

Pour un usage réflexif de R dans sa pratique des sciences sociales

Séminaire R à l'Usage en sciences sociales (RUSS)

3 décembre 2021 par E. Morand, B. Garnier et J. Gourmelin

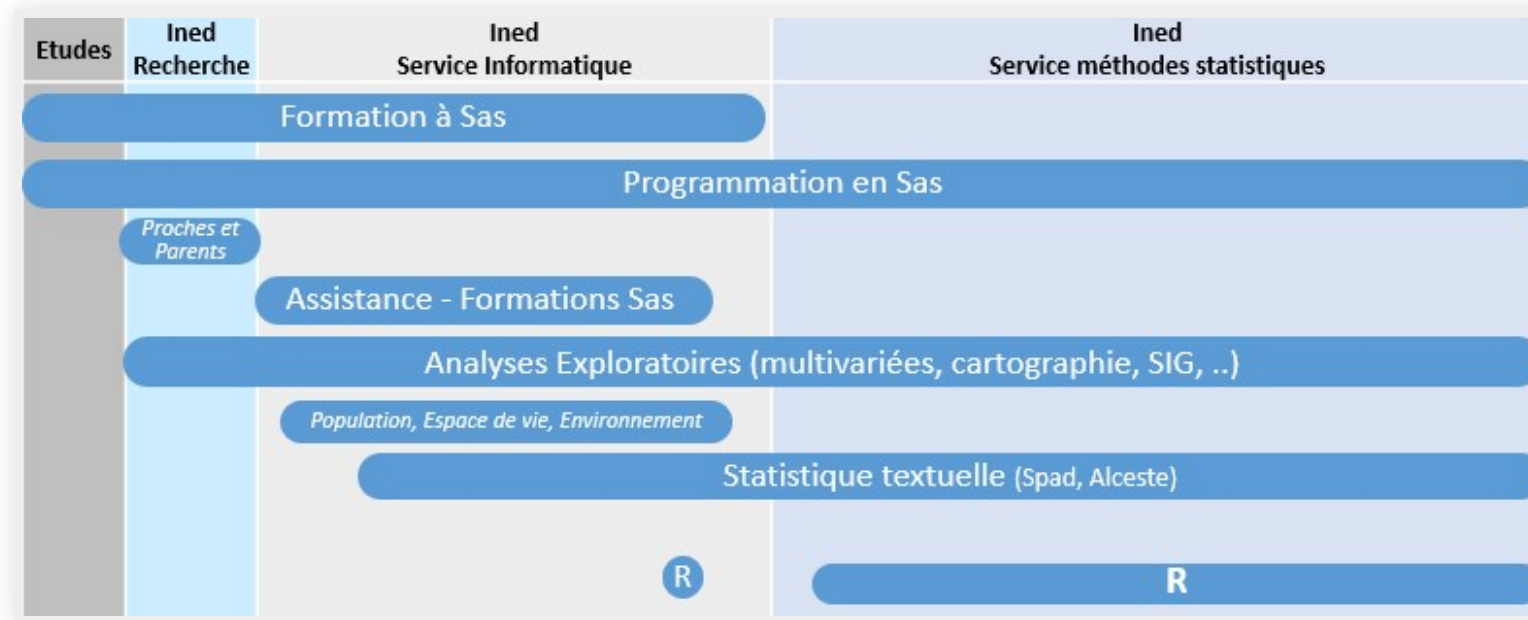


2021-2022 Nouvelle saison de RUSS

Un nouveau et un ancien RUSS

- Prolongation du format *à distance*
- Format hybride. Organismes à l'Ined
- un coup d'oeil dans le retro
 - Création en 2013
 - <https://russ.site.ined.fr/fr/seances-passees/jepudi-7-decembre-2017/>

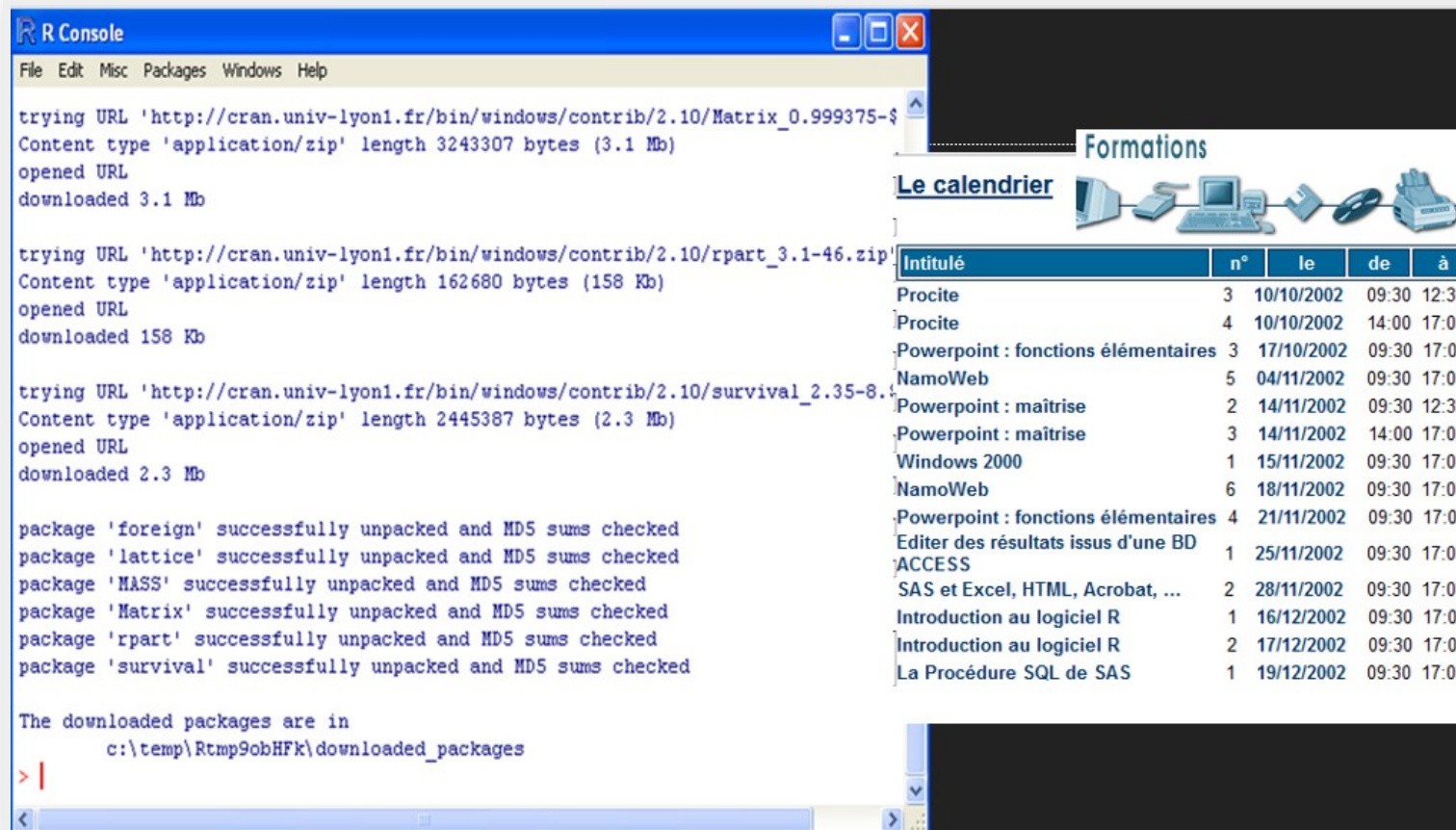
Apprentissage de Bénédicte



Beaucoup de logiciels payants. Première formation à R en 2002

Statistique textuelle (Lebart et Salem, 1994)

2002 - Formation Ined ... presque 20 ans



The screenshot shows an R Console window on the left and a 'Formations' calendar table on the right. The R Console window has a menu bar (File, Edit, Misc, Packages, Windows, Help) and a command line showing the installation of several packages from CRAN. The 'Formations' table lists various training sessions with columns for title, number, date, and time.

```
R Console
File Edit Misc Packages Windows Help

trying URL 'http://cran.univ-lyon1.fr/bin/windows/contrib/2.10/Matrix_0.999375-5.zip'
Content type 'application/zip' length 3243307 bytes (3.1 Mb)
opened URL
downloaded 3.1 Mb

trying URL 'http://cran.univ-lyon1.fr/bin/windows/contrib/2.10/rpart_3.1-46.zip'
Content type 'application/zip' length 162680 bytes (158 Kb)
opened URL
downloaded 158 Kb

trying URL 'http://cran.univ-lyon1.fr/bin/windows/contrib/2.10/survival_2.35-8.1.zip'
Content type 'application/zip' length 2445387 bytes (2.3 Mb)
opened URL
downloaded 2.3 Mb

package 'foreign' successfully unpacked and MD5 sums checked
package 'lattice' successfully unpacked and MD5 sums checked
package 'MASS' successfully unpacked and MD5 sums checked
package 'Matrix' successfully unpacked and MD5 sums checked
package 'rpart' successfully unpacked and MD5 sums checked
package 'survival' successfully unpacked and MD5 sums checked

The downloaded packages are in
  c:\temp\Rtmp9obHfK\downloaded_packages
> |
```

Intitulé	n°	le	de	à
Procite	3	10/10/2002	09:30	12:30
Procite	4	10/10/2002	14:00	17:00
Powerpoint : fonctions élémentaires	3	17/10/2002	09:30	17:00
NamoWeb	5	04/11/2002	09:30	17:00
Powerpoint : maîtrise	2	14/11/2002	09:30	12:30
Powerpoint : maîtrise	3	14/11/2002	14:00	17:00
Windows 2000	1	15/11/2002	09:30	17:00
NamoWeb	6	18/11/2002	09:30	17:00
Powerpoint : fonctions élémentaires	4	21/11/2002	09:30	17:00
Editer des résultats issus d'une BD ACCESS	1	25/11/2002	09:30	17:00
SAS et Excel, HTML, Acrobat, ...	2	28/11/2002	09:30	17:00
Introduction au logiciel R	1	16/12/2002	09:30	17:00
Introduction au logiciel R	2	17/12/2002	09:30	17:00
La Procédure SQL de SAS	1	19/12/2002	09:30	17:00

RConsole

2009 - Participation à un projet de recherche européen

Plus de 10 pays partenaires : France, Belgique, Portugal, Inde, Brésil, Chine, Cameroun, Tunisie, Hongrie, Suède, ...

Enquête sur la vision de l'Europe dans le monde avec une **question ouverte** : *"Quels sont les mots que vous associez le plus à Europe ?"* (5 max). 9340 étudiants interrogés dans 18 pays différents



Enquête sur la vision de l'Europe dans le monde

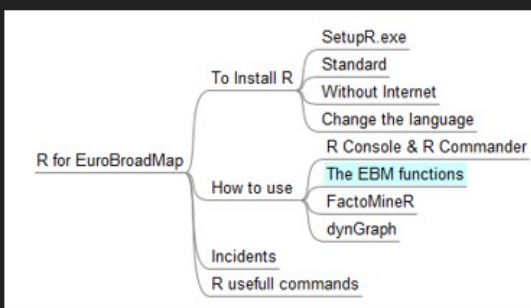
9340 étudiants enquêtés dans 18 pays et 43 villes

num	G_State	G_City	G_Study	A1_Gend	A10_Levin	EduM	PaysVis	nblang	VeuEtrange	D2_tot
1	BEL	BRU	ART	M	Inc3	4	4 Vis5	3lang	etra_N	economy politics development bureaucrat decree
2	BEL	BRU	ART	M	Inc2	4	4 Vis5	2lang	etra_O	coldness white not welcoming colonizer
3	BEL	BRU	ART	M	Inc3	2	2 Vis5	2lang	etra_N	freedom trade culture tolerance globalism
4	BEL	BRU	ART	F	Inc4	4	4 Vis5	3lang	etra_N	union power traditions
5	BEL	BRU	ART	M	Inc3	3	3 Vis5	3lang	etra_N	unity power social equality policy
6	BEL	BRU	ART	M	Inc4	4	4 Vis5	4+lang	etra_O	cultural rich
7	BEL	BRU	ART	M	Inc3	3	3 Vis5	2lang	etra_N	union policy agriculture energy brussels
8	BEL	BRU	ART	F	Inc2	3	3 Vis_1a4	3lang	etra_N	economy mutual aid
10	BEL	BRU	ART	F	Inc3	4	4 Vis5	3lang	etra_O	occident european union brussels
11	BEL	BRU	ART	M	Inc3	2	3 Vis5	4+lang	etra_O	together community euro border industrial
12	BEL	BRU	ART	F	Inc4	4	3 Vis_1a4	4+lang	etra_O	practice home
13	BEL	BRU	ART	M	Inc3	4	4 Vis5	3lang	etra_N	history
14	BEL	BRU	ART	F	Inc2	3	3 Vis5	3lang	etra_N	unity peace economy treaty travel
15	BEL	BRU	ART	F	Inc3	3	3 Vis5	4+lang	etra_N	euro travel dependency relation
16	BEL	BRU	ART	M	Inc4	4	3 Vis5	2lang	etra_N	old continent
17	BEL	BRU	ART	M	Inc4	4	4 Vis5	3lang	etra_O	continent economic power diversity
19	BEL	BRU	ART	M	Inc2	4	4 Vis5	2lang	etra_O	euro union
20	BEL	BRU	ART	M	Inc4	4	4 Vis5	4+lang	etra_N	union neighbor power similar
21	BEL	BRU	ART	M	Inc2	2	3 Vis5	3lang	etra_N	mediterranean colonizing country mix dynamic movement
22	BEL	BRU	ART	M	Inc3	4	4 Vis_1a4	4+lang	etra_O	union currency cultures
23	BEL	BRU	ART	M	Inc2	2	3 Vis_1a4	3lang	etra_N	collaboration importance politic economy ethnicity
24	BEL	BRU	ART	F	Inc3	4	4 Vis5	3lang	etra_N	union sharing exchange
25	BEL	BRU	ART	F	Inc3	4	3 Vis5	4+lang	etra_N	union euro transfer divergence
27	BEL	BRU	ART	M	Inc3	3	3 Vis5	4+lang	etra_O	paris london old continent
28	BEL	BRU	ART	F	Inc2	4	2 Vis5	4+lang	etra_N	exchange opening communication
29	BEL	BRU	ART	F	Inc2	3	3 Vis5	2lang	etra_N	economy anything great
30	BEL	BRU	ART	F	Inc3	4	4 Vis5	3lang	etra_N	union community economy plurality exchange

Package tm pour text minig (publié en 2008) : DocumentTermMatrix, stemmatisation, comptage

R "non " propriétaire, tous pays, tous environnements, avec ou sans internet

Plongée dans R



I. Function `fvocabulary()`

Creates and exports an Excel workbook named *Words_and_Frequencies.xls*, which contains all the words of the corpus with their frequencies.

Open R and write these commands (on the command line of the R Console or a script):

```
> fvocabulary(folder="D:\\myfolder\\", data_workbook_Excel = "mydatawb.xls", data_sheet_Excel = "sheet1", vector_variables_words = c("ID_D2_1", "ID_D2_2", "ID_D2_3"), byFreqDecr = T)
```

SetupR.exe (contains)

FILE	SIZE
car_1.2-16.zip	467 Ko
R-2.10.1-win32.exe	31 741 Ko
Rcmdr_1.5-4.zip	2 241 Ko
RcmdrPlugin.FactoMineR_1.0...	86 Ko
rJava_0.8-2.zip	522 Ko
RODBC_1.3-1.zip	650 Ko
RWeka_0.4-1.zip	6 449 Ko
slam_0.1-9.zip	34 Ko
tm_0.5-2.zip	500 Ko

And

fReturnTLA.r	5 Ko	Tinn-R
fReturnTLE.r	2 Ko	Tinn-R
fvocabulary.r	2 Ko	Tinn-R
fvocspec.r	4 Ko	Tinn-R

EuroBroadMap PROJECT - FP7- SHS- WP2

Textual Statistics Methods for exploring the Visions of Europe in the World (D2 question)

February 2010

R Functions for text mining

Parameters	Roles
folder = "...\\...\\..."	Data's folder
data_workbook_Excel = "nameworkbook.xls"	The data (in Excel)
data_sheet_Excel = "namesheet"	The sheet of the Excel workbook
vector_variables_words = c("variable1",...)	Vector containing the names of all the variables selected for the corpus
Optional : byFreqDecr = F (or T)	Option which determine the order of the words in the exported Excel workbook. By default, this option is False and the words are sorted by alphabetical order. If option=True then the words are sorted by decreasing frequencies.

Création de *fonctions sur mesure* (G. Taché et E. Morand, SMS Ined) et supports à disposition pour l'équipe : tutoriels, installation, erreurs fréquentes ...

Le code était “derrière”

```
# Function by Gwendoline TACHE -- Ined, Service Méthodes Statistiques -- 2009 December

fvocabulary <- function(folder, data_workbook_Excel,
data_sheet_Excel, vector_variables_words, byFreqDecr=F,
minLgWord=1) {

#####
library(RODBC)
chemin <- paste(folder,data_workbook_Excel,sep="/")
connexion <- odbcConnectExcel(chemin)
don <- sqlFetch(connexion, data_sheet_Excel)
odbcCloseAll()
library(tm)
library(RWeka)
#####
# vector of text
v_vsource <- NULL
for(nomcol in
vector_variables_words[1:length(vector_variables_words)]){
  v_vsource <-
paste(v_vsource,don[,which(colnames(don)==nomcol)],sep="
")}

# create the corpus :
lou <- Corpus(VectorSource(v_vsource))

# create the TLE :
tle <-
inspect(DocumentTermMatrix(lou,control=list(minWordLength
=minLgWord)))

#####

# create table giving all the words with their frequencies
tabwords <- data.frame(colnames(tle),colSums(tle))
colnames(tabwords) <- c("WORDS","FREQUENCIES")

# sort the words by frequencies if the option is required
if(byFreqDecr==T){ tabwords <- tabwords[
sort.list(tabwords$FREQUENCIES,decreasing=T), ] }

# export this table
chemin <- paste(folder, "Words_and_Frequencies.xls",
sep="/")
connexion <- odbcConnectExcel(chemin, readOnly=F)
sqlSave(connexion, tabwords, tablename="Words_and_freq",
rownames=F)
odbcCloseAll()
}
```

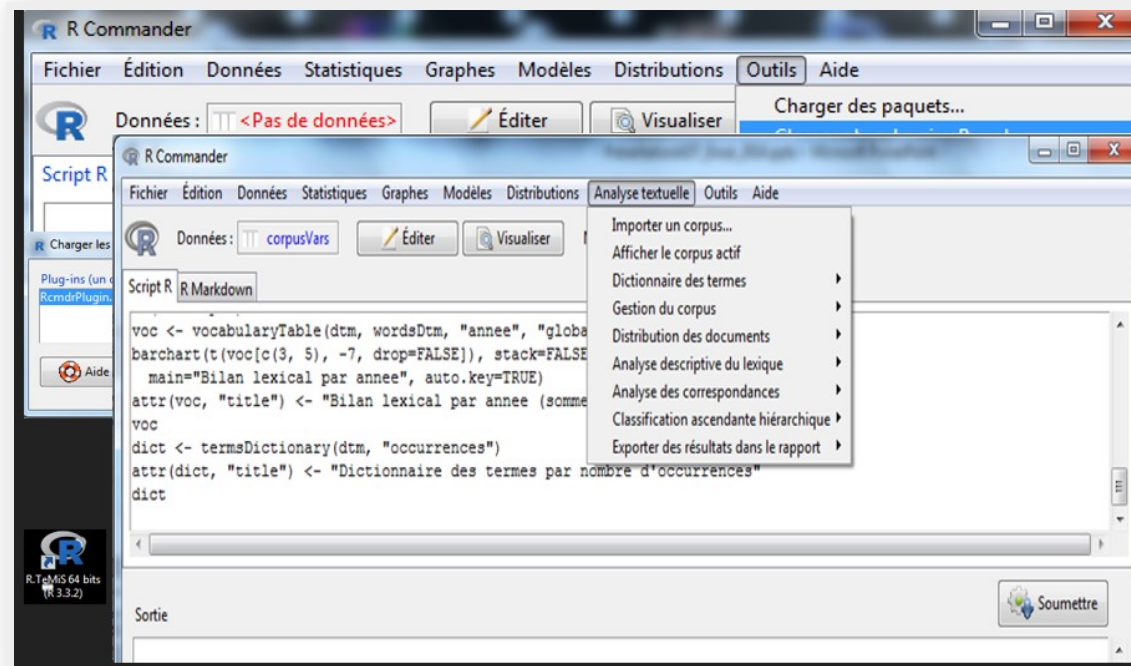
Packages tm, snowball, slam -> Tableau Lexical
(occurrences)

RcmdrPlugin.FactomineR -> Analyses factorielles
(cooccurrences)

Rassurée par les interfaces graphiques

RCommander pour importer/exporter les données, recoder des variables, statistiques univariées et bivariées, produire des graphiques.

Utilisation des *plug-ins* : *FactoMineR* pour l'analyse des données : ACP, AFC, ACM, classifications, ... et [*R.TeMiS*] (<https://rtemis.hypotheses.org/>) pour la statistique textuelle (2014) : M. Bouchet-Valat et G. Bastin



Initiation à la programmation dans R ?

Progression avec le séminaire RUSS depuis 2013

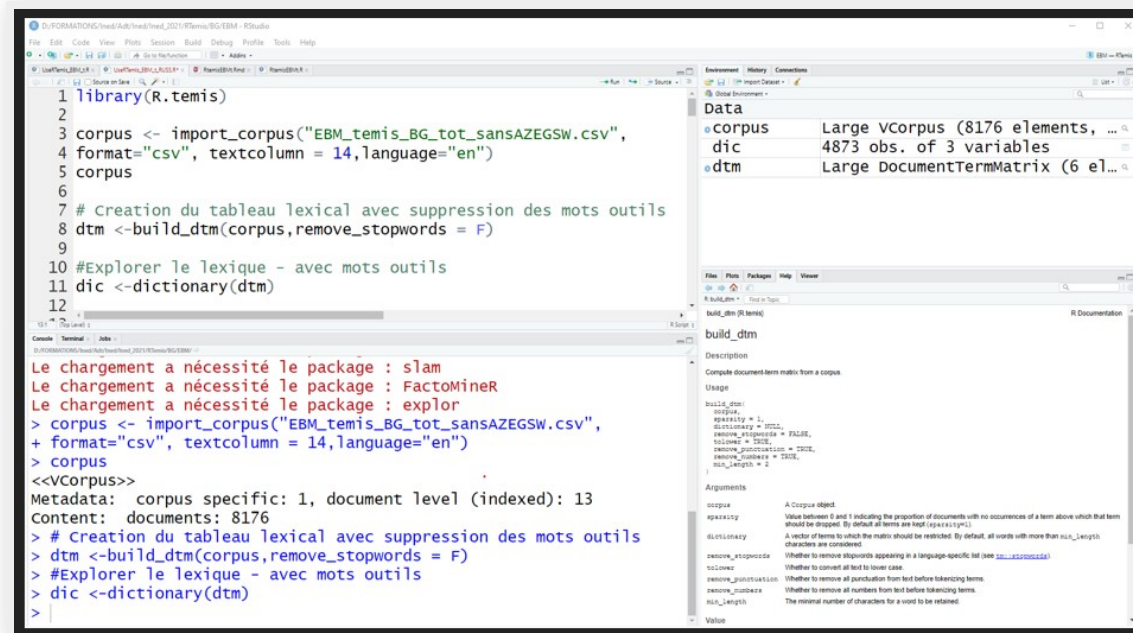
Présentations variées

- Introduction à R
- Analyses : de réseaux, de textes, de trajectoires, les forêts aléatoires, cartographie,
- Web scarping et APIs
- Shiny, Dplyr, Tidyverse
- Reproductibilité
- Retours d'expérience

2019 - RStudio

Enfin le codage facilité avec l'Interface **RStudio** 👍 et nouveau package **R.temis** (sans RCommander)

4 fenêtres principales : Code avec coloration syntaxique, auto complétion, Addins (questionr); Console avec messages et résultats ; Liste des objets dans l'environnement ; Aide sur la syntaxe, fenêtre graphique, installation/activation de packages



The screenshot displays the RStudio environment with the following components:

- Code Editor:** Contains R code for loading the `R.temis` package, importing a corpus, and building a document-term matrix (DTM).
- Console:** Shows the execution of the code, including package loading messages and the creation of the corpus and DTM objects.
- Environment:** Lists the objects in the current environment: `corpus` (Large vCorpus), `dic` (4873 obs. of 3 variables), and `dtm` (Large DocumentTermMatrix).
- Help Window:** Displays the documentation for the `build_dtm` function, including its description, usage, and arguments.

```
1 library(R.temis)
2
3 corpus <- import_corpus("EBM_temis_BG_tot_sansAZEGSW.csv",
4 format="csv", textcolumn = 14,language="en")
5 corpus
6
7 # Creation du tableau lexical avec suppression des mots outils
8 dtm <- build_dtm(corpus,remove_stopwords = F)
9
10 #Explorer le lexique - avec mots outils
11 dic <- dictionary(dtm)
12
```

Le chargement a nécessité le package : slam
Le chargement a nécessité le package : FactoMiner
Le chargement a nécessité le package : explor
> corpus <- import_corpus("EBM_temis_BG_tot_sansAZEGSW.csv",
+ format="csv", textcolumn = 14,language="en")
> corpus
<<vCorpus>>
Metadata: corpus specific: 1, document level (indexed): 13
Content: documents: 8176
> # Creation du tableau lexical avec suppression des mots outils
> dtm <- build_dtm(corpus,remove_stopwords = F)
> #Explorer le lexique - avec mots outils
> dic <- dictionary(dtm)
>

Environment:

Object	Class	Size
corpus	Large vCorpus	(8176 elements, ...)
dic	Dictionary	4873 obs. of 3 variables
dtm	Large DocumentTermMatrix	(6 e1...

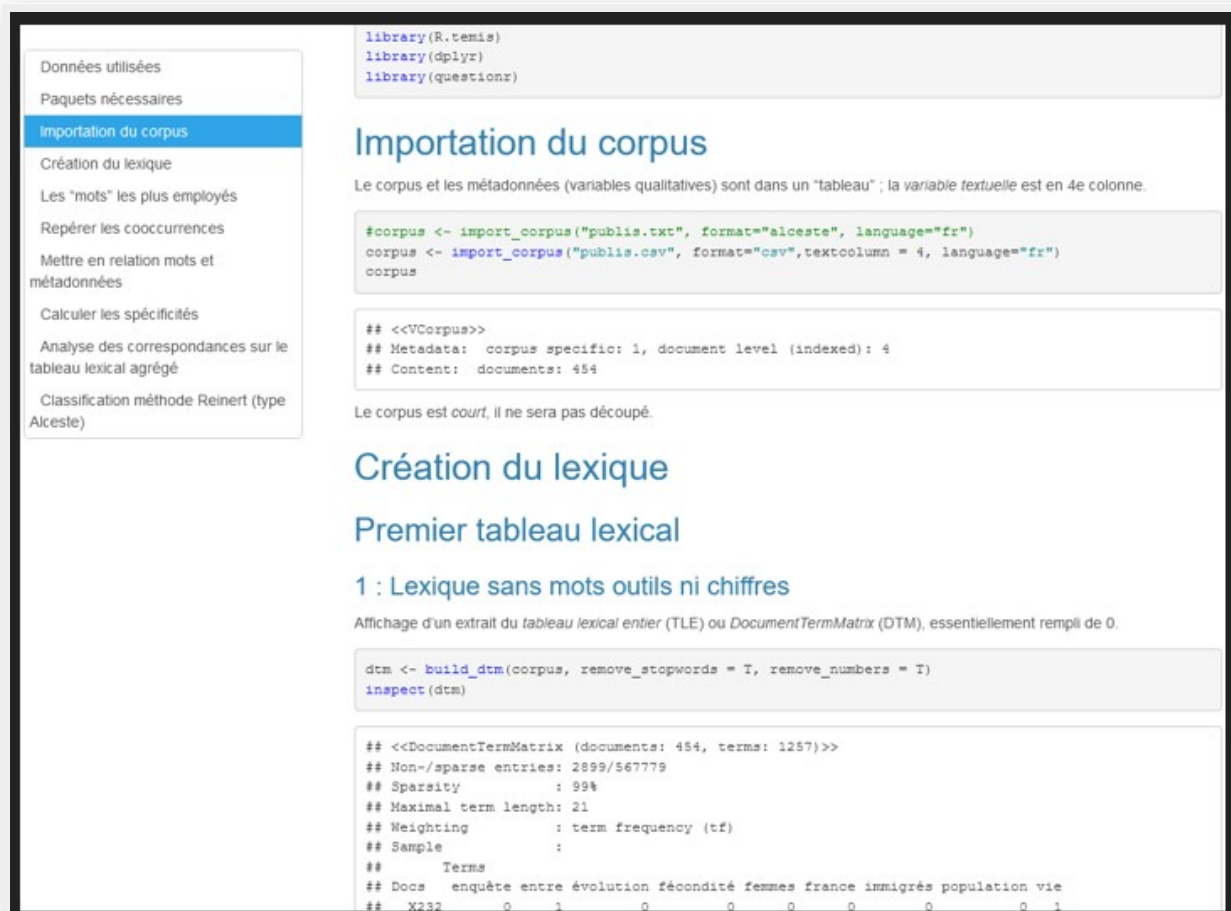
build_dtm
Description
Compute document term matrix from a corpus
Usage
build_dtm(
 corpus,
 remove_stopwords = FALSE,
 dictionary = NULL,
 remove_punctuation = FALSE,
 remove_numbers = FALSE,
 min_length = 2
)
Arguments
corpus A Corpus object
remove_stopwords Value between 0 and 1 indicating the proportion of documents with no occurrences of a term above which that term should be dropped. By default all terms are kept (remove_stopwords = 0).
dictionary A vector of terms to which the matrix should be restricted. By default, all words with more than min_length characters are considered.
remove_punctuation Whether to remove punctuation appearing in a language-specific list (see [get_punctuation](#)).
remove_numbers Whether to convert all text to lower case.
min_length Whether to remove all punctuation from text before tokenizing terms.
Value Whether to remove all numbers from text before tokenizing terms.
The minimal number of characters for a word to be retained.

Les supports de cours

Utilisation du **R Markdown** (dans R Studio) pour générer des documents en .html,.doc,.pdf

Texte avec balises "légères" pour gérer les niveaux de titres, gras, italique pour ajouter des commentaires ; code ("chunks"), résultats, images, ...

Facilite la *reproductibilité*



```
library(R.temis)
library(dplyr)
library(questionr)
```

Importation du corpus

Le corpus et les métadonnées (variables qualitatives) sont dans un "tableau" ; la variable *textuelle* est en 4e colonne.

```
#corpus <- import_corpus("publis.txt", format="alceste", language="fr")
corpus <- import_corpus("publis.csv", format="csv", textcolumn = 4, language="fr")
corpus
```

```
## <<VCorpus>>
## Metadata: corpus specific: 1, document level (indexed): 4
## Content: documents: 454
```

Le corpus est *court*, il ne sera pas découpé.

Création du lexique

Premier tableau lexical

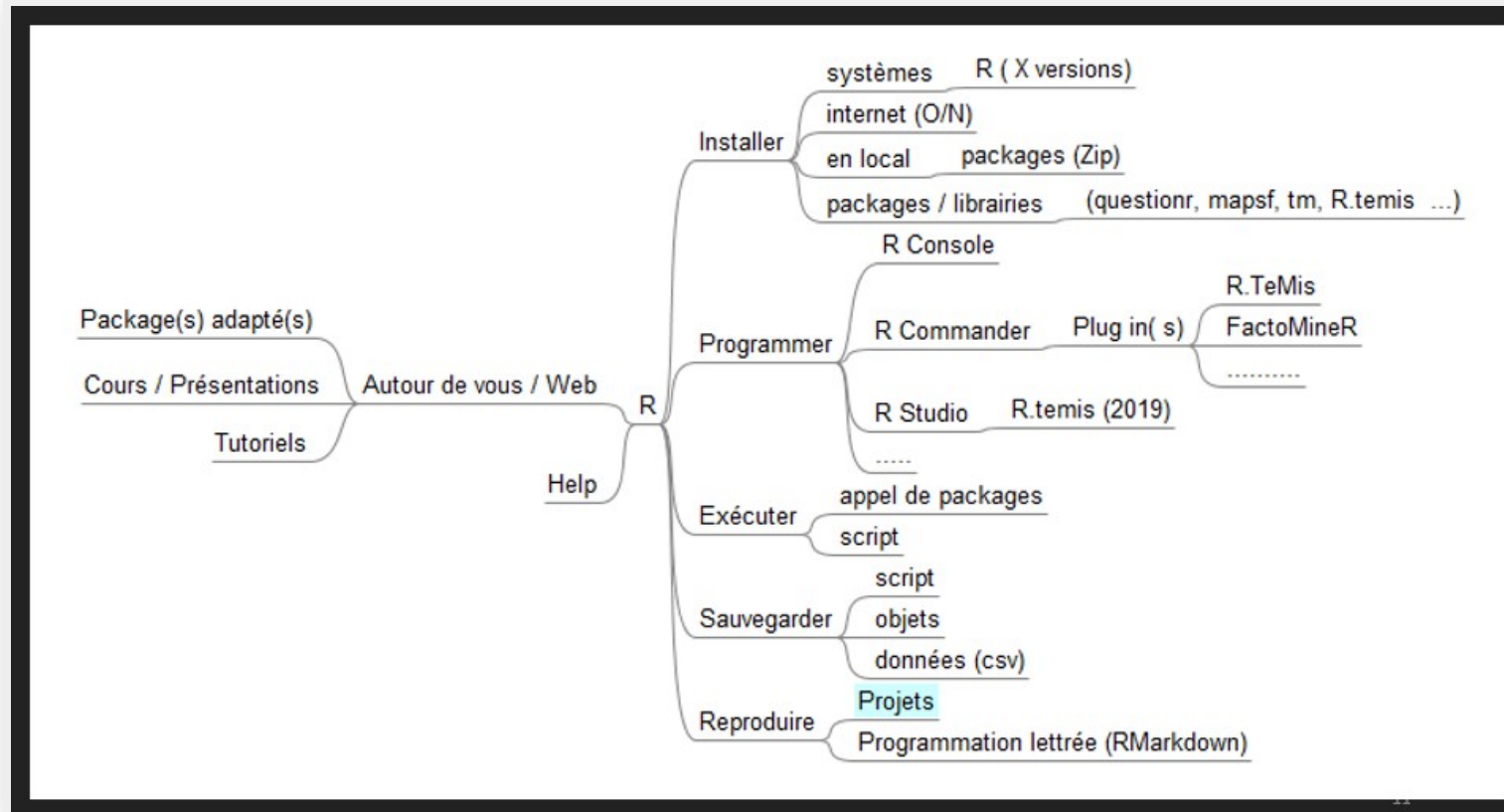
1 : Lexique sans mots outils ni chiffres

Affichage d'un extrait du *tableau lexical entier* (TLE) ou *DocumentTermMatrix* (DTM), essentiellement rempli de 0.

```
dtm <- build_dtm(corpus, remove_stopwords = T, remove_numbers = T)
inspect(dtm)
```

```
## <<DocumentTermMatrix (documents: 454, terms: 1257)>>
## Non-/sparse entries: 2899/567779
## Sparsity : 99%
## Maximal term length: 21
## Weighting : term frequency (tf)
## Sample :
## Terms
## Docs enquête entre évolution fécondité femmes france immigrés population vie
## X232 0 1 0 0 0 0 0 0 0 1
```

Ma pratique de R



Joyeux anniversaiRe



Le 29 février 2000, R v 1.0.0

<https://jozef.io/r921-happy-birthday-r/>

Apprendre R (avant)

- Pas de formation, Pas de cours
- "Le PaRadis"
- Les fiches ADE4

Apprendre R (après)

- “Le baRnier” (<https://juba.github.io/tidyverse/index.html> “juba”)
- “le grimoire” (https://perso.ens-lyon.fr/lise.vaudor/grimoireStat/_book/intro.html “Lise Vaudor”)
- etc...
- les cours en ligne
- les certifications (<https://silvia.rbind.io/blog/rstudio-instructor-certification-tidyverse/> “Rstudio”)
- webinr <https://larmarange.github.io/webin-R/>
- les presses universitaires de Rennes
- rladies <https://twitter.com/RLadiesGlobal>
- data shs

Interface de R

- au début était la ligne de commande ...(la console)
- on a enseigné avec Rcmdr (et les joies de Tcl/TK)
- en 2012 : **numero de JSS sur les graphical interface for R**

La programmation lettrée (1)

- Les scripts et leur commentaires et ...
- l'usages des éditeurs de textes avec colloration syntaxique
 - Xemacs
 - Tinn-R
 - Notepad+ +
 - ...

La programmation lettrée (2)

- Sweave
 - xtable
 - latex
 - cat & \
- presentation MNHN, 29 Mars 2012
- knitr

Faire un package

- Le guide initial [Writing R Extensions](#) (version actuelle) ;
- Heureusement il y a les users groupes [RUG MNHN](#)
- Le [livre](#)
- [faire un package en quelques minutes](#)
- [Fusen](#)

<https://www.r-bloggers.com/2021/08/fusen-is-now-available-on-cran/>

Faire des formations

Faire des formations

- Les interfaces

Faire des formations

- Les interfaces
 - Rcmdr

Faire des formations

- Les interfaces
 - Rcmdr
 - Jamovi

Faire des formations

- Les interfaces
 - Rcmdr
 - Jamovi
- swirl

Faire des formations

- Les interfaces
 - Rcmdr
 - Jamovi
- swirl
[lien](#)

Faire des formations

- Les interfaces
 - Rcmdr
 - Jamovi
- swirl
- [lien](#)
- learnr

Faire des formations

- Les interfaces
 - Rcmdr
 - Jamovi
- swirl
[lien](#)
- learnr
[lien](#)

Faire des formations

- Les interfaces
 - Rcmdr
 - Jamovi
- swirl
[lien](#)
- learnr
[lien](#)
- livecode

Faire des formations

- Les interfaces
 - Rcmdr
 - Jamovi
- swirl
[lien](#)
- learnr
[lien](#)
- livecode
[lien](#)

Faire des formations

- Les interfaces
 - Rcmdr
 - Jamovi
- swirl
[lien](#)
- learnr
[lien](#)
- livecode
[lien](#)
- et enfin les add-in

Dplyr

- L'enseignement et la formation
<https://russ.site.ined.fr/fr/seances-passees/19-novembre-2015/>
- Enseigner R en shs
<https://russ.site.ined.fr/fr/seances-passees/jeudi-5-avril-2018/>
- Pourquoi est-ce si difficile d'apprendre R
<https://thinkr.fr/r-difficile-a-apprendre/>

Les tableaux de bords

Dashboard

Shiny

“Flux de travail intégré et conception d’outils pour la recherche et l’enseignement avec R et shiny”

Les graphiques

la R graph gallery

<http://web.archive.org/web/20111219154500/http://addictedtor.free.fr/graphiques/>

La R graph gallery

<https://www.r-graph-gallery.com/>

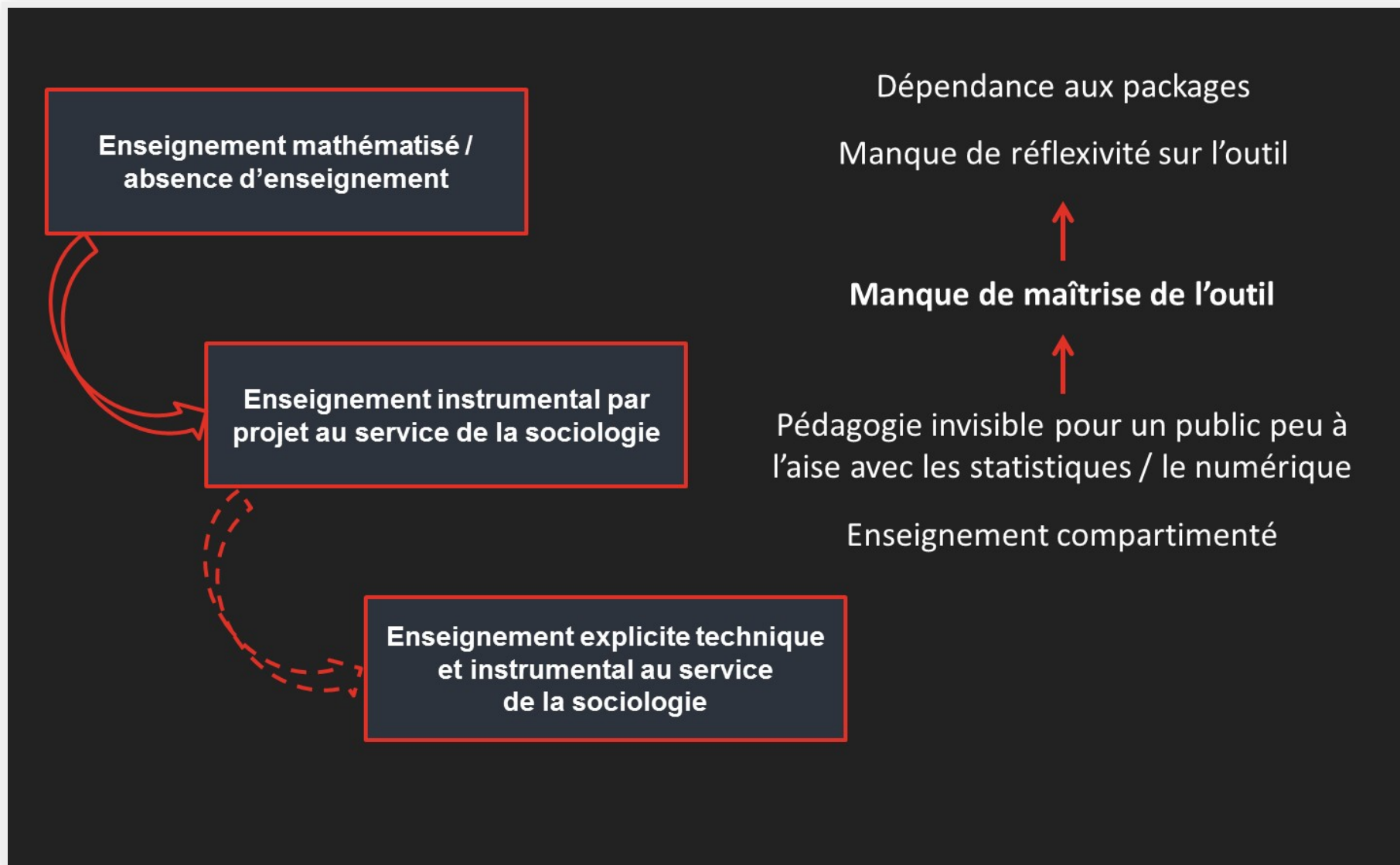
The end?!

R-Enseigné

R dans le curriculum formel de mon master

S1	Conception de questionnaire	<u>Soit sur l'ensemble du master</u>
	Démarche et épistémologie des méthodes quantitatives	
S2	Préparation des données	132 heures de méthodes quantitatives (18 % du volume horaire total)
	Environnement de R (R, RStudio, packages, RMarkdown)	
	Tris (à plat, croisés, trivariés) et test du χ^2	
	Recodages et création de sous-populations	
S3	Rédaction de 4-pages statistiques	Dont 78 heures sur R ou en lien avec (11 % du volume horaire total)
	Régressions logistiques	
	Analyses factorielles (essentiellement les ACM)	
S4	Classifications (essentiellement les CAH)	+ Analyse de réseaux (Hyphe)
	Vecteurs et data frames	
	Révisions des semestres précédents	
S4	Analyse secondaire d'une enquête de la statistique publique en autonomie	+ Analyse textuelle (R, IRaMuTeQ)

Des freins à l'apprentissage de R



Pour un enseignement explicite de R

Pédagogie explicite

1. Apprentissage des bases du langage sans application sociologique
2. Application des traitements usuels sur des bases (accent sur la préparation de données et les tris)
3. Application plus libre en intégrant des problématiques sociologiques
4. Initiation, auto-didaxie et recherche d'assistance sur des techniques avancées

Autonomie

Complexité

Usage sociologique et réflexivité

Tableau 3. Intensité du sentiment de pouvoir librement organiser son temps selon le degré de contrôle de l'organisation de son temps en population étudiante

	Très forte	Forte	Faible	Très faible	Total	Effectifs
Fort	25,2	45,6	18,4	10,7	100	103
Moyen	11,5	46,2	31,4	10,9	100	156
Faible	9,9	28,9	46,3	14,9	100	121
Nul	12,2	26,8	31,7	29,3	100	41

En ligne : degré de contrôle de l'organisation de son temps ; en colonne : intensité du sentiment de pouvoir librement organiser son temps

Lecture : 29,3 % des étudiants ne contrôlant pas du tout leur temps considèrent ne pas pouvoir du tout l'organiser librement

Valeur p associée au khi2 < 0,001



- Création d'une sous-population
- Recodage des variables
- Réalisation et affichage du tri croisé en pourcentages en ligne
- Test du χ^2

Quelques références bibliographiques

Beaud Stéphane, Weber Florence (2012), « Le raisonnement ethnographique », in *L'enquête sociologique*, Presses universitaires de France.

Bissonnette Steve, Richard Mario, Gauthier Clermont (2005), « Interventions pédagogiques efficaces et réussite scolaire des élèves provenant de milieux défavorisés », *Revue française de pédagogie*, n°150(1), p. 87-141.

Blavier Pierre (2021), « Quels enseignements dans les licences de sociologie françaises ? Essai de réponse empirique à partir du décompte des ECTS », *Socio-logos*, n°15, <https://journals.openedition.org/socio-logos/4879> (consulté le 07/10/2021).

Gros Julien (2017), « Quantifier en ethnographe : sur les enjeux d'une émancipation de la représentativité statistique », *Genèses*, n°108(3), p. 129-147.

Pons Xavier (2010), *Évaluer l'action éducative : des professionnels en concurrence*, Presses universitaires de France.

Renisio Yann, Sinthon Rémi (2014), « L'analyse des correspondances multiples au service de l'enquête de terrain. Pour en finir avec le dualisme "quantitatif" / "qualitatif" », *Genèses*, n°97(4), p. 109-125.

Serre Delphine (2019), « Une réflexivité pédagogique sous contraintes », *Socio-logos*, n°14, <http://journals.openedition.org/socio-logos/4164> (consulté le 21/04/2020).

