



Email in French,  
Natural Language  
Processing and

# Pêle-mél

Plateforme d'Evaluation, de Livraison et d'Exploration des méls

Platform of appraisal, delivery and  
exploration  
of the mails

project

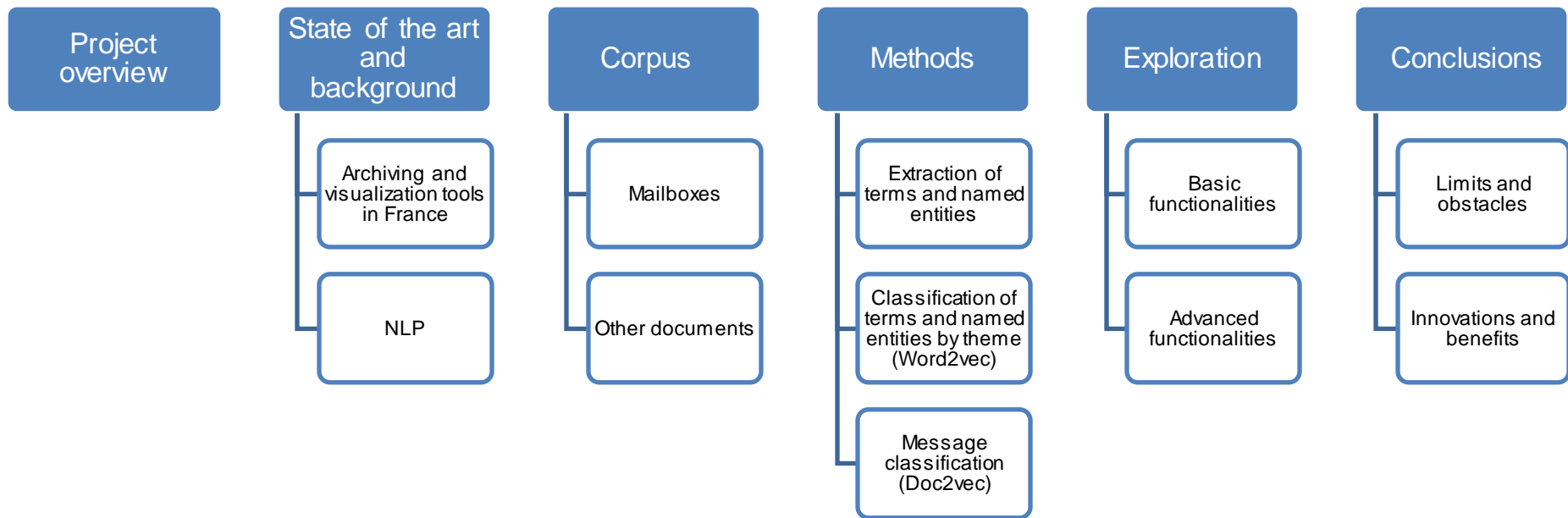
**Touria Ait el Mekki**  
**Bénédicte Grailles**



Soutenu par



# Plan





# Project overview

- Team
- Objectives

# Team & partners

## Team

- Angers University
  - Bénédicte Grailles (archival science, Temos)  
[benedicte.grailles@univ-angers.fr](mailto:benedicte.grailles@univ-angers.fr)
  - Touria Aït El Mekki (IT, Leria)  
[touria.aitelmekki@univ-angers.fr](mailto:touria.aitelmekki@univ-angers.fr)
    - Tsanta Randriatsitohaina (IT)
    - Chafik Akmouche (IT)
    - Taimane Zerez (IT)



## Partners

- Records and Archives office of the Ministry of Solidarity and Health
  - Anne Lambert (head of departement)
  - Chloé Moser (product manager Archifiltre)
- École nationale des Chartes
  - Edouard Vasseur (history and archival science, Jean-Mabillon center)



Temos (Temps, mondes, sociétés) is a Laboratory for Social Sciences joint with CNRS the French National Center for Scientific Research.

Leria is a Laboratory for Computer Science.

Project financed by the French Ministry of Culture in the framework of a call for projects "innovative digital cultural services"

Soutenu  
par



**MINISTÈRE  
DE LA CULTURE**

*Liberté  
Égalité  
Fraternité*

# Operational objectives



Testing different strategies for contextualizing boxes, correspondent networks and the information content of messages



Using Natural Language Processing (NLP) techniques adapted to French



Developing criteria to appraise the archival value of messaging systems and to help in the decision making process



Improving access to content



Developing prototype tools for exploring and visualizing electronic messages (beta version)

# Pêle-Mél

Platform of appraisal, delivery and exploration of the mails

Archiving

Natural  
language  
processing

Electronic mail



COOP. RÉG. des CH<sup>ts</sup> et DEUX-SÈVRES - SAINTES - Bureau du C



# Background and State of the art

- Concepts, practices and tools in France
- NLP

# Preliminary findings

Storage of mailboxes, for legal reasons and for a limited time, is within the production tools of these messaging service

The digital preservation of messaging systems = a blind spot in the collection of archives for heritage purposes

Two exceptions :

Mailboxes used by high-level management or expert functions

VIP mailboxes (special collections): writers for example

Few acquisitions : 3 %

Since the 2010s



# Evaluation and selection

Capstone approach validated by the interministerial electronic archiving program VITAM (2013)

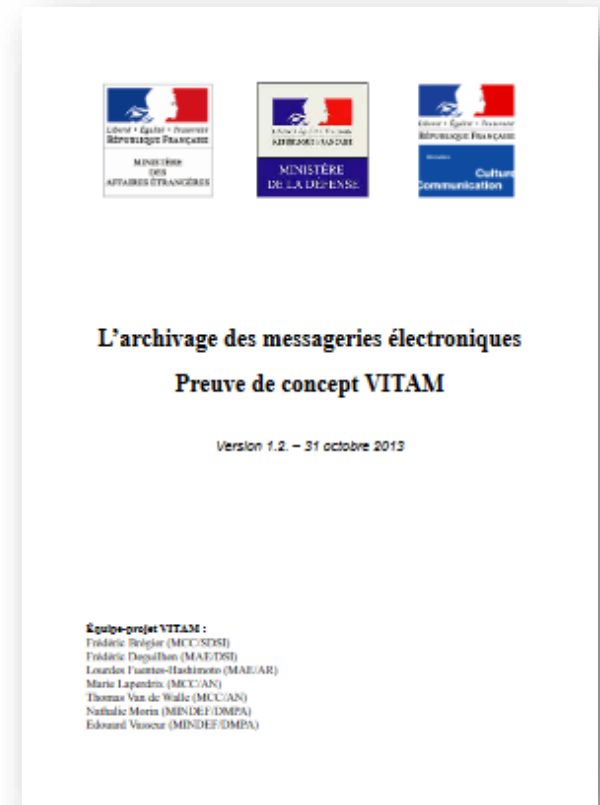
targeted collection from the e-mail addresses of key players in political and administrative decision-making processes (ministerial offices, directors, deputy directors)

Software testing (by a few of departments) for internal sorting :

- the ePADD tools (Stanford University)
- the RATOM tools (University of North Carolina)

Limited results

- Issue of French language (accented characters and cedillas)
- Individual mailbox

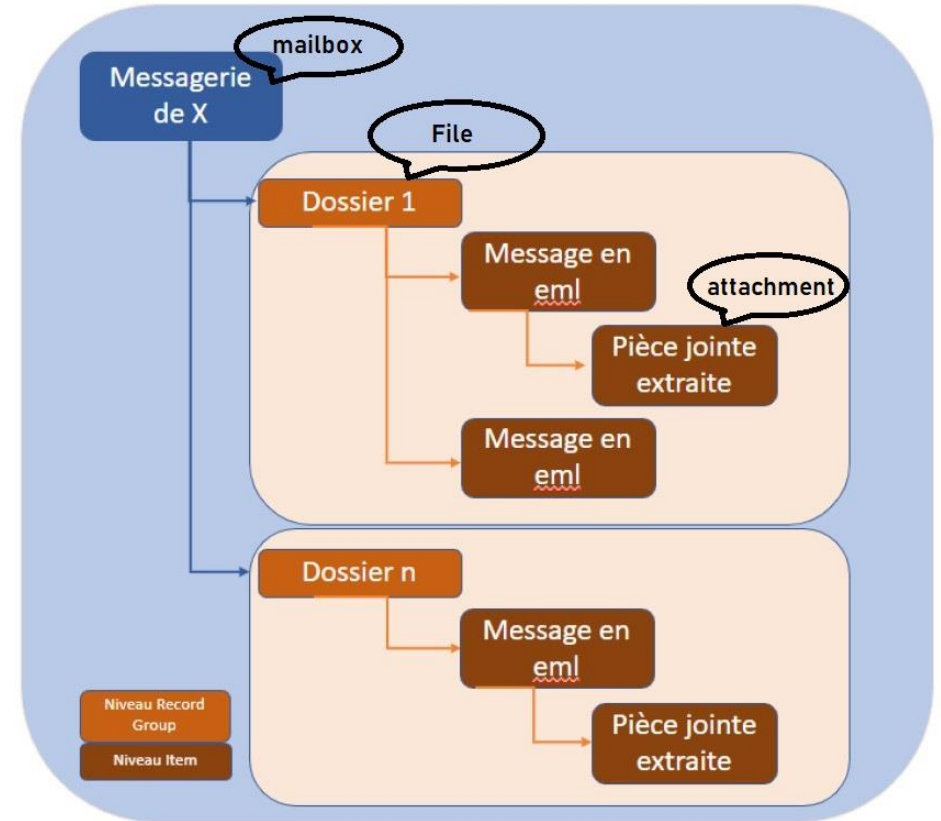


# French standards for capture

The interministerial electronic archiving program VITAM has developed a java Mail extract library, now available in the "SEDAlib toolbox"

- Extracting messages in eml format
- Each level of this hierarchy is provided with a Seda-compliant xml manifest (named ArchiveUnitMetadata)
- Preparing submission information packages

SEDA (data exchange standard for archiving) is a French standard transposed internationally (Depip, Data exchange protocol for interoperability and preservation, ISO 20614).



A#437-proiptc-2011	04/07/2022 16:57	Dossier de fichiers
_ArchiveUnitMetadata.xml	24/03/2021 15:43	XML
_BinaryMaster_1_-CB909FF5182D2C44B...	05/07/2022 21:48	Fichier EML

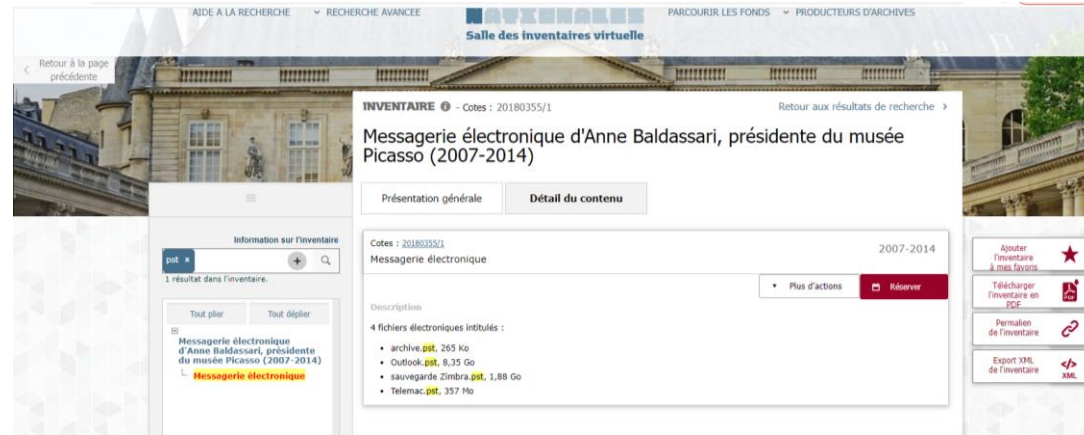
# Access to records and archives

Public services Mailboxes are "documents administratifs" and "archives publiques" (legal definition)

Anyone (judges, journalists, citizens, etc.) can ask to consult them.

- Example: requests for access to the Ministry of Health on subjects such as vaccination, masks, private consultancy firms, etc. and appeals to Commission for access to administrative documents

Today, however, access is not guaranteed because searches are too complex and time-consuming.



Mailbox descriptions (French National Archives) are "black boxes"

# Approaches for extraction of terms from corpus: term extraction and relation acquisition

Approaches based on traditional text-processing algorithms to analyze the sentences in a corpus and find the terms and relationships between them

==> good results, but all the rules and patterns must be defined beforehand.

Approaches based on artificial learning techniques using models trained on different corpora to identify terms and relationships between terms in a given corpus.

==> results depend on the amount of data and the degree of specialization of the trained models used

# Classification

- Approaches based on linguistic rules to classify messages:  
    limited by the complexity of the rules required.
- Machine learning-based approaches

Supervised classification :    require labeled training data

Unsupervised classification :

    Clustering (K-means)

    Dynamic Embedding /transformers (Bert )

    Word Embedding (Word2Vec,Doc2Vec)

# To conclude this state of the art

The mobilization of NLP methods is validated in English for the extraction of named entities

Knowledge extraction (who? what? when?) is possible using Artificial Neural Networks

Visualization improves approach and selection

There's an urgent need to find solutions to access problems, on pain of legal sanctions

But

1. Tools need to be adapted to the French language and script: accented characters, cedillas, representation of concepts and writing styles.
2. We need to be able to understand the networks that mailboxes create between themselves, so we can better appraise them.
3. We need to be able to respond to access requests: message content as well as attachments.
4. We therefore need to develop a strategy and relevant methods for extracting knowledge using artificial intelligence and Neural Networks.



Corpus

# Mailboxes overview



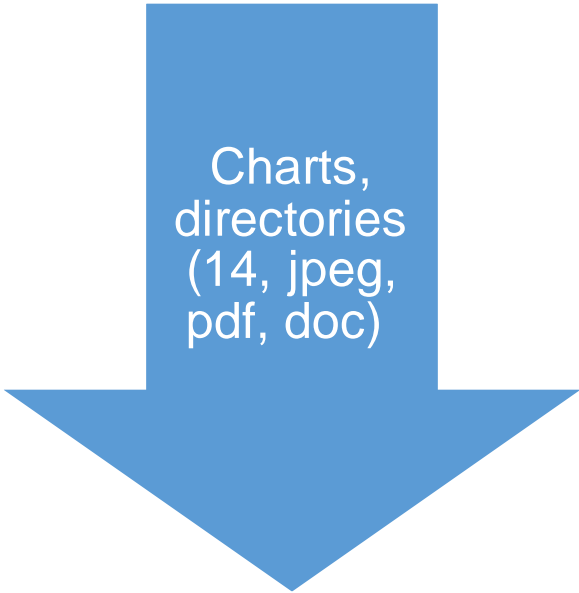


# Mailboxes overview



- Significant internal traffic
- Many distribution lists and functional addresses
- Volume of attachments and diversity of formats
- Well-written sentences, halfway between oral and written style (average number of words in a sentence = 15)

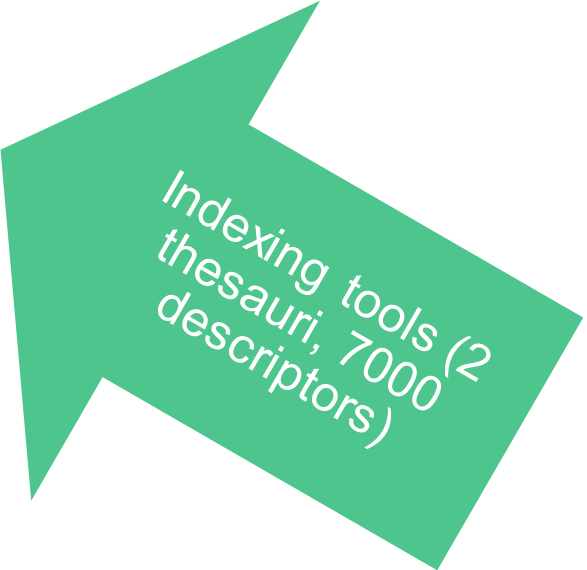
Other documents wanted : contextualization help



Charts,  
directories  
(14, jpeg,  
pdf, doc)



Minister's speech  
(810, doc, pdf)



Indexing tools (2  
thesauri, 7000  
descriptors)

# Methods



Represent each message unit  
[message + subject + attachment(s) + title]

Extract the list of terms and the list of named entities

Create term clouds by linking terms to each other  
and to named entities

Represent each theme with its term cloud with doc2vec

Find the most similar theme for each message

# Our approach to classifying messages

# Lexicon used

## Term

A term is a word or a group of words that univocally designates an object or a concept in a field (e.g. domestic violence, sexual violence, fight against violence etc.)

## Named entities

can refer to an object, a person, a location, a date or any other specific entities

## Semantic relationship

Link between two lexical entity (term, named entities)

## lexical-syntactic pattern

Ex.

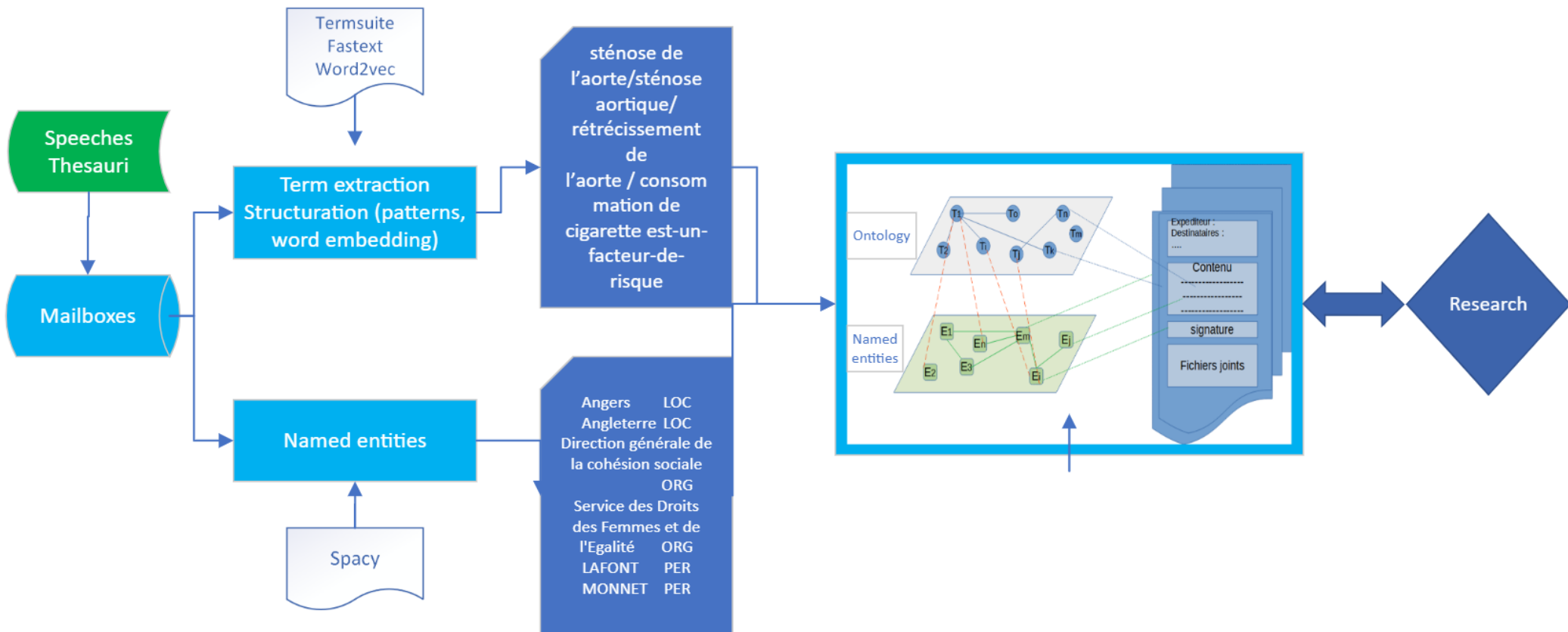
Term1 + be + det + Term2

Ex. AIDS is an incurable disease

-->

Hypernymy relationship (Aids and incurable disease )

# Extraction and structuration of terms and named entities



# Text processing TreeTagger

a probabilistic labeler  
grammatical categories  
morphosyntactic information  
lemmatization information  
Python: TreeTaggerWrapper

A tree where the root is the map  
of the whole world and each  
node is a quarter of its parent  
region.

TreeTagger

un	DET:ART	un
arbre	NOM	arbre
où	PRO:REL	où
la	DET:ART	le
racine	NOM	racine
est	VER:pres	être
la	DET:ART	le
carte	NOM	carte
du	PRP:det	du
monde	NOM	monde
entier	ADJ	entier
et	KON	et
chaque	PRO:IND	chaque
noeud	NOM	noeud
est	VER:pres	être
le	DET:ART	le
quart	NOM	quart
de	PRP	de
sa	DET:POS	son
région	NOM	région
parente	ADJ	parent
.	SENT	.

PÉLE-MÉL - Extraction de termes

Fichier Aide ?

**Sélectionner le corpus...** /home/etud/pele-mel-gitlab/pele-mel/workspace/presse/xaa

**Sauvegarder à...** /home/etud/pele-mel-gitlab/pele-mel/workspace/presse/Termes/termes.csv

Réduire les termes à leur racine :  Non

Méthode de Scoring :

Nombre min de mots dans un terme :

Nombre max de mots dans un terme :

**Lancer la recherche de termes**

4172 termes    Trier les termes par ordre :

N°	Terme	Score	Source
265	fédération hospitalière de france	0.0345939338286491	-
266	madame la ministre	0.0345939338286491	-
267	ministère de la santé	0.0345939338286491	-
268	service de pédiatrie générale	0.0345939338286491	-
269	soins à tarifs opposables	0.0345939338286491	-
270	élections législatives	0.03441557592697471	-
271	marges de progression	0.0343832893968011	-
272	membres du gouvernement	0.0343832893968011	-
273	nombre de choses	0.0343832893968011	-
274	organisation des soins	0.0343832893968011	-
275	réunion de travail	0.0343832893968011	-
276	sante ttp	0.0343832893968011	-
277	secrétaire d'etat	0.0343832893968011	-
278	services d'urgences	0.0343832893968011	-
279	rapport	0.03418999205224561	-
280	marche	0.03410432466611098	-
281	dimanche dernier	0.034084181410481584	-



Fichier Aide ?

Termes en attente de validation : /home/etud/pele-mel-gitlab/pele-mel/workspace/presse/Termes/en\_attente\_termes.csv

Termes validés : /home/etud/pele-mel-gitlab/pele-mel/workspace/presse/Termes/valid\_termes.csv

Termes supprimés : /home/etud/pele-mel-gitlab/pele-mel/workspace/presse/Termes/trash\_termes.csv

## Termes en attente de validation (3715)

Valider par termes de références

Valider par score

Valider par nombre de mots

Trier les termes par ordre :

Alphabétique des termes

Sauvegarder les modifications

N°	Terme	Score	Source
1	abandon	0.008019288673258906	--
2	abonnés absents	0.010805180404760504	--
3	absolue	0.010721760585547637	--
4	abstention très forte	0.010820142422937282	--
5	abus qu'ils constatent	0.010831250595117988	--
6	abus	0.009067912168631046	--
7	académie française	0.00866630377844585	--
8	accident	0.007567849669228233	--
9	accidents	0.008651878012343515	--
10	accompagnement	0.009723445067226865	--

## Termes validés

(413)

Trier les termes par ordre :

Décroissant des scores

N°	Terme	Score	Source
1	ministre de la santé	0.26932256443521285	--
2	président de la république	0.23051087346814003	--
3	modification de ce paramétrage	0.07719365443277432	--
4	centre périnatal de proximité	0.04563009372823321	--
5	communautés hospitalières de territoire	0.04563009372823321	--
6	maisons de santé pluridisciplinaires	0.04563009372823321	--
8	fédération hospitalière de france	0.0345939338286491	--
9	ministère de la santé	0.0345939338286491	--
10	service de pédiatrie générale	0.0345939338286491	--
11	soins à tarifs opposables	0.0345939338286491	--

## Corbeille

(12)

Trier les termes par ordre :

Décroissant des scores

N°	Terme	Score	Source
1	bonjour madame la ministre	0.0345939338286491	--
2	arrêt de la commercialisation	0.023087463457285157	--
3	actes en grand nombre	0.010831250595117988	--
4	action à votre service	0.010831250595117988	--
5	affichez un tarif moyen	0.010831250595117988	--
6	amis de mon entourage	0.010831250595117988	--
7	ancien ministre renaud dutreil	0.010831250595117988	--
8	ans dans le va	0.010831250595117988	--
9	apparence un peu bizarre	0.010831250595117988	--
10	applaudir ou les huer	0.010831250595117988	--

## Result of term extraction on the corpus

Messages with  
attachment(s):

**149** files (pdf, doc, txt, xml)

**14 563** sentences

**316 496** words'

---

number of terms proposed	<b>4493</b>
-----------------------------	-------------

number of correct terms	4136
----------------------------	------

number of incorrect terms	357
------------------------------	-----

---

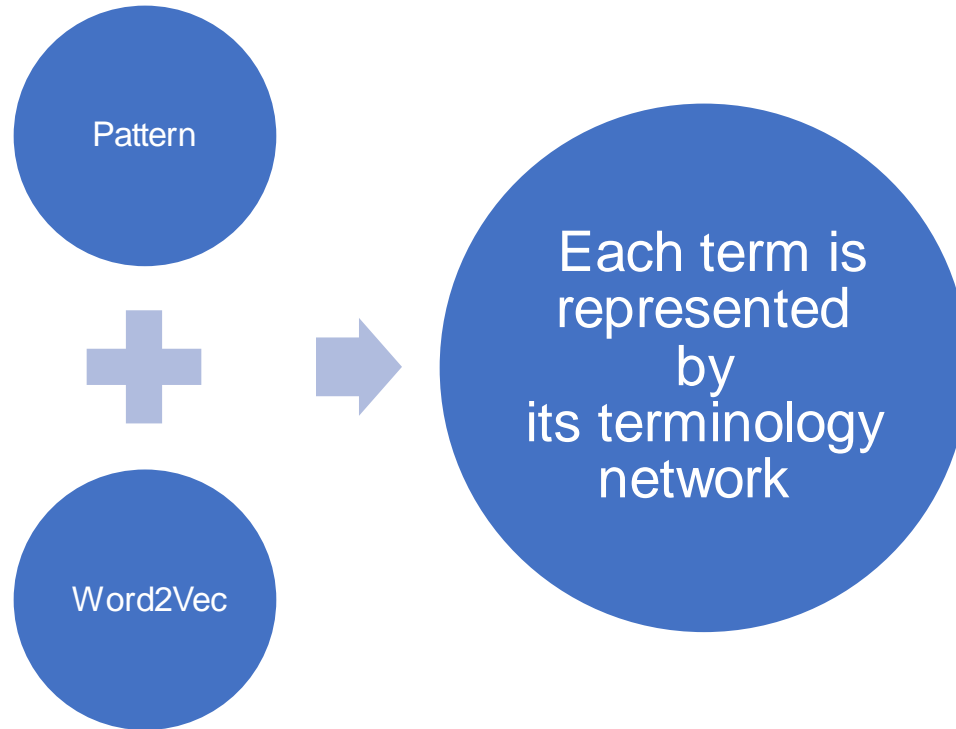
# Result of named entities extraction on the corpus

Messages with attachment(s):

**149** files (pdf, doc, txt, xml)

Named entities	Person	Organization
number of proposed	450	510
number of correct	424	366
number of incorrect	26	144

# Extraction semantic relationship



Fichier Relation Patron Aide ?

Sélectionner le corpus... /home/etud/pele-mel-gitlab/pele-mel/workspace/presse/sentences\_cleaned\_presse.txt

Sélectionner les termes... /home/etud/pele-mel-gitlab/pele-mel/workspace/presse/Termes/termes.csv

Sauvegarder à... /home/etud/pele-mel-gitlab/pele-mel/workspace/presse/Relations/reliations.csv

## Sélectionner les relations :

 hyperonymie  
 holonymie  
 meronymie  
 causalite-cause  
 causalite-effet  
 possession

## Sélectionner les patrons :

 hyperonymie\_qui\_etre\_adv\_det  
 hyperonymie\_qui\_etre\_adv\_det\_sorte\_de  
 hyperonymie\_qui\_etre\_adv\_det\_genre\_de  
 hyperonymie\_qui\_etre\_adv\_det\_genre\_du  
 hyperonymie\_qui\_etre\_adv\_de\_la\_famille\_det

Lancer l'extraction de relations

(38) relations

Valider les relations

N°	Terme 1	Terme 2	Relation
29	cause	tollé	hyperonymie
30	mensonge	tollé	hyperonymie
31	cause	tollé	hyperonymie
32	mensonge	tollé	hyperonymie
33	cause	tollé	hyperonymie
34	mensonge	tollé	hyperonymie
35	compte	crise financière	hyperonymie
36	projet	crise financière	hyperonymie
37	compte	crise financière	hyperonymie
38	projet	crise financière	hyperonymie

## PÊLE-MÉL - Détails de la relation

Terme 1 :  
projetTerme 2 :  
crise financièrePatron :  
entraînées parRelation :  
hyperonymie

Phrase :  
 et je ne reçois pas la critique du Parti socialiste parce que nous l'avons dans cette version du projet de loi de financement de la Sécurité Sociale tenu compte des modifications entraînées par la crise financière et la crise économique internationale.

# Result extraction semantic relationship ( Approach PATTERN)

Messages with PJ :  
**149** files (pdf, doc, txt, xml)

Type of relation	Number of relation extracted
Causality	483
Hyperonymy	416
Holonymy	31

# Word2VEC

Technique used to generate word embedding

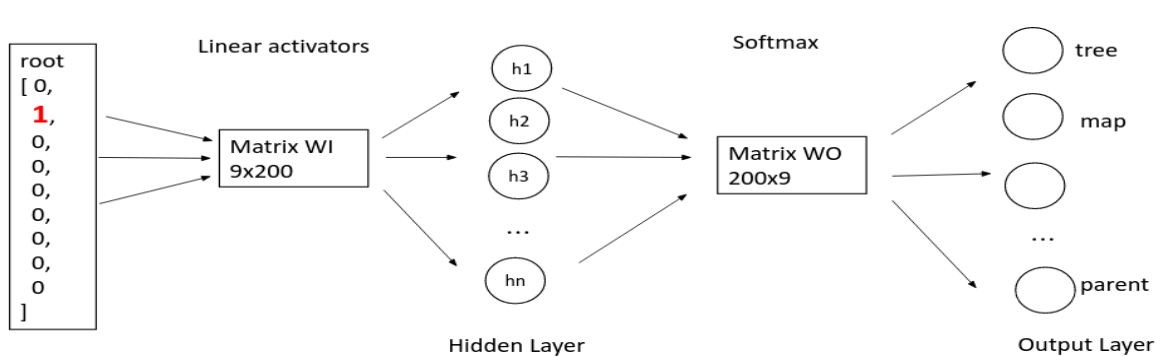
A two-layer artificial neural network trained to reconstruct the linguistic context of words

A ,tree, where ,the ,root ,is,  
the, map, of, the, whole, world,  
is, each, node, is, the, quarter,  
of, its, parent, region



[tree, root ,map, world, whole, node,  
quarter, region, parent]

Vector size = 200, Window size = 5 Vocabulary size = 9



	200
tree	[-0.00627418328076601 ... -5.979107227176428E-4 -0.025851745158433914 ]
root	[0.13695646822452545 ... -0.007662077434360981 0.009920799173414707]
map	[-0.034918226301670074 ... 0.023256413638591766 0.0065285274758934975 ]
word	[-0.006752261891961098 ... 0.02530670538544655 0.010089871473610401]
whole	[0.018857557326555252 ... -0.0714818686246872 0.014193015173077583]
node	[0.018857557326555252 ... -0.0714818686246872 0.014193015173077583]
quarter	[0.007662077434360981 ... -0.006752261891961098 0.014193015173077583]
region	[0.010404275730252266 ... 0.08519023656845093 -0.019239287823438644]
parent	[-0.03300181403756142 ... 0.027451258152723312 -0.04987586662173271]

PÊLE-MÊL - Word2Vec - Modèle entraîné - Recherche par thématiques

Fichier Aide ?

Sélectionner le modèle entraîné... /home/etud/pele-mel-gitlab/pele-mel/word2vec\_models/frWac\_no\_postag\_phrase\_500\_cbow\_c

Sélectionner la liste des thématiques... /home/etud/pele-mel-gitlab/pele-mel/data/thématiques.csv

Profondeur dans l'arbre de la recherche :

Sauvegarder les résultats à... /home/etud/pele-mel-gitlab/pele-mel/workspace/presse/Word2Vec/nuages\_mots.csv

Lancer la recherche

### Nuages de mots par thématique (10)

N°	Thématique	Nuage de mots
1	parentalité	fonction parental, parental, parents-enfant, reaaap, médiation familial, p
2	dépendance	dépendant, dépendance psychique, usage nocif, dépendance, persor
3	enfance	adolescence, adolescent, enfant, tendre enfance, âge adulte, famille,
4	santé publique	santé public, sante publique, épidémiologie, santé, léon bernard, conf
5	hôpital	hospitalier, établissement hospitalier, hôpitaux, centre hospitalier, assi:
6	prévention	santé, préventif, risque lier, sensibilisation, lutte contre, risque profess
7	assurance maladie	assurance-maladie, sécurité social, assuré social, assuré, assurance
8	assurance maternité	assurance maladie maternité, indemnité journalier, maladie maternité, &
9	insertion	insertion professionnel, insertion socioprofessionnel, plie plan, emploi
10	économie sociale	entreprendre autrement, économie solidaire, finance solidaire, solidaire

Fichier en cours de lecture : /home/etud/pele-mel-gitlab/pele-mel/workspace/presse/Word2Vec/nuages\_mots.csv

PÊLE-MÊL - Détails

Thématique :  
parentalité

Nuage de mots :

- fonction parental
- parental
- parents-enfant
- reaaap
- médiation familial
- prévention précoce
- conseil conjugal
- relation parents-enfant
- petit enfance
- conjuqalité



# Result (word2vec)

The screenshot shows a window titled "PÊLE-MÉL - Détails" with the following content:

Terme :  
abandon


Termes similaires :

- abandonner
- renoncement
- disparition
- abandon définitif
- désengagement
- abandon progressif
- renoncer
- irrémédiable
- dépossession
- contraindre

In the background, a list of terms is visible, including "accord", "accusation", and "accès". A search bar contains the text "Termes similaires".

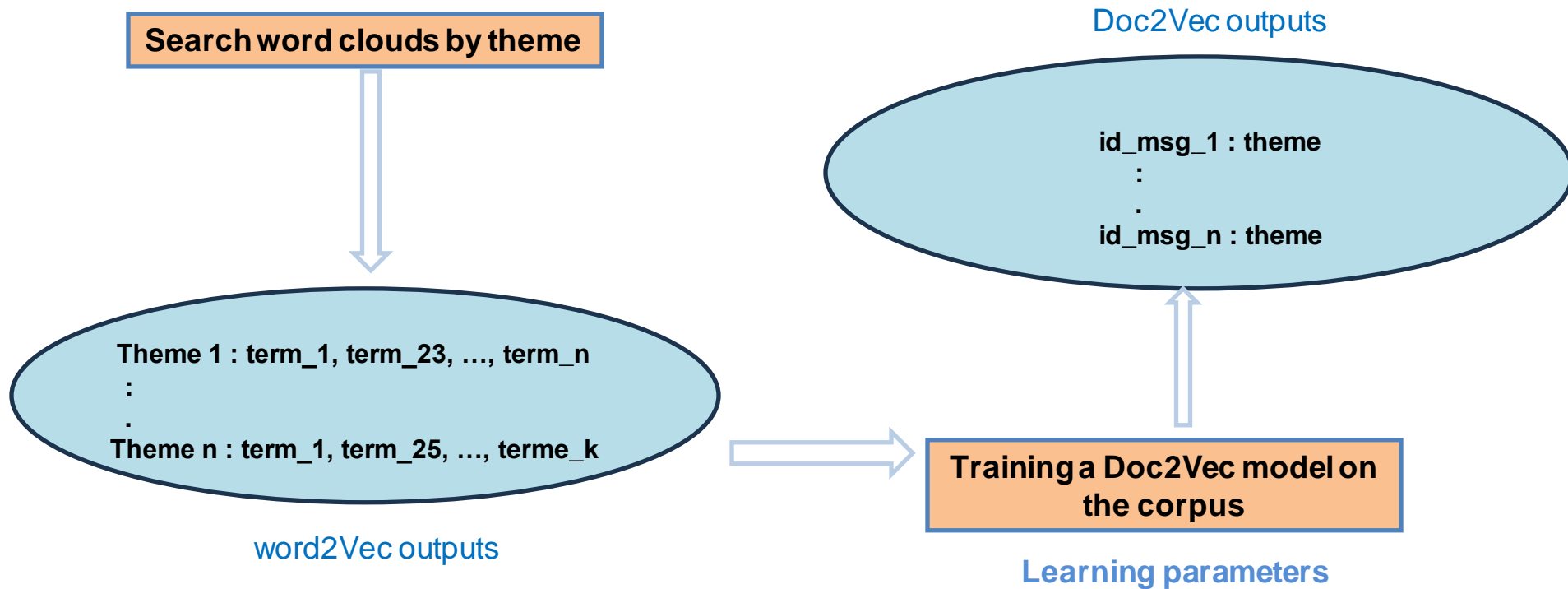


# Doc2vec

- Similar to word2vec
  - Word2vec vectors for each word
  - Doc2vec vector for each document + a "Paragraph Vector" to represent the entire document
  - the Doc2vec model will learn to associate each document with a numerical vector based on the context in which it appears in the corpus
- 

# Classification of messages by theme

Doc2Vec for classifying messages by theme



# Classification of messages by theme

Id message	Extract from the message	themes
EF1753BBA2A0A64C9763B0902E74 E0E1300263@AC005603.ac.intranet. sante.gouv.fr	<b>Sida, l'Etat dérape et fait silence...</b>	<b>Protection sociale</b>
EF1753BBA2A0A64C9763B0902E74 E0E1442F02@AC005603.ac.intranet. sante.gouv.fr	<b>Projet de réforme du système de santé... Sur les 10 ministres les plus appréciés des Français, six sont des femmes... Ce nouveau virus contient des gènes de plusieurs virus d'origine porcine</b>	<b>Grippe A</b>
...	...	...

# Output files

	A	B	C	D	E	F	G
1	Id	labels_niveau1	labels_niveau2	labels_niveau3			
2	<003301c7c3ed\$e9a13470\$23844251@maison>	Protection sociale	insertion	Autisme			
3	<AC005601cqttPJI7nAE00014a2c@AC005603.ac.intranet.sante.gouv.fr>	Handicap	Autonomie	Droit des malades en fin de vie			
4	<00b101c7b1e8\$596f1690\$c6304351@maison>	Femme	prévention	Grippe A			
5	<AC005602X2CTp2hFOWd00000e8b@AC005603.ac.intranet.sante.gouv.fr>	Protection sociale	médecine libérale	Aide à domicile			
6	<AC005601x9NxHyPuDDt00014f0f@AC005603.ac.intranet.sante.gouv.fr>	Famille	assurance maternité	Plan obésité			
7	<EF1753BBA2A0A64C9763B0902E74E0E1300263@AC005603.ac.intranet.sante.gouv.fr>	Famille	Offre de soins	Plan obésité			
8	<AC005601B07XOq3JkrK00001b5c@AC005603.ac.intranet.sante.gouv.fr>	Famille	Offre de soins	Aide à domicile			
9	<EF1753BBA2A0A64C9763B0902E74E0E1442F02@AC005603.ac.intranet.sante.gouv.fr>	Famille	Offre de soins	Médiateur			
10	<AC005601Qgl4rhYQYx300000286@AC005603.ac.intranet.sante.gouv.fr>	Handicap	insertion	Plan obésité			
11	<mnet3.1181654155.23611.beatrice.noellec@noos.fr>	Vie associative	Offre de soins	Médiateur			
12	<AC005601WoHLwo6vLxZ000080fb@AC005603.ac.intranet.sante.gouv.fr>	droits des femmes	insertion	Autisme			
13	<AC005601VMczi9yzu230001ad4b@AC005603.ac.intranet.sante.gouv.fr>	Famille	assurance maternité	Plan obésité			
14	<mnet3.1181036948.29516.beatrice.noellec@noos.fr>	Vie associative	Offre de soins	Aide à domicile			
15	<AC005601DG9I3tiA4el0001d3a5@AC005603.ac.intranet.sante.gouv.fr>	Famille	assurance maternité	Plan obésité			
16	<AC005601BGWvN2nSsNf0000ea89@AC005603.ac.intranet.sante.gouv.fr>	Famille	assurance maternité	Loi HPST			
17	<EF1753BBA2A0A64C9763B0902E74E0E1013E60D9@AC005603.ac.intranet.sante.gouv.fr>	Famille	lutte contre la pauvreté	lutte contre la pauvreté			
18	<EF1753BBA2A0A64C9763B0902E74E0E101CB1FDF@AC005603.ac.intranet.sante.gouv.fr>	Famille	assurance maternité	Loi HPST			
19	<AC005601qqJgSt51NDa0001d3a7@AC005603.ac.intranet.sante.gouv.fr>	droits des femmes	assurance maternité	Plan obésité			
20	<AC005601GRsq1bgIK780000856c@AC005603.ac.intranet.sante.gouv.fr>	Vie associative	Offre de soins	Aide à domicile			
21							
22							
23							
24							
25							
26							
27							
28							
29							

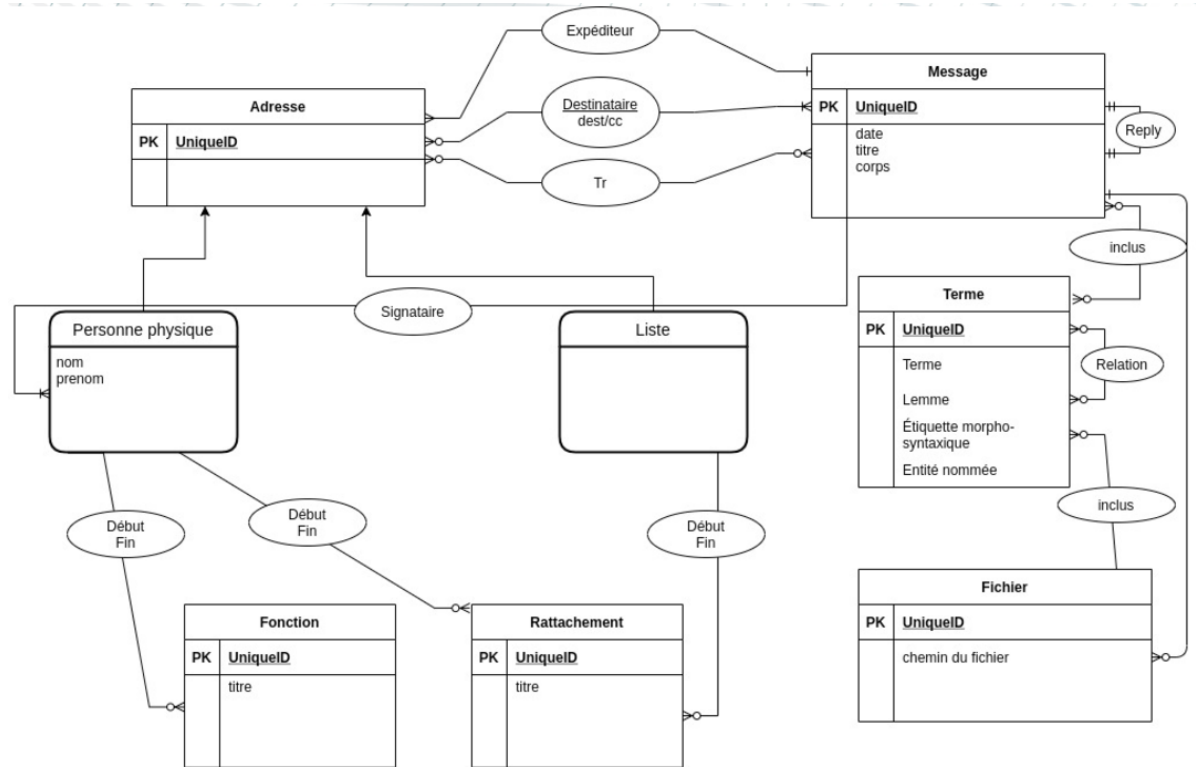
A black and white photograph of a busy post office sorting room. Several men in suits are working at long tables, sorting through large stacks of mail. The room is filled with rows of mailboxes on the right side, and the ceiling is supported by a complex network of pipes and lights. The overall atmosphere is one of organized activity.

# Exploration

# Two applications in one

A lightweight application that can be installed directly on the desktop by copying and pasting

A relational database that can be connected to open up further search possibilities and integrate classification output files



# Basic functionalities

You can query message metadata and content and filter results [1, 2]: name, subject, recipient, sender, date, CC, content (message and attachments). An advanced search is also possible.

The message list [3] can be displayed and edited in a window.

Message content [4] and attachments can be displayed in their archive format (pdf, image format, word).

The screenshot displays an email client interface with several key areas highlighted in yellow:

- 1**: The search bar labeled "Recherche par" with a dropdown menu showing options like "Nom", "Sujet", "Destinataire", "Expéditeur", "Date", "CC", "Contenu", and "Recherche avancée".
- 2**: The "Métadonnées" (Metadata) panel on the right, showing fields for "Nom", "Sujet", "De", "À", "Date", "CC", "Signature", and "Avis d'expert".
- 3**: The "Trier Liste Par" (Sort List By) dropdown menu, currently set to "Nom".
- 4**: The main message content area, which includes the text of the email and a list of attachments at the bottom.

The email content is as follows:

Bonjour à tous,

La consultation en ligne à propos de l'évaluation et la collecte des archives est disponible en avant-première à partir d'aujourd'hui 10 avril et jusqu'au 16 avril à l'adresse suivante : <https://assembl-civic.blunenove.com/archivespourdemain/debate/survey> <<http://assembl-civic.blunenove.com/archivespourdemain>>

Durant cette semaine, nous comptons sur vous pour contribuer à la première phase du débat afin que la plate-forme ne soit plus une page blanche lors de son ouverture officielle au public.

A partir du 16 avril, la plate-forme sera ouverte à tous à l'adresse suivante que vous pouvez d'ores et déjà diffuser : <http://assembl-civic.blunenove.com/archivespourdemain>. Nous vous encourageons à communiquer largement l'information autour de vous et à mobiliser votre réseau professionnel, par exemple :

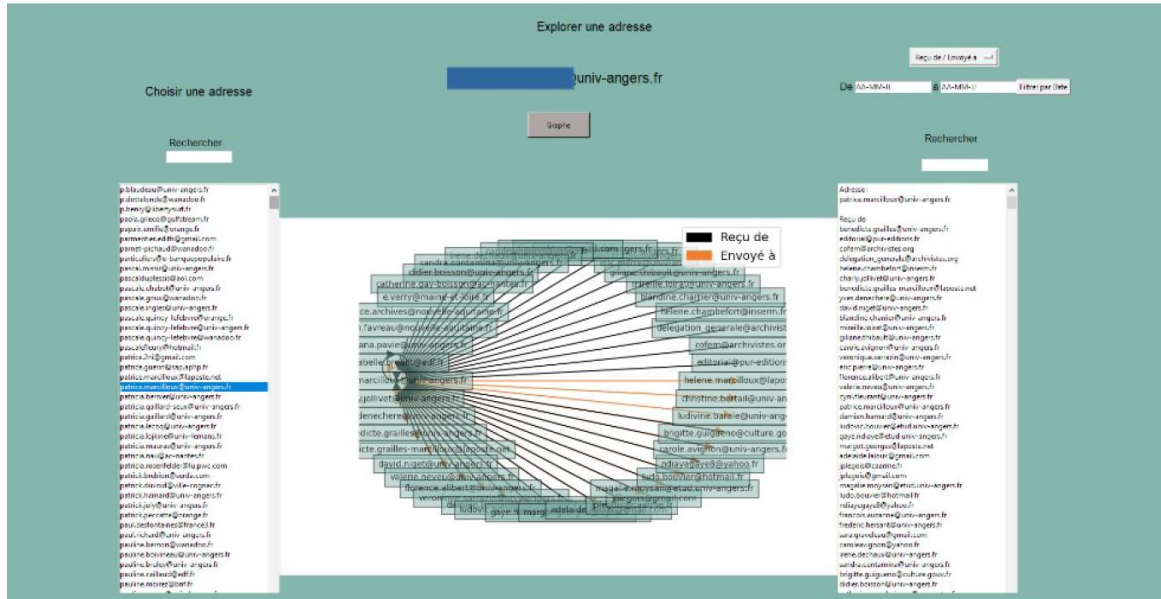
- en imprimant et affichant à l'attention du public qui fréquente vos locaux une affichette que vous trouverez en pièce-jointe ;
- en utilisant le communiqué de presse ci-joint pour publier une actualité sur votre site internet ou votre lettre d'information électronique ou papier ;
- en diffusant le communiqué de presse ci-joint à vos correspondants archives, votre réseau local d'

The attachments listed at the bottom are:

- Corps
- HTML
- PDF
- Images
- Microsoft Word



# Basic functionalities



You can visualize the relationships between an address and its correspondents (received and sent messages) in the form of a graph.

A dynamic, configurable graph can be generated.

Address and date filtering and search functions can be used to refine the graph, whose image can be exported.

## Advanced functionalities

If you want to exploit the classification previously performed on the first prototype, you'll need the database, the lists of themes used for classification and the output files classifying messages from doc2vec

Creating the database

Créer BDD	Extraire messages
Vider messages	Extraire PJs
Vider thèmes	Extraire thèmes
Vider classification	Ajouter classification

Extracting messages

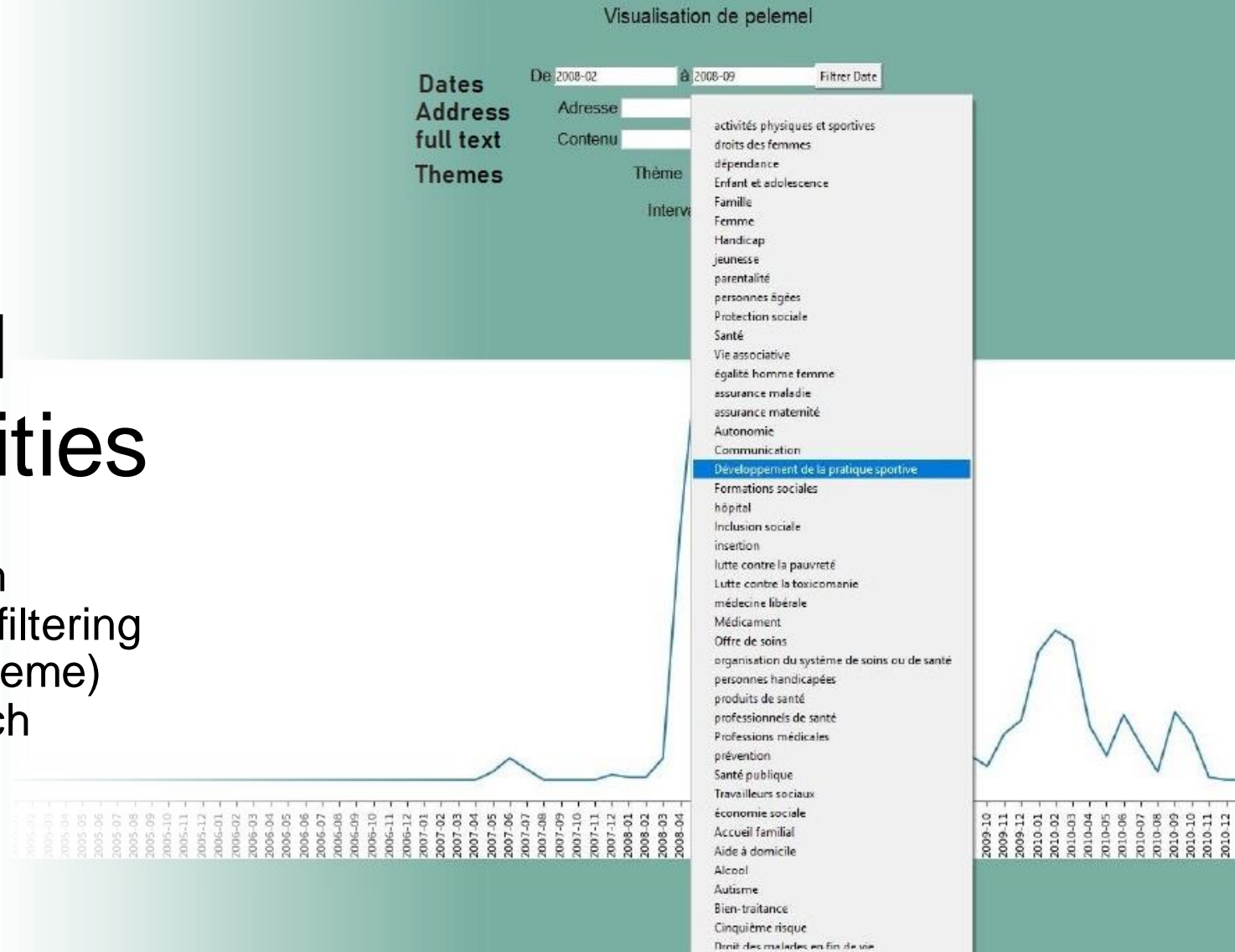
Extracting attachments

Loading files with themes

Loading message classification output files

# Advanced functionalities

Visualization: graph construction with filtering (date, address, theme) and full-text search capabilities



# Advanced functionalities

Via the database, the archivist can modify tables, annotate and correct classifications.

The screenshot displays a web application interface for managing a database of themes and terms. The interface is titled "Thèmes/Terms de pelemel" and is organized into four main columns, each with a list of items and a set of control buttons at the bottom.

- Thèmes de:** This column is currently empty.
- Thèmes:** This column contains a list of 20 themes, with "Professions médicales" highlighted in green. The list includes: personnes âgées, Professions médicales, assurance maladie, assurance maternité, Autonomie, Communication, Développement de la pratique sportive, Formations sociales, hôpital, Inclusion sociale, insertion, lutte contre la pauvreté, Lutte contre la toxicomanie, médecine libérale, Médicament, Offre de soins, organisation du système de soins ou de santé, personnes handicapées, produits de santé, professionnels de santé, Professions médicales, prévention, Santé publique, Travailleurs sociaux, and économie sociale.
- Messages:** This column contains a list of 11 messages, with message 6 highlighted in blue. The messages are numbered 3, 4, 6, 10, 17, 19, 22, 27, 32, and 33.
- Termes:** This column contains a list of 20 terms, including: accentuation, acompte, agregat, AIDE\_ALIMENTAIRE, apul, argus, assainir, ASSURANCE\_SOCIALE, AYANT, COMPTABILITÄ%\_NATIONALE, cotisant, CULTURE\_GÃ%NÃ%RALE, decomposition, decroitre, delibereront, depenses, deraper, differe, discrediter, esps, exemption, ffipsa, globalisation, interministerialite, mainmise, mecaniquement, MILIEU\_NATUREL, and minier.

At the bottom of each column, there are control buttons: "Ajouter", "Modifier", and "Supprimer" for the "Thèmes" and "Termes" columns, and "Plus", "Modifier", and "Supprimer" for the "Messages" column.

# Conclusions



# Limits and obstacles

The size of the corpus: the strategy should be tested on larger volumes. With many more mailboxes, the BERT option and its French variant, CamenBERT, might be appropriate.

The context specialized: the experiment should also be transposed to two other types of context, specialized and non-specialized.

Prototype: the ergonomics need to be improved.

Both archivists and the IT specialists who work with them need to become familiar with NLP, and this requires training.

The first batches of messages require a great deal of attention and time to correct, validate and improve the classification. This time spent is a long-term investment, but few departments see the immediate benefits.

# Innovations and benefits

Many elements can be replicated in another environment

We believe we have validated the proof of concept

These tools can be used to enrich archival thinking on box appraisal and internal sorting.

[benedicte.grailles@univ-angers.fr](mailto:benedicte.grailles@univ-angers.fr)

[touria.aitelmekki@univ-angers.fr](mailto:touria.aitelmekki@univ-angers.fr)

C  
o  
r  
p  
u  
s

Deux messages de conseillers de cabinet + pièces jointes

- Gouvernement, Cabinet, Direction générale de la santé, 2007, 2010 ; juin-oct 2011

Organigrammes et annuaires (cabinet, directions du ministère)

- Direction générale de la cohésion sociale

Deux thésaurus

- 14 documents
- 2014 & 2020
- 7000 descripteurs

Discours (corpus communicable)

- pdf ; doc
- 810 documents (nov. 2010 – mars 2012)
- doc / pdf

