Improving the archiving and contextualization of electronic messaging in French

|  |  |  |
| --- | --- | --- |
| **Bénédicte Grailles** | **Touria Aït El Mekki** | **Édouard Vasseur** |
| *Université d’Angers*  *France*  [*benedicte.grailles@univ-angers.fr*](mailto:benedicte.grailles@univ-angers.fr)  *0000-0003-2042-855X* | *Université d’Angers*  *France*  *touria.aitelmekki@univ-angers.fr* | *École nationale des chartes*  *France*  *edouard.vasseur@chartes.psl.eu*  0000-0003-1503-2075 |
|  |  |  |

**Abstract – The objective of the Pêle-mél program is to propose a prototype tool for exploring and visualizing electronic messages and to test different strategies for contextualizing and classifying electronic messages in French, using NLP techniques, artificial intelligence and learning. Beyond that, the aim is also to draw conclusions on archiving strategies and to theorize the scope of the external knowledge required to succeed in this type of project. This program, financed by the French Ministry of Culture, innovates on two points: the will to understand messaging systems in their reticular context and the deployment of techniques adapted to French. The first results confirm, thanks to a detailed analysis of the messages, the interest of large-scale archiving and classification. They also show the difficulties linked to the temporal context of the studied mailboxes and to the hybrid character of the information supports. The palliative strategies put in place are costly but possible.**

**Keywords – email, machine learning, classification, terminology**

**Conference Topics – innovation; resilience**

# Introduction

The objective of the Pêle-mél program is to propose a prototype tool for exploring and visualizing electronic messages and to test different strategies for contextualizing boxes, correspondent networks and the information content of messages using Automatic Language Processing (ALP) techniques adapted to French, in particular classification. It is also a question of developing criteria to evaluate the archival value of messaging systems and to help in the decision making process, in order to contribute to a relevant and reasoned selection. The current developments are based on mailboxes already collected by the Records and Archives office of the Ministry of Solidarity and Health and on knowledge external to these messaging systems [1]. The question of the cost of acquiring these criteria, the cost to provide or the actions to prioritize is also on the agenda. This project is financed by the French Ministry of Culture in the framework of a call for projects "innovative digital cultural services" [2].

# Context and state of the art

For more than 30 years now, electronic messaging has become an essential tool for the production and transmission of information. Like other countries, France is concerned by the phenomenon, even at the highest governmental levels. For members of the government and their direct collaborators, electronic messaging is an essential tool. In daily work, they have become the medium of strategic information and often the unique traces of decision-making processes [3].

Since the beginning of the 2010s, records managers and archivists of ministries systematically archive documents produced and transmitted by means of electronic messaging by ministers and their collaborators, on the occasion of the cessation of functions. In the social ministries (health, solidarity, labor), electronic messages constituted a significant part of the documents collected in 2020 and represented 45% of the size of electronic archives kept.

Archival appraisal and access to archived messages is also a growing challenge. This is a crucial issue, whether to respond to requests for access from administrations, judges and journalists, or to prepare for the transfer of documents to the National Archives by carrying out the necessary sorting operations. Facilitating the search for meaningful information in the midst of the mass of archival messages is becoming a strategic challenge.

The preservation of electronic messages has been widely studied for several years, both in France and abroad, with several recent publications such as those produced under the authority of Christopher J. Prom [4]. In France, the interministerial electronic archiving team Vitam has been particularly interested in the subject, both from a theoretical and practical point of view [5]. It has developed a java library, MailExtract, which allows the extraction of a tree structure of messages in .eml format from the raw exported files, taking into account the specificities of messages in French, notably accented characters.

On the other hand, the question of access is still little studied, except in the United States. Based on the results of the ePADD project led by Stanford University [6], the RATOM project led by the University of North Carolina has initiated the use of named entity recognition using NLP libraries, in order to facilitate the identification of messages, the publication and public access [7]. Furthermore, email processing is often used for information extraction: spam detection, email categorization, contact analysis, email network property analysis, and email visualization [8]. The value of NLP techniques combining rules and machine learning and using contextual information, has been shown [9]. The value of Topic Modeling to analyze big unclassified text is proven [10] and experiments with pre-annotation of named entities have been conducted [11].

However, standard clustering tools are not sufficient due to typography, the absence of a model or pre-trained corpus, the grammatical specificities to be taken into account in the models, or the semantic characteristics of terms when they contain several words or compound words, which require term extraction based on language.

Undertaking research to facilitate access to archived messages is therefore becoming a necessity in the French-speaking world, because of its linguistic specificities.

# Corpus and pre-processing

For this pilot project, we have at our disposal two electronic mailboxes from female advisors in the office of the Minister Roselyne Bachelot-Narquin, comprising 8,636 messages and their attachments, covering the period 2007-2011. It is these boxes that we seek to contextualize and interpret. The records and archives department also provided us with copies of paper directories and organization charts, some of which were native digital and others digitized. We also retrieved two thesauri used by the ministry's documentation center, one from 2014, the closest in date, the other from 2020 (7,000 descriptors) and a corpus of 810 speeches delivered by the minister between 2010 and 2012. These different sets have undergone a certain amount of pre-processing in order to be manipulated.

The messages were taken from Outlook via pst exports. They were processed by the MailExtract tool and are in the form of eml files documented by an xml file that complies with the data exchange standard for archiving [12]. Each message constitutes a separate silo.

The message metadata was retrieved from the xml files and separated from the attachments and the message body. Attachments, message objects, and message texts were morpho-syntactically tagged using the Natural Language Toolkit (NLTK) python software library [13]. The labels were also applied to the speech previously transferred from pdf to txt. Different term extractors, capable of categorizing the names of natural persons, legal entities, abbreviations, locations, etc. and supporting French, could be tested. Only those that could be installed locally were selected in order to preserve the confidentiality of the data. The corpora were also pre-processed to lemmatize and remove accented characters.

The organization charts and directories were entered manually into a spreadsheet. Indeed, these various documents did not present sufficient regularity to attempt an automatic recovery.

# First results

One third of the messages are in a thread (in reply or re-posting). Morpho-syntactic tagging shows that these messages are correctly written, in a relatively elaborate language. The median number of sentences per message is 6 and the median number of words is 228. These sentences include nouns, adjectives and verbs. They are well structured. Attachments are present in 30% to 70% of messages. It is therefore imperative to include these documents in the classification. In total, a network of more than 2,700 correspondents is involved. 30% of the addresses correspond to mailing lists that had to be located in the corpus thanks to regularities (capital letters, punctuation) and then a light manual cleaning. The real people concerned by these lists are not known; the groups have not been archived and cannot be reconstructed with certainty. Nevertheless, the names of the groups are quite transparent and the cross-referencing with directories and organization charts can be exploited. The statistics on recipients show that the information is mainly circulated internally within the minister's office and more broadly within the department.

The exploration tool is based on a relational database that relies on the initial metadata, the attached files and their naming, and the message body. Within the body of the message, signatures were identified and extracted to enrich a directory of individuals, their affiliation and their function. This capitalization is not sufficient in itself. The extraction of the named entities makes it possible to identify natural persons whose function and affiliation must be identified. The directories and organization charts make it possible to inject external knowledge to contribute to the identification.

One of the challenges of contextualizing messages is the identification and resolution of acronyms. The part-of-speech tagger offer an "abbreviations" category. To establish the most complete list possible, we relied on the descriptors of the thesaurus that we projected onto the messages, their subject and attachments, on the implementation of rules and on a dictionary of acronyms and acronyms of the administration crossed with the identification of named entities. More than 400 acronyms have been identified.

A unsupervised clustering was implemented with K-Means and Iramuteq [14]. The result was not very convincing. The thesaurus descriptors were projected onto the messages, their subject and attachments. 60% of the 7000 descriptors are used in the email corpus. Despite this figure, the thesaurus is not very helpful because it is not sufficient to discriminate between messages that turn out to be very similar.

It was decided to test word embedding models which represent words by vectors and document embedding model which, instead of vector representation of words in the documents present in the data, focuses on creating vector representations of documents regardless of its length: Fasttext [15], Word2Vec, Doc2Vec. Relationships are identified through generic and specific patterns. The results of this phase are currently being verified, but the results are correct.

# Conclusion

The program is now in the middle of the road. The classification gives usable results and the ontology is under construction. The visualizations are very advanced. The question of named entities still needs to be explored. It is still difficult to draw a conclusion. Strategies exist but they are costly. Relevant data is lost: hyperlinks point to nothing, LDAP directory are not kept; mailing lists have lost their contacts. Acquiring external knowledge is complex and requires manual rework at this pivotal time when information is hybrid, paper and electronic. However, it is equally true that there are interesting strategies that could be called upon depending on the level of results expected.

# REFERENCES

1. In addition to Touria Aït el Mekki, Bénédicte Grailles and Tsanta Randriatsitohaina (University of Angers), Anne Lambert and Chloé Moser (Ministry of Health and Solidarity) and Édouard Vasseur (École nationale des Chartes, Jean-Mabillon Center) are participating in this project.
2. List of winners of the call for innovative digital services 2020, Ministry of Culture. [<https://www.culture.gouv.fr/Thematiques/Innovation-numerique/Appel-a-projets-Services-numeriques-innovants-SNI/Laureats-de-l-appel-a-projets-Services-numeriques-innovants-2020>
3. S. Bretesché., B. de Geffroy, F. de Corbière., *E-bureaucratie: le travail emmailé des cadres*, Paris: Presses des Mines, 2018.
4. C. Prom *Preserving email*., 2nd ed., London: DPC, 2019. <https://www.dpconline.org/docs/technology-watch-reports/2159-twr19-01/file>
5. Programme VITAM, *L’archivage des messageries électroniques. Preuve de concept VITAM*, Paris: Ministère des Affaires étrangères/Ministère de la Culture/Ministère de la Défense, 2013.
6. ePADD Project Website Homepage, University of Standford. <https://library.stanford.edu/projects/epadd>
7. RATOM Project Website Homepage, University of North Carolina. <https://ratom.web.unc.edu/>
8. G. Tang., J. Pei, & W.S. Luk, “Email mining: tasks, common techniques, and tools”, *Knowledge and Information Systems*, vol. 41, no 1, pp. 1-31, 2014
9. R. Alghamdi., K. Alfalqi, “A survey of topic modeling in text mining”, *International Journal of Advanced Computer Science and Applications*, vol. 6.no 1, 2015.
10. D. Maynard, Y. Li, P. Wim, “NLP Techniques for Term Extraction and Ontology Population.” In Proceedings of the 2008 conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge. IOS Press, NLD, pp. 107–127, 2008.
11. P. Suárez, Y. Dupont., B. Muller, L. Romary, B. Sagot, “Establishing a New State-of-the-Art for French Named Entity Recognition” in LREC 2020 - 12th Language Resources and Evaluation Conference, May 2020, Marseille, France. [⟨hal-02617950v2⟩](https://hal.inria.fr/hal-02617950v2)
12. SEDA standard Homepage, Ministry of Culture. <https://www.francearchives.fr/seda/>.
13. Natural Language Toolkit Documentation. <https://www.nltk.org/>
14. Iramuteq Website. <http://iramuteq.org/>.
15. FastText Website. <https://fasttext.cc/>