ADMIN_EMAIL=admin@example.com" \
   -e "ADMIN_PASSWORD=password" \
   -p 8000:8000 chakkiworks/doccano
```

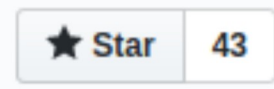
Dataturks: https://github.com/DataTurks/DataTurks



Installation : https://medium.com/@dataturks/dataturks-on-prem-a-fully-self-hosted-data-annotation-solution-86b455bf0634

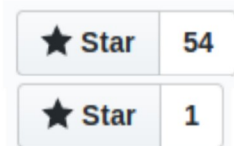
```
$ docker pull klimentij/dataturks:latest
$ docker run -d --rm --name dataturks \
   -p 80:80 klimentij/dataturks
```

BNOSAC CRFSuite: https://github.com/bnosac/crfsuite




```
rmarkdown::run(file = system.file(package = "crfsuite", "app", "annotation.Rmd"))
```

RQDA: https://github.com/Ronggui/RQDA
https://github.com/FrdVnW/dockerqda



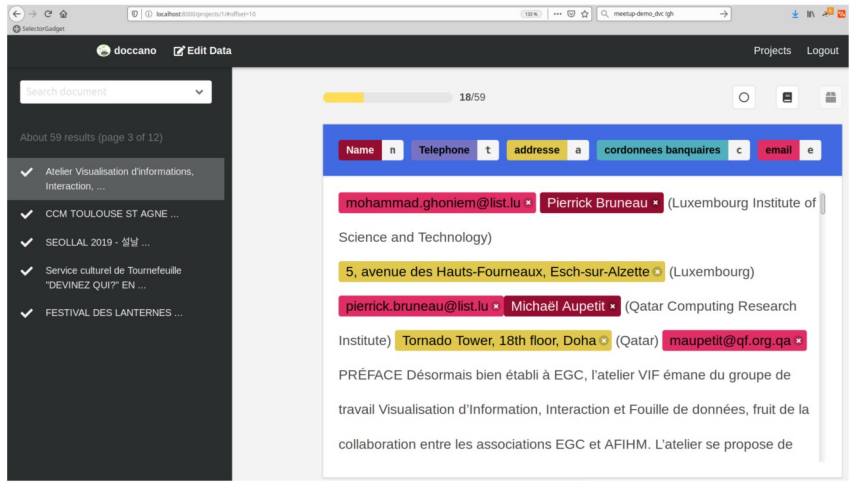

```
$ docker pull frdvnw/dockerqda
$ XSOCK=/tmp/.X11-unix
$ XAUTH=/tmp/.docker.xauth
$ xauth nlist $DISPLAY | sed -e 's/^..../ffff/' | $ xauth -f $XAUTH nmerge -
$ sudo docker run -it --volume=$XSOCK:$XSOCK:rw \
     --volume=$XAUTH:$XAUTH:rw \
     --env="XAUTHORITY=${XAUTH}" \
     --env="DISPLAY" \
     --name whirl_wheels \
     --workdir=/root/ \
     --volume=/WHERE/YOU/WANT/IN/YOUR/COMPUTER/dockerqda/:/home/dockerqda/ frdvnw/dockerqda:latest
```

Il faut ré-entrainer : Step 2: Importer le texte

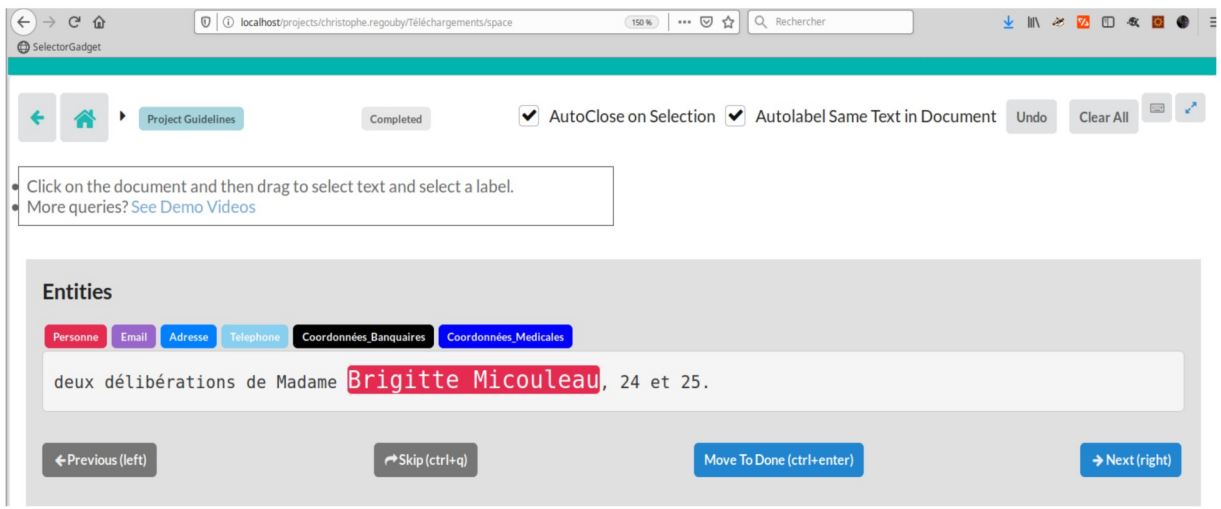
Corpus format	Specific	TSV	json	Plain text
Doccano :		ConLL	JSONL, resume annotation	1 doc per line
Dataturks:		X	Pre-annotated, resume annotation	
BNOSAC crfsuite	RDS			
RQDA:	R data ?			

Il faut ré-entrainer : Step 2: J’annote !

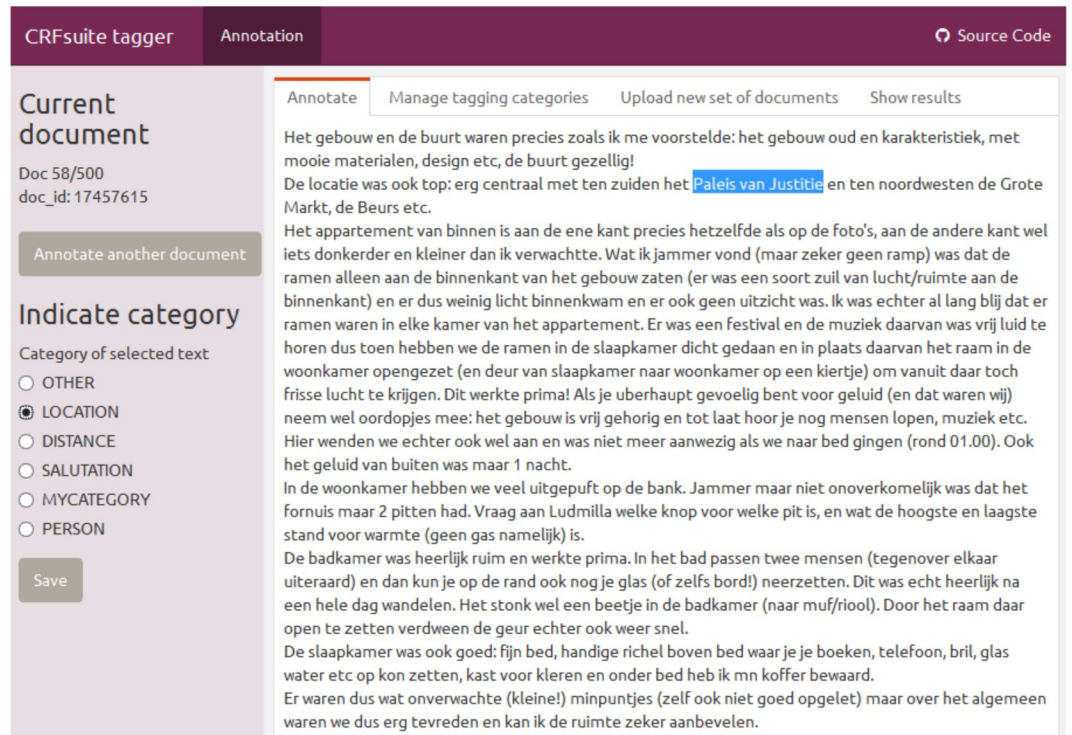
Doccano :



Dataturks:



CRFSuite :



Il faut ré-entrainer ! Step 3: Exporter les annotations


```
{"id": 19, "text": "BILLETERIE: CONCERT DU 13\n                                    AVRIL 2019 - GRANDE HALLE DE\n                                                                    L'UNION\n
                        samedi 13 avril 2019 21h00       Grande Halle\n
Rue du Somport\n                                                                  31240 L'Union\n
                      France\n                                                                  Christophe R\n
 DATE DE LA COMMANDE              10/04/2019\n                          N° DE LA TRANSACTION                    2342896  TARIF ADULTE\n
                                                Qté 1 - Tarif : 10€\n                          N° DE LA VENTE                          2591090\n
                                                                TARIF TOTAL : 10€\n                                DES QUESTIONS ?\n
            harmonie.h2o@gmail.com\n                                        +33662740728\n                          https://orchestreh2o.assoconnect.com/
billette\n                                  rie/offre/100990-s-billeterie-concert-\n                                du-13-avril-2019-grande-halle-de-l-union\n
                                                                VOUS FAITES PARTIE D'UNE ASSOCIATION ?\n
                      Découvrez AssoConnect, le logiciel des associations : site\n                          N° TICKET : 2591090GIFII
  internet, membres, gestion des adhérents & des dons,\n                                                                comptabilité, emailing.\n
                                                                En savoir plus sur www.assoconnect.com\nPowered by TCPDF (www.tcpdf.org)", "meta": {},
"annotation_approver": null, "labels": [[212, 419, "addresse"], [592, 604, "Name"], [1148, 1170, "email"], [1205, 1218, "Telephone"]]}
```


```
Pappus  Personne
;       O
mathématicien   O
grecque O
qui     O
vécut   O
au      O
IVe     O
Pappus  Personne
;       O
mathématicien   O
grecque O
qui     O
vécut   O
au      O
IVe     O

SCORRAILLE      Personne
,       O
Bertrand        Personne
```

	Corpus Format	TSV	json	Spacy json
Doccano :	Id Text		text-label	Yes (jsonl)
Dataturks:	Sentences	X	Complete	Via scipt.py
RQDA	Text files			
crfsuite	RDS			


```
{"content": "Pappus ; mathématicien grecque qui vécut au IVe                                    Pappus ; mathématicien grecque qui vécut au IVe",
    "annotation":[{"label":["Personne"],"points":[{"start":81,"end":86,"text":"Pappus"}]},{"label":["Personne"],"points":[{"start":0,"end":5,"text":"Pappus"}]}],
    "extras":null,
    "metadata":{"first_done_at":1574634784000,"last_updated_at":1574634784000,"sec_taken":0,"last_updated_by":"christophe.regouby@free.fr","status":"done","evaluation"
{"content": "SCORRAILLE, Bertrand SERP, Laurent LESGOURGUES, Evelyne NGBANDA OTTO, Samir HAJIJE,","annotation":[{"label":["Personne"],"points":[{"start":70,"end":81,"t
{"content": "° °","annotation":null,"extras":null,"metadata":{"first_done_at":1574634805000,"last_updated_at":1574634805000,"sec_taken":0,"last_updated_by":"christophe
{"content": "tradition. En ce qui me concerne, je vais être très court et j'invite d'ailleurs, ils feront ce qu'ils voudront bien sûr,","annotation":null,"extras":null
{"content": "territoriales, pour le moment les métropoles et dans pas longtemps les régions, c'est parce que ce sont les deux","annotation":null,"extras":null,"metadat
```

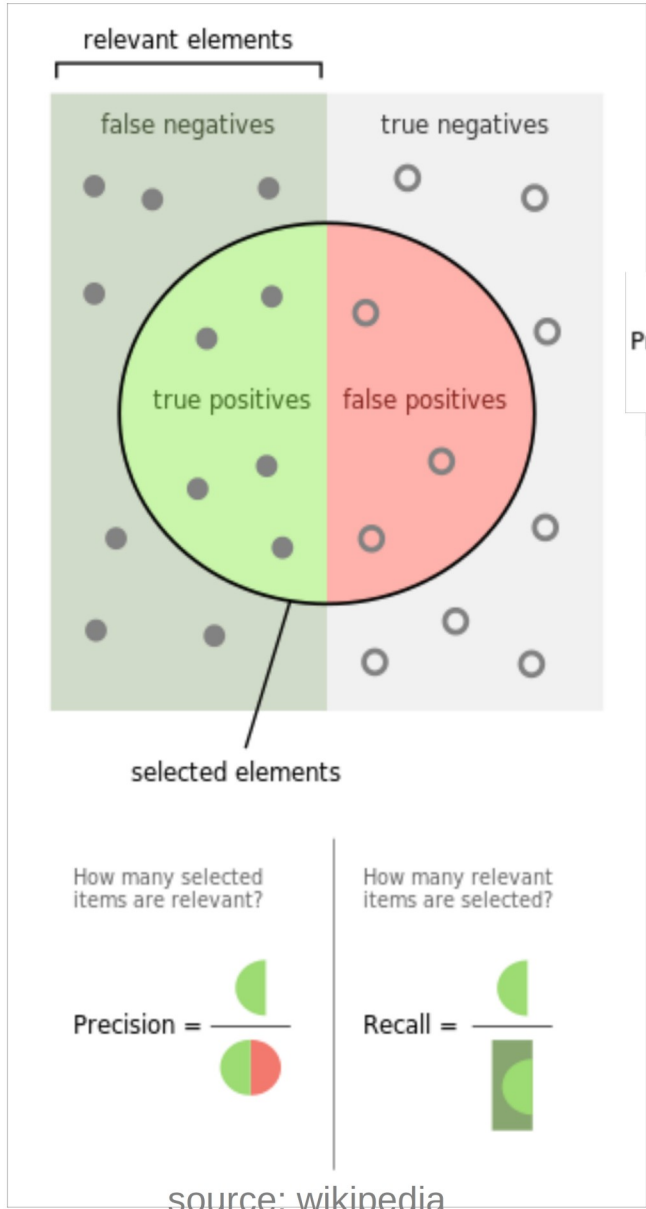
Il faut ré-entrainer !
Step 5 : Les Métriques d'évaluation

- Precision / Recall

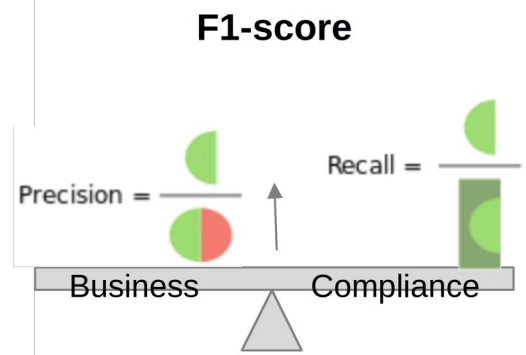
- In the context of PII we favor recall over precision and use **F2-score** (we give recall twice as much weight):

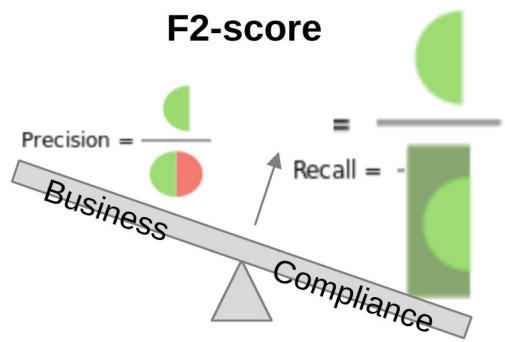
$$\frac{5 \times \text{precision} \times \text{recall}}{(4 \times \text{precision} + \text{recall})}$$

- We prefer to identify most PII even if this means removing other non PII tokens (as long as we can still use it for exploitation after)



source: wikipedia





Il faut ré-entrainer ! Step 3: Exporter les annotations

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
reticulate::use_condaenv("spacy")
library(cleanNLP)
# cnlp_download_spacy("fr-core-news-sm")
cnlp_init_spacy(model_name = "fr")
library(jsonlite)
library(tif) # from devtools::install_github("ropensci/tif")
library(fuzzyjoin) # requires BiocManager::install("Iranges") for interval_inner_join
```

Il faut ré-entrainer ! Step 3: Exporter les annotations

```
## Lecture du fichier de sortie des annotations

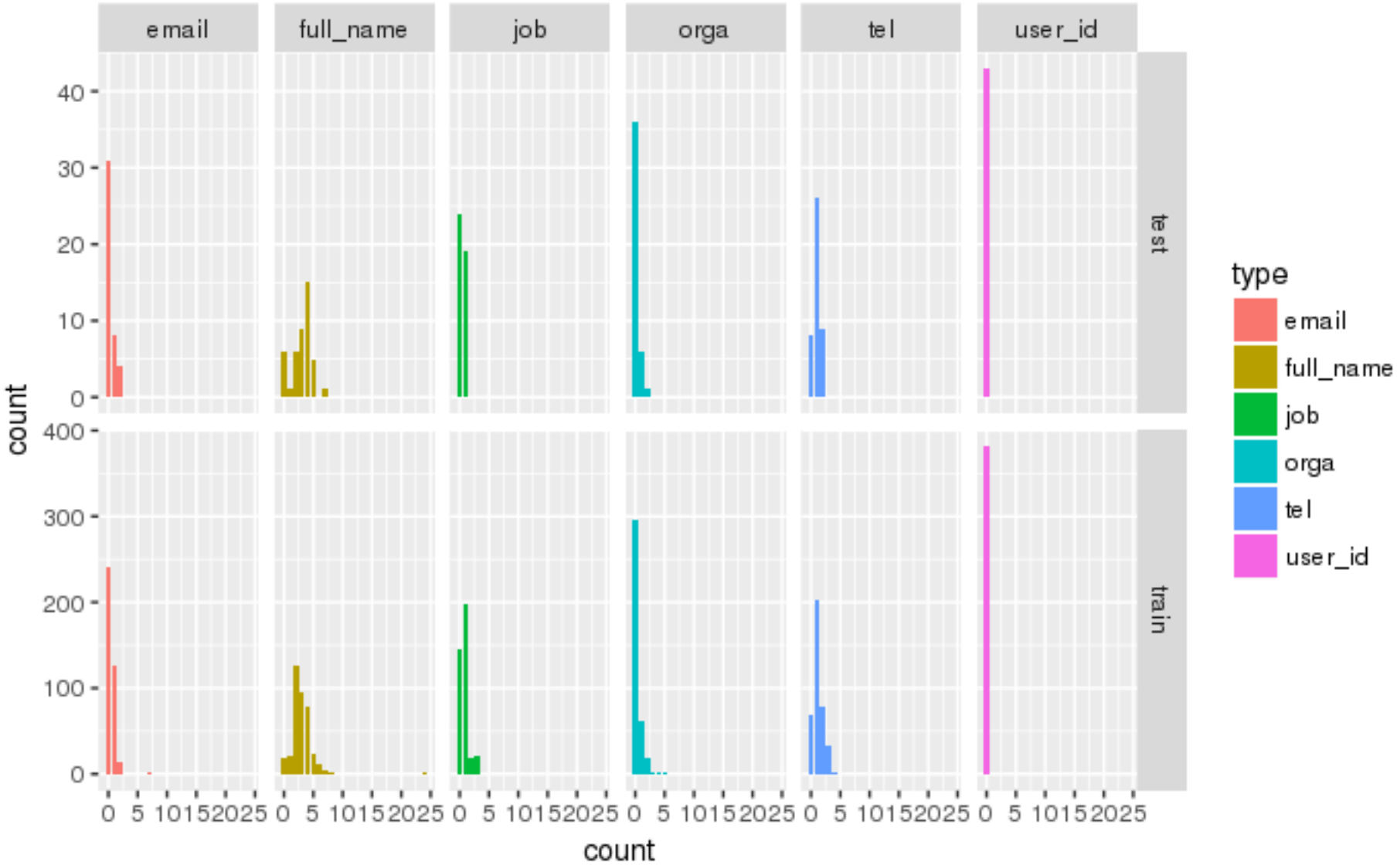
```{r read annotation}
jslite_annot <-jsonlite::stream_in(file(here::here("data/doccano_export_text_label.json")),verbose = T) %>%
  mutate(string_length = str_length(text))
```

extraction des entités annotées
```{r extract entities}
# DONOT USE map_dfr(as_tibble,.id="doc_id") as empty table are do not increment doc_id -> missalign doc_ids
starting @ 5
annot_entit <- jslite_annot$labels %>%
  map(as_tibble,.id="doc_id") %>% enframe() %>%  unnest(value) %>%
  transmute(doc_id=as.numeric(name), start=as.numeric(V1), end=as.numeric(V2), entity=as.factor(V3)) %>%
  group_by(doc_id)
(annot_entit %>% filter(doc_id==11))
# # A tibble: 122 x 4
# # Groups:   doc_id [8]
#    doc_id start  stop entity
#     <dbl> <dbl> <dbl> <fct>
#  1      1  2802  2808 Name
#  2      1  2850  2889 Name
#  3      2   531   547 Name
```
```

Il faut ré-entrainer ! Step 4 : Préparation des données

```
# tokenisation
```{r spacy tokenisation}
annot_lst  <- cnlp_annotate(input=jslite_annot$text , verbose = T) # could be long
# calculate token start position
annot_tok <- annot_lst$token %>%
  group_by(doc_id) %>%
  mutate(tok_ofs = cumsum(str_length(token_with_ws)),
         start = lag(tok_ofs) %>% replace_na(0L),
         end = start + str_length(token))%>%
  select(matches("id|token|start|end"))

(annot_tok %>% filter(doc_id==11) )
# # A tibble: 101,835 x 8
# # Groups:   doc_id [8]
#    doc_id   sid   tid token                    token_with_ws                tid_source start  stop
#     <int> <int> <int> <chr>                    <chr>                             <int> <int> <int>
#  1      1     1     1 610                      "610 "                                2    NA    NA
#  2      1     1     2 S                        "S "                                  0     4     5
#  3      1     1     3 "                      " "                          "          2     6    35
#  4      1     1     4 Devoir                   "Devoir "                             2    35    41
#  5      1     1     5 no                       "no "                                10    42    44
```
```

Il faut ré-entrainer ! Step 4 : Préparation des données

```
# Jointure tokens et annotations

Pour chaque document (doc_id), on utilise `fuzzyjoin::interval_left_join` entre tokens et entités avec une jointure sur `start` et `end` pour couvrir le potentiel espace precedent l'entité annotée.

```{r}
tok_entities <- map_dfr(attributes(annot_entit)[["groups"]]$doc_id,
              ~interval_left_join(annot_tok %>% filter(doc_id==.x) ,
                                  annot_entit %>% filter(doc_id==.x)  %>% ungroup %>% select(-doc_id),
                                  minoverlap = 2)
          ) %>% filter(!str_detect(token,"^\\s+$"))
```
```

Il faut ré-entrainer ! Step 5: Séparer le test du training



Il faut ré-entrainer ! Step 5: Séparer le test du training

On stratifie sur les entites pour équilibrer les 2 datasets. Ici une correction manuelle est nécessaire. Et on sauve au format TSV pour constituer le fichier d'entrée de Stanford coreNLP

```{r}
train_doc_id <- tok_entities %>%
  filter(!is.na(entity)) %>%
  group_by(doc_id, entity) %>%   summarise(num_rows=n()) %>%
  sample_frac(0.5, weight=num_rows) %>%
  ungroup %>%
  select(doc_id) %>%
  unique %>%
  filter(!doc_id==12) # manual intervention
train <- tok_entities %>% filter(doc_id %in% train_doc_id$doc_id) %>%
  ungroup %>%
  select(token, entity)
test <- tok_entities %>% filter(!doc_id %in% train_doc_id$doc_id) %>%
  ungroup %>%
  select(token, entity)
summary(train)
summary(test)
```

Il faut ré-entrainer ! Step 6: Entrainement

```
# l'Entrainement du modèle
```{r}
model <- crf(y = train$entity,
             x = train[, c("pos", "pos_previous", "pos_next",
                           "token", "token_previous", "token_next")],
             group = train$doc_id,
             method = "lbfgs", file = "tagger.crfsuite",
             options = list(max_iterations = 25, feature.minfreq = 5, c1 = 0, c2 = 1))
model
```
```

Il faut ré-entrainer !

Step 7: Mesurer la performance

```
library(caret)
overview <- confusionMatrix(crf_test$entity, crf_test$label, mode = "pre
c_recall")
overview$overall
overview$byClass[, c("Precision", "Recall", "F1")]
```

Results – Jan 19: F2 score on **Global targetted datasets**

`-total-` count here **only** refers to **anonymised entities**, so doesn't count for `TeamId` nor `Organisation` tokens.

entity <chr>	f2 <chr>	f1 <chr>	precision <chr>	recall <chr>	tp <int>	fn <int>	fp <dbl>
Email	88.8%	92.7%	100.0%	86.4%	19	3	0
Nominal_references	0.0%	0.0%	0.0%	0.0%	0	3	0
Organisation	92.5%	94.8%	99.0%	91.0%	1240	123	13
PersonName	95.8%	97.1%	99.3%	95.0%	2689	143	20
TeamId	95.5%	96.9%	99.4%	94.5%	811	47	5
TelNumber	97.0%	97.9%	99.6%	96.4%	667	25	3
UserId	97.7%	98.4%	99.6%	97.3%	755	21	3
-total-	96.3%	97.4%	99.4%	95.6%	4130	192	26

8 rows

RGPD : Notification des fuites de données 1/2 : Données tabulaires un extracteur de noms de colonnes !

```r
xls_files <-list.files("~/Download", pattern = "\\.(xls|XLS)\\w?$",
                       recursive = T, full.names = T) # 842 files

read_xls_colnames <- function(filename) {
  filename %>%
    excel_sheets() %>%
    set_names() %>%
    map(~ read_excel(.x, path = filename) %>% colnames)
}
xls_columns <- tibble(
  path = map_chr(xls_files, dirname),
  file = map_chr(xls_files, basename),
  sheet_cols = map(xls_files, possibly(
    read_xls_colnames, otherwise = list(NA_character_)
  ))
) %>%
  mutate(nb_sheets = map(.$sheet_cols, length))
```

RGPD : Notification des fuites de données 2/2 : Données documentaires

```
library(readtext)
DATA_DIR <- "~/Download/"
# read in all files from a folder
texts <-readtext(file=paste0(DATA_DIR, "/*.pdf"), docvarsfrom = "metadata",verbosity = 2)
# Reading texts from ~/Download/*.pdf
# PDF error: Invalid Font Weight
#...
# PDF error: Could not parse ligature component "folder" of "folder_close_alt" in parseCharName
# PDF error: Could not parse ligature component "close" of

# PDF error: Could not parse ligature component "level" of
# PDF error: Could not parse ligature component "down" of
#  ... read 108 documents.
# texts
# readtext object consisting of 108 documents and 0 docvar
# # Description: df[,2] [108 × 2]
#   doc_id                                  text
#   <chr>                                   <chr>
# 1 10.1038@s41598-017-12401-8.pdf          "\"
# 2 1703_WhyDoWeVisualiseData.pdf           "\"Lisa Charl
# 3 1910012156_TL-MR3020(EU)_V3_UG.pdf      "\"User Guide
```

VITA RADIUM SUPPOSITORIES

Actual Size of Suppository

OUR VITA RADIUM SUPPOSITORIES (HIGH STRENGTH) are one of the outstanding triumphs of Radium Science. These Suppositories are guaranteed to contain REAL RADIUM—in the exact amount for most beneficial effect. They are inserted per rectum, one each night, this being one of the several practical and successful ways of introducing Radium into the system.

After insertion, the Suppository quickly dissolves and the Radium is absorbed by the walls of the colon; then, within a few minutes, it enters the blood stream and traverses the entire body. Every tissue, every organ of the body is bombarded by its health-giving electric atoms. Thus the use of these Suppositories has an effect on the human body like recharging has on an electric battery.

And remember, Radium taken into the system remains for months, continuing its curative, restorative work. Thus, the effects are NOT merely temporary.

VITA RADIUM SUPPOSITORIES are guaranteed to be non-injurious—they are perfectly safe for anyone to use. Their action is due solely to the Radium contained therein.

Et vous, quel est votre outil ?

Repo du code : https://github.com/cregouby/RGPD_facile_avec_R.git

Thank you