



## Journée science ouverte CNRS 2023

Patrice Bellot,  
Délégué scientifique CNRS Sciences Informatiques  
Aix-Marseille Université (LIS)

[patrice.bellot@cnrs-dir.fr](mailto:patrice.bellot@cnrs-dir.fr)

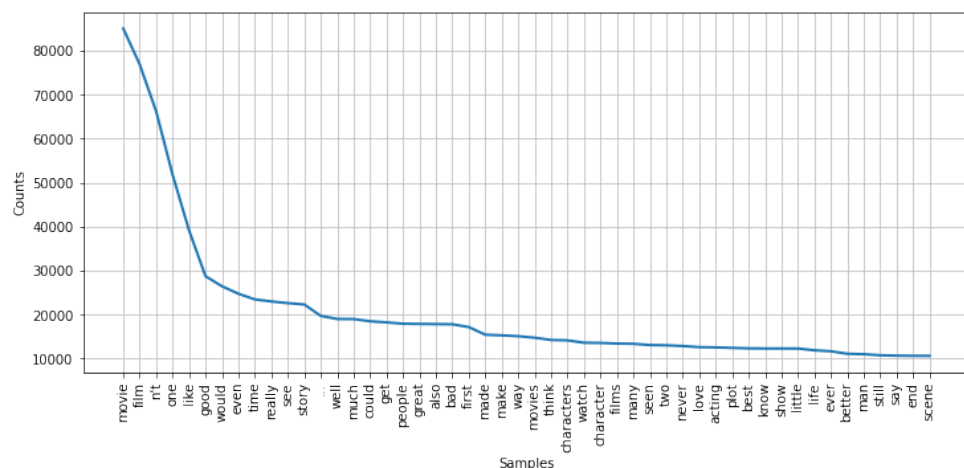
**Panorama des  
méthodologies  
pratiques et outils de  
fouille de textes**



## **Des tâches d'analyse et d'exploration de collections**

## Mots fréquents, co-occurrences, mesures d'association...

```
tokens = word_tokenize(reviews[:1000000])
from nltk.corpus import stopwords
stop_words = set(stopwords.words('english'))
tokens = [w.lower() for w in tokens if not w.lower() in stop_words and len(w)>2]
frequency_dist = nltk.FreqDist(tokens)
frequency_dist.plot(50)
```

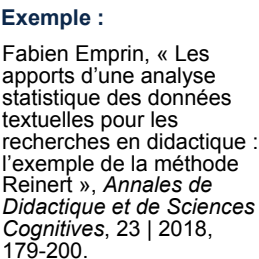


positive	Z_score	negative	Z_score
Love	14.31	Not	13.99
Good	14.01	Fuck	12.97
Happy	12.30	Don't	10.97
Great	11.10	Shit	8.99
Excite	10.35	Bad	8.40
Best	9.24	Hate	8.29
Thank	9.21	Sad	8.28
Hope	8.24	Sorry	8.11
Cant	8.10	Cancel	7.53
Wait	8.05	stupid	6.83

Table1. The first ten terms having the

$$Z_{\text{score}}(t_{ij}) = \frac{\text{tfr}_{ij} - \text{mean}_i}{\text{sdi}}$$

10

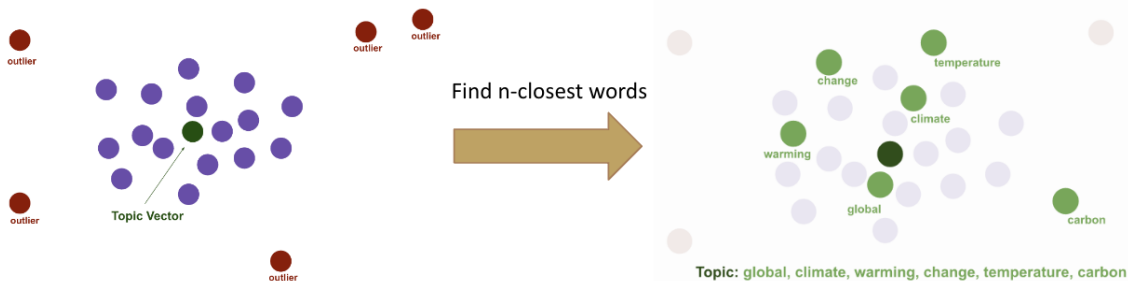
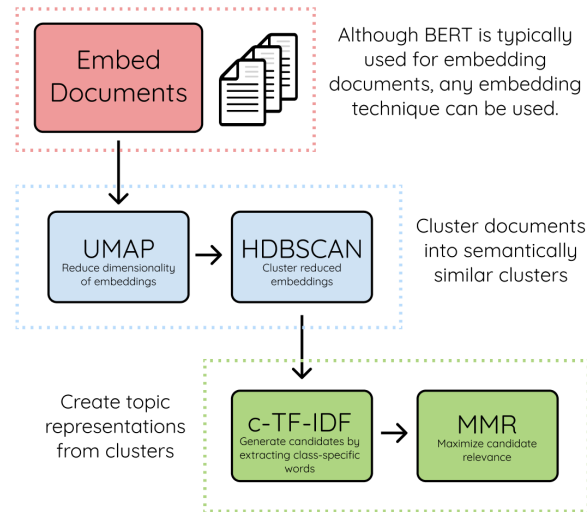
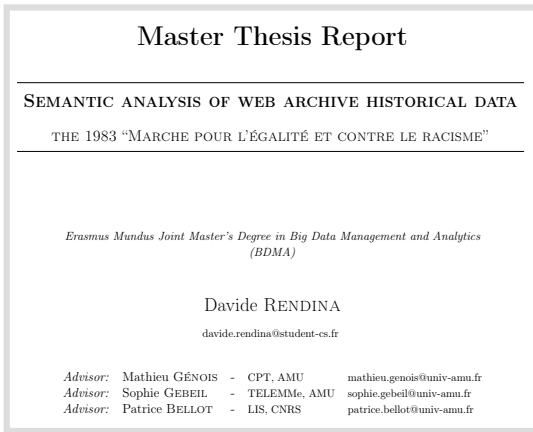


URL : <http://journals.openedition.org/adsc/458>  
DOI : <https://doi.org/10.4000/adsc.458>

1001



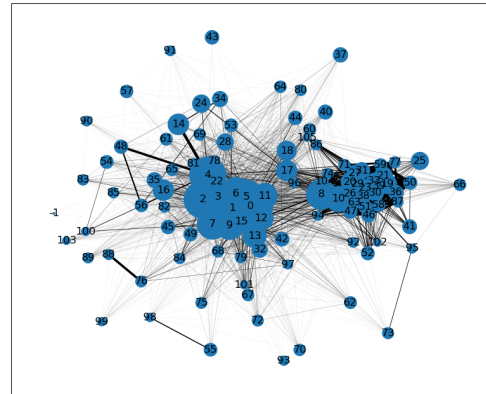
# Exemple d'analyse de thèmes et de partitionnement d'une collection



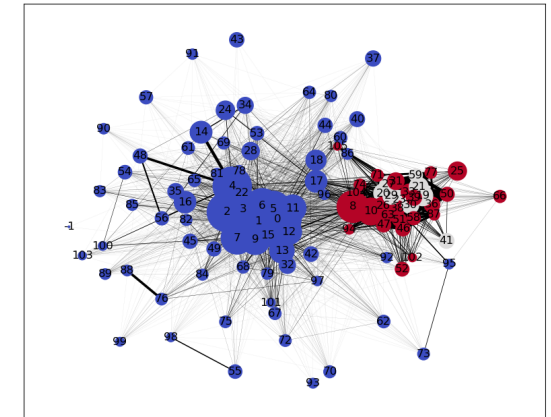
# Analyse de graphes pour la détection de communautés

## Représentations en réseaux (graphes)

Les nœuds du graphe peuvent être des documents, des thèmes ou des entités : si deux nœuds ont un nombre suffisant de documents, de thèmes ou de mots en commun, un lien est créé entre eux.



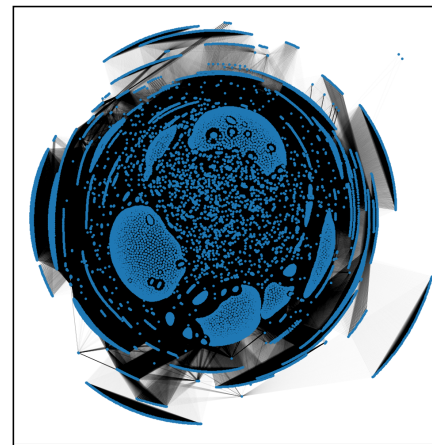
Chaque nœud est un thème ; le poids des liens entre deux thèmes est fonction du nombre de documents ayant ces deux thèmes en commun



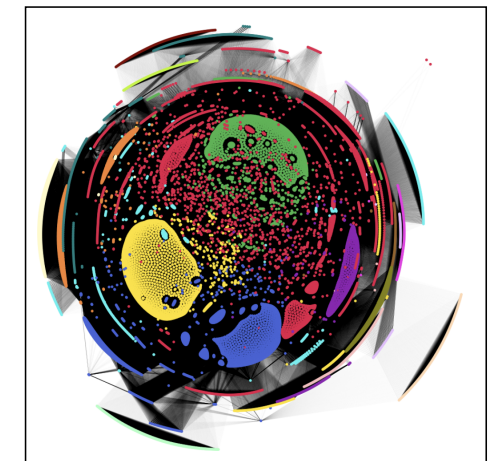
Application de l'algorithme de Louvain pour détecter des communautés de thèmes

Table B.5: Communities of topics extracted by Louvain

Topics	
Community 0	['Commemoration of Algerian War', 'Immigration and Memories', 'TV Guests and Political Entertainment', 'Politics', 'Commemoration and Racism', 'Current Events and News', 'Football and Politics', 'Website Cookies', 'Lobby, Politics and Social Issues', 'Politics and Public Figures', 'Election Campaign', 'Data Management Noise', 'Music Song', 'Television Drama', 'Religious and Historical Documentaries', 'Swiss Politics and Social Issues', 'Medical Breakthroughs', 'French Culture and Society', 'Politics and Criticism', 'Television News', 'Homophobic Speech', 'Film Archives and Interactive Events', 'Migration and Politics (2015)', 'Equality and Social Issues', 'Humanitarian Efforts and Social Issues', 'Online Blogs and Comments', 'Activism and Politics in India', 'Marche and Lyon (Start)', 'Music Genres and Programme', 'Political Commentary', 'Moroccan Community in France', 'Migration Crisis and Politics (2015)', 'Political Proposals and Advocacy', 'Streaming Video and Comments']
Community 1	['Outliers', 'Islam and Muslim Culture', 'Anti-Racism Activism in Marseille in 1983', 'Music, Television and Entertainment', 'Documentary and Film Festivals', 'Youth Education', 'Democracy and Politics', 'Police Violence', 'Immigration and Integration', 'Antisemitism', 'Terrorism and Fundamentalism', 'Miscellaneous', 'Race, Antiracism, and Political Involvement', 'Politics Left', 'Hotel Industry and Ads', 'Hip Hop and Music', 'La Marche Movie', 'Football', 'Political Propaganda', 'Politics and Socialism', 'Bonnets Rouge (Red Caps Movement)', 'Media and Political Controversy', 'Sport', 'Political Commentary', 'Politics Opposition', 'Holland and Elections', 'Music Performance', 'Music, Television and Entertainment', 'Historical Events and Exhibitions', 'Political Critique', 'Television Programming', 'Antiracism and Social Engagement', 'Immigration and Social Movements', 'Media and News', 'Political Statements and Controversial Remarks', 'Education and Gender Issues', 'Citizenship and Political Rights', 'Politics Far-right', 'Women Issues and Activism', 'TV Challenge and Commemoration of the Marche', 'Youtube Video', 'Terrorism and Freedom of Expression', 'Racism in Society', 'Homophobic Comments', 'Online Communication and Technology', 'Educational Resources', 'Radio and Culture', 'Political Controversial Commentary', 'Video and Commentary', 'Infrastructure and Public Projects', 'Miscellaneous', 'Webpage Prompts', 'History and Documentary War',



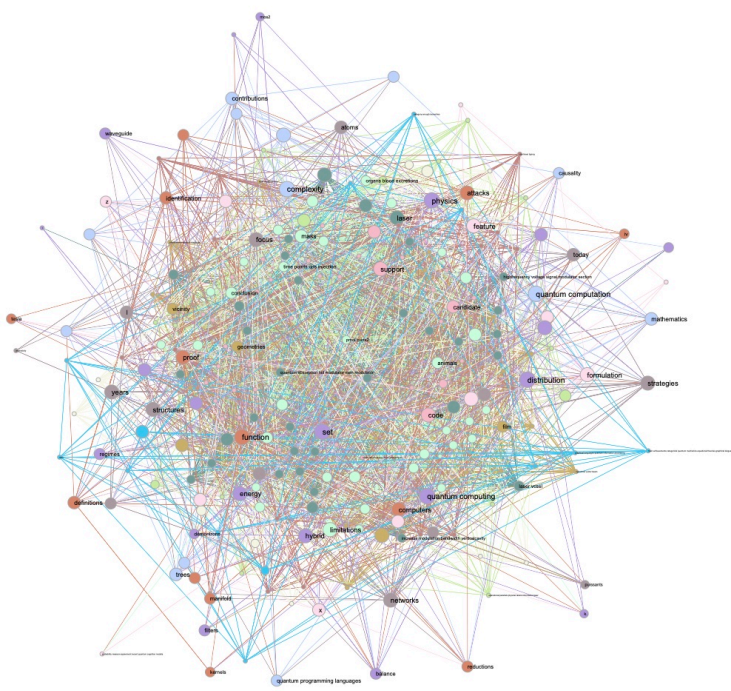
Chaque nœud est un document ; le poids des liens entre deux documents est fonction du nombre de thèmes qu'ils ont en commun



Application de l'algorithme de Louvain pour détecter des communautés de documents

1000000

Les résumés des publications sont extraits de HAL (API HAL, fichier CSV) puis analysés

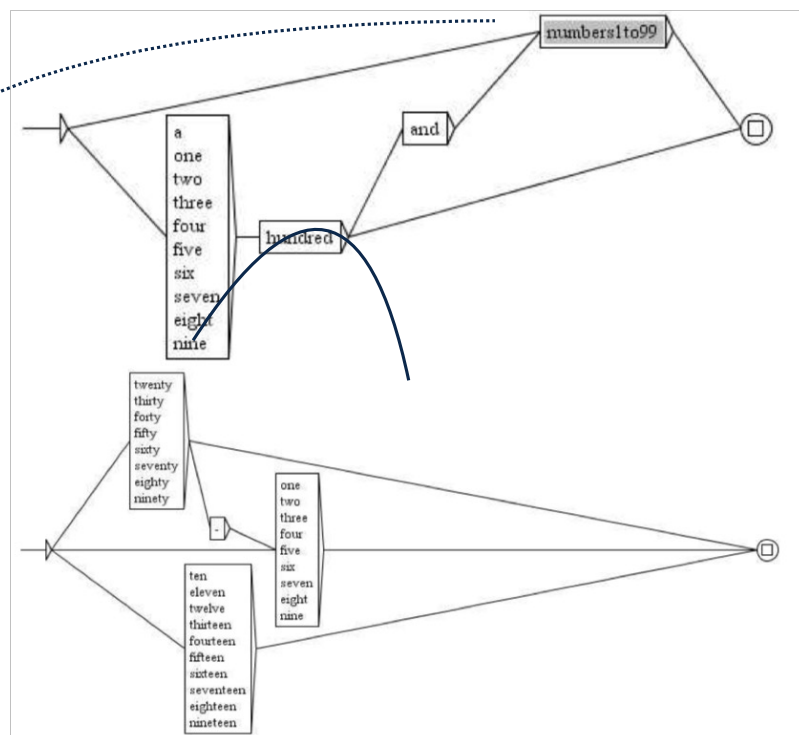


# **Des tâches de catégorisation ou d'annotation automatiques**

Des règles manuelles à l'apprentissage  
machine supervisé

## Approche manuelle par définition de grammaires (expressions régulières, règles symboliques ou automates à états finis)

### exemple : reconnaissance de nombres en anglais (jusqu'à 999)



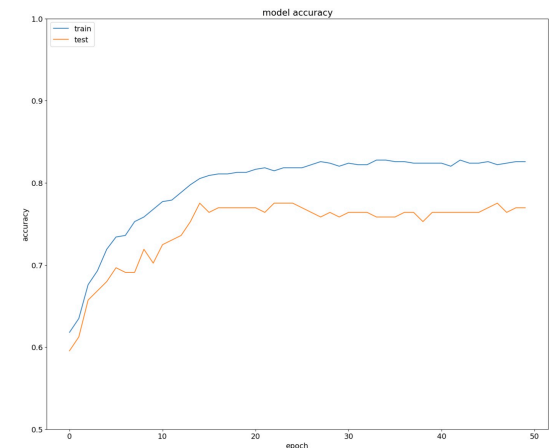
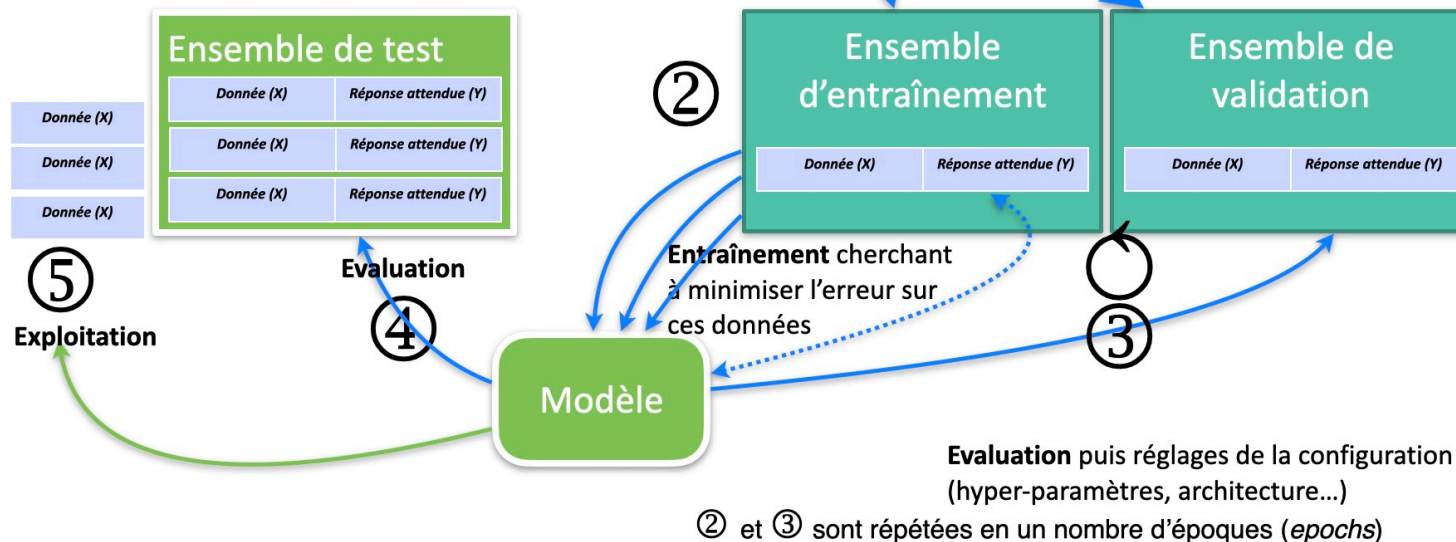
- Analyse manuelle des textes : l'humain détermine les règles qui formalisent les phénomènes linguistiques
- On écrit des ressources = par ex. des grammaires sous forme d'automates ou des requêtes sous forme d'expressions régulières
- La machine compile les automates et les applique à de nouveaux textes (analyse de corpus et de discours, annotation ou désambiguïsation...)

Eric Laporte. Symbolic Natural Language Processing. Lothaire. Applied Combinatorics on Words, Cambridge University Press, pp.164-209, 2005. hal-00145253

# Classification automatique de documents par apprentissage machine supervisé

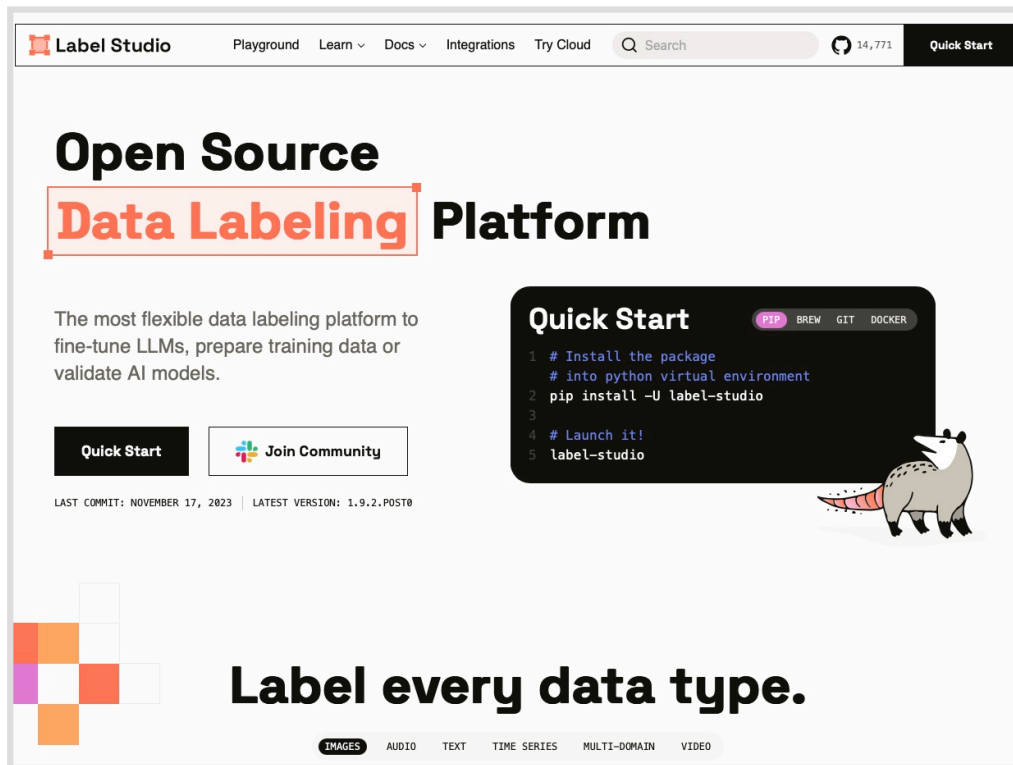
Ensemble d'exemples			
Donnée (X)	Réponse attendue (Y)	Donnée (X)	Réponse attendue (Y)
Donnée (X)	Réponse attendue (Y)	Donnée (X)	Réponse attendue (Y)
Donnée (X)	Réponse attendue (Y)	Donnée (X)	Réponse attendue (Y)

- On fournit des exemples résolus (données d'entraînement, de validation et de test)
- La machine apprend un modèle des données (représentation) et un modèle de la tâche à réaliser. Plusieurs approches sont possibles : arbres, neurones...
- Problèmes : combien d'exemples, quels exemples (variété des exemples et biais), capacité de généralisation sur des données nouvelles ?*



# Logiciels d'annotation

<https://labelstud.io/>  
<https://gguibon.github.io/ezcat/#/>



**Label Studio** Playground Learn Docs Integrations Try Cloud Search 14,771 Quick Start

## Open Source Data Labeling Platform

The most flexible data labeling platform to fine-tune LLMs, prepare training data or validate AI models.

**Quick Start** **Join Community**

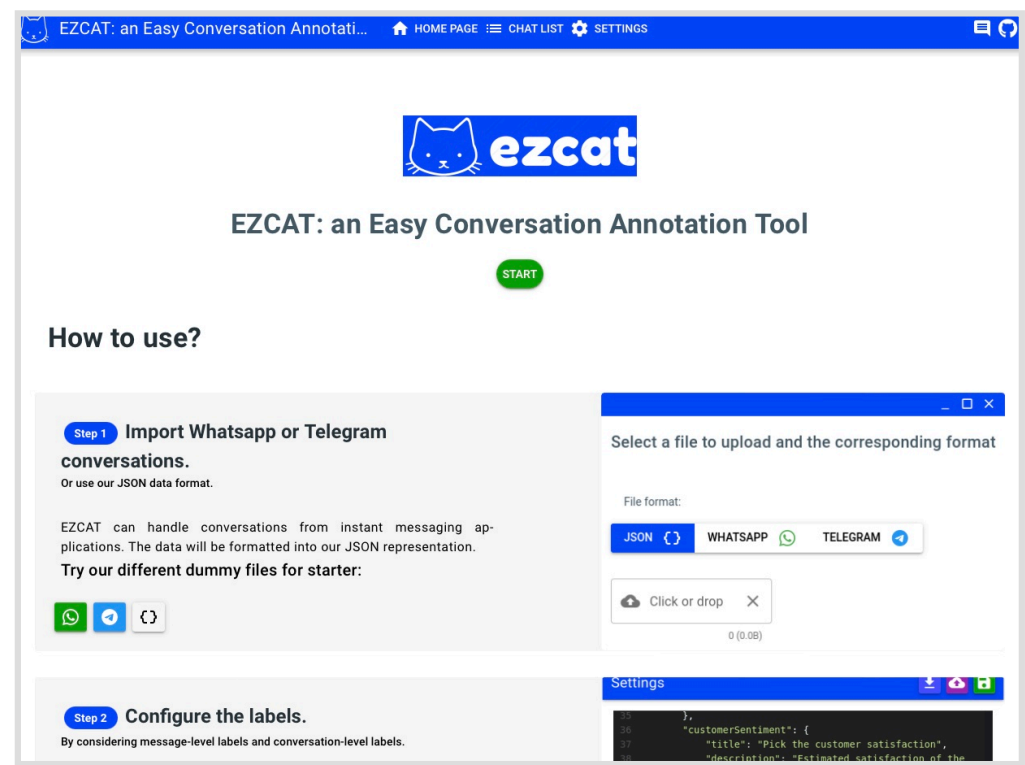
LAST COMMIT: NOVEMBER 17, 2023 | LATEST VERSION: 1.9.2.POST0

### Quick Start

1 # Install the package  
2 # into python virtual environment  
3 pip install -U label-studio  
4 # Launch it!  
5 label-studio

**Label every data type.**

IMAGES AUDIO TEXT TIME SERIES MULTI-DOMAIN VIDEO



**EZCAT: an Easy Conversation Annotation Tool**

**START**

### How to use?

**Step 1 Import Whatsapp or Telegram conversations.**  
Or use our JSON data format.

EZCAT can handle conversations from instant messaging applications. The data will be formatted into our JSON representation.  
Try our different dummy files for starter:

**Step 2 Configure the labels.**  
By considering message-level labels and conversation-level labels.

Select a file to upload and the corresponding format

File format:  
**JSON** WHATSAPP TELEGRAM

Click or drop 0 (0.0B)

**Settings**

```
},  
"customerSentiment": {  
  "title": "Pick the customer satisfaction",  
  "description": "Estimated satisfaction of the"
```

# Classification de textes par arbres de décision

ARFF-Viewer - /Users/Patrice/PycharmProjects/ANF2021/ANF/CorpusWekaResumes.csv

File Edit View

CorpusWekaResumes.csv

Relation: CorpusWekaResumes

No. 1: Resume (Nominal) 2: Categories (Nominal)

1 Structural biology is making significant contributions toward an understanding of molecular constituents and mechanisms underlying huma... cell  
2 The threat of infection by conventional transfusion-transmitted agents has been essentially eliminated from the blood supply in develop... hematology  
3 Not science fiction, but a technically feasible plan to probe our planet's inner workings. multidisciplinary  
4 Severe acute respiratory syndrome coronavirus (SARS-CoV) is the etiological agent of a newly emerged disease SARS. The SARS-CoV nucle... chemistry  
5 Objective: To understand the association between the SARS outbreak and the environmental temperature, and to provide a scientific basi... public  
6 For clinical diagnosis, a small number of targets (2-10 biomarkers) are often all that is required for disease assessment and accurate ear... chemistry  
7 Bacterial storage lipids including poly(hydroxyalkanoates), triacylglycerols and wax esters are biodegradable materials with applications i... microbiology  
8 The development of glycan arrays has enabled the high-sensitivity and high-throughput analysis of carbohydrate-protein interactions and ... chemistry

Weka GUI Chooser

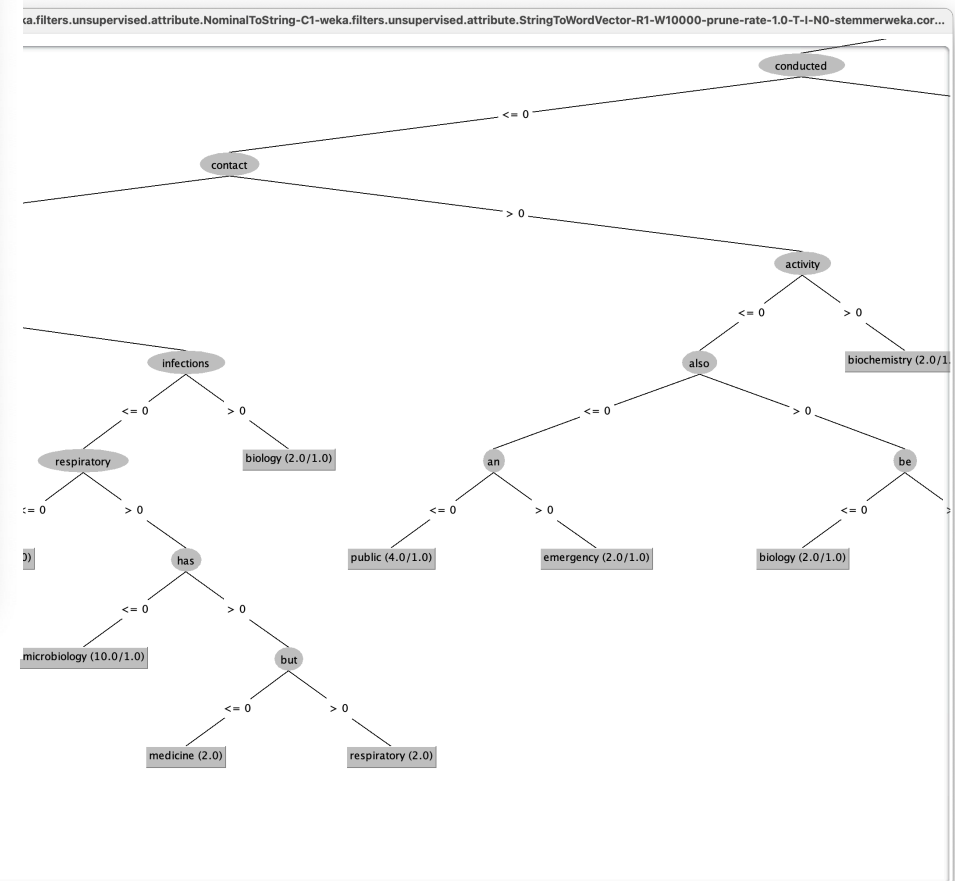
Program Visualization Tools Help

Package manager ^+U  
ArffViewer ^+A  
SqIViewer ^+S  
Bayes net editor ^+N

Applications

Explorer  
Experimenter  
KnowledgeFlow  
Workbench  
Simple CLI

Waikato Environment for Knowledge Analysis  
Version 3.8.5  
(c) 1999 - 2020  
The University of Waikato  
Hamilton, New Zealand



# Classification « neuronale » appliquée à l'analyse de sentiments (polarité)

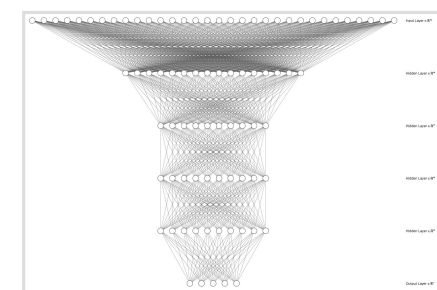
<https://github.com/pbellot/ANF-TDM>

This is a complex film that explores the effects of Fordist and Taylorist modes of industrial capitalist production on human relations. There are constant references to assembly line production, where workers are treated as cogs in a machine, overseen by managers wielding clipboards, controlling how much the workers leave exposed, and firing workers (Stanley) who meet all criteria (as his supervisor says, are always on time, are hard workers, do good work) but who may in some unspecified future make a mistake. This system destroys families - Stanley has to send his father to a nursing home (where he quickly dies) after Stanley loses his job. Iris' daughter is a single teen mother who drops out of high school to take a job in the plant. References are made to the fact that now, with declining wages, both partners need to work, the implication being that there's nobody left at home to care for the kids. Iris' husband is dead from an illness, and with the multiple references in the film about the costs of medical care, the viewer must wonder if he might have lived with better and more costly care. Iris' brother in law gets abusive after yet another unsuccessful day at the unemployment office when his wife yells at him for buying a beer with her savings instead of leaving it for her face lift and/or teeth job (even the working class with no stake in conventional bourgeois notions of perfection and beauty buy into them). The one reference to race in the film is through a black factory line worker who's husband is in jail (presumably, he's also black, and black men suffer disproportionately high incarceration rates). She remarks that he, like her, "is doing time" - her family is composed of a prisoner and a wage slave. Stanley, however, still believes in human relations and is therefore for most of the film outside of the system of Fordist capitalism. He cares for his father in spite of the fact that it was his father's traveling salesman job that resulted in his illiteracy - he has not yet reduced human relations to a purely instrumental contract, as Iris' brother in law does (suggesting that he married the wrong sister"). He does not, as Iris says, conform to the work-eat-sleep routine of everyone else; rather, he uses technology and the techniques of industrial production in an artisanal and creative way, in a sort of Bauhaus ideal. This was the dream of early modernists and 1920's socialists.

Exemple de critique de film dont il faut déterminer la polarité

Document + Modèle de langue

Représentation vectorielle du document



Réseau de neurones

## Résultats (% exactitude)

- Approche bayésienne « naïve » : **0,85** (temps de calcul : quelques secondes)
- Approche neuronale « plongements + couches denses » :
  - après quelques réglages et essais : **0,80** (18 epochs)
  - temps d'apprentissage : avec CPU seul 12 cœurs : environ 10s / epoch, soit 3 mn, avec GPU : environ 8s. / epoch soit ~ 2 mn
- Approche neuronale « plongements + réseaux récurrents »
  - meilleurs scores : **0,88 — 0,9** (20 epochs) (soit 3-5% de gain)
  - temps d'apprentissage : avec CPU seul : environ 3000 s. / epoch, soit > 24 h., avec GPU : environ 200 s. / epoch soit ~ 1h

# La question du format des documents

## Le cas idéal : les textes sont en XML

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0"
  xmlns:ns1="http://standoff.proposal"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:noNamespaceSchemaLocation="https://xml-schema.delivery.istex.fr/formats/tei-istex.xsd">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title level="a" type="main" xml:lang="en">The design and realization of CoViD: a system for
          design</title>
      </titleStmt>
      <publicationStmt>
        <authority>ISTEX</authority>
        <publisher ref="https://scientific-publisher.data.istex.fr/ark:/67375/H02-SWLMH5L1-1">Spring
        <pubPlace>London</pubPlace>
        <availability>
          <licence>Springer-Verlag London Limited</licence>
          <p scheme="https://loaded-corpus.data.istex.fr/ark:/67375/XBH-3XSW68JL-F">springer</p>
        </availability>
        <date type="published" when="2006">2006</date>
      </publicationStmt>
      <notesStmt>
        <note type="content-type"
          subtype="research-article"
          source="OriginalPaper"
          scheme="https://content-type.data.istex.fr/ark:/67375/XTP-1JC4F85T-7">research-article
        <note type="publication-type"
          subtype="journal"
          scheme="https://publication-type.data.istex.fr/ark:/67375/JMC-5WTPMB5N-F">journal</note>
      </notesStmt>
      <sourceDesc>
        <biblStruct>
          <analytic>
            <title level="a" type="main" xml:lang="en">The design and realization of CoViD: a syst
              design</title>
            <author role="corresp">
              <persName>
                <forename type="first">Wolfgang</forename>
                <surname>Stuerzlinger</surname>
              </persName>
              <affiliation>
                <orgName type="institution">York University</orgName>
                <address>
                  <settlement>Toronto</settlement>
                  <country key="CA" xml:lang="en">CANADA</country>
                </address>
              </affiliation>
            </author>
          </analytic>
        </biblStruct>
      </sourceDesc>
    </fileDesc>
  </teiHeader>

```



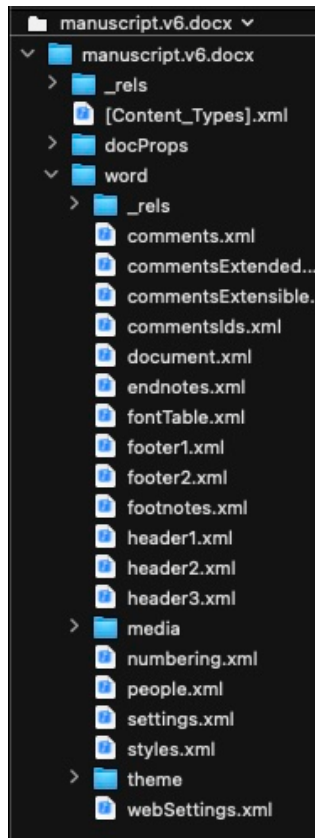
VS

Abstract Many important decisions in the design process are made during fairly early on, after designers have presented initial concepts. In many domains, these concepts are already realized as 3D digital models. Then, in a meeting, the stakeholders for the project get together and evaluate these potential solutions. Frequently, the participants in this meeting want to interactively modify the proposed 3D designs to explore the design space better. Today's systems and tools do not support this, as computer systems typically support only a single user and computer-aided design tools require significant training. This paper presents the design of a new system to facilitate a collaborative 3D design process. First, we discuss a set of guidelines which have been introduced by others and that are relevant to collaborative 3D design systems. Then, we introduce the new system, which consists of two main parts. The first part is an easy-to-use conceptual 3D design tool that can be used productively even by naive users. The tool provides novel interaction techniques that support important properties of conceptual design. The user interface is non-obtrusive, easy-to-learn, and supports rapid creation and modification of 3D models. The second part is a novel infrastructure for collaborative work, which offers a semi-immersive

W. Stuerzlinger (& amp;) L. Zaman A. Pavlovych  
York University, Toronto, Canada  
URL: <http://www.cs.yorku.ca/~wolfgang>  
URL: <http://www.cs.yorku.ca/~zaman>  
URL: <http://www.cs.yorku.ca/~andriyp>  
J.-Y. Oh  
University of Arizona, Tucson, AZ, USA  
e-mail: [jyoh@optics.arizona.edu](mailto:jyoh@optics.arizona.edu)

setup. It is designed to support multiple users working together. This infrastructure also includes novel pointing devices that work both as a stylus and a remote pointing device. This collaborative infrastructure forms a new platform for collaborative virtual 3D design. Then, we present our system against the guidelines for collaborative 3D design. Finally, we present results from a task on the new system. Keywords Collaborative design 3D design Collaborative virtual reality

## Les documents Microsoft Word et OpenOffice (.docx et .odt) : plusieurs XML



```
w:val="q4iawc"/><w:color w:val="auto"/><w:lang w:val="en"/></w:rPr></w:pPr><w:r  
w:rsidRPr="00E42DD1"><w:t xml:space="preserve">The variety and diversity of  
published content are currently expanding in all areas of science, with the  
simultaneous growth of interdisciplinarity. Powerful new tools and new technical  
infrastructures such as scientific knowledge graphs (SKG) have been developed  
[1], to help users navigate the flood of scientific information. However, the  
search experience requires more precision because.</w:t></w:r><w:r  
w:rsidRPr="00E42DD1"><w:rPr><w:rStyle w:val="q4iawc"/><w:color  
w:val="auto"/><w:lang w:val="en"/></w:rPr><w:t xml:space="preserve">retrieval  
systems do not benefit from a rich panel of content descriptors, in the context  
of model-based information retrieval allowing personalized answers to queries.  
At the same time, captured queries are rich in a.</w:t></w:r><w:r  
w:rsidRPr="00E42DD1"><w:rPr><w:rStyle w:val="q4iawc"/><w:color  
w:val="auto"/><w:lang  
w:val="en"/></w:rPr><w:lastRenderedPageBreak/><w:t>knowledge manifold that could  
be exploited to the benefit of a more personalized and efficient search. To  
achieve this result in the future we explore in this article three conditions:  
the first condition is to design a "cognitive community" to represent on  
knowledge graphs all the cliques of users of the same keyword; the second  
condition is to model, inside each community, a classifier of the interacting  
cliques, specifying each possible type of documentary need of the users of  
available items; the third condition is to optimize the efficient information  
use, by allowing all users of a keyword to access the mapping of all registered  
cliques to help them, if necessary, to refine or modify their  
choice.</w:t></w:r></w:p><w:p w14:paraId="544B7EB5" w14:textId="0B15CCDC"  
w:rsidR="00EB4121" w:rsidRPr="00E42DD1" w:rsidRDefault="00EB4121"
```

## Un cas plus difficiles : documents PDF

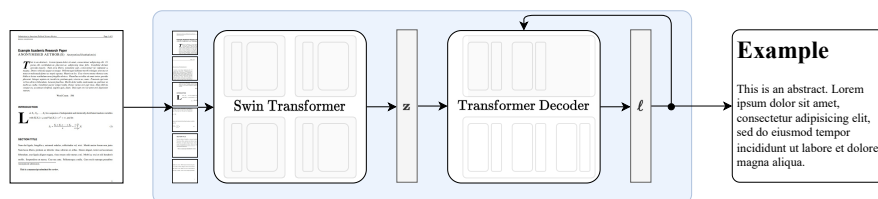
## GROBID (GeneRation Of Bibliographic Data)

extraction des titres et structuration des références bibliographiques  
(approches *deeplearning* ou CRF) + extraction du texte et de sa structure

## NOUGAT “Neural Optical Understanding for Academic Documents”

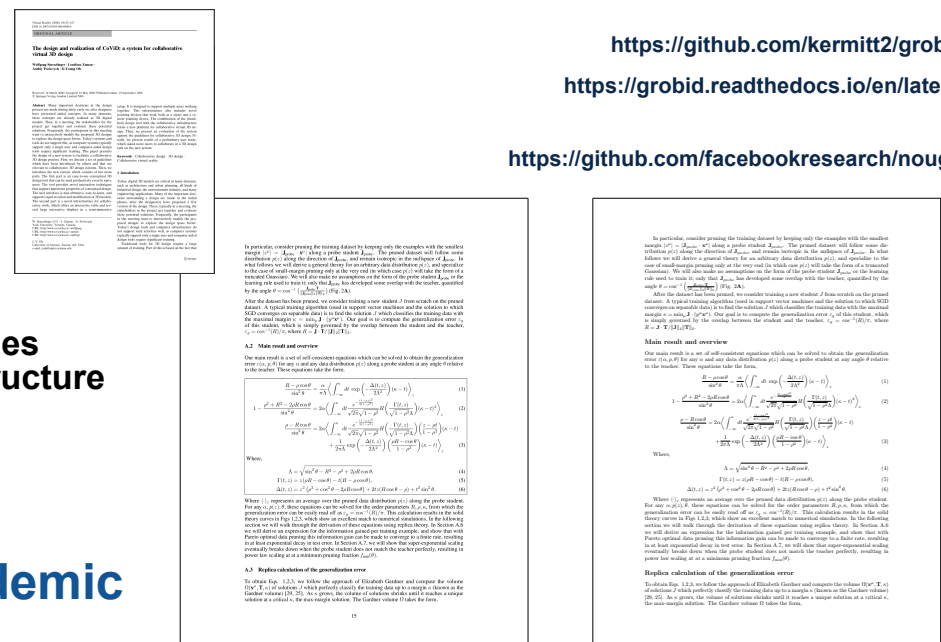
<https://arxiv.org/pdf/2308.13418.pdf>

créer des documents XML à partir de fichiers PDF sans étape d'OCR  
(architecture encodeur-décodeur *deep learning*, LLM BART)



**Figure 1:** Our simple end-to-end architecture following Donut [28]. The Swin Transformer encoder takes a document image and converts it into latent embeddings, which are subsequently converted to a sequence of tokens in a autoregressive manner

Visual Document Understanding (VDU)



**Figure 5:** Example of a page with many mathematical equations taken from [41]. Left: Image of a page in the document, Right: Model output converted to LaTeX and rendered back into a PDF. Examples of scanned documents can be found in the appendix B.

Method	Modality	Edit distance ↓	BLEU ↑	METEOR ↑	Precision ↑	Recall ↑	F1 ↑
PDF	All	0.255	65.8	82.1	77.1	81.4	79.2
GROBID	All	0.312	55.6	71.9	74.0	72.1	73.0
+ LaTeX OCR	Tables	0.626	25.1	64.5	61.4	80.7	69.7
	Plain text	0.363	57.4	69.2	82.1	70.5	75.9
	Math	0.727	0.3	5.0	11.0	8.6	9.7
Nougat small (250M*)	All	0.073	88.9	92.8	<b>93.6</b>	92.2	92.9
	Tables	0.220	68.5	78.6	75.0	79.8	77.3
	Plain text	0.058	91.0	94.3	96.1	95.3	95.7
Nougat base (350M*)	Math	0.117	56.0	74.7	77.1	76.8	76.9
	All	<b>0.071</b>	<b>89.1</b>	<b>93.0</b>	93.5	<b>92.8</b>	<b>93.1</b>
	Tables	0.211	69.7	79.1	75.4	80.7	78.0
	Plain text	0.058	91.2	94.6	96.2	95.3	95.7
	Math	0.128	56.9	75.4	76.5	76.6	76.5

**Table 1:** Results on arXiv test set. PDF is the text embedded in the PDF file. The modality “All” refers to the output text without any splitting. \*Number of parameters.

# Si les documents sont des images : OCR et traitement d'images

<https://github.com/mindee/doctr>

<https://github.com/tesseract-ocr/tesseract>

<https://cloud.google.com/vision/docs?hl=fr>

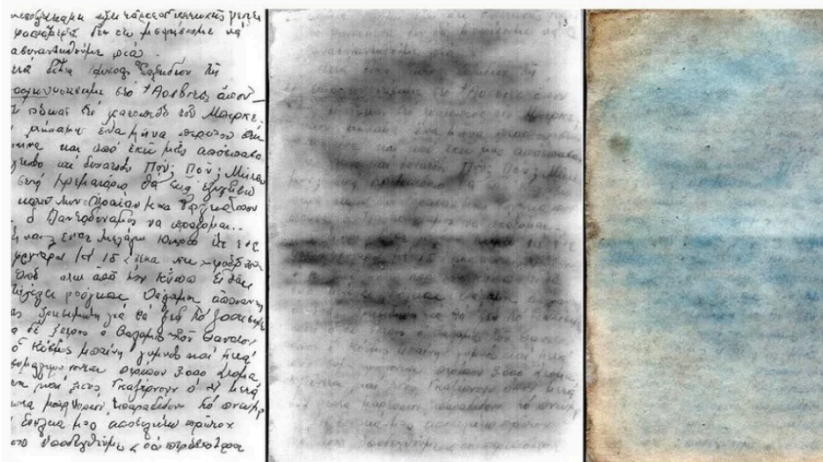
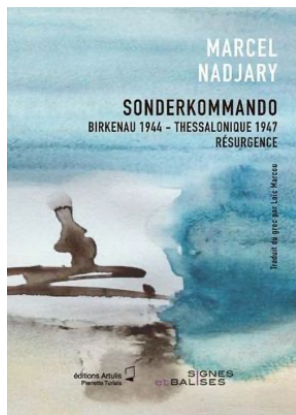
## Pour les formulaires, factures, textes sur photos... :

logiciels *open source* : Tesseract, docTR pour Python... (traitement « ligne à ligne »)

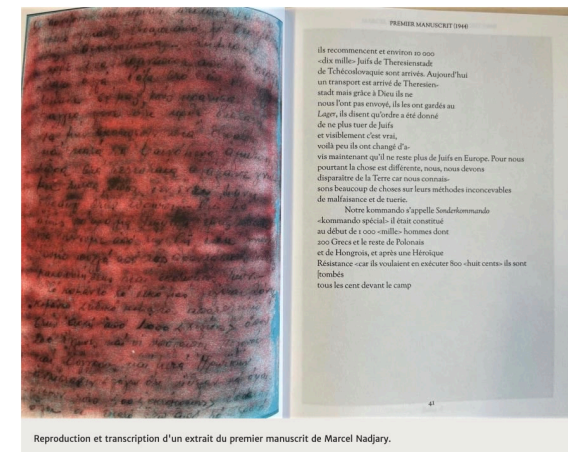
## Pour l'écriture manuscrite :

payants : Google API Vision, Microsoft Ignite, Amazon Textract, Abby FineReader...

## Pour les documents dégradés : traitement d'image avant OCR (par ex. imagerie multispectrale)



Seuls les progrès de l'optique et une caméra à haute technologie ont permis de déchiffrer, à partir des années 2010, le manuscrit de 1944 enfoui dans le sol d'Auschwitz. - Signes et Balises, édition française de Marcel Nadjary, 2023.



Reproduction et transcription d'un extrait du premier manuscrit de Marcel Nadjary.

Pavel Polian & Alexander Nikitjaev (2019) Deciphering a Mystery: Digital Technology and the Revelation of Handwritten Texts by Marcel Nadjari and Other Members of the Jewish Sonderkommando in Auschwitz-Birkenau, East European Jewish Affairs, 49:3, 220-229, DOI: 10.1080/13501674.2019.1721815

Journée science ouverte CNRS — Panorama Fouille de textes (P. Bellot)

## Conclusion : la fouille de textes...

### Nécessite :

- un corpus cible, un choix d'approches et de méthodologies
- d'intégrer et d'exploiter différents composants logiciels, modèles, APIs
- un scénario et une référence pour apprendre des modèles et évaluer des résultats

### Faisable si :

- les composants et données sont interopérables,  
les métadonnées compatibles, les ressources accessibles  
(principes FAIR)

La fouille de textes et de données à des fins de recherche : une pratique confirmée et désormais opérationnelle en droit français



L'ordonnance du 24.11.2021 *permet de reproduire des contenus protégés [sous réserve de les obtenir par moyen licite] par des droits de propriété intellectuelle dans le but de conduire des activités de fouille à des fins de recherche scientifique, sans avoir à recueillir d'autorisation préalable*

# Action Nationale de Formation | ANF TDM 2021 à 2023

## Exploration documentaire et extraction d'informations

<https://anf-tdm-2023.sciencesconf.org/>

<https://anf-tdm-2022.sciencesconf.org/>

<https://anf-tdm-2021.sciencesconf.org/>



### ANF TDM 2023 | Exploration documentaire et extraction d'informations

12-13 oct. 2023 Villejuif (France)

#### NAVIGATION

Accueil

Programme

Jour 1 (PM) - Conférence ▾

Jour 2 (AM) - Ateliers ▾

Jour 2 (PM) - Cas usage

Chaîne vidéo

#### SUPPORT

@ Contact

#### PRÉSENTATION

La production scientifique s'accroît chaque année avec un taux de croissance des articles publiés compris entre 5 et 6,5% pour l'année 2021. D'après le rapport du syndicat des éditeurs, la base de données Dimensions a enregistré plus de **4,7 millions d'articles publiés en 2020** dans plus de **35 000 journaux** d'éditeurs scientifiques.

Cette augmentation du volume de l'information produite par la communauté scientifique devient difficile à suivre par les chercheurs et chercheuses dans toutes les disciplines. L'apport massif de données et de publications, associé à la multiplication des canaux de diffusion, complexifie la veille et l'exploration de la littérature scientifique.

#### Objectif et profil

Cette formation nationale invite les communautés scientifiques à **exploiter les techniques numériques de la recherche d'information** et à **développer la fouille de textes de données dans leur activité professionnelle**.

Elle s'adresse aux **chercheurs et chercheuses, doctorant(e)s et ingénieur(e)s d'appui à la recherche** qui souhaitent se former aux techniques numériques pour mettre en place ou développer la recherche d'information et la fouille de textes dans leur activité professionnelle.

A la suite de cette formation, les participant(e)s pourront **identifier la nature du besoin**, sélectionner une **méthodologie logicielle adaptée** et réutiliser des outils et services en fonction des objectifs visés.

#### Lien d'inscription

Si vous **souhaitez vous perfectionner** ou si vous êtes **appelés à utiliser des logiciels numériques pour la recherche d'information**, il n'est pas nécessaire de savoir manipuler des bases de données ou d'avoir utilisé des logiciels de visualisation en amont. Néanmoins, une connaissance des enjeux et des méthodes de fouille de textes est fortement recommandée.

Par ailleurs, un **temps d'échanges en distanciel** sera organisé au mois de septembre pour **présenter les quatre logiciels** utilisés dans les ateliers pratiques.

Les participants devront **apporter leur matériel informatique** et certains logiciels nécessiteront d'être installés en amont de la formation. Pour les agents CNRS, les frais de transport et d'hébergement sont pris en charge par le service formation de leur délégation.

#### Highlight édition 2022

#### INFORMATIONS PRATIQUES

##### Dates

Jeudi 12 octobre 2023  
au vendredi 13 octobre 2023

##### Lieu

CNRS Délégation Villejuif  
7 Rue Guy Môquet,  
94800 Villejuif

[Lien itinéraire](#)

##### Accessibilité

Métro ligne 7  
(Arret Villejuif Paul Vaillant-Couturi)

Bus n°131  
(Arret Institut Gustave Roussy)

##### Vidéotheque

[Highlight 2022](#)  
[Interviews vidéos](#)  
[Focus logiciels](#)

En partenariat avec

**INRAE**

## Contacts, liens, exemples

# Ressources nationales et européennes

<https://www.ortolang.fr/>

<https://www.clarin.eu/>

<https://www.istex.fr/services-recherche/>

The screenshot shows the Ortolang website homepage. At the top, there is a navigation bar with links: Accueil, Ressources, Hébergement, Aide, and a language selector for English. The main header features the Ortolang logo and the tagline: "Plate-forme d'outils et de ressources linguistiques pour un traitement optimisé de la langue française". Below this, there is a search bar with the placeholder text "Chercher une ressource". The page is divided into three main sections: "DÉPOSER UNE RESSOURCE" (with a cloud upload icon and a description of the simple registration process), "EXPLORER LES RESSOURCES" (with a magnifying glass icon and a description of the available linguistic resources), and "HÉBERGER SON" (with a server rack icon). Each section has an orange button labeled "DÉPOSER", "EXPLORER", and "HÉBERGER" respectively.

The screenshot shows the ISTEX website page titled "Services à la recherche". The header includes the ISTEX logo and the tagline "Le socle de la bibliothèque scientifique numérique nationale". The navigation bar contains links: Base documentaire, Constitution de corpus, Services à la recherche (highlighted), Offre de formations, and Institutions adhérentes. The main content area has a green background with the title "Services à la recherche" and the subtitle "Les technologies et les outils de l'infrastructure Istex au service de votre projet de fouille de textes sur vos propres données". Below this, there is a section titled "Services à la recherche" with a breadcrumb trail "Accueil → Services à la recherche". This section lists "Services TDM" (Ressources terminologiques, Hébergement de corpus) and "Accès rapide aux services". It also features three callout boxes: "Contenu et accès aux publications intégrées dans votre espace de travail", "Démarrer votre projet de fouille de texte en constituant votre corpus à l'aide des ressources Istex", and "Les technologies et les outils de l'infrastructure Istex au service de votre projet de fouille de textes sur vos propres données".

## The research infrastructure for language as social and cultural data

CLARIN is a digital infrastructure offering data, tools and services to support research based on language resources.

## Le consortium Huma-Num « CORLI »



The screenshot shows the homepage of the CORLI website. At the top, there is a dark blue header with the CORLI logo (stylized orange and blue letters) and the text 'Consortium HN CORpus, Langues et Interactions'. Below this, there are logos for Huma-Num, CLARIN K CENTRE, and CNRS. A navigation bar with a dark blue background contains the following links: Accueil (highlighted in orange), Centre K CLARIN, Organisation, Actions, Réseau, Formations, Bonnes pratiques, Ressources, and Nous rejoindre, followed by a search icon. The main content area has a light blue background and contains the following text:

Bienvenue sur le site du consortium Huma-Num « CORLI »

CORLI est un réseau de laboratoires et de chercheurs travaillant sur les corpus de langage. Son but est d'offrir à tous des données, des outils, de la documentation et des formations autour de l'utilisation scientifique des corpus de langage en suivant les principes FAIR. CORLI est ouvert aux enseignants-chercheurs, chercheurs, ingénieurs, du monde entier et à l'étude de toutes les langues et le démontre particulièrement en étant labellisé Centre K par l'infrastructure européenne CLARIN.

Le consortium CORLI s'attache particulièrement à mettre en commun les ressources existantes des laboratoires du consortium, à aider la communauté scientifique à pérenniser ses données, à mieux les diffuser, et à donner des moyens complémentaires à des projets de réseau validés par les chercheurs en linguistique de corpus.

Plus particulièrement, CORLI propose:

- de la documentation sur les outils, les formats, les bonnes pratiques, les aspects juridiques
- des formations à l'utilisation des outils ou des formats
- une aide aux utilisateurs dans notre centre K.

CORLI travaille activement sur [trois projets qui répondent aux besoins de la communauté](#):

- l'annotation collaborative
- la citation de corpus ou d'extraits de corpus
- le Corpus Ouvert du Français: données de corpus et outils pour la langue française

# Le GDR TAL (Traitement Automatique des Langues)

<https://gdr-tal.ls2n.fr/>

**GDR** Groupement  
de recherche  
**TAL** Traitement automatique  
des langues

**cnrs**  
**SCIENCES  
INFORMATIQUES**

GDR TAL

Accueil

GdR TAL ▾

Actualités ▾

Doctorants ▾

Laboratoire

## Accueil

Le GDR TAL s'intéresse à la langue sous toutes ses formes : écrite, orale, signée. Il aborde les thématiques de la modélisation informatique de la langue, ses manifestations et ses applications. Les communautés centrales du GDR TAL sont celles du traitement automatique des langues, du traitement automatique du langage parlé et de la recherche d'informations. Les communautés secondaires sont le traitement automatique du document, le web sémantique, les neurosciences et les sciences cognitives, et toutes les communautés de recherche où la question linguistique est d'importance.

Le TAL a un fort impact sociétal au cœur des développements en sciences des données avec des applications dans le médical, l'éducation, le droit, le journalisme, le handicap.

Le but du GDR TAL est de conduire une prospection scientifique sur les grands enjeux du TAL, de réfléchir à l'animation scientifique de la communauté en vue d'améliorer sa stratégie scientifique, son attractivité et sa visibilité. Il vise aussi à structurer les activités de recherche des équipes des unités CNRS. Il définira son périmètre, ses interactions et coopération avec les sociétés savantes des domaines qu'il couvre : ATALA, l'association pour le traitement automatique des langues, ARIA, l'association francophone de Recherche d'Information et Applications et AFCP, l'association francophone de la communication parlée, ainsi que ses interfaces avec les autres communautés scientifiques de l'informatique représentées par les GDR RADIA, MADICS, ISIS, RO et avec l'INSHS. Il a aussi l'objectif d'être l'interlocuteur privilégié de l'INS2I ou de toutes instances nationales.

## Ses laboratoires

Groupe de Travail 1 –  
Apprentissage et modèles  
pour le TAL

Groupe de Travail 2 –  
Intermodalité,  
multimodalité

Groupe de Travail 3 –  
Multilinguisme, Multiplicité  
des langues

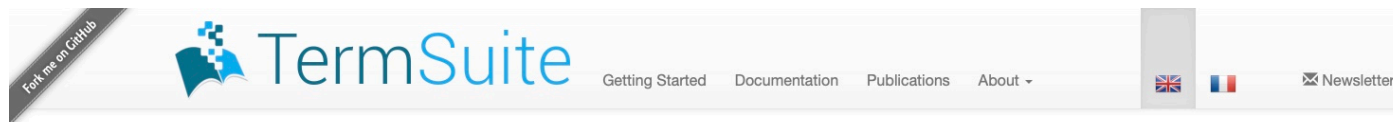
Groupe de Travail 4 – Accès à  
l'information et fouille de  
textes

Action 1 – Ressources



Journée science ouverte CNRS — Panorama Fouille de textes (P. Bellot)

# Extraction terminologique



TermSuite is a toolbox for **terminology extraction** and **multilingual term alignment**.

Multiword and compound term detection, morphosyntactic analysis, term variant detection, term specificity computation, etc. [See features](#)

Language Support



[Command Line](#) - [Graphical User Interface](#) - [Java API](#)



Current version of TermSuite is 3.0.10 [See Changelog](#)

Get it running !

Prepare your system for TermSuite, download, install and get it running on an example corpus quickly.

[Getting Started](#)

Documentation

List of all TermSuite's features, analysis engines, and configuration parameters. Java API.

[User Manual](#)

[Javadoc](#)

Developers

Build it from sources with Gradle, or use it as a maven dependency.

[View on Github](#)

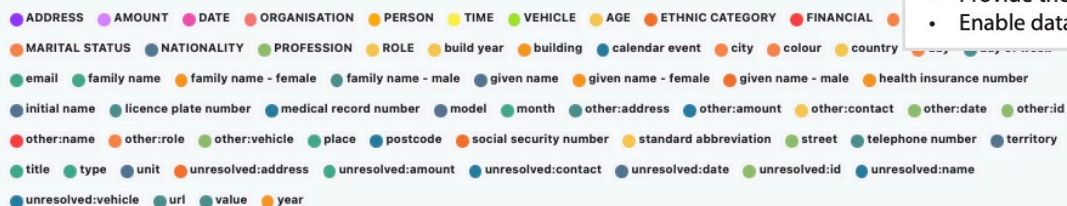
[Maven / Gradle](#)

#	type	p	pilot	spec	f			
1	T	NN	wind turbine	5,16	1852			
1	V	NNN	horizontal-axis wind turbines	3,52	42			
1	V	NNN	horizontal axis wind turbine	3,50	41			
1	V	NNN	vertical axis wind turbines	3,62	53			
1	V	NNN	modern horizontal-axis wind turbines	2,59	5			
1	V	NNN	smaller-scale wind turbines	2,20	2			
1	V	NNN	on-shore wind turbines	1,90	1			
1	V	NNN	pre-manufactured wind turbine	1,90	1			
1	V	NNN	repowred wind turbines	1,90	1			
1	V	NNN	conventional horizontal-axis wind turbines			1,90	1	
1	V	NNN	potential campus wind turbines	1,90	1			
1	V	NNN	typical horizontal-axis wind turbine	1,90	1			
1	V	NNN	unconventional horizontal-axis wind turbines			1,90	1	
1	V	NNN	hawts horizontal-axis wind turbines	1,90	1			
1	V	NNN	utility scale wind turbine	1,90	1			
1	V	NNN	lift type wind turbines	1,90	1			
1	V	NNN	domestic wind turbines	3,35	29			
1	V	NNN	wind turbine syndrome	3,21	21			
[...]								
2	T	N	rotor	4,82	848			
3	T	NN	wind energy	4,51	414			
3	V	NNN	californian wind energy	1,90	1			
3	V	NNN	offshore wind energy	3,56	47			
3	V	NNN	wind energy conversion	3,32	27			
3	V	NNN	wind energy conf	2,59	5			

# Anonymisation de textes



MAPA is a Connecting Europe Facility Action (2019-EU-IA-0013).  
Grant Agreement: INEA/CEF/ICT/A2019/1927065



## Interested in trying MAPA yourself?

MAPA is open-source and covered by an Apache-2 license. Click [here](https://github.com/INRIA/Mapa) to understand the clauses governing any use, re-use and results of MAPA.

You can download a copy of the project for your own purposes from the GitLab repository:  
[https://gitlab.com/MAPA-EU-Project/mapa\\_project](https://gitlab.com/MAPA-EU-Project/mapa_project)

## Enter or paste a text:

Enter a piece of text here and select the most suitable model to process it...

Available models: Multilingual

☒ Detect entities ☐ Obfuscate entities

Run it!

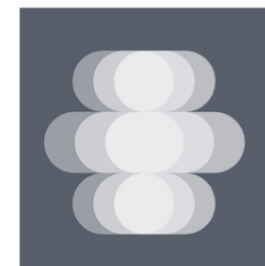
☐ Replace entities

# MAPA

## Multilingual Anonymization for Public Administrations

[www.mapa-project.eu](https://mapa-project.eu)

<https://mapa-project.eu/>



### OBJECTIVE

- Develop an open-source toolkit for effective and reliable text anonymisation in the medical and legal domains [de-identification of personal data]
- Provide the NLP tools needed for Public Administrations to be GDPR-compliant
- Enable data holders to free data that contains personally-identifiable information

## MAPA IS AN OPEN SOURCE PROJECT LED BY:



Partners:



SEAD  
(State Secretariat for Digital Advancement)



Centre National de la Recherche Scientifique

vicomtech  
your R&D partner for smart digital solutions

L-Università ta' Malta



# Mise en relation de pré-publications et de publications //

## Détection d'articles problématiques



<https://www.irit.fr/~Guillaume.Cabanac/covid19-preprint-tracker>

**COVID19 Preprint Tracker**  
Est. June 15<sup>th</sup>, 2020

This website tabulates Preprint-Publication links for a corpus of **1007 preprints related to COVID-19** curated by the Centre of Research in Epidemiology and Statistics and Cochrane France.

The software is being developed as part of [The COVID-NMA Project](#).

Harvesting and consolidating data from these APIs:

- ▶ bioRxiv
- ▶ Crossref
- ▶ Dimensions
- ▶ PubPeer

Data last refreshed on 10-OCT-2023 23:37:31.

[Quick presentation](#): problem tackled, method, and evaluation.

Stable URL for this website: <https://www.irit.fr/~Guillaume.Cabanac/covid19-preprint-tracker>

<https://www.irit.fr/~Guillaume.Cabanac/problematic-paper-screener>

**Problematic Paper Screener**  
Est. February 27<sup>th</sup>, 2021

Stable URL: <https://www.irit.fr/~Guillaume.Cabanac/problematic-paper-screener>

This website shows reports the daily screening of papers (partly) generated with:

- ▶ Automatic SBIR Proposal Generator
- ▶ Dada Engine
- ▶ Mathgen
- ▶ SCigen
- ▶ Tortured phrases
- ... and Citejacked papers 🔥

Harvesting data from these APIs:

- ▶ Crossref, now including the [Retraction Watch Database](#)
- ▶ Dimensions
- ▶ PubPeer

PPS in a nutshell: problem tackled and method used — video in English or in French.

Distinct problematic pap...  
Detectors: Annulled, Feet of Clay, Tortured, Suspect, Citejacked, Seek&Blastn, Journal Cases, SCigen, Mathgen, Problematic Cell Lines, SBIR