Etude

Analyse du besoin

Visa TM

Vers une infrastructure de services avancés de text mining

2017 / 2019

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR,
DE LA RECHERCHE
ET DE L'INNOVATION

Analyse du besoin

Livrable Etude – partie 1

| Définir concrètement les besoins de la communauté de recherche française en matière de fouille de textes afin d'y adapter une offre de service adéquate |

Description du Document

Analyse du besoin

Lot	Etude
Participants	INIST (CNRS) MaIAGE (INRA) DIST (INRA)
Date de livraison	31/10/2019
Nature : Rapport	Version : 1.0

Contributeurs

	Nom	Organisation
Rédaction	Claire Nédellec	MaIAGE (INRA)
	Fabienne Kettani	INIST (CNRS)
	Marie-Sophie Nourdin	INIST (CNRS)
Coordination	Fabienne Kettani	INIST (CNRS)
Relecture	Clément Jonquet	LIRMM (Université de Montpellier)
	Sophie Aubin	DIST (INRA)
	Ludovic Hamiaux	INIST (CNRS)

SOMMAIRE

AVERTISSEMENT 1
ACRONYMES ET SIGLES 2
RESUME PUBLIABLE 3
INTRODUCTION 4
CHAPITRE 1 CONTEXTE 5
1.1 Une maturation des technologies 5
1.2 Des politiques en faveur de la science ouverte, des fonds pour la soutenir et des lois pour « ouvrir des possibles » 6
1.2.1 Loi pour une république numérique 6
1.2.2 Plans Science ouverte 8
1.2.3 Directive Européenne sur le droit d'auteur 9
1.3 Un mouvement pour la Science citoyenne 9
1.4 Un environnement favorable 9
CHAPITRE 2 QUESTIONNAIRE 11
2.1 Méthodologie 11
2.2 Le questionnaire 11
2.3 Diffusion 11
2.4 Analyse des résultats 12
2.4.1 Premiers éléments 12
2.4.2 La fouille de textes dans vos activités 13
2.4.3 Compétences et formations 16
2.4.4 Freins et leviers 18
2.4.5 Fonctionnalités d'une infrastructure 19
2.4.6 Quelques verbatim à retenir 20
2.4.7 Profil des répondants 21
2.4.8 Synthèse 23
CHAPITRE 3 BESOINS EN FONCTION DES TYPES D'UTILISATEURS 25
3.1 Chercheurs/experts en TAL/fouille de textes 25
3.2 Chercheurs non TAL/fouille de textes et acteurs de l'appui à la recherche 26
3.3 Décideurs 27
3.4 Fournisseurs de contenus 28
3.5 Agrégateurs de contenus (ex : OpenAire, CORE, AGRIS) 28
CHAPITRE 4 BESOINS SPECIFIQUES EN FONCTION DES COMMUNAUTES 30

4.1 Agriculture/biodiversité 30
4.2 Médecine 30
4.3 Chimie 31
4.4 Psychologie de la mémoire 32
CHAPITRE 5 FOCUS SUR UN BESOIN PARTICULIER : LA CONSTITUTION DE CORPUS 34
5.1 Introduction 34
5.2 Motivation 35
5.3 Étapes de la conception de corpus 35
5.3.1 Données du problème 35
5.3.2 Moyens humains et matériels 36
5.3.3 Méthode 36
5.4 Besoins 43
5.4.1 Information sur les accès aux sources 43
5.4.2 Licences et abonnements 43
5.4.3 Documentation des formats 43
5.4.4 Centralisation et standardisation des métadonnées bibliographiques et d'accès 43
5.4.5 Uniformisation et centralisation des accès aux documents 44
5.4.6 Confidentialité et qualité de l'accès aux documents 44
5.4.7 Documentation et standardisation des formats 44
5.4.8 Partage 44
CONCLUSION 45
INDEX DES FIGURES 46

Avertissement

Ce document contient des descriptions des résultats du projet Visa TM. Certaines parties peuvent être soumises à des droits de propriété intellectuelle. Avant réutilisation du contenu, il est nécessaire de contacter le consortium pour approbation.

Acronymes et sigles

TAL	Traitement Automatique des Langues
IST	Information Scientifique et Technique
FAIR	Findable Accessible Interoperable Reusable
RDA	Research Data Alliance
CoSO	Comité pour la Science Ouverte

Résumé publiable

Le projet Visa TM vise à étudier les conditions de mise à disposition de services de fouille de textes pour les chercheurs français. En effet, la fouille de textes est devenue ces dernières années un sujet d'importance dans le sillage du mouvement pour la science ouverte mais si ses enjeux sont importants pour la recherche et si le projet européen OpenMinTeD a posé les jalons d'une première approche en proposant une infrastructure dédiée, il reste à initier une démarche similaire à un niveau national. Dans cette optique le volet Etude va permettre d'imaginer et décrire l'infrastructure technique et humaine nécessaire en se basant à la fois sur l'expérience OpenMinTeD et sur les besoins de la communauté de recherche nationale. Cela nous amènera à élaborer des recommandations après analyse des points forts et faibles des différentes approches possibles. Dans ce document nous avons donc tout d'abord présenté le contexte dans lequel s'inscrit le projet Visa TM, en mettant en avant les éléments favorables actuels puis nous avons développé les besoins des utilisateurs potentiels d'une plateforme, en distinguant les variations suivant les différentes typologies d'utilisateurs et les différentes communautés de recherche. Nous avons approfondi ces besoins en nous appuyant sur un questionnaire qui nous a permis de recueillir des avis sur les fonctionnalités attendues d'une e-infrastructure dédiée à la fouille de textes, sur les connaissances des répondants sur le sujet et l'étendue actuelle de l'utilisation de la fouille de textes dans les activités de recherche tout comme les freins à cette utilisation. Cela nous a permis également d'identifier les compétences existantes et les besoins en formation qui y sont liés. Enfin, nous avons mis en avant un besoin particulier et très partagé à savoir la constitution de corpus.

Dans la suite de notre étude, dans le document « Acteurs et organisation » nous nous attacherons à décrire les différents acteurs qui composent aujourd'hui le paysage de la fouille de textes et les différentes solutions organisationnelles pouvant paraître pertinentes pour la mise en place d'une infrastructure nationale. Puis nous nous pencherons plus précisément sur une solution choisie et déclinerons ses missions ainsi que les métiers qui y sont liés et les compétences qu'ils supposent. Enfin nous irons approfondir des aspects plus techniques avec un panorama des outils existant et les critères permettant de décider ou non de les intégrer dans la future plateforme selon les besoins énoncés et leur service rendu, leur interopérabilité avec d'autres.

Pour finir sur le volet Etude, nous essayerons d'esquisser selon quelles modalités les communautés de recherche peuvent interagir autour de la fouille de textes dans une optique de partage et de mise en commun des outils et des expériences afin d'enrichir et de faire évoluer collectivement la plateforme au service de tous.

Introduction

Le paysage actuel de la fouille de textes et de données en France (tout comme à un niveau plus large européen voire international) est assez disparate de plusieurs points de vue : laboratoires spécialisés en traitement automatique du langage, en fouille de textes, spécificités disciplinaires, utilisateurs divers allant du chercheur au décideur en passant par des fournisseurs de contenus ou d'outils. Il est donc aisé de comprendre que les besoins de ces différents acteurs puissent être assez variés. Nous les étudierons ici en nous inspirant de la démarche déjà suivie dans OpenMinTeD pour récolter ces besoins en leur apportant un éclairage plus national. Cette approche comporte plusieurs étapes :

- Elaboration d'un questionnaire destiné à qualifier les pratiques actuelles autour de la fouille de textes et les besoins exprimés
- Analyse des besoins spécifiques de certaines communautés disciplinaires
- Analyse des besoins en fonction des typologies d'utilisateurs
- Analyse d'un besoin particulier : la constitution de corpus

Mais avant cela nous allons voir dans un premier chapitre dans quel contexte s'inscrivent aujourd'hui ces besoins en fouille de textes, comment ils sont juridiquement et politiquement encadrés, éventuellement économiquement soutenus et comment les progrès de l'intelligence artificielle contribuent à une maturation des outils nécessaires à leur mise en œuvre.

Contexte

Plus personne aujourd'hui ne peut ignorer la croissance exponentielle de la circulation des données sous toutes leurs formes. Ces données massives, communément connues sous le nom de *Big data*, sont devenues à la fois inaccessibles à l'entendement humain mais aussi au traitement par des outils informatiques classiques. Elles nécessitent donc la mise en œuvre de moyens spécifiques pour en extraire les contenus et les rendre à nouveau exploitables pour leur analyse. Lorsqu'il s'agit de données textuelles, la fouille de textes est ainsi le moyen d'extraire des connaissances de ces énormes réservoirs en s'appuyant sur la linguistique, les statistiques, l'informatique et plus largement sur l'intelligence artificielle. Elle est soutenue aujourd'hui par une démarche d'ouverture s'appliquant en particulier au logiciel libre dont le domaine du traitement automatique du langage a pu bénéficier ces dernières années. Mais si cette inflation de données pose des défis techniques, elle suscite également des interrogations du point de vue juridique quant à leurs droits d'exploitation dans un souci de maintien de la protection des droits d'auteur, des droits des bases de données où elles sont entreposées et plus spécifiquement des droits des personnes, tout en favorisant la science ouverte. Enfin, ces activités doivent être encadrées par des politiques publiques adaptées et reposer sur un modèle économique adéquat, aussi bien au niveau national qu'international.

Nous exposerons donc brièvement dans ce chapitre quelques avancées majeures survenues dans ces domaines durant les dernières années et qui contribuent progressivement à rendre possible et encadrer l'activité de fouille de textes.

1.1 Une maturation des technologies

Le traitement des données massives est un enjeu technologique important tant en termes d'architecture à mettre en place que d'organisation des processus et de respect des cadres légaux qui encadrent ces traitements. Cet enjeu se concentre essentiellement sur trois aspects de la prise en charge des données : leur volume important, leur diversité et leur rythme de croissance exponentiel. Une bonne gestion de cette prise en charge conditionne la justesse de leur exploitation ultérieure et des résultats produits.

Les apports des technologies de l'intelligence artificielle ont permis de faire évoluer cette gestion de manière appropriée. Les traitements sont opérés par des plateformes souvent *open source* permettant la prise en compte de données structurées ou non, de les rendre interopérables avec d'autres et de garantir leur sécurité. Des moyens de calcul conséquents doivent aussi être dédiés à ce type de projet pour l'exploitation des données.

L'ensemble de ces nouvelles technologies mises en œuvre permet aujourd'hui de rendre plus accessible la fouille de données dont fait partie la fouille de textes.

1.2 Des politiques en faveur de la science ouverte, des fonds pour la soutenir et des lois pour « ouvrir des possibles »

Une fois les technologies mises à profit pour faire de la fouille de textes, la barrière de l'accessibilité de ces textes reste un frein à leur pleine exploitation. En effet, dans le domaine de la recherche, les grands éditeurs scientifiques avaient jusque récemment encore un monopole de droits sur les articles scientifiques en rendant leur diffusion payante. Ce fait a connu une évolution au cours des dernières années sous la pression d'un mouvement plus global d'ouverture des données connu sous le nom d'***Open Data*** et lancé en 2009 par Tim Berners-Lee lors d'une conférence TED sous l'injonction : « Raw data now » (« Des données brutes maintenant »). Le terme *Open Data* était né quant à lui en 1995 dans une publication du Committee on Geophysical and Environmental Data sous l'impulsion de chercheurs qui souhaitaient partager les résultats de leurs recherches.

Le partage et l'ouverture des données porteront dans un premier temps sur les publications scientifiques dans le cadre de l'*Open Access* (archives ouvertes) officialisé en 2002 par l'**Open Access Initiative de Budapest** puis en 2003 par la **déclaration de Berlin sur le libre accès**.

En 2004, une déclaration sur l'accès aux données de la recherche financée par des fonds publics est signée par les gouvernements membres de l'Organisation de Coopération et de Développement Economique (OCDE) dont la concrétisation et les principes se poursuivent dans le **programme européen Horizon 2020**. Des politiques nationales de partage de données sont élaborées et des incitations de publication des données de recherche conjointement aux publications scientifiques sont mises en place.

En 2013 est créée la **Research Data Alliance** (RDA) associant la Communauté européenne, l'American Science Foundation, l'American National Institute of Standards and Technology et l'Australian Department of Innovation et dont l'objectif est de permettre à des groupes de travail d'experts interdisciplinaires, définis sur des périodes de 18 mois, d'élaborer des recommandations pour le libre partage des données et leur interopérabilité. Depuis 2018, le CNRS est en charge du nœud **RDA France**.

Ces différentes évolutions ont amené les politiques de données ouvertes et d'archives ouvertes vers une politique plus large de science ouverte, partagée aussi bien au plan national qu'européen voire international, mais dont les applications restent encore différentiées suivant les pays concernés.

1.2.1 Loi pour une république numérique

En France, la **Loi pour une République Numérique**[1] parue en octobre 2016 va faire évoluer les droits sur les publications scientifiques et restreindre ceux des éditeurs, en particulier au travers de l'***article 30*** qui stipule :

[1] https://www.legifrance.gouv.fr/affichTexte.do;jsessionid=E42256AE6A34647613E5F0CEDD403E90.tplgfr44s_1?cidTexte=JORFTEXT000033202746&categorieLien=id#JORFARTI000033202841

« I- Lorsqu'un écrit scientifique issu d'une activité de recherche financée au moins pour moitié par des dotations de l'Etat, des collectivités territoriales ou des établissements publics, par des subventions d'agences de financement nationales ou par des fonds de l'Union européenne est publié dans un périodique paraissant au moins une fois par an, son auteur dispose, même après avoir accordé des droits exclusifs à un éditeur, du droit de mettre à disposition gratuitement dans un format ouvert, par voie numérique, sous réserve de l'accord des éventuels coauteurs, la version finale de son manuscrit acceptée pour publication, dès lors que l'éditeur met lui-même celle-ci gratuitement à disposition par voie numérique ou, à défaut, à l'expiration d'un délai courant à compter de la date de la première publication. Ce délai est au maximum de six mois pour une publication dans le domaine des sciences, de la technique et de la médecine et de douze mois dans celui des sciences humaines et sociales.

La version mise à disposition en application du premier alinéa ne peut faire l'objet d'une exploitation dans le cadre d'une activité d'édition à caractère commercial.

II- Dès lors que les données issues d'une activité de recherche financée au moins pour moitié par des dotations de l'Etat, des collectivités territoriales, des établissements publics, des subventions d'agences de financement nationales ou par des fonds de l'Union européenne ne sont pas protégées par un droit spécifique ou une réglementation particulière et qu'elles ont été rendues publiques par le chercheur, l'établissement ou l'organisme de recherche, leur réutilisation est libre.

III- L'éditeur d'un écrit scientifique mentionné au I ne peut limiter la réutilisation des données de la recherche rendues publiques dans le cadre de sa publication.

IV- Les dispositions du présent article sont d'ordre public et toute clause contraire à celles-ci est réputée non écrite. »

L'***article 38*** de cette même loi va modifier les clauses de la propriété intellectuelle en introduisant une exception au droit d'auteur ainsi qu'une exception au droit *sui generis* des producteurs de bases de données :

Le code de la propriété intellectuelle est ainsi modifié :

1° Après le second alinéa du 9° de l'article L. 122-5, il est inséré un 10° ainsi rédigé :

« 10° Les copies ou reproductions numériques réalisées à partir d'une source licite, en vue de l'exploration de textes et de données incluses ou associées aux écrits scientifiques pour les besoins de la recherche publique, à l'exclusion de toute finalité commerciale. Un décret fixe les conditions dans lesquelles l'exploration des textes et des données est mise en œuvre, ainsi que les modalités de conservation et de communication des fichiers produits au terme des activités de recherche pour lesquelles elles ont été produites ; ces fichiers constituent des données de la recherche ; »

2° Après le 4° de l'article L. 342-3, il est inséré un 5° ainsi rédigé :

« 5° Les copies ou reproductions numériques de la base réalisées par une personne qui y a licitement accès, en vue de fouilles de textes et de données incluses ou associées aux écrits scientifiques dans un cadre de recherche, à l'exclusion de toute finalité commerciale. La conservation et la communication des copies techniques issues des traitements, au terme des activités de recherche pour lesquelles elles ont été produites, sont assurées par des organismes désignés par décret. Les autres copies ou reproductions sont détruites. »

Ces deux exceptions sont de nature à pouvoir favoriser la fouille de textes mais laissent apparaître également des limites dans leur application : exclusion des fins commerciales (limites entre commercial et non commercial ?), distinction entre « source licite » et « accès licite », limitation au texte, etc. Elles n'encadrent pas non plus les limites à la fouille de textes pouvant être relatives à des clauses contractuelles (licences) ou des freins techniques (limitations de déchargement de publications, archivage et circulation des copies intermédiaires,…). Enfin, elles ne sont pas en cohérence avec les dispositions adoptées dans d'autres pays.

Un guide d'application de la Loi pour une République Numérique édité ultérieurement en 2018[2] par des chercheurs, juristes et professionnels de l'information scientifique et technique a permis de clarifier les modalités d'application de la loi à destination des chercheurs. Parallèlement un guide concernant l'ouverture des données de recherche est également publié[3].

1.2.2 Plans Science ouverte

Plan national pour la Science ouverte

En juillet 2018 la France édite un **Plan national pour la Science ouverte**[4] dans lequel « La France s'engage pour que les résultats de la recherche scientifique soient ouverts à tous, chercheurs, entreprises et citoyens, sans entrave, sans délai, sans paiement ».

Un Fonds National pour la Science Ouverte d'un montant de 3,1 millions d'euros complète ce plan national et est annoncé par Frédérique Vidal en avril 2019.

Plan S

L'initiative française trouve un écho dans le Plan S[5] lancé par Science Europe[6] pour la promotion de l'édition scientifique en libre accès sous l'égide de la Commission Européenne et de la « cOAlitionS » (consortium soutenu par le Conseil Européen de la Recherche et des

[2] https://www.ouvrirlascience.fr/guide-application-loi-republique-numerique-article-30-ecrits-scientifiques-version-courte/
[3] https://www.ouvrirlascience.fr/ouverture-des-donnees-de-recherche-guide-danalyse-du-cadre-juridique-en-france-v2/
[4] https://www.ouvrirlascience.fr/plan-national-pour-la-science-ouverte/
[5] https://www.ouvrirlascience.fr/plan-s/
[6] https://www.scienceeurope.org/our-priorities/open-access

agences de financement de la recherche de 12 pays européens). Un guide d'application de ce plan est également paru fin 2018[7].

1.2.3 Directive Européenne sur le droit d'auteur

En parallèle, et après deux ans de débats, la Directive Européenne sur le droit d'auteur a été votée par le Parlement européen en mars 2019 et entérinée par le conseil de l'Union européenne en avril 2019, amenant sa pierre à l'édifice des possibilités ouvertes à la fouille de textes. En effet, la directive exclut de son champ d'application le text et data mining rendant celui-ci possible pour la recherche car non assujetti au droit d'auteur, pour permettre aux chercheurs européens de sortir du « désavantage compétitif » dans lequel ils se trouvaient par rapport à certains de leurs homologues internationaux. Cette directive doit maintenant être transposée dans les droits des différents pays européens. La France a été le premier pays à démarrer cette transposition par un vote en juillet 2019 par le parlement concernant les « droits voisins ». Les mesures concernant la fouille de textes devraient suivre cette initiative.

1.3 Un mouvement pour la Science citoyenne

Le concept de sciences citoyennes ou sciences participatives a vu le jour dans les années 1970 aux Etats-Unis (*citizen science*).

En France en 2002 est créée la Fondation pour les Sciences citoyennes[8] qui est une association loi 1901 ayant pour objectif de « favoriser et prolonger le mouvement actuel de réappropriation citoyenne et démocratique de la science, afin de le mettre au service du bien commun ».

D'autres suivront ultérieurement comme Regards citoyens[9] (2009) ou encore Savoirscom1[10](2012) qui vont à leur échelle œuvrer pour la partage et l'ouverture des biens communs de la connaissance et contribuer ainsi à la mise en place de conditions plus favorables pour la fouille de textes au service d'une science ouverte et réutilisable par le citoyen.

1.4 Un environnement favorable

La convergence des avancées précédemment passées en revue permet à la fouille de textes de trouver aujourd'hui un terreau fertile à son expansion. Son utilisation dans le domaine de la recherche, en-dehors des experts du domaine, nécessite sans-doute de consentir des efforts

[7] https://www.coalition-s.org/feedback-on-the-implementation-guidance-of-plan-s-generates-large-public-response/
[8] https://sciencescitoyennes.org/
[9] https://www.regardscitoyens.org/#&panel1-1
[10] http://www.savoirscom1.info/

sur la sensibilisation à ses champs d'application, la formation à son utilisation et la mise en place d'infrastructures de services facilitant un accès centralisé à l'ensemble de ses composantes. Il nous reste dans la suite de notre analyse à mettre en lumière également les besoins exprimés par les divers acteurs selon leur connaissance du sujet, leur domaine de spécialité ou encore leur utilisation de la fouille de textes qui sont des éléments que nous avons recueillis en particulier au travers d'un questionnaire.

Questionnaire

2.1 Méthodologie

Le questionnaire élaboré dans le cadre du projet Visa TM s'est inspiré de celui ayant servi dans le projet OpenMinTeD. Il est destiné à identifier les opportunités et qualifier les besoins des acteurs de la recherche en France en matière de services de fouille de textes à destination des chercheurs. Il vise à qualifier les pratiques autour de la fouille de textes et à identifier les verrous et les moyens de les lever.

Il a été adapté pour correspondre aussi bien à une cible peu familiarisée avec la fouille de textes qu'à des chercheurs spécialisés dans ce domaine dont c'est l'objet de recherche au quotidien. Ceci afin de récupérer une variété plus grande de besoins et obtenir ainsi une vision plus large de ces derniers, la plateforme visée dans le projet Visa TM ayant pour vocation de répondre aux besoins du plus grand nombre et d'amener la fouille de textes sur le « bureau » du chercheur.

2.2 Le questionnaire

La fouille de textes aujourd'hui … et demain ?
Ce questionnaire est découpé en plusieurs parties :

- La fouille de textes dans vos activités
- Compétences et formation
- Freins et leviers
- Comment verriez-vous les fonctionnalités d'une infrastructure de fouille de textes ?
- Vous…

Un certain nombre de questions sont conditionnelles afin de guider de manière optimale le répondant en fonction de sa compréhension du sujet.

Suivant le niveau de connaissances sur le sujet, la durée de rédaction de la réponse peut varier de 5 à 20 mn.

2.3 Diffusion

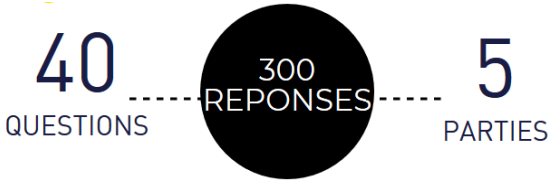
Plusieurs canaux de diffusion différents ont été mobilisés.

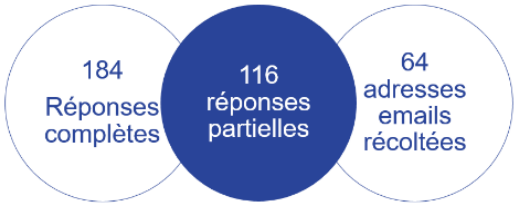
ACTEURS VISES	Formalisation du contact/de la diffusion (courriel simple, liste de diffusion,...)
ATALA	Liste de diffusion " LN", gérée par T. Hamon <thierry.hamon@lipn.univ-paris13.fr>
ADBU	tweet ciblé
EPST: CNRS, INSERM, IRSTEA, etc.	forum CoSO
EPRIST	tweet ciblé
INRA	tweet Sophie Aubin (IST INRA)
INIST	Blog Visa TM
	Twitter INIST
	Twitter CoSO
	Liste de diffusion LaLIST
Communautés spécifiques	Les Infos du Risc <Pourinfos@risc.cnrs.fr (sciences cognitives) par exemple
Recherche en ingénierie des connaissances	liste de diffusion écrire à info-ic@listes.irisa.fr ((https://sympa.inria.fr/sympa/info/info-ic))
Recherche en apprentissage automatique	écrire à : à liste-proml@lists.lri.fr (http://lists.lri.fr/cgi-bin/mailman/listinfo/liste-proml)
Masse de données, GDR Madics	utiliser le formulaire https://www.madics.fr/diffuser/annonce/
Bio-informatique	écrire à : bioinfo@sfbi.fr (http://listes.sfbi.fr/wws/info/bioinfo)
IA	bulle-I3, écrire à : bull-i3@irit.fr (https://www.irit.fr/GDR-I3/Abonnement.html)
Extraction et gestion de connaissance	liste-egc@polytech.univ-nantes.fr
Web sémantique	web.semantique@lists-sop.inria.fr
GDR TAL	mail
Direction Régionale du CNRS	CNRS Hebdo

2.4 Analyse des résultats

2.4.1 Premiers éléments

Compte-tenu de la diffusion somme toutes assez restreinte du questionnaire, le nombre de réponses obtenues nous paraît plutôt intéressant. La répartition de ces réponses (184 réponses complètes et 116 réponses partielles) est a priori le reflet de la possibilité de répondre aux questions selon deux parcours distincts : un parcours s'adressant plutôt à des personnes ayant une certaine maîtrise (voir expertise) du sujet et un parcours plus adapté à un interlocuteur novice en la matière. Par contre, les répondants ont peu utilisé la possibilité de déclarer nominativement qui ils étaient et dans quelle structure ils travaillaient (seulement 64 adresses mail récoltées) ce qui aurait pu être une opportunité pour les recontacter et envisager avec eux d'approfondir certains éléments de réponse.



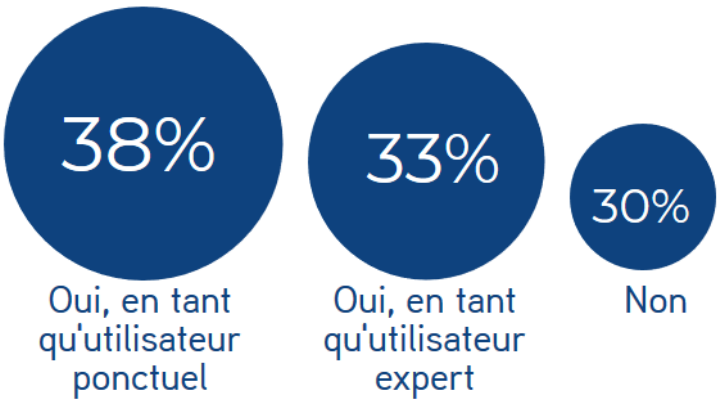


Nous constatons que les réponses complètes sont majoritairement le fait de trois catégories de personnes: les chercheurs, les enseignants-chercheurs et les fonctions d'appui. Compte-tenu des réponses trouvées dans la catégorie « Autres » nous pouvons supposer que la catégorie « Fonctions d'appui » n'a pas été cochée par un certain nombre de répondants qui auraient pu y figurer (documentalistes, ingénieurs d'étude et de recherche).

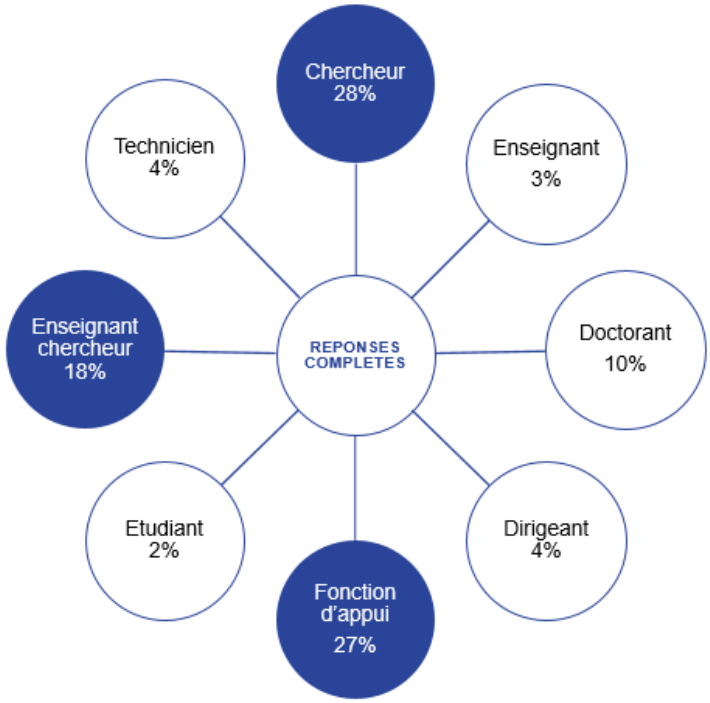
Après un aperçu rapide sur ces constations très générales, nous détaillons les réponses aux différentes thématiques abordées par le questionnaire.

2.4.2 La fouille de textes dans vos activités

Je pratique la fouille de textes…



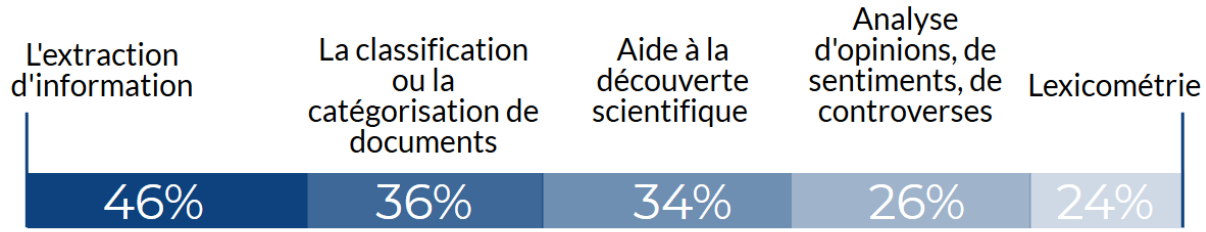
On note une très légère prépondérance de réponses du côté des utilisateurs ponctuels de la fouille de textes (38%) par rapport aux experts (33%) sachant que sur cette question seules 273 réponses ont été recueillies sur les 300.



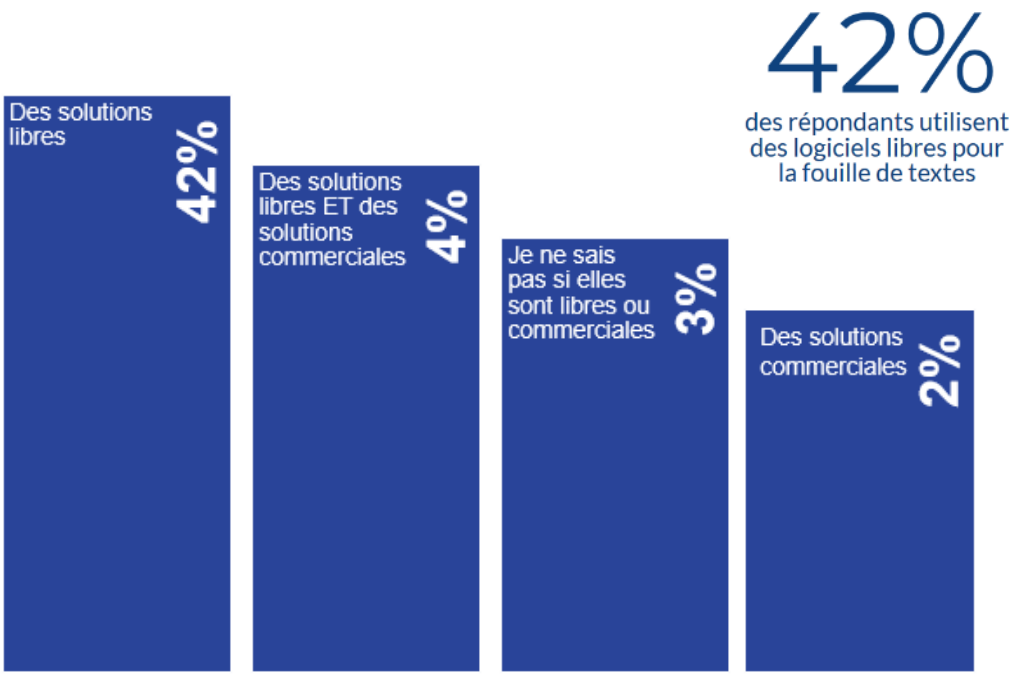
184

Réponses complètes

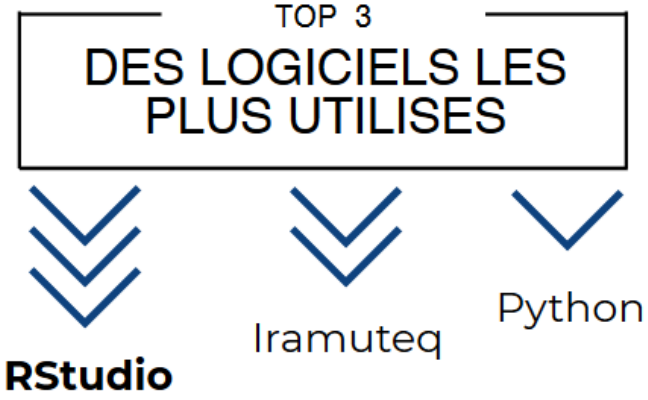
Les utilisateurs ponctuels sont surtout des chercheurs (20%)/enseignants-chercheurs (15%) ainsi que des fonctions d'appui (17%) alors que les experts utilisateurs se recrutent majoritairement parmi les chercheurs (35%)/enseignants-chercheurs (16%) et que les non-utilisateurs sont principalement des fonctions d'appui (28%). Il serait intéressant de creuser si les experts sont essentiellement issus du domaine de la recherche sur la fouille de textes ou s'il y a d'autres catégories de chercheurs concernés.

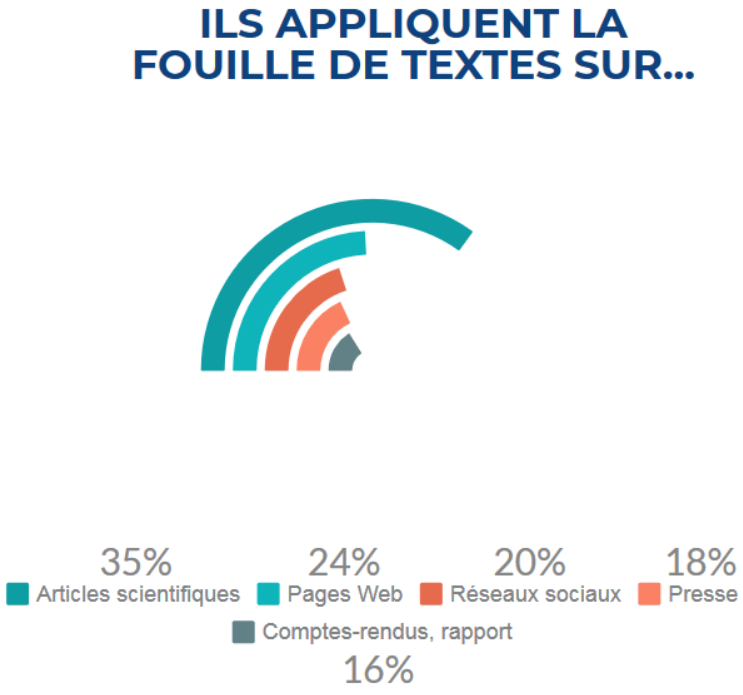


Les trois principales utilisations de la fouille de textes se concentrent autour de l'extraction d'information, la classification/catégorisation de documents et l'aide à la découverte scientifique. Là encore, ces utilisations sont, de façon générale, le fait de chercheurs et enseignants-chercheurs.



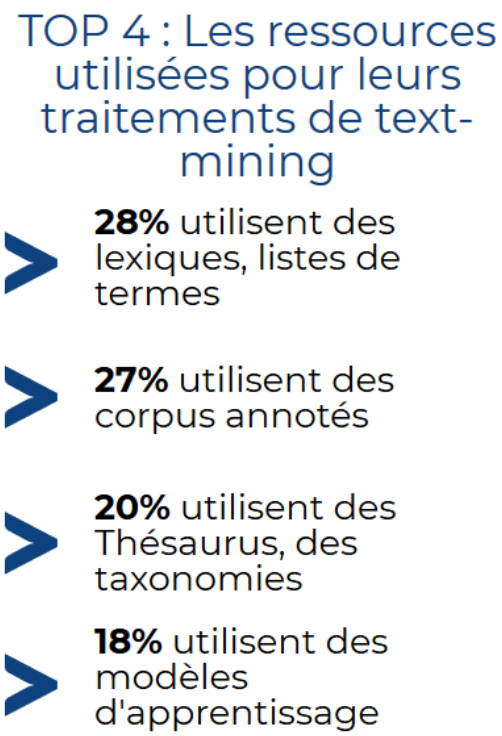
Ils utilisent à cet effet très majoritairement des outils libres et essentiellement les plus connus dans la communauté. A la question de savoir qui développe ses propres outils, peu de répondants se déclarent (19% de réponses « oui ») et surtout 75% des interrogés ne répondent pas du tout à cette question. Par ailleurs ils ne fournissent aucune description ou lien vers une documentation dans les cas où ils développent eux-mêmes. A noter que certains citent tout de même des plateformes de dépôt d'outils (GitHub) ou des langages (Python) plutôt que des outils.





Ces documents sur lesquels la fouille de textes est appliquée sont ceux qui constituent le support de leurs activités. On notera que les articles scientifiques sont loin d'être la seule ressource mise en jeu (les pages web et les réseaux sociaux sont bien mis à contribution) et ceci qu'il s'agisse de chercheurs/enseignants-chercheurs ou de fonctions d'appui.

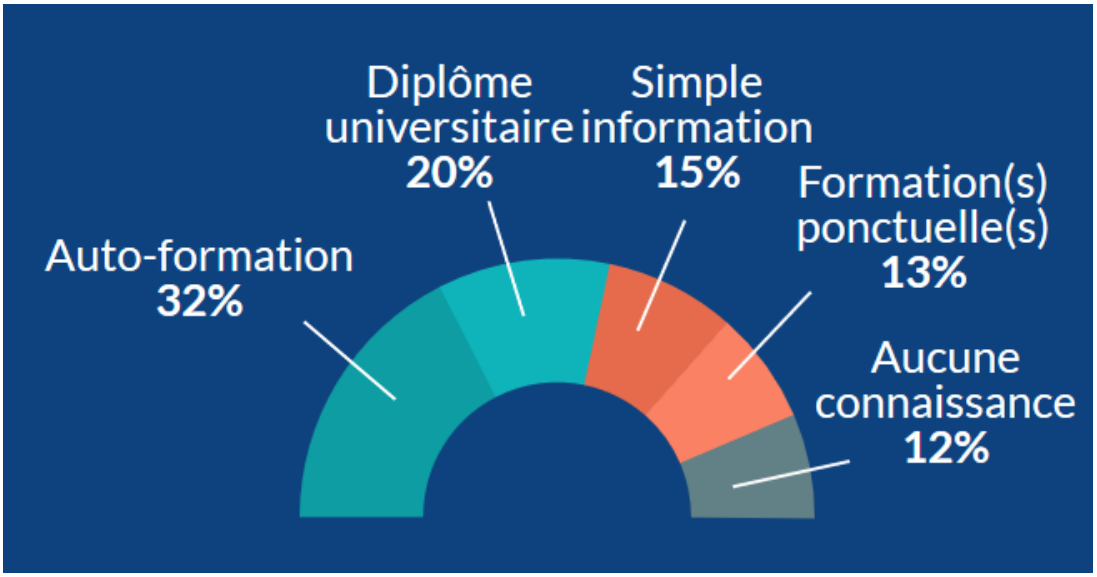
Le taux élevé de non réponse à la question de savoir si des prétraitements sont appliqués nous fait supposer que beaucoup de répondants n'ont pas une idée très précise de ce dont il s'agit. Ces prétraitements sont essentiellement de type prétraitements linguistiques et la réponse « oui » à leur utilisation est à 100% le fait d'utilisateurs experts.



En termes de ressources utilisées pour faire de la fouille de textes, les répondants ont recours à des corpus annotés mais également beaucoup à des ressources sémantiques (si on cumule tous types de ressources de ce type). Cet élément est intéressant car ce n'est pas nécessairement un point auquel on peut penser spontanément dans le cadre des besoins de chercheurs/enseignants-chercheurs. Certains développent eux-mêmes ce type de ressources: corpus annotés, terminologies/ontologies et modèles d'apprentissage en majorité. Ici aussi, la typologie des ressources développées est non documentée et sujette à confusion avec des plateformes de dépôt de code (Github) ou de ressources sémantiques (BioPortal).

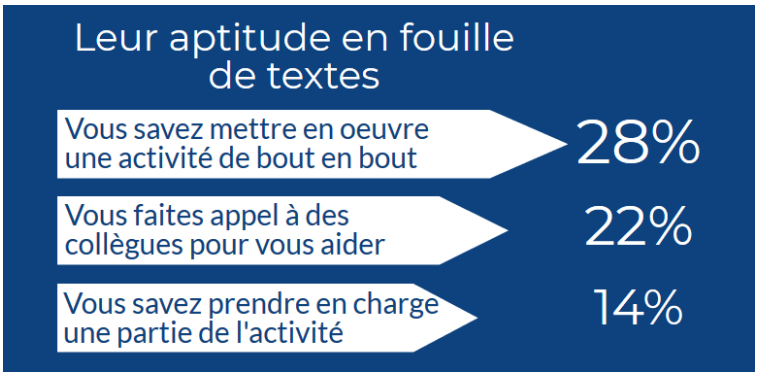
Pour ce qui est du partage des productions, on note un nombre très important de non réponses et un nombre assez faible de partage effectif (18%).

2.4.3 Compétences et formations

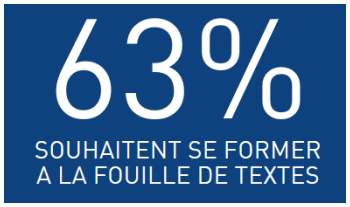


Le niveau de formation en fouille de textes des répondants met en avant essentiellement de l'autoformation pouvant être associée ou non à une formation plus conventionnelle. Le diplôme universitaire est plutôt l'apanage des chercheurs, alors que les formations ponctuelles concernent tout aussi bien les fonctions d'appui.

En matière de droits en fouille de textes, seuls 17% des répondants ne les connaissent pas du tout (64 réponses) ce qui est une information importante à exploiter dans le contexte actuel de valorisation de la Science ouverte. La connaissance de ces droits est sensiblement identique entre les chercheurs/enseignants-chercheurs et fonctions d'appui mais est très faible dans les autres catégories de répondants.



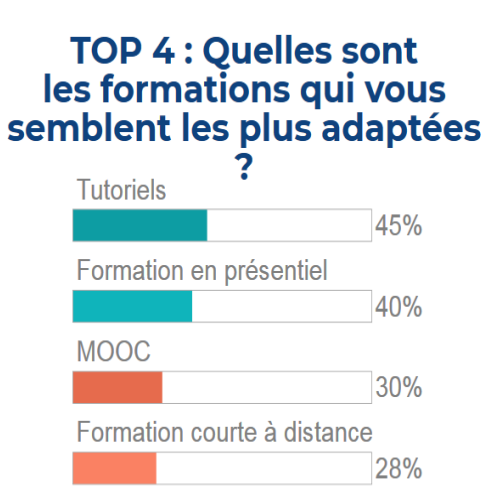
Il est étonnant de constater qu'un nombre non négligeable de répondants se déclare apte à mettre en œuvre une activité de fouille de textes en faisant appel si besoin aux compétences de collègues. Ceci est essentiellement le fait des chercheurs et, dans une moindre mesure, d'enseignants-chercheurs et corrélé à un niveau élevé de formation, de type universitaire.



Une surprise? Une volonté clairement exprimée de se former !

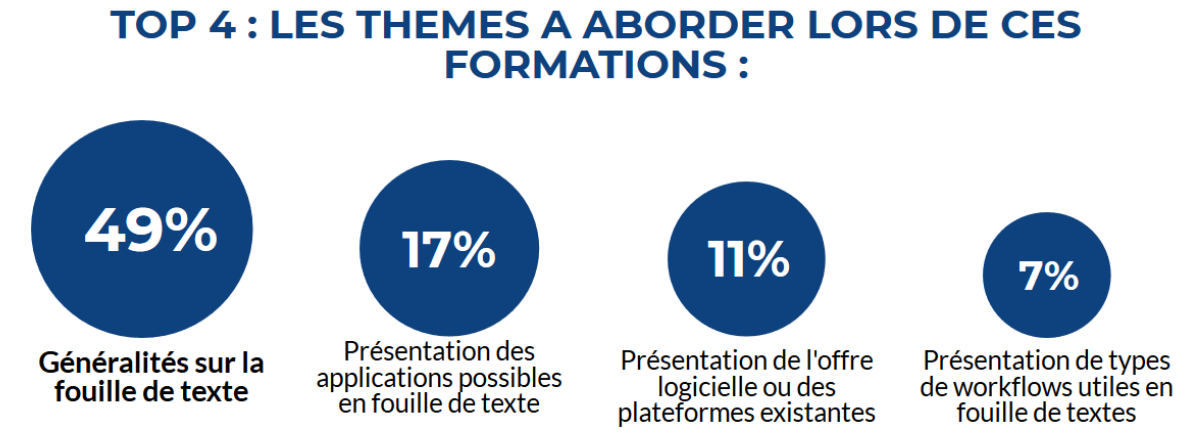
On peut supposer que c'est là le reflet d'un vrai besoin et il est intéressant de cerner mieux les attentes de ces personnes, aussi bien en matière de contenus de formation qu'en termes de typologie de formations attendues.

On notera l'absence de préférence tranchée pour une modalité particulière de formation et on relèvera également le fait que pas mal de répondants plébiscitent la lecture de la littérature sur le sujet.

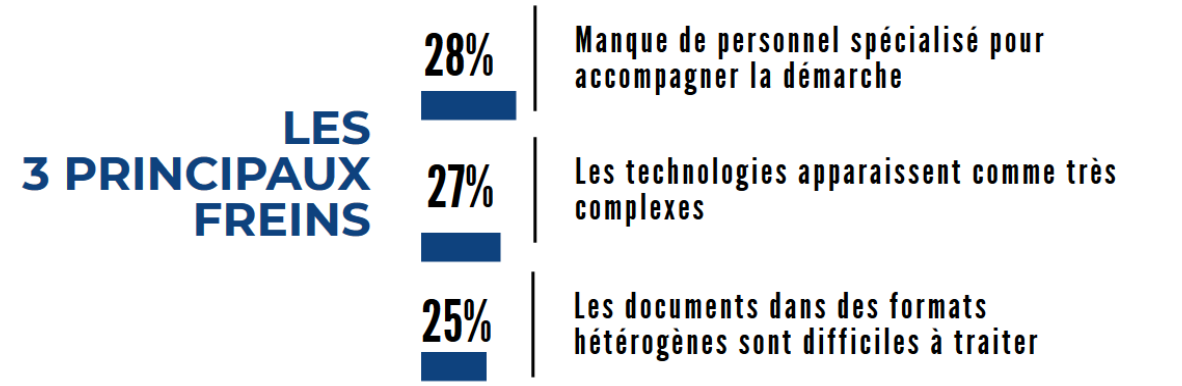


Un autre résultat un peu inattendu : un taux de demande de formations en présentiel encore à 40% alors qu'en particulier les chercheurs ont souvent la réputation d'être peu disponibles. Ce résultat peut être mis en parallèle aussi avec le taux très faible (6%) de personnes déclarant n'avoir pas de temps pour s'investir dans ce domaine ou trouvant cette problématique hors de leur champ de mission (5%) ou de leurs besoins (3%). A noter la possibilité de formation par co-développement professionnel.

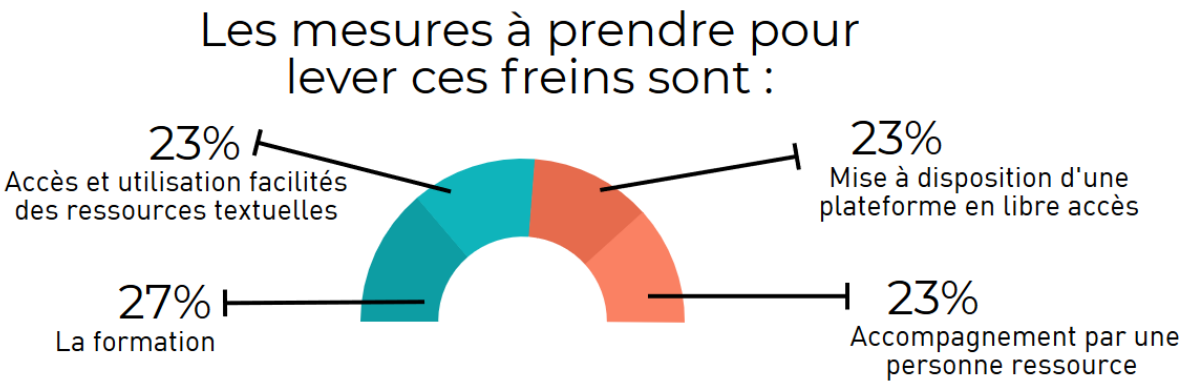
Sur le contenu des formations, les répondants semblent avoir par contre peu d'avis tranchés laissant à penser qu'ils manquent singulièrement de connaissances sur le sujet. Les propositions majoritaires sont généralistes et théoriques (généralités sur la fouille de textes, quelles applications ? quels logiciels ? quelles ressources ?). Seuls les chercheurs formulent plus de demandes pouvant porter sur des formations à des applications spécifiques.



2.4.4 Freins et leviers



Lorsqu'on se penche sur les freins actuels à la fouille de textes, l'ensemble cumulé de tous les aspects techniques/technologiques/informatiques arrive en premier lieu suivi ensuite par le manque de personnel spécialisé.



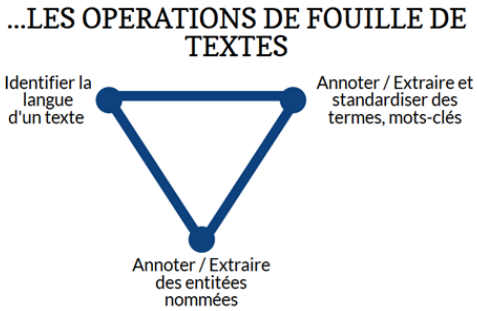
La formation est l'un des moyens d'y remédier tout comme le recours à des personnes qualifiées ressources, à côté du libre accès à des outils/ressources.

2.4.5 Fonctionnalités d'une infrastructure

La partie du questionnaire « Fonctionnalités d'une infrastructure » ne s'adressait qu'à des personnes se reconnaissant comme familières de la fouille de textes.

Les fonctions de base d'une plateforme de fouille de textes qui sont privilégiées sont l'export de résultats et la configuration de processus de traitement suivis de toutes les possibilités de calcul, stockage et traitement de gros volumes de données.

Dans les opérations indispensables, on note essentiellement l'annotation/extraction de termes ou d'entités nommées, suivie par la classification/catégorisation. Presque autant plébiscitées : analyse morphologique et syntaxique, analyse du discours, etc.



… L'INTEGRATION

Disposer de fonctionnalités de base à travers une API documentée

Pour ce qui est des possibilités d'intégration, elles ont suscité un intérêt plus mesuré sans-doute parce que nécessitant un niveau de technicité déjà assez important.

… LE CORPUS

Les répondants privilégient la possibilité de charger leurs propres corpus et jugent utile de pouvoir les traiter, en particulier même si les documents sont dans des formats hétérogènes. Les ressources sémantiques associées sont quant à elles jugées utiles pour traiter ces corpus aussi bien dans le sens d'une adaptation des traitements grâce à ces ressources que dans le sens d'une extraction de ressources utiles grâce aux traitements.

…LE COLLABORATIF

Disposer d'un espace personnel de travail sécurisé

Le collaboratif est vu comme un élément indispensable. Le partage d'outils, de traitements, de ressources est vu comme utile tout comme les éléments d'interaction entre utilisateurs (forum de discussion par exemple)

… LES INTERFACES

Visualiser les résultats sous forme brute

Une demande forte :

… LE MODULE AIDE

Est plébiscité, l'accès à une documentation claire et complète aux outils et ressources disponibles et adaptée suivant les profils utilisateurs.

2.4.6 Quelques verbatim à retenir

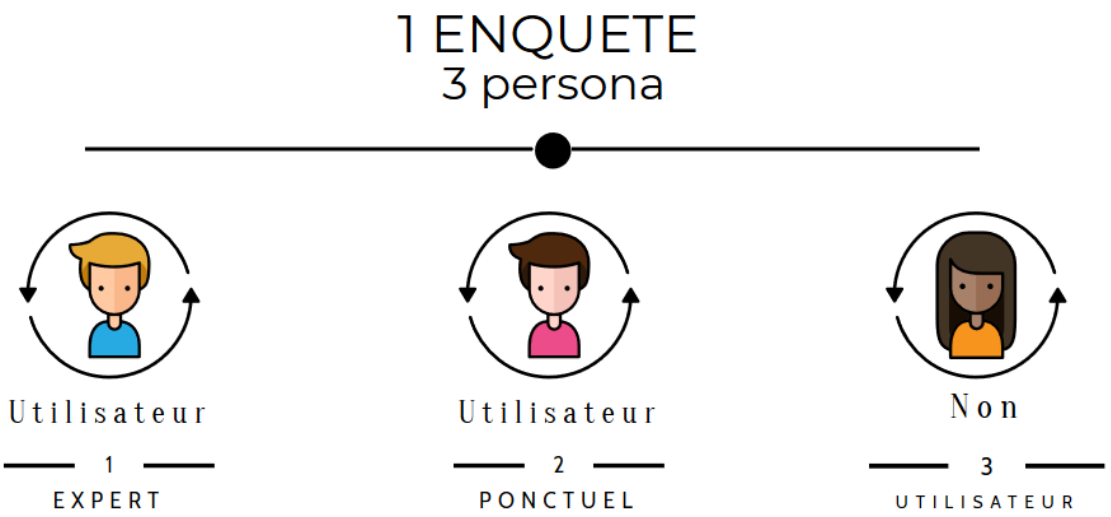
Nous notons que très peu de commentaires ont été formulés. Parmi ceux-ci nous en avons retenus deux qui soulignent le besoin de mettre la fouille de textes au service, non seulement de spécialistes du domaine de la fouille de textes, mais aussi d'utilisateurs plus novices. Ce constat répond parfaitement aux objectifs visés dans le projet Visa TM.

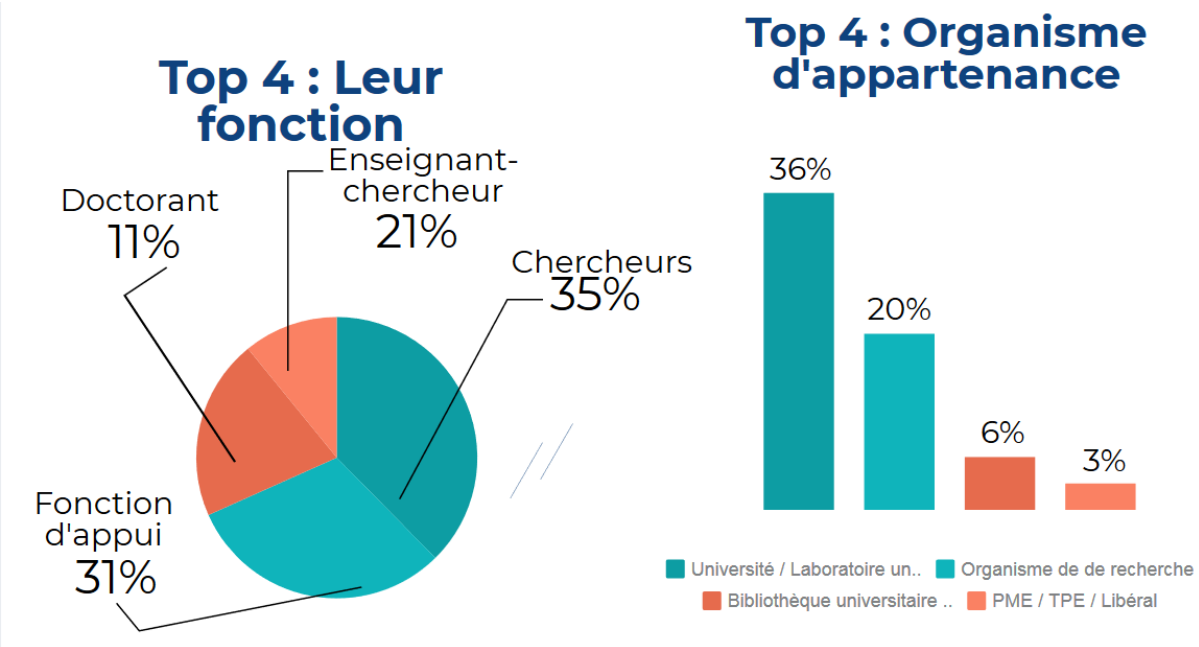
> Les outils doivent **être à la portée d'utilisateurs** comme le sont les nouvelles générations d'applications.

Pas spécialiste en analyse des données textuelles (*text mining*) j'aimerais **avoir à ma disposition un outil aussi facile d'emploi qu'un moteur de recherche**, mais qui gère non pas seulement des mots-clés mais des concepts (ex : concept de résilience) ou des réseaux de concepts qui se forment et se déforment au cours du temps, avec des relations d'étymologie, de proximité, d'inclusion, et peut-être d'autres auxquelles je n'ai pas pensé.

2.4.7 Profil des répondants

Comment les reconnaitre ?





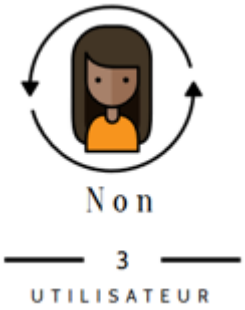
Les répondants de notre questionnaire sont en grande majorité issus des universités et secondairement des organismes de recherche. Ils viennent essentiellement du domaine des SHS et en second lieu des STIC. Cette répartition ne doit pas être lue de façon brute car elle est aussi le reflet des milieux de diffusion du questionnaire. On notera par exemple le peu de représentativité du domaine biomédical qui est pourtant fortement impliqué depuis de nombreuses années dans la recherche et l'utilisation de la fouille de textes. Il est probable aussi qu'un certain nombre de répondants n'ont pas su se placer dans l'une de nos catégories (en particulier des personnes issues du TAL et de l'informatique).

Ces répondants occupent essentiellement des postes de recherche tous profils confondus (chercheur, enseignant-chercheur et doctorant) suivis par les fonctions d'appui.

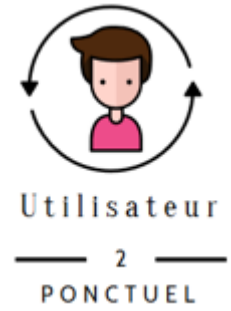
Les 3 persona[11] partagent des points communs:

Ils appliquent ou souhaitent appliquer la fouille de textes sur…	Ils utilisent ou souhaitent utiliser la fouille de textes pour…	Ils souhaitent une plateforme disposant …
• Les articles scientifiques • Les comptes-rendus et les rapports • Les pages webs et les réseaux sociaux	• L'extraction d'informations / annotation • La catégorisation de documents • La constitution de corpus et le nettoyage	• D'un espace personnel de travail sécurisé • D'une aide pour accèder à une documentation claire et complète, adaptée à chaque profil • D'un accès à une documentation claire et complète des outils et ressources disponibles

Mais se différencient également sur certains points :

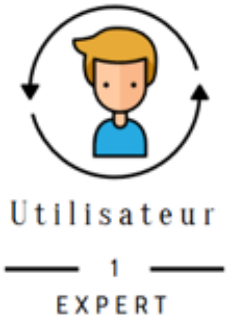


Le (la) non utilisateur (trice) n'a qu'une simple connaissance ou information sur la fouille de textes. Il est prêt à se former, essentiellement à des généralités et des présentations d'outils/plateformes existants et devant la complexité des technologies et le défi informatique, il juge nécessaire de se faire accompagner dans sa démarche.



L'utilisateur (trice) ponctuel (le)s'est auto-formé et/ou fait appel à des collègues, il utilise des logiciels libres qu'il n'a pas développé lui-même. Il est freiné dans son approche de la fouille de textes essentiellement par la complexité des technologies et le manque de personnel d'accompagnement spécialisé mais est disposé à se former plus amplement au moyen de formations en présentiel ou de tutoriels.

[11] https://hal.archives-ouvertes.fr/hal-01376762



L'utilisateur (trice) expert(e) s'est auto-formé ou a une formation universitaire. Il sait mettre en œuvre une procédure de fouille de textes dans sa globalité, et développe ses propres outils/ressources qu'il rend accessibles à d'autres et utilise des logiciels libres. Ses demandes de formation sont plus précises et il est freiné prioritairement par un manque d'accès aux ressources et à leur mauvaise qualité.

2.4.8 Synthèse

Ce questionnaire destiné à un public large et diversifié nous a permis d'appréhender un peu mieux le profil des utilisateurs potentiels de la fouille de textes et leurs attentes dans le milieu de l'enseignement supérieur, de la recherche et de l'innovation. Quelques acteurs du privé n'ont pas hésité à se joindre aux répondants.

Son analyse nous montre que les réponses les plus complètes (184 sur 300) sont essentiellement le fait de chercheurs, enseignants-chercheurs ou de fonctions d'appui, appartenant majoritairement à des universités, des laboratoires universitaires ou des organismes de recherche. Ces acteurs sont souvent impliqués eux-mêmes dans des activités de fouille de textes, qu'il s'agisse de l'utilisation de celle-ci ou de travaux de recherche sur le sujet. Ils sont de fait en capacité de mettre en œuvre cette fouille de textes au profit de l'exploration scientifique et ont une formation adaptée de type universitaire ou se sont auto-formés par ailleurs.

D'autres répondants ont apporté des réponses moins précises et fouillées (116 sur 300). Ce sont plutôt des utilisateurs ponctuels de la fouille de textes ou ne l'utilisant pas du tout. Ils n'ont que peu de formation, se sont quelquefois auto-formés ou font ponctuellement appel à des collègues pour leur venir en aide.

La fouille de textes est utilisée majoritairement pour de l'extraction d'information, de la classification/catégorisation de documents, de l'aide à la découverte scientifique mais aussi pour une part non négligeable d'analyse d'opinions/de sentiments.

Elle repose très largement sur le recours à des outils libres (42%) - dont certains plus ou moins reconnus dans les communautés - qui sont appliqués sur des documents de type « Articles scientifiques » mais également sur des ressources issues du web et des réseaux sociaux. Les prétraitements de ces ressources sont essentiellement linguistiques et donc le fait d'utilisateurs experts. Une aide est apportée par des corpus annotés, mais aussi beaucoup par des ressources de type sémantique (lexiques, thésaurus, taxonomies, etc.) ainsi que par le recours à des modèles d'apprentissage.

Nous notons une volonté clairement affichée de creuser le sujet de la fouille de textes au travers de formations (63% des répondants se déclarent favorables) aussi bien en présentiel qu'à l'aide de tutoriels, MOOC ou formations courtes à distance. Sans surprise, ces formations sont essentiellement demandées sur des généralités sur la fouille de textes et les applications

ou les outils disponibles par des répondants peu familiers du sujet et plutôt sur des applications plus spécifiques par les plus experts.

Les freins technologiques, la complexité d'approche des outils en particulier, ainsi que le manque de personnel spécialisé et l'hétérogénéité des formats de documents à traiter sont les principaux obstacles à une « démocratisation » et une utilisation plus extensive de la fouille de textes aujourd'hui.

La formation est un levier efficace pour faire progresser ce constat tout comme la mise à disposition de personnels dédiés et un accès facilité et documenté aux différents types de ressources utiles comme ISTEX pour les ressources bibliographiques numériques ou Agroportal pour les ressources sémantiques par exemple.

Concernant les fonctionnalités attendues d'une infrastructure de fouille de textes telle qu'envisagée dans le projet Visa TM, les résultats nous ont apporté un éclairage que nous allons détailler. Les fonctions de base d'une telle infrastructure comprennent la configuration et l'adaptation de processus de traitement existants ainsi que l'export des résultats dans des formats standards, sans oublier les possibilités de calcul, stockage et traitement de grands volumes d'information. Les opérations jugées indispensables se concentrent autour de l'annotation/extraction de termes ou d'entités nommées et de la classification/catégorisation. Les besoins d'intégration se focalisent autour d'une API documentée. Une demande forte est la possibilité de charger des corpus y compris personnels et surtout dans des formats hétérogènes qui pourront être traités à l'aide de ressources sémantiques associées. Un espace personnel sécurisé permettra de conserver les traitements effectués et les ressources mais devrait aussi être ouvert à des possibilités de collaboration/partage avec d'autres utilisateurs. Une demande émerge pour une visualisation des résultats sous forme brute. Enfin les répondants attendent une aide documentée, claire et complète sur les ressources et outils disponibles, si possible adaptée au profil de l'utilisateur.

Ce questionnaire a, en tous les cas, suscité un intérêt certain, même si l'appropriation du sujet reste très partielle pour certains répondants. Nous mettrons également les résultats à disposition sur le blog dédié au projet pour répondre aux demandes de retour sur ces résultats exprimées par les répondants.

Par ailleurs, nous sommes conscients que les diverses fonctionnalités de la plateforme n'ont été évoquées que partiellement et que des aspects techniques plus spécifiques pourraient être étudiés (algorithmes, apprentissage automatique,…) tout comme des besoins particuliers inhérents aux contraintes de certaines communautés (anonymisation de données,…).

Pour une analyse plus détaillée il sera utile de se reporter au document en annexe : « Résultats de l'enquête Visa TM ».

Besoins en fonction des types d'utilisateurs

Comme analysé au chapitre précédent, les services envisagés par l'infrastructure visée par le projet Visa TM sont destinés à différentes familles d'acteurs :

- les chercheurs en TAL qui souhaitent exposer et partager leurs travaux et outils
- les chercheurs non TAL, les acteurs de l'appui à la recherche
- les décideurs qui souhaitent s'approprier les technologies du TAL pour répondre à leurs besoins de fouille de l'information

Ce chapitre poursuit l'analyse des parties prenantes en l'étendant au-delà des utilisateurs et de leurs besoins. Les services envisagés impliquent également :

- les fournisseurs de services, d'outils de fouille de textes et de contenus qui souhaitent valoriser leur production et la mettre à la disposition de la communauté de la recherche
- les fournisseurs d'infrastructures de calcul, stockage et transfert, susceptibles d'accueillir l'infrastructure

Ces différents acteurs peuvent avoir des besoins très différents auxquels la fouille de textes devrait répondre. Nous essayerons ici de formaliser ces divers besoins. Pour trois catégories d'entre eux nous avons mené une réflexion sur les besoins avant d'avoir les résultats du questionnaire et nous l'avons formalisée sous forme de cartes heuristiques.

3.1 Chercheurs/experts en TAL/fouille de textes

Ils se répartissent selon différents niveaux de connaissances, du doctorant au chercheur senior.
Ils sont utilisateurs de contenus textuels sous forme de corpus (articles scientifiques essentiellement, mais aussi réseaux sociaux, informations, métadonnées…) ainsi que de ressources terminologiques, linguistiques. Pour eux, l'enjeu est donc la mise à disposition sur une plateforme de ces contenus de façon simultanée et à long terme. Le recours à des corpus annotés comme moyen d'entraînement et d'évaluation des résultats est particulièrement pertinent.
Elle est une opportunité de mettre en ligne les outils issus de leurs recherches et de bénéficier de ceux des autres dans une optique de réutilisation même si de prime abord l'interopérabilité n'est pas une de leurs principales préoccupations.
Elle peut leur apporter une aide à une intégration de ces outils et un accès facilité aux différents workflows qu'il est possible de construire avec eux, tout en améliorant le partage par une documentation systématique et détaillée et en permettant de faire des tests de ces processus.

Enfin ils ont des besoins en stockage et puissance de calcul pour pouvoir faire des traitements de données massives et pouvant être rejoués de façon itérative. Un accès à différents algorithmes de traitement est également essentiel pour en expérimenter les différents résultats.
La moitié d'entre eux est plus ou moins familière avec les problématiques de licence.

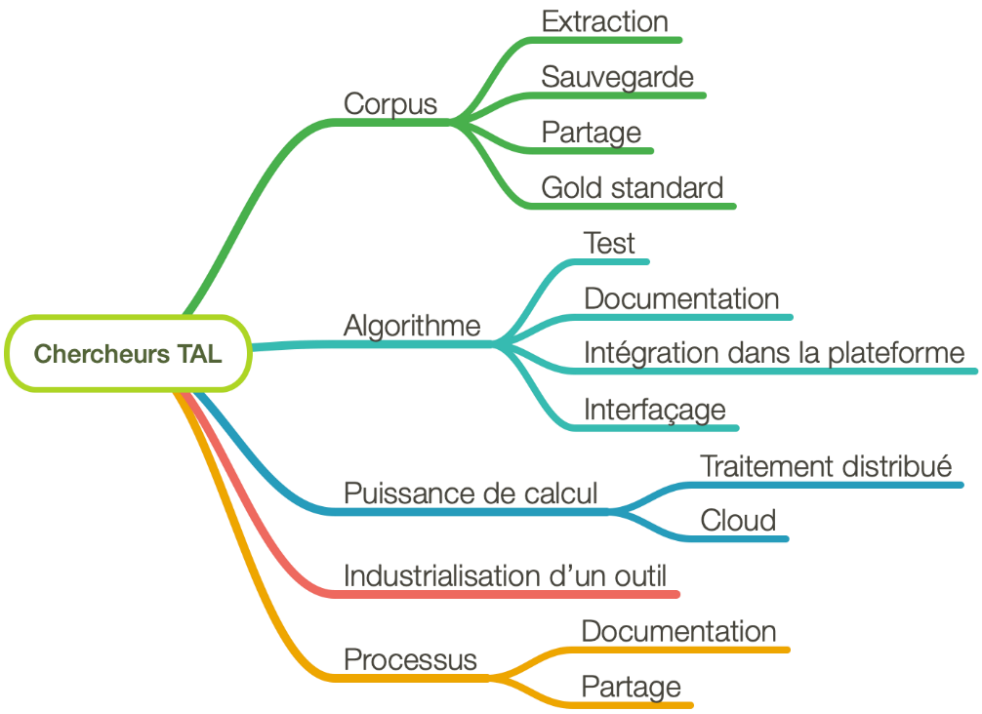


Figure 1. Réflexion sur les besoins possibles des chercheurs TAL

3.2 Chercheurs non TAL/fouille de textes et acteurs de l'appui à la recherche

On peut inclure dans cette catégorie d'autres partie prenantes comme les décideurs, les personnels IST, les organismes de subventionnement tout en gardant à l'esprit que celles-ci peuvent avoir des besoins spécifiques tout comme il y a des spécificités de domaines également. Les contenus utilisés sont ici les mêmes que ceux précédemment décrits pour les chercheurs TAL. Les ressources terminologiques ont une importance particulière du fait de leur spécificité de domaine. L'objectif principal de la fouille de textes est ici de repérer rapidement le/les sujets d'un document/corpus de documents et d'exclure ceux qui ne présentent pas d'intérêt.
Ici un besoin prépondérant est celui d'une assistance à la mise en œuvre d'une tâche de fouille de textes : choix d'un outil adéquat et aide à son utilisation, choix d'un workflow, aide à l'interprétation d'un résultat. L'infrastructure devra donc être en mesure d'apporter cette

aide, soit sous forme d'une documentation adaptée, soit par un accompagnement par des personnels d'appui compétents dans le domaine.

Figure 2. Réflexion sur les besoins possibles des chercheurs non TAL

3.3 Décideurs

Pour les décideurs, la fouille de textes est avant tout un moyen de pilotage et d'évaluation. L'objectif ici est de puiser dans les données textuelles des informations permettant de faire de la prospective, d'envisager de nouveaux partenariats sur des sujets scientifiques, d'envisager les innovations de demain. Dans cette optique, leur besoin est surtout concentré autour d'une aide à la fois au lancement de traitements adéquats sur des corpus choisis mais aussi sur l'interprétation des résultats fournis par ces traitements. Cela peut, tout comme pour les chercheurs non TAL passer par la nécessité d'un accompagnement direct par des personnels d'appui capables de mettre en valeur les résultats prégnants.

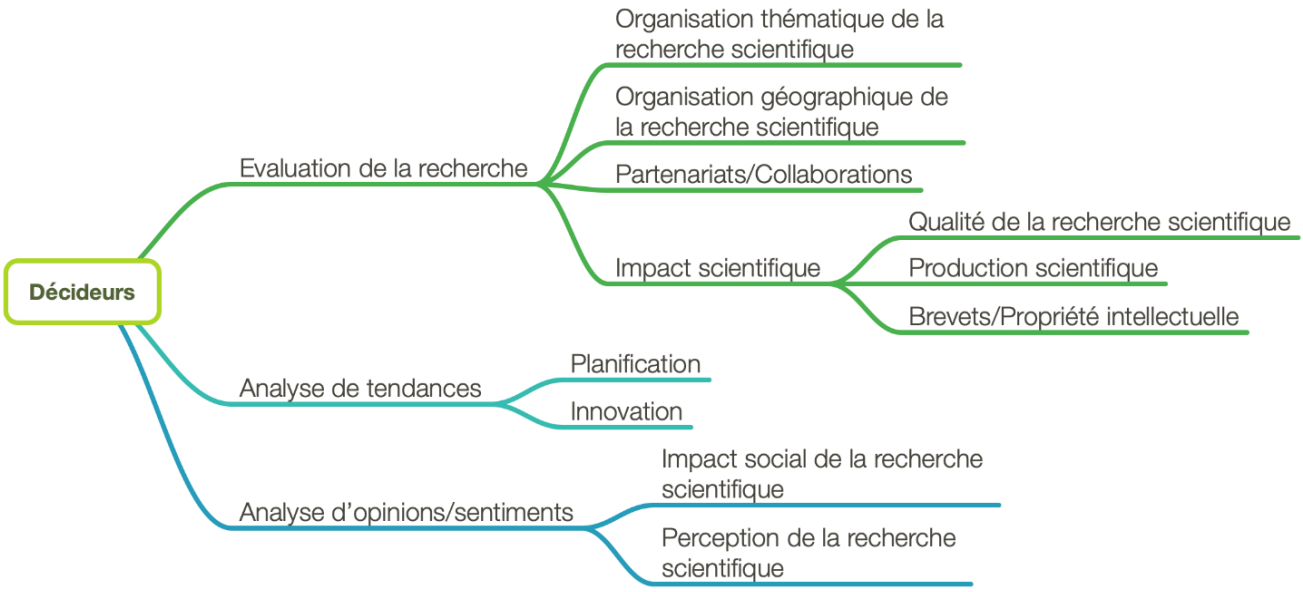


Figure 3. Réflexion sur les besoins possibles des décideurs

3.4 Fournisseurs de contenus

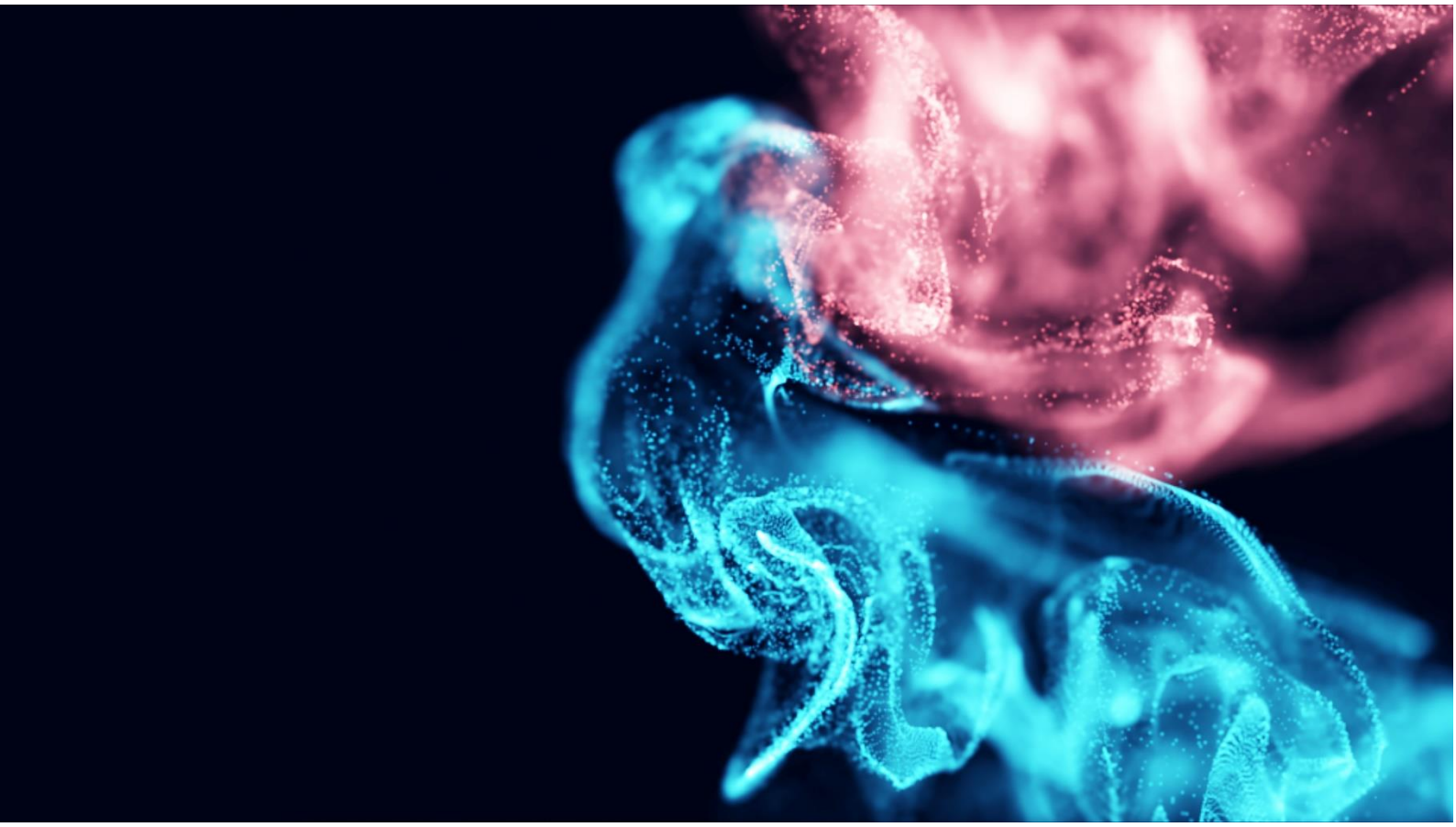
Ces contenus peuvent être de type bibliothèques numériques ou ressources sémantiques. On trouvera donc dans cette catégorie des éditeurs, des bibliothécaires, des ingénieurs de l'information, des gestionnaires de l'information, des curateurs etc.
Il s'agit de personnes habituellement plutôt familières avec les aspects juridiques (licences par exemple). Les besoins sont ici la mise en commun d'outils, la possibilité d'utiliser la fouille de textes à visée d'enrichissement automatique de métadonnées ou de terminologies. Elles ont donc un besoin d'accès le plus large possible à différentes ressources ainsi qu'à la possibilité de connecter ces ressources à la plateforme pour y permettre un accès facilité et partagé. Une fonctionnalité de cette plateforme devrait donc être un accompagnement sur l'intégration des ressources à la fois sous forme documentée mais aussi sous forme d'appui technologique, en rendant possible une interopérabilité des ressources entre autres.

3.5 Agrégateurs de contenus (ex : OpenAire[12], CORE[13], AGRIS[14])

Il s'agit de l'ensemble des structures, organismes contribuant à l'ouverture des données et dont les demandes peuvent se rapprocher de celles des fournisseurs de contenus mais se

[12] https://www.openaire.eu/
[13] https://core.ac.uk/
[14] http://agris.fao.org/agris-search/index.do

concentrent aussi sur l'homogénéité des métadonnées et un accès possible au texte plein auquel ces infrastructures s'ouvrent au travers d'API par exemple.



Besoins spécifiques en fonction des communautés

Le choix des communautés présentées ici à titre d'exemple repose à la fois sur les contributions et études déjà produites dans le projet OpenMinTeD mais aussi sur les connaissances dans certains champs disciplinaires des partenaires du projet Visa TM. Il n'est évidemment pas exhaustif et n'a de but que de mettre en lumière une diversité de besoins selon les acteurs considérés dont il faudra tenir compte dans les préconisations issue de ce projet.

4.1 Agriculture/biodiversité

Cette communauté et ses besoins a été décrite dans le livrable OpenMinTeD D4.2 "White Paper on OpenMinTeD Community Requirements"[15] suite à une enquête en ligne, des entretiens, des groupes de discussion et des ateliers de travail avec des représentants des différents interlocuteurs concernés, en particulier des chercheurs en fouille de textes et des utilisateurs finaux.
Elle se subdivise en un certain nombre de sous-communautés (AGRIS, sécurité alimentaire, biodiversité en microbiologie, cultures agricoles) dont les besoins peuvent différer. Ainsi pour la sécurité alimentaire, dont les objectifs principaux répondent à des impératifs de santé publique, un des besoins majeur est de collecter un maximum de données à partir de publications scientifiques ou d'alertes institutionnelles sur des pathogènes afin d'établir des modèles prévisionnels de risques.
Pour les chercheurs en biodiversité en microbiologie le challenge est plutôt de compiler les connaissances issues de diverses bases de données avec celles de la littérature scientifique ou d'écrits non structurés. Ils ont donc un grand besoin de normalisation et de catégorisation des ressources (taxonomies d'organismes, ontologies d'habitats etc.)
Une plateforme de fouille de textes devra donc être en mesure de répondre à ces différents besoins en mettant à disposition les outils de constitution de corpus de données textuelles et de traitement de ces données tout comme les ressources sémantiques utiles à des catégorisations d'entités nommées par exemple.

4.2 Médecine

La fouille de textes dans le domaine biomédical est indispensable pour gérer les gros volumes de données. Elle s'applique aux publications scientifiques mais également à d'autres données

[15] http://openminted.eu/wp-content/uploads/2016/12/OpenMinTeD_D4.2_CommunityRequirementsAnalysisReport-whitepaper_v2.pdf

textuelles non structurées, telles que les données cliniques renseignées par les médecins et autres soignants dans le cadre du suivi du malade (observations médicales et paramédicales, prescriptions médicales, comptes rendus hospitaliers, …) et figurant dans les dossiers médicaux informatisés, mais aussi des entretiens retranscrits de patients parlant de leur maladie, des réponses à des questionnaires de santé, des textes libres issus des messages postés sur internet dans les forums de santé et sur les réseaux sociaux et bien sûr les ressources du Système national des données de santé (SNDS). Tout ce qui est de l'ordre de données cliniques doit être anonymisé, dans le respect des lois (RGPD de 2018 et Loi pour une république numérique de 2016), ce qui nécessite la mise en place de traitements particuliers sur les textes et un niveau de sécurité renforcé au niveau de la plateforme de stockage et d'analyse.

Leurs besoins relèvent de différentes techniques de fouille de textes qui permettraient l'annotation, la construction d'ontologies en vue de leur exploitation par un moteur de raisonnement ou encore l'extraction d'information.

Tous les acteurs impliqués dans le système de soins ont des besoins auxquels la fouille de textes peut répondre :

- mettre en évidence les facteurs de risque de maladies ou définir des groupes de malades
- aider à la prescription ;
- mettre en évidence des signaux d'émergence de nouvelles maladies ;
- faciliter le recueil des données pour la qualité et la continuité des soins ;
- transformer les données textuelles en connaissances utiles pour les chercheurs ;
- analyser les productions des patients dans les médias sociaux pour, par exemple, établir des recommandations d'experts, analyser des pratiques sexuelles à risque, suivre des personnes à risque suicidaires, …
- aider à la prise de décision concernant la gestion des hôpitaux et la maîtrise du coût des soins ;
- construire des ressources termino-ontologiques pour servir de support au codage médico-économique hospitalier ;
- détecter des abus ou des fraudes pour les assureurs.

En démultipliant le pouvoir des données médicales et en abolissant les frontières entre disciplines, métiers, recherches et pratiques, ces techniques permettent d'établir des corrélations et favoriser l'avènement d'une nouvelle médecine, à la fois prédictive, préventive, personnalisée et participative.

4.3 Chimie

L'utilisation de la fouille de textes par les chercheurs travaillant dans le domaine de la chimie est principalement tournée vers la recherche et l'annotation des composés chimiques, qui peuvent être présents dans les textes sous différentes formes, y compris graphique.

Le premier objectif des chercheurs consiste donc à repérer et indexer les composés cités dans des documents non structurés, pour obtenir des informations structurées et interrogeables, permettant en outre d'associer un composé à ses méthodes de synthèse ou à ses propriétés (physicochimiques, cristallographiques, toxicité, activité biologique, etc.).

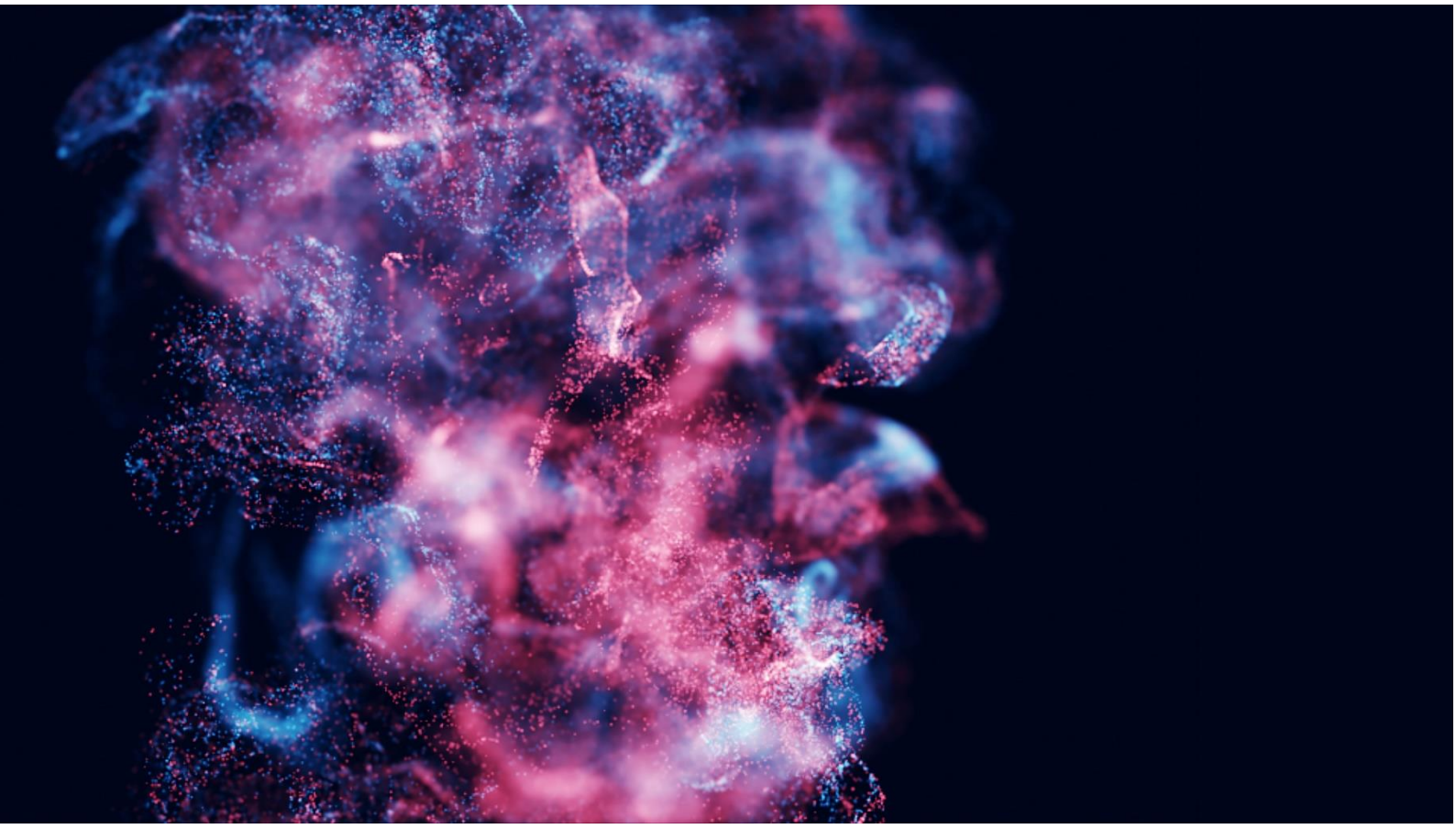
Les outils sont le plus souvent destinés à la reconnaissance d'entités nommées, parfois couplés avec des programmes de conversion nom-structure ou image-structure. Ils utilisent des lexiques et des patrons spécifiques au domaine. La nomenclature des composés chimiques faisant un usage très spécifique de caractères comme les espaces, les virgules, les tirets, les guillemets, les points, les parenthèses..., l'utilisation de tokenizers spécifiques est indispensable. L'application d'ontologies ou de nomenclatures spécifiques permet de gérer les liens structure-propriété.

Nous avons là encore l'illustration de besoins très spécifiques auxquels doit répondre une plateforme de fouille de textes si elle doit s'adresser à des communautés de chercheurs très diverses.

4.4 Psychologie de la mémoire

Nous avons choisi le domaine de la « Psychologie de la mémoire » qui est un sous-domaine de la psychologie, car il est l'illustration des besoins d'une communauté de chercheurs spécifique regroupée aujourd'hui dans un Groupement De Recherche (GDR) auquel participe l'INIST-CNRS en tant que représentant sur les aspects terminologiques. Or, comme cela a été souligné dans les réponses au questionnaire, les ressources sémantiques sont importantes pour la fouille de textes. Au sein de la communauté des chercheurs travaillant sur la mémoire : la fouille de textes est encore balbutiante et n'est pas aujourd'hui structurée de manière suffisamment explicite. Néanmoins, dans ce domaine, certaines techniques de fouille de textes sont connues, ont déjà été utilisées dans le passé, et le sont toujours, mais surtout comme moyen de modélisation de processus ou systèmes mnésiques. C'est le cas, par exemple, de l'analyse sémantique latente comme modèle de la mémoire sémantique (mémoire des connaissances décontextualisées sur le monde) ou encore des réseaux de neurones artificiels (apprentissage, reconnaissance des visages et d'objets, mémoire de travail, etc.)

Le potentiel pour la fouille de textes dans cette communauté est important : analyse de retranscriptions d'entretiens (voir, par exemple le programme 13 novembre pour l'analyse des souvenirs des attentats du 13 novembre 2015 à Paris), la détection dans les articles scientifiques des coordonnées d'activations cérébrales au cours de tâches de mémoire, l'annotation de bases de données bibliographiques à l'aide d'ontologies, l'utilisation du topic modeling pour dégager des thèmes dans les articles et les aligner ensuite avec les activations cérébrales correspondantes, l'analyse de sentiments et d'opinions dans des retranscriptions d'entretiens, la détection des tendances dans la recherche sur la mémoire (« hot topics ») etc. Les outils d'apprentissage automatique ou semi-automatique d'ontologies à partir de textes sont au service de ces différentes utilisations.



Focus sur un besoin particulier : la constitution de corpus

Dans l'ensemble du panorama des besoins, déclinés par acteurs et par domaines spécialisés dans les chapitres précédents, le besoin de corpus structurés sur lesquels appliquer la fouille de textes revient constamment, pourtant le nombre de documents en ligne n'a jamais été aussi grand. Ce chapitre a pour objectif de détailler plus finement ce que sont les besoins de conception de corpus des utilisateurs de la fouille de textes pour identifier précisément les verrous et conduire à des recommandations à prendre en compte dans la future plateforme.

5.1 Introduction

La conception d'un corpus d'articles scientifiques pour un besoin d'extraction d'information est une tâche complexe et peu automatisée. La diffusion d'applications de fouille de textes auprès des scientifiques suppose que cette étape soit aussi simple que possible.

Le besoin d'automatisation de l'extraction d'information trouve sa source dans la recherche d'informations spécifiques que le chercheur est capable de spécifier pour son domaine mais auquel les outils classiques ne répondent pas parce que les sources d'information sont trop nombreuses, ou dispersées, ou les informations elles-mêmes ne sont pas identifiables par une recherche bibliographique simple à base de mots-clefs.

La traduction de ce besoin en un corpus de documents porteurs de ces informations se heurte à de très nombreuses difficultés méconnues parce qu'apparaissant comme techniques, insolubles ou secondaires. La conséquence en est que la majorité des corpus d'extraction d'information n'a pas été constituée par rapport au besoin malgré l'intention initiale, mais en fonction de la disponibilité des sources et de leur facilité d'accès. Par exemple, la grande majorité des applications en sciences de la vie utilise la base de références bibliographique PubMed (titres et résumés) et non les textes complets des articles. Les applications d'extraction d'information qui utilisent des textes complets exploitent les sources les plus accessibles, souvent une seule (ex. PMC, Istex) ou si elles sont plusieurs ce sont en général, les six principales plateformes interrogeables par API. La réalité est que les informations sont disséminées dans des documents d'une bien plus grande diversité. La section "6.3 Bilan sur les données ouvertes et la réutilisation " du document "Application pilote pour la recherche : exemple de l'écologie microbienne" en donne un exemple détaillé.

Cette section détaille les nombreuses étapes de la conception de corpus d'articles scientifiques pour en dégager les principes. Nous identifions un certain nombre de verrous d'ordre technique, juridique et liés aux pratiques pour lesquels des pistes d'automatisation, d'évolution des pratiques et des standards et des évolutions juridiques sont proposées.

5.2 Motivation

La conception d'un corpus documentaire pour un projet scientifique consiste à rassembler physiquement une copie de l'ensemble des documents nécessaires à la tâche et à les caractériser en extension et en intention de telle sorte que le rôle de chaque document dans l'ensemble documentaire soit défini.

Une fois le périmètre caractérisé en fonction du besoin, la conception du corpus répond à de nombreuses contraintes sur l'identification des documents pertinents et leur disponibilité dont le documentaliste et le spécialiste du domaine sont traditionnellement les acteurs. De nombreuses évolutions techniques ont profondément modifié le processus et permis l'automatisation partielle du processus de conception de corpus, principalement, les moteurs de recherche bibliographique et la disponibilité des documents en ligne.

La fouille de textes apporte une nouvelle dimension : l'analyse du corpus elle-même est automatisée. L'ensemble du processus, conception du corpus et analyse doit pouvoir être formalisé premièrement, pour permettre l'évaluation de la pertinence de la démarche, assurer la reproductibilité et éventuellement permettre la réutilisation de certaines données pour d'autres tâches. La formalisation du processus est également un enjeu plus général pour identifier les verrous, rationaliser, automatiser et plus généralement pour l'évolution des applications de fouille de textes.

Dans les faits, les travaux en application de fouille de textes décrivent généralement de façon brève la démarche de conception du corpus, à quelques exceptions près qui mettent en évidence certaines difficultés comme celle de l'accès. Les travaux sur les plateformes de fouille de textes décrivent les possibilités de conception de corpus sous l'angle technique et quantitatif : sont concevables par les plateformes les corpus dont les documents sont dans les archives ou bibliothèques numériques accessibles par la plateforme. Partant du besoin du chercheur et de la tâche, nous voulons montrer ici que le processus de conception de corpus comporte un certain nombre d'étapes générales et obligées sur lesquelles les deux perspectives, applicatives et plateforme font l'impasse, l'une en raison de sa dimension illustrative limitée, l'autre par son absence de perspective utilisateur.

5.3 Étapes de la conception de corpus

5.3.1 Données du problème

Les projets de fouille de textes, recherche bibliographique à des fins de veille ou d'évaluation, projets d'extraction d'information constituent des bases bibliographiques ou des ensembles d'information structurée aussi représentatives que possible du besoin, y compris et surtout les signaux dits "faibles", c'est-à-dire les informations peu répétées. Dans cette section nous nous concentrons sur les articles scientifiques dans des actes de congrès et revues majoritairement internationales comme sources principales. Les sources varient en fonction de leur spécificité, dans le cas général les sources possèdent une couverture plus large ou différente du besoin, et les articles doivent être sélectionnés, plus rarement les sources sont

dédiées au domaine précis correspondant au besoin et dans ce cas tous les articles sont pertinents.

5.3.2 Moyens humains et matériels

La conception du corpus regroupe dans le meilleur des cas une équipe projet dont (1) les experts du domaine qui sont aussi souvent représentants des utilisateurs de l'application de fouille de textes, et (2) pour l'analyse du besoin documentaire et la définition du cahier des charges : les documentalistes et les informaticiens spécialistes de la fouille de textes. Ces derniers apportent la méthode et l'outillage nécessaire. Nous nous plaçons dans le cadre où le projet scientifique est développé par un organisme de recherche public bénéficiant de ce fait des accès aux bases bibliographiques en accès libre ou sur abonnement selon les licences.

5.3.3 Méthode

Les différentes étapes de la conception du corpus sont itératives, c'est-à-dire que l'examen du résultat conduit à modifier la tâche et à la répéter jusqu'à convergence vers un résultat satisfaisant. Nous décrivons ici le processus de façon linéaire par souci de simplification, puis nous préciserons les points incrémentaux.

Le processus est divisé en cinq étapes

1. Définition de requêtes bibliographiques pour produire une liste de références
2. Identification des points d'accès aux documents
3. Formalisation des contraintes au niveau des plateformes (ordre de préférence), des journaux (accès technique, accès légal, pertinence du journal, anticipation du format),
4. Téléchargement des documents
5. Conversion des formats des articles

Les étapes deux à quatre ne sont pas nécessaires si le projet bénéficie d'une base bibliographique, éventuellement plus large que le besoin.

Requête bibliographique

La requête bibliographique exécutée sur une base de références bibliographiques large comme Web of Science ou PubMed en Science de la Vie permet d'identifier les sources pertinentes (actes de conférence, journaux) et le nombre d'articles pour chacun d'entre eux. Elle s'exprime sous la forme plus ou moins riche de critères : combinaison de mots clefs de la base ou des auteurs, domaines, date de publication, langue, type de support.
Dans certains cas, il n'est pas possible d'interroger la base de référence par programme (API, interface de programmation d'application) et il faut plus ou moins manuellement cliquer sur les pages de listes de références pour les télécharger et les traiter.

Classiquement, l'examen des références correspondant aux requêtes permet d'affiner les requêtes sur la base de références bibliographiques pour mieux les faire correspondre aux besoins. La période de publication en est un bon exemple en Science de la Vie. Il est fréquent que les articles anciens soient sous forme de photo (scan) complexes à traiter et que les informations qu'ils contiennent soient répétées dans des sources plus récentes.

Plus le besoin est transversal plus cette étape est répétée et complexe. L'enjeu est ici de ne retenir sans en éliminer que les articles pertinents car le coût de leur traitement est ensuite très élevé. Sous-estimer cette étape conduit à une liste de référence inutilement large - et le coût de traitement en est augmenté d'autant- ou restreinte - et des informations seront manquantes à la fin du processus. Les modes d'interrogation des bases de référence et leur qualité varient très vite.

Lister les sources

De la liste des références pertinentes est extraite la liste des sources (actes, journaux) et le nombre d'articles par source. Cette opération en apparence simple suppose d'être capable d'uniformiser les titres des sources : les titres changent au cours du temps ou sont mal orthographiés, le numéro ISSN n'est pas toujours fourni. Le numéro ISSN est un code de 8 chiffres servant à identifier les journaux, revues, magazines, périodiques de toute nature et sur tous supports, papier comme électronique[16]. Les journaux papier et électronique ont deux ISSN distincts pas toujours faciles à relier. Ce traitement ne peut pas être purement automatique. Lister les sources pertinentes et accessibles est le premier verrou de l'automatisation de la conception de corpus. A la fin de de cette étape, dans le meilleur des cas, un numéro ISSN a pu être attribué à chaque source. Les sources sont disponibles et regroupées sur des sites ou plateformes, soient ceux des éditeurs (publishers) ou des agrégateurs (ex. OpenAire, PMC). L'interrogation d'une base bibliographique ne permet pas toujours de savoir quel est le site ou la plateforme.

Sélectionner les sources et les normaliser

La sélection des sources répond à plusieurs critères et résulte d'un compromis coût / besoin.

L'importance de la source pour le projet de recherche : en général le coût de traitement est trop important au regard du nombre de sources. Les experts chercheurs doivent les ordonner en fonction de leur importance : indispensables, utiles, secondaires, inutiles. Il est prudent de comparer les avis de plusieurs spécialistes.

Le coût d'accès : l'accès automatique à une source impose l'utilisation d'un logiciel connecteur (décrit plus bas). Si le connecteur est déjà disponible, le coût est nul mais s'il faut développer le connecteur d'accès à la source, ce coût doit être pris en compte dans la sélection. Selon les cas, ce développement peut avoir un coût très élevé, à anticiper autant que possible à cette étape. Il recouvre le temps de documentation : comprendre le mode d'accès pour l'utiliser et le temps de développement et de test du connecteur. Si le corpus doit être maintenu dans le temps, il faut inclure le temps de mise à jour du connecteur en cas d'évolution du mode d'accès. Dans le cas le pire où tous les connecteurs sont coûteux et le corpus pas essentiel, la conception du corpus s'arrête à cette étape. Dans la réalité ce coût est très difficile à anticiper et les connecteurs malheureusement très peu mutualisés entre développeurs d'applications de text mining.

16 https://www.issn.org/fr/comprendre-lissn/quest-ce-que-lissn/

La connaissance anticipée des connecteurs aux sources existants par site ou plateforme est le deuxième verrou de l'automatisation de la conception de corpus.

L'accès licite : certaines sources ne sont pas licitement accessibles, ou seulement partiellement :

- L'organisme de recherche n'a pas souscrit les abonnements,
- La source retient les droits d'accès durant une période d'embargo.
- L'usage peut être restreint à un usage non commercial, ou non

Pour être réalisable, cette étape devrait être automatisée, les sources envisagées sont fréquemment de plusieurs centaines, voire milliers de sources, même pour un corpus modeste. Certaines sources peuvent apporter peu d'articles pertinents mais ils sont d'un intérêt considérable. En fait, l'information sur l'accès licite est très difficile à avoir a priori en raison des conditions de variation de date d'embargo, parce que les licences varient pour une même source en fonction de la rubrique dans laquelle l'article est publié (note, préface, etc.) et parce que les licences ne sont pas accessibles et ne sont pas lisibles automatiquement. La connaissance des articles licitement accessibles par source est le troisième verrou de l'automatisation de la conception de corpus.

Les organismes de recherche proposent souvent une interface qui permet à un chercheur de vérifier manuellement qu'un titre de revue est licitement accessible ou non, mais rares sont les interfaces utilisables par des machines et l'information est donnée au niveau du journal, pas au niveau des articles individuellement. Le service SFX payant (ex Inra)[17] propose un accès par programme ou manuel par la lettre du début du titre du journal. Un exemple éclairant sur la variabilité des licences par journal est donné par la table publiée par PubMed sur la disponibilité des articles[18] : sur 3371 journaux référencés, un tiers (1080) propose certains articles en OA, mais pas tous (avec la mention "some"). Pour beaucoup de sources, la seule façon de savoir si on peut accéder à un article est de tenter de le télécharger en faisant l'hypothèse que s'il est accessible techniquement, il est aussi licite d'accéder. C'est la ligne proposée par le document de recommandation du Groupe du Comité pour la Science Ouverte sur la fouille de textes.

Dans certains cas très exceptionnels, des plateformes comme PMC[19] par exemple fournissent des listes des articles avec leur licence. Dans le cas de PMC, c'est une liste des articles identifiés par leur PMID relevant des trois types de licence, CC-BY, utilisation commerciale, utilisation non commerciale. Cette liste est mise à jour et dans un format lisible par une machine. Cette information est très utile pour sélectionner ou non un journal comme source pour la conception d'un corpus, en fonction du nombre réel d'articles accessibles. Choisir un journal comme source impose ensuite de développer le connecteur d'accès à ce journal. Ce développement a un coût élevé. Plus on sait à l'avance quelle est la pertinence d'une source

17 https://sfx-33inra.hosted.exlibrisgroup.com/33inra/az
18 https://www.ncbi.nlm.nih.gov/pmc/journals/#csvfile
19 https://www.ncbi.nlm.nih.gov/pmc/tools/ftp/

en fonction du nombre exact d'articles réellement accessibles, plus on évite des développements coûteux et inutiles.

Un autre critère important du choix des sources est le format des articles. Selon l'application de fouille de textes prévue, le format va avoir un impact très important sur la mise en œuvre, le coût et la qualité du résultat. Dans l'idéal, les formats des articles à sélectionner sont connus et cette information est un élément important de la décision. Dans la réalité, le format relève en général de l'expertise du concepteur de corpus. La connaissance des formats des articles par source est le quatrième verrou de l'automatisation de la conception de corpus.

Sélectionner les articles

A cette étape, nous voulons sélectionner précisément les articles à télécharger. Les agrégateurs proposent des licences différentes pour une même source (ex. arXiv.org/Scopus). Un même article peut être accessible par différents moyens qui ne sont pas équivalents. Nous avons expliqué à l'étape précédente le coût variable de développement des connecteurs aux sources. Il est donc essentiel pour un article donné de décider à l'avance de la source choisie. Plus encore, il faudra éviter de tenter de télécharger des articles non accessibles. Dans l'énorme majorité des cas, la seule façon automatique (ou manuelle !) d'avoir l'information si un article est licitement accessible est de tenter de le télécharger. Ces tentatives inutiles ont également un coût : comme on va voir à l'étape suivante, les tentatives de téléchargement doivent être espacées. Plus on inclut des sources inutiles dans le plan de conception de corpus, plus la conception est longue.
Le choix de la source pour un article donné suppose d'être capable de normaliser les journaux et les identifiants des articles. Les articles publiés sur papier et en ligne ont 2 DOI distincts alors que le contenu est identique. L'absence de numéro ISSN et de DOI rend aussi la tâche compliquée. Il est parfois nécessaire de dédoublonner les articles en comparant leurs métadonnées bibliographiques. Celles-ci ne sont pas toujours normalisées, les titres et les auteurs sont variables, en particulier s'ils contiennent des diacritiques et des caractères non alphabétiques (ex. tiret).

Le dédoublonnage des articles est le cinquième verrou de l'automatisation de la conception de corpus.

La localisation et l'accès

Une fois la liste des articles établie, il faut localiser chaque article. Dans le meilleur des cas, la plateforme bibliographique qui a permis d'obtenir les références, a également fourni soit une URL de téléchargement ou la mention d'une API, ou un DOI. Selon les bases bibliographiques interrogées, les références contiendront ou non un DOI. Celles qui ne contiennent pas de DOI peuvent contenir un identifiant (ex PMID) qui peut parfois être converti en DOI mais pas toujours. Les articles de PMC avec PMID sont localisables sur l'interface de PMC par API. Plusieurs outils permettent d'associer un DOI à une localisation. Le plus utilisé est doi.org. La localisation se présente sous la forme d'un lien vers une page web ou vers une API. L'outil Crossref.org[20] interrogeable par programme, ou en ligne, associe à un identifiant DOI les liens vers les différentes sources et leur format (texte, XML, etc.). Il est maintenu par une

20 https://search.crossref.org/

association à but non lucratif dont l'objectif est de pérenniser et uniformiser les citations académiques entre éditeurs. Les éditeurs payent un abonnement pour maintenir le service en ligne, et ce sont les éditeurs qui nourrissent la base. Dans certains cas, les outils fournissent également la licence associée, ce qui permet d'éviter de tenter de télécharger des articles non OA.
La localisation des articles souhaités est le sixième verrou de l'automatisation de la conception de corpus.

Le téléchargement

Le choix du mode de téléchargement dépend du format souhaité en fonction de la tâche de fouille de textes et de la qualité attendue et en fonction des traitements dont l'équipe dispose déjà pour télécharger et traiter les formats. Les formats les plus courants sont pdf, html et xml. A chacun de ces formats génériques correspond une grande variété d'implémentations qui a un impact majeur sur la conception du corpus. Ces formats sont détaillés ci-dessous.

Le téléchargement par URL

Quand la localisation de l'article est l'URL d'une page web dite d'atterrissage, il s'agit d'une page conçue pour le lecteur humain, pas pour un logiciel de conception de corpus. Plusieurs cas se présentent, la page web propose

- Directement l'article complet au format pdf
- Directement l'article complet au format html
- Diverses informations composites dont l'article complet au format html
- Diverses informations composites dont des boutons (liens) vers l'article complet aux formats html ou au format pdf.

Dans le cas (1) format pdf

Si le format souhaité est html et non pdf, il faut connaître la façon de générer automatiquement l'URL d'accès à l'article au format html pour la source considérée pour y accéder de manière automatisée, par exemple en concaténant le DOI à l'URL. Cette génération n'est pas toujours possible.

Dans des cas de plus en plus nombreux, seul le pdf est accessible, dont malheureusement la majorité des publications en Open Access, bibliothèques des organismes de recherche (congrès, arXiV, HAL, etc.) et cette proportion grandit.

Dans le cas (2) format html

Si le format souhaité est pdf et non pas html, il faut développer un analyseur pour la page du journal qui permette d'isoler l'URL derrière le bouton de téléchargement du pdf. Cet analyseur est coûteux et hasardeux à développer. Les formats de page évoluent très rapidement et les pages contiennent de très nombreux liens. Les mentions de téléchargements sont très variables et difficiles à isoler automatiquement (pdf download, PDF, view pdf).

Si le format souhaité est bien html, l'article entier peut être très difficile à obtenir. De plus en plus de sites n'affichent pas l'intégralité de la page, mais déclenchent son affichage par partie en fonction de l'ascenseur, une hypothèse étant la lutte contre la fouille de textes. C'est assez étonnant sachant que par ailleurs le connecteur n'a accès à la source que licitement. Dans ce cas, il faudrait mimer par programme l'appel à l'affichage des sous-parties pour pouvoir les télécharger. C'est un frein considérable à la conception de corpus pour des coûts raisonnables.

Dans le cas (3) page composite dont html

Il faut développer un analyseur de la page web pour isoler la partie qui contient l'article. Comme dans le cas précédent, cet analyseur est coûteux et hasardeux à développer. Les formats de page évoluent très rapidement et les pages contiennent de très nombreuses sous-sections encodées contenant du JavaScript et difficiles à interpréter.

Dans le cas (4) page web avec boutons de téléchargement

Il faut développer un analyseur pour la page du journal qui permette d'isoler l'URL derrière le bouton d'accès au document au format souhaité, html ou pdf, comme dans le cas (2). Dans le domaine des sciences de la vie, nos expériences montrent qu'environ un tiers des articles souhaités pour un corpus ne sont accessibles que par un lien vers une page web (URL) et répartis dans un grand nombre de journaux. Le coût de développement est en général prohibitif et le corpus final se limite à quelques journaux ciblés dont l'importance est critique à l'exclusion des autres. C'est en particulier le cas des journaux édités par des associations ou société savantes qui n'ont pas développé d'API. Cette situation favorise la visibilité des informations publiées par les grandes plateformes au détriment des petits éditeurs.

Le téléchargement par API

L'accès par API est proposé par la plateforme pour donner accès à des articles par programme. En principe, l'API est simple d'utilisation, dans la réalité, plusieurs écueils freinent son utilisation :

- Il faut développer un programme qui interroge l'API dans le format qui convient aux besoins. Il faut trouver la documentation, la comprendre, développer et tester le connecteur.
- La plateforme peut retenir toutes les informations sur les requêtes envoyées à l'API et l'identité du requêteur (numéro IP de la machine, clef d'accès). Dans le cas de recherche sensibles, il est recommandé d'utiliser un autre moyen.
- Les plateformes imposent en général des limites sur le nombre d'articles téléchargeables dans un temps donné. Cette information est souvent très difficile à trouver et même pas toujours accessible. Le non-respect de la limite entraîne un blocage pour toutes les machines derrière la même IP ce qui peut avoir des conséquences imprévisibles et importantes pour les collègues.
- Les API ne renvoient pas toujours exactement les réponses demandées sans explication, ni signalement d'erreur. Par exemple, l'envoi d'une liste de DOI d'articles disponibles sur une plateforme ne permet pas toujours de récupérer la totalité des articles.

Le développement de connecteurs permettant de télécharger les articles au format souhaité est le septième verrou de l'automatisation de la conception de corpus et un des plus

importants, avec celui de la conversion de format. Les connecteurs décrits ci-dessus ont un fonctionnement imprévisible pour les raisons détaillées. Il faut toujours vérifier après téléchargement que les réponses sont conformes. Les quantités effectivement téléchargées peuvent être assez différentes de celles attendues. Il faut pouvoir les caractériser (période, journal) pour indiquer à l'utilisateur les informations absentes.

Le téléchargement effectif des articles au format souhaité et la caractérisation des documents manqués est le huitième verrou de l'automatisation de la conception de corpus.

Le traitement du format

Le traitement du format pdf

Le format pdf est plus ou moins facile à convertir en texte. Si le pdf est un scan d'image, il faut utiliser un logiciel d'OCR avec des résultats variables selon la qualité de la photo et l'organisation des pavés de texte dans l'image. Notre expérience est que la qualité est rarement suffisante pour faire ensuite plus que du comptage de mots. Le découpage en sections reste compliqué par exemple.

Si le fichier contient du texte, la récupération du texte correct dépend de nombreux paramètres. Notre expérience est que de nombreux éléments perturbent le fil du texte : multi-colonnes, l'insertion d'éléments extérieurs (pieds de pages, en tête, figures et légendes en travers des colonnes). Les notations particulières, (ex. caractères grecs, indices, etc.) sont rarement correctement repris. Il existe de très nombreux convertisseurs, mais aucun ne donne complètement satisfaction pour des phénomènes pourtant fréquents. Si l'objectif de la fouille de textes est l'extraction d'information basée sur une analyse profonde, il est préférable d'éviter ce format

Le traitement du format html

Le format html est un format destiné à l'affichage. Il contient néanmoins beaucoup d'informations de structure (titres, sous-titres, tableau, légendes, etc.), et typographiques (italique pour les noms de gènes et d'espèce) très utiles au traitement de fouille de textes pour le choix des sections etc. Le schéma n'est pas standard, il faut un convertisseur html -> texte pour chaque journal. Certaines plateformes utilisent le même format html pour tous les journaux (ex. Springer, Elsevier) ou pas (ex. Wiley). La présence de code JavaScript complexifie de plus en plus cette conversion mais le résultat obtenu est de meilleure qualité que la conversion de pdf en texte. Il n'existe pas d'outils clef en main pour le réaliser. Certaines entreprises proposent ce service.

Le traitement du format xml

Le format xml est le format préférable pour la fouille de textes, mais le terme xml recouvre une très grande variété de schémas différents et utilisés de façon combinée. Ces formats sont très rarement bien documentés et il faut comme pour l'html prévoir de développer un convertisseur adapté pour chaque source. Ces formats évoluent rapidement et il faut faire évoluer le convertisseur au cours du temps en cas de mise à jour du corpus.

Le développement de convertisseurs permettant de transformer les articles au format pdf, html ou xml en format texte est le neuvième verrou de l'automatisation de la conception de corpus et un des plus importants.

5.4 Besoins

5.4.1 Information sur les accès aux sources

En réponse au premier verrou de l'automatisation de la conception de corpus : "Lister les sources pertinentes et accessibles", le concepteur de corpus doit avoir accès à une information standardisée et lisible par une machine sur les sources (journal, acte, monographie) licitement accessibles compte tenu de son appartenance à un organisme de recherche donné. Il n'y a pas de verrou technique ou légal à ce besoin.

En réponse au deuxième verrou de l'automatisation de la conception de corpus : "la connaissance anticipée des connecteurs aux sources", le concepteur de corpus doit avoir accès à une information standardisée et lisible par une machine sur les possibilités mise à disposition par les plateformes, de type d'accès à chaque source : API, lien URL, autre.

5.4.2 Licences et abonnements

En réponse au troisième verrou sur la connaissance des articles licitement accessibles par source, le concepteur de corpus doit avoir accès à une information standardisée et lisible par une machine sur le droit de télécharger licitement chaque document en fonction de son organisme de recherche d'appartenance, quelle que soit sa date de publication, sa rubrique ou sa plateforme.

5.4.3 Documentation des formats

En réponse au quatrième verrou sur la connaissance des formats proposés pour les articles licitement accessibles, le concepteur de corpus doit avoir accès à une information standardisée et lisible par une machine sur les types de format et sur leur sémantique dans le cas du xml.

5.4.4 Centralisation et standardisation des métadonnées bibliographiques et d'accès

Le dédoublonnage des articles est avec le dédoublonnage des sources, le cinquième verrou de l'automatisation de la conception de corpus. Il est nécessaire que les sources et documents soient identifiés par un identifiant unique de type ISSN et DOI et systématiquement attaché à chaque document téléchargeable au même titre que les autres données bibliographiques. Les variations typographiques dans les titres et noms d'auteurs doivent être évitées. Les données bibliographiques doivent être uniformisées et centralisées de telle sorte qu'il soit aisé pour un document donné de lister automatiquement l'ensemble des sources et plateformes.

Pour répondre au sixième verrou sur la localisation des articles, les métadonnées doivent inclure dans un format lisible par machine les adresses du document et son mode d'accès. Les outils actuels renvoient des adresses souvent inexactes.

5.4.5 Uniformisation et centralisation des accès aux documents

En réponse au septième verrou sur le développement de connecteurs permettant de télécharger les articles au format souhaité, il serait souhaitable de mettre en place des mécanismes de mutualisation des connecteurs développés par les concepteurs de corpus pour télécharger les documents.

Le développement de dépôts (type Panist) ou d'interfaces transparentes pour l'accès aux documents sont un pas de plus pour répondre à l'énorme diversité des sources et lever le sixième verrou sur la localisation et le septième verrou sur le téléchargement.

5.4.6 Confidentialité et qualité de l'accès aux documents

Les concepteurs de corpus recherchent la confidentialité de leurs accès aux plateformes. La confidentialité passe par le droit d'utiliser les modes de téléchargement qu'ils souhaitent (dans le respect de l'intégrité des systèmes) sans que soient imposés des API. Il n'y a pas de raison technique d'empêcher les URL qui de façon similaire aux API, peuvent être configurées par les fournisseurs pour empêcher le piratage. Aucun moyen technique ne garantit la confidentialité, seul un engagement formel des éditeurs peut le garantir.

La qualité des téléchargements doit être garantie en lien avec le huitième verrou. Il est nécessaire que les plateformes fournisseuses de contenu mettent en place les mécanismes nécessaires pour s'assurer que les réponses aux requêtes sont conformes et dans le cas contraire, permette au concepteur de corpus de communiquer une erreur selon un mécanisme simple et efficace.

5.4.7 Documentation et standardisation des formats

En lien avec le neuvième verrou, la standardisation des formats xml dans un format cible unique, par exemple TEI, est un enjeu majeur de l'utilisation de la fouille de textes. Aujourd'hui, la conversion à partir de milliers de sources dans des formats différents et parfois non conforme est un frein déterminant, comme montré par le projet Istex. Heureusement on assiste à une adoption très progressive de standards : JATS (EPMC, Springer OA), TEI (ISTEX).

5.4.8 Partage

Les besoins décrits ici nécessitent pour une part une organisation, une standardisation et des changements de pratique. Dans l'intervalle, la possibilité de partager les corpus si coûteusement acquis de gré à gré ou travers des dépôts doit être autorisée et organisée. Il est contre-productif qu'une si grande partie des projets de fouille de textes soit consommée en ingénierie de corpus non réutilisable, ou inversement que la recherche se prive de l'accès à l'information par manque de rationalisation du partage de corpus. Les mécanismes tels que la fédération des identités permettent aujourd'hui techniquement à un dépôt de vérifier la licité des accès à des plateformes en fonction de l'identité, c'est-à-dire concrètement de vérifier qu'une personne donnée a accès à un document en fonction de la licence et des abonnements souscrits.

Conclusion

Nous avons dans ce document fait un tour d'horizon de l'ensemble des besoins potentiels de la communauté de recherche à laquelle s'adresse notre étude sur les potentialités d'une e-infrastructure de fouille de textes à une échelle nationale.

Nous avons pu mettre en évidence que ces besoins étaient extrêmement divers, à l'image des nombreuses potentialités de la fouille de textes et donc à son rôle majeur dans une accélération de l'innovation et de la recherche.

Nous avons pu constater en particulier que ces besoins étaient différents suivant les types d'utilisateurs, allant du chercheur en TAL qui privilégie un accès à des ressources (corpus de notices bibliographiques par exemple ou ressources sémantiques) ainsi qu'à des outils de traitement de ces données textuelles, au décideur qui a besoin d'une assistance au traitement de ses données et une mise en forme explicite de celles-ci, en passant par le chercheur non spécialiste de la fouille de textes qui aura besoin à la fois d'un accompagnement sur des choix de ressources et d'outils mais aussi une aide à la constitution des chaînes de traitement tout en conservant un regard sur les résultats et leur interprétation.

Les besoins se différencient également en fonction des communautés de recherche dont certaines ont des demandes toutes particulières comme l'anonymisation de certains textes dans le domaine médical.

Enfin des besoins plus précis peuvent être liés aux résultats attendus d'un traitement de fouille de textes dans un objectif précis : ainsi nous avons décrit plus spécifiquement tous les besoins liés à l'activité d'extraction de corpus dans un fonds bibliographique.

Le dépouillement du questionnaire établi afin de consolider les attentes de la communauté de recherche nous a montré que l'étude Visa TM répondait à une attente et suscitait un intérêt certain. Les besoins exprimés répondent plutôt bien à ceux que nous avions imaginés, même si des non spécialistes peuvent encore s'interroger sur les apports de la fouille de textes dans leur quotidien ou se trouver peu armés pour s'attaquer seuls à ce type de traitement. Il apparaît donc que la fourniture de fouille de textes ne peut s'envisager sans un accompagnement sur les services proposés et que la mise en place d'une infrastructure nationale est un moyen de partager des expériences et compétences autour de technologies complexes mais qui nécessite aussi une implication forte et collaborative des différents acteurs partie prenante que nous avons détaillés dans le document « Acteurs et organisation ».

Index des figures

Figure 1. Réflexion sur les besoins possibles des chercheurs TAL 26
Figure 2. Réflexion sur les besoins possibles des chercheurs non TAL 27
Figure 3. Réflexion sur les besoins possibles des décideurs 28