Pêle-mél

Plate-forme d'exploration, de livraison et d'évaluation des méls

Rapport d'évaluation des usages

Autrice : Bénédicte Grailles (Université d'Angers, Temos)

31 mars 2023

Soutenu par

Abstract

This report presents an assessment of the Pèle-mél project, which aime to improve access to electronic mailboxes held by archives. The project resulted in two prototypes, one allowing the extraction and classification of a corpus of mailboxes, the other allowing the exploration of these same boxes. The evaluation of uses was based on focus groups with archivists involved in archiving projects or who had already archived mailboxes and on an sample survey. Several observations were made: the need to adapt professional representations, the requirement for training and support, the real complexity of taking charge, but also innovative perspectives, likely to change practices, by allowing real access to content.

Keywords : electronic messaging, supervised automatic classification, natural language processing, archiving, access, word embeddings, document embeddings, uses

Résumé

Ce rapport propose un bilan du projet Pèle-mél visant à améliorer l'accès aux messageries électroniques conservées par les services d'archives. Le projet a abouti à deux prototypes, l'un permettant les extractions et la classification de corpus de boîtes méls, l'autre l'exploration de ces mêmes boîtes. L'évaluation des usages s'est appuyée sur des groupes de discussion (focus groups) réunissant des archivistes engagés dans des projets ou des collectes d'archivage de messageries et sur un questionnaire. Plusieurs constats ont été formulés : la nécessité d'adapter les représentations professionnelles, des besoins de formation et d'accompagnement, une réelle complexité de prise en charge mais aussi des perspectives innovantes, susceptibles de faire évoluer les pratiques, en permettant un réel accès au contenu.

Mots clés : messagerie électronique, classification automatique supervisée, traitement automatique de la langue naturelle, archivage, accès, méthode de plongement lexical, méthode de plongement de documents, usages

MINISTÈRES SOCIAUX Liberté Égalité Fraternité
PSL
LERIA
temos cnrs temps · mondes · sociétés · UMR 9016
université angers
Soutenu par MINISTÈRE DE LA CULTURE Liberté Égalité Fraternité

Remerciements

Pêle-mél est un projet exploratoire qui propose d'importer des méthodologies nouvelles dans l'univers des archives. Il s'est nourri d'échanges préalables avec plusieurs archivistes. Qu'elles et ils soient toutes et tous remercié·es, avec une pensée particulière pour Aurélien Conraux, administrateur ministériel des données délégué au ministère de la Culture. Le programme a été accompagné par le Service interministériel des archives de France. Notre gratitude va à Françoise Banat-Berger, cheffe du Service interministériel des Archives de France, Violette Levy, cheffe du bureau de l'expertise numérique et de la conservation durable, Mélanie Rebours, cheffe du bureau du contrôle, de la collecte, des missions et de la coordination interministérielle, Dominique Naud, experte en archivage numérique. Lauréat de l'appel à projets Service numérique innovant du ministère de la Culture en 2020, il a bénéficié du suivi bienveillant d'Ariane Faraldi. Merci à elle. Merci aussi au laboratoire Temos, pour son soutien sans faille à l'archivistique et son aide dans la gestion et le montage du projet. Merci à Yves Denéchère, directeur, et Mireille Loirat, gestionnaire administratrice et aide au pilotage.

Ce projet n'aurait pas vu le jour sans notre partenaire culturel, la mission Archives des ministères sociaux, Anne Lambert, cheffe de la mission, et Chloé Moser, cheffe de produit Archifiltre. Qu'elles soient ici vivement remerciées. Nos remerciements vont également à Édouard Vasseur, École nationale des Chartes, pour son expertise jamais démentie, son implication et ses contributions tout au long de nos travaux.

Ce projet ne pouvait être imaginé sans corpus. Pour cela, il fallait obtenir de la ministre en exercice aux dates concernées l'autorisation d'accéder à des messageries et de les manipuler. Cette dérogation nous a été accordée sans aucune difficulté. Nous remercions Roselyne Bachelot-Narquin d'avoir accepté de nous ouvrir le fonds du cabinet.

Il nous faut aussi remercier tout·es les archivistes qui ont participé aux ateliers et bien voulu consacrer du temps à notre projet. Merci également à Patrice Marcilloux, professeur d'archivistique à l'Université d'Angers, membre de Temos, qui a assumé le rôle d'observateur à l'occasion des groupes de discussion et à tou·tes les contributeur·rices qui ont mis leur talent au service du programme : Chafik Akmouche, Tsanta Randriatsitohaina, Taïmane Zerez.

Sommaire

MINISTÈRES SOCIAUX Liberté Égalité Fraternité
PSL
LERIA
temos cnrs temps · mondes · sociétés · UMR 9016
université angers
Soutenu par
MINISTÈRE DE LA CULTURE Liberté Égalité Fraternité

Introduction

Le programme Pêle-mél [https://alma.hypotheses.org/programmes-de-recherche/pele-mel-plate-forme-dexploration-de-livraison-et-devaluation-des-mels] a pour objectifs de fournir de nouveaux outils d'exploration de corpus de messageries pour satisfaire les demandes d'information en sélectionnant les messages pertinents afin de répondre aux enjeux d'accès et de mise en conformité avec les obligations de communication des documents administratifs et des archives (Code du patrimoine, Code des relations entre le public et les administrations). L'originalité du projet réside dans la mobilisation de technologies de traitement automatique de la langue naturelle en français.

Les objectifs principaux sont d'élaborer une base de connaissance recensant les métadonnées internes des messages, les contenus, les signatures, les pièces jointes et leurs métadonnées ; d'en extraire les entités nommées et les termes ; de les relier à des thèmes ; de catégoriser et classifier des messages en utilisant du clustering et du machine learning et de proposer des fonctions de visualisation et de recherche via des requêtes simples et expertes, des fonctions de filtre. Le projet convoque donc un certain nombre d'outils et de techniques de traitement automatique de la langue naturelle (Taln).

Les partenaires sont l'université d'Angers (Laboratoires – Temos Temps, Mondes, Société UMR CNRS [https://temos.cnrs.fr/] – et Leria – Laboratoire d'étude et de recherche en informatique d'Angers [https://leria.univ-angers.fr/]), le ministère de la Santé et des Solidarités (mission archives) et l'École nationale des chartes (centre Jean-Mabillon). Le portage du projet a été assuré par le laboratoire Temos.

Contacts : Bénédicte Grailles (benedicte.grailles[at]univ-angers.fr),
maîtresse de conférences en archivistique
[https://temos.cnrs.fr/grailles-benedicte/]
Touria Aït el Mekki (tourial.aitelmekki[at]univ-angers.fr),
maîtresse de conférences en informatique

1. Contexte

C'est à partir des années 1990 que les messageries électroniques se généralisent pour devenir un médium central de circulation de l'information dans les sphères personnelles et professionnelles. Dans le travail quotidien, elles sont devenues le support d'informations stratégiques et souvent les traces uniques de processus décisionnels (Bretesché S., Geffroy B., de Corbière F. 2018).
Depuis une dizaine d'années, les missions archives des ministères procèdent à la collecte systématique des boîtes mél des ministres et de leurs collaborateur.rices direct·es lors des remaniements gouvernementaux. Au sein des ministères sociaux (santé, solidarités, travail), les messages électroniques constituent en 2020 un part importante des documents collectés et représentent 45 % du volume des archives électroniques conservées. 200 comptes de messagerie ont déjà été collectés depuis 2012, la plus volumineuse représentant 70 Go de données. En revanche, dans les services d'archives territoriaux et chez les opérateurs, la collecte des messageries reste limitée et parfois inexistantes. En cause, divers obstacles techniques et méthodologiques mais aussi une certaine réserve de la part des archivistes : crainte de conflits autour du caractère privé de certaines correspondances ou certains usages, difficulté à prioriser cette collecte faute d'une vision claire des usages et exploitations possibles.

1.1. Archivage des courriels

La pérennisation des courriels intéresse les archivistes depuis plusieurs années. Pour preuve, on trouvera plusieurs publications récentes (Prom C. 2019) et exemples de projets à l'étranger (ePADD [https://library.stanford.edu/projects/epadd], RATOM [https://ratom.web.unc.edu/]). En France, le programme interministériel d'archivage numérique Vitam [http://www.programmevitam.fr] a développé une réflexion (Programme Vitam 2013) et des outils notamment une librairie java, MailExtract, permettant d'extraire une arborescence de messages au format .eml des fichiers bruts exportés, et tenant compte des spécificités de la langue française (caractères accentués). La recherche s'est focalisée sur la préservation, ne s'intéressant que marginalement à la question de l'accès et de la restitution de l'information.
L'archivage des courriels s'inscrit dans le cadre de la norme ISO 14721:2012 ou norme OAIS et du standard d'échange des données pour l'archivage (SEDA) promu par le service interministériel des Archives de France.
En août 2022, une recherche dans la salle de lecture virtuelle des Archives nationales permet d'identifier environ 130 messageries essentiellement issues du ministère de la Culture, mais de nombreuses messageries collectées par les missions n'ont pas encore été transférées. Une recherche plus large sur les sites des services départementaux d'archives fait apparaître très peu de résultats.

Les organisations opèrent des choix en s'appuyant sur le niveau de décision du titulaire de la messagerie. Ainsi le ministère des Affaires étrangères ne collecte que 3 à 4 % des messageries existantes. La mission des ministères sociaux collecte quant à elle les

messageries du cabinet, des directeur.rices et sous-directeur.rices d'administration centrale. Le tri interne est en général laissé aux utilisateurs.

1.2. Traitement automatique de la langue naturelle (Taln) et courriels

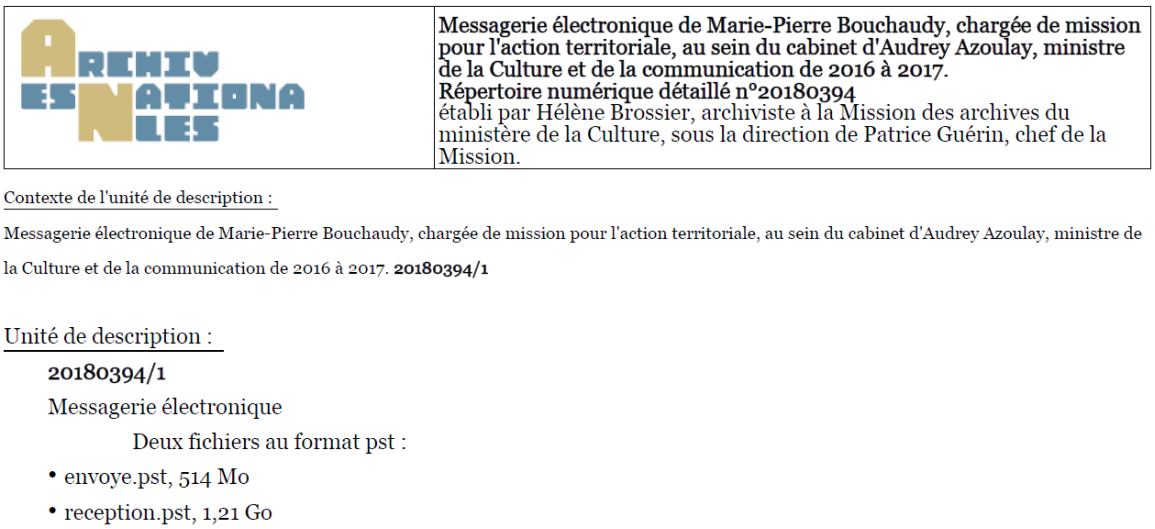
L'explosion du volume de méls a rendu leur traitement automatique indispensable, notamment pour détecter les pourriels (Tang et al. 2013). L'exploration repose sur différentes méthodes. Elles peuvent utiliser des règles (Xia, 2020), de l'apprentissage (Nadjate et al., 2020) ou la combinaison des deux. L'efficacité des extracteurs automatiques des termes (Nazarenko et al. 2009) nécessite une évaluation. La catégorisation des courriels a été mobilisée dans le but de les organiser.
La question des contacts a aussi été explorée et touche à l'analyse du contenu : catégorisation de contacts ayant le même centre d'intérêt (Johansen, 2007) ; identification de contacts appartenant à une même communauté (Tyler et al., 2003).
Les interactions par méls pouvant être appréhendées comme un réseau, elles ont été explorées, en se fondant sur les statistiques des contributions de chaque contact au sein du réseau (Karagiannis, 2009) ou par l'apprentissage de l'objectif des méls envoyés, pour informer, enquêter et planifier (Lockerd, 2003). La détection d'évènements à partir des courriels a aussi été prospectée permettant l'identification de ceux (recrutement, discours, événement social) mentionnés dans les textes avec les détails comme la date, l'heure, le lieu. Des méthodes de reconnaissance d'entités nommées (Suárez et al. 2020) sont utilisées pour extraire ces informations qui sont ensuite soumis à validation (Nair et al., 2020).
Côté archives, le projet RATOM (université de Caroline du Nord) a travaillé à l'extraction d'entités nommées au moyen de librairies Taln dans un double but : identifier des informations potentiellement sensibles et préparer la diffusion. Notons que cette question de l'accès aux contenus archivés – recherche au sein des corpus et communication à la demande – a été peu abordée, sauf aux États-Unis.

1.3. Analyse des besoins et solutions proposées

Les messageries des ministères et collectivités sont généralement produites via Microsoft Outlook et peuvent être exportées au format propriétaire pst. Si on souhaite ultérieurement consulter ces pst, il faut les réimporter dans Outlook, sous licence commerciale, ou utiliser une visionneuse pst, ce qui limite considérablement l'accès et l'utilisation de ces conteneurs. En octobre 2022, Archifiltre© a développé une proposition pour visualiser les pst Outlook uniquement à l'unité : Archifiltre Mail©.

Une fois les boîtes migrées dans un format plus pérenne (eml), aucune solution n'est actuellement disponible. De même les visualisations sont restreintes à une et une seule boîte mél.

Les messageries sont de fait archivées dans des silos séparés. Un format conteneur type pst une fois intégré dans un SAE est une boîte noire : impossible de savoir ce qu'elle contient réellement (fig. 1).



Messagerie électronique de Marie-Pierre Bouchaudy, chargée de mission pour l'action territoriale, au sein du cabinet d'Audrey Azoulay, ministre de la Culture et de la communication de 2016 à 2017.
Répertoire numérique détaillé n°20180394
établi par Hélène Brossier, archiviste à la Mission des archives du ministère de la Culture, sous la direction de Patrice Guérin, chef de la Mission.

Contexte de l'unité de description :

Messagerie électronique de Marie-Pierre Bouchaudy, chargée de mission pour l'action territoriale, au sein du cabinet d'Audrey Azoulay, ministre de la Culture et de la communication de 2016 à 2017. **20180394/1**

Unité de description :

20180394/1

Messagerie électronique

Deux fichiers au format pst :

- envoye.pst, 514 Mo
- reception.pst, 1,21 Go

Figure 1: Exemple d'une description de messagerie [en ligne] sur le site des Archives nationales (consulté le 19 août 2022)

Or sans accès aux contenus des messages et des pièces jointes, il s'avère impossible :

- d'évaluer l'intérêt à long terme de chaque boîte mél comme des catégories de message qu'elle contient
- de jauger de l'originalité d'une boîte mél dans le réseau de méls de son organisation ou de son rôle pivot, autant d'éléments déterminants dans le processus de patrimonialisation des courriels
- d'effectuer un tri interne efficient
- de garantir l'exclusion des messages à caractère personnel inévitablement présents dans des messageries pourtant à caractère professionnel
- de proposer une description précise et une indexation pertinente
- d'analyser le flux informationnel d'une organisation
- de répondre aux besoins de recherche à caractère interne ou externe

De fait, les dispositifs socio-techniques déjà mis en place assurent la préservation mais ne prennent pas en compte l'accès au contenu des messages et pièces jointes. C'est une des raisons qui freine l'archivage des messageries.

La démarche proposée allie des méthodes d'apprentissage automatique pour la classification et la catégorisation des contenus textuels et de l'intervention humaine dans un processus coopératif. Nous mettons en œuvre une méthode composée des principaux modules suivants (fig. 2) :

- Un module de prétraitement : découpage de courriels, élaboration de réseaux de contacts, extraction des informations comme la fonction, le rattachement des personnes physiques ou morales à partir des adresses méls à relier avec les annuaires et les signatures. Une base de données qui contient l'ensemble de ces informations a été construite, ce qui permet de mettre en évidence les liens entre les personnes, les fonctions et les services et facilite aussi différentes visualisations graphiques.
- Un module d'analyse fine de textes : extraction de termes, d'entités nommées.
- Un module de classification thématique afin de parcourir les messages selon les différents thèmes. Il permet d'avoir une vue plus globale de l'ensemble des messageries.

Figure 2: Méthodologie suivie

Deux interfaces ont été développées dans le cadre de ce projet :

1. une interface de classification
2. une interface de visualisation avec des fonctions de base et des fonctions avancées

L'interface de classification est une application desktop en local sous Linux. Elle effectue les tâches suivantes :

- Pré-traitement et segmentation en phrases
- Extractions d'entités nommées et de termes, validations automatiques et manuelles

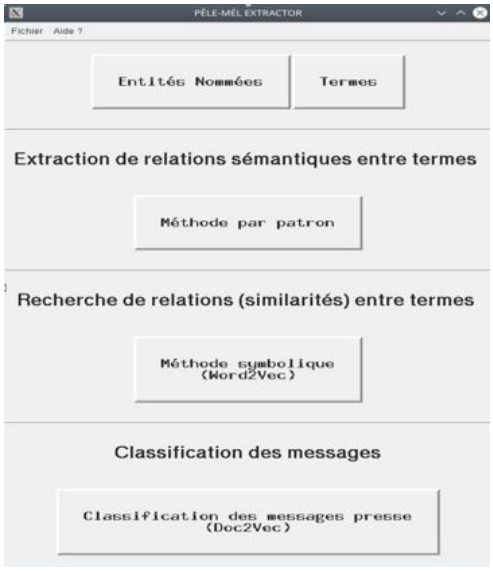


Figure 3: Interface de classification

- Création de relations entre thèmes et termes et classification avec un modèle pré-entraîné : création des nuages de mots
- Association des messages avec des thématiques
- Recherche de similarités en s'appuyant sur une représentation mathématique de vecteurs de mot ou de message

L'interface de visualisation (fig. 4) est une application desktop en local sous Linux ou Windows. Elle permet de s'orienter dans les messages et pièces jointes d'une ou de plusieurs messageries. Il est possible d'interroger les métadonnées des messages, leur contenu, filtrer les résultats à partir de nombreux critères, visualiser les relations entre une adresse et ses correspondants sous forme de graphe dynamique, visualiser et s'orienter dans la classification. Via la base de données, l'archiviste peut modifier les tables, annoter et corriger les classifications.

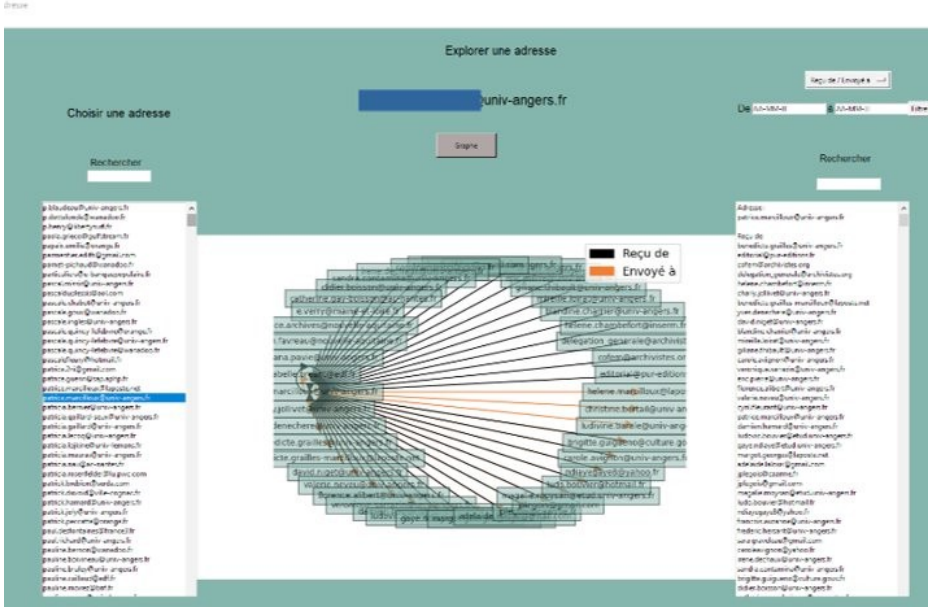


Figure 4: Interface d'exploration

2. Méthodologie d'étude des usages

Pour évaluer les usages et anticiper de futurs développements, nous avons fait le choix d'une étude qualitative auprès d'un public-cible d'utilisateur.rices potentiel.les, intermédiaires entre les ressources archivistiques (les messageries) et les publics demandeurs.

2.1. Approche qualitative et quantitative

Nous avons volontairement restreint notre champ à des archivistes déjà engagés dans des réalisations ou des projets d'archivage électronique, effectuant des collectes de messageries ou ayant à l'étude des collectes de messageries. Ces utilisateur.rices n'étaient pas familiarisés avec les traitements automatiques de langue naturelle ni avec des méthodes de classification supervisée ou non supervisée.

La méthodologie choisie est celle des focus groups ou groupes focalisés ou groupes de discussion. Il s'agit d'une technique d'entretien de groupe semi-structurée, modéré par une animatrice en présence d'un observateur. Elle permet de recueillir des données sur un nombre limité de questions en explorant et stimulant des points de vue par la discussion. Les participant·es peuvent s'exprimer directement, produire des idées diverses, parfois contradictoires, controversées ou inattendues. La méthode est utile pour évaluer des expériences, besoins, attentes mais aussi des représentations. Elle permet d'approfondir la compréhension d'une question complexe difficile à mesurer par des indicateurs objectifs grâce à un éventail d'idées et de réactions personnelles. Chaque groupe va réagir différemment et le déroulé d'un groupe ne peut s'appliquer à un autre groupe. On recherche donc un effet de saturation.

À l'issue de ces ateliers, un questionnaire (annexe 1) a été envoyé aux participant·es pour recueillir leur appréciation générale sur le projet et son utilité. 10 questions ont été posées visant à faire le point sur l'outillage professionnel, la place du projet dans cet arsenal, sa réception et la perception des prototypes.

2.2. Organisation des groupes de discussion

Les ateliers, chacun d'une durée de trois heures, se sont déroulés en trois temps.

Une présentation générale était d'abord proposée. Elle visait à transmettre des éléments de connaissances et d'état de l'art, d'exposer certains concepts et notions complexes :

- entités nommées
- termes
- lemmatisation
- étiqueteur grammatical
- extracteur d'entités nommées ou de termes
- clustering / machine learning
- régularités et patrons
- plongement lexical ou plongement de mots
- plongement de documents
- modèle pré-entrainé
- méthode de scoring (notamment TF-IDF)

Puis les interfaces et prototypes étaient présentées, à l'aide de vidéos et d'une démonstration directe.

Enfin, un temps d'échange concluait l'atelier. Ce temps était structuré à l'aide de diapositives pour faire réagir les différents groupes sur les mêmes points tout en laissant la parole et les discussions libres entre participant·es. Les points abordés étaient les suivants :

- adéquation de la proposition aux besoins
- adaptation des pratiques actuelles en fonction des résultats acquis
- conséquences sur l'évaluation et la collecte
- conséquences sur le service de référence (aide à la recherche)
- modalités d'amélioration des instruments de recherche
- développements ultérieurs à privilégier

Quatre ateliers ont été organisés, trois en présentiel, un en visioconférence, en février et mars 2023. Ils ont réuni 36 personnes couvrant tous types de services d'archives, de l'échelon national à l'échelon local, services d'archives intermédiaires et services d'archives définitives, services de l'État, de collectivités territoriales, d'établissements publics (établissement public à caractère scientifique, culturel et professionnel ; établissement public à caractère administratif ; établissement public à caractère industriel et commercial).

3. Usages

L'appréhension des usages est fortement corrélée au degré d'expertise et à la place du service dans la chaîne de traitement archivistique (service d'archives intermédiaires ou service d'archives définitives).

3.1. Une acculturation nécessaire

Les archivistes disposant déjà d'un système d'archivage électronique (SAE) ou étant en cours d'acquisition de ce système ou en cours de test, semblent avoir perçu plus directement et plus concrètement les usages possibles. Mais ce constat semble amoindri quand la politique du service vise à prioriser le développement de profils d'archivage Seda et que l'expérience de collectes manuelles est restreinte ou différée.

Le prototype de classification qui enchaîne plusieurs étapes rappelées dans la figure 2 (prétraitement, extractions de termes et d'entités nommées, création des nuages de termes, classification des messages) et s'appuie sur des concepts et des opérations mobilisant des méthodes de Taln, a globalement paru complexe d'autant que l'ensemble des participant·es, à deux exceptions près, n'a pas d'expérience de l'environnement linux. Une des participantes a d'ailleurs exprimé ainsi ses réserves : « Tout ce qui concerne les préparations des corpus dans l'outil, ça fait un peu peur ». La question des compétences informatiques internes concourt à ces réticences. Plusieurs échanges sont allés dans ce sens, de même que les réponses à la question 3 (*Diriez-vous que vous êtes armé·e sur le plan technique pour aborder cette question ?*, questionnaire en annexe 1). Mais, paradoxalement, alors que les porteur·es de ce projet pensaient que les spécificités

techniques, les présupposés linguistiques et les méthodologies de classification pouvaient être un obstacle, si ils ont été perçus comme complexes, ils n'en ont pas moins suscité plus d'intérêt et de curiosité que de craintes.

On notera qu'à la question 5 (*La présentation proposée par l'université d'Angers vous a-t-elle paru:*, questionnaire en annexe 1), aucun·e répondant·e n'a choisi parmi les trois réponses suivantes :

- Peu utile : cela m'a paru sans rapport avec les questions soulevées par les messageries
- Décourageant : trop de complexité

et un·e seul·e :

- Pertinente mais non applicable dans mon type de service (taille)

La complexité, en l'occurrence réelle, a été interprétée comme inhérente aux objectifs poursuivis et donc justifiée.

De fait, à la question 6 (*Le prototype de classification (sous Linux) vous a paru*, questionnaire en annexe 1), les deux réponses les plus choisies sont :

- Difficile à mettre en œuvre : il me faudrait l'appui d'un·e informaticien.ne
- Stimulant : des technologies existent pour aborder de gros volumes de données

Alors que les réponses à une question comparable sur le prototype d'exploration (question 7, questionnaire en annexe 1) se sont orientées vers les options suivantes :

- Concret : j'aimerais me lancer dans une expérimentation
- Faisable : après quelques essais, je pense pouvoir maîtriser la démarche
- Stimulant : des technologies existent pour aborder de gros volumes de données

À l'issue des différents ateliers, il est apparu que les participant·es avaient évolué au cours de la séance dans leur approche. Par exemple, certain·es d'entre eux et elles n'étaient pas convaincus a priori et de l'intérêt de conserver les messageries et de l'existence avérée ou potentielle de demandes d'accès. Le fait de voir que certains services étaient plus avancés sur ce sujet et de prendre acte par exemple de la collecte massive de messageries du gouvernement Castex (500, soit 10 TO) comme de la publication de plusieurs saisines de la Commission d'accès aux documents administratifs (CADA) et de l'accélération de celles-ci, a modifié leur regard, comme les échanges à bâton rompu. Par exemple :

> Participant 1 : pour beaucoup d'acteurs, le mél, c'est pas important. En même temps, ils disent que c'est là qu'on trouve les décisions.
>
> Participant 1 : C'est vrai que [les agent·es] n'ont pas conscience de ce qu'ils écrivent.
> Participant 2 : ça voudrait donc dire que c'est intéressant en fait.

Participant 1 : j'avais pensé archiver les boîtes du président et du DGS, mais finalement on pourrait tout récupérer.

3.2. Une approche déstabilisante

Le programme a mis en lumière des manques dans les collectes d'archives pour une contextualisation réelle des messageries et des messages :

- absence de traçabilité des listes de diffusion ou des adresses fonctionnelles

- pas d'archivage systématique des annuaires Ldap

- des hyperliens pointant dans le vide sur des fichiers initialement stockés sur des serveurs mais détruits ou déplacés, ou sur des sites internet

- importance des contenus diffusés via les listes

- usage des « copies à » comme forme de validation.

Il a montré en revanche sa capacité à exploiter les contenus des pièces jointes.

Or, beaucoup de services d'archives recommandent d'utiliser des hyperliens par exemple, pour alléger le flux et les volumes en circulation. La plupart n'avaient jamais envisagé la question de la conservation des annuaires Ldap et n'avaient jamais discuté avec leur DSI de ce point précis.

De même, le programme a montré la limite dans la capacité de classement et d'organisation des messages par l'agent·e. Les services d'archives recommandent pour certains l'enregistrement des messages dans des dossiers d'affaire. Mais, de fait, cette injonction ne peut se traduire que de manière limitée.

Par ailleurs, beaucoup de services cherchent à sensibiliser les producteurs sur la nécessité d'adopter des comportements électroniques plus sobres, d'éviter les redondances et de dégager de l'espace en détruisant plus systématiquement. Dans le cadre de ces *cleaning day*, les boîtes méls sont souvent prises en exemple avec des préconisations d'élimination, par exemple des messages liées à des listes ou des messages reçus en copie.

Enfin, il apparaît que leur mode de sélection vise à répondre à des besoins probants : « s'occuper du probant, c'est déjà pas mal ». Des discussions se sont engagées sur la notion de messages engageants. En effet, on recommande aux producteurs de conserver les messages engageants et/ou de les sauvegarder. Or les participant·es ont abouti au constat que cette notion est appréciée de manière subjective. Pour certain·es agent·es, tous les messages sont engageants : « cela veut dire se couvrir. » Une des participantes relève que « l'outil proposé apporte de la neutralité et de l'objectivité. Engageant, c'est subjectif. Cela permet d'aller au-delà. »

Pour beaucoup de participant·es, les boîtes méls étaient vues comme des chronos de courrier. Or les boîtes méls sont un ensemble plus complexe qu'un chrono de courriers, les messageries répondant à des usages variés au-delà du simple échange de correspondance. Une des participantes s'interroge et met en avant une approche par processus plutôt que par outil. Elle a l'impression de tomber dans un logique d'outil en raisonnant à l'échelle des boîtes méls. Or, il existe des usages spécifiques aux messageries (comme l'auto-archivage ou une modalité de transferts entre environnements technologiques), internalisés dans l'outil et invisibles par les autres processus.

Les outils de classification peuvent contribuer, par une appréhension à une échelle macroscopique, à une évolution du système de représentations professionnelles qui orientent et guident les conduites professionnelles. A l'issue des séances, plusieurs participant·es ont indiqué s'interroger dès lors sur les consignes à donner. D'autres ont trouvé intéressant « ce qu'on peut tirer de la séance en termes de recommandations en amont ».

La question 9 (questionnaire en annexe 1) demandait d'expliciter en quoi l'atelier avait amené le ou la répondant·e à reconsidérer ses pratiques. Le verbatim qui suit recouvre l'essentiel des réactions :

- « Cet atelier m'a amené à revoir mes premières intentions de collecte qui voulaient que la collecte seule des boîtes méls de dirigeants était suffisante »
- « Va me permettre de mieux aborder la question de la gestion quotidienne des mails et la mise en place de bonnes pratiques »
- « L'approche qui consiste à aborder les boîtes mails en mode réseaux pour repérer des boîtes mails "pivots" me semble très pertinente et complémentaire de l'approche par profils (et la collecte des boîtes mails des décideurs : DGS, DGA, directeurs, cheffes de projets stratégiques par exemple) »

3.3. Des dimensions stimulantes

Les participant·es ont compris et admis le côté expérimental du projet qui se traduit dans des interfaces austères, et aimeraient bien évidemment des interfaces plus conviviales et surtout un développement de l'interface de classification sous Windows. D'une manière générale, Linux est inaccessible à l'écrasante majorité des services et une proposition s'appuyant sur des solutions techniques de type docker devrait pouvoir répondre à la demande.

Plusieurs participant·es ont également indiqué avoir effectué des tests avec des outils développés en langue anglaise et en anglo-américain (ePADD par exemple) et avoir constaté leur inefficacité sur des corpus en langue française. Il y a donc bien un réel besoin d'approches linguistiques spécifiques.

Les propositions quant à l'amélioration possible des instruments de recherche via la création d'annexes à celui-ci comme un dictionnaire des correspondants ou des graphiques de réseau, de fréquence simple ou avancée, et intégrant les thématiques

abordées, ont recueilli un vif intérêt. Certain·es ont souligné la nécessité de disposer de modalités de recherche innovantes. La proposition a été perçue comme un mode nouveau d'indexation de corpus.

Les quelques services qui ont déjà été confrontés à des demandes d'accès ont été sensibles à la possibilité d'utiliser le module de classification à la demande. Autrement dit, au lieu de lancer une classification à l'aide d'une liste de thèmes conçue a priori à partir des missions/attributions/procédures, il est possible de constituer un nuage de termes à partir de la soumission des mots clés de la demande d'accès et d'identifier les messages qui lui sont reliés.

Certain·es participant·es ont souligné le temps nécessaire à la préparation des boîtes méls pour la collecte comme pour répondre aux demandes d'accès. Un participant a reconnu l'intérêt de changer d'échelle en passant d'une appréhension à la boîte au réseau de boîtes. Mais la démarche lui semble lourde. Un autre ajoute : « c'est vraiment une grosse évolution ».

Le fait de pouvoir exporter les identifiants des messages se rapportant à une thématique est apparu comme efficient pour pouvoir répondre à des demandes de communication, y compris au titre des documents administratifs et préparer le corpus à consulter.

Plusieurs services ont demandé à bénéficier des prototypes pour pouvoir les tester. D'autres ont indiqué vouloir recommander notre démarche.

4. Les perspectives

Au terme de ce programme, nous avons pu identifier quelques facteurs de réussite mais aussi les freins existants et envisager des évolutions possibles ou utiles.

4.1. Facteurs clés de réussite et freins

Divers freins vis-à-vis des messageries ont pu être identifiés sur le plan technique, organisationnel mais aussi sur le plan des représentations.

Le fait de disposer d'outils de classification et de visualisation peut contribuer à faire tomber ces freins, en confirmant la validité de la démarche de sélection, en assurant ensuite l'accès et la recherche et la qualité des instruments de description archivistique.

La réplicabilité de la démarche est un point essentiel. Lors des ateliers, nous avons pris soin de distinguer ce qui était, à notre sens, réplicable de ce qui était spécifique à notre partenaire culturel.

Concernant la classification en elle-même, deux points ont été soulevés :

- du côté des services intermédiaires, comment procéder à l'égard des fonctions transversales (DRH, finances par exemple) ? L'établissement d'une liste de thématiques paraît un investissement considérable.

- du côté des services d'archives définitives, il paraît indispensable qu'une partie de la mise en œuvre soit assurée par les services d'archives intermédiaires.

Les échanges sur ces deux points ont mis en avant les éléments suivants :

- il existe déjà des outils comme les tableaux de gestion, les circulaires interministérielles de tri et d'élimination, les plans de classement de type *Records management*, qui permettent d'identifier des thématiques

- les fonctions transverses sont communes à tous les types de service, un travail collaboratif entre services peut être envisagé

- la connaissance fine du fonctionnement et des missions de l'organisation productrice est en effet un atout considérable et la présence de services d'archives intermédiaires, sans être une nécessité absolue, est un avantage et une garantie de qualité.

Un autre facteur de réussite est le travail collaboratif entre DSI et services d'archives, en proximité.

Un besoin de formation et d'accompagnement a également émergé. Il est global sur l'ensemble de la chaîne archivistique, depuis la sensibilisation des services à la collecte des messageries jusqu'à leur accès. Il apparaît également dans les réponses au questionnaire (annexe 1). Ce point est très important car un des freins identifiés est les représentations professionnelles. Or, elles sont susceptibles d'évoluer rapidement puisqu'un atelier de 3 h a permis de faire bouger les lignes auprès d'un public déjà sensibilisé.

4.2. Développements futurs ?

D'une manière générale, le public ciblé étant des archivistes, les discussions ont principalement porté soit sur des questions archivistiques, soit sur des questions d'ergonomie.

Une participante a résumé les enjeux de cette manière : « sur la question de l'accès, on ne peut plus attendre, cela devient urgent. » Ce constat est confirmé du côté des archives publiques mais aussi des archives privées. Une autre participante a souligné le caractère chronophage de la prise en charge puis de la recherche de messages. Elle indique un temps de préparation pour un référent archives de deux messageries par an, avec un traitement au dossier. Un des participants a indiqué avoir déjà eu à traiter plusieurs fois des demandes de communication : « on ne trouve jamais rien ». Dans ce contexte, toute aide est bienvenue et toute amélioration bonne à prendre.

Pour ce qui est des interfaces, les voies d'amélioration sont à chercher du côté de l'environnement système (unifier les environnements pour Windows). Il faut donc trouver des solutions techniques pour rendre accessible le prototype de classification sous Windows, sachant que les briques libres utilisées, des bibliothèques python, n'ont pas été développées ou n'ont pas d'équivalent pour cet environnement. Les porteur·es du projet ont évoqué une plateforme docker, qui permettrait de lancer des applications après les avoir empaquetées dans des conteneurs et augmenterait leur portabilité.

Certains participant·es aimeraient une étude comparative entre messageries, Ged et arborescence bureautique pour choisir une stratégie de collecte.

D'autres insistent sur la difficulté pratique d'identifier les messages d'ordre privé oubliés dans les messageries professionnelles. Des essais de classification avec le premier prototype pourraient être envisagé dans cette direction.

Un autre participant aimerait pouvoir traiter des boîtes méls de personnes qui occupent plusieurs fonctions et être en capacité de repérer ce qui relève de telle ou telle fonction. Il serait intéressant de tester le prototype avec cette focale. De même il voudrait un outil capable de rechercher le pourcentage de similitudes entre différentes boîtes méls.

Nous avons également proposé les pistes suivantes : création de résumés automatiques sur un lot de méls, amélioration des fonctions de recherche en générant une classification par plongement lexical puis plongement de documents, à partir de mots clefs proposés par l'utilisateur.rice (et en jouant sur un arbre de profondeur). Ces propositions ont suscité une forme d'étonnement : les participant·es n'avaient pas jusque là imaginé ce type d'approches comme possibles et applicables à leur domaine.

L'expérimentation s'est développée sur un terrain particulier mais les résultats sont généralisables à d'autres contextes et d'autres environnements.

Conclusion

Le dernier item du questionnaire (annexe 1) proposait aux répondant·es de citer des mots clefs en rapport avec le projet (fig. 5).

Figure 5: Réponse à la question 10. *A l'issue de cet atelier, quels sont les mots clefs que vous mettriez en avant ? (merci de saisir à la suite quelques mots)*

Le nuage de termes créé à partir de ces réponses (fig. 5) résume parfaitement le contenu du projet (boites méls pivots, indexation, facettes, apprentissage supervisé, langage, classement, classification) comme sa perception (stimulant, complexe, simplicité, efficacité, pertinent, utile, innovant, intéressant).

Les mots, placés au centre, sont innovation et recherche.

Les participant·es ont été sensibles aux apports du programme Pêle-mél.

Tout d'abord, dans les méthodes proposées. Il s'agit de la première adaptation de méthodes symboliques utilisant les réseaux de neurones au cas de corpus de méls en français. Il s'agit aussi d'une expérimentation concrète aboutissant à une stratégie précise de mise en œuvre.

Ensuite, par la manière d'appréhender la question des méls, non à l'échelle de l'unité de la boîte mais à celle du réseau des boîtes d'une organisation.

Enfin, par le choix d'utiliser la classification pour permettrre l'accès au contenu des méls. L'unité sur laquelle il a été choisi de réfléchir est constitué du message et des pièces jointes et pas du message seul. Autre innovation, le thème est délimité non par un mot clef ou des mots clefs mais bien par un nuage de termes préalablement extraits et reliés par une recherche de similarités via une méthode de plongement lexical.

Cet intérêt a été partagé par la communauté internationale. Le programme a donné lieu à deux communications dans des congrès internationaux :

- Touria Aït El Mekki, Bénédicte Grailles, Tsanta Randriatsitohaina. *Création d'une base de connaissances à partir de messageries spécialisées pour améliorer l'exploitation et l'archivage des méls*. 1, Vadistat press; Editzioni Erranti, pp.52-59, 2022, JADT 2022 ¨Proceedings of the 16th International Conference on statistical analysis of textual data [16e conférence internationale Statistical Analysis of textual Data]. ⟨hal-03914993⟩

- Bénédicte Grailles, Touria Aït El Mekki, Édouard Vasseur. *Improving the archiving and contextualization of electronic messaging in French*. p.374, 2022, Proceedings iPres [International Conference on Digital Preservation] 2022 Glasgow 12—16 September 2022. ⟨hal-03837632⟩

Glossaire

Apprentissage	Non supervisé (clustering), supervisé (machine learning)
Apprentissage non supervisé	Situation d'apprentissage automatique sans données étiquetées. L'objectif est d'extraire des classes d'individus partageant des caractéristiques communes.
Apprentissage supervisé	Situation d'apprentissage automatique à partir d'exemples annotés et de données étiquetées. Capacité d'apprendre une fonction de prédiction.
Clustering	Apprentissage non supervisé
Entité nommée	Mot ou groupe de mots de noms de personnes, noms d'organisations ou d'entreprises, noms de lieux, quantités, distances, valeurs, dates, etc.
Étiqueteur morpho-syntaxique	Associe aux mots d'un texte les informations grammaticales correspondantes comme la partie du discours, le genre, le nombre, etc.
Extracteur de termes	Outil établissant une liste de candidats-termes dans une masse au préalable indifférenciée d'unités lexicales
Fréquence « brute »	Nombre d'occurrences d'un terme dans un corpus. Voir TF-IDF
Machine learning	Apprentissage supervisé
Plongement de document	Permet de créer une représentation numérique d'un document par un vecteur de nombres réels. Des documents utilisés dans un contexte similaire seront représentés par des vecteurs proches. L'approche est prédictive : un document prédit son contexte et un contexte prédit un document.
Plongement lexical	Permet de créer une représentation numérique de chaque mot d'un corpus par un vecteur de nombres réels. Des mots utilisés dans un contexte similaire seront représentés par des vecteurs proches. L'approche est prédictive : un mot prédit son contexte et un contexte prédit un mot. On parle également de plongement de mots.
Regex	Regular expressions ou expressions régulières ou motifs, permettent de décrire selon une syntaxe précise ou règles un ensemble souhaité de chaînes de caractères
Réseau de neurones	Modèle informatique dont la structure en couches est similaire à la structure en réseau des neurones du cerveau, avec des couches de

	nœuds connectés. C'est une structure capable d'apprendre par l'expérience, qui peut être entraînée à reconnaître et à classer
Terme	Tout mot ou groupe de mots représentant un concept spécifique d'un domaine
TF-IDF	*term frequency-inverse document* frequency. Schéma de pondération qui mesure l'importance d'un terme dans l'ensemble du corpus. Des termes moins fréquents sont considérés comme plus discriminants et améliore la pertinence.

Orientation bibliographique

Avertissement : ne sont mentionnées ici que les publications citées dans le présent document et celles des contributeur·rices du projet. Elle ne prétend donc pas à l'exhaustivité.

Aït el Mekki T., Grailles B., Randriatsitohaina T. (2022), Création d'une base de connaissances à partir de messageries spécialisées pour améliorer l'exploitation et l'archivage des méls. 1, Vadistat press; Editzioni Erranti, 2022, JADT 2022 ¨Proceedings of the 16th International Conference on statistical analysis of textual data [16e conférence internationale Statistical Analysis of textual Data]. 〈hal-03914993〉

Akmouche C. (2022). PELE-MEL : Plate-forme d'exploration, de livraison et d'évaluation des méls. Extraction et classification de connaissances à partir d'une base de connaissances et un corpus de messageries. Rapport de stage de master 2 Informatique – Intelligence décisionnelle, Université d'Angers, 40 p.

Bretesché S., de Geffroy B., de Corbière F. (2018). E-bureaucratie: le travail emmailé des cadres, Paris: Presses des Mines.

ePADD Project Website Homepage, University of Standford. [https://library.standford.edu/projects/epadd]

Grailles B., Aït el Mekki T., Vasseur É. (2022). Improving the archiving and contextualization of electronic messaging in French. 2022, Proceedings iPres [International Conference on Digital Preservation] 2022 Glasgow 12—16 September 2022. 〈hal-03837632〉

Grailles B., Aït el Mekki T. (2022). Pêle-mél. Plate-forme d'exploration, de livraison et d'évaluation des méls. Rapport de recherche, Angers, 64 p..

Johansen L., Rowell M., Butler K. R., and McDaniel P. D. (2007). Email Communities of Interest. In CEAS The Fourth Conference on Email and Anti-Spam, 2-3 August 2007, USA.

Karagiannis T., and Vojnovic M. (2009). Behavioral profiles for advanced email features. In Proceedings of the 18th international conference on World Wide Web, pp. 711-720.

Lockerd A., and Selker T. (2003). DriftCatcher: The Implicit Social Context of Email. In INTERACT'03, pp. 813-816.

Nadjate S., Adi K. and Allili M. (2020). Semantic Representation Based on Deep Learning for Spam Detection. Foundations and Practice of Security, pp. 72-81.

Nair A. M., Justus A. A., Ramesh A., and Rajan B. (2020). Event Extraction from Emails. International Journal of Computer Applications, vol. 176, n°41, pp. 1-8.

Nazarenko A., Zargayouna H., Hamon O. and Puymbrouck J. (2009). Évaluation des outils terminologiques : enjeux, difficultés et propositions. Revue TAL, ATALA, vol. 50, pp. 257-281.

Programme VITAM (2013). L'archivage des messageries électroniques. Preuve de concept VITAM, Paris: Ministère des Affaires étrangères/Ministère de la Culture/Ministère de la Défense.

Prom C. (2019). Preserving email. 2nd ed., London: DPC. [https://www.dpconline.org/docs/technology-watch-reports/2159-twr19-01/file]

RATOM Project Website Homepage, University of North Carolina. [https://ratom.web.unc.edu/]

Suárez P., Dupont Y., Muller B., Romary L., Sagot B. (2020). "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]

Tang G., Pei J., and Luk W. S. (2014). Email mining: tasks, common techniques, and tools. Knowledge and Information Systems, vol. 41, pp. 1-31.

Tyler J.R., Wilkinson D.M., Huberman B.A. (2003). Email as Spectroscopy: Automated Discovery of Community Structure within Organizations. In Huysman M., Wenger E., Wulf V. (eds) Communities and Technologies. Dordrecht, pp. 81-96.

Xia T. (2020). A Constant Time Complexity Spam Detection Algorithm for Boosting Throughput on Rule-Based Filtering Systems. IEEE Access, vol. 8, pp. 82653- 82661.

Zerez T. (2022). Stage Apprentissage, traitement et visualisation des messageries. Rapport de stage de M1 Informatique, Université d'Angers, 20 p.

Annexe – Questionnaire diffusé aux participant·es

Atelier messageries

Chères et chers collègues,

Vous avez participé à un atelier autour de la question des messageries électroniques proposé par l'université d'Angers où vous ont été présentés les résultats des projets Pêle-mél (Plateforme d'évaluation, de livraison et d'exploration des méls) et CARAméls (Comprendre les administrateurs et leur rapport à leurs méls).

Nous aimerions recueillir votre avis.

Ce questionnaire ne vous prendra que quelques minutes. Aucune donnée personnelle n'est collectée.

Contact : benedicte.grailles[at]univ-angers.fr

Début : 1 / 2

1. La question de la collecte des messageries vous semble-t-elle d'actualité ?

- [ ] Forte actualité
- [ ] D'actualité mais parmi d'autres priorités
- [ ] Pas prioritaire
- [ ] Accessoire

2. La question de l'accès aux messageries (recherches internes ou demandes externes) vous semble-t-elle d'actualité ?

- [ ] Forte actualité
- [ ] D'actualité mais parmi d'autres priorités
- [ ] Pas prioritaire
- [ ] Accessoire

3.Diriez-vous que vous êtes armé.e sur le plan technique pour aborder cette question ? (plusieurs réponses possibles)

- [ ] Oui : on a tous les outils nécessaires
- [ ] Oui mais il reste des difficultés
- [ ] Oui mais j'aurais besoin de compléter ma formation
- [ ] Non : trop de complexité
- [ ] Non : pas assez d'appui DSI ou de compétences internes informatiques
- [ ] Non : je ne suis pas formé.e

4. Diriez-vous que vous êtes armé.e sur le plan archivistique pour aborder cette question ? (plusieurs réponses possibles)

- [ ] Oui : on a tous les outils nécessaires
- [ ] Oui mais il reste des difficultés
- [ ] Oui mais j'aurais besoin de compléter ma formation
- [ ] Non : trop de complexité
- [ ] Non : pas assez d'appui DSI ou de compétences internes informatiques
- [ ] Non : je ne suis pas formé.e

5. La présentation proposée par l'université d'Angers vous a-t-elle paru (plusieurs réponses possibles)

- [ ] Utile : cela a fait écho à des questions que je me posais ou des problèmes déjà rencontrés
- [ ] Intéressante : j'ai découvert des comportements utilisateurs ou cela a recoupé des observations que j'avais déjà faites
- [ ] Innovante : j'ai découvert des technologies de traitement nouvelles
- [ ] Peu utile : cela m'a paru sans rapport avec les questions soulevées par les messageries
- [ ] Décourageant : trop de complexité
- [ ] Pertinente mais non applicable dans mon type de service (taille)
- [ ] A approfondir quand je disposerai d'un SAE en production
- [ ] Applicable : certains points peuvent déjà être mis en oeuvre

6.Le prototype de classification (sous Linux) vous a paru (plusieurs réponses possibles)

- [ ] Complexe : il faut maîtriser beaucoup de concepts nouveaux
- [ ] Concret : j'aimerais me lancer dans une expérimentation
- [ ] Inadapté : je ne vois pas quel saut qualitatif il me permet de franchir
- [ ] Faisable : après quelques essais, je pense pouvoir maîtriser la démarche
- [ ] Stimulant : des technologies existent pour aborder de gros volumes de données
- [ ] Difficile à mettre en oeuvre : il me faudrait l'appui d'un.e informaticien.ne

7. Le prototype d'exploration (sous Windows) vous a paru (plusieurs réponses possibles)

- [ ] Complexe : il faut maîtriser beaucoup de concepts nouveaux
- [ ] Concret : j'aimerais me lancer dans une expérimentation
- [ ] Inadapté : je ne vois pas quel saut qualitatif il me permet de franchir
- [ ] Faisable : après quelques essais, je pense pouvoir maîtriser la démarche
- [ ] Stimulant : des technologies existent pour aborder de gros volumes de données
- [ ] Difficile à mettre en oeuvre : il me faudrait l'appui d'un.e informaticien.ne

8. Cet atelier vous a-t-il amené à reconsidérer vos pratiques actuelles ?

- [ ] Oui
- [ ] Non
- [ ] Je ne sais pas encore

9. Pourriez-vous préciser ?

10. A l'issue de cet atelier, quels sont les mots clefs que vous mettriez en avant ? (merci de saisir à la suite quelques mots)